Christopher Clayton
John E. Estes
Department of Geography
University of California
Santa Barbara, CA 93106

Image Analysis as a Check on Census Enumeration Accuracy

Image interpretation procedures provide a count of residential structures of higher absolute accuracy than the Federal Census.

Introduction

The federal bureau of the Census is moving rapidly away from field enumeration methods to the use of a mailout census. This method of enumeration requires the existence of a locational addressing system whereby each address can be uniquely identified with the smallest areal unit of aggregation used in a particular environment. Such a system, known as the Address Coding Guide (ACG), was first used in a number of selected areas in 1970 although the area under consideration here was not one of them. This geographical referencing system has since been incorporated into the Dual Independent Map Encoding (DIME) procedure.[1]

The DIME file contains a detailed map of an area stored in the form of a collection of digitized nodes (point references such as street intersections) and lines (linear features such as streets, railroad tracks, and creeks). This geo-coded information allows one to construct polygons of varied size and shape using reference nodes and lines and, when interfaced with the ACG, aggregate statistics to any areal level greater than or equal to that of the census block.

Imagery can serve an extremely important function in the preparation of the DIME files. The construction of these files relies upon existing map sources covering the area of concern, and usually a number of independently derived map sources are used simultaneously in either file creation or update. However, such sources often contain errors concerning correct street naming, street location and length, and even street existence. This is especially true in areas which are undergoing rapid or continual change. Under such circumstances imagery provides a contemporaneous source of locational information that can enhance the accuracy of the DIME file.

Remotely sensed data from aircraft and satellite platforms, both in digital and nondigital format, have successfully been used to provide information concerning socioeconomic and demographic character-

Abstract: *The research reported herein demonstrates that high altitude color infrared photography constitutes a data source capable of being employed as a check on the accuracy of certain types of information presented by the Federal Bureau of the Census. Such imagery is used to provide an accurate estimate of the number of single-family residential structures contained in census blocks; important data which can serve in many types of socioeconomic and demographic studies as a surrogate for areally distributed population parameters. This image-derived data pinpoints the magnitude and location of errors contained in the Federal Census tabulations and enables one to suggest how some of the errors might have been made. Errors associated with the image-based data are less in both magnitude and variability than those of the Federal Census at both the census tract and block levels.*

Photogrammetric Engineering and Remote Sensing,
Vol. 46, No. 6, June 1980, pp. 757-764.

istics of populations. These data sources, which complement alternative sources such as those collected by the Federal Bureau of the Census, have a number of characteristics that have proved useful in public and private decision-making. In general it can be said that remotely sensed data (1) have the capacity to provide information not regularly collected by other means, e.g., land-use data; (2) allow a rapid inventory and/or enumeration to be made of numerous phenomena, e.g., the amount of land in a given use category, or the number of residences within a given area; (3) facilitate the frequent and efficient updating of data files on important socioeconomic and demographic parameters, allowing changes to be detected between successive time periods; and (4) provide a physical and long-lived record of the state of an environment against which alternative or subsequent data sources may be checked with relative ease.

This study, an outgrowth of the authors' work testing a conceptual model for generating residential energy demand estimates, reports on research which evaluates the utility of imagery-derived estimates of single family residential structures at the census block level of aggregation.[2] As our research progressed it became apparent that analysis of high altitude color infrared aerial photography could provide an excellent method for checking the accuracy of census data. Major findings in this regard are that (1) the magnitude of discrepancy between imagery- and census-derived counts at the tract level is relatively small; (2) image analysis data discrepancy levels at the block scale are exceptionally small compared with those of the census; (3) the dispersion of error is much greater in the case of the census counts; and (4) reasons for errors in the census data can often be inferred from the analysis of the remotely sensed data. The ability to count accurately the number of structures in any areal unit is of considerable importance. Through the application of average values of socioeconomic and demographic characteristics extracted from the census, it is possible to derive accurate estimates of a number of attributes for areas that would not otherwise be available.

