Applied Geography 66 (2016) 100–108

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida

Peng Jia [a, *], Andrea E. Gaughan [b]

[a] Department of Epidemiology and Environment Health, University at Buffalo, Buffalo, NY, USA
[b] Department of Geography and Geosciences, University of Louisville, Louisville, KY, USA

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form
8 November 2015
Accepted 9 November 2015
Available online 11 December 2015

Keywords:
Dasymetric mapping
Land cover
Disaggregation
Parcel
Population
GIS

ABSTRACT

Spatial techniques and fine-scale geographic data may be combined in a variety of innovative ways to serve high-resolution population modeling efforts at local scales, which has been further facilitated by growing computation power and access to open-source spatial data. Previous work has highlighted the importance of a dasymetric approach to produce a parcel-based high-resolution gridded population surface (HGPS). In this study, we investigate the application of land-cover data integrated with the parcel-based HGPS to further improve the accuracy of the HGPS. Consideration is given to twelve combinations made by three land cover strategies (1- no land cover class, 2- five separate classes, and 3- three combined classes) and four property type strategies (1- seven types from an empirical study, 2- eight residential types, 3- seventeen types within Alachua County, and 4- twenty-five types within Florida). Results from different strategies are statistically compared with the most significant combination identified as three combined land-cover classes (heavy vegetation, 0–50% and >50–100% impervious surface) and with seven property types from the empirical study (single family, mobile family, multi-family (≥10 and <10 units), condominiums, mobile homes parks, and homes for the aged). A final data set named the Enhanced HGPS (E-HGPS) is created for Alachua County, Florida, with a distribution of population counts at the scale of individual housing units. This study highlights an innovative approach to incorporating land-cover and parcel data for the purpose of spatial population modeling, and holds potential to broaden the E-HGPS to a state or regional scope.

1. Introduction

The demands for spatially-explicit population products in various fields continue to increase (Gaughan, Stevens, Linard, Jia, & Tatem, 2013; Linard, Gilbert, Snow, Noor, & Tatem, 2012; Tatem et al., 2013) and as a variety of finer geographical data become available to the public, a wider range of options for working with Geographic Information Systems (GIS) in conjunction with census data, the most reliable and authoritative source of demographic facts, closely follows. One common approach for creating gridded population products involves dasymetric mapping, which re-distributes census counts bounded at an administrative level onto higher-resolution spatial units (Jia, Qiu, & Gaughan, 2014; Martin, 2011; Mennis, 2009). Traditionally, choropleth maps were commonly used to visually highlight differences in population counts at the administrative unit level. However, dasymetric mapping has improved on traditional choropleth maps by increasing the spatial variation and accuracy in which data are mapped to a surface (Mennis, 2009). Dasymetric mapping techniques may incorporate various ancillary data and range in sophistication from the use of areal weighting to more involved statistical approaches (Martin, 2011; Mennis, 2009). Continually refining and improving on these methods is important for creating spatially-explicit information about variables of interest (i.e. human population counts) that subsequently inform studies on populations at risk (Tatem et al., 2012), transportation patterns (Linard et al., 2012), healthcare resource allocation (Jia, Xierali, & Wang, 2015), and emergency management (Goodchild & Glennon, 2010).

For applications at regional or global scales, where spatial resolution of the gridded products are typically 100 m or greater, available data sets include the Gridded Population of the World (GPW) (Balk et al., 2006), Global Rural Urban Mapping Project (GRUMP) (CIESIN, 2004), LandScan Global (Dobson, Bright,

* Corresponding author.
E-mail addresses: jiapengff@hotmail.com (P. Jia), aegaughan@gmail.com (A.E. Gaughan).

http://dx.doi.org/10.1016/j.apgeog.2015.11.006
0143-6228/

Coleman, Durfee, & Worley, 2000), LandScan USA model (Bhaduri, Bright, Coleman, & Urban, 2007), and the WorldPop Project (Stevens, Gaughan, Linard, & Tatem, 2015; Tatem et al., 2013). However, for local-scale studies, it may be more appropriate to generate a site-specific gridded population data set that takes advantage of novel data sources such as parcel data (Jia et al., 2014; Maantay, Maroko, & Herrmann, 2007; Xie, 2006).

Parcel boundaries, are a valuable independent data source for revealing the underlying population distribution and assisting with population re-distribution due to its direct relationship with population density (Jia et al., 2014). In a recent study, Jia et al. (2014) produced a High- Gridded Population Surface (HGPS) based on fine-scale parcel data for Alachua County, Florida. The study demonstrates the viability of using fine-scale parcel data to increase the accuracy of the distribution of population counts for specific, local-scale objectives. However, land cover is still recognized as one of the most useful sources of ancillary information for many population products (Leyk, Buttenfield, Nagle, & Stum, 2013; Reibel & Agrawal, 2007; Zandbergen, 2011; Zandbergen & Ignizio, 2010). Accessibility and typically strong correlation of various land covers (e.g. urban/built) to population distributions increases the appeal for integrating land cover as an ancillary data source with dasymetric mapping approaches (McKee, Rose, Bright, Huynh, & Bhaduri, 2015). In this study, it is hypothesized that the combination of parcel information and land cover data may further increase the accuracy of a gridded population surface.

Considerations of estimation error still exist due to variation of the population density on the same type of parcel and perhaps the inherent error in census and/or parcel data collection, leading to uncertainty in accurately redistributing census counts. In other words, once an appropriate HGPS is created from census block groups and aggregated over blocks that are spatially nested within block groups, the aggregated estimates within blocks will have an uncertain level of error. It is important to know how the accuracy of population redistribution may be associated with the known information, which might enable us to better know the limitations of the final population grid, and to potentially avoid or overcome these limitations in similar products in the future.

