

Raster-Based Automated Dasymetric Mapping

Torrin Hultgren

University of Colorado, Boulder

Hultgren@Colorado.edu

Introduction

One of the major trends in urban GIScience today is data integration and synthesis through augmentation. Types of data that traditionally were regarded as unrelated, such as remotely sensed imagery and demographic statistics, are now being used complementarily, each to refine and improve the usage of the other. For example, census and zoning data have been integrated into automated image classification decision trees to push beyond the limitations of a purely spectral method (Hutchinson 1982, Mesev 1998). The utility of data integration, however, is an avenue that runs both ways. Demographic data are frequently aggregated into areal units designed to be homogeneous with respect to population characteristics, economic status, and living conditions. Accompanying goals for and restrictions upon these boundaries, such as optimum unit populations regardless of areal extent, the preservation of boundaries over time, and the requirement that units perfectly subdivide larger arbitrary units (such as U.S. census tracts nesting in counties) significantly degrade the ability of such areal units to accurately reflect demographic distribution. Dasymetric mapping utilizes ancillary data to redistribute population data from such arbitrarily delineated enumeration districts into units of increased homogeneity in order to better represent the actual underlying statistical surface.

Applications for such maps include intercensal population estimates, resource management in rapidly growing third world megacities, and even the reconstruction of

demographic distributions from historic air photos or archaeological records. “If remote sensing data are integrated or used in conjunction with other sources of socioeconomic, administrative, and regulatory data, their potential applicability to both research and policy understanding of the urban environment increases significantly.” (Miller *et al*, 2003). Improved resolution of demographics could in turn assist in more accurate classification of remotely sensed images, bringing full circle a feedback loop of refinement.

Most censuses will only publish demographic statistics that have been aggregated across certain areas, both for obvious privacy reasons, and because the near constant change at the housing-unit level makes ensuring the accuracy of finer resolution data nearly impossible. The challenges that these census groupings into blocks and tracts create, however, are numerous and formidable. In outlining the boundaries, the census attempts to form areas that are relatively homogenous in terms of demographic characteristics (socioeconomic racial and housing type) at the time of establishment, as well as approximately equal in population. While these units can be quite functional for certain studies, the assumption of population homogeneity across often very large and arbitrary areas can be both misleading and erroneous.

Numerous methods have been explored to redistribute census counts. Some researchers have employed interpolation techniques to develop a population surface based on weighted distance from tract centroids (Bracken, 1993; Harris and Longley, 2000; Harvey, 2003). While this method does create maps of some utility, it is based on several assumptions that upon further inspection appear shaky. For instance, the assumption that population will decay in any sort of predictable fashion with decreasing distance from a centroid or cluster of centroids might be confounded by, say, the clustering of population along a waterfront at the very edge of an enumeration district. Corrections for such exceptions would likely need to be handled manually

on a district-by-district basis based on first-hand knowledge of the area, a labor intensive process.

To reduce this sort of labor and reliance on uncertain assumptions, researchers have looked to incorporate ancillary sources of data on which to base redistribution, a technique known as dasymetric mapping. Although the technique was first described in the 1930s, recent advancements in geographic information systems as well as increased availability of digital datasets have revitalized the concept. Despite the renewed interest, however, dasymetric mapping lacks a standardized methodology.

One difficulty, for example, has been finding ancillary data that corresponds well with population distribution. Night-time city lights imagery, for example, has been shown to demonstrate a reasonable correlation both with population density and several other measures of socioeconomic conditions, and has the advantage of very high temporal resolution (Sutton, 2003). The disadvantages, though, include very low spatial resolution (1km² pixels for the Defense Meteorological Satellite Program), and brightly-lit uninhabited areas such as parking lots and car dealerships which can skew data at higher resolutions.

To incorporate greater spatial resolution as well as more reliable characterization of population distribution, many researchers turn to land-cover maps, such as those produced from satellite imagery. With various land-cover classes, however, the key question becomes how to distribute the population among those classes. The most basic technique is known as binary classification, wherein all classes are designated as either inhabitable or non-inhabitable, and the population is distributed by areal weighting into the inhabitable areas of each enumeration district. This simple method has been shown to improve areal interpolation accuracy by almost

33% over choropleth mapping (Langford, 2002), however, further refinement is most certainly warranted.

