Frankenstein's people: pseudo-individual populations from census data

Christoph Fink
Doctorate College GIScience
Dept. of Geoinformatics Z_GIS
University of Salzburg, Austria
christoph.fink@sbg.ac.at

Introduction

Most geographical research methods are sensitive to scaling. The *Modifiable Areal Unit Problem* (MAUP), a term coined by Openshaw (cf. 1984), describes a specific problem which arises from the inevitable arbitrariness of delineating geo-statistical spatial units.
Koch & Carson (2013) proposed widening the concept to the temporal and social dimensions (*MTUP*, *MSUP*) while looking into scaling issues in simulation models.
This research is set within a larger project which tries to quantify the effects of scaling on agent-based models of social actions and collective human behaviour. For the aimed-at stepwise aggregation, **datasets in the highest "sensible" resolutions** on all aforementioned dimensions **are to be synthesised**. This is implemented in a transferable and re-usable way; the aim is to use whichever data sources the probed model uses.

Methods

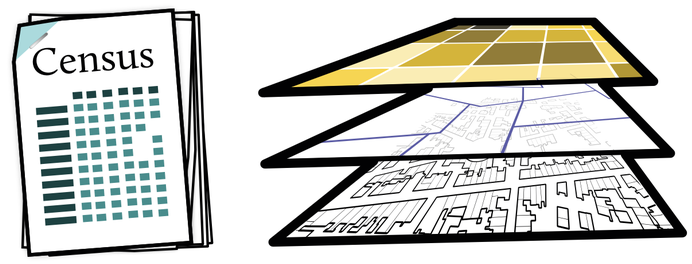
Input data used are a fine mesh population count grid, census tract polygons with demographic and socio-economic variables attached, and building polygons. These are available for most countries.
First, buildings are filtered using a local areal size threshold (building block's median±3σ) to eliminate none-residential buildings.
Next, the population counts from the grid dataset are distributed over the buildings.
Then, a **modified inverse distance weighting algorithm** is applied, calculating local values in each building for each of the available columns of the census data:

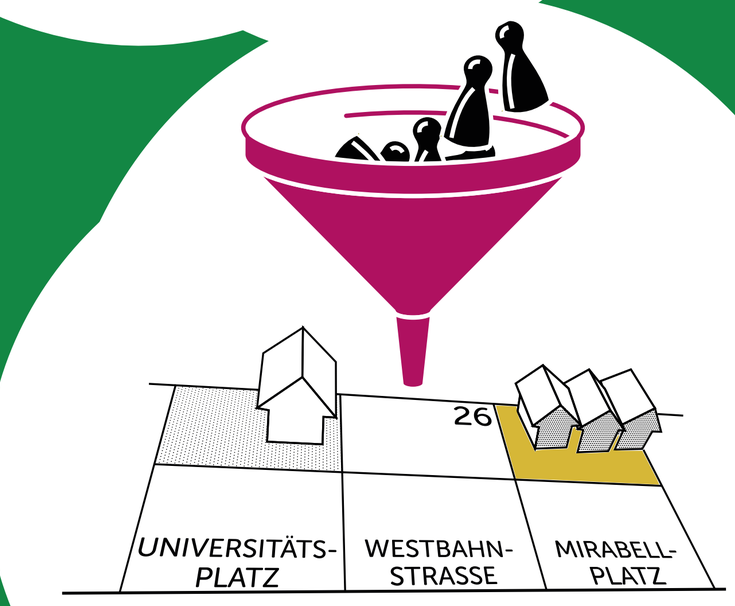
$$u(B_x) = \left(\sum_{i=0}^{N} \frac{u(S_i) \cdot \left(\frac{1}{d(B_i, S_i)} \right)}{\left(\sum_{j=0}^{N} \frac{1}{d(B_i, S_j)} \right)^p} \right) \cdot \frac{1}{u(S \ni B_x)} \cdot \frac{s(B_x)}{\sum_{B_k \in S_i} s(B_k)}$$

$u(A)$.....value of u in polygon A
p.....const.
$d(A,B)$.....euclidean distance between the centroids of the polygons A and B
$s(A)$.....initial "seed" population (weighting) in polygon A (typically a building)
B.....building polygon
S.....census tract polygon

