Using Big (Synthetic) Data to Identify Local Housing Market Attributes.

A.Yair Grinberger and Daniel Felsenstein
Department of Geography, Hebrew University of Jerusalem

1. Introduction

Local housing market areas are notoriously difficult to identify. Short of in-depth case studies, little is known about local housing market attributes. Furthermore, housing markets are inherently local phenomena. Potential buyers tend to have limited search areas (van Ommeren, Rietveld and Nijkamp 1997) and potential builders tend to have local preferences for construction (Beenstock and Felsenstein 2015). If supply and demand can be met within a given geographic area without any spatial adjustment processes (such as migration or excess commuting) a local housing market (LHM) can be said to exist. But just how local is local? A dominant strand in the literature deals with defining LHMs and fine-tuning the levels of 'containment' or 'closure' needed to qualify as such (Jones, Coombes and Wong 2012). This approach posits that buyers will bid for a housing unit up to a distance that does not involve residential re-location. Consequently, migration or commuting patterns will determine the outer bounds of the LHM.

The housing market posited under these assumptions resembles the classic Alonso-Muth-Mills (AMM) model in which households with perfect information bid for residences whose equilibrium price (per sqm) declines monotonically from the city center. This is of course the long-term equilibrium and spatial arbitrage involves a trade-off between accessibility and size of residential unit. Over the short term however, households and house builders have (local) preferences, the housing stock is differentiated in terms of quality and spatial adjustment processes (such as commuting) need to operate in order to ensure a smooth-functioning market (Alonso 1964, Mills 1978). Under these conditions it would seem more pertinent to define the LHM in terms of substitution within a given area than in terms of self-containment (Pryce 2013). In addition, the adjustment mechanisms at the core of LHM definitions, operate at different spatial and temporal scales. Migration to a LHM represents a longer-term adjustment process and can signify the outer boundary of the LHM. Shorter-term and shorter distance spatial arbitrage can be achieved though commuting, short distance

searches and moves based on housing 'improvement' consideration. Finally, there are housing submarkets within LHMs that are characterized by socio-economic considerations such as population composition, housing types and tenure. These cannot be identified by arbitrary or calibrated containment or cut-off measures. Given these reservations, the elusive search for an applicable level of self-containment for LHM definition may belie the question whether a one-size-fits-all definition can really capture local conditions and idiosyncracies.

One approach to housing submarket identification involves clustering hedonic housing characteristics and searching for spatial heterogeneity (Bourassa, Hoesli and Peng 2003; Bhattarcharjee, Castro, Maita and Marques 2016). The basic thinking is that if hedonized house prices are highly correlated over time, they should belong to the same market and the submarket therefore corresponds to a local equilibrium between supply and demand. Another view is that within local housing markets, sub areas should be perfect substitutes. This can be tested by estimating a discrete choice model for locational choices which indicates the level of substitutability or by estimating cross price elasticity of prices across pairs of housing characteristics (Pryce 2013).

In contrast to current practice, we harness recent advances in data disaggregation and the generation of synthetic spatial microdata to propose an approach for identifying LHM attributes (rather than defining LHMs). This is an important input into the study of urban change in general and housing market analysis in particular. We show how traditional public sector 'small' data can be disaggregated to yield big (synthetic) spatial microdata. Specifically, we illustrate how micro-level housing market information can be derived by combining big (synthetic) socio-economic data with house price data using a three stage analysis. Initially, an allocation algorithm is used to attach synthetic socio-economic attributes to residential buildings. Administrative census tract data is fused with a detailed buildings GIS layer and a national residential dwellings dataset to generate an accurate synthetic spatial representation of individuals and households occupying dwelling units in both single and multi-unit buildings. The result is a national-level database comprising millions of households cross-tabulated with the attributes of the residences they occupy and the synthetic socio-economic characteristics that they represent.

