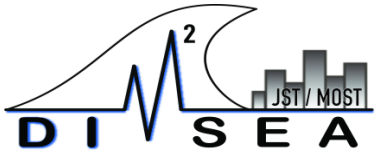
AASDC: An Allocation Algorithm for Data Disaggregation and Synthetic Database Construction

Daniel Felsenstein, Peleg Samuels and Yair Grinberger

WP 02/16

Increasing Urban Resilience to Large Scale Disasters: The Development of a Dynamic Integrated Model for Disaster Management and Socio-Economic Analysis (DIM2SEA)

funded by
JAPAN Science and Technology Agency (JST) and
Ministry of Science, Technology and Space, Israel (MOST)



1. Introduction

This working paper describes a procedure for assigning socio-economic attributes to households and individuals and spatially distributing them to dwelling units. We develop a dedicated allocation algorithm for data disaggregation and the generation of synthetic spatial microdata (AASDC). This is a necessary pre-requisite for any simulation exercise and allows for achieving a detailed representation of household and individual agents. The AASDC automates the allocation process and generates data at a national scale rather for an individual region, city or neighborhood. As data is down-scaled to the level of the household occupying a dwelling unit within a geocoded building, this affords the potential for creating spatial units at multiple level of resolution and for any spatial scale. The approach calls on combining census tract (CT)[1] level data with GIS buildings layers in order to generate synthetic spatial microdata. The resultant synthetic database is both detailed and accurately represents the spatial distribution of buildings, dwelling units, households and inhabitants within the urban area. Each entity in the database (building, household, individual) can be used as an entity for further simulation. The synthetic database thus generates baseline conditions from which a simulation starts.

The paper reviews some of the approaches to generating synthetic spatial microdata (section 2). A short theoretical exposition as to why synthetic micro data should yield unbiased estimators of the true data is presented in section 3. The mechanics of the data disaggregation and allocation processes are formalized generically in section 4. Section 5 provides applies the process to actual data. The generation of synthetic spatial microdata raises questions of reliability and verification of the estimates produced (Hermes and Poulsen 2012). Therefore in section 6 we present an external validation exercise by comparing our synthetic spatial distributions with those derived from an independent survey yielding micro (individual) level data. Section 7 concludes with some implications for future work.

[1] CT's in Israel are uniform administrative spatial unit defined by the Israeli Central Bureau of Statistics (CBS) and called 'statistical areas'. They have relatively homogenous populations of roughly 3,000 persons. Municipalities of over 10,000 population are subdivided into multiple CT's. See below (Section 5) for an example.

2. Literature Review

Simulations invariably call on household and individual level data using spatial units not generally covered by standard sources. Even if the spatial units are standard they virtually never provide a level of disaggregated detail needed for the simulation. Thus a trade off usually exists between the spatial resolution of units and the level of detail on individuals and households. To circumvent this impasse various approaches have been suggested to generate synthetic microdata relating to the population and spatial unit of interest. The literature offers a number of data disaggregation techniques such as population gridding (Linard, Gilbert and Tatum 2010) areal interpolation (Reibel and Bufalino 2005), dasymetric representation (Eicher and Brewer 2001, Mennis 2003) iterative proportional fitting (Beckman, Baggerly and McKay 1996, Pritchard and Miller 2001) and dynamic population modeling (Bhaduri, Bright, Coleman and Urban 2007; Martin, Cockings and Leung 2015) . Much of this work is concerned with disaggregation whereby spatial data collected at one set of areal units is allocated to a different set of units. This transfer can be effected in a variety of ways such as using spatial algorithms, GIS techniques, weighting systems etc (Reibel and Bufalino 2005). Alternatively a 'fusing and matching' approach can be used where local administrative count data can be used to by-pass some of the rigidities of census data (Harper and Mayhew 2012). This microdata can also be used to supplement existing census data although it is not clear to just what extent is can be recombined into non-standard spatial units and whether this exercise can be effected on a national scale.

Another approach is to generate synthetic populations by selecting households and individuals from random samples such that the joint distribution of the critical attributes of interest in the synthetic population, match known aggregate distributions or control totals available from some institutional source such as the national census. The standard technique for estimating joint distributions from a set of control variables is iterative proportional fitting (IFP). This involves generating synthetic households by random draws from a sample and then adjusting them according to the marginal distributions from a census (Beckman et al 1996). However, while this approach may ensure that synthetic households match iteratively determined joint distributions, it lacks a mechanism to ensure that this regularity continues through to the individuals that comprise the households. The resulting synthetic population is therefore inconsistent: it is comprised of households whose distribution matches the marginal distribution and individuals whose distribution does not.

The various broad approaches to generating small area data for microsimulation are presented by Rahman, Harding ,Tanton and Liu (2010). They map the small area estimation landscape arguing that the geographic approach to generating spatially granular microdata is both robust and preferable to the statistical approach. The former is concerned with creating synthetic data for small areas or individuals while the latter uses explicit statistical models suitable for small area estimation using BLUE or Bayesian methods.

Hermes and Polsen (2012) present a typology of approaches for generating synthetic populations for spatial units. They distinguish between 'synthetic reconstruction' methods described above which involve an element of generating data artificially and 'reweighting' methods in which individuals or households from survey data are assigned weights that flag their representativeness in the spatial unit. These weights are sequentially adjusted until the known marginal distribution of the population of the spatial unit matches the weighted survey data. Thus both methods use IPF to ensure that synthetic/survey data correspond to the known marginal distribution of the population.

Our initial work in this area used simple proportional allocation with no marginal adjustments. In Lichter and Felsenstein (2012) socio-economic values of households rae allocated to buildings in proportion to their floor-space. As the following equation illustrates for population size, individual buildings values are calculated by multiplying CT-level densities with buildings-level floor-space:

$$Pop_b = fs_b * \frac{Pop_c}{fs_c}$$

where *Pop* is population size, *fs* is floor-space, *b* is individual building and *c* is census tract.

Floor-space is calculated according to aerial footprint and building height. Floor heights of residential buildings is taken as 5m for non-residential buildings as 7m. These figures are derived by comparing the calculated sum total of floor-space over all buildings by use with total national built floor-space. This proportional allocation process necessarily entails a loss of data due to the division of integers (e.g. population) by fractions (e.g. floor-space). The SQL-based allocation algorithm compensates for this by adjusting the floating point figures rounding threshold for each variable separately. In this manner, the algorithm verifies that CT control totals are met.

At the second stage, each individual is given a unique id tied to a specific building and randomly located within the building. Next, demographic values (e.g. age, disability, workforce participation) are allocated to individuals so that the entire set of residents within a building represents the distribution of socio-economic variables within it. This distribution corresponds to the CT distribution from the previous stage. Under this allocation system the socio-demographic structure of households in multi-unit buildings is homogenous while for single household units it is variable. Due to this inconsistency, current research (see below) uses a more refined down-scaling method based on adjustments to representative marginal distributions grounded in the national census rather than in floor-space area.

