HIGH-RESOLUTION CENSUS DATA: SIMPLE WAYS TO MAKE THEM USEFUL

Itzhak Benenson* and Itzhak Omer

Dept of Geography and Human Environment, Environmental Simulation Laboratory, University Tel Aviv, Ramat-Aviv, Tel-Aviv, 69978, Israel
Email: bennya@post.tau.ac.il, omery@post.tau.ac.il.

ABSTRACT

Recent population censuses have brought about arrays of high-resolution explicitly geo-referenced socio-economic data stored in the framework of Geographic Information Systems. Geography and social science are not prepared for these new urban databases, and this paper considers their potential for investigating residential distribution, based on the data of the 1995 Israeli Census of Population and Households. We focus on the methodological problems: understanding the phenomena, formal analysis, and statistical inference. The methods for mapping high-resolution data, establishing spatial relationships between them, analyzing neighborhood structure, and exploring the significance of the results are proposed and illustrated by examples of the cities of Tel-Aviv (pop. 350,000) and Ashdod (pop. 100,000).

Keywords: Population census, GIS, Thematic mapping, Spatial patterns, Spatial statistics

1. INTRODUCTION

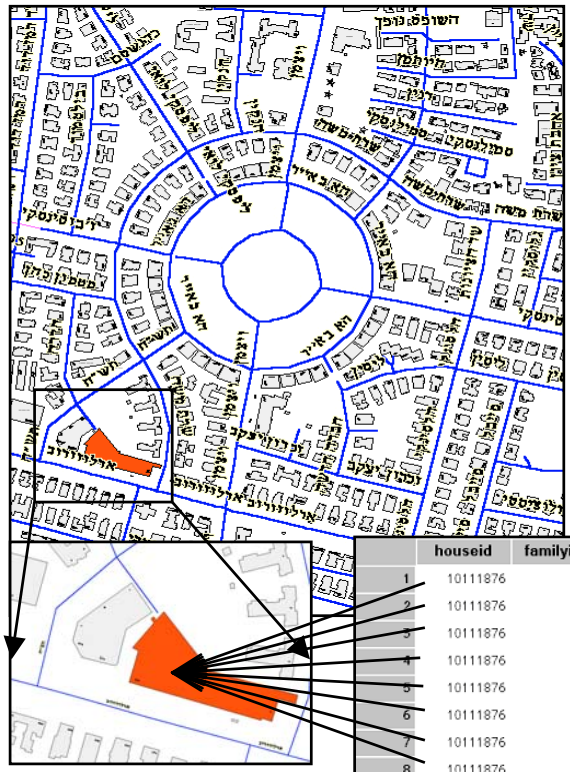
The new generation of European population censuses has brought about a revolution in the analysis of urban population data – long awaited in the field of social science (National Statistics, 2002). Before mid 1990s, population data were available at an aggregate levels only, for social, economic, ethnic, or other groups over the units of predetermined administrative partitions of a city. In the 21st century, demographic information is no longer aggregated; instead, individual and family records include the precise geographic location, and are stored in the framework of a high-resolution Geographic Information System (GIS). The layers of the high-resolution GIS include roads, open spaces and topography, which makes it possible to characterize explicitly the spatial distribution of millions of individuals at a resolution of separate buildings, an idea that seemed fanciful just a decade ago.

The 1995 Israeli Census of Population and Households (ICBS, 2000) follows the new census framework and is a remarkable example of the census revolution. For each settlement with a population of over 2,000, two layers of census GIS, representing the streets and buildings bases, are constructed. Within the GIS framework, the records containing information on individuals and families are related to the building in which they reside. Consequently, each person living in an Israeli settlement, for three months prior to the census, is precisely *geo-referenced* (Figure 1). It is worth noting that, according to different estimates, the reliability of the census data is very high.

Data of a new kind creates a new situation. As recently described by M. Batty: “Small scale studies of individual locations and individuals within urban location have never been able to develop theory of sufficient generality … All this is changing. Quite suddenly, so it appears, a new kind of fine-scale geography is beginning to emerge from data which are sufficiently intensive to detect detailed patterns and morphologies, but also sufficiently extensive to enable these patterns to be generalized to entire metropolitan areas.” (Batty, 2000).

