NOTICE

AVIS

ISBN 0-315-53235-1

<u>DATA CLASSIFICATION FOR CHOROPLETH MAPPING</u>

by

Sheila Currie

Submitted to the School of Graduate Studies in partial fulfillment of the requirements of the Master of Arts program in Geography.

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Table of Contents

List of Figures

List of Tables

List of Appendices

Acknowledgements

I would especially like to thank Dr. Gordon Deecker for his encouragement, expert advice and optimism throughout the course of this research.

I also want to express my gratitude to Dr. Peter Johnson for his guidance and continued support of this study.

In addition, I would like to thank the Geocartographics division, Statistics Canada, for the use of their facilities.

Last but not least -- Thanks to Peter Leighton, for supporting all my endeavours.

Abstract

In this study, six computer programs proposed by K. Chang, G. Jenks, H. Moellering and M. Wasilenko, M. Monmonier, C. Youngmann, and the SAS Institute were evaluated on their ability to classify data for choropleth mapping. The evaluation took into consideration both the mathematical accuracy of the classed data sets obtained by each program, and the visual complexity of the maps produced. The effects of data set size and data distribution were included as additional factors in the testing.

Although all of the iterative classing techniques performed well, the results of the study clearly indicated that the "variance" option of the program proposed by G. Jenks is the best classing method overall. It consistently produced good results both in terms of mathematical accuracy and visual map characteristics.

Résumé

Cette étude évalué six programmes informatiques, proposés par K. Chang, G. Jenks, H. Moellering and M. Wasilenko, M. Monmonier, C. Youngmann, et l'Institut SAS. Les six programmes ont été évalués selon leur capacité à classer des données statistiques pour fin de cartographie choroplèthe. L'évaluation s'est penchée sur deux aspects des programmes, soit l'exactitude mathématique des classements de données produits par chaque programme, ainsi que l'aspect visuel des cartes générées par les différents programmes. Deux facteurs additionnels ont été incorporés dans l'évaluation, soit le nombre de données à classer, et la répartition des données à classer.

Quoique toutes ces méthodes itératives aient bien effectué le classement des données, les résultats de la présente étude démontrent clairement que le programme proposé par G. Jenks, utilisé avec l'option "variance", est supérieur, de façon générale, aux autres programmes. Ce dernier a régulierement donné de bons résultats, que l'on considère l'exactitude mathématique où la qualité visuelle des cartes qu'il génere.

I. INTRODUCTION

1.1 Geography as the Study of Spatial Phenomena

1.1.1 Introduction

Geography as a discipline encompasses such numerous and diverse sub-disciplines that scholars have had difficulty agreeing upon a definition of "geography", and on the concepts underlying it.

The pedologist studies the texture and chemical composition of soils, while environmental planners are interested in the complex interactions between man and his environment. Human geographers study the spatial variation of man's cultural and socioeconomic characteristics.

The cartographer attempts to enhance the understanding of geographical information by displaying it in the form of a map. Other sub-disciplines of geography include urban and regional planning, climatology, geographic information systems, biogeography, geomorphology, economic geography, historical geography, and various combinations of these and other disciplines.

The common thread holding all of these sub-disciplines together, and making them part of "geography", is that they all study spatial phenomena.

By classifying spatial phenomena into forms and measuring the magnitude or quality of their existence, geographers translate spatial phenomena into spatial data. This translation process, or data collection, is accomplished by the use of satellite, field expedition and survey, among other methods.

Geographers do many things with spatial data. They compile, analyze, and summarize it with the use of laboratory equipment, geographical modelling, statistics, and computers. They display it in tabular, graphic, and mapped form, and then use the displayed data in the formulation of theory, recommendations, and policies.

1.1.2 The Display of Spatial Data

Spatial data can be displayed in tabular, graphic and mapped form (Figure 1.1). Tables are often used during data collection to record each individual value of a data set and, to a lesser extent, to display data sets. In addition, tables are useful for storing raw data in preparation for analysis, graphing or mapping.

