2004

Mapping population density using a dasymetric mapping technique

Rachel R. Trusty

San Jose State University

MAPPING POPULATION DENSITY USING A DASYMETRIC MAPPING TECHNIQUE

A Thesis

Presented to

The Faculty of the Department of Geography

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Rachel R. Trusty

December 2004

UMI Number: 1425487

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

APPROVED FOR THE DEPARTMENT OF GEOGRAPHY

Richard Taketa

Dr. Richard Taketa

Gary Pereira

Dr. Gary Pereira

Leonard Gaydos

Dr. Leonard Gaydos, U.S. Geological Survey

APPROVED FOR THE UNIVERSITY

Thea I. Williamson

ABSTRACT

MAPPING POPULATION DENSITY USING A DASYMETRIC MAPPING TECHNIQUE

by Rachel R. Trusty

Demographic data are commonly represented by using a choropleth map, which aggregates the data to arbitrary areal units, causing inaccuracies associated with spatial analysis and distribution. In contrast, dasymetric mapping takes quantitative areal data and attempts to show the underlying statistical surface by breaking up the areal units into zones of relative homogeneity. This thesis applies the dasymetric mapping method to the 1990 US Census block-group populations of Alameda County, California, using the US Geological Survey's 1992 National Land Cover Data Set and other ancillary land-cover sources to redistribute the block-group populations into a 30-m grid based on categorical zones relative to population distribution. To test the accuracy of the dasymetric approach, census block populations were compared with the dasymetric mapping distributions; the results yield high correlation coefficients (between 0.80-0.88), indicating that the dasymetric mapping method produced more accurate population distributions than the choropleth method relative to the census block.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 STATEMENT OF PROBLEM

Demographic data and socioeconomic information are commonly displayed cartographically using choropleth mapping techniques. For example, the choropleth map is used to display US Census data, a geographic standard for demographics, and is used as a medium by virtually all geographers and many non-geographers (Slocum and Egbert 1993). The choropleth map spatially aggregates data into geographic areas or areal units (e.g., county, census tract, block-group). If the spatial units are too large, the data's spatial variation tends to be reduced or averaged out. Since the value in the enumeration unit is spread uniformly throughout the areal unit, continuous geographic phenomena cannot be displayed. Dorling (1993) noted that choropleth maps of population by administrative areal unit give the notion that population is distributed homogeneously throughout each areal unit, even when proportions of the region are, in reality, uninhabited. This discrepancy is greatest in areas with mixed urban, undeveloped, and agricultural land uses.

The US Census Bureau (2001) collects demographic data at the block geographic level. Census blocks are areas bounded on all sides by visible features, such as streets, roads, streams, and railroad tracks, and by invisible boundaries, such as city and county limits. In areas where there is a tight road network, census blocks are generally small in area. However, census blocks in sparsly populated areas may be large and contain many square miles. The block group is the next geographic level of US Census delineations, consisting of a cluster of census blocks generally containing from 600 to 3,000 persons

(US Census Bureau 2001). The population of each block group is an aggregate of the cluster of blocks. The boundaries of the census delineations are chosen on the basis of linear features and administrative boundaries, causing discrepancies between the enumeration units and the relevancy of population distributions.

Dasymetric mapping is a potential solution for the dilemma of portraying population data that have been aggregated to areal units. Eicher and Brewer (2001, 125) stated that, "Dasymetric mapping depicts quantitative areal data using boundaries that divide the mapped area into zones of relative homogeneity with the purpose of best portraying the underlying statistical surface." This type of mapping has been described as an intelligent approach to choropleth mapping in an attempt to improve area homogeneity. Thus, new zones are created that directly relate to the function of the map, which is to show spatial variations in population density. Land-cover data can indicate residential areas for the delineation of new homogeneous zones. The census block-group populations can be redistributed to the new zones, resulting in a more accurate portrayal of where people live within an administrative boundary.

This study explores a surface-based representation of population, using a dasymetric mapping technique that incorporates land-cover classifications as a means to redistribute the original census block-group population value into a surface grid based on levels of urbanization and undeveloped land. Through areal interpolation, the distribution is depicted semicontinuously, where multiple datasets redefine the populated surface. The hypothesis argues that this method will provide a more accurate representation of where people live in Alameda County, California, within a given block group than would choropleth maps of the same area. The greatest improvements in the accuracy of population-density values should be seen in block groups with various land-cover types and significant amounts of undisturbed land. In block groups that are heavily urbanized, the dasymetric mapping method may not show much difference from the choropleth method owing to the smaller size of block groups and the better correlation of the block-group boundaries in these areas to the actual population distribution.

To determine the accuracy of the dasymetric mapping technique, the 1990 US Census blocks were analyzed to see how closely the population density of the urbanization zones by block-group match the populations of the census block. The hypothesis concurs that the block groups on the dasymetric map results will show a statistically superior match to the census-block populations over those on the choropleth map.

CHAPTER 2: LITERATURE REVIEW

2.1 DASYMETRIC MAPPING APPROACHES

Adequate research has been done on the dasymetric map, but the complexity of this method compared to the choropleth method deters cartographers and nongeographers from using the dasymetric map. Even though the nature of a population distribution is more realistically represented with a dasymetric map, the choropleth map is more commonly used due to the complex methodology of the dasymetric map (Charpentier 1997).

The earliest study documented is by John K. Wright, written in 1936, about the methods he used to map population density of Cape Cod, Massachusetts. Wright (1936) mentions the term dasymetric as a map type Russian geographers coined meaning, "density measuring." Wright used a dasymetric mapping method to show population density of Cape Cod by excluding uninhabited areas and by weighting the original 1930 census values through interpolation. He argued that his method was based on many assumptions, such as the delineation of uninhabited areas and "rough-and-ready" methods of dividing the townships into tracts of different densities. His "rough and ready" assumptions for the new zones were centered on USGS topographic maps and personal recollection. Following the uninhabited-area delineations, density measurements were given to different parts of the township, using "controlled guesswork" as a subjective method. Wright's method allocated estimated densities by interpolation to areas whose subdivisions had no statistical data available.

An essential step in dasymetric mapping is the creation of zones within the areal unit that correspond to the variable being mapped. To create intraunit zones of relative homogeneity among population, ancillary data must be used to interpret relative levels of habitation. Past approaches have focused on using ownership records, topography, and land-cover classifications to identify and mask uninhabited areas. Holloway et al. (1997) used multiple datasets to detect and remove uninhabited lands from the area of analysis. Four types of area were ruled out, including census blocks with zero population, all lands owned by local, State, or Federal government, all corporate timberlands, and all water or wetlands, as well as all open and wooded areas with elevation data that have a slope at most 15% (Holloway et al. 1997). To redistribute the census population to the ancillary feature classes, a predetermined percentage was assigned to each class. The subjectivity and accuracy of this percentage assignment (e.g., 80% of the population to urban polygons, 10% to open polygons, and 5% to agriculture and wooded polygons) can be argued because of the absence of empirical evidence.

In contrast, Mennis (2003) used a three-tier raster classification of urban land cover derived from the Landsat Thematic Mapper (TM) as ancillary data. Within the remotely sensed land-cover data, urban features were put into three classes of high density, low density, and nonurban, with no distinction of wooded areas, agriculture, or slope. Initially, all census data were converted into a 100-m raster surface that was used for areal interpolation. The population statistics derived from the 1990 Census block-group data were distributed via areal interpolation to each 100-m grid cell on the basis of two factors: "the relative difference in population densities among the three urbanization classes…and the percentage of total area of each block group occupied by each of the three urbanization classes" (Mennis 2003, 36). An empirical sampling of population density between urbanization classes helps determine what percentage of the census block-group population should be assigned to each urbanization class. Also, an area-based weighting addresses the relative differences between each urbanization class within the census block group.

2.2 DATA MODELS and REPRESENTATION

Mapping techniques and models can differ based on the structure and theme of a spatial dataset. For example, a choropleth map is the preferred map model to use when the geographic theme portrays data that occur within well defined enumeration units, for example, statistical administrative boundaries showing a rate or ratio representative of the given boundary. However, if the variable changes within the enumeration unit the choropleth map cannot detect this change. To map average annual temperature, or elevation data, due to the gradual increase or decrease in the nature of the data, an isopleth map would be appropriate because isolines connect the points of equal value (e.g. contour lines). MacEachren (1994) situated the dasymetric map in the continuum between isopleth and choropleth maps, suggesting that dasymetric maps represent data half way between smooth and stepped statistical surfaces. A stepped statistical surface would represent events that occur in isolated areas separated by areas where the phenomenon is not present (MacEachren 1994). A smooth statistical surface would represent data that has a continuous nature, like elevation. To track change within an enumeration unit the dasymetric mapping model can detect the inherent principal statistical distribution.

The relative merit of object versus field models for quantitative representation is a subject of ongoing debate in the fields of cartography and geography. Michael Goodchild (1992) has written extensively about object versus field models in a geographic information system (GIS). In the object model, features are generally represented as points, lines, or polygons, and so this mode is known as the vector data model. The field model, which typically represents square features as a set of uniform-sized cells, is known as the raster data model. The advantages and disadvantages for visualization and quantitative representation in both models have become evident and depend on the scale and quality of the data. Mennis (2003) determined that a field representation of population data, where the data are modeled onto a continuous surface, works well with the transformation of population data from census block groups.

For this study, a combination of the object and field models was used. If the object represents the same area as the field, then the two models can be used interchangeably. For example, in a vector representation of points where each point represents 30-m, the points can be converted to 30-m pixels without a loss of data. This conversion meets the ideal of the pycnophylactic property (Tobler 1979), where "the summation of population data to the original set of areal units is preserved in the transformation to a new set of areal units" (Mennis 2003, 32). Therefore, the Modifiable Areal Unit Problem is avoided during the areal interpolation.

2.3 MODIFIABLE AREAL UNIT PROBLEM (MAUP)

The MAUP is a potential source of error that can affect spatial studies which utilize aggregate data sources (Unwin 1996). The MAUP is most prominent in socio-economic studies where areal data cannot be measured at a single point, but within an areal boundary. The effects of MAUP can be divided into two components: scale and zonation. The scale factor is the variation in results when data is grouped into different levels of spatial resolution (e.g., regions, cities, census tracts, census block-groups). The zonation factor is the variation in results due to the aggregation of smaller units into larger units (e.g., census blocks aggregated into census block-groups). By using a dasymetric mapping method, zones are created that have a closer spatial relationship to the objective of the map, than the given boundaries. When original block-group census values have to be redistributed to the new intraunit zones, as in the dasymetric mapping method the zonation factor becomes important because the accuracy and relative qualities of the zones chosen have a fundamental effect on the statistical outcome. In other words, if the data were grouped into alternate zones, there may be excessive spatial and statistical variability in the results. Therefore, the quality and accuracy of the ancillary data highly influences dasymetric mapping results.

The dasymetric mapping technique tends to remove the ecological fallacy associated with enumeration units. Ecological fallacy is a situation that can occur when an analyst makes an inference about an individual based on aggregated data for a group. For instance, if a census block-group had a population density of 1000 persons/per sq. mi. it would be true and accurate to say that this value is the average population density for the given block group. On the other hand, if the analyst were to pick out a particular square mile within that block-group and apply the same ratio of 1000 persons/per sq. mi. to that area, the statement becomes inaccurate which introduces the ecological fallacy.

2.4 NATIONAL LAND COVER DATASET (NLCD)

Land-cover data has become widely available due to the increase of GIS users and the demand for land-cover change analysis. Land-cover data can be collected and mapped in a variety of ways; however, remotely sensed imagery seems to be the most popular source for classifying land cover. The US Geological Survey (USGS) provides a land-cover dataset derived from Landsat TM imagery known as the NLCD. This 1992 dataset was created by classifying 30-m resolution Landsat TM into 21 land-cover categories based on the Anderson classification (Vogelmann et al. 1998; Anderson 1976). By using NLCD as the primary ancillary dataset for a dasymetric map, the process of image classification and accuracy assessment have been completed with a well defined error matrix. However, the geographer needs to be aware of the spatial and spectral accuracy of the land-cover data being used. The accuracy assessment of NLCD has been conducted region-by region using a scientifically rigorous approach, and meets USGS data requirements for applications at the regional to continental scale (Vogelmann et al. 2000). This geo-spatial dataset is also available nationwide to download, free of charge. Although the process of creating urbanization zones becomes simplified when using NLCD as the input land-cover dataset, the accuracy of the data at the scale of the study becomes arguable. In this study, NLCD is under assessment at the county level factoring into the quality and accuracy of the final dasymetric map.