3. Theoretical Foundation

The AASDC algorithm generates synthetic data for spatial units of different levels of resolution and geometries. In this section we provide a short exposition to illustrate that synthetic estimators can be unbiased estimators of socio-economic attributes in given spatial units. Large volume synthetic attributes are similarly created by increasing the number of spatial units.

Let us assume a certain geographic polygon, P, which encompasses a set amount of individuals, with Y properties. Individuals create discrete sets of households f, with each such set occupying a specific geographical unit. Entities can be either households or individuals. We further assume that for each geographical polygon certain statistical measurements and distributions are given for each property, as well as detailed housing information.

From each distribution it is possible to generate a data set, which describes the given distribution for any number of individuals greater than one. Assume that within P, real population 'a' is taken from distribution A for a given property of the real population. We further denote α a set of possible populations that could be drawn from A. By definition, $a \in \alpha$.

We now define function $f(x, A)$ which generates a mapping of values from A to a synthetic population of given x entities. We require f to use true random assignment. Function f will in general be termed the "allocation function".

We now propose that $(x_a, A_a) = \alpha_i$, where α_i is a possible population in α.

Let us denote $\alpha_{i,j}$ as the total number of values within population α_i for property j. Remembering that A was reconstructed from a:

$$E(\alpha_{i,j}) = A_j = a_j \rightarrow E(\alpha_{i,j}) - a_j = 0$$

$$\sum_{j=1}^{n}\left[E(\alpha_{i,j}) - a_j\right] = E\left(\sum_{j=1}^{n}\alpha_{i,j}\right) - a = 0$$

$$E(\alpha_i) - a = 0$$
$$E(\alpha_i - a) = 0$$

Therefore, we expect the population given by f to be an unbiased estimator for a.
We now define function $g((f(x, A)) = \alpha_i, B) = \beta_i$ to map values from distribution B to population α_i, similar to function f. We generate such a function for each property, to a total of all Y properties. Each time, we use a former set of entities. Thus we construct function F, which maps all Y properties to the initial synthetic population x such that $F(x, Y) = \pi_p$, which is a final, unbiased, synthetic estimator for polygon P (x being a 'clean' array of entities, the size as the real population of P).

4. The Mechanics of Data Set Disaggregation and Allocation in AASDC: A Generic Description

The first step of disaggregation is the simultaneous creation of two synthetic (non-spatial) data sets – households and individuals. These data sets are assigned various properties, as described above. For each given polygon, P, a blank data set is created for each type of entity within the given number of entities – either households or individuals.

Let H be the synthetic household population generated, and I be the synthetic individual population generated for P. $z = f(x, y)$ is an allocation function, assigning attributes from distribution y to population x, generating a new subset of data – z.

Household size and individuals:

$H_{Size} = f(H, P_{Size})$

Where P_{Size} is the distribution of household sizes within P. As shown above, re-aggregation of H_{Size} will yield P_{Size}.

After generating a blank synthetic population for the size of the population within P, we assign individuals to households within H.

$I_{i,Household} = f(I, H)$

At this point, the combined data sets contains linked entities, representing the distribution with polygon P and totaling to the same marginal distribution as the true population for P.

Social & Economic attributes

The second step focuses on assigning social and economic attributes to the entities. These attributes are now described:

Age:

Let $P_{Children}$ be the number of children distribution within P.

$H_{Children} = g(H, P_{Children})$

Function g being an allocation function, with the addition criteria enforced:

$H_{i,Size} - H_{i,Children} \geq 1, for\ every\ household\ i$

This condition prevents the existence of children-only households, and guarantees at least a single adult per household.

P_{Age} denotes the general age distribution within P. We define 'a' as an age group. For 0<a<17:

$I_a = f(I \in H_{Children}, P_{Age,a})$

Function f now allocates ages for age group 0 to 17 to individuals connected to households with children, according to the number of children.

Assignment of senior citizens (65+) living on their own is generated in a similar fashion. Denote the group of households $H_i \in H_{Size=1}$ as $H_{Elderlies}$. We now assign the number of lone senior citizens randomly to this group. By definition, $H_{Elderlies} \leq H_{Size=1}$

For age group 17<a<64 we assign a_1, a_2, a_3 as partial age groups. We now redefine H as the group of households which are not $H_{Elderlies}$.

$I_{a_1} = f(I \in H, P_{Age,a_1})$

Assuming $P_{a_2}, P_{a_3} > 0$, and remembering that by definition, $H_{a_1} \leq H$, the allocation exhausts all allocable individuals in age group $a_1$before all households are allocated.

We now redefine available H as households where not all individuals were previously assigned an age group.

$I_{a_2} = f(I \in H, P_{Age,a_2})$

The process is repeated for age group a_3 omitting households previously assigned. The result is three subsets of households, the number of each corresponding to the general

age distribution. Moreover, each household exhibits inherent age homogeneity among adults.

Finally, for $a_4 > 64$:

$I_{a_4} = f(I \in H, P_{a_4})$

With H being all households in which not all individuals were assigned age.

As $a_1 + a_2 + a_3 + a_4 = Adult\ Population \rightarrow$

$P_{Age,a_1} + P_{Age,a_2} + P_{Age,a_3} + P_{Age,a_4} = 1 - P_{Children} - P_{Eldelies} \rightarrow$

$I_{a_1} + I_{a_2} + I_{a_3} + I_{a_4} = Population - I_{Children} - H_{Elderlies} \rightarrow$

$I_{a_1} + I_{a_2} + I_{a_3} + I_{a_4} + I_{Children} + I_{Elderlies} = Population$

Gender:

Let $P_{Gender,a}$ be the distribution of gender for age group a within polygon P. As gender is a binary attribute, we can state that the total number of females to be allocated for age group a, F_a, within P is:

$F_a = P_{Gender=female,a} \cdot Population_a$,

M_a, the total male population to be allocated is equal to $Population_a - F_a$.

For age group 0<a<17, we use the general allocation function:

$I_{Gender,a} = f(I_a, P_{Gender,a})$, for all individuals in age group a

For age group 17<a<64 we use the same a_1, a_2, a_3 as partial age groups given above. Denote $\dot{H}_{a_i}$ as a subset of households which satisfy the following conditions:

- $H \in H_{Size \geq 2}$
- $H \in H_{a_i}$

We now draw the smaller of either M_{a_i} or F_{a_i}, , households from $\dot{H}_{a_i}$. We randomly assign a gender attribute to two individuals within these households. All further individuals within a_i are assigned as either male or female, depending on the larger, such that all individuals $I_{a_i} \in \dot{H}_{a_i}$ are assigned a gender.