Suggested methods for improvement include a density regression for the different land covers, field sampled density values, or a standardized set of density fractions to apply to all districts, such as 80% built-up area, 15% agricultural land covers, and 5% to other land covers excluding water bodies, used either as a distribution method or a limiting variable. (Donnay and Unwin, 2001; Eicher and Brewer, 2001). All of these methods are flawed in some way, though. Clearly no standardized ratio will be appropriate for all cities and these are reliant on either researcher assumptions or field sampling of density which is time consuming and difficult. On the other hand, a density regression can predict negative population densities and is apparently quite sensitive to classification error. What seems necessary is a distribution technique that mines the available data to make decisions about class distributions.

Mennis (2003) proposes such a methodology. Density values for each land cover class are calculated based upon the density of all census block groups that lie entirely within that class. These values can be calculated across the entire study area, or different relative ratios can be calculated for smaller areas, such as counties. This avoids the problems mentioned above and customizes the dasymetric distribution using the population values that one is already assuming are reliable. One difficulty that this method faced, however, is that as the spatial resolution of the remotely sensed land cover areas improves, it will become increasingly difficult to find even block groups that lie entirely within one land cover type. A potential solution is that the selection technique could be modified by designating population source units as representative of a land cover if that land cover comprised a certain percentage of the total source unit area, i.e. 90-95%. In addition, if certain classes remain unsampled, a class population estimate can be

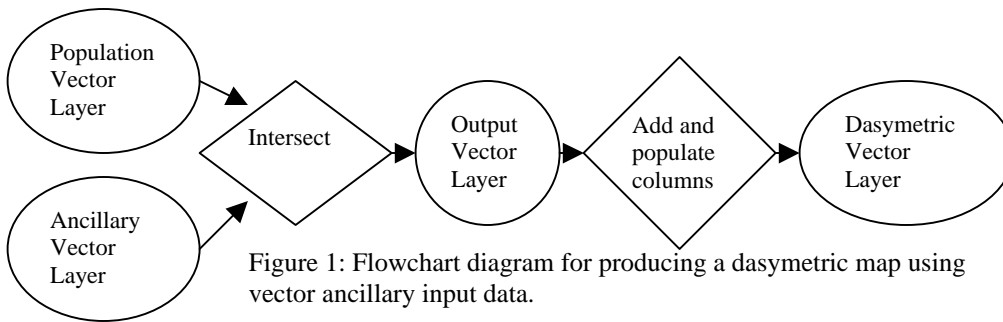
calculated using mean density values for the sampled classes to estimate the unknown population of each tract. This methodology ought to function at all levels of spatial resolution and offers the advantage that no assumptions about the distribution are required. The sheer numbers of steps and intermediate values involved in the calculation as well as the need for repetition to determine the ideal sampling threshold have suggested the benefits of automating this procedure.

Another key issue to be addressed in the context of dasymetric mapping is the format of the input data. Population enumeration units almost quintessentially fit the definition of vector data – polygons with explicitly defined borders and attached attributes. Although one might find rasterized population units, their native format is inarguably vector. Ancillary data for population disaggregation is not so consistent. At one time land use and land cover data was exclusively derived by hand from air-photo images by defining uniform areas and assigning attributes, thereby producing vector data. This remains the most accurate, if also the most time-consuming, method for deriving land use and land cover data from remotely sensed images. More recently, maps such as the National Land Cover Dataset (NLCD) have been produced by automatic and supervised classification of satellite image pixels based on their electromagnetic reflectance properties. Although this method is extremely efficient for large areas, these data are of very questionable accuracy. (Wickham, *et al*, 2004)

A third alternative is emerging, whereby images are automatically divided into regions using pattern recognition software. Although homogenous polygonal units are produced with some attempts at smoothed borders, these data are still fundamentally classified pixels. This dual nature should give researchers the freedom to choose the appropriate data format based on other factors. This paper will demonstrate that when dividing population data dasymetrically between nominal ancillary categories, raster format is far more efficient than vector.