CHAPTER 3: METHODOLOGY

3.1 STUDY AREA

The Alameda County, California study area, which is part of the greater San Francisco Bay region, was chosen because the familiarity of the area enhances the depth of understanding that can be brought to the spatial relations being analyzed (Figure 3.1). As Eicher and Brewer (2001, 125) pointed out: "The cartographer generates dasymetric zones by using ancillary information. This information can be both objective and subjective, depending on other available data and the cartographer's knowledge of the area." Furthermore, Alameda County has a widely varying topography and a mix of land-cover types from undeveloped and agricultural to heavily urbanized. Alameda County is home to the city of Oakland, the eighth-largest city in California, with a 2000 US Census population of 399,484, and the Port of Oakland, one of the major container ports on the west coast of the United States. Many city centers within Alameda County contribute to the greater San Francisco Bay region, such as

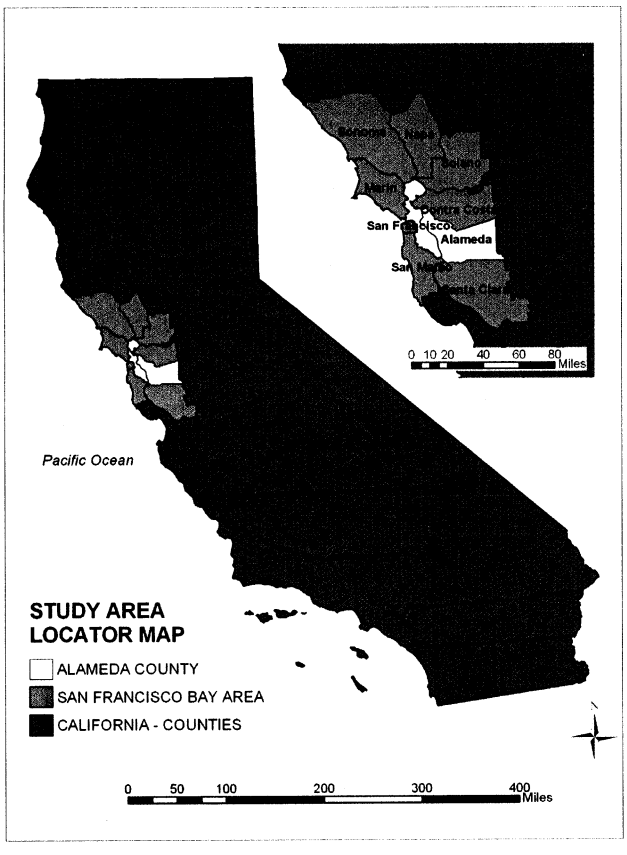
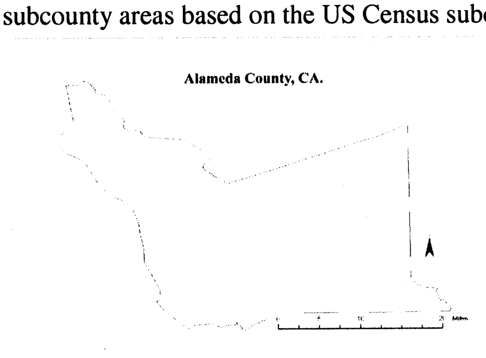


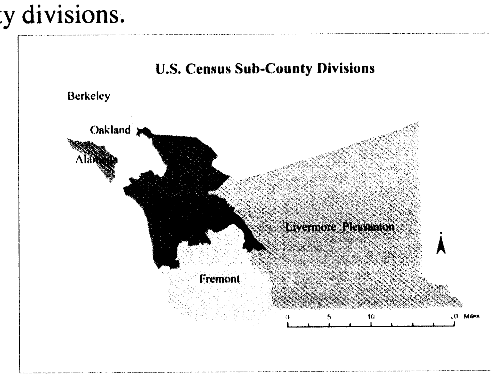
Figure 3.1. California, showing location of Alameda County study area.

Livermore, Pleasanton, Berkeley, Alameda, Hayward, and Fremont. Conversely, the network of East Bay Regional Parks snakes through Alameda and neighboring Contra Costa counties, preserving 94,500 acres of open space. Although the west side of Alameda County is heavily populated and urbanized, the land cover changes drastically toward the southeast into rugged hills and toward the east into the agricultural landscape of the California Central Valley. Alameda County, which typifies the urban-rural fringe of the San Francisco Bay region, is an important but complex area to understand and portray demographically.

This study divides Alameda County into the US Census subcounty divisions to treat each urban core area as a separate entity and to model the comparative population clusters in the different parts of the county. By calculating each subcounty division separately, a conclusion can be drawn pertaining to the correlation between land-cover types and this dasymetric mapping technique. Within Alameda County, there are six subcounty divisions: Alameda, Berkeley, Oakland, Hayward, Fremont, and Livermore-Pleasanton (Figure 3.2).

Figure 3.2. Scale of research. Alameda County, Calif. is 738 sq. miles. Alameda is broken up into subcounty areas based on the US Census subcounty divisions.





3.2 DATA PREPARATION

This approach combined the methodologies of Mennis (2003) and Holloway et al. (1997) by choosing four land-cover classes, using a three-tier urbanization classification and adding an excluded class representing zero population. The 21-class National Land Cover Data Set (NLCD) (Volgelmann et al. 2000) were recoded into four classes; high-intensity residential, low-intensity residential, nonurban, and water: the nonurban class consists of all remaining 18 classes, representing all lands that are not residential but may be sparsely populated. The undeveloped layer, contributed by GreenInfo Network (2003), incorporates lands that have some recreational, open-space, habitat-protection, or agricultural-protection value in the San Francisco Bay region. These lands either are owned by a public agency or a nongovernmental organization (NGO), or have an easement on them held by a public agency or an NGO. The uninhabited layer was merged with the recoded NLCD layer to produce a comprehensive land-cover layer. From this dataset, the classes were merged and reconfigured into classes of high-intensity residential, low-intensity residential, nonurban and exclusion; the exclusion class is a combination of all water and undeveloped areas (see Table 3.1). The advantage of incorporating an exclusion class is to more accurately display population density by weeding out large areas of the areal interpolation, allowing the visual depiction of population to be strictly within those areas that are actually populated.

After the recoding process, the new zones of relative homogeneity were in a raster format. Before the areal interpolation, the raster land-cover data were converted to points for two reasons: (1) to provide an easy way to spatially join the land-cover dataset to the census block data and (2) to efficiently create and calculate new fields in a tabular

Table 3.1. Combining land-cover layers for recoding.

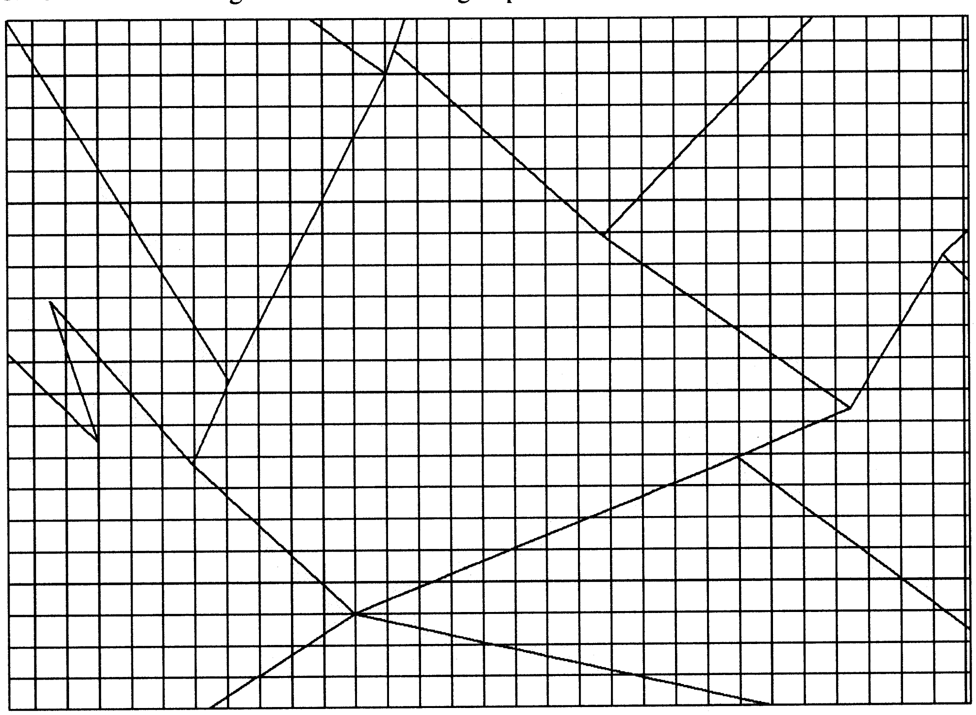
Class Code	Class Definitions for Re-coding	
0	No Data	
1	High Intensity	
2	Low Intensity	
3	Non Urban	
4	Water	
10	Open Space	
11	Open Space + High Intensity	
12	Open Space + Low Intensity	← Classes 10-14 re-coded to class code (4) after summation.
13	Open Space + Non Urban	
14	Open Space + Water	

structure. As mentioned above, this method allows the use of both raster and vector data interchangeably. At this stage, each point representing a 30-m pixel has an associated land-cover code but no census block-group information. To attach the census data, each point also needs an associated block-group identifier. This step is required for the dasymetric mapping approach because each calculation is performed on a block-group-by-block-group basis. The census polygon data are joined spatially to the land-cover points, and the data are now ready for areal interpolation.

3.3 RASTER /VECTOR PROBLEM

A potential problem can occur in the assignment of land-cover points to census block-groups. When the spatial join occurs, the land-cover points are assigned to census block-groups if the point falls completely inside the census block-group boundary. Since each point will ultimately represent the centroid of each 30-m pixel, the points that fall close to the block-group boundaries, when converted to pixels, may overlap with the census block-group boundary. The result causes slivers along the boundaries, making it unclear which block-group the full 30-m pixel belongs to (Figure 3.3). To avoid erroneous results, the centroid rule is observed throughout the project which indicates that if the center of the point falls within a given block group, the entire value of that point is attributed to that block-group. This geo-spatial rule, sometimes referred to as the centroid rule, becomes important when the 30-m points are converted back to 30-m pixels for the final demonstration.

Figure 3.3. Portions of the raster cells fall within adjacent census block-groups causing "slivers" of data falling in erroneous block-groups.



Within the USGS, Land Cover Trends Project, a similar problem occurs with the sample block boundaries and the ecoregion boundaries. "If a 20 km block contains more than one ecoregion, the block is assigned to the stratum identified at the block's center"

(Stehman et al. 2003, 3). The 30-m resolution pixel is much finer than the 20 km sampling block used in the Land Cover Trends project and when using the centroid rule, this only leaves up to 15-m slivers that do not necessarily belong to the assigned census block-group. In terms of population distribution, 15-m is a very small chunk of real estate compared to the overall distribution. An area calculation determined that 88% of the pixels fall completely within their assigned census block-group. Due to the census' generalized boundaries this shift becomes even less significant (Table 3.2).

Table 3.2. Sliver Problem

	Total # of Cells	# of "Slivers"	% of Potential Error
	198054	23848	12%
	Area of BG	Area of Cells Completely w/in BG	% Accuracy
Large BG			
60014090009	16168210.14	15728400	97%
Medium BG			
60014042003	1769698	1624500	91%
Small BG			
60014026002	25824	15300	51%
		Area of Cells 50%- 100% w/in BG	
Large BG			
60014090009	16168210.14	16086633	99%
Medium BG			
60014042003	1769698	1744194	98%
Small BG			
60014026002	25824	15300	85%

3.4 AREAL INTERPOLATION

Areal interpolation, which is the process by which data from one set of source polygons are redistributed to another set of overlapping target polygons, is used primarily when a project contains data from various sources covering the same area but with differing internal boundaries. This study adapted the four equations from Mennis (2003) (see Appendix A) to address the addition of a zero-population zone or exclusion class. The removal of the spatial area of the exclusion class from the total possible area of population distribution should contribute more accurate results overall relative to Mennis' approach because of the addition of a fourth class covering all areas of zero population, such as water.

To quantify the urban land-cover variable within each subcounty subdivision, a sampling method is used to calculate the relative difference in population density among urbanization classes within each unit (Mennis 2003). Three representative block groups were selected for each urbanization class (block groups that clearly had a majority of high, low, or non-urban points within them). The total population and area were calculated for each urbanization class sample and resulted in an aggregated population density (see Table 3.3).

The population density fraction is then calculated for each urbanization class within each subcounty. This number indicates the percentage of the block-group population that should be assigned to each urbanization class within the given block group (see Appendix A for equations). In order to alter the population density fraction values, an area ratio value for each block group was calculated showing the proportion of area that each of the three urbanization classes occupies within a given block group. This calculation was performed on every urbanization class within every block group. At this point, the exclusion class was inserted back into the analysis as the entire areas of each land-cover class within the block-group total were summed into area ratios including the exclusion class points.