Following this optimistic claim, the detached observer might think that social scientists would leap on these enormous data sets, but surprisingly, the reality is different. As it turns out, geography and social science are not prepared for these new urban databases - general concepts of urban structure are only loosely linked to individual data and the methods of analysis of high-resolution socio-spatial data are underdeveloped. The

reason is evident - the quantitative analysis of high-resolution spatial data was never in the mainstream of local-scale intensive socio-geographic studies (Sayer, 1985), just because they were not available ever before.



	houseid	familyid	v2	v3	v4	v5	v6	v7	v48	v49	v50	v51	v52
1	10111876	3	Tel-Aviv	Tel	5000	9	2	1	Stud	9-10	Did	Prim	Doe
2	10111876	28	Tel-Aviv	Tel	5000	9	2	6	Nev	0(did	Stud	Nev	Unk
3	10111876	28	Tel-Aviv	Tel	5000	9	2	6	Stud	11-1	Did	Post	Live
4	10111876	28	Tel-Aviv	Tel	5000	9	2	6	Stud	11-1	Did	Prim	Live
5	10111876	28	Tel-Aviv	Tel	5000	9	2	6	Stud	1-4	Did	Did	Doe
6	10111876	32	Tel-Aviv	Tel	5000	9	3	5	Stud	5-8	Did	Did	Doe
7	10111876	32	Tel-Aviv	Tel	5000	9	3	5	Stud	5-8	Did	Prim	Unk
8	10111876	35	Tel-Aviv	Tel	5000	3	4	4	Stud	11-1	Stud	Matr	Unk
9	10111877	49	Tel-Aviv	Tel	5000	5	1	2	Stud	11-1	Did	Matr	Unk
10	10111877	57	Tel-Aviv	Tel	5000	9	3	7	Stud	9-10	Stud	Did	Live
11	10111877	57	Tel-Aviv	Tel	5000	9	3	7	Stud	9-10	Stud	Did	Live
12	10111877	57	Tel-Aviv	Tel	5000	9	3	7	Stud	11-1	Stud	Did	Live
13	10111877	57	Tel-Aviv	Tel	5000	9	3	7	Stud	11-1	Stud	Sec	Live

Figure 1. The structure of the GIS of the 1995 Israeli Census of Population and Households

In this paper, we illustrate the potential of high-resolution socio-spatial data for investigating urban residential distribution. We focus on methodological problems only and are concerned with three aspects of data investigation:

- *The initial grasp of phenomena*: Thematic mapping of high-resolution census data;
- *Analysis of spatial patterns:* Formal description and measurement of urban residential distributions;
- *Inference*: Statistical inference regarding residential patterns.

We illustrate our conclusions on the basis of geo-referenced census data on population of Tel Aviv (pop. 350,000 in 1995) and Ashdod (pop. 100,000 in 1995), available for supervised study at the Israeli Bureau of Statistics.

2. THEMATIC MAPPING OF HIGH-RESOLUTION CENSUS DATA

Thematic mapping is the necessary first step toward understanding and explaining any geographic phenomena. The typical situation a reasearcher faces when utilising standard mapping tools is presented in Figure 2. The map of the population distribution presented there, directly follows GIS logic - census data

on householders are, by definition, linked to dwelling units – and because of that reason the map is clumsy and not very informative regarding the spatial variation of the income level. It is so, because dwelling units are *always* sparsely distributed. The map in Figure 2 reflects the standard situation - and applying standard GIS techniques, we easily ascertain that in Israeli cities, even in completely built-up areas, buildings cover, at most, 40 percent of the urban area. Consequently, at a distant zoom, the non-informative part of the map dominates and obscures the meaningful population information. Closer zooms, at resolution of several buildings, are not helpful here - there are just too many trees in a forest of urban buildings!

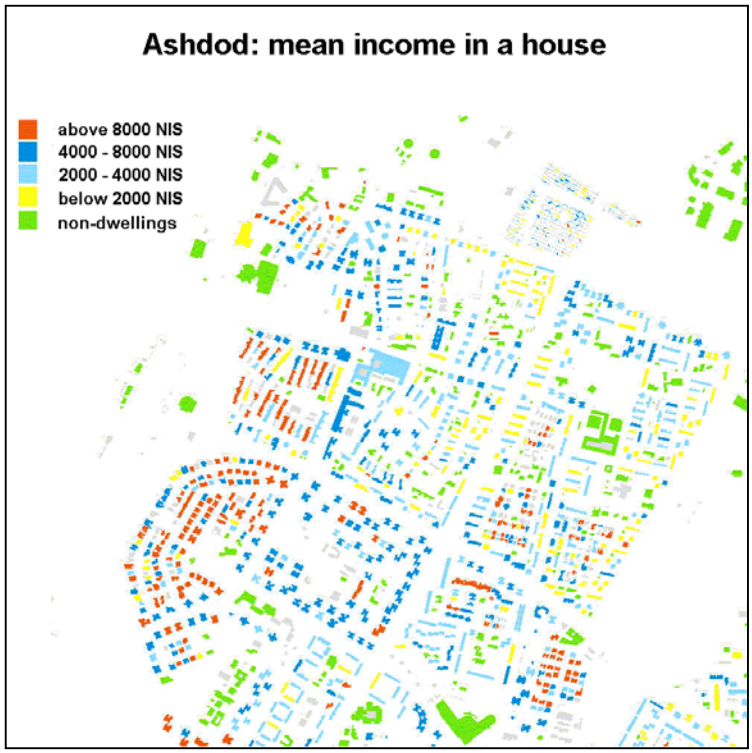


Figure 2. Standard thematic map, constructed on the basis of individual census data