The second stage deals with identifying inconsistencies between housing values and the ascribed socio-economic attributes of the resident population. This involves analyzing the residuals of the house price-house occupier relationship identifying incipient 'hot spots' and applying measures of spatial clustering. In the third stage, the clusters identified are typologized using the synthetic socio-economic data coupled with building attributes data. This yields information on housing market attributes such as segmentation and patterns of change such as gentrification. The approach is operationalized for the entire stock of residential units and households in Israel. It can be easily reproduced in other national contexts providing the flexible levels of spatial resolution needed in local housing market analysis.

2. Generating Big (Synthetic) Microdata

We use a dedicated allocation algorithm for data disaggregation and the generation of synthetic spatial microdata in order to identify local housing market attributes. The algorithm generates data at a national scale that can then be spatially downscaled to the level of the city, neighborhood, building and even household occupying a dwelling unit within a geocoded building. This affords the potential for creating spatial units at multiple levels of spatial resolution. We combine socio economic data available at the 'Statistical Area (SA)[1] level with a national GIS buildings layer and with a national real-estate transaction database for the period 1997-2014 to generate synthetic spatial microdata. In this way we allocate over 7m individuals that recombine into 2.3 households and distribute them spatially to over 800,000 buildings comprising 1.4m dwelling units. The application is national in scale and comprises three main stages (Grinberger, Lichter and Felsenstein 2016). In the first stage, a dis-aggregation procedure is applied to aggregate SA 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

[1] SA's in Israel are uniform administrative spatial units defined by the Israeli Central Bureau of Statistics (CBS) and conform to census tracts. They have relatively homogenous populations of roughly 3,000 persons. Municipalities of over 10,000 population are subdivided into multiple SA's.

synthetic representation that closely represents the socio economic fabric of the SA. The third stage is concerned with the spatial allocation of households to dwelling units.

The allocation procedure uses the 'synthetic reconstruction' approach (see Hermes and Poulsen 2012) for artificially generating data and iterative proportional fitting (IPF) for sequentially adjusting the synthetic data so that it corresponds to the known marginal distribution of the population. The data analysis and processing procedures are written in Python and SQL and are fully automated. This enables updating the database as new data become available. It also facilitates the adjustment of the database and its variables according to the changing needs of the application.

2.1 Disaggregating and Spatially Allocating Synthetic Data

Figure 1 depicts the process of generating and allocating the synthetic microdata[2]. This process is then embedded in the workflow for identifying LHM attributes as depicted in Figure 2 (data generation). The starting point in the data disaggregation task is creating household and individual level data. We use aggregate SA counts from each of the 3000 SA's nationally. These pertain to households and household sizes as well as population counts. We create disaggregated discrete data in which each household and individual in a given SA is represented as a separate entity. The result is two non-spatial data sets: one of households (2.3 m observations) and one of individuals (7.0 m observations). The next stage is to allocate socio-economic attributes 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 SA.

The first allocated attribute relates to age. Individuals in each household are assigned an age category that is iteratively adjusted to represent the age distribution of households in each SA. The code ensure no anomalies arise and thus no households are comprised entirely of children and each household comprises at least one adult. Other than these restrictions, the process is based on a random allocation procedure. Consequently senior citizens and the working age population are assigned households with marginal adjustments to ensure that control totals are not exceeded. The algorithm takes each household in turn and assigns the individuals in the household unit an age

[2] For a formal description of the allocation algorithm and synthetic database construction, see Felsenstein, Samuels and Grinberger (2016)

until each age category is exhausted. Most of the adults in each household are members of the same age category unless the category is exhausted. In this way, age homogeneity is introduced into the adult age distribution of each household.

Gender is another key allocation variable. In contrast to the homogeneity in the adult age distribution allocation, the gender allocation procedure aims at producing gender heterogeneity. Allocation of gender to the 0-17 age category is done randomly while for the adult population the algorithm introduces heterogeneity in households by selecting the adult members of each household and assigning them male and female attributes interchangeably. 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. Variables relating to workforce participation, occupation, industry of employment, disabilities, education and car ownership are assigned to households in very similar manner. In most cases this is according to age and gender marginal distributions. In contrast earnings are treated differently. They are distributed to households based on a Mincer-type earnings regression that relates to the marginal contributions of age, education, gender, occupation and industry of employment.