Table 3.3. Sampled population-density fractions for the subcounties

Urban-Class	Subcounty	Area(mi²)	Populated Area (mi²)	Population Density (persons/mi²)	Sum of Density	Population Density Fraction
High	Oakland	0.0516	1,863	36,104.651	57,150.053	63.18
Low	Oakland	0.080782	1,180	14,607.214	57,150.053	25.56
NonUrban	Oakland	0.0817	526	6,438.188	57,150.053	11.26
High	Berkeley	0.108026	2,389	22,115.046	38,097.222	58.05
Low	Berkeley	0.070769	944	13,339.174	38,097.222	35.01
NonUrban	Berkeley	0.18199	481	2,643.002	38,097.222	6.94
High	Alameda	0.27888	2,715	9,735.37	25,626.627	38
Low	Alameda	0.068526	842	12,287.307	25,626.627	48
NonUrban	Alameda	1.09158	3,934	3,603.95	25,626.627	14
High	Hayward	0.106925	1,127	10,540.0981	19,209.8786	54.87
Low	Hayward	0.129875	1,116	8,592.8777	19,209.8786	44.73
NonUrban	Hayward	5.82553	448	76.9028	19,209.8786	0.4
High	Fremont	0.126942	1,070	8,431.6459	25,341.69808	33.27
Low	Fremont	0.08667	1,436	16,572.40106	25,341.69808	65.4
NonUrban	Fremont	6.35665	2,146	337.65112	25,341.69808	1.33
High	Livermore-Pleasanton	0.179661	1,029	5,727.453371	13,911.78012	41.17
Low	Livermore-Pleasanton	0.105797	865	8,176.035237	13,911.78012	58.77
NonUrban	Livermore-Pleasanton	12.42234933	103	8.291507285	13,911.78012	0.06

Next, the population-density fraction was multiplied by the area ratio to give the fraction of the original block-group population that was distributed to each urbanization class within each block group; then that result was divided by the sum of the same expression of all three urbanization classes. The total fraction is the underlying solution to the interpolation. Once the total fraction was calculated, part of the original block-group population could be assigned to each point within the block group according to its urbanization class. This calculation was the final step in the areal interpolation, resulting in a surface-based representation of population density (see Figure 3.4). The final distribution was completed by multiplying the total fraction of an urbanization class by the total block-group population and then dividing the result by the number of points within the urbanization class (see Appendix A). The utilization of a point-feature class in a GIS made the areal interpolation efficient because each new field could be created and calculated in a semiautomated mode, using the GIS field calculator.

Figure 3.4. Aggregated Census Value vs. Surface Population Relative to Spatial Land-Cover

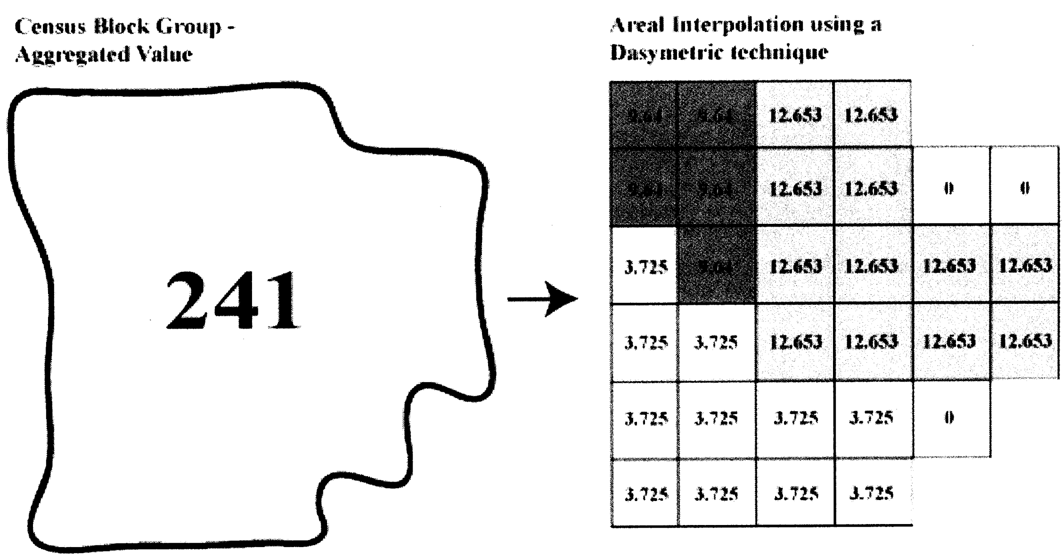


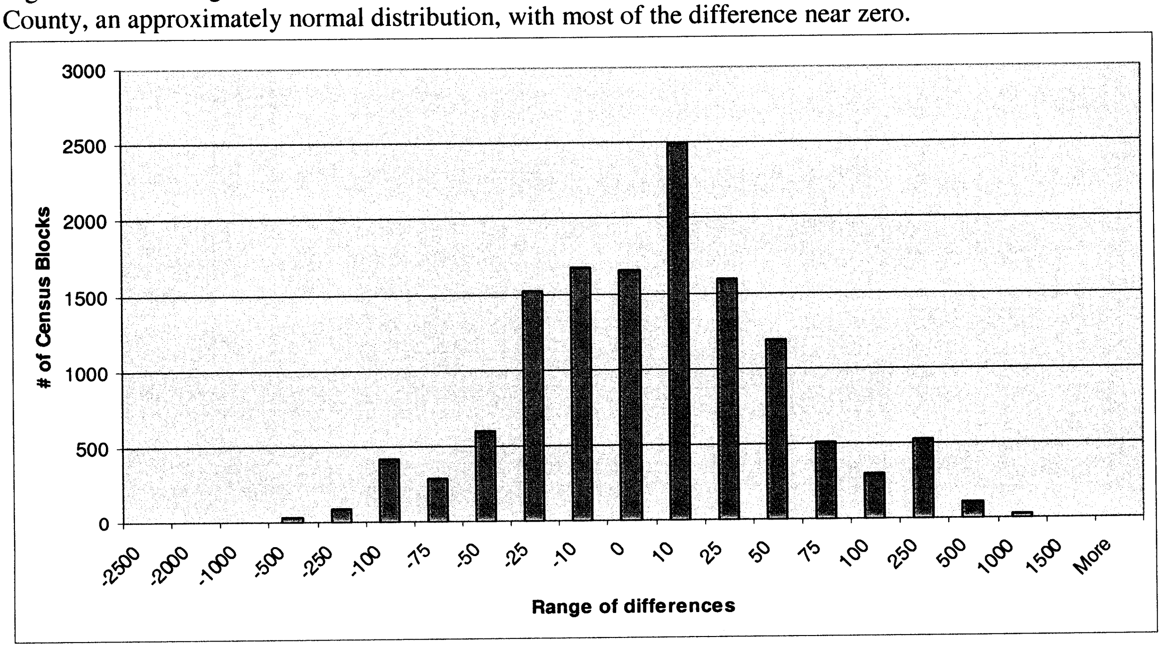
Figure 3.4. This graphic is an example of the input and output values of areal interpolation used in this study. The dasymetric technique uses land-cover urbanization classes and relative area and density to redistribute the block-group population to a surface. Dark gray = high-intensity urban, medium gray = low-intensity urban, light gray = non-urban, and white = exclusion or population zero.

CHAPTER 4: ANALYSIS

4.1 COMPARING CENSUS BLOCK POPULATIONS TO DASYMETRIC MAPPING RESULTS

The hypothesis is that the dasymetric mapping method will more accurately represent where people live within a given block group. To test the precision of the dasymetric map distribution, the census block totals were evaluated as an indicator of how well the population was distributed within the block groups. The dasymetric map consists of 30-m points, each representing a population total for that area. To obtain a value comparable to the block populations, a sum of the dasymetric points that are located within a given block was generated. Figure 4.1 is a histogram of the absolute differences between the two variables, calculated by subtracting the block populations from the summation of all the dasymetric points for each block. If the difference between

Figure 4.1. Showing difference between block populations and dasymetric comparative values for Alameda County, an approximately normal distribution, with most of the difference near zero.



the block populations and the dasymetric comparative values equals zero, the results indicate that the dasymetric map distributions equals the census-block population. The histogram shows an approximately normal distribution, with most of the difference values near zero, affirming the claim that the dasymetric map preserves an accurate block-level summation of census population data. Appendix B includes the histograms and scatterplots for each subcounty comparison.

4.2 CORRELATION COEFFICIENTS

To further test a positive association between the block-population totals and the dasymetric mapping distributions a correlation analysis was conducted. The correlation coefficient, denoted by *r* of the pairs (*x*, *y*) is calculated as

$$r = \frac{\Sigma d_x d_y}{\sqrt{(\Sigma d_x^2 \Sigma d_y^2)}}$$

Here the strength of the relation between the estimated dasymetric population per block and the observed block population is tested by using a bivariate or simple correlation analysis (Burt and Barber 1996). The hypothesis requires a positive correlation between the two arrays, which would indicate that "the relationship between *x* and *y* is such that small values of *y* tend to go with small values of *x* and large values of *y* tend to go with large values of *x*" (Freund and Simon 1997, 528). The correlation coefficients for each subcounty are high, ranging from 0.80 to 0.88 (see Table 4.1). This statistic can be interpreted as a standardized measure of areal association and the degree of similarity of the two maps in the individual statistics (Burt and Barber 1996).

Another correlation coefficient was calculated to compare the choropleth map of block-level summations derived from block-group population densities to the actual census-block population densities. The results support the initial hypothesis. The subcounties that scored the lowest in the choropleth-to-block comparison were Alameda, Fremont, and Livermore-Pleasanton (Table 4.2). These three subcounties also have a lower ratio of urbanized to nonurban and undeveloped land cover (see Table 4.3). The hypothesis is that the dasymetric mapping method would be more effective in areas with more land-cover variation and less concentrated urbanization as was true for all but one subcounty (Hayward), which is highly urbanized but has some large undeveloped areas. The large undeveloped area in Hayward may have contributed to the lower percentage of residential land cover.

Table 4.1. Correlation coefficients for each subcounty

Subcounties	
Berkeley	0.84
Oakland	0.82
Alameda	0.88
Hayward	0.87
Fremont	0.87
Livermore_Pleasanton	0.80
Alameda County	0.85

Table 4.2. Correlation coefficient comparisons by Alameda's subcounties

	Dasymetric : Block	Choropleth : Block
Berkeley	0.84	0.83
Alameda	0.88	0.67
Oakland	0.81	0.79
Hayward	0.87	0.79
Fremont	0.87	0.56
Livermore-Pleasanton	0.8	0.57
ALAMEDA COUNTY	0.85	0.7

Table 4.3. Area ratio between residential and non-urban/undeveloped by subcounty (in pixels)

SubCounties	Residential	Total	RATIO	% Residential
Berkeley	27482	33691	0.815707	81.57
Alameda	14575	36407	0.400335	40.03
Oakland	94215	174206	0.540825	54.08
Hayward	103602	380763	0.272091	27.21
Fremont	94850	338532	0.28018	28.02
Livermore-Pleasanton	62553	1206282	0.051856	5.19
ALAMEDA COUNTY	397277	2169881	0.183087	18.31

4.3 F-TEST TWO SAMPLE VARIANCES

Theoretically all census blocks nest perfectly within their associated block-groups indicating that the statistical mean of the two arrays should be the same. After looking at the block population mean and the dasymetric mean for all sub-counties, it is clear that they are almost identical (Table 4.4). The noticeable difference between the observed block populations and the estimated block populations was thus seen in the variance of the data. The F-Test is designed to test if the two sample variances are equal. The null hypothesis is that the variances equal one another at the $\alpha/2= 0.025$ significance level.

Table 4.4.

Statistical MEAN	Dasymetric Population Mean	Block Population Mean	Differences (Est-Obs) Mean
Berkeley	85.86756541	86.23277737	-0.365211965
Oakland	87.71772569	87.59026048	0.12746521
Alameda	103.7963762	103.8713018	-0.074925592
Hayward	108.141221	108.1579912	-0.016770198
Fremont	134.0055249	134.0862944	-0.080769492
Livermore_Pleasanton	79.10778605	78.98947368	0.118312368
ALAMEDA COUNTY	97.52577375	97.55012281	-0.024349064

The alternative hypothesis is that the variances do not equal each other at the $\alpha/2= 0.025$ significance level. The results showed that only two subcounties out of the six meet the null hypothesis which concludes that those areas have equal variances at the $\alpha/2=0.025$ significance level (Table 4.5).

Table 4.5.

$H_0: \sigma^2 1 = \sigma^2 2$			
$H_a: \sigma^2 1 \neq \sigma^2 2$			
Reject H_0 if $F > F_{\alpha/2}$			
F-Test Two Sample Variances $\alpha/2= .025$	$F_{\alpha/2}$ - Critical Value	F - Statistic	Reject/Accept: H_0
Berkeley	1.11	1.29	Reject
Oakland	1.06	1.01	Accept
Alameda	1.16	1.18	Reject
Hayward	1.07	1.05	Accept
Fremont	1.09	1.23	Reject
Livermore_Pleasanton	1.1	1.27	Reject
ALAMEDA COUNTY	1.03	1.15	Reject

Oakland and Hayward had the lowest variances to the block populations indicating that the dasymetric mapping technique demonstrated the finest results in these areas. The hypothesized estimation asserted that the best results would be seen in areas that have higher land-cover variation and low levels of urbanization, which is not the case in Oakland or Hayward. Livermore-Pleasanton has the most variation in land-cover, with the most undeveloped and uninhabited areas, but showed the highest variances. The comparative statistics showed that smaller blocks seen in highly urbanized areas produce closer estimations to the block boundaries because the areal range of distribution is smaller allowing less distribution error.

