Identifying Residential Land in Rural Areas to Improve Dasymetric Mapping

Stefan Leyk*, Matthew Ruther*, Barbara P. Buttenfield*, Nicholas N. Nagle and Alexander K. Stum***

* Geography Department, University of Colorado, Boulder, Colorado 80309 USA, {stefan.leyk, matthew.ruther, babs, alexander.stum}@colorado.edu
** Department of Geography, University of Tennessee, Knoxville, TN 37996 USA, nnagle@utk.edu

Abstract. In most landcover datasets, the classification of residential land in urban areas demonstrates higher accuracy, relative to that in rural areas. This research focuses on identifying dasymetric relationships between residential land and different ancillary (related) variables in rural regions to improve the classification of residential developed land. Seven ancillary variables are tested including slope and terrain roughness, distance to water and road density. Results show that for this rural region, the two most informative ancillary variables are slope and distance to roads.

Keywords: daysmetric mapping, small area estimation, rural areas, landcover

1. Introduction

Estimates of demographic characteristics are often needed at finer spatial scales than are provided by a national census. Finer resolution demographic estimates may be obtained through dasymetric mapping (Semenov-Tian-Shansky, 1928; Wright, 1936), a special type of areal interpolation that makes use of ancillary data. Dasymetric mapping partitions the population surface of the study area into homogeneous zones, minimizing variation in demographic characteristics within each zone and showing steepest changes in values at zone boundaries (Mennis, 2009). Dasymetric models have been used successfully in crime analysis (Poulsen and Kennedy, 2004), environmental risk assessments (Maantay and Maroko, 2009; Giordano and Cheever, 2010), and environmental health research (Maantay et al., 2008).

Descriptions of various kinds of ancillary data include landcover (Mennis, 2003), road density (Reibel and Bufalino, 2005), remote sensing data (Yuan et al., 1997), parcel data (Tapp, 2010) and address points (Zandbergen and Ignizio, 2010). Two basic types of ancillary variables are commonly considered: Limiting variables, which eliminate from the study area regions that are unpopulated (e.g. bodies of water, parks); and related variables, which delineate complex relationships within population data to curtail or amplify demographic estimates at finer spatial resolution.

Dasymetric mapping results for urban areas demonstrate higher accuracy, relative to results from rural areas (Eicher and Brewer, 2001; Zandbergen and Ignizio, 2011). This is particularly true when using ancillary data such as the National Landcover Database (NLCD) (Fry et al., 2011), which traditionally reflects urban residential areas quite reliably (Mennis, 2003) but provides less reliable identification of residential locations in rural regions. For example Wickham et al. (2013) report an overall accuracy for NLCD 2006 of 78% (with a regional range from 60% to 90%) with varying classification accuracy across developed classes. There are two reasons for this shortcoming of the NLCD.

First, the detection of rural developed land (small farms or residences) in remote sensing data is difficult due to mixed pixels and a lack of discriminatory power in existing detection approaches, and therefore shows higher misclassification rates (Smith et al., 2002). Second, census data are highly aggregated in rural areas due to low population densities. These limitations make dasymetric refinement less effective in rural areas, introducing bias into studies of landcover change and rural residential patterns that often rely on patterns directly derived from landcover databases (Irwin and Bockstael, 2007). To advance knowledge of human development and rural occupancy at very fine scales, accurate fine-resolution data are in urgent demand.

This paper will focus on dasymetry in rural areas by examining the predictive power of different ancillary (related) variables to improve the spatial precision of population estimation. A set of variables are spatially overlaid to examine their potential for refining residential land in rural regions. Based on this exploratory analysis statistical models are developed that predict the most probable locations of residential developed land in rural areas, to augment and improve dasymetric mapping techniques. A baseline approach is presented to potentially improve small area population estimates and dasymetric mapping, through the identification of areas where the rural population is likely to reside.