We further denote $\ddot{H}$ as all households which satisfy $H \notin H_{a_i}$ for all previously assigned age groups a_i:

$I_{Gender,a} = f(I_a \in \ddot{H}, P_{Gender,a})$

This algorithm satisfies the randomness for each individual I, but creates heterogeneity within a majority of households.

Workforce participation:

Denote $P_{Work,Gender}$ as the probability of being employed, by gender, within P. For both genders this probability recreates the work force participation distribution. We denote a_{Work} as the age group comprising the work force (in Israel this corresponds to ages 17-65). We wish to prioritize participation within this age group.

$I_{Work} = f(I \in I_{a_{Work},Gender}, P_{Work,Gender})$

For both genders, if $\sum I_{Work,Gender} < Population_{Gender} \cdot P_{Gender}$, then we randomly select individuals within $I_{a>65,Gender}$, and allocate them, until such point where

$\sum I_{Work,Gender} = Population_{Gender} \cdot P_{Gender}$

Occupation:

Denote O_i as a field of occupation, and $P_{O_i,Gender}$as the probability of a worker from a given gender being employed in a given occupation. For each occupation and gender:

$I_{Occupation,Gender} = f(I \in I_{Work,Gender}, P_{O_i,Gender})$

Suppose $\sum I_{O_i,Gender} < Population_{O_i} \cdot P_{O_i,Gender}$, then we continue to assign

$I_{Occupation,Gender} = f(I \in I_{NotWork,Gender}, P_{O_i,Gender})$ until such point that

$\sum I_{O_i,Gender} = Population_{O_i} \cdot P_{O_i,Gender}$

Industry:

Denote ι_i as an industrial sector and $P_{\iota_i,Gender}$as the probability of a worker from a given gender being employed in a given sector. For each industrial sector and gender:

$I_{Industry,Gender} = f(I \in I_{Work,Gender}, P_{\iota_i,Gender})$

Suppose $\sum I_{\iota_i,Gender} < Population_{\iota_i} \cdot P_{\iota_i,Gender}$, then we continue to assign

$I_{Industry,Gender} = f(I \in I_{NotWork,Gender}, P_{\iota_i,Gender})$ until such point that

$\sum I_{\iota_i,Gender} = Population_{\iota_i} \cdot P_{\iota_i,Gender}$

Disability

Let D be a binary variable, indicating disability for a person, and P_D the probability of being disabled within P.

$I_{Industry,Gender} = f(I \in I_{NotWork}, P_D)$

Suppose $\sum I_D < Population \cdot P_D$, then we continue to assign

$I_D = f(I \in I_{Work}, P_D)$ until such point that

$\sum I_D = Population \cdot P_D$

Education

Let E_{Gender}be the distribution of the total years of education for a given gender. For a given age group a_i,a segment of $E_{Gender}$is selected, and denoted $\dot{E}_{Gender}$

$\dot{E}_{Gender} \leq \frac{a_i}{\beta}$, where $\beta > a_i$ so as to enforce total years of education to represent a certain proportion of age.

$I_{Education,a_i} = f(I_{a_i}, \dot{E}_{Gender})$ for both genders and all age groups.

Earnings

As this is a pivotal variable for distributing households to buildings and thereby fixing their spatial allocation, we offer two methods. The first is a scoring and ranking method that assigns individuals to an earnings quantile based on attributes that presumably influence earnings (i.e. gender, occupation, education, age and industrial branch). The second approach estimates a standard Mincer-type earnings equation.

Method A:

Let A be a national (as in for all polygons) array of the following attributes: age, gender, education, occupation and industry which correlate with earnings distribution, ϕ. A_i is a possible combination of these attributes, yielding an earnings distribution ϕ_i. We find A_j such that:

$\bar{\phi}_j \geq \bar{\phi}_i$ for every $i \neq j$.

We map all A_i in this fashion, finding the relative hierarchy of attribute groups such that if

$j > i$ than $\bar{\phi}_j \geq \bar{\phi}_i$.

Let Q be the known earnings distribution[2] within P, divided into quantiles. Denote Q_k as earnings distribution for quantile k.

$I_{A_i} = f(I \in A_i, Q_K)$

[2] Each earnings quantile includes the median income and the corresponding standard deviation, for that particular polygon. For different polygons, the first quantile will be different, as will other quantiles.

Where ‘i’ is within the ‘k’ quantile of A.

Method B:

Assuming that detailed earnings distribution is not available for P, we propose a ‘Mincer-type’ regression allocation, based on other available data.

Let A be an array of the following attributes: age, gender, education, occupation and industry.

Function g coalesces these attributes in to a single score for a single individual, denoted as S_i

$S_i = g(A_i)$

$S_i = \beta_1 \cdot Age + \beta_2 \cdot Gender + \beta_3 \cdot Occupation + \beta_4 \cdot Industry$ is one such function. Each parameter is drawn from a normal distribution:

$\beta_{\mathrm{i}} = N(\hat{\beta}_i, sd(\hat{\beta}_i)^2)$

Where $\hat{\beta}_i$ denotes a calculated regression estimator for field i, based on national-level data, and $sd(\hat{\beta}_i)$ denotes the standard deviation of the estimator. This is done in order to introduce some hetrogenity within earnings between polygons. If any additional information is available for polygon P’s earnings data it is possible to refine this:

$\beta_{\mathrm{i}} = N(\hat{\beta}_i + \alpha_1, sd(\hat{\beta}_i)^2 + \alpha_2)$

Where α_1 and α_2 are weights calculated from this additional information. For instance, suppose earnings distribution at the polygon level is available, consisting earnings quantile mean or median values. Let Q_i be earnings quantile i, and δ_{Q_i} be the regional earnings average for a given polygon[3] for quantile i. We produce α_1 such that

$\overline{I_{Q_\iota}} = \overline{S_{Q_\iota}} = \overline{\beta_1 \cdot Age_{Q_\iota}} + \overline{\beta_2 \cdot Gender_{Q_\iota}} + \overline{\beta_3 \cdot Occupatıon_{Q_\iota}} + \overline{\beta_4 \cdot Industry_{Q_\iota}} =$

$(\hat{\beta}_1 + \alpha_1)^4 \cdot \overline{Age_{Q_\iota}} + (\hat{\beta}_2 + \alpha_1) \cdot \overline{Gender_{Q_\iota}} + (\hat{\beta}_3 + \alpha_1) \cdot \overline{Occupatıon_{Q_\iota}} + (\hat{\beta}_4 +$

$\alpha_1) \cdot \overline{Industry_{Q_\iota}} = \overline{S_{Q_\iota}} + 4\alpha_1 \xrightarrow{Demand\ that} \delta_{Q_i}, \alpha_1 = {}^{\delta_{Q_i} - \overline{S_{Q_\iota}}}\Big/_{4}$)

It should be explicitly noted, that α_1 as calculated above represents the local deviation from the national regression. It should come as no surprise that this equals the difference

[3] Consisting primarily of a mean and a variation value.

