REMODELING CENSUS POPULATION WITH SPATIAL INFORMATION FROM LANDSAT TM IMAGERY

Yew Yuan,*†,1 Richard M. Smith† and W. Fredrick Limp*
*Center for Advanced Spatial Technologies, University of Arkansas, Fayetteville, Arkansas 72701, U.S.A.
†Department of Geography, University of Arkansas, Fayetteville, Arkansas 72701, U.S.A.

ABSTRACT. In geographic information systems (GIS) studies there has been some difficulty integrating socioeconomic and physiogeographic data. One important type of socioeconomic data, census data, offers a wide range of socioeconomic information, but is aggregated within arbitrary enumeration districts (EDs). Values reflect either raw counts or, when standardized, the mean densities in the EDs. On the other hand, remote sensing imagery, an important type of physiogeographic data, provides large quantities of information with more spatial details than census data. Based on the dasymetric mapping principle, this study applies multivariable regression to examine the correlation between population counts from census and land cover types. The land cover map is classified from Landsat TM imagery. The correlation is high. Census population counts are remodeled to a GIS raster layer based on the discovered correlations coupled with scaling techniques, which offset influences from other than land cover types. The GIS raster layer depicts the population distribution with much more spatial detail than census data offer. The resulting GIS raster layer is ready to be analyzed or integrated with other GIS data. © 1998 Elsevier Science Ltd. All rights reserved

INTRODUCTION

Since geographic information systems (GIS) were introduced in the 1960s, they have become an important technology for the integration of data from different sources and expertise from different disciplines (Burrough, 1986). Difficulties have been encountered, however, in fully integrating two types of important complementary GIS source data: socioeconomic data and physiogeographic data (Martin & Bracken, 1993).

One important type of socioeconomic data, census data, identifies and records all members of a population and their related socioeconomic variables with regular series,
complete spatial coverage, clearly defined boundaries, and easily understood variables. However, to protect the identity of individual population elements for confidentiality and to lower the cost of geographic coding, the data are reported in an aggregated tabular format within enumeration districts (EDs). The value for any variable in a particular ED reflects the aggregate total or mean value for the entire spatial extent of the district without any internal variation. The boundaries of EDs to provide the field enumerators with well delimited areas are chosen without a statistical basis which would serve to reflect the nature of the distribution of population and other socioeconomic phenomena. They are arbitrary at all levels and vary greatly in size and shape. Certain parts of an ED can cover some areas with substantial population and others that have no residents. The resulting measures, however, apply to the entire ED and can result in unpopulated areas which appear populated. In short, for very appropriate reasons, census data are often deficient with regard to spatial referencing, making spatial disaggregation difficult (Robinson, Morrison, Muehrcke, Kimerling, & Guptill, 1995; Martin & Bracken, 1993; Bracken, 1991; Langford, Maguire, & Unwin, 1991; Langford & Unwin, 1994). The most common cartographic representations of census data are derived by choroplethic mapping methods, which are used by virtually all geographers and many non-geographers (Slocum & Egbert, 1993). These choropleth maps and their underlying concepts are used in GIS analysis as well. Errors in choropleth maps are inherent and obvious to cartographers (Jenks & Caspall, 1971), but may not be evident to others. Among spatial interpolations, smoothing methods may initially be viewed as an adjunct to choroplethic methods, but they may result in spurious impressions of precision since many socioeconomic phenomena are discontinuously distributed and have sharp discontinuities between adjacent areas within an enumeration district (Lain, 1983; Arbia, 1989).

Remote sensing data, particularly that provided by the current LandSat and SPOT sensors, are widely recognized by today’s GIS users as one of the most important physiogeographic data sources, providing large quantities of timely and accurate spatial information. These remote sensing data are like many other physiogeographic data as they can very clearly provide information on the locations of socioeconomic phenomena including many human activities and residential areas. Conversely, they are unable to provide other socioeconomic characteristics, such as are available in census data (Martin & Bracken, 1993).

Integrating these two complementary data sets clearly would be valuable, but difficulties are encountered because these two types of data are captured differently, structured differently, based on different underlying conceptual models, and constructed for different application needs. The successful integration of socioeconomic and physiogeographic data could clearly provide better sources of information to many GIS applications, and should enhance GIS analytical power (Martin & Bracken, 1993).