The relation of the high-resolution census data to discontinuous and sparse coverage of buildings triggers one more problem – one of 'neighborhood' definition - vital for studying the social composition of the individual's home area (Aitken & Prosser 1990; Pacione, 1983). Standard approaches consider as neighbors individuals located in the same unit of the partition that defines the aggregation of the data (statistical areas, for example). The detailed information in Figure 2 is, intuitively, more adequate for estimating neighborhood relationships, but how we should translate this potential into an analytical procedure?

To resolve the problems entailed by the sparseness of high-resolution census data, we have proposed employing Voronoi polygons, constructed on the basis of building coverage (Benenson, Omer & Hetna, 2002; Omer & Benenson, 2002). To remind the reader, each polygon of a Voronoi partition corresponds to the building and includes the building and the surrounding area, given the building is the nearest for every point in this area (Halls, Bulling, White, Garland & Harris, 2001). Two buildings are considered neighbors if their Voronoi polygons have a common boundary. In addition, a common boundary between two Voronoi polygons usually means the corresponding buildings are visible, at least partially, to each other.

Standard GIS tools make it possible to combine the adjacency-based definition of the neighboring relationships with metric relations and, in addition, account for infrastructure elements, such as roads and open spaces. In Benenson et al, (2002), we define buildings as neighboring based on all three characteristics. Namely, two buildings **A** and **B** are neighbors if:

1. Voronoi polygons of **A** and **B** are adjacent
2. The distance between the centroids of **A** and **B** is less than 150 m.
3. **A** and **B** are not considered as neighbors if they are located on different sides of a main street or there is a public place between them.

Figure 3 illustrates this definition

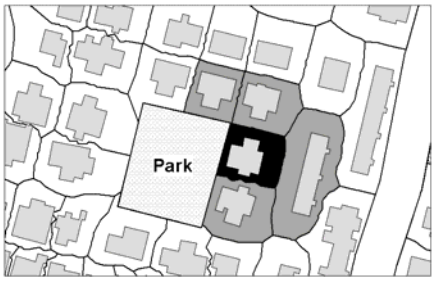


Figure 3. Voronoi partition of the plane and neighboring relations defined on its basis

The use of Voronoi polygons enables comparison of the new high-resolution census data to a standard aggregate presentation. This comparison demonstrates important differences. The statistical area outlined by a dashed red line in Figure 4 (constructed for Yaffo residential area of Tel Aviv, pop. 30,000) is an example: the residents of two spatially separated groups of houses, almost exclusively populated by Arabs or Jews, appear to be living in an ethnically mixed statistical area. Consequently, the level of segregation of the Arab and Jewish population in Yaffo is much higher than one can conclude on the basis of aggregated data.