[4] The average of a parameter produced by random distribution is its expected value (E)

between the national earnings for an average individual within Q_i at P, and the observed average for a given local quantile.

Spatial Allocation:

Let δ be a special set of dwelling units within P, including location coordinate and floor within a building. We do not assume the existence of a known spatial-distribution of households within P. Hence, household allocation to dwelling units is executed using a method different to that previously described.

The first stage consists of allocating attributes in a to β:

Denote β as a spatial information layer, consisting of the buildings within P. The layer includes location and physical properties of the building.

δ is assigned the attributes of β by a spatial join algorithm. For each location (x,y) within P $\delta_{\beta,(x,y)} = \beta_{(x,y)}$

Price information is given by a database, π, which contains past transaction information for dwelling units. δ_π, the documented price of each dwelling is allocated[5]

$\delta_{i,\pi} = \pi_i$

where i is every dwelling unit which has a recorded price within π[6].

It is expected that $\sum \delta < \sum H$, as a result of dwelling unit data paucity. Therefore, it is necessary to generate synthetic dwellings. As β is expected to be more complete, the process will generate synthetic units within buildings relatively empty in units, in accordance to local characteristics of dwelling unit density within buildings.

Total floor space is calculated by building:

$\beta_{i,f} = BH \cdot FH \cdot \bar{\delta}_{i,f}$

where BH is the building height, FH is the average floor height and $\bar{\delta}_{i,f}$ is the average dwelling unit floor space.

$\hat{\beta}_i = \frac{\sum \delta_{i,f}}{\beta_{i,f}}$

[5] Reference between the tables is done by a single-value code representing each dwelling unit, shared by both δ and π.

[6] Dwellings with no recorded transactions are allocated the average price per m^2 for P

$\hat{\beta}_i$ is now the floor space density within building i.

$\hat{\beta}_P = \sum \hat{\beta}_i$ is the average floor space density

An allocation distribution, is now generated which governs the allocation of new units into buildings, relative to their density. The probability of a building receiving a new unit will fall linearly as its current density approaches $\hat{\beta}_P$.

The probability function is therefore:

$$f_W(\hat{\beta}_i) = \begin{Bmatrix} \frac{2}{\hat{\beta}_P} - \frac{2}{\hat{\beta}_P{}^2} \cdot \hat{\beta}_i, \hat{\beta}_i \leq \hat{\beta}_P \\ 0, \quad o.w. \end{Bmatrix}$$ [7]

We produce $H - \sum \delta$: $\dot{\delta}_\beta = f(\beta, W)$, where $\dot{\delta}_\beta$ is a set of artificially generated dwelling units assigned to buildings.

Let D_β be the size distribution of apartments for each building

$\dot{\delta}_{f,\pi} = f(\dot{\delta}, D_\beta)$

Populating Dwelling Units:

Let R_H be a ranking system of households and R_δ a ranking system of dwelling units. It should be noted that the previous construction of data set δ guarantees that the sizes of these data sets is exactly the same. Household and dwelling unit scores are calculated:

$SR_{H,i} = H_{Income} \cdot 0.35 + H_{Size} \cdot 0.35 + U \cdot 0.3$

$SR_{\delta,i} = \delta_\pi \cdot 0.35 + \delta_f \cdot 0.35 + U \cdot 0.3$

Where U is a random component $U \sim Uniform(0,100)$ $\pi = x \cdot p - x \cdot C$

Household with rank $R_H = x$ is assigned the dwelling unit for which $R_\delta = x$.

[7] From the demand for linearity in the probability function: $f_W(\hat{\beta}_i) = \begin{Bmatrix} X - \frac{X}{\hat{\beta}_P} \cdot \hat{\beta}_i, \hat{\beta}_i \leq \hat{\beta}_P \\ 0, \ o.w. \end{Bmatrix}$ where X is the maximum value of the linear function, for buildings where density =0. Knowing that the a probability function's sum must equal 1:

$$F(\hat{\beta}_i) = \int_{-\infty}^{\infty} f_W(\hat{\beta}_i) d\hat{\beta}_i = \int_0^{\hat{\beta}_P} (X - \frac{X}{\hat{\beta}_P} \cdot \hat{\beta}_i) d\hat{\beta}_i = (X \cdot \hat{\beta}_P - \frac{X\hat{\beta}_P{}^2}{2 \cdot \hat{\beta}_P}) = 1 \rightarrow X = \frac{2}{\hat{\beta}_P}$$

This could also be derived through a geometric consideration of the probability function.

5. An Application Using Real Data

This section describes the application of the generic method outlined in the previous section using Israeli CT data from the national census 2008. For purposes of illustration, Figure 1 shows a map of CT's for the city of Haifa and its contiguous suburbs (the Israeli study area in DIM2SEA). Similar spatial units exist for all built-up areas in Israel. The application however is national in scale and comprises three main stages (Figure 2). In the first stage, a dis-aggregation procedure is applied to aggregate CT data. This results in the creation of discrete household and individual level data sets. In the second stage, households and individuals in the data sets are embellished with socio-economic attributes. Each attribute assignment iteration builds on its predecessor to create a synthetic representation that closely represents the socio economic fabric of the CT. The third stage is concerned with the spatial allocation of households to dwelling units.

This process relies both on the socio-economic attributes of households and the individuals comprising them and on the attributes of the dwelling units. The data analysis procedures used in the different stages and data processing stages are written in Python and SQL and are fully automated. This enables the automatic update of the database and the input of new and updated data as they become available. It also enables the adjustment of the database and the its variables, according to the changing needs of the application. Figure 2 depicts the specific connections and dependencies between the various attributes in the allocation process.