A goal of integration and interpolation of these two data sets was identified some time ago but implementation was without much success (Martin & Bracken, 1993). However, using traditional dasymetric mapping methods (Dent, 1996) and recently introduced statistical methods in a GIS setting, it is now possible to obtain encouraging results (Langford et al., 1991). Dasymetric mapping can approximate the true character of surface variations more accurately than simple polygon overlay methods and smoothing methods. Although errors are introduced in each step of the dasymetric process, they can be partially offset by additional information. In the past, dasymetric mapping programs have been more difficult to develop than choroplethic mapping programs (Robinson, Sale, & Morrison, 1984). However, with the development of GIS technology dasymetric methods
are now becoming more popular. Using GIS based methods it is now possible to integrate a variety of other spatial information to create new, relatively more homogeneous target units within an original ED (Robinson et al., 1995). For example, Flowerdew and Green (1989) regression methods allow for interpolation of two socioeconomic data sets collected by different unit systems. The results were much closer to the true values than the standard polygon-overlay areal interpolation approach. Langford et al. (1991) derived population estimation models by integrating British census data for 49 wards in the Leicestershire area with LandSat TM imagery. The results were encouraging, but failed in globally applying the regionally derived model. The problem of the unexplainable intercept in their statistical models was recognized but not addressed. In solving negative estimates given by the statistics, arbitrary parameter reduction was then applied (Langford et al., 1991). The resultant simplified model is insensitive to classification errors (Fisher & Langford, 1996).

In this study, 1990 U.S. census population counts are remodeled by integrating spatial information from a land cover map. The land cover map is derived from 1992 LandSat 5 Thematic Mapper imagery. The study area consists of four counties in central Arkansas: Faulkner, Lonoke, Pulaski, and Saline (Figure 1). The size of the entire study area is 7779.2 km². After initial regression, the land cover type aggregation alternatives (Langford et al., 1991) are assessed using statistical T-tests. The tests do not support certain types of aggregation. In order to maintain the maximum precision of the estimates and solve the negative estimate problem, we have not chosen to regionally derive a globally applicable model, but rather to derive regional regression models followed by local fitting with scaling techniques. The Flowerdew and Green (1989) scaling technique offsets influences other than land cover types, and a newly introduced scaling technique removes the negative

FIGURE 1. Study area — Faulkner, Lonoke, Pulaski, and Saline counties.
estimates. The intercept is preset to “zero” in regression since where there is no land, there are no people.

HARDWARE AND SOFTWARE

This study is performed on a Sun 690a at the Center for Advanced Spatial Technologies (CAST) and a SPARCcenter 2000 at the University of Arkansas, Fayetteville. The Geographical Resources Analysis Support System (GRASS) (U.S. Army Corps of Engineers, Construction Engineering Research Laboratory, 1993) is chosen to provide the GIS working environment in this study. GRASS is public domain UNIX-based GIS software originally developed by the U.S. Army Corps of Engineers, Construction Engineering Research Laboratory. GRASS contains many raster and vector modules and strongly supports raster/vector integration and spatial analysis. The transparent data format and structure and the full capability to interact with the UNIX shell environment make it very powerful and efficient when integrated with shell programs, though it requires the operator to have a certain level of programming skills. Statistical analyses are performed with SAS (SAS Institute, Inc., 1989), which includes a wide range of mathematical computation modules and many statistical analysis modules. In SAS, these modules embrace regression procedures (REG), general least-squares-fitting procedures (GLM), etc. (SAS Institute, Inc., 1989).

DATA

The 1990 census data from the U.S. Bureau of the Census used in this study include two parts, the Summary Tape Files (STFs), the tabular data, and the Topologically Integrated Geographic Encoding and Referencing system (TIGER), the spatial data. The link between these two are census geographic entity codes. When the U.S. Bureau of the Census collects census data, the housing unit is the means of locating people and their variables in space. In the final form, all socioeconomic variables are summarized in EDs and presented in tabular format (U.S. Bureau of the Census, 1991a).