By adopting different ways of combining parcel with land cover classes, calculating the weight of the combined classes, and based on which disaggregating population into different parcels, the best strategy for improving the parcel-only HGPS (Jia et al., 2014) is statistically identified. If land cover data are included in the best strategy, our first hypothesis could be supported that integration of land cover with parcel data improves the accuracy of the HGPS. In addition, we examined underlying factors associated with the differences between dasymetric results and census counts over blocks, testing our second hypothesis that different demographic and/or parcel proportions within blocks may correlate with varying degrees of population redistribution error.

2. Study region, datasets and methods

Alachua County is located in the north part of Florida, a state that comprises the southeastern panhandle of the U.S (Fig. 1). There are a total of 155 block groups and 7382 blocks in Alachua County. The total population in the county is 247,336, among which 82.1% are over the age of 18, 69.6% are White, 20.3% are Blacks, 8.4% is Hispanic, and 5.4% are Asian (U.S. Census Bureau, 2011).

2.1. Datasets

The U.S. Decennial Census data comprise three spatial aggregation levels based on administrative units and the total population counts within them (census tract, block group, and block). The most recent 2010 census population counts at the block and block group levels are used. The parcel data contains the boundaries of parcels in all 67 counties of Florida with associated tax information including the property types of parcels (Florida Department of Revenue, 2010). Although the parcel data in 2012 are available, the data in 2010 are used for a temporal match with the Census 2010.

The National Land Cover Database (NLCD), with a spatial resolution of 30 m, is the most commonly used derived land-cover classification data source in the United States (Fry et al., 2011), where four land-cover categories are defined by the percentage of impervious surfaces, to depict most of the populated areas, including open space (<20%), low intensity (20–50%), medium intensity (>50–80%), and high intensity (>80–100%). Open space regions include areas with some mixed constructed materials, but mostly with vegetation in the form of lawn grasses, such as single family homes, golf courses, parks, etc. Low and medium intensity areas are both composed of constructed materials and vegetation with various extents, where single family is the major type of housing units. High intensity areas primarily consist of highly developed areas associated with increased population densities. The spatial patterns of four land-cover classes in Alachua County are showed in Fig. 1.

2.2. Modeling approach

We adopt the empirical sampling procedure described by Mennis (2003), where census blocks that are covered by primarily one type land cover are used as population density training samples. Due to lack of census blocks completely covered by one property-type, which is required by traditional empirical sampling, a concept of *eligible mono-type block* is used for this analysis. With eligible mono-type blocking, the area proportion of only one dominating property-type is larger than 10% and the total proportion of the remaining property-types is less than 0.1% (Jia et al., 2014). In this study, land-cover classes from the NLCD are used to refine the existing parcel-based HGPS, which means that land-cover categories are used to subdivide given property-types into several subtypes. Here a *subtype* is defined as a combination of a given property-type and a given land-cover category, which applies to all subsequent appearances of the word “subtype” in the rest of this paper. If the centroids of all the parcels in an eligible mono-type block are located in the same land-cover category, that eligible mono-type block is further defined as an eligible mono-*sub*type block. This level of detail in designating potential surface areas for population density training is important for improving model estimates at a fine scale.

Seven property-types presented in Alachua County (Jia et al., 2014), including single family, mobile family, multi-family (≥10 and <10 units), condominiums, mobile homes parks, and homes for the aged, are predefined as residential property types, and from these eligible mono-type blocks are selected. Representative centroids of all residential parcels are extracted and superimposed on the NLCD layer, so as to assign a land-cover category to each residential parcel centroid. Four land-cover categories are initially defined as populated and coded as Class 21 (open space), 22 (low intensity), 23 (medium intensity) and 24 (high intensity). It is worth noting that while matching parcel centroids with land-cover categories, the centroids of some residential parcels, mostly single family, are located in other natural land cover classes such as shrub and woody wetlands rather than the four populated land-cover categories. This might be attributed to either the unavoidable misclassification in NLCD, heavy coverage of vegetation around the parcels, or mismatching between parcel and land cover data. All these parcels are allocated as the fifth populated land-cover