2. Data and Data Processing

2.1. Independent Variables

The National Elevation Data (NED), available at 30m resolution from the U.S. Geological Services (USGS), are used to compute slope, terrain roughness and curvature. The USGS National Hydrography Database, compiled for use at a scale of 1:24,000, is used to calculate distance to water. Road network data from the Colorado Department of Transportation (also compiled for use at 1:24,000) are used to compute path distance to roads as well as road density measures (within a 250x250m neighborhood). These seven spatial measures encompass the initial set of related ancillary variables. The study area includes rural land that is privately owned in Boulder County, Colorado, a university town roughly 70 km northwest of Denver. Private lands were identified using federal land (included in USGS' *National Map*), open space extents and city limits (both included in Boulder County data). These data layers serve as limiting variables, delineating rural areas where no residential parcels can be expected, prior to model fitting (*Figure 1*). The study encompasses a total of 605.1 square kilometers (233.6 square miles).

2.2. Dependent Variables

Parcel data from Boulder County for the year 2008 were obtained from the Boulder County Assessor's Office, and include geometry and attributes. The data are filtered by the 'land use' attribute to generate a parcel dataset consisting solely of residential units. A parcel is identified as residential if it is categorized as residential-use or multi-use residential. In rural areas, nearly all parcels are defined as single-family residences and can have considerable spatial extents. Parcel data are not available for all areas in the U.S., and are often expensive to obtain and labor-intensive to process. But they often represent the most reliable data source for identifying residential land in rural regions.

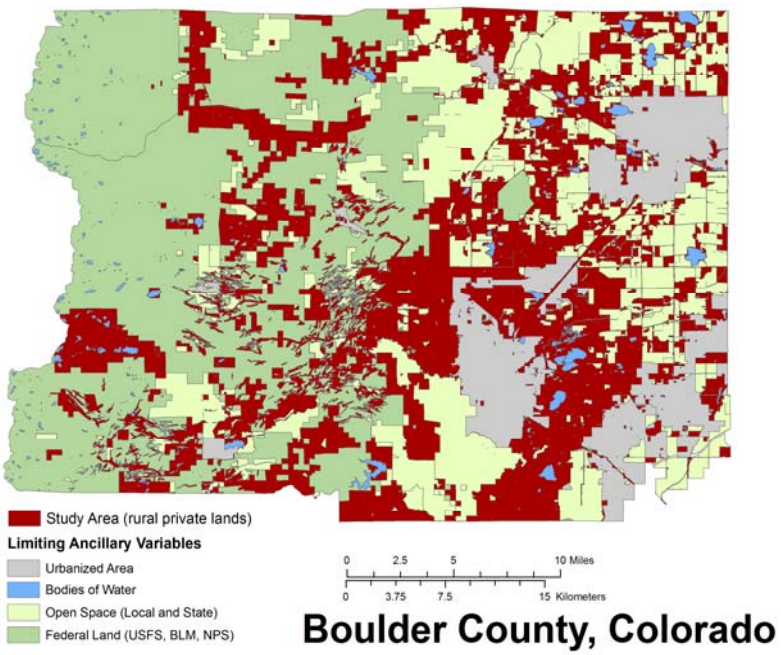

Figure 1. The study area is rural privately owned land parcels.

3. Methods

A two-step approach is carried out in this paper. In the first step the spatial variables described above are systematically explored to examine whether they may be spatially related to residential parcels. This exploratory step is based on the creation of binary criteria grids for each of the independent variables. These criteria are subsequently included in different spatial overlay operations that generate outcome distributions using composite criteria combining the spatial variables. The results of these overlay operations are then linked with residential parcel data in order to compute the proportion of parcels that intersect with the identified residential area. In addition, these binary criteria grids are used to calculate the proportion of total land in the study area that may be suitable for residential development. Together, these two measures are used to assess the benefit of each spatial variable in constraining the search for residential land. The second step will fit a statistical model using the same variables to examine whether the spatial associations identified in the first step can be successfully used to establish statistically significant relationships that would make it possible to predict locations of residential parcel units.

3.1 Exploratory Spatial Overlay

In the first step, frequency distributions of parcel units are plotted as a function of the different independent variables and used to identify relevant ranges of these independent variables in which most residential parcels are located. These ranges are used as attribute criteria for creating the binary grids. For each variable the pixels inside each range are assigned a one (criterion fulfilled). They are assigned a zero at all other locations (criterion not fulfilled), creating Boolean sets in grid form. For example if the majority of parcel units are found at locations between 0 and 16 degrees of slope then this range will be used as the slope-related criterion to create the corresponding Boolean set.