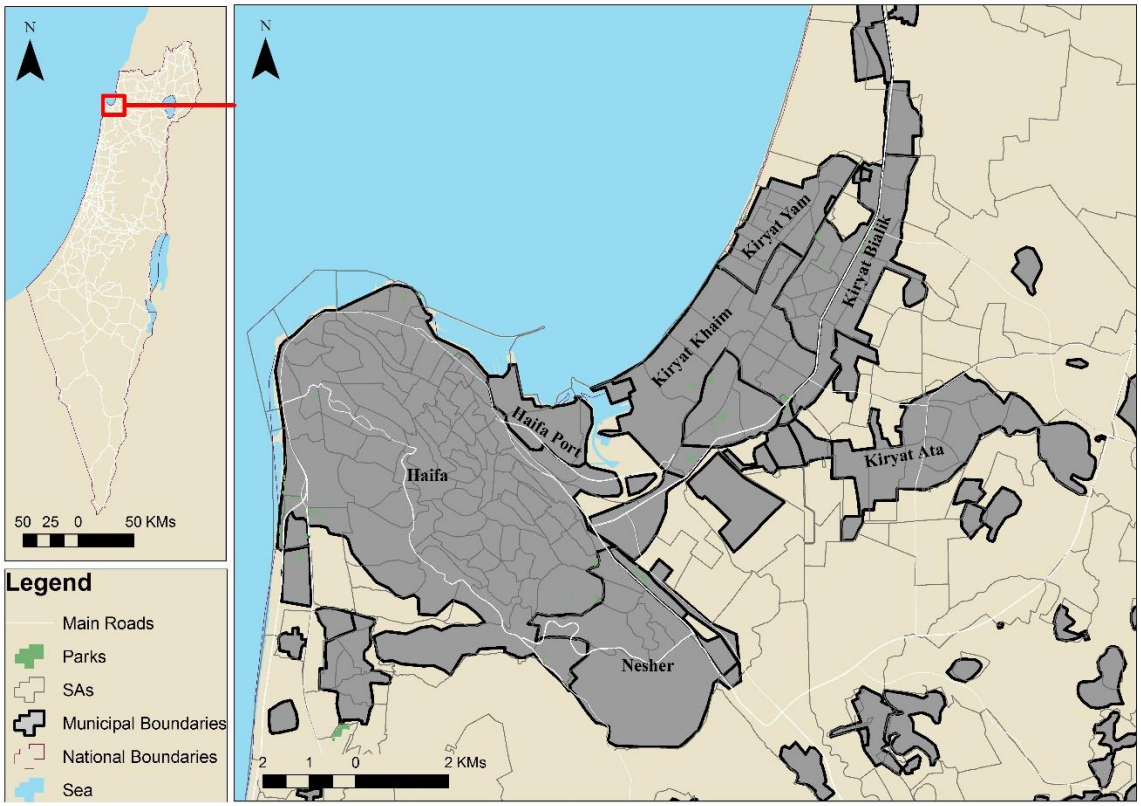


Figure 1: Map of CT's in the city of Haifa and contiguous suburbs.

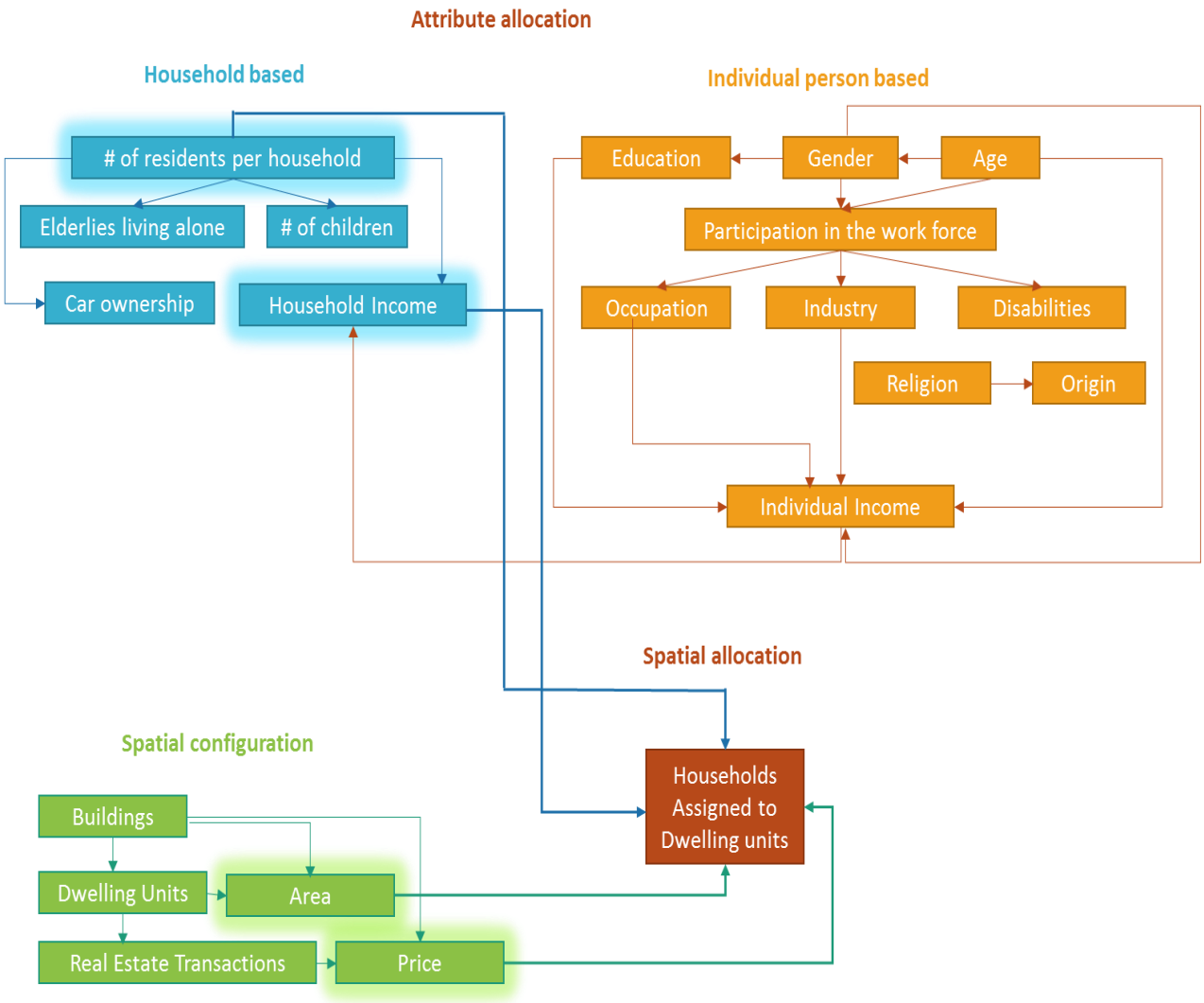


Figure 2 : The three stages of database development, input data (blue rectangles) and expected output (red square).

Note: religion and country of origin are not utilized in the database construction

.

5.1 Stage 1: Household and individual level data disaggregation

The starting point is the Israeli Census of Population conducted by CBS (Central Bureau of Statistics) in 2008. This includes over 3,000 CTs comprising over 2.3 million households and over 7.3 million inhabitants. We use aggregate CT level counts of households and household sizes, as well as population counts in order to create a dis-aggregated discrete dataset in which each household and each individual in a given CT is represented as a separate entity.