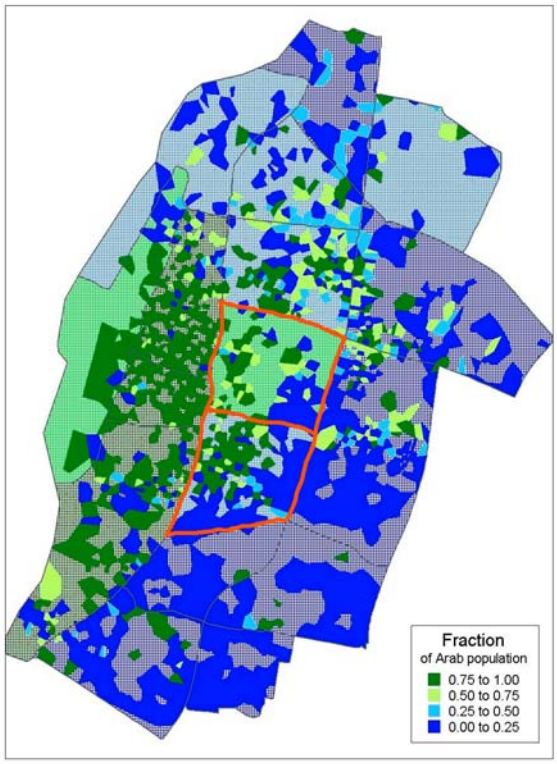
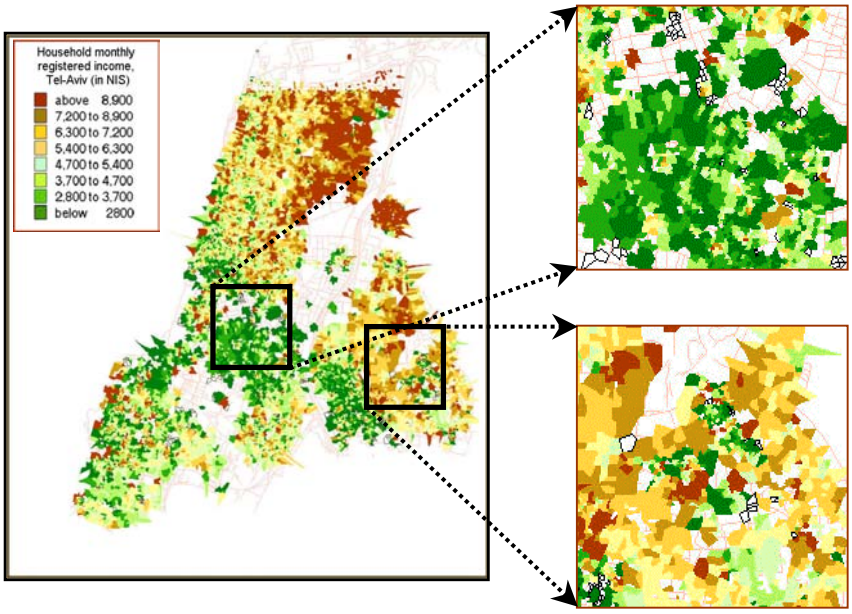
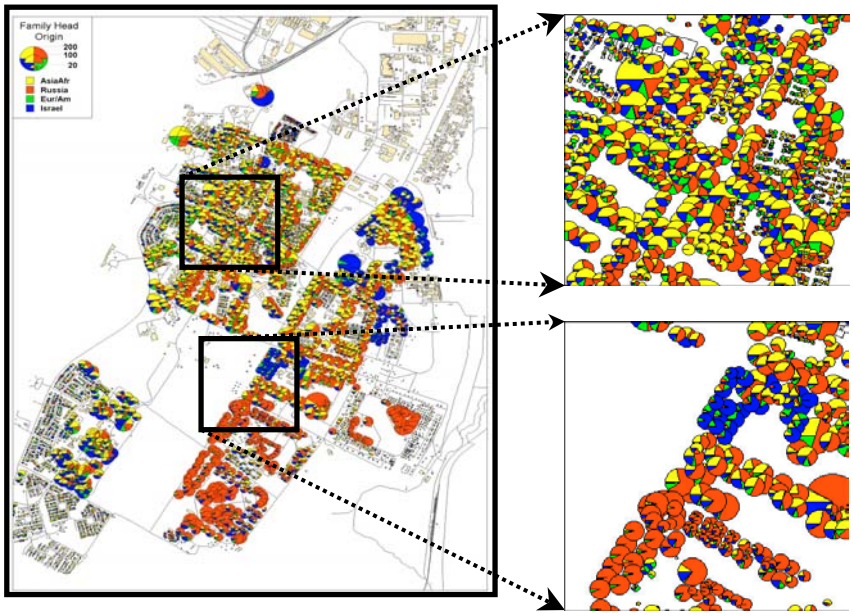


Figure 4. Thematic maps of ethnic distribution in Yaffo, Tel-Aviv, at resolution of separate buildings and statistic areas

Choropletic mapping on the basis of Voronoi coverage provides the basic view on high-resolution population data and indicates that other ways of expanding the presentation beyond dwelling units can be useful. Besides Voronoi polygons, this expansion can be done by means of another standard GIS tool - pie charts, which are especially useful when several characteristics should be presented simultaneously (see examples in Figure 5).



a



b

Figure 5. Examples of high-resolution thematic maps: (a) Salaried income in Tel-Aviv presented on the basis of Voronoi coverage; (b) Pie-chart map of the family head origin in Ashdod.

3. FORMAL DESCRIPTION AND MESUREMENT OF URBAN RESIDENTIAL DISTRIBUTION

The maps in Figures 4 and 5 clearly demonstrate that the local structure of the residential distribution essentially varies over urban space. To make use of high-resolution data, we have to estimate this variation. The estimate is important in two respects: Firstly, in order to understand the validity of standard aggregate partitions used until recently. In Israel, for example, the standard partition of population data is based on 'statistical areas,' established in the 1950s, and containing, at that time, about 3,000 individuals. Secondly, the social composition is one of the main factors of residential choice and neighborhood perception (Aitken & Prosser 1990; Pacione, 1983). In order to validate any theory of urban population formation, we have to estimate it on the basis of the characteristics of a resident and his/her neighbors.