Methods

Conceptually, there is no dramatic difference between performing dasymetric mapping using vector ancillary data and using raster ancillary data. Figure 1 illustrates the basic steps to

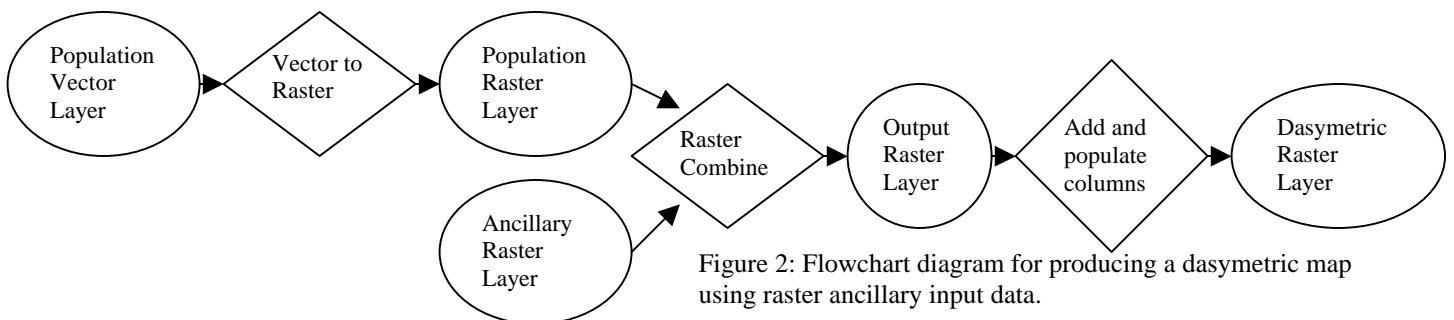


produce a dasymetric map using vector data. The two input layers are combined using a GIS INTERSECT tool,

and selections and calculations are then performed on the attribute table of the output layer to produce a dasymetric population distribution among the intersected polygons. Naturally, additional steps would be necessary if one initially had ancillary data in raster format.

Conversion from raster to vector, depending on the quantity, resolution, and variation of the data, can potentially be very time consuming, and the resultant converted data is often inefficient both in terms of storage and required processing time.

In contrast, Figure 2 illustrates the same function using raster ancillary data. It is assumed that the native format of the population source layer is vector, and thus a vector to raster



conversion is a required, and not an optional, part of the program, as the opposite conversion might be for the ancillary layer. The two raster layers are then intersected using a GIS

COMBINE tool, and just as in the preceding example, selections and calculations are performed on the attribute table to dasymmetrically redistribute the population. While this appears on the surface to be nearly identical, there is a critical difference between the vector INTERSECT and the raster COMBINE functions. The attribute table for combined raster layers contains a single record for each combination of input values, thus a single record for each ancillary class in each population unit. The attribute table for intersected vector layers contains a single record for each output polygon, thus each population unit could contain tens or hundreds of polygons for the each ancillary class. As resolution and complexity of the ancillary classes inevitably increase, this intersected attribute table can become exceedingly large, while the attribute table for combined raster layers will always have a maximum number of records R_{MAX}

$$R_{MAX} = N_a \cdot N_p$$

where N_a is the number of ancillary classes and N_p is the number of source population units.

This number of records in the output table has a significant impact on processing times for the subsequent calculations.

An outline of these calculations is shown below, most of which are described in greater detail by Mennis (2003), with the notable addition of the smart areal weighting procedure in step 3. These steps are fundamentally the same for both data formats, and it should also be clear why one would wish to minimize the number of records affected by each of the many select or update queries. The results section will show how dramatically different the processing times for each method can be.

1. For each ancillary class, select representative population units and calculate the (representative population density) = (sum of representative population) / (sum of representative inhabited area).

2. For each sampled or preset class, calculate a preliminary population estimate by simply multiplying the representative population density (sampled or preset) by the area of each output unit.
3. For each population unit, sum all of the preliminary population estimates of its subsidiary output units, compare to the actual population, and distribute any remaining population areally to the remaining inhabited subsidiary output units. Calculate (the representative population densities for all of the unsampled ancillary classes) = (sum of estimated population for all class output units) / (total ancillary class area). This is referred to as smart areal weighting.
4. Recalculate a secondary population estimate by again multiplying all representative population densities (sampled, preset, or smart areal weighted) by the area of each output unit.
5. To maintain pycnophylactic integrity for each population unit, find the sum of the secondary population estimates and calculate (a distribution ratio) = (output unit secondary population estimate) / (total estimated population for the specified population unit).
6. Calculate (the final population estimate) = (the initial population of the source unit) * (the output unit's distribution ratio).