The final stage involves allocating households to buildings to obtain a discrete spatial dwelling location distribution. The national dwelling unit dataset is spatially joined to the national 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. As not all dwelling units are listed in the national data set, a shortage of dwelling units relative to households occurs. We therefore generate synthetic dwelling units in residential buildings using SA averages. Households are allocated to dwelling units via a coupled weighted ranking mechanism. On the household side, each household is ranked 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 SA are then assigned to dwelling units based on their corresponding ranks.

3. Identifying Local Housing Markets Attributes in Israel

3.1. Clustering of LHMs

The processing of the synthetic households database is depicted in Figure 1 as the data generation step. Once the total national population of households is assigned and socio-economic attributes are spatially allocated to buildings and dwelling units within them, the issue becomes one of identifying whether these households are located in the type of housing one would expect given their socio-economic characteristics. To this end we empirically test for house price elasticities with respect to earnings.

As noted above, the allocation of synthetic households to dwelling units is partially stochastic, producing different sets of allocations each time. In order to identify repeated house price-earnings incompatibilities, we run the allocation procedure 54 times, each run producing a different dataset of average earnings per building. Median residual values resulting from a log regression of average earnings on average per meter price for each dataset are computed for each building [3]. The average cross-price elasticity for this regression is 0.229 with a standard error of 0.003. A Getis-Ord G_i-based hot spot analysis (Getis and Ord, 1992) is used to identify buildings for which this value is significantly low or high in relation to surrounding buildings[4]. The data is made spatially continuous by averaging the "hot spot" value (-1 – low/'cold', 0 – non significant, 1 –high/'hot') for buildings within each cell in a 50X50m grid vector layer (Table 1). Cold spots are buildings where populations with low earnings capacity occupy high price buildings and hot spots are the reverse. This comprises the clustering step in Fig 1. The criterion for inclusion is that an LHM comprise a minimum of 10 buildings. Table 2 shows that 180 LHMs meet this minimum requirement nationally. Many small clusters of buildings are discarded (~5700).

[3] Excluding buildings which are non-residential, had no socio-economic data or presented average values outside the range 1,000-100,000 NIS/m2. Final dataset size = 106,120 buildings.
[4] The radius of analysis is set according to the global Moran's I score first peak (3090m) and an inverse-distance-based measure is used thereafter.

Table 1: Clustering results

LHM Type	Number	Avg LHM size (s.d.) – number of buildings	Avg. Earnings (s.d.) over all runs	Avg. Price/m^2 (s.d.)	Correlation	Avg. Residual (s.d.)
Hot (1)	101	27.76 (29.72)	2,204.60 (3,564.80)	1,054.78 (3,705.89)	0.13	0.22 (0.18)
Cold (-1)	79	97.51 (366.91)	592.04 (157.44)	1,216.10 (604.20)	0.64	-0.26 (0.24)

3.2. Classification of LHMs

Having clustered buildings on the basis of a house price-earnings relationship, the classification step (Figure 2) generates a typology of LHMs was using Grouping Analysis - a non-spatial k-means clustering procedure. The group analysis procedure uses normalized average values of different socio-economic variables for the LHM's (Table 2). The source of these is the synthetic values generated by the database in the first step. Thus the generation of synthetic big data contributes inputs to both the initial and final stage of the analysis (Fig 2). The analysis is performed separately for each LHM type and excludes LHMs that fail to meet the basic criterion for inclusion, i.e. 10 buildings (Table 1). Two generic clusters are identified for both ‘cold’ and ‘hot’ LHM's.