Different overlay operations were tested to relate these seven Boolean grids in different ways. Various combinations of logical union and intersection operators were investigated to mimic different possible relations between the independent variables, each of which may result in different delineation of rural residential areas. For each overlay outcome the proportion of residential parcels overlapping with the area "inside" (combined criteria fulfilled) is calculated. If at least one inside-pixel (of value 1) overlaps with a residential parcel unit this parcel is included in the proportion calculation as a hit. Each parcel identifier is only counted once even if several pixels with the same identifier overlap, as we are interested in the proportion of parcel units, not pixels (or area).

This first exploratory step will inform selection of the predictor variables, how they relate to each other, and whether these relations can help in determining where residential parcels may be located. The advantage of this exploratory approach is that it is data driven and is robust against the existence of non-linear relationships. It thus allows an in-depth examination of locational correspondence between residential parcel area and various combinations of independent variables.

3.2 Sampling and Statistical modeling

The selected range-graded variables will then be tested in a statistical model as independent variables. The statistical model can be conceptually understood as a prediction of the most probable locations of residential land (the dependent variable) using the explanatory power of terrain and distance variables. A Generalized Linear Model (GLM) predicts the likelihood of each rural location inside the study area being residential land. A random spatial sample of 10,000 points was collected from within the study area. Each point was labeled as

present (inside a residential parcel unit = 1) or absent (outside any parcel units = 0) thus creating a binary dependent variable.

GLMs are mathematical extensions of ordinary least-square (OLS) regression models that do not force data into arbitrary scales but allow for non-linearity and non-constant variances (McCullagh and Nelder, 1989). This is an important prerequisite, as the dependent variable used here is bounded and cannot be modeled using OLS. In GLMs the linear combination of the independent variables is related to the mean of the dependent variable μ through a link function which transforms them to linearity and maintains the predicted values within the range of coherent values. In this study the logistic link (often termed logit regression model) is used; it is the logarithm of odds, log(μ/(1-μ)), a model widely used for binomial data (Dobson, 2002).

The general logistic regression model has the form:

$$\text{logit} = \log\left(\frac{\mu}{1-\mu}\right) = \alpha + X^T\beta$$

where α is the regression intercept, X represents a vector of p independent variables (X_1,...,X_p) of any possible power T, and β denotes a vector of p regression coefficients (β_1 ,..., β_p) which are determined for each predictor. The predicted values will be back-transformed (using the inverse logistic transformation) to values between 0 and 1 representing probabilities that a considered pixel belongs to a residential parcel. Thus the models will produce probability surfaces for rural regions of the study area. Coefficients are tested for significance ($p<0.05$) using the Akaike Information Criterion (AIC) to compare different models. The D^2 (or pseudo-R^2) value (percent deviance explained) will be used to evaluate model fit.

4. Results and Discussion Points

Criteria for locations that could be suitable for residential land were first derived from parcel frequency distributions which spanned the observed range of values found for each potential explanatory variable within residential land (parcels). Each explanatory variable was binned, to reduce the influence of outliers or rare values. Criteria values of 1 were assigned to bins with frequency density greater than the mean frequency density over all bins for each variable. In this way, the most common (or most frequently found) conditions for residential parcels were selected. The ranges of values for different explanatory

spatial variables are shown in *Table 1*, along with the percentage of pixels that fall within these ranges and are inside residential parcel units.

Variable	Criteria Derived Based on Frequency Density	% Pixels in Range and Inside Parcels
Elevation	1525 - 1700, 1950 - 2000, 2300 - 2400, 2450	69.1
Slope	0 - 19	82.8
Curvature	-0.25 - 0.25	77.3
Terrain Roughness	0.6 - 3.2	82.1
Distance from Road	0 - 275	88.1
Road Density	50 - 300	84.5
Distance from Water	0 - 1200	89.1

Table 1. Criteria defined for various explanatory spatial variables.

The ideal explanatory variables will identify with high probability the majority of parcels in the study area (high proportion of unique parcel identifiers), while simultaneously minimizing the total amount of land which falls within its criteria range, thus optimizing the 'refinement effect' for residential land. Therefore the area refined will also be computed for subsequent combined criteria.