The original CT level data is disaggregated into households and individuals preserving the distribution of their attributes as represented in the census data. This process is comprised of two steps:

- Creating a Household Level Non-Spatial Dataset The total number of households of each size (number of residents) is calculated using the total household count per CT and the distribution (in %) of household sizes in each CT (items 1,2 in Table 1). Each household is of a different size and all the households in each CT represent the marginal distribution of household sizes in the original census data. This is achieved by enforcing an automated IFP procedure of control total adjustment. This ensures that the transition from household sizes given in % and yielding floating point calculations, will eventually fit the total household count in each CT. The result is a dataset containing a unique representation of more than 2.3 million households nationally.
- Creating an Individual Level Non-spatial Dataset Households are further broken down to represent the number of residents in each household as distinct entities in the dataset (items 2,3 in Table 1). This dataset contains over 7.3 million entities representing individuals tied to (members of) households.

5.2 Stage 2: Allocation of socio-economic and demographic attributes to households and individuals

Having represented households and individual residents within them as separate entities in the database, socio-economic attributes are allocated to each entity (households or individuals). At this stage, each household in the database is composed of a certain number of individuals reflecting the distribution of household size in each CT.

Age Allocation

Individuals in each household are assigned an age attribute according to a number of variables and adjusted to eventually represent the age distribution of households in each SA.

- Initially, the distribution of the number of children per household in each CT (% of households with 1, 2, 3, 4, 5+ children, see item 4 in Table 1) is evoked. This is used to allocate an age attribute of 0-17 to residents in each household ensuring representation of the original distribution of the number of children in a household in each CT. An automated code is written to ensure no households are composed entirely of children and that each household comprises at least one adult. In this way children, for example, are not allocated to households with only one resident. Other than these restrictions, the process is based on a random allocation procedure.
- In the second stage, the number of senior citizens (65+) in each CT living on their own (item 5 in Table 1) is used in order to allocate this age attribute residents of single-resident households until the quota of number senior citizenss living on their own in each CT is filled.
- Finally, the entire remaining (non-allocated) age distribution of the population in each CT is allocated to individuals with no age attribute in the database (item 6 in Table 1). This includes the remaining population of the 65+ category and the middle age group (17-64 years of age) which is further divided into three sub-groups. This is based on a sorted age group allocation. The algorithm takes each household in turn and assigns the individuals in the household unit an age until each age category is exhausted. In this way, most of the adults in each household are members of the same age category unless the category is exhausted in the middle of the allocation to one household. In this way, age homogeneity is introduced into the adult age distribution of each household.

Gender allocation

In contrast to the homogeneity in the adult age distribution allocation, the gender allocation procedure aims at producing gender heterogeneity in the adult population.

The gender allocation procedure is based on the age distribution in the population as gender distribution in each CT is given per age groups (item 7 in Table 1). Hence, gender is allocated to individuals in each CT according to their distribution in the age group to which they belong. Allocation of gender to the 0-17 age category is done on a random basis until CT quotas are exhausted. For the adult population, an algorithm is applied to introduce heterogeneity in household by selecting the adult members of each household and assigning them male and female attributes interchangeably (according to the CT counts). This does not prevent the existence of two members of the same gender in a household but creates a preference mechanism by which the occurrence of gender heterogeneity is more probable.

Workforce participation

This is a Boolean variable indicating the distribution of adult individuals participating in the work force according to gender (item 8 in Table 1). It is allocated to adult individuals in each CT until the quota working individuals of each gender in each CT is filled. A ranking preference is applied to first allocate positive participation to adults in age groups under 65 years old. If the quota is not filled, positive participation is allocated to individuals in the 65+ age group until the quota is filled.

Occupation

We aggregate occupations as they appear in the census into 3 categories: academic and management, administration, sales and services; and agriculture, industry and construction. The distribution of workers by occupation (item 9 in Table 1) is done according to the gender distribution of the adult population. A ranking preference is given to individuals participating in the work force.

Industry of Employment

We aggregate industries of employment as they appear in the census into 4 categories: commerce and communications; public sector; agriculture, industry, infrastructure and construction; domestic services. The distribution of workers by Industry (item 10 in Table 1) is done according to the gender distribution of the adult population. A ranking preference is given to individuals participating in the work force.

Disabilities

We aggregate 5 disability categories (disabilities in hearing, seeing, walking, dressing, memory) into one binary category (disabled or not). A ranking preference is given to individuals participating not in the work force. Otherwise this attribute is allocated to individuals on a random basis according to the distribution in each CT (item 12 in Table 1).

Religion

A religion Boolean attribute (Jewish or non-Jewish) is assigned to individuals in each CT to reflect the religion composition in that CT (item 13 in Table 1). Individuals in each CT are ranked according to household groupings so that in most (though not all) cases, the resultant household religion composition is homogeneous. The allocation procedure is based on a sorted religion category allocation: the algorithm takes each household in turn and assigns individuals in the household unit a religion category until the category is exhausted. This results in homogenous households in term of religion. For the most part, members of the same household are of the same religion category (unless the category is exhausted in the middle of allocation to a household). In this way, we preserve religion homogeneity in households.

Education

The distribution of education level categories builds on the gender and age distributions of the population. The census provides data on education based on gender (item 15 in Table 1). We use the age categories previously allocated to individuals in order to prevent young adults from being assigned an unreasonable education category. For example, individuals under the age of 30 are not likely to be assigned an education category of 16+ years (an exception can theoretically occur if the all other education categories quotas are exhausted).

Car Ownership

Relates to the number of cars in a households (none, one or two+). Allocated to households on a random basis until the CT quota is filled (item 16 in Table 1).

Earnings

The main data set used in this allocation process is the percentage of persons in each national earnings quantile in the CT (item 17 in Table 1). However, since earnings are a function of other factors we use ancillary data on a national scale in order to characterize the probability of an individual belonging to a certain earnings quantile based on the attributes already assigned to that individual. The attributes are used in order to give each individual a personal score based on these attributes. The attributes used are: gender, occupation category (builds on participation in the work force), Industry of employment builds on participation in the work force), education category and age. Our data includes the average income in the latter four categories all based on gender (items 18-22 in Table 1). We calculate a score based on the percentage of the earnings of each category in relation to the highest category. Each individual in the database is assigned a score and the allocation process sorts the individuals in each CT by score in a descending order and assigns each of them an earnings quantile (also by descending order), until each quantile's quota is filled. Table 2 illustrates the scores per each category. The ranking and scoring approach corresponds to Method A in section 4 (above).