Most of the socio-geographic tools offer non-spatial estimates of residential patterns. The standard indices of 'dissimilarity', 'evenness', 'clustering', 'exposure', are all the results of averaging over the entire urban area (Massey & Denton, 1988). Social science was never satisfied with these indices, because they do not reflect the spatial variability of residential distribution. In addition, research on the 'Modifiable Aerial Unit Problem' clearly demonstrates that the conclusions reached on the basis of aggregate datasets change significantly when the same data are considered over different partitions (Openshaw, 1984).

With high-resolution data, differences between the individual and his/her neighbors are the focus of attention (Benenson & Omer, 2002). The characterization of the residential pattern as a whole is a by-product of this view: for example, the residential distribution is homogeneous in an area where the differences between the individual and neighbors are low at each location. Based on Voronoi polygons we can define neighboring buildings. But how can we define neighbors influencing, for example, habitat choice?

Our approach to defining neighboring relationships (Benenson & Omer, 2002) is based on placing the individual within a *hierarchy of neighborhood*, which begins with the building itself (neighbors of the zero order), where the closest neighbors are located. The householders, who populate adjacent buildings, are considered as neighbors of the first-order. Second-order neighbors are those populating buildings closest to the neighbors of the first-order, and so on. Figure 6 presents three levels of the proposed hierarchy for the case of neighborhood relations between buildings defined via Voronoi polygons.

The next step is to estimate the structure of the neighborhoods of different orders. The relationships between spatially located data are studied by geostatistics (Cressie, 1993). The basic assumption of geostatistics is the stationarity of spatial distribution, which, informally, means that the same spatial relationship (spatial autocorrelation) is held over the entire area studied. Until now, the stationarity of urban residential distribution could not be validated. It became feasible with the new generation of census data, but is still awaiting its application. For now, the common-sense arguments are against the assumption of stationarity, because the residential distribution within an area *depends on its history*, and the structure of the neighborhoods in, say, rapidly developing areas differs from that in a steadily developing one.

To avoid the problem of non-stationarity, the properties of the spatial distribution should be estimated at every point anew. The series of 'local indices of spatial association,' analogous to the *local* average, variance, autocorrelation and coefficient of variance (CV) were recently proposed in geography and ecology (Anselin, 1995; Wu & Sui, 2001). The above definitions of neighborhood relationships and the neighborhood hierarchy provide the background for local calculations, and in recently published articles, (Benenson & Omer, 2002; Omer & Benenson, 2002) we have discussed the applications of indices of spatial association in length; here we present the simplest form of four of them:

Index G* of Getis (Anselin, 1995) is analogous to the moving average:

$$G^*_{i,k} = \Sigma_{j \in Uk(i)} w_{ij}(f_j - \langle f \rangle)/(W_i s) \qquad (1)$$

Index K of Geary (Anselin, 1995) is analogous to the standard deviation:

$$K_{i,k} = \Sigma_{j \in Uk(i)} w_{ij}|f_j - f_i|/(W_i s) \qquad (2)$$

Index I of Moran (Anselin, 1995) is analogous to autocorrelation:

$$I_{i,k} = (f_i - \langle f \rangle)\Sigma_{j \in Uk(i)} w_{ij}(f_j - \langle f \rangle)/(W_i s^2) \qquad (3)$$

Index L of Lacunarity (Wu, Sui, 2001) is analogous to squared CV:

$$L_{i,k} = Var_{Uk(i)}/(Mean_{Uk(i)})^2 \qquad (4)$$

In formula (1) – (4), f_i represents the value of trait f in location i (for example a building); $U_k(i)$ represents the neighborhood of location i of the order k (k = 0, 1, …), and the weight w_{ij} defines the influence of the neighbor j on i. $Mean_{Uk(i)} = \Sigma_{j \in Uk(i)} f_j$ and $Var_{Uk(i)} = \Sigma_{j \in Uk(i)}(f_j - Mean_{Uk(i)})^2$ are the local mean and variance of trait f over $U_k(i)$ and $W_i = \Sigma_{j \in Uk(i)} w_{ij}$, $\langle f \rangle$ represents the average and s^2 the variance of f over the entire urban space.