Background

Substantial precedent exists for the use of remotely sensed information sources in estimating socioeconomic and demographic attributes for relatively small areas such as census tracts and blocks. Such information is frequently used in conjunction with census information, the latter being used to derive structural parameters that can then be applied to data derived solely from the imagery, e.g., number of particular types of dwelling units per given area, and the relationships that exist between socioeconomic and land-use characteristics. Such studies include the now-classic works of Green[3], Binsell[4], and Mumbower and Donoghue[5], and later studies by Eyre *et al.*[6], Hsu[7], Lindgren[8], and Kraus[9].

Recent work, using multi-spectral Landsat imagery, has concentrated on (1) constructing accurate land-use classification schemes in urban and suburban area; (2) the detection of land-use change on the edge of metropolitan communities; and (3) deriving population estimates in areas undergoing substantial growth. See the works of Christensen[10], General Electric[11], Friedman *et al.*[12], and Landini[13] for aspects of such applications. A major impetus for this work is the need on the part of the Federal Bureau of the Census to acquire accurate population estimates and urbanized area delineations on a frequent basis after 1980. Landsat-derived data possesses considerable importance here because of the timeliness and spatially extensive nature of the coverage.

Study Area

The study area chosen in the Goleta Valley of the Santa Barbara Standard Metropolitan Statistical Area (SMSA), California, is one of a largely suburban, residential nature in which the vast majority of dwellings are single family detached houses. Of the approximately 2000 year-round housing units in the chosen census tract, 95 percent are in one unit (single family) structures and 60 percent of the land area is in residential use. The tract contains a total of 65 census blocks which are defined by the census as, "a well-defined rectangular piece of land bounded by streets and roads. However, it may be irregular in shape or bounded by railroad tracks, streams or other features."[14] Small and homogeneous as the study area may be, it is important to consider the fact that it is exactly in such environments that the majority of population redistribution is occurring.

Data Sources

Three independently derived enumerations were used in the study. First, the number of what were judged to be single family dwelling structures was enumerated employing standard manual photographic interpretation techniques from 1:63,360 scale high altitude color infrared aerial photography (Plate 1). Second, housing count