Table 2: Variables used in the clustering stage

Variable	Details
HH size	Number of members
% children	% of HH members whose age below 18
Car ownership	Number of cars – 0,1,2+
% employed	% of HH members who are employed
% academic and management	% of employed HH members whose occupation is academic or management & sales
% disability	% of HH members which are disabled
% Jewish	% of HH members which are Jewish
Immigration	Weighted average of number of immigrants by immigration period
% education > 12	% of HH members with more than 12 years of education
% earners	% of HH members with registered earnings
Total income	Total monthly income for a HH
Average income	Average income according to a HH's total income and size

The final stage in the empirical analysis involves inductively assigning typologies to the clusters. In this respect, empirical testing is used as proof-of-concept. If emergent cluster represent plausible LHM's i.e. those that reflect actual real-world patterns of housing change, we can conclude that the data processing has identified salient local housing market attributes.

Characterizing the LHM's is therefore based on socio-demographic attributes and local knowledge. In 'ethnic enclave' LHM's socio-demographic characteristics (Table 3) correlate with those of the low level earners and large families occupying expensive housing - a key attribute of the ultra orthodox Jewish populations that cluster in both Jerusalem and the satellite cities around Tel Aviv (Shilhav 1993). The other 'cold' spot typology identifies 'rental dominated' housing markets. These include mainly inner-city LHMs within metropolitan cores, where relatively poor populations overcome high housing prices through renting and doubling -up rather than owning property. These groups include student-dominated areas in Jerusalem, Tel Aviv and Be'er Sheva and migrant worker clusters in Tel Aviv (Figure 3). These LHM's are characterized by low numbers of earners and small families.

The 'hot' LHM clusters represent markets where the supply-demand mismatch (i.e. stronger population living in under-priced dwellings) is probably due to instability and housing market dynamics. Such disequilibrium can be found in 'new construction markets' such as the new city of Modiin where construction began less than 20 years

ago and the western neighborhoods of the city of Rishon Letzion south of Tel Aviv that has experienced a building boom over the last 10-15 years (Table 3, Figure 3). A second 'hot' cluster includes mainly inner-city neighborhoods in cities near metropolitan cores where average construction year is 1986 (Figure 3). These LHM's are highly accessible but also have deteriorating infrastructure and represent typical gentrification or 'transitional' markets. The socio-economic characteristics of their residents fit this profile: i.e. educated and employed couples with almost no children (Table 3).

Table 3: Characteristics of LHM groups

LHM Type	**Group**	**Variables – absolute value (% of HH members)**		
		HH size	Children	Earners
Cold	Ethnic Enclaves	4.54 (100)	2.14 (47.03)	1.43 (31.54)
	Rent Dominated	2.18 (100)	0.46 (20.94)	1.33 (61.00)
		Children	Employed	Education (> 12 yrs)
Hot	New Construction Markets	1.47 (35.19)	2.73 (65.24)	1.36 (32.51)
	Zones in Transition	0.84 (23.56)	2.74 (76.82)	1.34 (37.62)