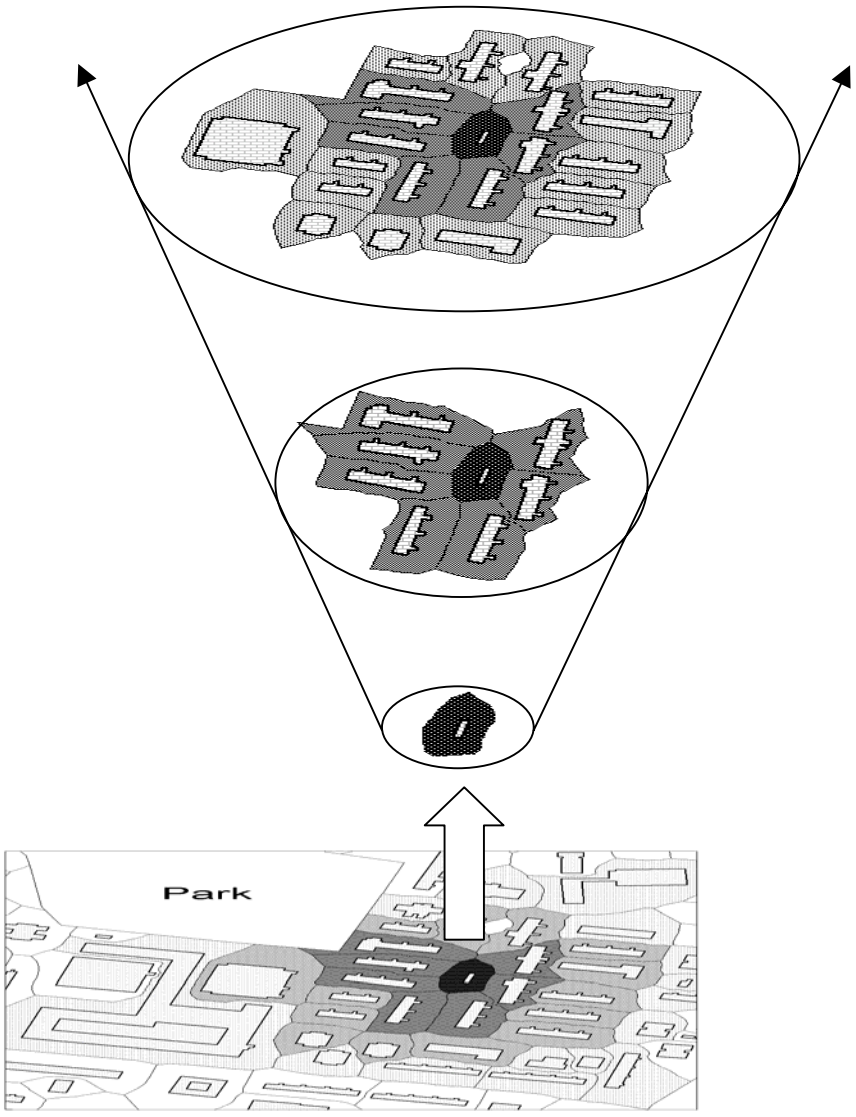
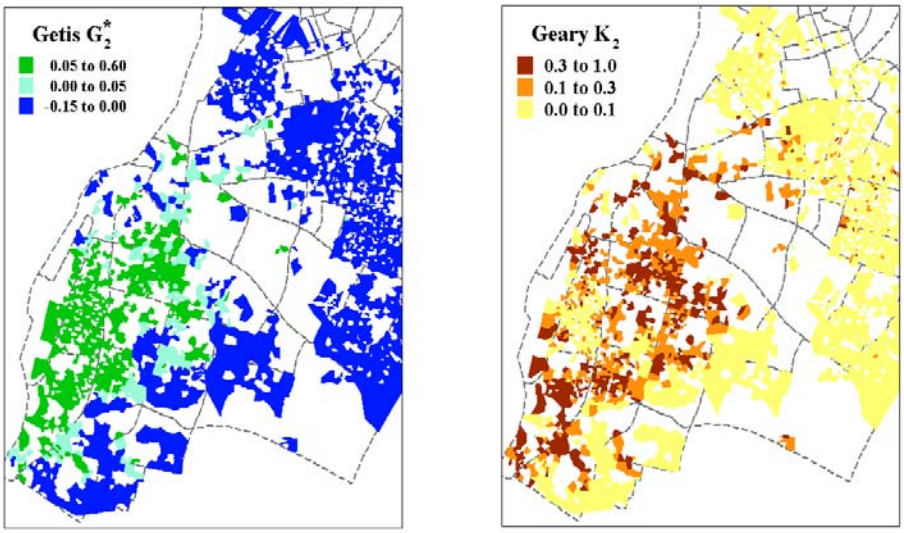


Figure 6. Hierarchy of the neighborhoods, constructed on the basis of the Voronoi coverage.

In application, the values of the local indices are calculated for every building in the urban area over neighborhoods of increasing sizes. Figures 7a and 7b show maps of the Getis and Geary indices for neighborhoods of the second-order (which correspond to the 'home area' of the householder – Omer &

Benenson, 2002) calculated to estimate the ethnic residential pattern in Yaffo. According to Figure 7, one can easily see where in Yaffo an individual feels him/herself to be in a homogeneous Jewish area (marked in blue in Figure 7a and yellow in Figure 7b) or homogeneous Arab area (marked in green in Figure 7a and yellow in Figure 7b) and where his/her neighbors are of varying ethnicity (marked in red in Figure 7b). Detailed discussion of the local indices, their comparison with standard aggregate indices and their application in other aspects of the spatial-social segregation in Yaffo can be found in Benenson & Omer (2002) and Omer & Benenson, (2002).