Both scripts were written in Visual Basic for Applications utilizing the ArcObjects 8.3 object-oriented library and available at <http://ucsu.colorado.edu/~hultgren/AutoDasy.html>. It is anticipated that both shall be rewritten using ModelBuilder in Arc 9.0 in the near future, however, this is primarily to improve usability and flexibility, and the relative processing times are not expected to be affected.

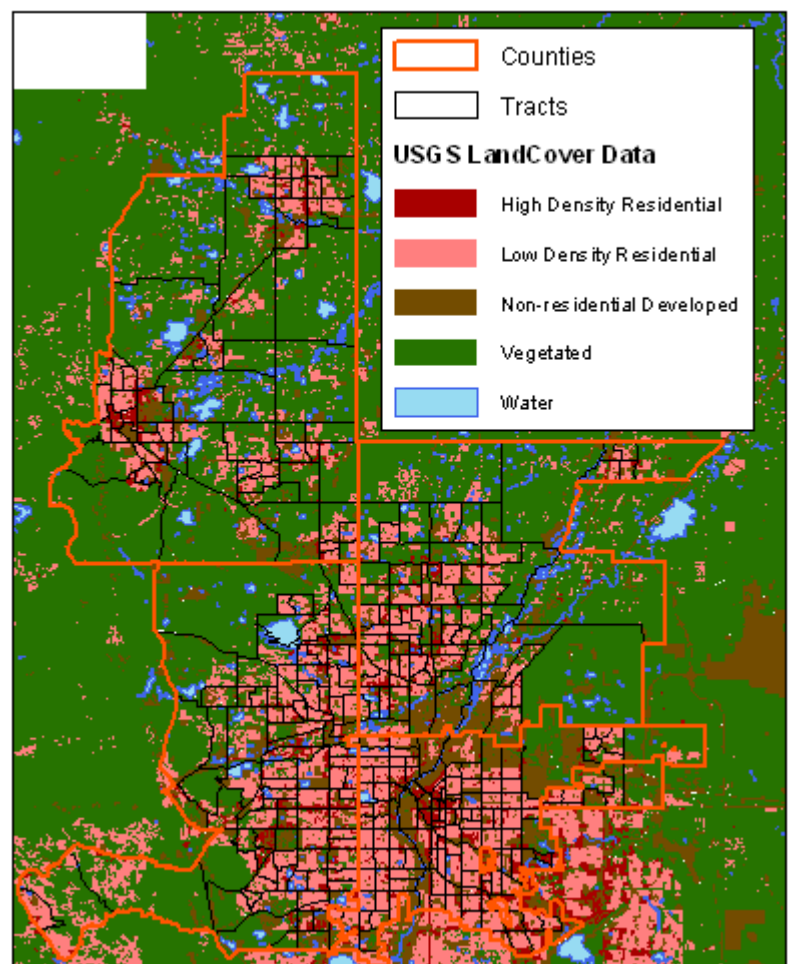
To ensure an accurate comparison, it was necessary to select a consistent dataset for both methods. Because of the disadvantages associated with available raster ancillary datasets, including accuracy concerns and conversion factors discussed, a vector dataset was chosen. The data were compiled by the United States Geological Survey (USGS) using a combination of

airborne and satellite imagery and represent the northern Front Range of Colorado from south Denver to Fort Collins in the year 1996. The accuracy of borders within the dataset is nominally 5m, with a minimum mapping unit (MMU) of 2.5 acres. The data are available to the public here: <http://rockyweb.cr.usgs.gov/frontrange/datasets.htm> in ArcInfo Coverage format. The data are classified using a standard Anderson hierarchical schema, and were aggregated for this study into the following classes of particular relevance for population distribution mapping:

- High Density Residential
- Low Density Residential
- Non-Residential Developed
- Vegetated
- Water

A map of the data is included in Figure 3. The Water and Non-Residential Developed classes were preset to zero population density, while the densities of the other three classes were determined by the script. This has the effect of beginning with a binary classification into inhabited and uninhabited classes, a technique with well-established reliability and effectiveness which is then refined further using sampling and smart areal weighting. For use as raster input, the data were converted using the Spatial Analyst conversion tool in ArcGIS 8.3 using an output cell size of 5m.