Graphs transform tabular data into a form which depends more upon the visual/perceptual skills of the user in order to impart information. Ideally, graphs are employed to depict changes and/or trends in data sets placing less emphasis on the exact values of each number

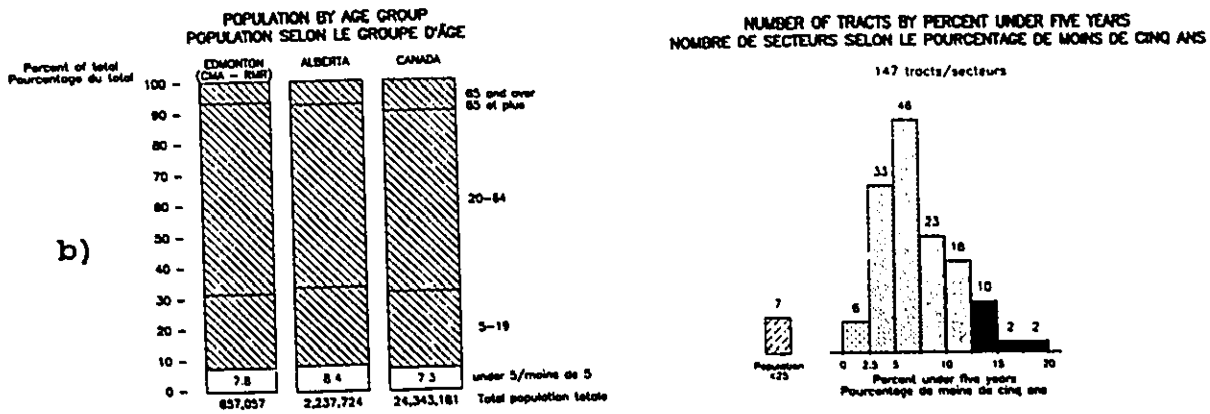
EDMONTON - POPULATION LESS THAN FIVE YEARS OF AGE

a)

AREA	POP. <5	TOTAL POP.	% POP. <5
2142	330	5500	6.0
2143	270	6135	4.4
2147	200	4445	4.5
2148	230	6210	3.7

etc.

b)



c)