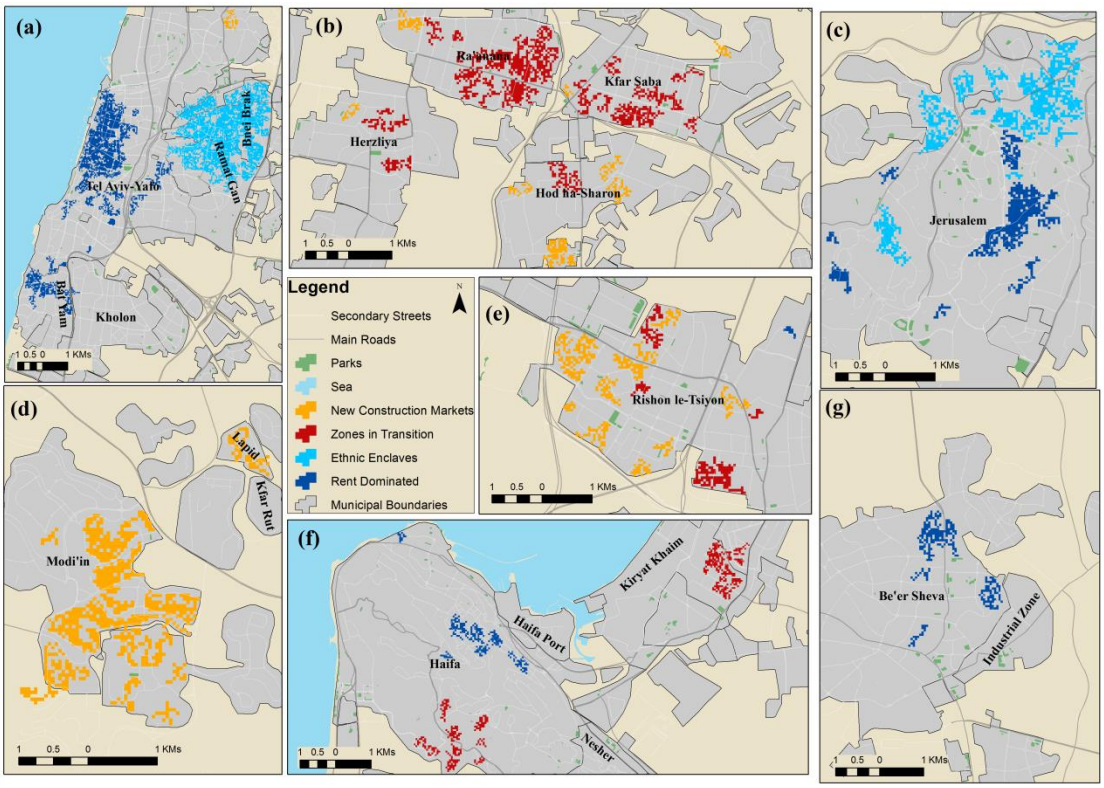

Figure 3: Characteristic LHM's for a Selection of Cities in Israel: (a) Tel Aviv and adjacent suburbs, (b) Tel Aviv's northern suburbs, (c) Jerusalem, (d) Modi'in, (e) Rishon le-Tziyon (A southern suburb of Tel Aviv), (f) Haifa, (g) Be'er Sheva.

4. Conclusions

Local housing market change is an area of traditional interest in regional science. This paper illustrates how recent advances in generating big (synthetic) microdata can be harnessed to yield new insights into local housing market dynamics. The core sources on which this analysis rests are national data sets from conventional government databases such as the national census, national GIS building layer and real estate transactions from the tax authorities. These are becoming increasingly available in both volume and quality hitherto experienced. This is part of an 'open government' movement worldwide that releases structured open data that can be further manipulated (Arribas-Bel 2014). The IPF-type algorithm that underpins the data disaggregation process described here is illustrative of the computational developments directly spawned by this growing availability of public sector data. These developments allow us to generate big (synthetic) data from open institutional data

The analysis presented above has been data-exploratory and broadly descriptive. Regional science is traditionally concerned with the inferential. With the growth of open-government-generated big data, the challenges become more complex. As traditional 'tall data' (where n>k) makes way for 'fat data' (where k>n), this can lead to more explanatory variables than needed and subsequent model over-fitting. This calls for the development of auxiliary tools capable of guiding selection for appropriate modeling under these conditions (Varian 2014). These are but some of the challenges that regional science faces as it moves from a data-scarce to a new data-rich environment and from broad spatial aggregates to high level spatial resolution.

References

Alonso W (1964*) Location and Land Use*, Harvard University Press, Cambridge. Evans

Arribas-Bel D (2014) Accidental, open and everywhere: Emerging data sources for the understanding of cities, *Applied Geography* 49, 45-53

Beenstock M and Felsenstein D (2015) Estimating spatial spillover in housing construction with nonstationary panel data, *Journal of Housing Economics* 28, 42-58

Bhattacharjee A Castor E, Maiti T and Marques J. (2016) Endogenous Spatial Regression and Deliniation of Submarkets: A New Framework with Applications to Housing Markets, *Journal of Applied Econometrics*, 31, 32-57.