Figure 3: Dasymetric Input Data



Population data at the tract level were obtained from the U.S. Census in ArcInfo Shapefile format using the 2000 counts and areas. Tracts were selected from four counties in the Denver Metro Area (Denver, Boulder, Jefferson, and Adams) that were entirely contained within the ancillary dataset. No counties are entirely contained within the ancillary dataset, but the 373 selected tracts do represent a cross section of local land uses, including heavily urbanized, suburban, agricultural and natural areas. The total population of the dataset is just over 1.5 million, and the tracts encompass an area of about 2,000 square kilometers, which is a relatively modest size for typical dasymetric analyses.

Results

The output table for the raster combine operation on the Colorado dataset has a total of 1565 records, which is well under the calculated potential maximum of 1865. In contrast, the output table for the vector intersect operation has 8615 records, about 5.5 times as many. Thus for our dataset, there are an average of about 23 records for each source tract, or less than 5 polygons for each ancillary class within each tract. When one considers the typical complexity of the urban landscape, and the fact that remote sensing satellites have had sub-meter resolution for several years now, 23 land cover polygons per census tract seems like a rather conservative number that will only increase in years to come.

Both programs have been run numerous times with different settings on a Pentium 4 2.79GHz processor with 1.00 GB of RAM with no other non-system processes running. Although a rigorous benchmark test was not performed, the processing times are generally consistent. The table below shows typical processing times for each of the computational steps in the dasymetric framework. As the times are nearly an order of magnitude apart, it was

deemed unnecessary to perform a rigorous statistical analysis to determine the mean processing times for either.

Computational Step	Vector Time (s)	Raster Time (s)
Conversion of population layer from vector to raster	---	13
Intersection/combination of the two layers	57	82
Creation of new fields/columns	701	218
1. Selection of representative source units	973	3
2. Preliminary population estimate		
3. Smart Areal Weighting	803	128
4. Secondary population estimate		
5. Distribution ratio calculation	485	89
6. Final population estimate		
Total processing time	3019	533

Although the vector times are clearly much greater than the raster times in all categories following the intersection, the greatest disparity between the two lies distinctly in the first calculation step, where the vector program requires about 16 minutes and the raster program a negligible three seconds. The selection of representative source units can be easily accomplished using a simple SQL query in the raster attribute table, but requires a loop to sum the ancillary class areas for every source zone in the vector attribute table. Interestingly, although the difference between the programs varies at each step, the total processing time for the vector data is about 5.6 times the total time for the raster data, a very similar proportion as that between the number of records in the two data tables.

Discussion

There are two potential methods for optimizing the vector data that were not explored here but nonetheless merit serious discussion. The first is the impact of so-called sliver polygons in the intersection of the two vector datasets. This occurs when there are small registration errors between the two input layers, producing tiny sliver polygons when lines that should coincide do

not do so perfectly. These polygons introduce only marginal population error into a dasymetric map, but because each one represents a new record in the attribute table, if they occur with sufficient frequency they can contribute significantly to the final processing time. Workstation ArcInfo does offer a tool for eliminating sliver polygons, but this tool was not available within ArcGIS 8.3 at the time of this research. Since the average number of ancillary polygons per population unit appeared conservative based on anticipated future trends in land use and land cover data, it was decided that allowing the sliver polygons to remain was appropriate. Eliminating sliver polygons will become more important as minimum mapping units for datasets continue to shrink.

A second omission from the vector program could potentially have a much greater impact. ArcGIS 8.3 has a tool that allows a user to dissolve vector data based upon a chosen attribute, that is, all polygons with the same value in a column of the attribute table are combined into a single row, and their individual geometries are combined, even if the polygons themselves are spatially disjoint. While this would appear to be the perfect solution for reducing the number of rows in the vector attribute table, unfortunately the current implementation of the dissolve tool in ArcGIS does not allow multidimensional analysis of nominal data. One can dissolve on either the ancillary or the population data, but it is not presently possible to preserve both. A more robust OLAP tool would be required to preserve population unit attributes while dissolving on the ancillary categories. If such a tool were developed, the resultant attribute table would be identical to the raster attribute table, and the only difference in processing time between the two data formats would be that necessary to perform the OLAP dissolve.