For example, one of the highest variances within the Livermore-Pleasanton subcounty was within block group '060014507222'. The population of the block group is 4558 people (see Figure 4.2a). Within this block group lie 16 blocks with different populations all adding up to 4558 people. The population of 4558 was distributed according to the areal interpolation method, relying heavily on the land-cover data. For the reason that the block-group population is so high (4558), the distributions made could result in blocks that have a low population. In Figure 4.2b the block population of the highlighted block is 18 people, and the estimated dasymetric population is 902 people. This difference yielded a high variance which is why some of the subcounties did not pass the variance hypothesis testing. This error in the population distribution can be attributed to the accuracy of NLCD at the block level. This is one major drawback of working with NLCD when testing results with something as fine as the census block level geography.

4.4 DISCUSSION

Figure 4.1 showed that the deviations between block and block-group totals aggregated via dasymetric techniques were approximately normally distributed indicating minimal difference between the two datasets. Also, all of the correlation coefficients exceeded 0.80. The correlation coefficients between the choropleth map and the block populations ranged from 0.56 to 0.83, unambiguously lower than for the dasymetric map confirming the hypothesis that the dasymetric mapping method of representing block-group population density was more accurate than the choropleth mapping method.

The dasymetric map also produces a superior visual enhancement of the data, a fact most evident when focusing on the water features (see Figure 4.3 and Figure 4.4). For instance, the city of Alameda is an island on the north western part of the county that is adjacent

FIGURE 4.2a. Block group with population of 4558. Population is interpolated and distributed to the land-cover classes within the block group.

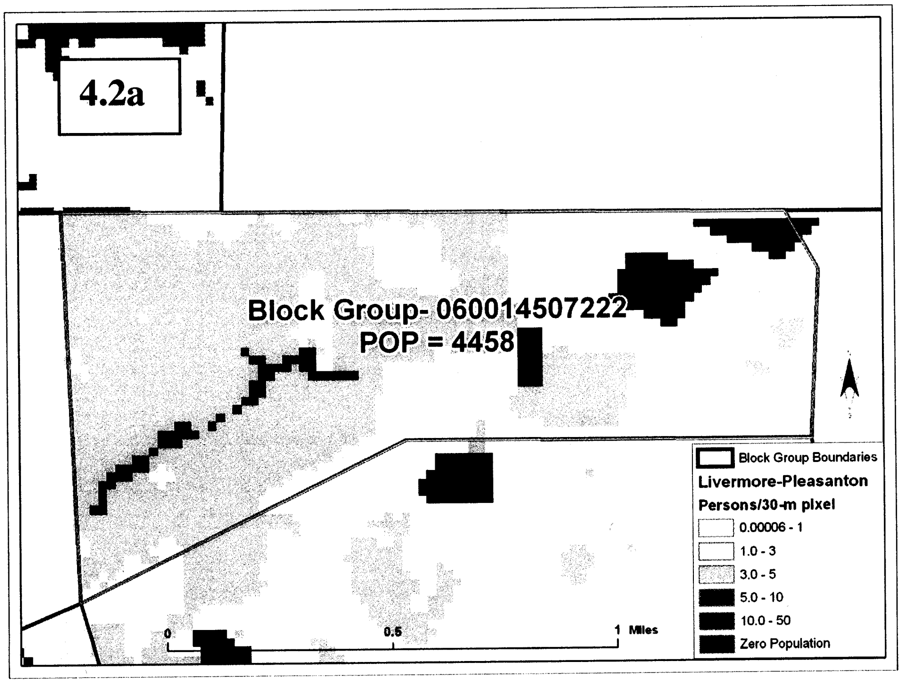
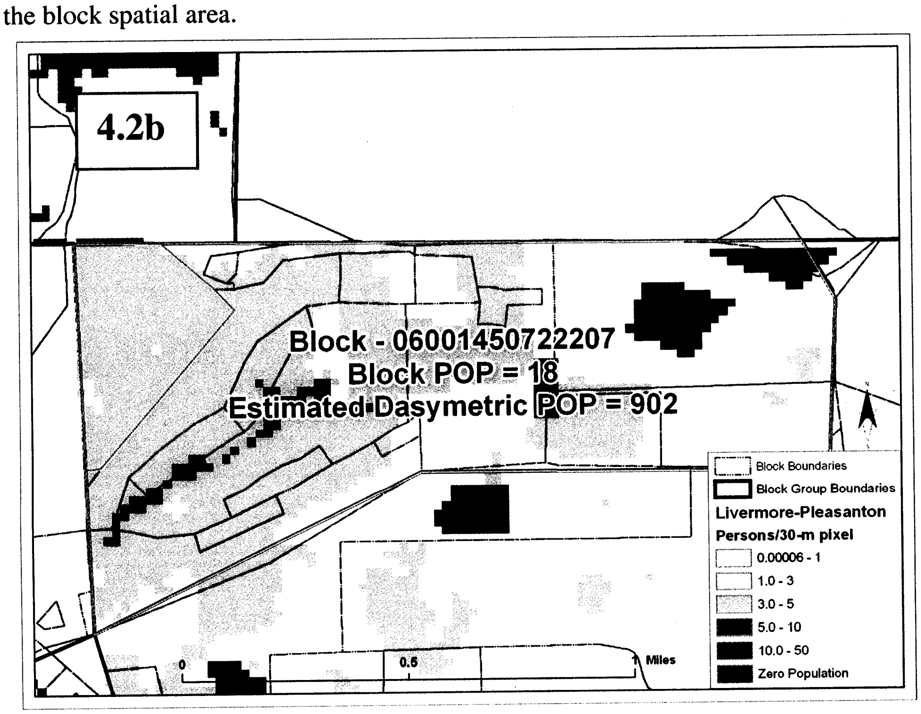


FIGURE 4.2b. If land-cover is not accurate enough, the population is not distributed to the correct areas resulting in high variances when tested against the block populations. The highlighted block in the upper left has a population of 18 people whereas the dasymetric mapping technique interpolated 902 people for the block spatial area.



to San Francisco Bay, which the choropleth map conceals entirely and also shows non-zero population levels where there's water. Also, Lake Merritt, the major urban water feature of Oakland, appears to be populated in the choropleth map but the dasymetric map correctly identifies this area as uninhabited. Other features, such as parks, have been designated as uninhabited areas as well, adding to the overall visual realism of the dasymetric map.

Figure 4.3. Diagram (A) shows the choropleth map of the Oakland area (top) and the Livermore-Pleasanton area (bottom) and diagram (B) shows the dasymetric map of the same area.

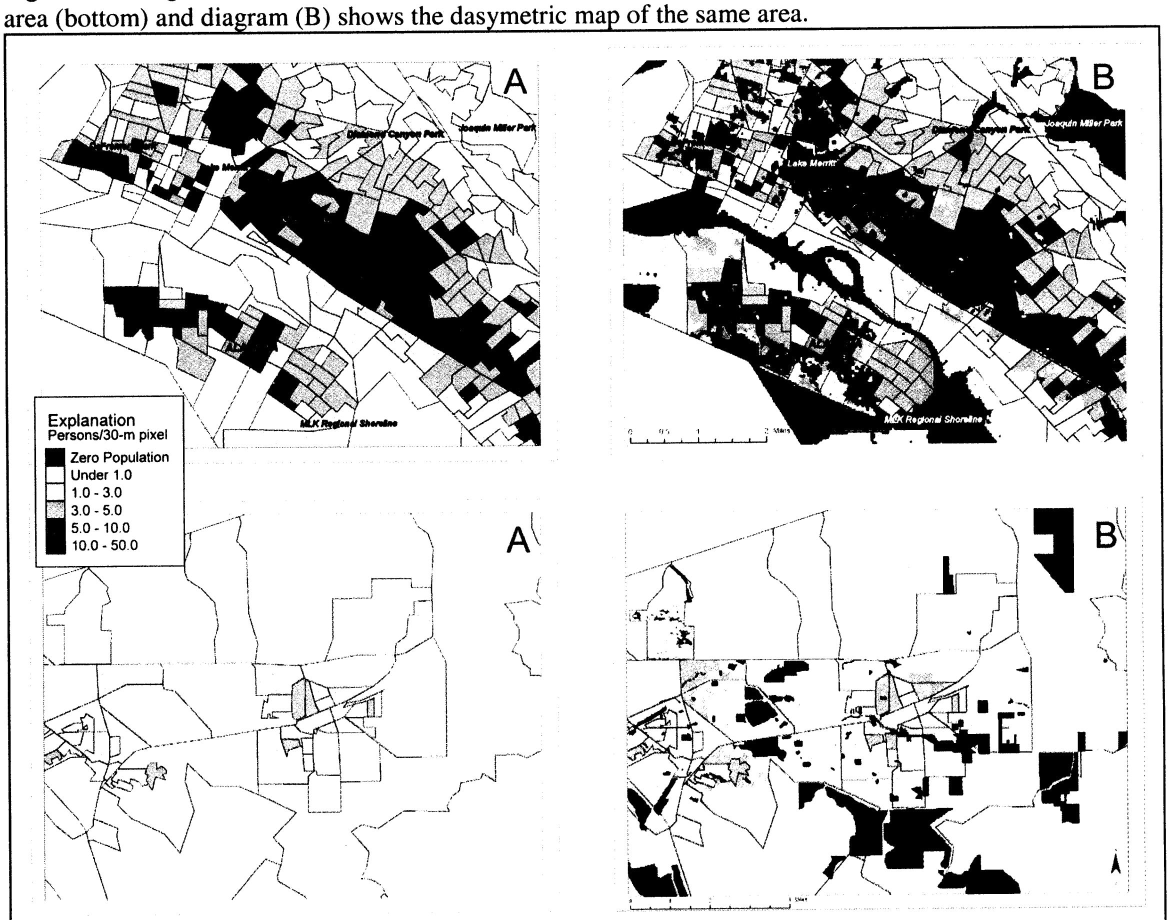
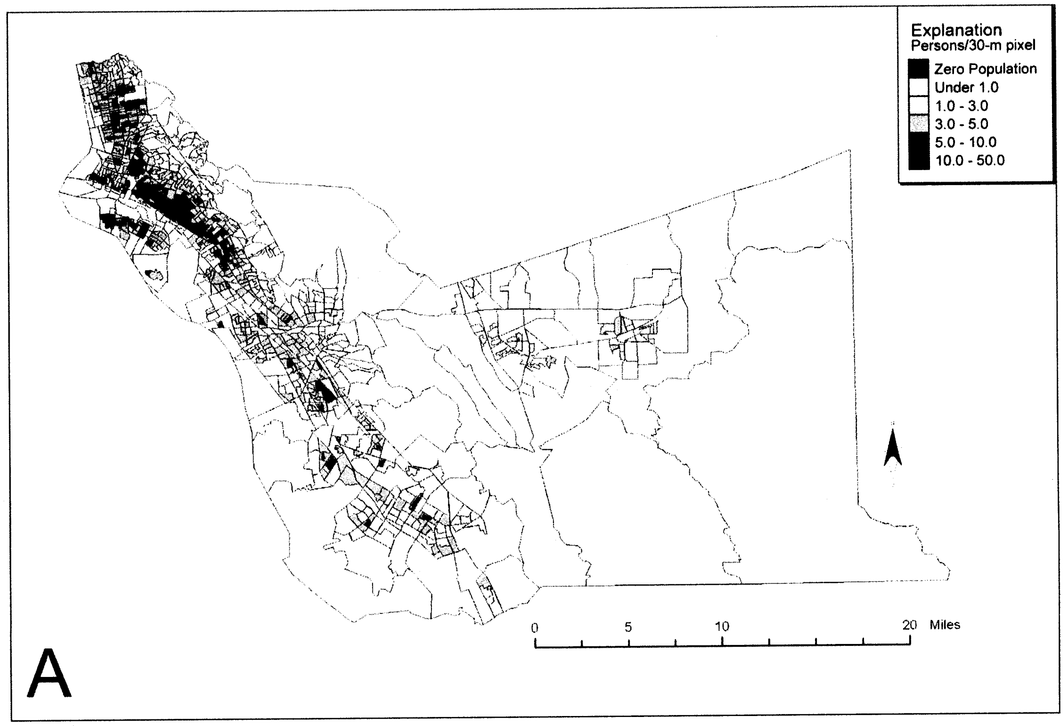
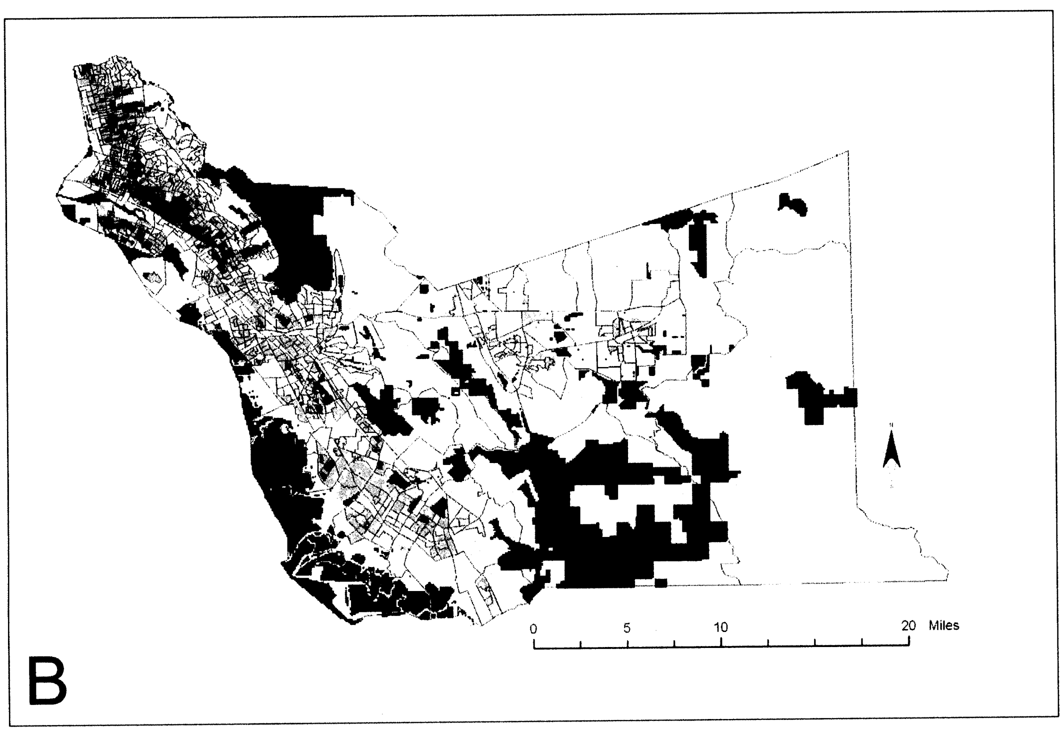