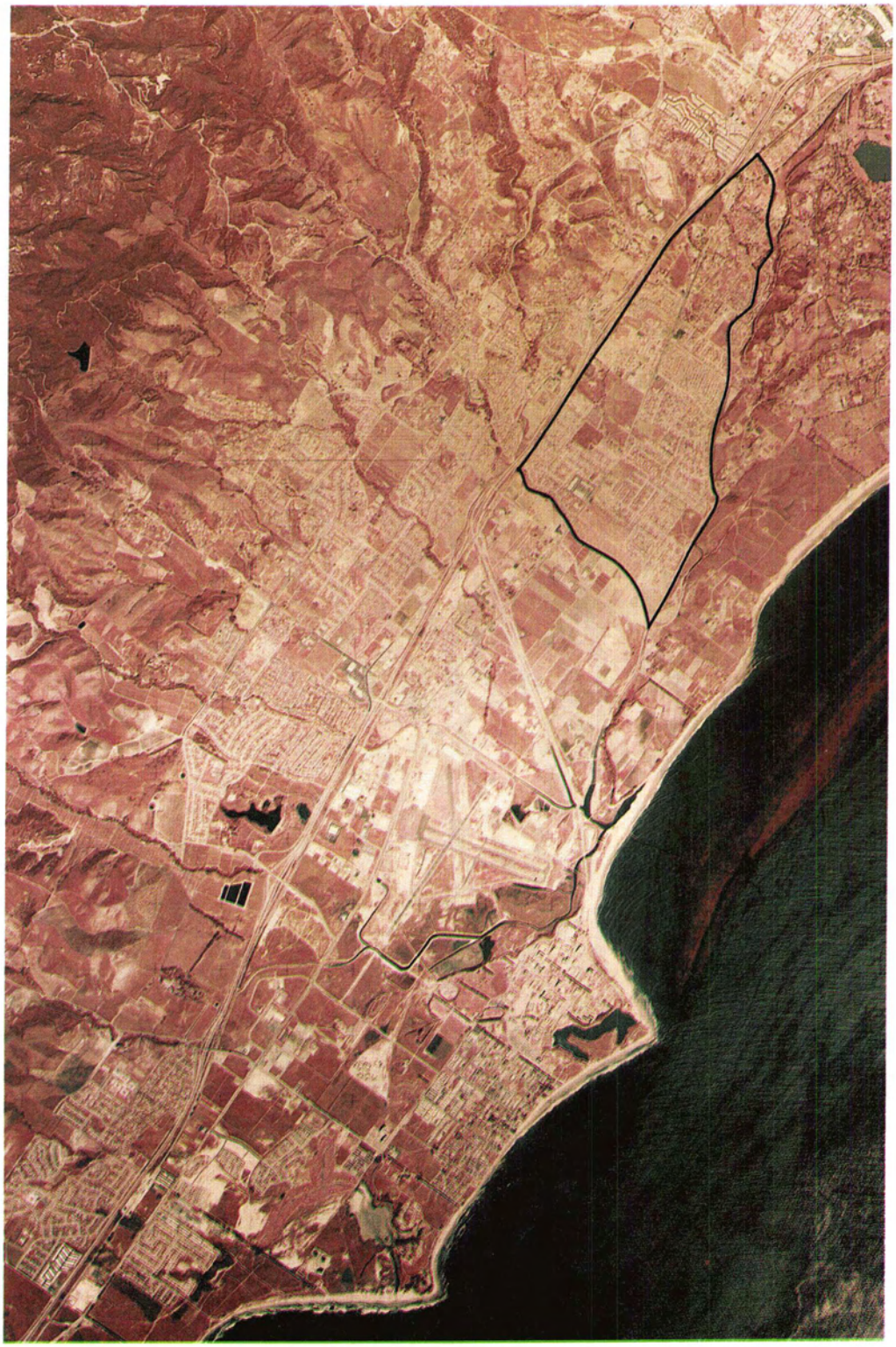


PLATE 1. The location of the study area, census tract 30.03 in the Santa Barbara, California, Standard Metropolitan Statistical Area, has been outlined on the NASA high altitude color infrared aerial photograph taken in April 1971 at a contact scale of 1:63,360. This photo is illustrative of the type of imagery from which our estimates of numbers of residential units were made.

data were obtained from the appropriate official Federal Census Report for 1970.[15] Third, housing counts were also made from detailed land-use maps compiled by the Santa Barbara County Planning Department. This latter source was also checked with the aid of low altitude aerial photographic coverage (at a scale of 1:7200) of the study area provided to the county by a corporation which commercially markets photographic and map atlases of urban and suburban areas.[16] From our analysis of the data from each of the three sources, it was our opinion that the detailed data from this latter source contained the most reliable information and as a result it was this County Planning Department generated information that acted as the standard against which other enumerations were compared. These maps are updated on a relatively frequent basis, usually once a year.

Analysis

For the total of 65 blocks contained in the census tract of interest the estimate of the number of single family type residential structures varied from 2108 for the imagery, 2132 for the land-use map, and 2223 for the census. The discrepancy between the census-derived and land-use map estimates is substantially greater than between the imagery-derived and land-use map estimates. The significance of this is confirmed by a chi-square goodness-of-fit statistic: 8.26 in the case of the imagery versus 737.48 in the case of the Federal Census.

Such discrepancies can be due to both under- and over-counting error. The magnitude of under- and over-counts is calculated from a comparison with the counts derived from the county landuse map. Discrepancies of all kinds are substantially less in the case of the imagery-derived figures (62 errors versus 558). It is interesting to note that, although the absolute number of miscounts is nine times as great on the part of the Federal Census when compared to the imagery, the net error is only three times as large (−32 versus +90). The difference in percentage error at the tract level between both sources is just over 5 percent of the total number of units present. The census had 38 percent more over- than under-counts whereas the imagery showed 213 percent more under- than over-counts. Thus, the results of this study indicate that relying on imagery as a data source is most likely to produce conservative estimates because of the relative preponderance of undercounting. Image interpretation procedures provided an estimate of the total number of dwelling units present that was 2.7 percent more accurate than the census (irrespective of sign). The above results are summarized in Table 1.