The EDs defined by the U.S. Bureau of the Census are named as census geographic entities, which have a structured hierarchy. There are ten basic levels and a few supplementary levels. The basic levels are (from lower to higher): Block, Block Group (BG), Tract and Block Numbering Area (CTBNA), Place, County Subdivision, County, State, Division, Region, and Nation (U.S. Bureau of the Census, 1991a; Myers, 1992). In this study the second lowest level, Block Group (BG), was chosen as the working level.

The tabular format reports are distributed in book form and in computer readable, digital format, and are referred to as STFs data. Some STFs are derived from 100% of all respondents in the census, such as STF1s and STF2s, and others from samples of respondents, such as STF3s and STF4s (Myers, 1992). The STF1a was used in this study.

The TIGER was initiated by the U.S. Bureau of the Census as a means to facilitate decennial census counts. It is designed to automate the mapping and related geographic activities which are required to support the tabular reports of the decennial census. It allows users to efficiently produce maps of census geographic entities at all levels. Each census geographic entity on these maps has its unique code which agrees with the code
used in STFs. A census geographic entity map can be linked to STFs to graphically present any of the socioeconomic variables of STFs (U.S. Bureau of the Census, 1991b).

The land cover map used in this study was derived by Gorham (1995) from Landsat 5 Thematic Mapper (TM) images (scene 24/36, October 19, 1992) with a 30 m x 30 m spatial resolution. The data were processed by applying a tasseled cap transformation (Crist & Cicone, 1984) on images of the bands 1–5 and 7 of the TM imagery, producing three output indices, brightness, greenness and wetness. The three indices were isoclastered to 159 classes with subtle detailed differences between land cover types. Each of the 159 classes was placed into one of the Anderson Level 2 (Avery & Berlin, 1992) categories. There are a total of 16 categories in the initial land cover map (Gorham, 1995).

**METHODOLOGY**

The work flow in this study is illustrated in Figure 2. The first step is to reconstruct the census geographic entities and retrieve population data. The second is to reclassify the land cover map. The third is to overlay the census population map and the reclassified land cover map. The fourth is the statistical regression, modeling, testing, and model choice. The final step is to locally fit the estimates from regression to each county and each land cover type with scaling techniques.

**Census Population Map**

Using appropriate software modules and shell programs, a vector format census geographic entity boundary map at BG level is constructed and each BG in the map from TIGER is linked to the appropriate census socioeconomic variables by combination codes from TIGER and key columns in STFs. A population distribution map is constructed and used in subsequent steps.

**Reclassification of the Land Cover Map**

The next step is reclassification of the land cover map. For a statistical regression, fewer categories and more observations generally provide more reliable results. It is clear that some categories in the initial land cover map may not be significantly different with respect to population distribution. Based on the similarities of possible population densities, the original land cover map (Gorham, 1995) is reclassified to include only seven categories:

1. **Residential** [11];
2. **Commercial–industrial**, including commercial [12], industrial [13] and mixed urban or built-up land [16];
3. **Recreational**, including other urban or built-up land (consists mostly of parks and golf courses) [17];
4. **Agricultural**, including crop land and pasture [21];
5. **Other agricultural**, including other agricultural land [24];

---

2 The numbers in brackets are category codes for Anderson level 2 of the Land Use and Land Cover Classification System for Use with Remote Sensor Data.
(6) *Forested*, including deciduous forest [41], evergreen forest [42], and mixed forest [43], and including transitional areas [76];

(7) *Uninhabited*, including streams [51], lakes [52], reservoirs [53], other water bodies [54], forested and non-forested wetlands [61, 62], sandy areas other than beaches [73], and strip mines, quarries, and gravel pits [75].

According to the reclassified land cover map (Figure 3), 4.82% of the study area is residential; 1.65% is industrial and commercial; 0.23% is recreational; 37.93% is agricultural; 48.87% is forested; and 6.5% is uninhabitable.

**Regression Models**

Population distribution within an ED is not homogeneous, but may be expected to relate to the land cover types since land cover properties can be expected to be one factor in the distribution process (Flowerdew & Green, 1989). If average population densities ($b_j$)
and total areas ($x_j$) for each land cover type within an ED are known, the total population ($y$) of the ED would be the sum of the products of the population densities ($b_j$) and total areas ($x_j$) for every land cover type:

$$y = \sum_{j=1}^{n} (b_j \cdot x_j)$$

where $y$ is the total population in an ED, $b_j$ is the population density for the $j$th land cover type, $x_j$ is the total area of the $j$th land cover type within the ED, and $n$ is the number of land cover types occurring in the ED.