Table 1: Datasets from the National 2008 Census used in stage 2

#	**Variable**	**Aggregation level**	**Source**	**Dataset table name**	**Allocation to**	**categories**	**Data Format**
1	Total number of households	CT	CBS census 2008	House holds	Buildings		Total count in thou.
2	Household Size (#of residents)	CT	CBS census 2008	House holds	Buildings à Households	1,2,3,4,5,6,7+	% of the total Households
3	Total population	CT	CBS census 2008	Demog_Yishuv, Sex Age Religion	Individuals		Total count in thou.
4	Number of children per household	CT	CBS census 2008	House holds	Households à Individuals (age)	1,2,3,4,5+	% of the total Households
5	Senior citizens living on their own	CT	CBS census 2008	Age 65+	Households (size 1)à		

					Individuals (age)		
6	Age distribution	CT	CBS census 2008	Demog_Yishuv, Sex Age Religion	Individuals	age groups 0-17, 18-19, 20-29, 30-64, 65+	% of the total population
7	Gender distribution by age	CT	CBS census 2008	Sex Age Religion	Individuals	age0-17m, age0-17f, age18-64m, age18-64f, age65+m, age65+f	% of the total population
8	Work force participation by gender	CT	CBS census 2008	Civil Labor Force	Individuals	age15+m, age15+f	% of the population above the age of 15 by gender
9	Occupation by gender	CT	CBS census 2008	Occupations	Individuals	academic and management (f, m); administratio	% of the population above the age of 15 by gender
11	Industry of employment by gender	CT	CBS census 2008	Industries	Individuals	commerece and communicati ons (f, m); public sector (f, m);	% of the population above the age of 15 by gender
12	Disabilities	CT	CBS census 2008	Disabilities	Individuals	aggregated to a binary category	% of the population above the age of 15
13	Religion	CT	CBS census 2008	Demog_Yishuv, Sex Age Religion	Individuals	Jewish, Non-Jewish	% of the population
14	Year of Immigration	CT	CBS census 2008	Origin	Individuals	prior to 1960; 1961-1989; 1990-2001; 2002+	% immigs in the Jewish population by year of immigration
15	Education distribution by gender	CT	CBS census 2008	Education	Individuals	up to 8 years; 9-12 years; 13-15 years; 16+ years	% of the adult population by gender
16	Car ownership	CT	CBS census 2008	Durable goods	Households	0; 1; 2+	% of the total Households
17	Earnings	CT	CBS census 2008		Individuals	10 national quantiles	
18		National		gross earnings per employee by occupation and gender	Individuals	earnings by occupation and Gender	Average monthly income in NIS
19		National		gross earnings per employee by industry and gender	Individuals	earnings by Industry and Gender	Average monthly income in NIS
20		National		gross earnings per employee by number of years of schooling and	Individuals	earnings by education and Gender	Average monthly income in NIS
21		National		gross earnings er employee by age and gender	Individuals	earnings by age and Gender	Average monthly income in NIS

Table 2: Scores used in the earnings allocation process

Var4rank	VarSub	VarPopulation	meanWage	rankValue
Ind	ind	u	9026.2666	60.28407
ind	ind	m	9612.5428	64.199655
ind	ind	w	7181.6732	47.964513
Ind	com	u	8579.9639	57.303331
ind	com	m	10043.16	67.075635
ind	com	w	6744.4099	45.044147
Ind	gov	u	7831.746	52.306179
ind	gov	m	10419.421	69.588585
ind	gov	w	6564.1322	43.840119
Ind	home	u	3422.7	22.859317
ind	home	m	3844.7	25.677744
ind	home	w	3357.9	22.426534
Occ	AcdMng	u	12006.653	80.189288
Occ	AcdMng	m	14972.889	100
Occ	AcdMng	w	9070.6575	60.580545
Occ	SrvSls	u	6038.4917	40.329504
Occ	SrvSls	m	7470.8366	49.89576
Occ	SrvSls	w	5324.9409	35.563885
Occ	IndAgr	u	5999.4035	40.068444
Occ	IndAgr	m	6589.9152	44.012317
Occ	IndAgr	w	3865.1455	25.814294
Edu	8	u	4462.7	29.805204
Edu	8	m	5145.1	34.362775
Edu	8	w	3355.2	22.408502
Edu	9-12	u	6173.1116	41.228595
Edu	9-12	m	7225.7239	48.258717
Edu	9-12	w	4866.0189	32.498865
Edu	13-15	u	8137.5	54.34823
Edu	13-15	m	10183.9	68.0156
Edu	13-15	w	6315.2	42.177566
Edu	16	u	11768.7	78.600064
Edu	16	m	14735.4	98.413876
Edu	16	w	9066.4	60.55211
Age	18-29	u	6008.2293	40.127389
Age	18-29	m	6940.3954	46.353082
Age	18-29	w	5062.2533	33.809464
Age	30-64	u	9886.2448	66.027639
Age	30-64	m	12002.281	80.160093
Age	30-64	w	7697.4097	51.408983
Age	65	u	8422.7	56.253007
Age	65	m	10082.4	67.337708
Age	65	w	4893.2	32.680401

5.3 Stage 3: Allocating Households to buildings to obtain a discrete spatial dwelling location distribution

At this stage, each household in the disaggregated households dataset (broken down into individuals) is allocated to a residential building in its corresponding CT. A national dwelling unit dataset is spatially joined to the building layer, containing data regarding the number of dwelling units in a building, their floorspace (area in m^2), and their respective floor in the building. Three further adjustments are made:

Prices: Since we only have prices for assets sold during 1998-2016, we assign the average price per m2 in a building to dwelling units with no transactions. If no transactions were made in a building, we assign the CT mean price per m2 to the dwelling unit.

Dwelling units and households: Since not all dwelling units are listed in the national data set, where a shortage of dwelling units relative to households occurs, we produce synthetic dwelling units in residential buildings. This is done by calculating the total residential floor space in a CT using the height of residential buildings and the average floor height in the CT. We then insert synthetic dwelling units into residential buildings according to the missing number of units relative to the total floorspace of a building and the floorspace of the dwelling units already occupying each building.

Households: are allocated to assets according to a coupled weighted ranking mechanism: On the household side, each household in the database is ranked according by size (35%) income (the median wage of all of its earning members 35%) and a random component (30%). On the building and dwelling unit side, each unit is ranked by area (35%), price per m^2 in 2009 prices (35%) and a random component (30%). Households in each CT are then assigned to dwelling units based on the corresponding rank.

6. Validation of the data disaggregation and allocation process

Verification of results is an issue endemic to the creation of synthetic spatial microdata (Hermes and Poulsen 2012). Consequently we conduct an external validation exercise. This involves comparing our synthetic spatial distributions with those derived from the Household Travel Survey (HTS) conducted in 2010 by the Jerusalem Transport Masterplan Team (JTMT; Oliveira et al., 2011). This is an independent source that yields microdata. For our purposes, the survey contains data on total population, households, age distribution and car ownership by Aggregated Transportation Area Zones (ATAZs). The HTS is based on representative samples of the population in the Jerusalem metropolitan area. The sampled population is multiplied by an expansion factor to represent the entire population in an ATAZ and their demographic attributes.