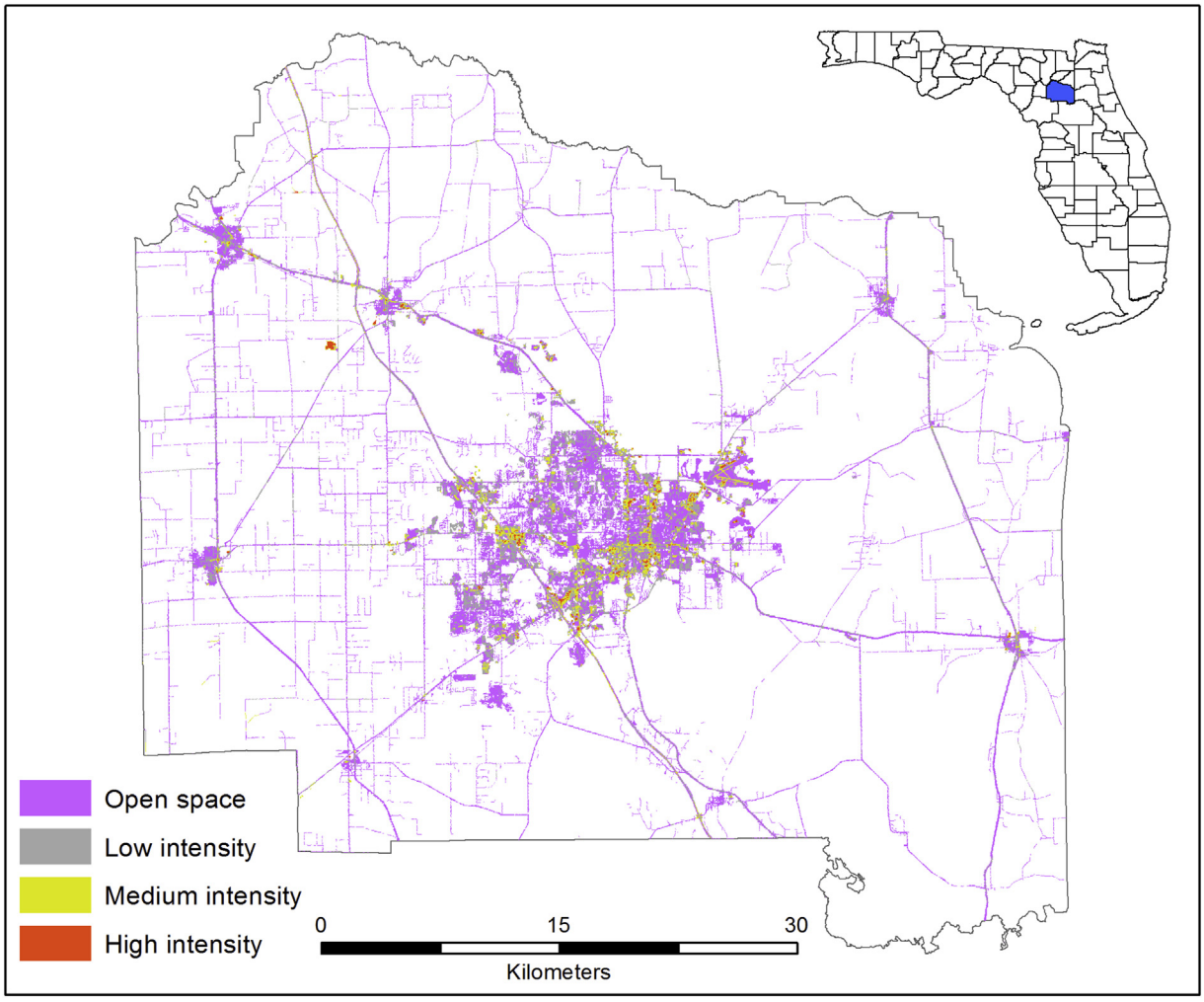

Fig. 1. Alachua County, Florida. Four land-cover categories are defined by the percentage of impervious surfaces: open space (<20%), low intensity (20–50%), medium intensity (>50–80%), and high intensity (>80–100%).

category, coded as Class 25.

Each of five populated land-cover categories are initially considered as separate categories. Each eligible mono-type block with all included parcel centroids located in the same land-cover category is assigned to that subtype (*eligible mono-subtype block*). Three eligible mono-subtype blocks are considered as the minimum number for calculating the population density for any subtype. We also consider the aggregation of the four initial populated categories, merging the data into two coarse-level categories, Class 21-22 and Class 23-24, in order to enhance the numbers of eligible mono-subtype blocks for more trustworthy population density counts over subtypes.

The population density for each residential subtype with at least three eligible mono-subtype blocks or merged subtype is calculated:

$$\rho_v = \sum_{b \in B} P_{vb} \Big/ \sum_{b \in B} A_{vb} \quad (1)$$

where ρ_v = aggregate population density of (merged) subtype v, P_{vb} = population count in block b dominated by (merged) subtype v, A_{vb} = total area of the parcels in (merged) subtype v within block b, and B encompasses all eligible mono-subtype blocks for (merged) subtype v across the Alachua County. The entire process of the dasymetric population redistribution is outlined in Fig. 2. More details can be found in the study of Jia et al. (2014).

2.3. Sensitivity testing

For testing the sensitivity of the results to the selection of population density and residential property-types, we 1) increase the minimum sufficient number of eligible mono-subtype blocks to five and ten versus three, 2) use eight officially defined residential property-types populated in Alachua County (Florida Department of Revenue, 2010), 3) use 17 property-types with at least three eligible mono-type blocks in Alachua County, and lastly, 4) use 25 property-types with at least three eligible mono-type blocks for all of Florida (Table 1). Empirical sampling is undertaken statewide for calculating the population density for those property-types without three eligible mono-type blocks within Alachua County. The property-types and land-cover categories involved in different strategies are summarized in Table 2.

The population counts in resulting grid cells from different strategies are each re-aggregated at the block level and compared to the original census population counts for each block. The absolute value of the difference between re-aggregated and original census counts over blocks is termed as absolute raw error (ARE). The results from different strategies are compared with one