If the population density for each land cover type within each ED is unknown but the total population for each ED and the total area for each land cover type within every ED is known, and if population densities are related to land cover types (Flowerdew & Green, 1989), then a regression model can be written to estimate average population densities ($\bar{b}_j$) for each land cover type as (Langford et al., 1991; Langford & Unwin, 1994):

$$y_i = \sum_{j=1}^{n} (b_j \cdot x_{ij} + \epsilon_i)$$

where dependent variable $y_i$ is the total population count for the $i$th ED, $b_j$ is the average population density for the $j$th land cover type, and independent variables $x_{ij}$ are the total areas for the $j$th land cover type within the $i$th ED. If the number ($m$) of EDs is greater than or equal to the number ($n$) of land cover types, a regression process can find estimates ($\hat{b}_j$).
If \( m \) is much greater than \( n \), the estimates \((b_j)\) will, generally, be more reliable. Also, certain statistical tests can give goodness of fit of the model, reliability of the model, and reliability of the estimates.

**Overlaying Operation**

In a GIS environment, the census population map at the BG level is overlain on the land cover map. When the census population map, considered as the source unit system, is overlain by the reclassified land cover map, considered as the target unit system, each source unit is broken to smaller pieces. These small pieces are called subunits. These subunits are labeled by the combination of source unit code \((i)\) and the target unit code \((j)\). The total areas \((x_{ij})\) for a certain land cover type within a source unit are the sum of all subunits labeled by the same code \((ij)\), where the \(i\) is the \(i\)th BG, and \(j\) is the \(j\)th land cover type. The result of the overlay operation is twofold: one is a spatial record, a map with the subunit system; the other is a data matrix. There are \(m\) rows by \(n + 1\) columns, where \(m\) is the number of source units in a working area and \(n\) is the number of land cover types. Each element of the matrix from the second column onwards is the total area \((x_{ij})\) of the \(j\)th land cover type within the \(i\)th BG. The first column is the total population counts \((y_i)\) of BGs. In this study, the \(m\) number varies from county to county, 256 for the whole study area, Faulkner 34, Lonoke 31, Pulaski 149, and Saline 42. The \(n\) number is seven for testing General Model I, and five for the testing General Model II.

**Regression Model Testing and Refinement**

In SAS, the GLMs (general linear models) with the “non-int” (non-intercept) option are chosen to find estimates and to test the models. The option “non-int” is preset because there should be no population if there is no area on the Earth surface. The General Model I, which has seven land cover types, can be written as:

\[
y_i = \sum_{j=1}^{7} (b_j \cdot x_{ij} + \epsilon_i)
\]

The data matrix \((256 \times 8)\) for the whole study area is input into to SAS, and analyzed by a GLM. SAS reported estimates \((b_j)\) of population densities for each land cover type (Table 1), other test results (Table 2), and predicted values \((\hat{p}_i)\) for each BG. The multiple coefficient of determination \((R^2)\), 0.799, indicates that the goodness of fit for General Model I on the whole study area is quite high. Almost 80% of the total sum of squares of deviation in the population distribution in BGs over all of the study area can be explained by this model. The results of the \(F\)-test of the null hypotheses for estimates, \(H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0\), further confirm the whole model by indicating that the relationship between population distribution and land cover map exists and is strong (Table 3).