Figure 4.4. Alameda County, Calif., (A) showing the choropleth map of population density and (B) showing the dasymetric map of population density





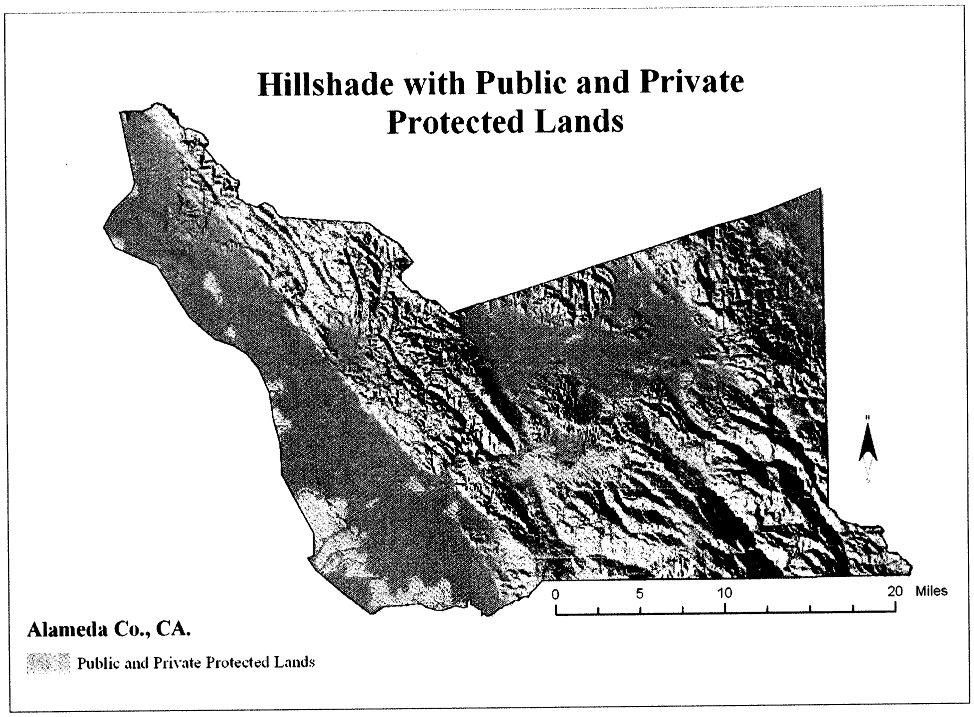
CHAPTER 5: CONCLUSION

5.1 CONCLUSION

Considering the overall statistics compiled (histograms, scatter plots, correlation coefficients and the F-Test), the dasymetric results tested well against the block populations. The histograms showing the differences between the two testing datasets are approximately normally distributed, and the correlation coefficients are very high exceeding 0.80. The correlation coefficients of the choropleth map and the block populations yielded results in the range of 0.56-0.83, which is far lower than the dasymetric map. This statistic proves that the dasymetric mapping method of block-group population density demonstrates a more informative population distribution then the choropleth mapping method.

The segment of the hypothesis stating that the greatest improvements in accuracy would be seen in areas that have a variety of land-cover types is unclear when looking at the variance in data distribution. The highest distribution error within this dasymetric mapping technique occurred within block-groups that had a higher variability in land-cover types. However, the choropleth map, when compared to the block populations demonstrated weak correlation coefficients in the subcounties with high variability in land-cover types. Therefore, the dasymetric map is a superior technique to use for overall population distribution, and the high variances observed are a factor of the data limitations relative to NLCD at this scale. To overcome this limitation, the increase in ancillary data is necessary to better define urbanization zones and to add to the exclusion class. Slope is a variable that should be explored and would pare down the areas that are classified as nonurban. Figure 5.1 shows how the combination of the uninhabited layer with the elevation model could add much more to the exclusion class, which would weed out high variances in these error-susceptible areas.

Figure 5.1. Open Space and Elevation data demonstrate the potential for a stronger exclusion class. Notice all of the topography in the south-eastern part of the county.



The study has shown that the dasymetric mapping technique is a viable approach for defining the underlying statistical surface from a spatial data that is aggregated and attributed to large areal units. The process may seem laborious to some geographers for mapping population density due to the fact that urban core areas typically show the same distributions seen in a choropleth map. However, in large block groups with sparse population (see Figure 4.3) the dasymetric map demonstrates an intuitive and more informative distribution. Processing time was minimized with the GIS field calculator, but the overall task of areal interpolation is time consuming, although could be automated into a programming interface if desired. The inclusion of enhanced ancillary data could produce higher accuracy within all land-cover types due to the identification and elimination of all lands with zero population. The decision to use a dasymetric mapping technique should be made based on the purpose and the audience of the map.

APPENDIX A: DETAILED METHODOLOGY

1) **Gather all necessary datasets.**
   a) National Landcover Dataset 1992. Available for download at http://landcover.usgs.gov/. Downloaded California North half, and subset the data by county.
   b) US Census Bureau, 1990 Census Block-groups, Population; Subcounty Cartographic Boundary Files. Available for download at http://www.census.gov/.
   c) Vector Layer showing open space, public land ownership (Federal, State, and City). Received from Green Info. Network. http:/www.greeninfo.org
   d) Landsat ETM+ 1990 Path44, Row34
2) **Re-project all datasets into common projection and datum.**
   a) UTM, Zone 10 N, WGS_84, Meters
3) **Create Choropleth Map of Population Density, by census block-group.**
4) **Processing NLCD**
   a) *Neighborhood* command :
      i) Using Erdas Imagine, use the *Neighborhood* command. Neighborhood functions are specialized filtering functions that are designed for use on thematic layers. Each pixel is analyzed with the pixels in its neighborhood. The number and location of the pixels in the neighborhood are determined by the size and shape of the filter, which you define. In this case, the 3x3 pixel filter was used. Each filtering function results in the center pixel value being replaced by the result of the filtering function. This function was performed to filter the NLCD for pixels that were classified incorrectly, due to the fact that NLCD is a national dataset, and this study is at the county scale.
   b) Raster Recode
      i) The NLCD needed to be recoded into the classes chosen for this study. Out of the twenty-one NLCD classes, only three classes are identified, and the remainders are classified as non-urban. Below the three classes are highlighted, Open Water, Low Intensity Residential, and High Intensity

Residential. The residential classes are the urban classes, and the open water signifies an area that can later be eliminated, or given a value of zero population for the interpolation.

c) Recode NLCD into classes 0, 1,2,3,4.
   (a) 0 = No Data
   (b) 1 = High Intensity
   (c) 2 = Low Intensity
   (d) 3 = Non Urban (18 classes)
   (e) 4 = Water (this is separate from non-urban because later it will be part of the exclusion class).

5) Process Open Space Layer

a) Create raster layer from vector open space layer.
   i) ArcMap-Spatial Analyst – Convert – Features to Raster
b) Raster Recode
   i) Erdas Imagine – Recode into two classes 0, 10.
      (a) 0 = No Data
      (b) 10 = Open Space
c) Combine Two Rasters to produce the sum of all values.
   i) Erdas Imagine- Utilities – Two Input Operators
   ii) Output :
      (1) 0 = No Data
      (2) 1 = High Intensity
      (3) 2 = Low Intensity
      (4) 3 = Non Urban
      (5) 4 = Water
      (6) 10 = Open Space
      (7) 11 = Open Space + High Intensity
      (8) 12 = Open Space + Low Intensity ↔ Recode all Values to
      (9) 13 = Open Space + Non Urban 4 = Exclusion

(10)14 = Open Space + Water

iii) Water becomes part of the exclusion class as well as all of the areas that are excluded due to the fact that we know that there is no population in these areas.

6) **Creating Point Feature Class Representing Land-cover ID**

a) New LandCover Table

i) 0 = No Data

ii) 1 = High Intensity

iii) 2 = Low Intensity

iv) 3 = Non Urban

v) 4 = Exclusion

b) Convert Raster to Feature using Spatial Analyst, outputting POINTS.

i) For every point in the new vector layer, there is a grid_code attribute which represents the thematic land-cover code.

ii) Each Point Represents 30 meters on the ground, relative to the input pixel size.

iii) The output point file for Alameda County created approximately 4.7 million points. In order to process the statistics with efficiency, the data needs to be subset into a smaller, more manageable area. To emulate Mennis' study, the population density of Alameda should be calculated by subcounty, because the counties on the west coast are much larger than the east coast, where Mennis conducted his research. In addition, the large point file is subset into the POINTS the *have their center within*, each subcounty.

7) **Join Land-cover data (points) with Census Block-group data (polygons)**

a) Each Point needs to have the Block-group Unique ID associated to it.

b) To do this, conduct a *Spatial Join* (Polygons to Points)

c) Now each point has all block-group attributes associated to it creating a layer that will be the final statistical table.

d) Using the Livermore-Pleasanton Subcounty Division as an example, the dataset should now contain around 1.2 million points.

8) Areal Interpolation

a) Before the field calculations can take place, population density values for the subcounty needs to be established. This is a sampling method which calculates the relative difference in population density among urbanization classes (Mennis, 2003). To sample this, three block-groups were selected for each urbanization class (block-groups that had a majority of high, low, non-urban points). If possible, a block-group that was entirely high, low or non-urban points would be the best sample. The total population and area were calculated for each sample and an aggregated population density was the outcome. (See Table 3.1) The sampled population density was calculated for each subcounty, independently because the relative difference in population density varies within the county due to the fact that population density is going to be much higher in the urban core compared to the urban fringe of the county where there lies a lot of open space.

b) New fields can be added to the table. A POINT feature class represents the foundation for the interpolation. All of the calculations were performed with the *Field Calculator* in ArcMap.

c) The first field added was the Population Density Fraction which will be calculated by equation (1) of Mennis' (2003) study.

$$d_{uc} = P_{uc}/(P_{bc} + P_{lc} + P_{nc}) \qquad (1)$$

where d_{uc} = population density fraction of urbanization class u in county c,

P_{uc} = population density (persons/900 m^2)of urbanization class u in county c,

P_{bc} = population density (persons/900 m^2) of urbanization class h (high) in county c,

P_{lc} = population density (persons/900 m^2) of urbanization class l (low) in county c, and

P_{nc} = population density (persons/900 m^2) of urbanization class n (nonurban) in county c.

This fraction will be used throughout the interpolation.

d) The next calculation is to determine the Area Ratio, which focuses on the difference in area between the urbanization classes, within each block group.

$a_{ub} = (n_{ub}/n_b)/0.33$ (2)

where a_{ub} = area ratio of urbanization class *u* in block-group *b*,

n_{ub} = number of grid cells of urbanization class *u* in block-group *b*, and

n_b = number of grid cells in block-group *b*.

e) After the Area Ratio is calculated the Total Fraction is calculated which determines the percentage of the block-group population goes to each urbanization class. This is the most crucial value of the interpolation.