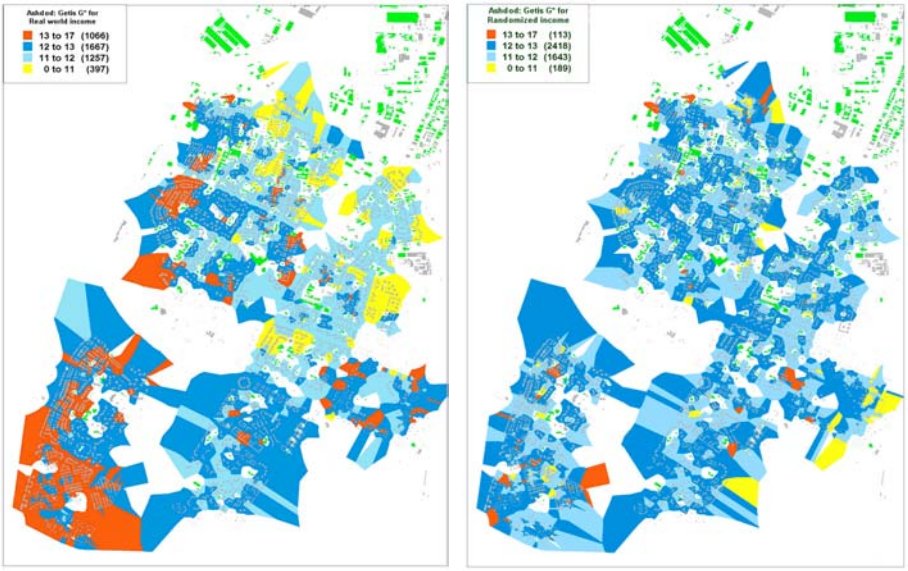
a b

Figure 7. Thematic maps of the local indices of spatial association, constructed for Yaffo ethnic distribution on the basis of neighborhoods of the second-order (a) Getis $G^*_{i,2}$ index; (b) Geary $K_{i,2}$ index

4. Statistical inference regarding residential patterns

High-resolution maps reveal extended urban areas, where residential patterns essentially vary (Figure 5). How can we distinguish between the areas where these patterns are accidental and those, where they demonstrate some order? Available analytical methods provide confidence intervals for local indices of spatial association on the basis of maximum likelihood (Anselin, 1995) and their power is low. For distributions in Figure 5 and indices (1) – (4) above, almost all the neighborhood patterns do not differ from the random ones at 95% significance level. A "bootstrap" idea (Davidson & Hinkley, 1997) works better here. The bootstrap approach is conceptually simple and with regard to residential distribution is as follows. First, to construct the spectrum of possible random distribution, redistribute individual data randomly over the same houses many, say 1000, times and each time calculate indices (1) – (4) at each location. Second, compare the actual value of the index at a certain location with the characteristics of the generated distribution of indices at the same point. If the probability of obtaining the actual value of an index is low then the local structure at a point is non-random.

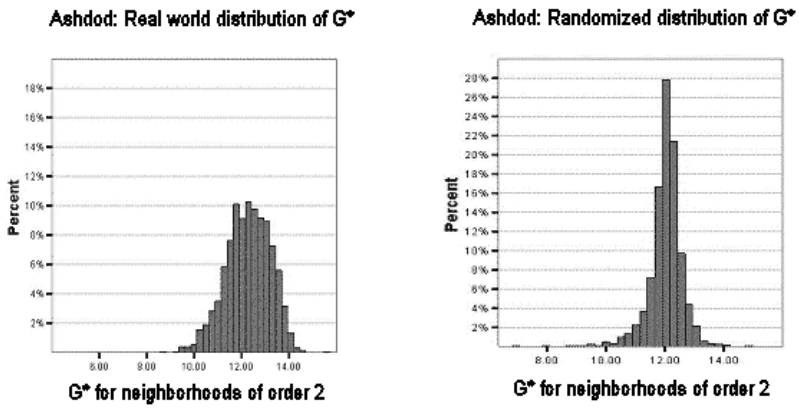
We do not go into the details of the bootstrap approach but illustrate it briefly by the distribution of registered household income in Ashdod. Figure 8a represents actual distribution of the Getis G* index (analogous to a moving average) of the household income in a building over the neighborhoods of the second-order, while Figure 8b represents the same index calculated for the randomized distributions. To obtain the latter, all Ashdod families were randomly redistributed between the buildings, taking into account the number of apartments in each.



a b

Figure 8. The maps of $G^*_{i,2}$ constructed for actual (a) and randomized (b) distribution of salaried income in the city of Ashod.