The results of the \(T\)-test, which evaluates each parameter \((b_j)\) independently in a model with the null hypotheses \(H_0: \beta_j = 0\), indicate that some of the parameters in General Model I are good, but others may not be. The \(T\)-values for each null hypothesis \(H_0: \beta_j = 0\), the \(T\)-statistic values, and the probabilities of wrong rejections \(\text{Prob} > |T|\) for each estimate for General Model I are listed in Table 1. The \(T\)-value for the residential parameter, 24.258, is much greater than the \(T\)-statistic value, 2.358. For the commercial—industrial and
Remodeling Population with Satellite Imagery

Table 1. Estimates and Results of the T-test for General Model I on Whole Study Area

| Land cover type      | Parameter estimate | Standard error | T for $H_0$ | T-statistic $\alpha = 0.01$ | Prob > |T| Significance |
|----------------------|--------------------|----------------|-------------|-----------------------------|--------|--------------|
| Residential $b_1$    | 0.928045           | 0.038257       | 24.258      | 2.358                       | 0.0001 | Yes          |
| Corn-indus $b_2$     | 0.170972           | 0.055183       | 3.098       | 2.358                       | 0.0022 | Yes          |
| Recreation $b_3$     | 0.542633           | 0.184892       | 2.935       | 2.358                       | 0.0036 | Yes          |
| Agriculture $b_4$    | 0.002362           | 0.001150       | 2.054       | 2.358                       | 0.0410 | No           |
| Other-agr $b_5$      | 0.005211           | 0.003747       | 1.390       | 2.358                       | 0.1656 | No           |
| Forested $b_6$       | -0.003988          | 0.002017       | -1.976      | 2.358                       | 0.0492 | No           |
| Uninhabit $b_7$      | -0.010576          | 0.009838       | -1.075      | 2.358                       | 0.2834 | No           |

recreational land cover types, the T-values are also greater than their T-statistic values. These three $T$-tests confirm the strong correlation between population distribution and each of these three types of land cover. However, the remaining four land cover types in General Model I all failed their T-tests. If all of these estimates are applied in remodeling, the population distribution in correspondent areas may not as reliable as desired. In general, the results of these $T$-test suggest the overall model has merit but should be refined.

General Model I is simplified to General Model II and tested. Agricultural land ($b_4$), other agricultural land ($b_5$), and forested areas ($b_6$) in General Model I are clustered together and become a new land cover type labeled as others ($b_4$) in General Model II. Uninhabitable remains, but its parameter symbol becomes $b_5$ in General Model II:

$$y_i = \sum_{j=1}^{5} (b_j \cdot x_{ij} + \varepsilon_i).$$

The results of the regression and $F$-test are listed in Tables 2 and 3, and the results of $T$-tests are listed in Table 4. The multiple coefficient of determination ($R^2$) and the result of the $F$-test are similar to General Model I. The results of the $T$-test for the first three parameters are also similar to General Model I. However, the result of the $T$-test for the
Table 4. Estimates and Results of the T-test for General Model II

<table>
<thead>
<tr>
<th>Land cover type</th>
<th>Parameter estimate</th>
<th>Standard error</th>
<th>T for $H_0$</th>
<th>T-statistic $a = 0.01$</th>
<th>Prob &gt;</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential $b_1$</td>
<td>0.917207</td>
<td>0.038198</td>
<td>24.012</td>
<td>2.358</td>
<td>0.0001</td>
<td>Yes</td>
</tr>
<tr>
<td>Com-indus $b_2$</td>
<td>0.201261</td>
<td>0.055652</td>
<td>3.616</td>
<td>2.358</td>
<td>0.0004</td>
<td>Yes</td>
</tr>
<tr>
<td>Recreation $b_3$</td>
<td>0.486003</td>
<td>0.175296</td>
<td>2.772</td>
<td>2.358</td>
<td>0.0060</td>
<td>Yes</td>
</tr>
<tr>
<td>Others $b_4$</td>
<td>0.000705</td>
<td>0.000909</td>
<td>0.775</td>
<td>2.358</td>
<td>0.4390</td>
<td>No</td>
</tr>
<tr>
<td>Uninhabit $b_5$</td>
<td>0</td>
<td>0.000000</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

new land cover type, others, is now worse than the original results in General Model I. This implies that the correlations between population distribution in each of these three land cover types are different and these land cover types should not be clustered but discriminated. The correlations of these land cover types to the population distribution apparently have not yet been better defined, or the correlations between population distribution and these land cover types are not uniform over space, or there may be some other process operating. This differs somewhat from Fisher and Langford's findings (Fisher & Langford, 1996), and suggests that some problems may become hidden when aggregating land cover types.