$f_{ubc} = (d_{uc}*a_{ub})/$

$[(d_{hc}*a_{hb}) + (d_{lc}*a_{lb}) + (d_{nc}*a_{nb})]$ (3)

where f_{ubc} = total fraction of urbanization class *u* in block-group *b* and in county *c*,

d_{uc} = population density fraction of urbanization class *u* in county *c*,

a_{ub} = area ration of urbanization class *u* in block-group *b*,

d_{hc} = population density fraction of urbanization class *h* (high) in county *c*, d_{lc} = population density fraction of urbanization class *l* (low) in county *c*, d_{nc} = population density fraction of urbanization class *n* (nonurban) in county *c*,

a_{hb} = area ratio of urbanization class *h* (high) in block-group *b*,

a_{lb} = area ratio of urbanization class *l* (low) in block-group *b*, and

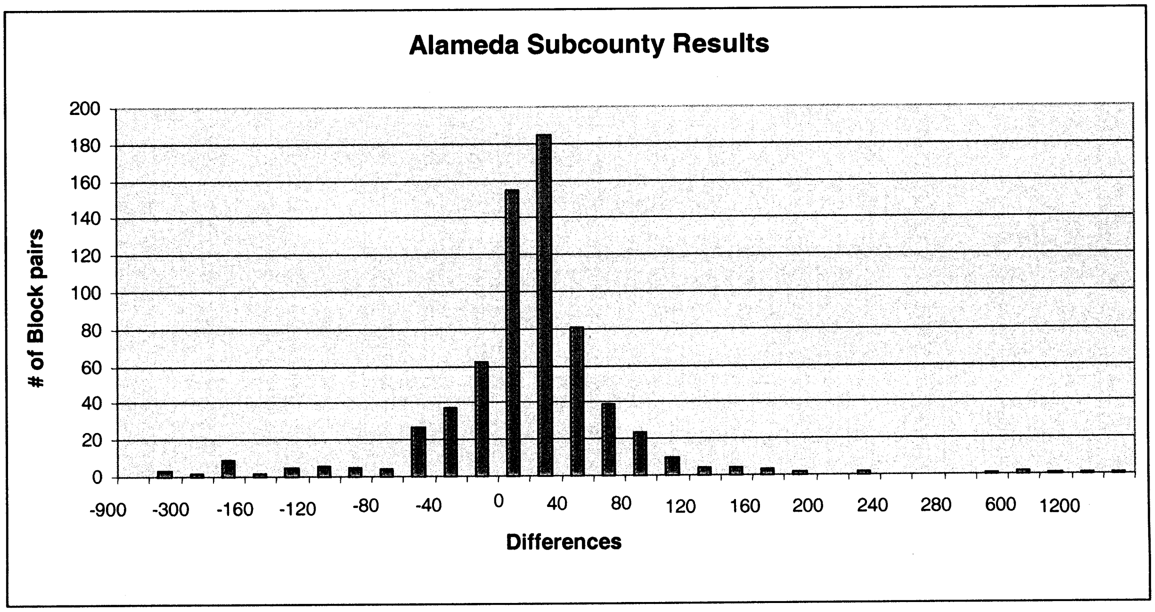
a_{nb} = area ratio of urbanization class *n* (non-urban)in block-group *b*.

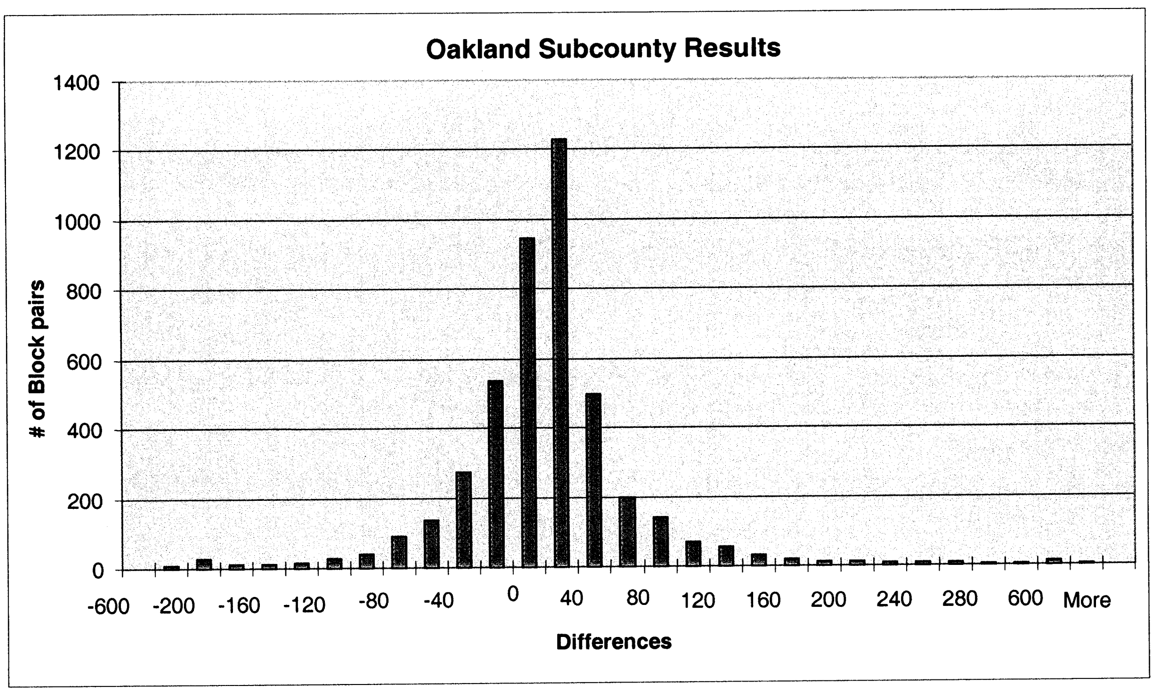
f) Finally the block-group population can be distributed to the points within each block-group.

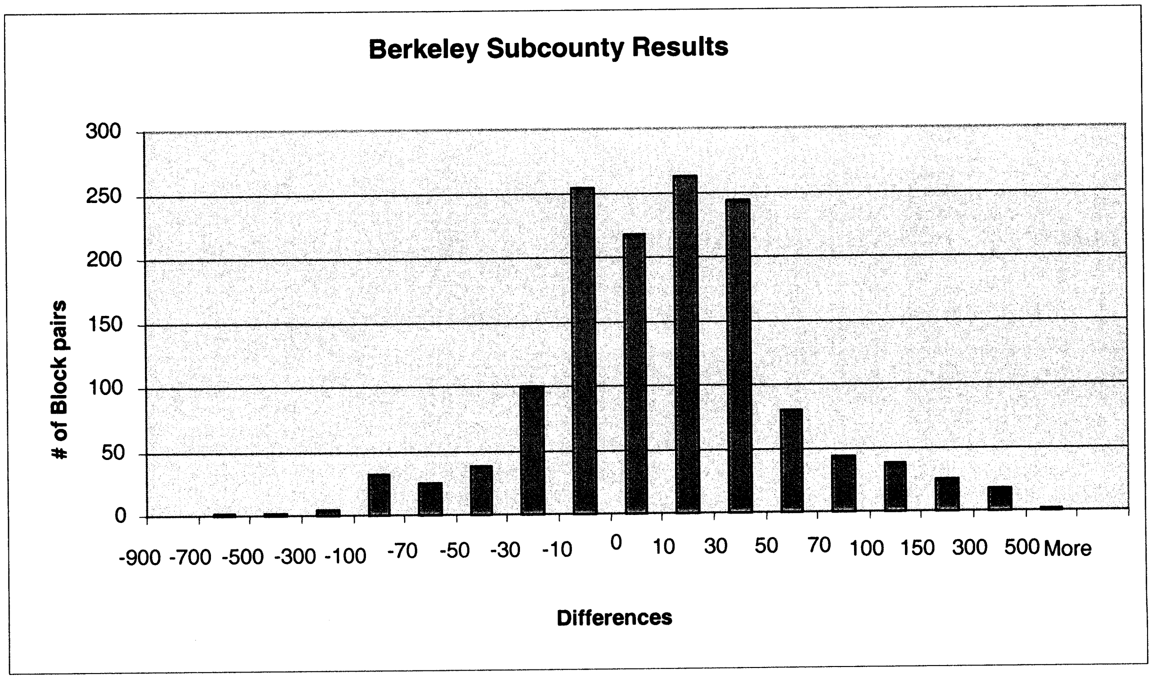
$pop_{ubc} = (f_{ubc}*pop_b)/n_{ub}$ (4)

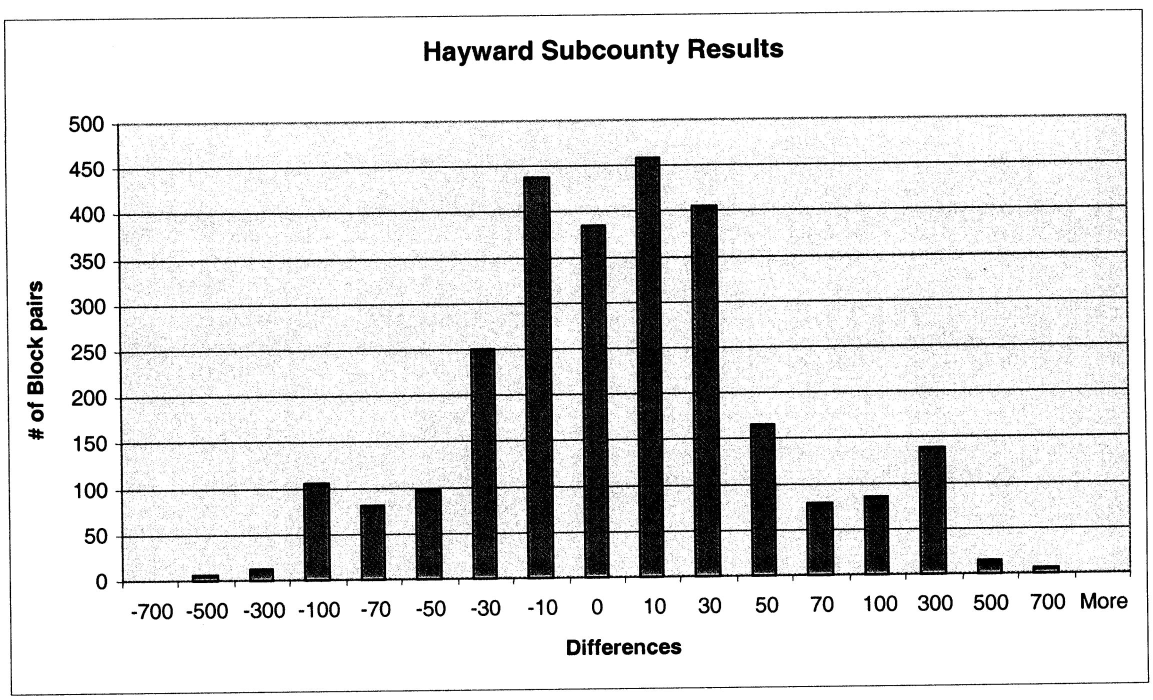
where pop_{ubc} = population assigned to one grid cell of urbanization class *u* in block-group *b* and in county *c*, f_{ubc} = total fraction for urbanization class *u* in block-group b and in county c, pop_b = population of block-group b, and n_{ub} = number of grid cells of urbanization class u in block-group b.