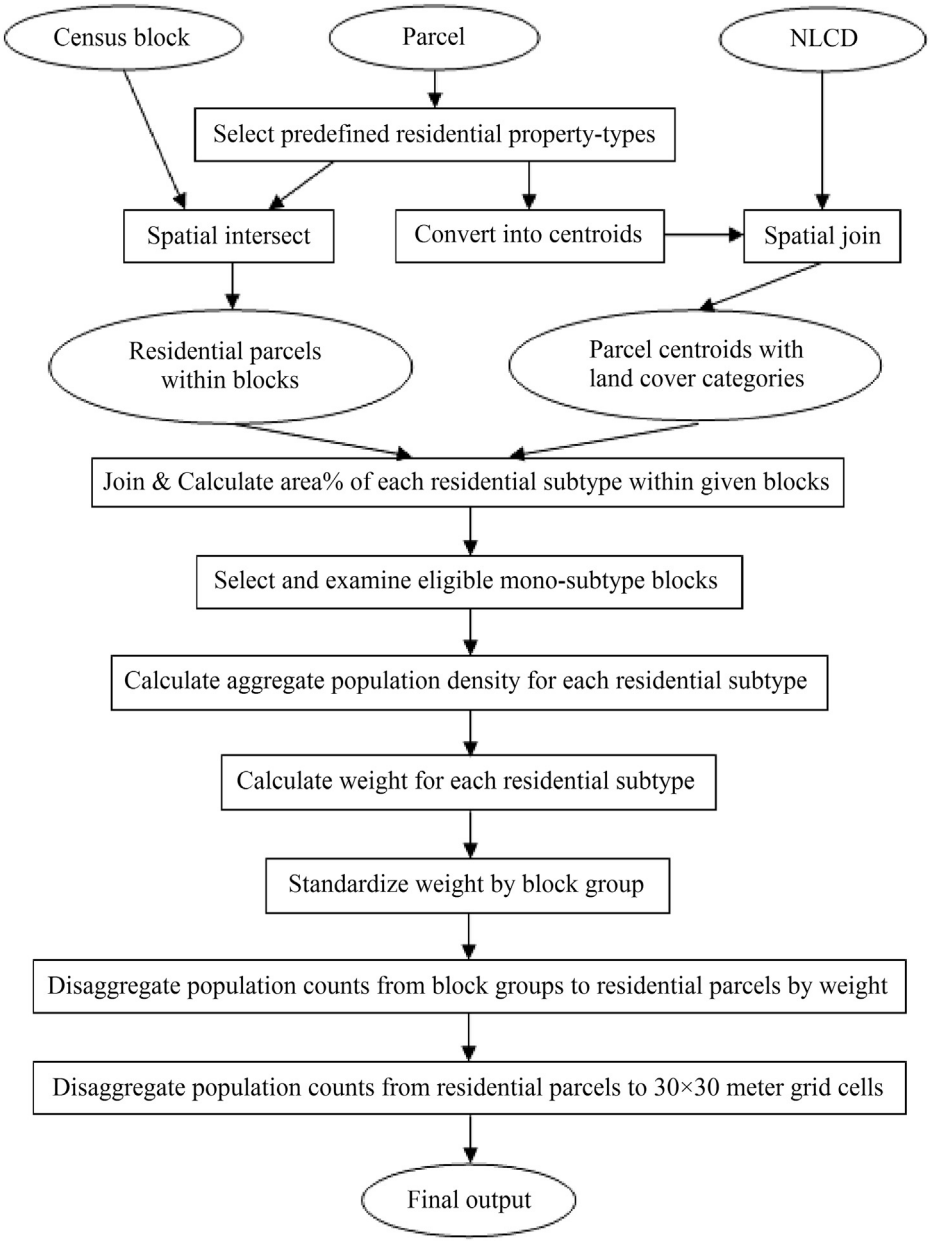

Fig. 2. A flowchart of dasymetric modeling showing all the steps of population being disaggregated from block groups to 30 × 30 m grid cells.

another by calculating and comparing total absolute error (TAE) over blocks, and root mean square error (RMSE) and coefficient of variance (CV) over block groups (Jia et al., 2014). Lower values for TAE, RMSE and CV represent better results. After a natural log transformation is conducted to alleviate the skewness of the distribution of two groups of raw CVs, a *t*-test is used for comparing the best resulting output with the original HGPS, in order to determine if the improvement by integrating NLCD is significant, or if the null hypothesis (H_0) of no difference between the two groups of CVs can be rejected (Jia et al., 2014; Tapp, 2010). An alternate hypothesis (H_1) is that the log-transformed CVs of the E-HGPS are lower than those of the HGPS.

2.4. Factors associated with population redistribution

In this study we hypothesize that varying degrees of the error of population re-distribution could be accounted for by the demographic and/or parcel proportions within different blocks. The relative degree of the error of redistribution is the response variable, calculated by dividing absolute raw error (ARE) by census counts within blocks over 4504 populated blocks in Alachua County (with non-zero census counts). The selection of explanatory variables is based on availability of demographic features in the U.S. 2010 Census, including the percent ethnicity, percentage of the population over the age of 18, and occupied housing units within blocks (Table 3). Other explanatory variables calculated during the dasymetric process include the area ratio of each property-type over blocks, which may shed light on the degree of stability/instability of population density over some types of property. To explore the correlations between the response variable and each explanatory variable, we use the Spearman's rank correlation coefficient (Daniel, 1990). The Spearman's rank correlation is a measure of the directional association between two variables and was deemed appropriate due to the non-Gaussian distribution of the

Table 1
Basic statistics and aggregate population density for all the property types involved in all 16 strategies.

Code	Property type (subtype)	Eligible *mono-subtype* block	Summed population	Summed area (km^2)	Aggregate density[e] (persons/ha)
001[a]	Single family	904	28,183	17.13	16
	001&21-22	787	26,644	14.58	18
	001&21	395	9966	6.27	16[d]
	001&22	41	1004	0.41	24[d]
	001&23-24	4	114	0.03	45
	001&23	4	114	0.03	45[d]
	001&24	–	–	–	[45][d]
	001&25	103	1090	2.41	5[d]
002[a]	Mobile home	43	272	1.08	3
	002&21-22	8	134	0.26	5
	002&21	7	113	0.25	4[d]
	002&22	–	–	–	[5][d]
	002&23-24	–	–	–	[3]
	002&23	–	–	–	[3][d]
	002&24	–	–	–	[3][d]
	002&25	35	138	0.82	2[d]
003[a]	Multi-fml (≥10)	132	12,657	1.50	85
	003&21-22	22	2167	0.30	72
	003&21	5	594	0.13	46[d]
	003&22	16	1485	0.16	91[d]
	003&23-24	15	2150	0.28	78
	003&23	14	1338	0.24	55[d]
	003&24	1	812	0.03	[78][d]
	003&25	94	8154	0.91	90[d]
008[a]	Multi-fml (<10)	12	461	0.09	52
	008&21-22	11	458	0.09	52
	008&21	4	103	0.03	39[d]
	008&22	2	60	0.01	[52][d]
	008&23-24	–	–	–	[52]
	008&23	–	–	–	[52][d]
	008&24	–	–	–	[52][d]
	008&25	1	3	0.001	[52][d]
004[a]	Condominiums	21	1134	0.06	176
	004&21-22	16	684	0.05	144
	004&21	7	314	0.02	152[d]
	004&22	2	29	0.002	[144][d]
	004&23-24	–	–	–	[176]
	004&23	–	–	–	[176][d]
	004&24	–	–	–	[176][d]
	004&25	3	281	0.01	272[d]
028[b]	Mobile parks	63	2505	0.77	32
	028&21-22	10	456	0.14	33
	028&21	4	106	0.04	29[d]
	028&22	6	350	0.10	34[d]
	028&23-24	–	–	–	[32]
	028&23	–	–	–	[32][d]
	028&24	–	–	–	[32][d]
	028&25	53	2049	0.63	32[d]
074[b]	Aged Home	9	862	0.40	22
	074&21-22	2	156	0.15	[22]
	074&21	2	156	0.15	[22][d]
	074&22	–	–	–	[22][d]
	074&23-24	1	117	0.007	[22]
	074&23	1	117	0.007	[22][d]
	074&24	–	–	–	[22][d]
	074&25	6	589	0.24	25[d]
000[a]	Vacant	4	34	0.04	8
007[a]	Miscellaneous	2	13	0.02	6
009[a]	Undefined	11	75	0.05	16
012[b]	Mixed use	4	14	0.02	8
071[b]	Church	6	23	0.06	4
080[b]	Undefined	8	599	0.26	23
084[b]	College	13	3707	3.36	11
	084&21-22	2	583	0.32	18
	084&21	2	583	0.32	18
	084&25	11	3124	3.04	10
086[b]	County	9	178	1.74	1
	086&21-22	4	167	0.05	36
	086&21	2	48	0.02	25
	086&22	1	29	0.008	37
	086&25	5	11	1.70	6
087[b]	State	14	3069	3.30	9
	087&21-22	2	1952	0.66	30
	087&21	1	3	0.49	0.1
	087&25	12	1117	2.65	4