We aggregate our discrete household and person-level data (spatially allocated to residential buildings) into the HTS-based, spatial ATAZ units. The rationale for this is to check whether the results of our allocation method match data obtained from a different source when re-aggregated into different spatial configurations.

We compare the results of our model obtained from disaggregated statistical areas (CT's) to those of the HTS. As the HTS relies on a sampled population, it represents the population in aggregates at the ATAZ level and does not spatially represent discrete spatial individual distribution throughout the ATAZ. Therefore, those ATAZs covered by the HTS but not fully covered by our distribution model (for example, ATAZs in East Jerusalem) are excluded from the analysis. Only those ATAZs completely covered by the model coverage are used. Validation is performed for the following attributes: number of households, number of persons, persons ages 0-18, 19-64 and number of cars.

Table 1 shows NRMSE (normalized root mean square error)[8] statistics for two subsets of the original ATAZ dataset. Because of the aggregate nature of the ATAZ's, some of them are fully sampled in the HTS (i.e. the survey includes households from

[8] $NRMSE_y = \frac{\sqrt{\sum_{a \in ATAZ} (y_a - \hat{y}_a)^2 / n}}{\max(y) - \min(y)}$ where y and $\hat{y}$ are the HTS-sample and allocation-based values of a variable respectively and n is number of ATAZs.

all the travel area zones (TAZ's) included within the ATAZ) and some contain much lower sampling coverage. This can cause some misrepresentation of certain populations and demographics in some ATAZs. Here we compare the results of aggregating our synthetic database with those ATAZs with complete coverage and those with a sampling coverage above 70%. Mean errors between the two sources for most variables record a level of about 10 percent. This level of error may be the result of lack of congruence in size between the spatial units of the simulated data (CT's) and the units used for comparison (ATAZ's). Other studies have attributed similar levels of error to this cause (Edwards, Clarke, Thomas and Forman 2011). Furthermore, NRMSE could be over-estimated as they reflect the unknown sampling errors of both the HTS and CT datasets.

Table 1: Validation Results of Data Disaggregation Process for Key Variables

Variable		**Fully Sampled ATAZs**		**ATAZs sampling rate >70%**	
		Number of ATAZs	**NMRSE**	**Number of ATAZs**	**Number of ATAZs**
Population size		39	13.26%	65	8.52%
Number of households		39	9.95%	65	10.23%
Population size by age	0-18	39	11.51%	65	7.05%
	19-64	39	12.90%	65	10.92%
Number of cars		39	10.75%	65	10.21%

7. Conclusions

The methodology outlined above extends existing practice in a number of ways. First, in terms of scope, the approach presented is national. We are concerned with the automated allocation of all the individuals and households in Israel to buildings and dwelling units within them. This is a scale not hitherto attempted. The AASDC algorithm is capable of allocating over 7m individuals that recombine into 2.3 households and distributing them spatially to over 800,000 buildings comprising 1.4m dwelling units. Most data generation and allocation exercises of this kind do not operate at this level of magnitude. Obviously, country size is a key factor here and for larger countries this would be intractable. However, it is precisely in the small countries whose national space may comprise only two or three distinct regional housing and labor markets, where this kind of simultaneous allocation is most needed.

Second, the AASDC calls on extensive GIS resources for spatial allocation. This is a relatively unique feature. Disaggregated individuals are assigned to both buildings and dwelling units, i.e. separate GIS layers. Necessary adjustments need to be made to ensure the correct balance between the allocation of households to buildings such that there are enough dwelling units in a building to accommodate the households allocated to them. Our approach goes the extra mile in trying to ensure an accurate spatial distribution of households within CTs adjusting for an exogenous stock of buildings and dwelling units.

Finally, given the scale and scope of the exercise and the spatial accuracy of the allocation, the possibilities for creating flexible spatial configurations is almost endless. While the CT may be the springboard for the AASDC procedure, it is certainly not the final destination. This confers possibilities for spatial analysis at a level of resolution using high quality (synthetic) socio-economic data. The analytical opportunities that this creates are very exciting.

References

Beckman, R. J. Baggerly, K. A.and McKay, M. D (1996). Creating synthetic baseline populations. *Transportation Resesearch Part A. Policy and Practice., 30 (6),* 415–429

Bhaduri, B. Bright, E. Coleman, P.and Urban, M. L (2007) LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal, 69,* 103–117.

Edwards, K.L, Clarke, G.P., Thomas, J., and Forman, D. (2011) Internal and External Validation of Spatial Microsimulation Models: Small Area Estimates of Adult Obesity. *Applied Spatial Analysis and Policy, 4(4),* 281-300

Eicher, C. L.and Brewer, C. A. (2001) Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographical Information Science. 28,* 125–138.

Harper, G.and Mayhew, L (2012). Applications of population counts based on administrative data at local level. *Appl. Spat. Anal. Policy, 5,* 183–209.

Hermes, K., & Poulsen, M. (2012). A Review of Current Methods to Generate Synthetic Spatial Microdata Using Reweighting and Future Directions., *Computers Environment and Urban Systems, 36(4),* 281-290

Lichter, M and Felsenstein, D (2012). Assessing the costs of sea-level rise and extreme flooding at the local level: A GIS-based approach. *Ocean and Coastal Management, 59,* 47–62.

Linard, C. Gilbert, M. and Tatem, A. J (2010). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal 76,* 525–538.

Martin, D.,Cockings, S. and Leung, S. (2015) Developing a Flexible Framework for Spatiotemporal Population Modeling. *Annals of the Association of American Geographers, 105,* 754–772.

Mennis, J. (2003) Generating Surface Models of Population Using Dasymetric Mapping. *Professional Geographer.55,* 31–42

Oliveira, M.G.S., Vovsha, P., Wolf, J., Birotker, Y., Givon, D., & Paasche, J. (2011). Global Positioning System-Assisted Prompted Recall Household Travel Survey to Support Development of Advanced Travel Model in Jerusalem, Israel. *Transportation Research Record: Journal of the Transportation Research Board, 2246(1),* 16-23.

Pritchard, D. R and Miller, E. J (2011) Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation, 39,* 685–704.

Rahman, A.H., Tanton R and Liu S., (2010),Methodological Issues in Spatial Microsimulation Modelling for Small Area Estimation, *International Journal of Microsimulation,* 3(2), 3-22

Reibel, M. and Bufalino, M (2005). E. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning. A, 37,* 127–139.

.