Table 2 shows the proportion of parcel units overlapping as well as the percentage area refined for different spatial overlays i.e., various logical combinations between the described criteria that could be meaningful in the context of residential land detection. The results vary greatly depending on the overlay selected. Slope, curvature and roughness show very similar results as single criteria as well as in combination. These variables are highly correlated, and thus in most cases identify the same locations as suitable residential land. Distance from road shows the highest explanatory power for a single variable, as the percent of parcel units identified is closest to 100%. In addition, the percentage of land refined using the derived criteria is only 71%, among the lowest values for all single criteria.

Among the combined criteria the intersection between slope and distance from road resulted in very high proportions of residential parcel units and the greatest refinement (62%) among all tests run. Enforcing that both criteria have to be fulfilled largely maintained the high accuracy shown by both single criteria but could reduce the refined land proportion considerably. Most other combined criteria did not result in considerable improvements compared to the single criteria.

Criteria	% of Parcel Units Identified	% Land Value=1
Elevation = 1	84.4	65.6
Slope = 1	98.6	83.3
Curvature = 1	97.4	80.2
Terrain Roughness = 1	93.5	80.3
Distance from Road = 1	99.6	71.7
Road Density = 1	62.9	79.0
Distance from Water = 1	93.8	92,2
Slope =1 AND Distance from Road = 1	98.2	61.8
Slope =1 OR Distance from Road = 1	100.0	93.2
Distance from Road = 1 AND Road Density = 1	62.5	63.0
Slope = 1 AND Terrain Roughness = 1	93.1	77.6
Slope = 1 OR Terrain Roughness = 1	99.0	86.0

Table 2. Identification of unique parcels by various spatial overlays of criteria grids.

The statistical model results are shown in Table 3. *Figure 2* shows the predicted probabilities from the statistical model overlain with parcel boundaries. Based on the above results and additional correlation testing some variables were excluded to avoid collinearity and overfitting. The model was thus fitted using only elevation, slope, distance to roads and road density as explanatory variables. The coefficients for all variables are highly significant but show unexpected directions for slope and elevation, which is surprising given the results from the spatial overlays described above. A possible explanation is that parcel units in rural regions can become very large and the high degree of variability of the explanatory variables within these large parcels may result in these variables exhibiting less strength in statistically estimating the coefficients.

Variable	Odds Ratio	Standard Error	z-score
Elevation	***1.0005	0.0001	8.57
Slope	***1.0192	0.0029	6.78
Distance to Road	***0.9977	0.0001	-15.60
Road Density	***1.0033	0.0002	14.05
N	10,000		
pseudo R-squared	0.1063		
*** Significant at p<0.001			

Table 3. Odds ratios from logistic regression of spatial explanatory variables on parcel outcome (residential parcel=1, non-residential parcel=0).

To test this theory, the effect of parcel unit size was examined by fitting an identical model restricted to predicting parcels less than 12,190 m^2 in area. The coefficients in this second model show the theoretically expected directions (negative for slope, elevation, and street distance, positive for street density) and are all significant. In addition, the pseudo R-squared, which measures the amount of deviance explained by the model, increases substantially in the size-constrained model. This provides a strong indication that the issue of parcel unit size (area) is critical in the modeling process. This relates to the issue of scale sensitivity of the analytical unit. The smaller the size of a unit, the more representative the underlying environmental conditions are in reflecting suitability for residential land. As mentioned above, large parcels exhibit too much variability in the independent variables and the explanatory power of these variables is thus very weak.

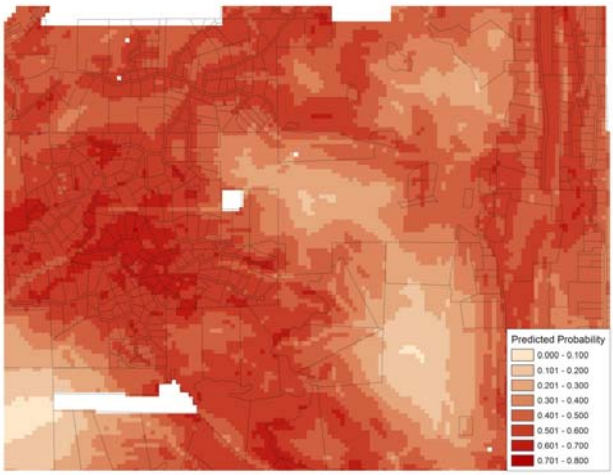

Figure 2. Predicted (Modeled) Probabilities for Residential Parcel Locations.