Table 1 (*continued*)

Code	Property type (subtype)	Eligible *mono-subtype* block	Summed population	Summed area (km^2)	Aggregate density[e] (persons/ha)
089[b]	Municipal	3	166	0.04	40
010[c]	Vacant	39	1097	0.45	24
011[c]	Store	15	1011	0.2	51
018[c]	Office building	5	176	0.09	19
050[c]	Agricultural	124	535	21.15	30
072[c]	Private school	7	1817	0.21	87
075[c]	Charitable	1	1	0.0002	46
083[c]	Public school	7	162	0.24	7
088[c]	Federal	79	4345	18.93	2

[ ] Population density replaced with that of the coarser-level class.
[a] Residential property-types officially defined and also populated in Alachua County.
[b] Non-residential property-types but populated in Alachua County.
[c] Non-residential property-types populated in Florida instead of Alachua County.
[d] Population density of each subtype used in the subgroup 1.3.
[e] Aggregated density (unit: persons/ha) = summed population/(summed area × 100).

Table 2
Descriptions of 16 strategies for calculating population density.

Group	Description
1.1	Using 7 residential property-types[a]
1.2	Using 7 residential property-types and 5 land-cover categories; the minimum sufficient number of eligible mono-subtype blocks is 3
1.3	Using 7 residential property-types and 3 merged land-cover categories; the minimum sufficient number of eligible mono-subtype blocks is 3
2.1	Using 7 residential property-types (same as 1.1)
2.2	Using 7 residential property-types and 5 land-cover categories; the minimum sufficient number of eligible mono-subtype blocks is 5
2.3	Using 7 residential property-types and 3 merged land-cover categories; the minimum sufficient number of eligible mono-subtype blocks is 5
3.1	Using 7 residential property-types (same as 1.1)
3.2	Using 7 residential property-types and 5 land-cover categories; the minimum sufficient number of eligible mono-subtype blocks is 10
3.3	Using 7 residential property-types and 3 merged land-cover categories; the minimum sufficient number of eligible mono-subtype blocks is 10
4.1	Using 8 residential property-types[b]
4.2	Using 8 residential property-types and 5 land-cover categories
4.3	Using 8 residential property-types and 3 merged land-cover categories
5.1	Using all 17 populated property-types in Alachua[c]
5.2	Using all 17 populated property-types in Alachua and 5 land-cover categories
5.3	Using all 17 populated property-types in Alachua and 3 merged land-cover categories
6.1	Using all 25 populated property-types in Florida[d]
6.2	Using all 25 populated property-types in Florida and 5 land-cover categories
6.3	Using all 25 populated property-types in Florida and 3 merged land-cover categories

[a] 001, 002, 003, 004, 008, 028 and 074 in Table 1.
[b] Eight property-types with a superscript of 1 in Table 1.
[c] All property-types with a superscript of 1 or 2 in Table 1.
[d] All property-types in Table 1.

data. The statistical test identifies the strength of a relationship between the error of redistribution and various explanatory variables. We also explored the use of the alternative technique (Kendall tau rank correlation coefficient) which produced similar results and thus present findings only for the Spearman's rank test.

3. Results

3.1. Population density

Table 1 shows that integrating land-cover categories with residential property-types has the largest influence on the population density in single family (ranging from 16 persons/ha in open space to 45 persons/ha in medium intensity), multi-family with more than 10 units (ranging from 46 persons/ha in open space to 91 persons/ha in low intensity), and condominiums (ranging from 144 persons/ha in low intensity to 272 persons/ha in other natural land types). Most of the parcels in mobile home, multi-family (<10 units), mobile parks and aged home are located in low intensity regions. Therefore, the population density in these property-types is stable over all land-cover types.

When the number of eligible mono-subtype blocks in a given subtype is less than three, the population density of its merged subtype (the coarser-level class) is substituted. If there are still insufficient eligible mono-subtype blocks for the merged subtype, the general population density of that property type, without being split into land-cover categories, is substituted. For example, in Table 1, the population density in multi-family (≥10 units) in Class 23-24 (003 & 23-24) was substituted for that in multi-family (≥10 units) in Class 24 (003 & 24), as there were less than three eligible mono-subtype blocks in that class. An exception, however, was that if less than three eligible mono-subtype blocks existed in the mobile home Class 23–24 (002 & 23-24), then the population density in mobile home (002), regardless of land-cover classes, was substituted for that in Class 23 (002 & 23), 24 (002 & 24) and 23–24 (002 & 23-24).

3.2. Statistical comparison among outputs

The first subgroup in each group (*x*.1), generated based on property-types without mixing with any land-cover classes, is considered as a “control” group in contrast to the other “case”

Table 3
Strength of the relationship between estimation error and different demographic and parcel components (explanatory variables) over 4504 populated blocks.