The relationships discussed immediately above are displayed in Figure 1. As can be seen, the scatter of observations is quite small and the correlations between both the imagery and census derived counts and that from the land-use map are both statistically significant at the $p = 0.001$ level. However, the r^2 value for the imagery is 0.99 while that for the census is 0.78. This confirms that the total number of units present in a given area using either estimation procedure is acceptable; yet the image interpretation method proved to be superior should highly site-specific data be required.

A detailed analysis was then performed on the data to examine the question concerning the statistical and spatial distribution of the tract level errors referred to above. Figure 2 illustrates the frequency distribution of the absolute discrepancies on a block by block basis for all 62 blocks for which data were uniformly reported. Information on three blocks was absent from the Federal Census report. Several points are worthy of note. The average census error is an overcount of 1.47 units per block while that of the imagery is an undercount of 0.39 units per block. The dispersion of values about these respective means is substantially different: almost twice as high for the census (see Figure 2). It can be seen that the accuracy of estimates at the tract level noted earlier between the two sources is explained in the

Table 1. Tabulation of Under- and Over-Counts by Information Source

	Imagery	Federal Census	Land-use Map
Total Number of Units	2108	2223	2132
Number of Over-counts	15	324	
Number of Under-counts	47	234	
Absolute Discrepancy	62	558	
Net Discrepancy	− 32	+ 90	
% of County Map Total	− 1.5	+ 4.2	

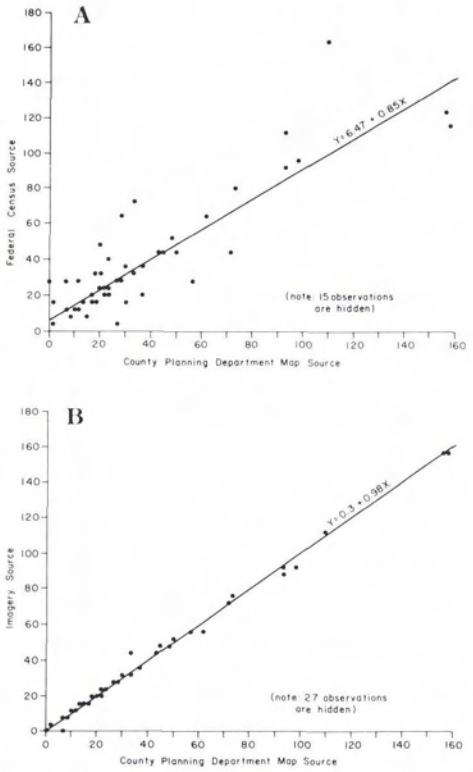


FIG 1. Bivariate plots of (A) number of structures as reported by Federal Census vs. number reported by county planning department; and (B) number of structures as interpreted from high altitude color infrared imagery vs. number reported by county planning department. Scatter of observations about the respective regression lines is substantially greater in the case of the Federal Census.

case of the Federal Census by the relatively even spread of errors above and below the mean. Thus, positive discrepancies are cancelled out by negative ones. Also, the chance of deriving extremely inaccurate estimates is much less with the use of imagery than with the Federal Census information source.

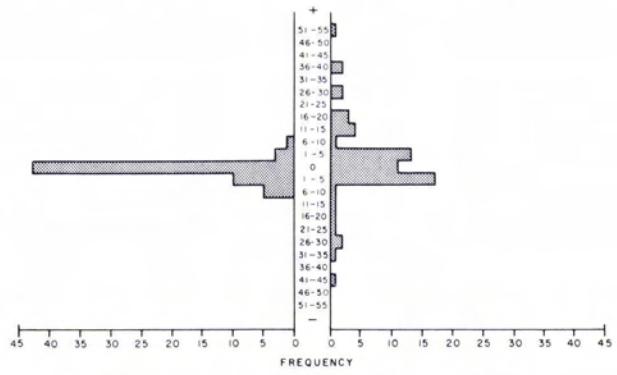
The next point of interest concerns the spatial distribution of the discrepancies. Figures 3A and 3B illustrate the situation with choropleth maps for both data sources. It is immediately evident that the areal extent of high levels of inaccuracy is substantially greater for the Federal Census data as opposed to the data interpreted from the imagery. A comparison between the maps shows relatively little spatial correspondence. One note of particular interest, however, is that certain types of blocks have high errors associated with them. These are mainly the irregularly shaped blocks which act as spatial infillers and often contain structures only on their outer boundaries with much of the interior area devoid of structures. Federal Census data appear especially prone to this type of error. Figures 4A and 4B emphasize the areal extent of the highest discrepancies. There is correspondence only in the case of one block.