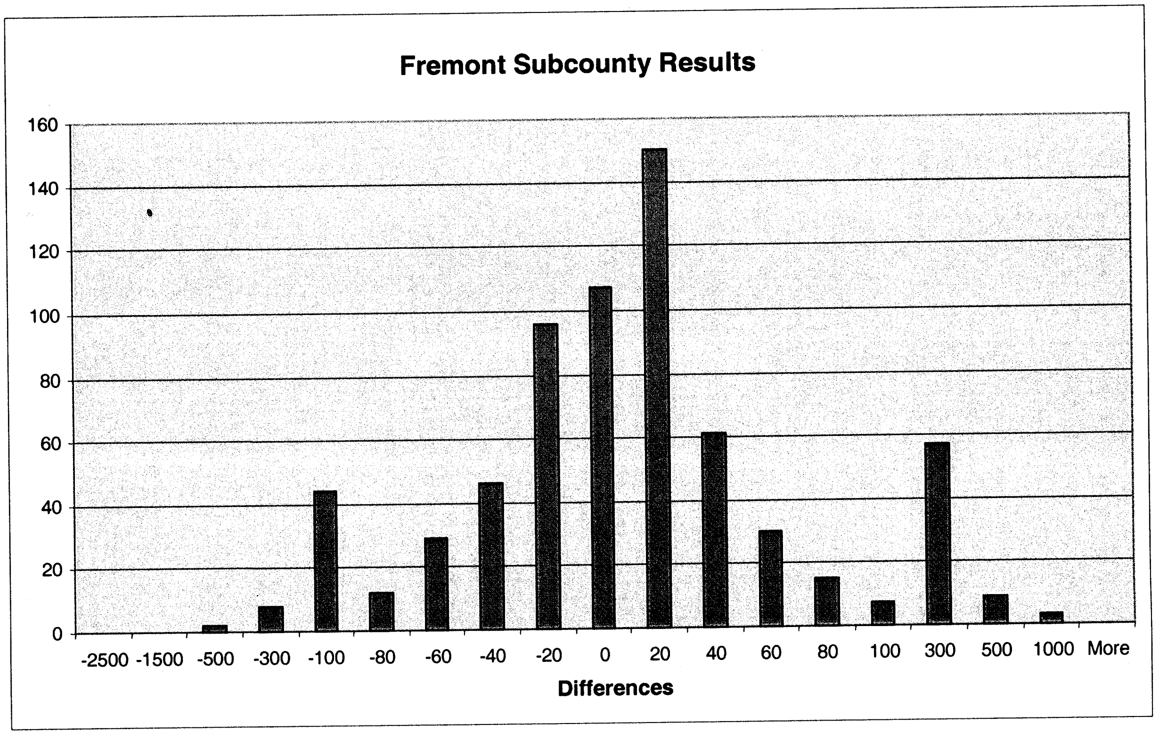
APPENDIX B: CHARTS AND GRAPHS



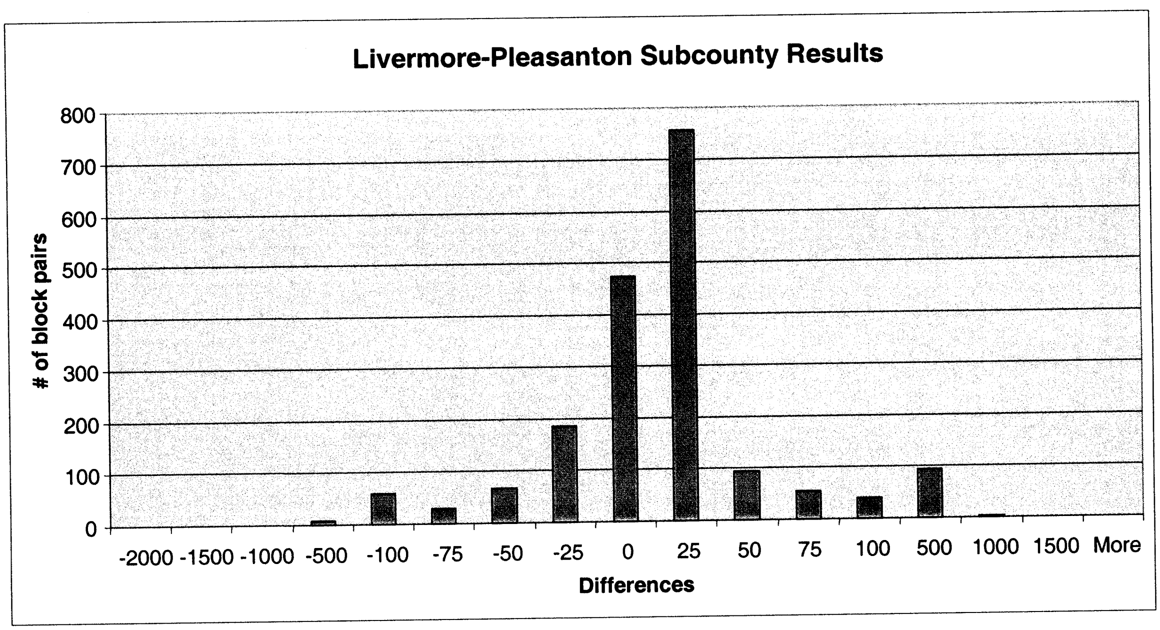



Berkeley Subcounty Results
300
250
200
150
100
50
0
# of Block pairs
-900 -700 -500 -300 -100 -70 -50 -30 -10 0 10 30 50 70 100 150 300 500 More
Differences


Hayward Subcounty Results
500
450
400
350
300
250
200
150
100
50
0
# of Block pairs
-700 -500 -300 -100 -70 -50 -30 -10 0 10 30 50 70 100 300 500 700 More
Differences

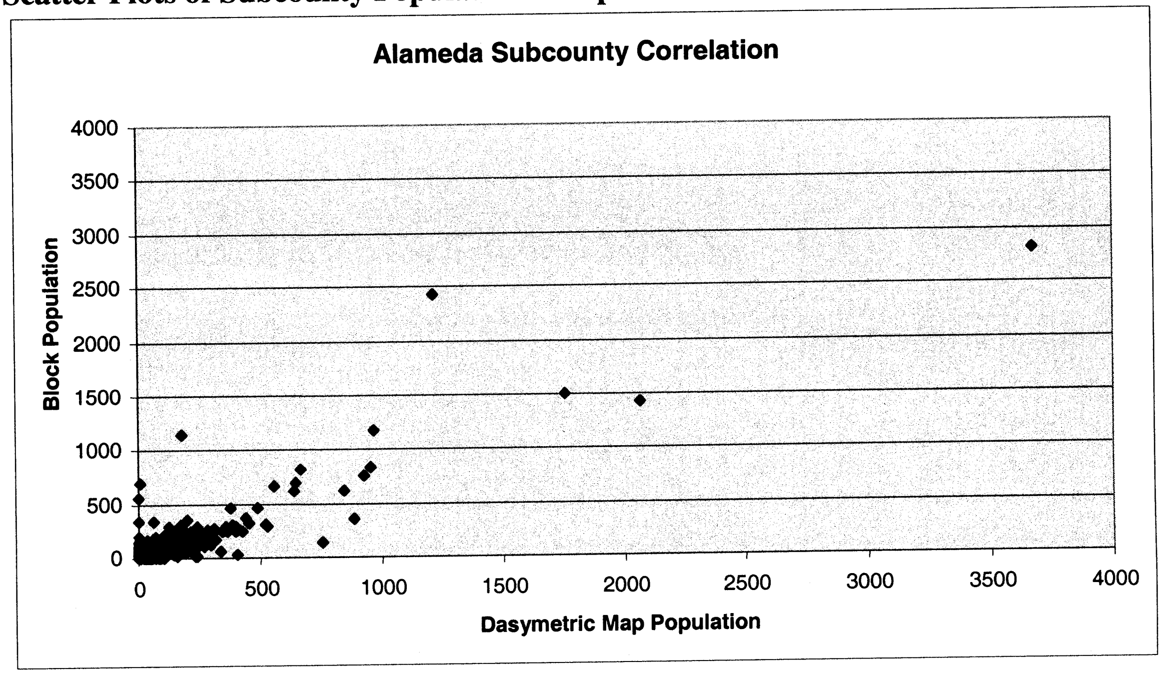


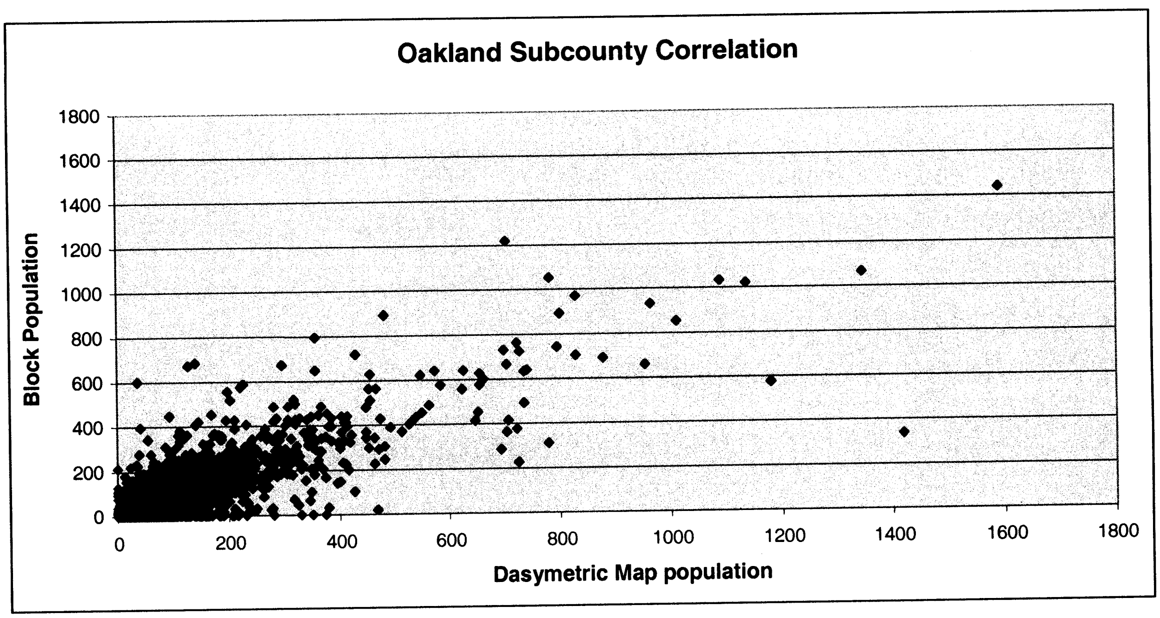
Fremont Subcounty Results
160
140
120
100
80
60
40
20
0
-2500 -1500 -500 -300 -100 -80 -60 -40 -20 0 20 40 60 80 100 300 500 1000 More
Differences

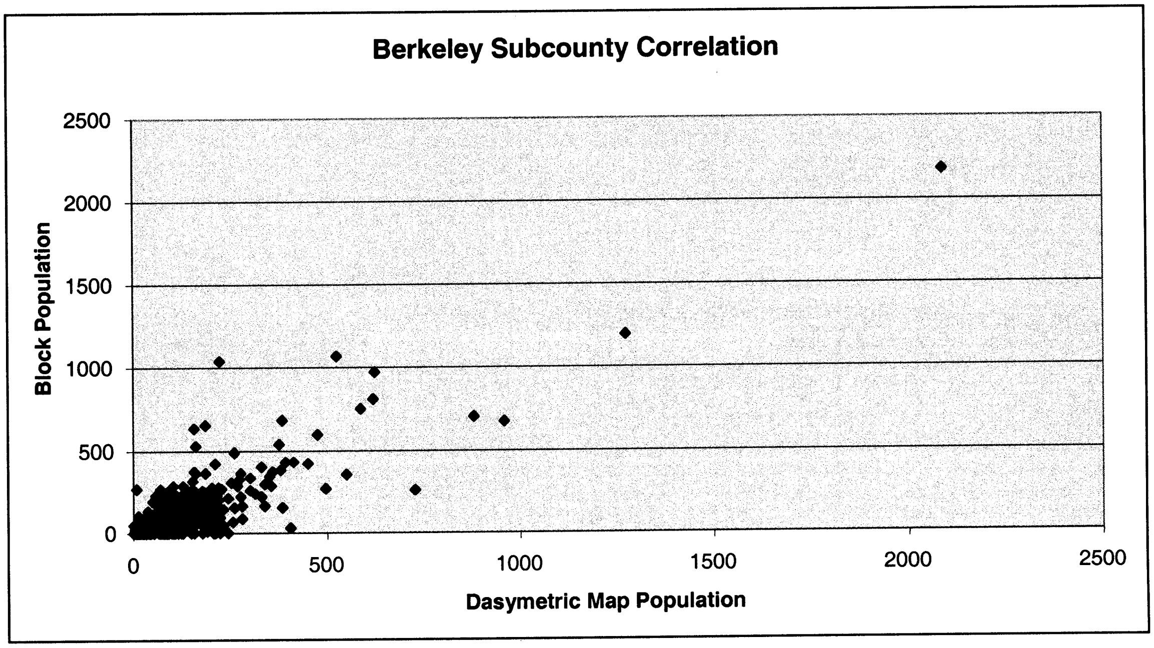


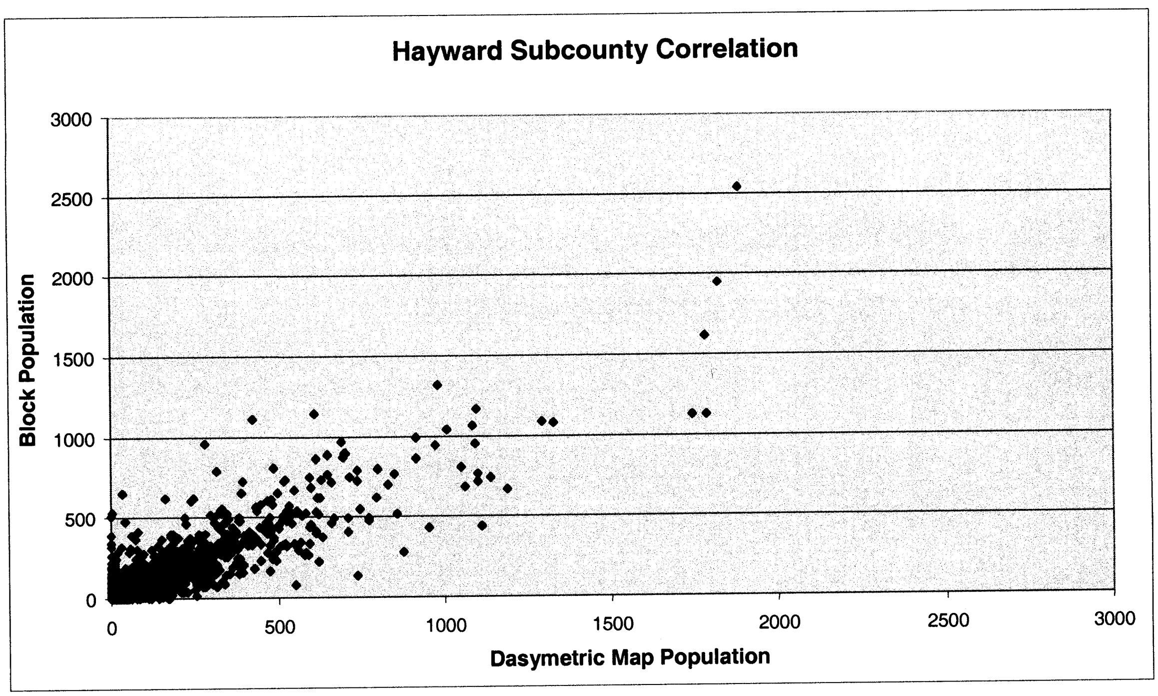
Livermore-Pleasanton Subcounty Results
800
700
600
500
400
300
200
100
0
# of block pairs
-2000 -1500 -1000 -500 -100 -75 -50 -25 0 25 50 75 100 500 1000 1500 More
Differences