Variable	Description	Spearman's
Block (**obtained from census**, **unit**: **%**)		
Pct_over18	Percentage of the population over the age of 18	0.113***
Pct_occu	Percentage of housing units occupied	0.059***
Ethnicity (**obtained from census**, **unit**: **%**)		
Pct_WHITE	Percentage of Whites	0.065***
Pct_BLACK	Percentage of Blacks	−0.086***
Pct_HISPANIC	Percentage of Hispanics	−0.156***
Pct_ASIAN	Percentage of Asians	−0.082***
Pct_AMERI	Percentage of American Indians/Alaska Natives	−0.083***
Pct_HAWN	Percentage of Hawaiians/Pacific Islanders	−0.036*
Ratio_SGL	Area ratio of single family	−0.283***
Ratio_MLT	Area ratio of multi-family (≥10 units)	0.005
Ratio_MLTL	Area ratio of multi-family (<10 units)	−0.104***
Ratio_CDMN	Area ratio of condominiums	−0.022
Ratio_MBL	Area ratio of mobile homes	0.045**
Ratio_MBLP	Area ratio of mobile home parks	0.013
Ratio_RTM	Area ratio of retirement homes	−0.009
Ratio_AGE	Area ratio of homes for the aged	−0.013
Ratio_MSLN	Area ratio of miscellaneous residential property	−0.009
Ratio_UDF	Area ratio of undefined residential property	0.037*
Ratio_VCTR	Area ratio of vacant residential property	0.001
Ratio_VCTC	Area ratio of vacant commercial property	0.013
Ratio_VCTD	Area ratio of vacant industrial property	0.027
Ratio_VCTT	Area ratio of vacant institutional property	0.015
Ratio_VCTG	Area ratio of government property	0.013
Ratio_STOR	Area ratio of stores	−0.019
Ratio_MIX	Area ratio of mixed use buildings	−0.014
Ratio_OFC	Area ratio of office buildings	−0.002
Ratio_CHUR	Area ratio of churches	−0.018
Ratio_PRIS	Area ratio of private schools	−0.014
Ratio_PUBS	Area ratio of public schools	−0.010
Ratio_COLG	Area ratio of colleges	0.072***
Ratio_PRIH	Area ratio of private hospitals	0.020
Ratio_PUBH	Area ratio of public hospitals	−0.001
Ratio_CHAR	Area ratio of charitable property	−0.020
Ratio_SANI	Area ratio of sanitariums	−0.026
Ratio_AGRI	Area ratio of agricultural property	−0.012
Ratio_COUT	Area ratio of county property	−0.008
Ratio_STAT	Area ratio of state property	0.035*
Ratio_FEDE	Area ratio of federal property	0.006
Ratio_MUNI	Area ratio of municipal property	−0.003

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

subgroups (*x*.2 and *x*.3) in the same group, where *x* represents 1–6 (Table 4). The second subgroup (*x*.2) uses five individual categories (Class 21, 22, 23, 24, and 25) while the third subgroup (*x*.3) uses three merged categories (Class 21–22, 23–24, and 25). The descriptive statistics and aggregate population density for all the property types and subtypes (with three or more eligible mono-subtype blocks in Alachua County) involved in all 16 strategies are demonstrated in Table 1.

Sixteen strategies for population density settings are separately used to generate the HGPSs in Alachua County, among which Group 1.1 is the original HGPS from Jia et al. (2014). After disaggregating the population counts from block groups to grid cells and re-aggregating them on block level, the TAEs between re-aggregated and original census counts over blocks are calculated and compared with one another under five categories, including all the blocks (7382), unpopulated (2878), populated (4504), residential (5203) and non-residential blocks (2179) (Table 4).

Comparing six groups as a whole, we find that the TAEs in Group 1–3 are generally smaller than Group 4–6, among which the subgroup 1.3 outperforms all other subgroups in all types of blocks. An exception is when compared to the original HGPS (Group 1.1) in unpopulated blocks. This confirms that using the seven property-types is a better selection than using all officially defined residential property-types (Group 4) that are additionally superior to the inclusion of non-residential property-types (Group 5 and 6). Substituting statewide population density for missing countywide population density further increases the error in Group 5, compared to Group 6. The second (*x*.2) and third subgroups (*x*.3) both generate smaller TAEs than the first subgroups (*x*.1) in all six groups, which suggests that land-cover categories do bring improvement to the parcel-based population products. The third subgroups particularly outperform the second subgroups in all six groups, which indicates merging land-cover categories is a better way to integrate the land-cover categories from NLCD with parcel data than individual land-cover categories. The mean RMSEs and CVs consistently demonstrate the same findings (Table 5). The final output of dasymetric mapping using the method with the best accuracy is subgroup 1.3 and is named the *Enhanced HGPS* (E-HGPS).

The log-transformed CVs of the E-HGPS (subgroup 1.3) were statistically compared with those of the original HGPS (subgroup 1.1) by a simple comparison of means *t*-test. There is a difference in the log-transformed CVs for the E-HGPS ($M = -0.227$, $SD = 0.785$) and HGPS ($M = -0.122$, $SD = 0.823$); $t(308) = 1.357$, $p = 0.176$ (two-tailed). A one-tailed significance level of 0.088, although not significant enough at a 95% confidence level, still allows us to reject the null hypothesis, and accept the alternate hypothesis at a 90% confidence level that the log-transformed CVs of the E-HGPS are

Table 4
Total absolute error (TAEs) between re-aggregated and original census counts.