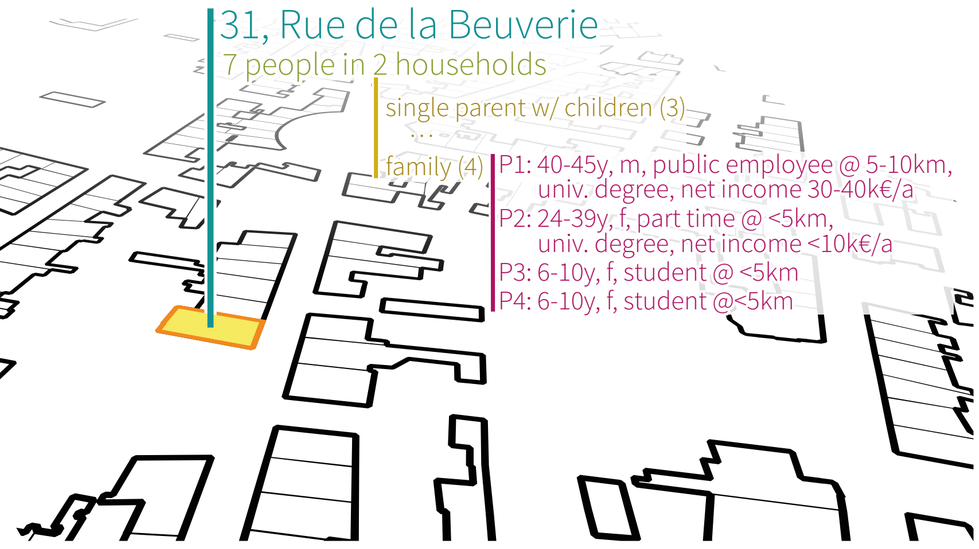
Finally **individuals** are **grouped into households** (household sizes and types are available from the census data), which in turn are distributed over the buildings.

Taking socio-economic and demographic datasets, as well as multiple layers of geographical data, as inputs, ...





... this Python/GDAL-OGR algorithm almalgamates high resolution population numbers with coarser data of more sophisticated indices ...



... and disaggregates it into individuals with all original attribute variables attached, who belong to households, which in turn live in buildings.

Results

The described method was applied using **sample data** from France, the city of Bordeaux was chosen as a test area.
The population grid (200×200 m) used is the *données carroyées* dataset of the *Institut national de la statistique et des études économiques* (INSEE 2012b); also the census data (in *IRIS* census tracts) is available from INSEE (2012a). The building polygons were obtained from the *Institut national de l'information geographique et forestiére* (IGN 2012).
The **output** of the calculations are datasets of **(a) buildings with whole-number population counts**, **(b) buildings with** decimal and whole-number values in all 350 columns of the **census dataset**, and **(c)** a set of **interrelational tables of individuals, households, and buildings**. The data are stored in a SpatialLite data base, and are linked to their respective spatial positions and extents.

Discussion

The disaggregation for the sample data **worked as expected**. Predictibility is neither sensible for this kind of algorithm nor required for the use case of benchmarking agent-based simulation models. Thus, true ground-truthing is not carried out. As an **intrinsic validation**, data is re-aggregated and compared to the input data. The mismatch in roughly 3.5% of all values is attributed to rounding errors in either the aggregation or the disaggregation processes.
It might be suggested that a purely artificial population were sufficient. This can be rejected: the general aim is to re-use the original dataset of the agent-based models under examination – special focus is laid upon models of urban residential mobility. Such models typically use statistically refined population data of existing cities. The **next step** in this research project is to **apply the algorithm to** data from actual, **existing models**. This will point out its strengths and weaknesses, and its robustness and transferability.

source code & more :-)

http://chri.stoph.at/blog/modelling/

References:
IGN (2012): BD PARCELLAIRE®. – available from http://professionnels.ign.fr/bdparcellaire (last visited 2013-09-16)
INSEE (2012a): Bases de données du recensement de la population 2008. – available from http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensement/resultats/2008/donnees-detaillees-recensement-2008.htm (last visited 2013-09-16)
INSEE (2012b): Données carroyées de la population. – available from http://insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=donnees-carroyees (last visited 2013-09-16)
Koch, A. & D. Carson (2012): Spatial, Temporal and Social Scaling in Sparsely Populated Areas – Geospatial Mapping and Simulation Techniques to Investigate Social Diversity,
in: T. Jekel, A. Car, J. Strobl & G. Griesebner (eds.), GI Forum 2012, (pp. 2–5).
Openshaw, S. (1984): The Modifiable Areal Unit Problem. CATMOG, 38.

Z_GIS GISCIENCE DOCTORAL COLLEGE UNIVERSITY of SALZBURG FWF