Approaching solutions differently to Langford et al. (1991) and Langford and Unwin (1994), General Model I is regionally regressed on each of the four counties in the study area one by one. The results of regression and F-test are listed in Tables 2 and 3. The estimates for each land cover type of each county are listed in Table 5. The coefficients of determination ($R^2$) are 0.925 for Faulkner, 0.866 for Saline, and 0.843 for Pulaski. All three of these are higher than those of the models regressed over the whole study area. Lonoke County's coefficient of determination ($R^2$) is about 0.042 lower than the models regressed over the whole study area. The F-values are all high. The T-test values for each estimate are similar to or slightly better than those from the model regressed over the whole study area except those for Lonoke, which are worse. The residential estimates ($b_1$) for Faulkner and Saline are close to but slightly lower than the average. The estimate for Pulaski is higher than the average. All three of the estimates are considered reasonable. However, the residential estimate for Lonoke is far below the average, whereas this is not true for that area. The problem may be caused by errors of classification. More specifically, the potential problem is that some small areas of harvested crop land might be misclassified as residential areas based on the fall season 30 m resolution satellite imagery.

Models regionally regressed may obtain more precise results (such as Faulkner, Saline, and Pulaski), or indicate original land cover map errors (Lonoke). Therefore, those models regionally derived from General Model I are chosen to remodel the census population distribution with scaling techniques.

Table 5. Estimates from General Model I by Regional Regression

<table>
<thead>
<tr>
<th>Land cover type</th>
<th>Faulkner</th>
<th>Lonoke</th>
<th>Pulaski</th>
<th>Saline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential $b_1$</td>
<td>0.901071</td>
<td>0.552624</td>
<td>1.065059</td>
<td>0.905725</td>
</tr>
<tr>
<td>Com-indus $b_2$</td>
<td>0.350876</td>
<td>0.544438</td>
<td>0.101640</td>
<td>0.002716</td>
</tr>
<tr>
<td>Recreation $b_3$</td>
<td>1.471982</td>
<td>0.373849</td>
<td>0.332006</td>
<td>0.595889</td>
</tr>
<tr>
<td>Agricult $b_4$</td>
<td>0.016745</td>
<td>-.007219</td>
<td>0.014246</td>
<td>0.067863</td>
</tr>
<tr>
<td>Other-agr $b_5$</td>
<td>0.019921</td>
<td>-.005925</td>
<td>0.017750</td>
<td>-.003610</td>
</tr>
<tr>
<td>Forested $b_6$</td>
<td>0.012544</td>
<td>0.025462</td>
<td>0.001486</td>
<td>0.000000</td>
</tr>
<tr>
<td>Uninhabit $b_7$</td>
<td>0.003693</td>
<td>-.040052</td>
<td>-.003157</td>
<td>0.045478</td>
</tr>
</tbody>
</table>
Applying Scaling

There are two reasons to further modify the estimates obtained from the previous steps. First, some of the estimates of certain counties from the regression are negative numbers, though only slightly less than zero. Secondly, it is obvious that population distribution is not explained solely by land cover type. For example, the residential population density near an urban center may be higher than residential areas far from urban centers. Transportation systems, hydrology, natural environment, and socioeconomic settings all influence population distribution.

To begin to address the negative estimates and the influence of factors other than land cover type, primitive but useful scaling techniques can be introduced. First, negative estimates are adjusted. All estimates of a county in which one or more negative estimates are reported are raised by the absolute value of the lowest negative estimate ($b_k$) of that county:

$$b_j' = b_j - b_k$$

where the $b_j'$ are the adjusted estimates, $b_j$ are the estimates from the regression, and $b_k$ is the lowest negative estimate. The lowest estimate ($b_k'$) will now be zero, which means there is no population on that land cover type. All other adjusted estimates ($b_j'$) are positive and the ratio between them remains same as before the adjustment.

