Areal interpolation of population in the USA using a combination of national parcel data and a national building outline layer

E. M. Weber*[1], J. J. Moehl[1], and A. N. Rose[1]

[1]Oak Ridge National Laboratory

*Email: weberem@ornl.gov

Abstract

We build on the recent literature on the use of parcel data in areal interpolation of population by incorporating high resolution building outline data. Our results show that models that constrain the operating footprint of the areal interpolation not just to parcel boundaries but the outlines of the buildings themselves have the potential to further reduce the areal interpolation error beyond what has already been demonstrated in previous literature. We also demonstrate that these principles apply not just to residential populations but to worker populations as well.

Keywords: population, areal interpolation, dasymetric, cadastral, parcel, poisson, buildings.

1 Introduction

Census data typically tabulate residents at an aggregate level within some set of enumeration areas, and when the spatial precision of the available aggregations is too coarse for a particular need, it is common for researchers to employ intelligent areal interpolation methods (often referred to as "dasymetric" methods) in an attempt to spatially refine the mapped distribution of population. Land cover and land use data, address points, street networks, and many other sources of information have been used to better concentrate the mapped population in the land actually containing homes.

When available, parcel datasets (i.e., property data, tax lots, cadastral data) can be a valuable source of information about the segmentation and use of land within an enumeration area, and so have seen some recent use in the areal interpolation literature. In some parcel datasets, the residential floor area and/or the number of residential units is recorded for each parcel, which allows for very precise disaggregation of residential census counts (Maantay et al., 2007). Even without these attributes, Tapp (2010) showed that simply populating each residential parcel as if it contains a single average

*Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

household results in a more accurate modeled distribution than dasymetric techniques using national land cover data (though with the downside that the populations of multi-family housing in urban areas tend to be underestimated). Jia et al. (2014) adapted dasymetric techniques based on empirical sampling (Mennis, 2003) to parcel data in a county in Florida, demonstrating improvement over land cover-based methods, and showed further improvements when using parcel data and land cover in a combined model (Jia and Gaughan, 2016).

Here, we extend the use of parcel data in areal interpolation to a scenario in which both a national (USA) parcel database and a national map of building outlines are available. The national parcel database provides a common set of land use attributes across most of the contiguous United States (CONUS), and the building outlines can potentially allow more precision in both the estimation of densities as well as the final assignment of populations to small areas.

2 Data and methods

2.1 National parcel layer

Parcel polygons and their associated land uses were obtained from the CoreLogic ParcelPoint database. There are 276 different land use classes represented by a three-digit number where the first digit represents larger super-classes (for example, a first digit of 1 indicates a residential land use). Many parcels, however, do not have an associated land use (this class is signified by "Null") and there are many patches of land that are not covered by parcels at all. We consolidated the land uses to a simpler set of categories to use in the regression modeling described below. The specific categories and the land uses assigned to each are defined in table 1.

2.2 National building layer

The building outline layer is a national layer representing building area, derived from NAIP (National Agricultural Imagery Program) aerial images using a supervised convolutional neural network classifier (Yang et al., 2018). The classifier produces a binary output at the resolution of the source imagery (1 m), which we converted to polygons and stored in a PostgreSQL database extended with PostGIS.

2.3 Regression modeling

A variety of options for using regression to estimate population densities from multiple land classes have been demonstrated in the literature. An exhaustive assessment of the different approaches is outside the scope of this paper, but see section 4 for a discussion of potential options to explore in future work. In the current work, we use generalized linear models with a poisson data distribution, an identity link function, and a zero intercept, following the work of Flowerdew and Green (1989). After applying the model coefficients to estimate populations for every intersection of building and parcel (which we refer to as building "pieces"), we also implement a standard scaling step to ensure that the estimates for the pieces sum to the actual tract populations.

Census tracts serve as our observational units for both residential and worker populations, and the corresponding source populations are derived from two surveys from the U.S. Census Bureau.