Group	All	Unpopulated	Populated	Residential	Non-residential
1.1	133,641	7208	126,434	120,325	13,316
1.2	120,096	7323	112,774	106,780	13,316
1.3	**118,133**	**7263**	**110,870**	**104,817**	**13,316**
2.1	133,641	7208	126,434	120,325	13,316
2.2	120,182	7333	112,849	106,866	13,316
2.3	118,372	7284	111,089	105,056	13,316
3.1	133,641	7208	126,434	120,325	13,316
3.2	119,864	7354	112,509	106,548	13,316
3.3	118,600	7287	111,313	105,284	13,316
4.1	137,417	7631	129,786	120,586	16,831
4.2	127,426	8336	119,090	110,595	16,831
4.3	125,413	8233	117,180	108,582	16,831
5.1	153,040	19,797	133,243	129,157	23,883
5.2	148,112	21,008	127,104	123,635	24,477
5.3	146,053	20,860	125,193	121,608	24,445
6.1	161,809	22,450	139,359	136,134	25,675
6.2	159,037	23,925	135,112	132,459	26,578
6.3	156,793	23,716	133,077	130,321	26,472

The Group 1.3 produced the minimum total absolute error (TAE) of population estimation.

Table 5
Comparison of the accuracy of high-resolution population products generated by different strategies.

Group	Mean RMSE	Mean CV	Median CV
1.1	64.40	1.22	0.90
1.2	60.26	1.10	0.85
1.3	59.22	1.08	0.72
2.1	64.40	1.22	0.90
2.2	60.15	1.10	0.85
2.3	59.27	1.08	0.86
3.1	64.40	1.22	0.90
3.2	59.93	1.10	0.86
3.3	59.31	1.08	0.86
4.1	65.51	1.24	0.95
4.2	61.50	1.16	0.89
4.3	60.59	1.14	0.89
5.1	69.92	1.49	1.13
5.2	68.80	1.48	1.16
5.3	67.53	1.46	1.13
6.1	66.85	1.43	1.11
6.2	64.67	1.41	1.08
6.3	63.53	1.39	1.04

lower than those of the HGPS.

3.3. Error analysis

The signs and values of Spearman's coefficients indicate how and to what extent each explanatory variable is correlated with the estimation error (Table 3). The *area ratio of single family* demonstrates the strongest correlation across all the explanatory variables tested, where the negative sign means the blocks with a higher percentage of single family units normally have a lower percentage of estimation error. Other significant factors showing a relatively stronger correlation with estimation error (Spearman's coefficient >0.1) include the percentages of Hispanics (−0.156) and the population over the age of 18 (0.113), as well as the area ratio of multi-family (<10 units) (−0.104). That is to say, the blocks with a larger percentage of the population over the age of 18 tend to have a higher percentage of estimation error, while the blocks with a higher percentage of Hispanics and a higher area ratio of multi-family (<10 units) normally have a lower percentage of estimation error. The factors showing a weak but significant positive correlation include the percentages of whites, housing units occupied, and the area ratios of mobile homes, colleges, undefined residential property, and state property. The percentages of other races (blacks, Asians, American Indians/Alaska Natives, and Hawaiians/Pacific Islanders) all show a weak but significant negative correlation with estimation error.

4. Discussions

The findings in this study support our two hypotheses that 1) the integration of land cover and parcel-based data further increases the overall accuracy of the gridded population surface, and 2) some demographic and parcel components within blocks are significantly correlated with the error resulting from population redistribution. The best estimated model is subgroup 1.3 (E-HGPS) which includes seven residential property-types (single family, mobile family, multi-family (≥10 and <10 units), condominiums, mobile homes parks, and homes for the aged), three merged land-cover categories (heavy vegetation, 0–50% and >50–100% impervious surface), and requiring a minimum of three eligible blocks for training the population density for each combination of land cover and parcel categories. The study area of this study, comprising 155 block groups in Alachua County, limits to some degree the statistical comparison of the E-HGPS and HGPS due to a small sample size. While the improvement of the E-HGPS over the HGPS is not significant at 95% confidence level, the significance at the 90% confidence level still gives a directional sense to the results., In addition, the findings corroborate the approach to determining the residential property-types involved in the previous study (Jia et al., 2014).

According to the results from the non-parametric correlation approach, the percentage of the population over 18 influences the ratio of estimation error in the positive direction. It might imply that the population per unit of living area for adults is more heterogeneous across the county due perhaps to uneven geographical distribution of population, varied family sizes and/or housing occupancy rate, making it difficult to capture by county-wide empirical sampling. In addition, higher occupancy rates for housing properties increase the uncertainty in population estimation which is not unexpected. Likewise, the area ratio of some property-types over blocks (mobile homes, college, undefined residential property, and state property) is positively correlated with the error ratio, which means that the population density in these property-types is relatively unstable across the county. Added error could also be from larger blocks which include more parcels with more complexity in pattern distribution of population counts. Results from the correlation analysis provide many potential avenues for exploration in future work on both modeling population distribution and inform considerations in collection and processing efforts of both census and parcel data. Some property-types in the sense of utility are not supposed to be habitable, such as private/public school and hospital. However, non-zero population counts have been assigned to a certain number of blocks completely and only including each of those property-types. Hence, despite being excluded from a list of residential property-types, they still negatively influence the assessment of our final product by adding untrue error to the overall error (*Type I error*). Overcoming these issues requires mutual efforts and coordination among relevant governmental agencies.

The accuracy of classification in NLCD and parcel data is a limitation in this study, with the accuracy of the NLCD well noted in other studies (Jia et al., 2014; Smith, Zhou, Cadenasso, Grove, & Band, 2010). Methodologically, assigning all parcels located outside the four primary populated land-covers to one class (Class 25) may result in lower accuracy in the population density of that mixed class, given that a considerable number (42.7%) of the habitable parcels in Alachua County are classified into Class 25. In addition, individual land-cover categories with an approximate percentage of impervious surface are merged at the expense of losing detail in land-cover type. This is due primarily to insufficient numbers of eligible mono-subtype blocks for population density training in Alachua County and could contribute to error in the modeling process that relies on five individual land-cover categories. As such, future work that is able to incorporate a larger number of eligible mono-subtype blocks for each combination of individual land-cover and parcel categories may provide more robust results than the aggregated approach found in this study.