Secondly, in removing the influence of factors other than land cover type, it is assumed that the population counts ($y_i$) for BGs from the census are highly reliable. The estimates ($b_j$) of population density derived from the regression or adjusted estimates ($b_j'$) from the previous step can be scaled to more refined estimates ($b_0$). The ratio of the predicted population ($p_i$) and the census counts ($y_i$) is used to adjust the estimates ($b_j$ or $b_j'$) (Flowerdew & Green, 1989). The mathematical expression for scaling the population densities ($b_j$) is:

$$b_j = \left( \frac{y_i}{p_i} \right) \times b_j \quad \text{or} \quad b_j = \left( \frac{y_i}{p_i'} \right) \times b_j'$$

where $b_j$ is the refined population density for the $j$th land cover type within the $i$th BG; $y_i$ is the population count of the $i$th BG from the census; $p_i$ is the predicted total population of the $i$th BG, the sum of products of estimates ($b_j$) and total areas ($x_{ij}$), or $p_i'$ is the adjusted predicted population for the $i$th BG, the sum of products of the adjusted estimates ($b_j'$) and the total areas ($x_{ij}$); and $b_j$ is the estimate for the $j$th land cover type for the whole county from the regression, and $b_j'$ is the adjusted estimate from the previous step.

It can be seen that the refined estimates ($b_0$) are the original estimates ($b_j$) scaled down for those BGs in which the populations are overestimated ($p_i > y_i$ or $p_i' > y_i$). Conversely, the refined estimates ($b_0$) are the original estimates ($b_j$) scaled up for those BGs in which the populations are underestimated ($p_i < y_i$ or $p_i' < y_i$). After scaling, the sum of the products of the refined population density ($b_0$) and the total areas ($x_{ij}$) for the $i$th BG can be mathematically proven to be equal to the population counts ($y_i$) reported in the census:

$$y_i = \sum_{j=1}^{n} (b_0 \cdot x_{ij})$$

Linking all adjusted and refined population densities ($b_0$) to the map of subunits is the final step of the remodeling. The final output is the remodeled population distribution map.
(Figure 4). It is a GIS raster map layer with $30 \times 30 \text{m}^2$ spatial resolution. Comparing this GIS map layer with the population distribution map composed by the classless choroplethic method solely based on census counts (Figure 5), it is obvious that the remodeled map layer carries much more spatial information, and can be seen as an alternative to smoothing methods.

CONCLUSION

In summary, this study applies a modified method for identifying correlations between socioeconomic variables and physiogeographic data, uses these correlations to remodel census population counts, and develops a method of regional modeling with local fitting techniques to offset influences other than land cover types. The results are subjected to statistical analyses in the hope of demonstrating the reliability of the remodeled socioeconomic variable. Results of the analyses show that the correlations between population distribution and land cover types are high and reliable. This is similar to results obtained by Langford et al. (1991). However, the results of the $T$-tests suggest that the simplification method used in previous studies (Langford et al., 1991; Langford & Unwin, 1994) and in this study may hide problems and make some errors of image classification undetectable (Fisher & Langford, 1996). We believe that the alternative local fitting via scaling (Flowerdew & Green, 1989) and other techniques as used in this study have
advantages over simplification in reducing model parameters, though the theory of local fitting is not yet robust.

This resulted in the construction of a raster GIS map layer of population distribution with much more spatial detail than can be provided by the census data and their conventional representations. Such a GIS map layer is readily integrated with other GIS data, and could be beneficial to a variety of GIS applications. The methodology developed in this study for remodeling population distribution also can be applied to remodeling other census socioeconomic variables. The widespread application of this method and subsequent production of raster GIS map layers of socioeconomic variables could allow better utilization of abundant socioeconomic information from census reports.

Further research may lead in three directions. One direction is to improve the resolution of the results. A second is to remodel socioeconomic variables other than population. Finally, methods to assess the reliability of the results can be improved. Improving the resolution of results can be accomplished in several ways. Researchers might obtain better land cover maps by using finer resolution remote sensing imagery, or by conducting image classifications which use methods and band combinations that are focused on criteria related to the socioeconomic variable which is being remodeled. Results of such a classification might include either traditional categories of land cover, which can be clustered later, or “cover types” based on imagery classes more closely linked to socioeconomic landscape characteristics. Methods to assess reliability could include the use
of field measures to test the model as well as photo-interpretation of high resolution aerial photography.

ACKNOWLEDGEMENTS—Portions of the work reported here were conducted with support from the National Aeronautics and Space Administrations under grant NAGW 4117.

REFERENCES