The spatial overlay experiments and the issue of scale provide important guidance for future improvement of the statistical models and subsequent small area estimation.

5. Further Work

The spatial overlays will be further developed and more advanced weighting schemes will be implemented to allow for more flexible differentiation of

criteria. The statistical models have to be revised and the conceptual approach needs improvement to better reflect the relationships we are looking for and to resolve the issue of scale sensitivity i.e., the area of the analytical units. Upon further refinement of the statistical modeling, these results will be incorporated into dasymetric maps, to examine how small area estimation and dasymetric mapping can be improved in rural regions.

Acknowledgements

This research is funded by the National Science Foundation: “Collaborative Research: Putting People in Their Place: Constructing a Geography for Census Microdata”, project BCS-0961598 awarded to University of Colorado - Boulder and University of Tennessee - Knoxville. A portion of this work is supported by USGS-CEGIS grant # 04121HS039, "Generalization and Data Modeling for New Generation Topographic Mapping".

References

Dobson A. J. (2002) *An Introduction to Generalized Linear Models*. Second Edition, Chapman and Hall/CRC, New York

Eicher, C.L. and Brewer, C.A. (2001) Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science* 28, 2:125-138

Fry, J. Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., and Wickham, J. (2011) Completion of the 2006 National Land Cover Database for the Conterminous United States. *Photogrammetric Engineering and Remote Sensing* 77(9), pp. 858–864

Giordano, A. and Cheever, L. (2010) Using Dasymetric Mapping to Identify Communities at Risk from Hazardous Waste Generation in San Antonio, Texas. *Urban Geography* 31, 5, pp. 623-647

Green, D. M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics*. Wiley, New York

Irwin, E. G. and Bockstael, N. E. (2007) The evolution of urban sprawl: Evidence of spatial heterogeneity and increasing land fragmentation**.** *Proceedings of the National Academy of Sciences of the United States of America* 104, 52, pp. 20672-20677

Maantay, J. A. and Maroko, A. R. (2009) Mapping urban risk: Flood hazards, race, and environmental justice in New York. *Applied Geography* 29, 1, pp. 111-124

Maantay, J. A. Maroko, A. R. and Porter-Morgan, H. (2008) Research Note—A New Method for Mapping Population and Understanding the Spatial Dynamics of Disease in Urban Areas: Asthma in the Bronx, New York. *Urban Geography* 29, 7, pp. 724-738

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. Second Edition, Chapman and Hall, London

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

Mennis, J. (2009) Dasymetric Mapping for Estimating Population in Small Areas." *Geography Compass* 3, pp. 2:727-745

Poulsen, E. and Kennedy, L. W. (2004) Using da symetric mapping for spatially aggregated crime data. *Jounal of Quantitative Criminology* 20 (3), pp. 243-62

Reibel, M. and Bufalino, M. E. (2005) Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A* 27, pp. 127–139

Semenov-Tian-Shansky, B. (1928) Russia: Territory and Population: A Perspective on the 1926 Census. *Geographical Review* 18, pp. 4:616-640

Smith, J. H. Wickham, J.D., Stehman, S. V. and Yang, L. (2002) Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. *Photogrammetric Engineering and Remote Sensing* 68, pp. 65–70

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

Wickham, J. D. Stehman, S.V., Gass, L., Dewitz, J., Fry, J. A., Wade, T. G. (2013) Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sensing of Environment* 130, 15, pp. 294-304

Wright, J.K. (1936) A Method of Mapping Densities of Population: With Cape Cod as an Example. *Geographical Review* 26, 1:103-110

Yuan, Y. Smith, R. M. and Limp, W. F. (1997) Remodeling census population with spatial information from LandSat TM imagery. *Computers, Environment and Urban Systems* 21, pp. 245–258

Zandbergen, P. A. and Ignizio, D. A. (2010) Comparison of Dasymetric Mapping Techniques for Small-Area Population Estimates. *Cartography and Geographic Information Science* 37, 3, pp. 199-214