Land use	Description	Res vs. Wrk	Category
118	Frat/Sorority House	res	groups
119	Residence Hall/Dormitories	res	groups
155	Group Quarters	res	groups
156	Orphanage	res	groups
157	Nursing Home	res	groups
103	Apartment/Hotel	res	multi
106	Apartment	res	multi
111	Cooperative	res	multi
112	Condominium	res	multi
113	Condominium Project	res	multi
115	Duplex	res	multi
116	Mid Rise Condo	res	multi
117	High Rise Condo	res	multi
131	Multi Family 10 Units Plus	res	multi
132	Multi Family 10 Units Less	res	multi
133	Multi Family Dwelling	res	multi
134	Mixed Complex	res	multi
148	PUD	res	multi
151	Quadruplex	res	multi
165	Triplex	res	multi
245	Office & Residential	res	multi
281	Stores & Residential	res	multi
100	Residential (Nec)	res	single
102	Townhouse/Rowhouse	res	single
135	Mobile Home Lot	res	single
136	Mobile Home Park	res	single
137	Mobile Home	res	single
138	Manufactured Home	res	single
160	Rural Homesite	res	single
163	SFR (Single-Family Residential)	res	single
509	Ranch	res	single
511	Farms	res	single
2xx	Commercial (65 three-digit codes)	wrk	com
6xx	Public (35 three-digit codes)	wrk	pub
xxx	Other (144 three-digit codes)	wrk	other

Table 1: Consolidation of land use codes from CoreLogic parcel data to model categories. All 276 land use codes are assigned to a category (res-single, res-multi, res-groups, wrk-com, wrk-public, or wrk-other). "2xx" indicates all codes with a first digit of 2; "xxx" indicates "everything else" (all codes not explicitly specified elsewhere in the table). This category contains recreational, transportation, and agriculture uses, among others.

Worker populations are obtained from the LEHD (Longitudinal Employer-Household Dynamics) Origin-Destination Employment Statistics (LODES), while residential population estimates are obtained from the American Community Survey (ACS) 5-year estimates. These populations are the outcome variables in the regression models. The predictors are the areas (in m^2) of the consolidated land use classes. We fit the models only with census tracts having a non-zero population, having a non-zero area for more than one of the three categories, and in which at least 90% of the building area in the tract is within parcels with non-null land use values.

We performed a total of four regressions. First we fit a pair of regression models (one for residential and one for workers) using all eligible tracts (those meeting the criteria defined above) across CONUS and with predictors representing building area of the land use categories. (We will refer to these models as *res-bld* and *wrk-bld*.) We then fit two additional models for the same set of CONUS tracts, but using predictors representing land area of the land use categories rather than building area. (We will refer to these models as *res-land* and *wrk-land*.) The latter two models are more akin to the models demonstrated in previous studies, in which densities are estimated in terms of land area rather than building area.

3 Results

The coefficients of the regressions can be interpreted as estimated population densities for each of the consolidated classes (Table 2). It can be seen that for every category, the densities of the *res-land* and *wrk-land* models are lower than the corresponding densities of the *res-bld* and *wrk-bld* densities. This is as expected, because the *res-land* and *wrk-land* densities are in terms of total parcel land area rather than the more concentrated building area in the *res-bld* and *wrk-bld* models. But regardless of the difference in magnitude, it can also be seen that the relative order of the categories from highest to lowest density is consistent between the two residential models and between the two worker models (i.e., *multi* > *groups* > *single*, and *com* > *public* > *other*).

Model	Category	Density (people per square meter)
res-bld	groups	0.04580401
	multi	0.06268017
	single	0.01518875
wrk-bld	com	0.02825485
	public	0.01185636
	other	0.00305430
res-land	groups	0.00741439
	multi	0.01229686
	single	0.00006058
wrk-land	com	0.00150508
	public	0.00000154
	other	0.00000019

Table 2: Estimated population densities (poisson regression coefficients) for the three residential and three worker categories from the four regression models.