A visual comparison of the two maps in Figure 8 is sufficient to recognize areas of significantly high and significantly low income in the city. Quantitatively, a general comparison of this kind can be done on the basis of distributions of G* for actual and randomized distributions (Figure 9). The variance of the actual distribution of G* (0.882) is 2.68 times higher than the variance of the G* for randomized one (0.329) and, according to χ^2 test, this difference is significant at $p<0.001$. It is possible to proceed to statistical inference regarding a specific location, but we will not do that here.



a b

Figure 9. Distributions of $G^*_{i,2}$ obtained for actual (a) and randomized (b) distribution of salaried income in the city of Ashod.

5. DISCUSSION

Despite the high potential of individual geo-referenced census data, their usage in social science, including social geography, is still far below expectation. We assert that the reasons are not only due to technical and bureaucratic constraints related to access to these data, but also methodological.

In this paper, we have illustrated the potential of high-resolution data for investigating census data on residential distribution and have suggested operative solutions for the methodological problems that arise with these data. We have presented methods of mapping individual data, establishing spatial relationships between them, analyzing neighborhood structures, and exploring the significance of the obtained results. The approach is applied to social residential distributions, constructed on the basis of the 1995 Israeli population census.

Many other techniques, beside those presented, can be considered within the proposed methodology. For example, neighborhoods can be defined on the basis of the three-dimensional visibility of houses or pedestrian accessibility, randomized redistribution of the individual data can account for their income versus housing price, and so on.

Let us note, in conclusion, that the analysis of high-resolution Israel census data made according to the methodology offered (which will be presented elsewhere), essentially alters our understanding of the urban residential distribution. Namely, the classic models of urban ecology suggest that a city consists of "homogeneous patches separated by mixed boundaries," (Godfrey, 1988). The analysis of the Israeli census data does not confirm this view. The residential distributions in big Israeli cities, just as examples of Tel Aviv and Ashdod (Figures 5 and 8) demonstrate, are characterized by *high heterogeneity* over the majority of residential areas; only over few areas (as in Yaffo, Figure 4) can residential distribution be considered as homogeneous.

6. REFERENCES

Aitken S., & Prosser R. (1990) Residents' spatial knowledge of neighborhood continuity and form, *Geographical Analysis 22*(4), 1990, 301- 325.

Anselin, L. (1995) Local Indicators of Spatial Association – LISA, *Geographical Analysis 27*(2), 93–115.

Batty, M. (2000) The new urban geography of the third dimension, *Environment and Planning B: Planning and Design, 27*(4), 483-484.

Benenson, I., & Omer, I. (2002) Individual-based approach to measuring spatial segregation: Formal representation and the case study. In Schnell, I. (Ed.) *Patterns of Segregation and Desegregation*, Burlington, Ashgate, 11-38.

Benenson I., Omer I., & Hetna E. (2002) Entity-based modeling of urban residential dynamics – the case of Yaffo, Tel Aviv, *Environment and Planning B: Planning and Design*, *29*(4), 491–512.

Cressie, N.A.C. (1993) *Statistics for Spatial Data*, NY: Wiley & Sons.

Davison, A. C., & Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*, Cambridge, UK: University Press.

Halls P J, Bulling M, White P. C. L., Garland L, & Harris S. (2001) Dirichlet neighbors: Revisiting Dirichlet tessellation for neighborhood analysis, *Computers, Environment and Urban Systems 25*, 105-117.

Godfrey B.J. (1988) *Neighborhoods in Transition*, University of California Publications in Geography.
ICBS (Israeli Central Bureau of Statistics). (2000) Socio-economic characteristics of population and households in localities and statistical areas, *Pub. No 8 in the 1995 Census of Population and Housing series,* Jerusalem: State of Israel, Central Bureau of Statistics Publications, 2000.

Massey, D., & Denton, N. (1988) The dimensions of residential segregation, *Social Forces 76*, 1988, 281-315.

Omer, I., & Benenson, I (2002) Investigating fine-scale residential segregation by means of local spatial statistics, *Geographical Research Forum*, *22*, 41-60.

Openshaw, S. (1984) The modifiable aerial unit problem, *Concepts and Techniques in Modern Geography*, 38, Norwich: GeoBooks.

Pacione, M. (1983) The temporal stability of perceived neighborhood areas in Glasgow, *Professional Geographer, 35*(1), 1983, 66-73.

Sayer, A. (1985) The difference that space makes. In Gregory D., Urry J. (Eds.) *Social Relations and Spatial Structure.* London: Macmillan, 45-66.

National Statistics (2002) UK Census 2001. Retrieved January, 10, 2003 from the National Statistics website: http://www.statistics.gov.uk/census2001/default.asp

Wu X. B., & Sui D. Z. (2001) An initial exploration of a lacunarity-based segregation measure, *Environment and Planning B: Planning and Design*, *28*(3), 433 – 446.