The last question to be raised is whether or not there appears to be any systematic manner in which positive and negative discrepancies are related. As noted earlier, although the absolute number of discrepancies on the part of the census is high, the net frequency is relatively low. This implies that housing units which are being incorrectly included in one block are also being incorrectly excluded from their true block location.

Since no record is available concerning the techniques and methodologies of this specific enumeration conducted by the Federal Census takers, any explanation attempting to pinpoint the cause of error and its probable location is somewhat speculative. The boundaries of the census blocks are typically streets or other identifiable physical objects such as railroad lines and creek beds. The number of residential structures can be either under- or over-enumerated because of two main types of error: (a) structures on both sides of the block boundary are counted rather than on the interior side alone; and (b) substantial areas are inappropriately included or excluded from the block. Some possible examples of such errors are illustrated in Figure 5. Cases where counting on both sides of the boundary appears to have occurred are F, G, and H, with the rest of the cases suggestive of the inclusion of inappropriate areas from adjacent blocks. Although cases such as these appear plausible, no systematic pattern of under- and over-counting was detected. Based on our analysis of the data from this study, then, it does not appear that errors in enumeration have any discernible and systematic occurrence.

SUMMARY AND CONCLUSION

The major findings of this study can be summarized as follows: First, manual image analysis of high altitude aerial photography can provide a more accurate estimate of the number of single family dwelling structures than can alternate sources of information. Second, counts derived from imagery will most likely be on the conservative side, i.e., under-counting will prevail. Third, the errors associated with imagery-derived data



IMAGERY		FEDERAL CENSUS
-0.39	*Mean*	1.47
2.22	*Standard Deviation*	15.49
569%	*Coefficient of Variation*	1056%

FIG. 2. Frequency distribution of under- and over-counts at the census block level for both imagery (on the left) and Federal Census (on the right). Note the smaller variance in the error associated with the imagery-derived counts shown on the left.

can be less in both magnitude and in variability. Fourth, independently derived estimates may provide similar degrees of accuracy at one spatial scale and highly dissimilar estimates at another scale. Last, although suggestions can be made concerning how the errors were made in this particular case, no systematic spatial bias in the error was discernible.

It must not be inferred from the above

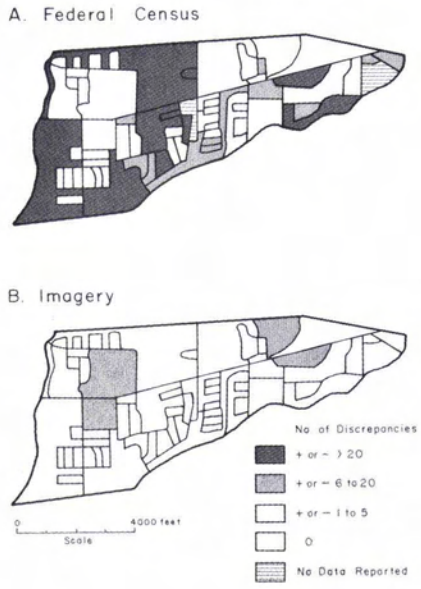


FIG. 3. Spatial distribution of the discrepancies for both Federal Census (A) and imagery (B) on a block by block basis. The number of errors enumerated and their areal coverage of the tract is significantly higher in the case of the Census.