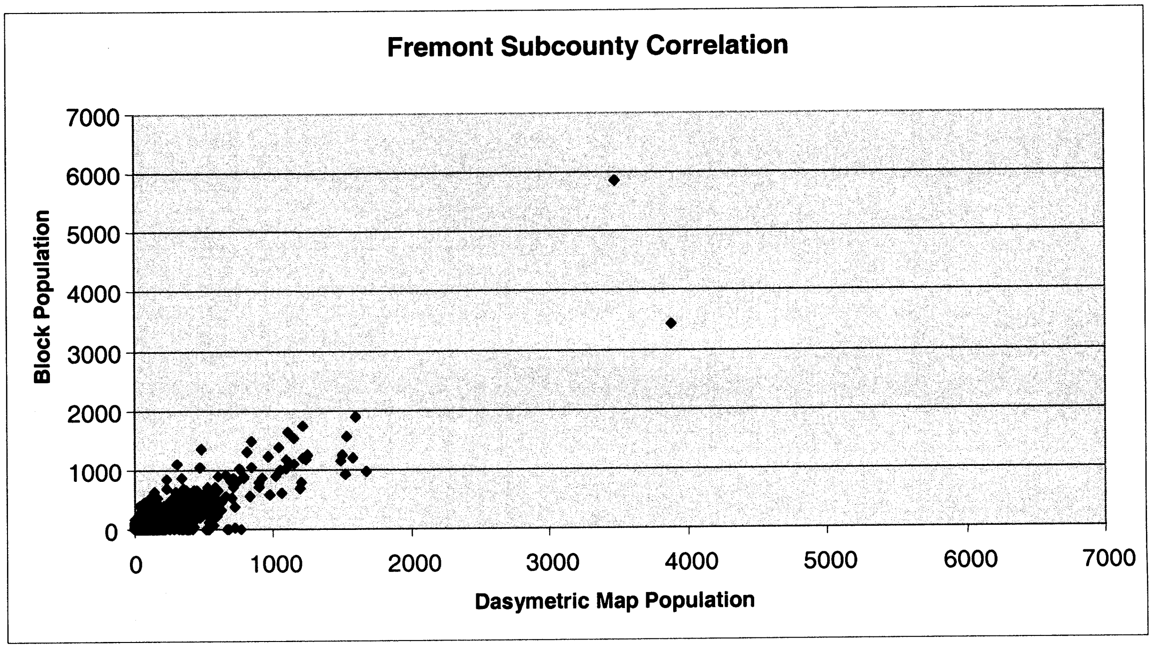
Scatter Plots of Subcounty Population Comparisons

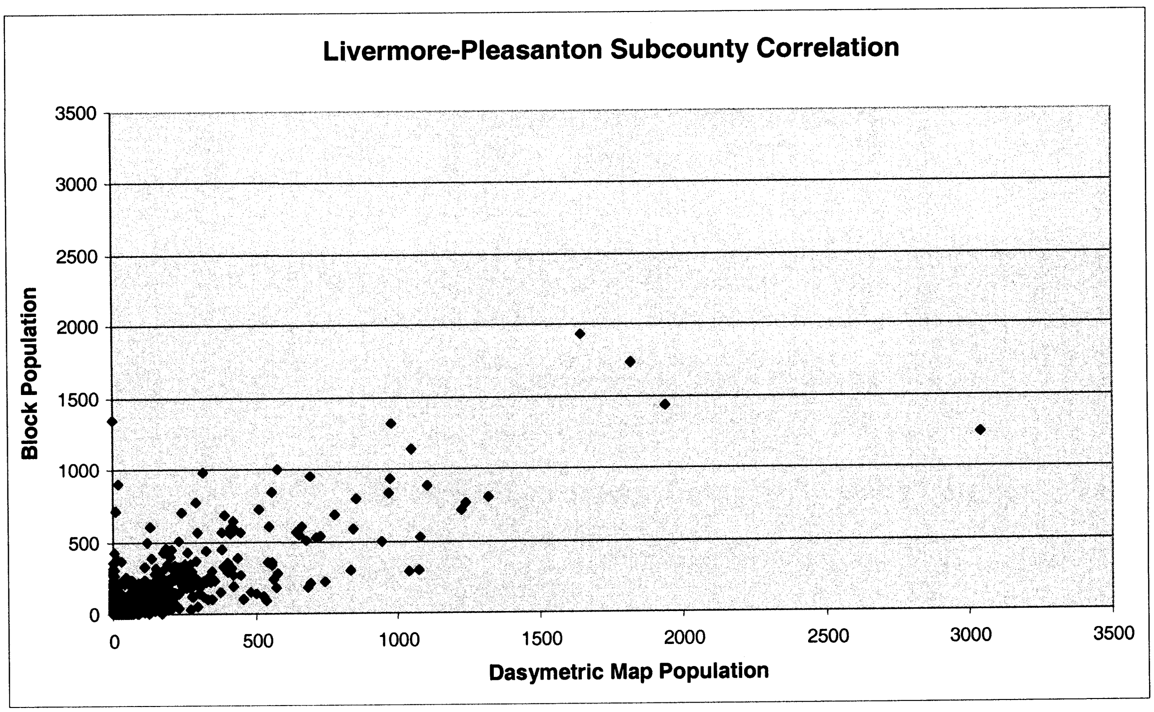


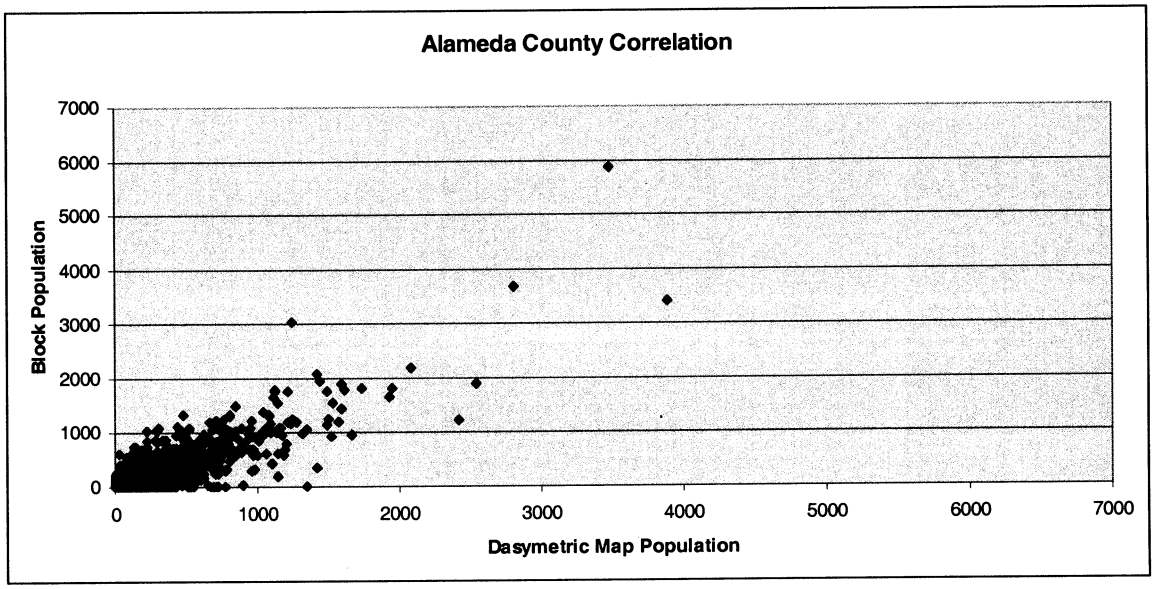



Berkeley Subcounty Correlation
2500
2000
1500
1000
500
0
Block Population
0
500
1000
1500
2000
2500
Dasymetric Map Population


Hayward Subcounty Correlation
3000
2500
2000
1500
1000
500
0
Block Population
0
500
1000
1500
2000
2500
3000
Dasymetric Map Population


Fremont Subcounty Correlation
7000
6000
5000
4000
3000
2000
1000
0
Block Population
0
1000
2000
3000
4000
5000
6000
7000
Dasymetric Map Population


Livermore-Pleasanton Subcounty Correlation
3500
3000
2500
2000
1500
1000
500
0
Block Population
0
500
1000
1500
2000
2500
3000
3500
Dasymetric Map Population


Alameda County Correlation
Block Population
0
1000
2000
3000
4000
5000
6000
7000
Dasymetric Map Population

WORKS CITED

Anderson, J.F., E.E. Hardy, J.T. Roach, and R.E. Witmer. A land use and land-cover classification system for use with remote sensor data. *Professional Paper 964*. Washington, DC: US Government Printing Office, 1976.

Burt, J. E., and Gerald M. Barber. 1996. *Elementary Statistics for Geographers*. New York: The Guilford Press.

Charpentier, M. 1997. The dasymetric method with comments on its use in GIS. Paper presented at the 1997 University Consortium for Geographic Information Science annual assembly and summer retreat, June 15-21, Bar Harbor, ME.

Dorling, D. 1993. Map design for census mapping. *The Cartographic Journal* 30:167-183.

Eicher, L. C. and C. A. Brewer. 2001. Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science* 28 (2):125-38.

Freund, J. and G. Simon. 1997. *Modern Elementary Statistics*. New Jersey: Prentice Hall, Inc.

Goodchild, M. 1992. Geographical data modeling. *Computers and Geosciences* 18(4):401-8.

GreenInfo Network. 2003. Bay area public and privately protected lands. Shapefile.

Holloway, S., J. Schumacher, and R. Redmond. 1997. People & place: Dasymetric mapping using Arc/Info. *Cartographic Design Using ArcView and Arc/Info*. High Mountain Press, NM.

MacEachren, A.M. 1994. *Some truth with maps: A primer of symbolization and design*. Washington DC: Association of American Geographers.

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

Slocum, T. A., and Egbert, S. L. 1993. Knowledge acquisition from choropleth maps. *Cartography and Geographic Information Systems* 20(2):83-95.

Stehman, S.V., T.L. Sohl., and T.R. Loveland, 2003. Statistical sampling to characterize recent United States land-cover change. *Remote Sensing of Environment* 86(4):517-529.

Tobler, W. 1979. Smooth pyncnophylactic interpolation for geographic regions. *Journal of the American Statistical Association* 74:519-30.

Unwin, D. J. 1996. GIS, spatial analysis and spatial statistics. *Progress in Human Geography* 20(4):540-441.

US Census Bureau. 2001. Census 2000 geographic definitions. http://www.census.gov/geo/www/geo_defn.html (accessed August 10, 2004).

Volgelmann, J.E., T. Sohl, P.V. Campbell, and D.M. Sahw. 1998. Regional land cover characterization using landsat thematic mapper data and ancillary data sources. *Environmental Monitoring and Assessment* 51(1-2):415-428.

Vogelmann, J.E., S.M. Howard, L. Yang, C.R. Larson, B.K. Wylie, and N. Van Driel. 2000. Completion of the 1990s national land cover dataset for the conterminous United States from landsat thematic mapper data and ancillary data sources. *Photogrammetric Engineering and Remote Sensing* 67(6):650-661.

Wright, J. K. 1936. A method of mapping densities of population with Capecod as an example. *Geographical Review* 26:103-10.