Because our model used census tract populations as the source populations, it is possible to assess the estimation error at the finer block group level (for which population data is available from both the ACS and LODES sources). (It is important to note that both the ACS and LODES populations are estimates from surveys and so should not be considered "ground truth" in the same way full decennial census counts would be.) At the block group level, we assess the estimates on a suite of metrics proposed by Sridharan and Qiu (2013). Two well-known metrics, root mean square error (RMSE) and mean absolute error (MAE), are shown in Table 3. Because these metrics are sensitive to size (block groups with larger populations will tend to have larger errors), we also include four standardized metrics (adjusted RMSE, mean absolute percentage error (MAPE), median absolute percentage error (MedAPE), and population-weighted mean absolute error (PWMAE)), all of which divide each error value by the observed value. The two building area-based models have lower error values across all metrics than the corresponding land area-based models, which demonstrates the value that high resolution building maps can bring to parcel-based areal interpolation.

Model	RMSE	MAE	Adj. RMSE	MAPE	MedAPE	PWMAE
res-bld	461.110	298.210	1.350	0.250	16.250	24.310
wrk-bld	641.520	201.320	8.310	1.200	37.900	104.660
res-land	998.580	683.930	2.320	0.540	39.110	64.230
wrk-land	791.600	272.680	11.290	1.550	54.190	198.740

Table 3: Error metrics for residential and worker model predictions at the block group level.

4 Conclusion and future directions

We have demonstrated some initial steps toward incorporating national parcel data and national building data into an areal interpolation model for mapping population, and we have done so not just for residential population but also for worker populations. We consider the work described here to be merely a first proof of concept of using parcel data in combination with high resolution building outline data. From here, a number of lines of inquiry may follow. We intend to work with other modeling approaches beyond the global poisson regression approach described above. Especially interesting to consider will be approaches that do not estimate global coefficients, but allow locally varying coefficients. Applications of quantile regression and geographically weighted regression to the areal interpolation problem have been demonstrated (Cromley et al., 2012, 2013; Lin and Cromley, 2015) but have only been applied using coarse land cover data. Therefore, it would be worthwhile to explore extending these approaches in the context of parcel-level land use and precise building outlines. Various geographical stratification approaches and/or multilevel/hierarchical models could also be explored in this context. Our treatment of the parcel land uses deserves some additional scrutiny as well. We chose in this analysis to consolidate the land uses rather aggressively—down to just three populated categories for each model. This topic was explored by Jia and Gaughan (2016), who demonstrated that overall error was reduced by consolidating the 25 property types in Alachua County, Florida down to 7 coarser classes. The optimal number of classes will probably depend on the particular data context and modeling approach; ideally, this question can be further explored in conjunction with explorations of different modeling approaches.

5 References

Cromley, R. G., D. M. Hanink, and G. C. Bentley
2012. A Quantile Regression Approach to Areal Interpolation. *Annals of the Association of American Geographers*, 102(4):763–777.

Cromley, R. G., D. M. Hanink, and J. Lin
2013. Developing Choropleth Maps of Parameter Results for Quantile Regression. *Cartographica: The International Journal for Geographic Information and Geovisualization.*

Flowerdew, R. and M. Green
1989. Statistical methods for inference between incompatible zonal systems. *The accuracy of spatial databases*, Pp. 239–247.

Jia, P. and A. E. Gaughan
2016. Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. *Applied Geography*, 66:100–108.

Jia, P., Y. Qiu, and A. E. Gaughan
2014. A fine-scale spatial population distribution on the High-resolution Gridded Population Surface and application in Alachua County, Florida. *Applied Geography*, 50:99–107.

Lin, J. and R. G. Cromley
2015. A local polycategorical approach to areal interpolation. *Computers, environment and urban systems*, 54:23–31.

Maantay, J. A., A. R. Maroko, and C. Herrmann
2007. Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science*, 34(2):77–102.

Mennis, J.
2003. Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer*, 55(1):31–42.

Sridharan, H. and F. Qiu
2013. A Spatially Disaggregated Areal Interpolation Model Using Light Detection and Ranging-Derived Building Volumes. *Geographical Analysis*, 45(3):238–258.

Tapp, A. F.
2010. Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources. *Cartography and Geographic Information Science*, 37(3):215–228.

Yang, H. L., J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri
2018. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2600–2614.