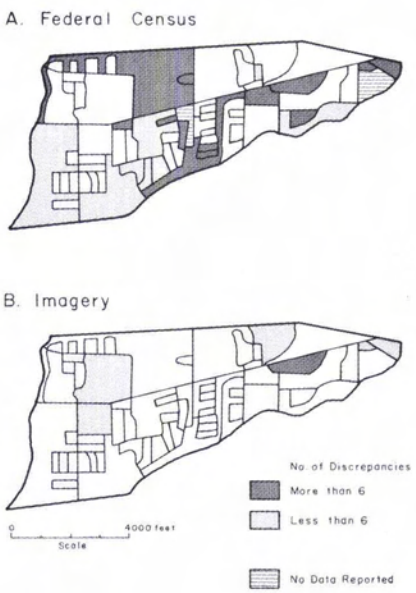


FIG. 4. Spatial distribution of excessive discrepancies for Federal Census (A) and imagery (B). As in Figure 3, the number and areal extent of Census errors is greater. The spatial coincidence between the discrepancies is low.

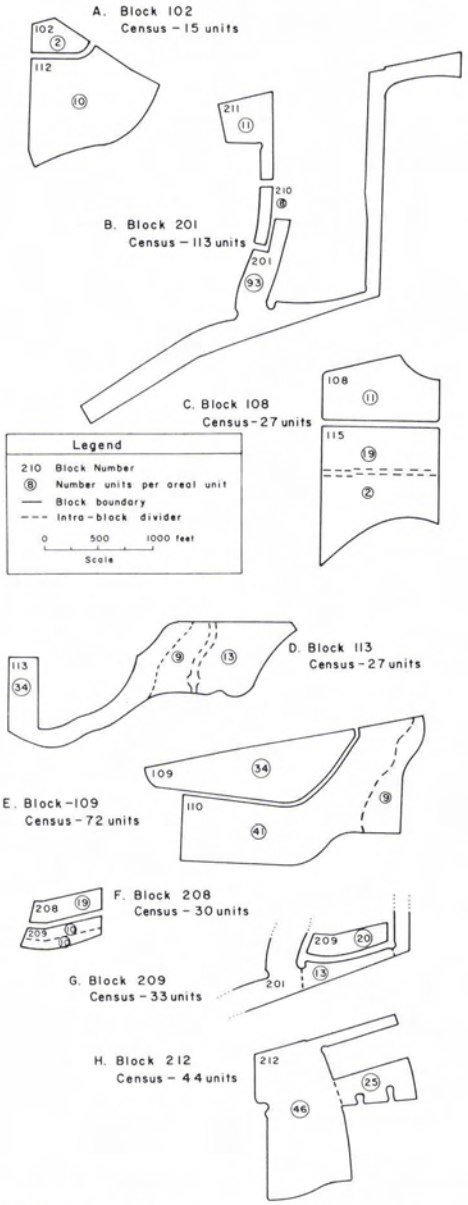


FIG. 5. Examples of some possible causes of errors present in the Federal Census enumeration. Cases F, G, and H illustrate errors that could have been made by incorrect inclusion of dwelling units on both sides of boundary streets. Other cases illustrate possible inclusion or exclusion of inappropriate areas.

study that the authors advocate the superiority of image-based data over census data for all uses. However, it certainly appears that, with regard to one type of information, namely, the number of residential structures per given spatially defined unit (in this case the census tract and block), image interpretation procedures are superior in providing a count of higher absolute accuracy than the Federal Census. The importance of this from a remote sensing standpoint, however, should not be underestimated because it is virtually impossible and practically infeasible to count individuals on remotely sensed data, and the role of surrogates for population estimates per given area becomes most important. For studies which employ remote sensing, the ability to count accurately the number of dwelling unit structures is extremely important because it allows estimates of socioeconomic and demographic parameters to be made. We suggest that imagery can be used successfully as a verification tool, a check that can indicate the presence and location of excessive error. It thus can provide a unique complementary data source of great value. This study has reported findings from a detailed analysis of all blocks contained within a single census tract. However, the types of error described above had previously been identified throughout six census tracts in the same general vicinity. It is true that this study focused on one type of urban environment alone. However, it is exactly in such environments where the vast majority of nonurban to urban land-use conversion is taking place and where population is growing most rapidly. For this reason the findings reported above take on considerable significance.