It is worth mentioning that the currently up-to-date NLCD 2011 was not available at the time of conducting this study, but is available now. The differences between NLCD 2006 and 2011 in Alachua County were examined by Kappa statistics (Viera & Garrett, 2005), and they were found to be minimal. Following the superior strategy in this study to consider Class 21 and 22, Class 23 and 24, and all other classes outside of these four classes as three independent categories, 94.7% of 87,753 cells have been stable over time with a weighted kappa of 0.74. According to the percentage of unchanged habitable areas and the associated kappa value, the variations in Alachua County between 2006 and 2011 are assumed not to significantly affect the results in this study.

The substitution of population density from state-wide empirical sampling for missing population density in some property-types in Alachua County lowers the overall accuracy of estimation. This implies that the population density over similar property types across Florida may vary county by county due to geographic location and socioeconomic status. Therefore, despite the potential effectiveness of using a state-wide sampling approach, it should be used with caution. This also applies to considerations in comparing and extending the approach to other states than Florida. For example, property taxes in Florida are some of the highest in the country, which might affect population density over the same acre.

Despite these considerations, this study lays a solid groundwork of better understanding how land cover and parcel data can best be integrated for producing gridded population datasets at a fine scale. Future work will build on this by investigating an increased number of land-cover categories as associated with various residential property-types. In addition, increasing the geographic range of the study will enable a large sample size across combinations of land cover and parcel types. Alternatively, the approach could be tested in a larger urban environment that has a large range in land cover types. Final products from these efforts may be assessed relative to

other datasets at comparable spatial resolutions and updated with more contemporary derived data products as they become available.

In summary, we incorporate land-cover with parcel data through a dasymetric modeling approach to better understand population density distributions across Alachua County. The 30 × 30 m E-HGPS produced in this study arguably increases the accuracy of spatially disaggregated population from using parcel data exclusively, and provides a positive step forward towards better identifying the nuances of intra-class variation within population density counts. The potential uses of this product in practical applications (e.g. emergency management, health resource allocations, etc.) make small increases more important. For example, knowing how many people live in individual housing units is a necessary precondition for measuring health disparities in *per capita* accessibility to health/hospital resources in a road network-based real world. The U.S.-wide coverage of products such as the NLCD product and the growing availability of parcel data across States provide the necessary framework to expand the approaches outlined in this paper to future work at regional and potentially national scales.

References

Balk, D. L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S. I., & Nelson, A. (2006). Determining global population distribution: methods, applications and data. *Advances in Parasitology, 62*, 119–156.

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal, 69*(1–2), 103–117.

CIESIN, C. f. I. E. S. I. N.. (2004). Global rural–urban mapping project (GRUMP), alpha version: urban extents. In *Center for international earth science information network (CIESIN)*. New York: Columbia University.

Daniel, W. W. (1990). Spearman rank correlation coefficient. In *Applied nonparametric statistics* (2nd ed., pp. 358–365). Boston: PWS-Kent.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., & Worley, B. A. (2000). LandScan: a global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing, 66*(7), 849–857.

Florida Department of Revenue. (2010). *User's guide for 2010 department property tax data files*.

Fry, J. A., Xian, G., Jin, S., Dewitz, J. A., Homer, C. G., Yang, L., et al. (2011). Completion of the 2006 national land cover database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing, 77*(9), 858–864.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High resolution population distribution maps for southeast Asia in 2010 and 2015. *PLoS One, 8*(2), e55882.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth, 3*(3), 231–241.

Jia, P., Qiu, Y., & Gaughan, A. E. (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.

Jia, P., Xierali, I., & Wang, F. (2015). Evaluating and re-demarcating the hospital service areas in Florida. *Applied Geography, 60*, 248–253.

Leyk, S., Buttenfield, B. P., Nagle, N. N., & Stum, A. K. (2013). Establishing relationships between parcel data and land cover for demographic small area estimation. *Cartography and Geographic Information Science, 40*(4), 305–315.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS One, 7*(2), e31743.

Maantay, J. A., Maroko, A. R., & Herrmann, C. (2007). Mapping population distribution in the urban environment: the cadastral-based expert dasymetric system (CEDS). *Cartography and Geographic Information Science, 34*(2), 77–102.

Martin, D. (2011). Directions in population GIS. *Geography Compass, 5*(9), 655–665.

McKee, J. J., Rose, A. N., Bright, E. A., Huynh, T., & Bhaduri, B. L. (2015). Locally adaptive, spatially explicit projection of US population for 2030 and 2050. *Proceedings of the National Academy of Sciences, 112*(5), 1344–1349.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*(1), 31–42.

Mennis, J. (2009). Dasymetric mapping for estimating population in small areas. *Geography Compass, 3*(2), 727–745.

Reibel, M., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review, 26*(5–6), 619–633.

Smith, M. L., Zhou, W., Cadenasso, M., Grove, M., & Band, L. E. (2010). Evaluation of the national land cover database for hydrologic applications in urban and suburban Baltimore, Maryland 1. *JAWRA Journal of the American Water Resources Association, 46*(2), 429–442.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS One, 10*(2), e0107042.

Tapp, A. F. (2010). Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science, 37*(3), 215–228.

Tatem, A. J., Adamo, S., Bharti, N., Burgert, C. R., Castro, M., Dorelien, A., et al. (2012). Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation. *Population Health Metrics, 10*(1), 8.

Tatem, A. J., Gaughan, A. E., Stevens, F. R., Patel, N. N., Jia, P., Pandey, A., et al. (2013). Quantifying the effects of using detailed spatial demographic data on health metrics: a systematic analysis for the AfriPop, AsiaPop, and AmeriPop projects. *The Lancet, 381*, S142.

U.S. Census Bureau. (2011). *2010 census of population and housing, demographic profile summary file: Technical documentation*.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine Journal, 37*(5), 360–363.

Xie, Z. (2006). A framework for interpolating the population surface at the residential-housing-unit level. *GIScience & Remote Sensing, 43*(3), 233–251.

Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS, 15*(s1), 5–27.

Zandbergen, P. A., & Ignizio, D. A. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science, 37*(3), 199–214.