Bourassa SC, Hoesli M, and Peng VC (2003) Do housing submarkets really matter? *Journal of Housing Economics* 12(1): 12–28.

Felsenstein D, Samuels P and Grinberger Y. (2016) *AASDC: An Allocation Algorithm for Data Disaggregation and Synthetic Database Construction*, WP 02/16, DIM2SEA, The Development of a Dynamic Integrated Model for Disaster Management and Socio-Economic Analysis, Japan Science and Technology Agency (JST) and Ministry of Science, Technology and Space, Israel (MOST).

Getis, A., and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24(3)*: 189-206.

Grinberger, A. Y., Lichter, M., and Felsenstein, D. (2016). Dynamic Agent Based Simulation of an Urban Disaster using Synthetic Big Data, in Thakuria P, Tilahun N and Zellner M (eds) *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics*, Springer (forthcoming)

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

Jones C, Combes M and Wong C (2012) A system of national tiered housing-market areas and spatial planning, *Environment and Planning B*, 39, 518-532.

Mills E (1972) *Studies in the Structure of the Urban Economy*, Johns Hopkins Press for Resources for the Future, Baltimore MD.

Ommeren, J.N. van, Rietveld, P. & Nijkamp, P. (1997) Commuting: In Search of Jobs and Residences. *Journal of Urban Economics, 42* (3), 402-421.

Pryce G. 2013. Housing submarkets and the lattice of substitution. *Urban Studies*,50, 2682–2699.

Shilhav, Y. (1993) The emergence of ultra-orthodox neighborhoods in Israeli urban centers.pp 157-189 in Ben-Zadok E (Ed.) *Local Communities and the Israeli Polity: Conflict of Values and Interests* State University of New York Press, Albany, NY.

Varian H. R (2014) Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28 (2), 3-28

Figure 1. Work-Flow for Identifying LHM Attributes

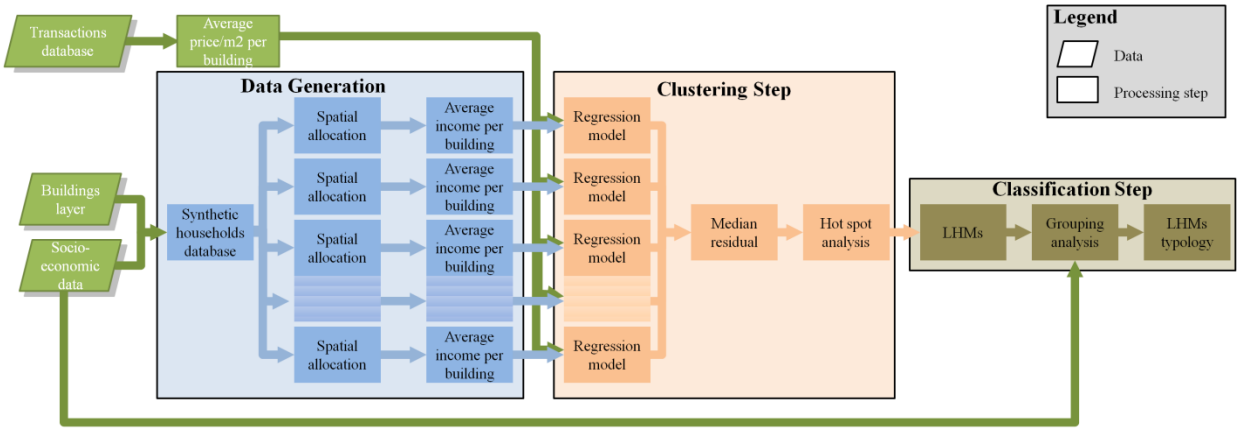

Figure 2: Data Disagregation and Synthetic Database Construction