REFERENCES

1. United States Bureau of the Census, *Census Use Study, Report No. 4 - the DIME Geocoding System*, 1970
2. Clayton, C., and J. E. Estes, Distributed Parameter Modelling of Urban Residential Energy Demand, *Remote Sensing Quarterly*, Vol. 1, No. 1, 1979, pp. 106-115.
3. Green, N. E., Aerial Photographic Interpretation and the Social Structure of the City, *Photogrammetric Engineering*, Vol. 23, No. 1, March 1957, pp. 89-96.
4. Binsell, R., Dwelling Unit Estimation from Aerial Photography, in Westerlund, F. W., *Remote Sensing for Planning: A Bibliography and Review of Literature*, Department of Urban Planning, University of Washington, Seattle, 1972, pp. 39-40.
5. Mumbower, L., and J. Donoghue, Urban Poverty Study, *Photogrammetric Engineering*, Vol. 33, No. 6, June 1967, pp. 610-618.
6. Eyre, L. A., Census Analysis and Population Studies, *Photogrammetric Engineering*, Vol. 36, No. 5, May 1970, pp. 460-466.

7. Hsu, Shin-yi, Population Estimation, *Photogrammetric Engineering*, Vol. 37, No. 5, May 1971, pp. 449-454.
8. Lindgren, D. T., Dwelling Unit Estimation with Color-IR Photos, *Photogrammetric Engineering*, Vol. 37, No. 4, April 1971, pp. 373-378.
9. Kraus, S. P., L. W. Senger, and J. M. Ryerson, Estimating Population From Photographically Determined Residential Land Use Types, *Remote Sensing of Environment*, Vol. 3, 1974, pp. 35-42
10. Christensen, J. W., and H. M. Lachowski, *Urban Area Delineation and Detection of Change Along the Urban-Rural Boundary as Derived from Landsat Digital Data*, Preprint X-923-77-245, Goddard Space Flight Center, Greenbelt, Maryland, October 1977.
11. General Electric Corp., *Preliminary Design Requirements for Census/Urbanized Area Applications Systems Verification and Transfer*, Final Report prepared for NASA, Goddard Space Flight Center, Contract No. NAS5-23412, June 1977.
12. Friedman, S. Z., G. L. Angelici, and N. A. Bryant, The Detection of Urban Expansion from LANDSAT Imagery, Paper presented at the Annual Meeting of the Association of American Geographers, New Orleans, April, 1978.
13. Landini, A. J., and R. McLeod, City of Los Angeles Landuse Inventory and Population Updating, Paper presented at the Annual Meeting of the Association of American Geographers, New Orleans, April, 1978.
14. United States Bureau of the Census, *1970 Census of Housing, Block Statistics Report HC(3)-26*, p. iv.
15. *Ibid.*
16. Real Estate Development Corp., Inc. Imagery Scale is 1:7200.

(Received 23 May 1978; revised and accepted 18 December 1979)

Notice to Contributors

1. Manuscripts should be typed, double-spaced on $8\frac{1}{2} \times 11$ or $8 \times 10\frac{1}{2}$ white bond, on *one* side only. References, footnotes, captions—everything should be double-spaced. Margins should be $1\frac{1}{2}$ inches.
2. Ordinarily *two* copies of the manuscript and two sets of illustrations should be submitted where the second set of illustrations need not be prime quality; EXCEPT that *five* copies of papers on Remote Sensing and Photointerpretation are needed, all with prime quality illustrations to facilitate the review process.
3. Each article should include an abstract, which is a *digest* of the article. An abstract should be 100 to 150 words in length.
4. Tables should be designed to fit into a width no more than five inches.
5. Illustrations should not be more than twice the final print size: *glossy* prints of photos should be submitted. Lettering should be neat, and designed for the reduction anticipated. Please include a separate list of captions.
6. Formulas should be expressed as simply as possible, keeping in mind the difficulties and limitations encountered in setting type.

Journal Staff