Nevertheless, both of these methods for optimizing the vector data would also require additional processing time that would have to be included. With respect to usability, it is

unlikely that average users would have the patience to wait 50 minutes for a simple tool to process. Larger datasets have required more than 6 hours for processing. Naturally this time will decrease as processing hardware improves, but it is also likely that the complexity of the input data (if not the desired volume) will increase at a similar rate. As has already been demonstrated, increasing data complexity and resolution has a much greater effect on vector data in this context, with raster data remaining limited only by the number of ancillary classes and the number of source polygons. Unless there is a particular barrier to either selecting raster data or converting vector data to raster, it would seem that raster data is inherently optimized for this type of categorical dasymetric analysis.

Conclusion

Geographic Information Scientists have been searching for an effective dasymetric mapping methodology for years now. As methods are refined and improved, they also continue to increase in complexity. The creation of a binary dasymetric map requires merely a few simple GIS operations, but more advanced distributions demand extensive automated tabular calculations. This in turn has created the need to reduce processing times by various methods of optimization, so that techniques can be repeatedly studied and refined. For example, an empirical methodology proposed by Jeremy Mennis has already proven to be significantly more accurate than previous techniques, but processing very large vector input datasets often requires more than 24 hours, and frequently crashed before completion. The identical operation using rasterized input data has been shown to be almost an order of magnitude faster, and has already lead to further refinements in the methodology. Eventual standardization of this program would be an extremely beneficial tool in a scientist's GIS toolbox.

The United Nations projects that the majority of the world's population will be living in cities by 2005, and that almost all of the global population growth expected in the next 30 years will be concentrated in urban areas. Concerns about how to study and manage this growth abound, particularly in the developing world, where most of the growth is occurring and the fewest resources exist to cope with it. The continued emergence of remote sensing as an effective and inexpensive tool for many types of urban analysis will clearly be driven by the needs of this expanding global urban population. If an empirical methodology can be firmly established, dasymetric mapping has the potential to emerge as a robust tool for integrating remotely sensed data with traditional GIS data to quickly and accurately address these needs.

References:

- Bracken, I., 1993. An extensive surface model database for population-related information: concept and application. *Environment and Planning B: Planning and Design*, 20:13-27.
- Donnay, Jean Paul, and David Unwin., 2001. *Modelling Geographical Distributions in Urban Areas*. Remote Sensing and Urban Analysis. London: Taylor and Francis, 205-224.
- 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.
- Harris, Richard J, and Paul A Longley, 2000. New Data and Approaches for Urban Analysis: Modeling Residential Densities. *Transactions in GIS*, 4:217-234.
- Harvey, Jack T., 2003. *Population estimation at the pixel level: developing the expectation maximization technique*. Remotely Sensed Cities, London: Taylor and Francis, 181-206.
- Hutchinson, C. F., 1982, Techniques for combining Landsat and ancillary data for digital classification improvement. *Photogrammetric Engineering and Remote Sensing*, 48:123-130.
- Langford, Mitchel, 2003. *Refining methods for dasymetric mapping using satellite remote sensing*. Remotely Sensed Cities, London: Taylor and Francis, 137-156.
- Mennis, Jeremy, 2003. Generating Surface Models of Population Using Dasymetric Mapping, *The Professional Geographer*, 55:31-42.
- Mesev, V., 1998, The use of census data in urban image classification. *Photogrammetric Engineering and Remote Sensing*, 64:431-438.

- Miller, Roberta Balstad and Christopher Small, 2003. Cities from space: potential applications of remote sensing in urban environmental research and policy. *Environmental Science and Policy*, 6:129-137.
- Sutton, Paul, 2003. *Estimation of human population parameters using night-time satellite imagery*. Remotely Sensed Cities, London: Taylor and Francis, 301-334.
- Wickham, J.D., S.V. Stehman, J.H. Smith, L. Yang, 2004. Thematic accuracy of the 1992 National Land-Cover Data for the western United States. *Remote Sensing of Environment*, 91:452-486.