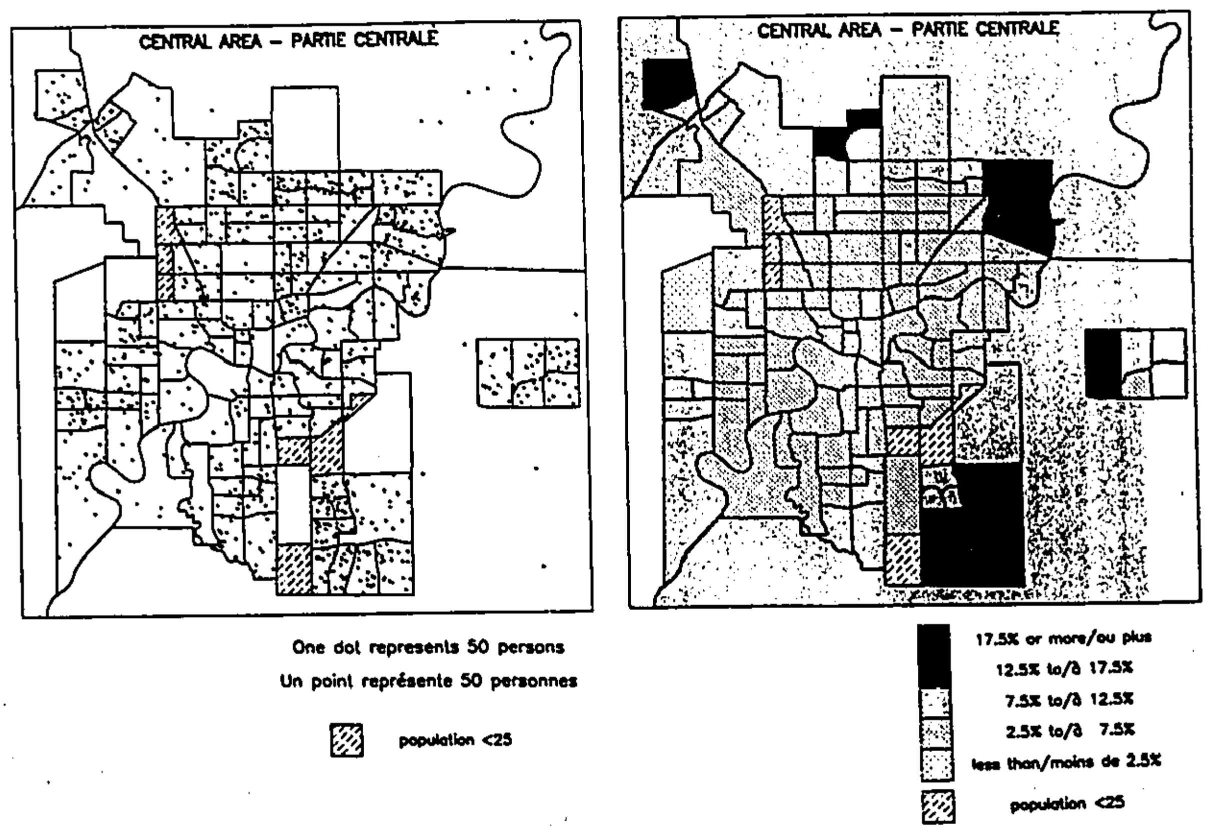


Figure 1.1 **A data set displayed in a) tabular, b) graphic and c) mapped form. Metropolitan Atlas Series - Edmonton. Statistics Canada, 1984.**

in the raw data set. In the process of portraying trends rather than individual values, graphed data have often been compiled, symbolized and/or summarized in a procedure which geographers often refer to as "generalization" (Jenks, 1963; Jenks and Caspall, 1971; Robinson et al., 1984).

Similar to graphs, maps display data which have been transformed. However, in the case of maps, the transformation produces a display of the data that highlights the spatial interrelationships amongst the data values. The generalization procedure sometimes applied in the formulation of graphs is used extensively in the creation of maps. This generalization procedure, also termed "cartographic abstraction" (Muehrcke, 1981), gives maps the power to summarize vast amounts of data, and to present it as information to map users who can gain an understanding of the mapped phenomena by viewing the map.

Tables and graphs are essential tools that geographers employ to gain and provide an understanding of spatial data. However, because maps have the ability to provide users with an understanding of immense amounts of data by instantly highlighting data values and their spatial distribution, they are the main form of communication for geographers. Muehrcke (1981, p. 13) states that:

> "The geographer's map is a magical tool. It assists in data collection, provides a convenient storage medium for geographical information, and through its immense abstractive and transformational powers, it provides creative views of the environment which would otherwise be unattainable."

Cartographic generalization includes one or more of the following functions: simplification, classification, symbolization, and induction (Robinson et al., 1984). Simplification of spatial data involves selecting only the most important data values or attributes. Classification differs from simplification in that the data is modified to reflect all values in the data set, rather than eliminating some of the data itself. Symbolization is the process of assigning graphics to the summarized data. Induction requires human involvement in making logical inferences about the data. For example, induction includes making associations about the values associated with various shadings, or extrapolating additional information from mapped information.

This study examines in detail the generalization process of classification for choropleth mapping.

1.2 Definition of the Problem

1.2.1 Choropleth Mapping and Classification

Choropleth maps are used to display relative data such as rates, fractions, percentages and other ratios using shaded area symbols. Generalization of the data is accomplished by classifying it into groups, and using one shading pattern per group on the map. Cartographic and associated perceptual research has resulted in the wide acceptance of certain standards concerning the number of classes to use and how to apply graded series in the selection of shading patterns. However, as evidenced by recent literature (Paslawski, 1984; Smith, 1986; Lu, 1987), controversy still exists concerning the methodology of data classification for choropleth mapping.

1.2.2 The Nature of the Problem

The problem of data classification for choropleth mapping has been researched and documented in geographic and statistical literature for several decades. There is an increasingly wide variety of classification methods to choose from, and no method is universally accepted.

The data classification problem is augmented by the fact that numerous factors may affect the applicability and/or accuracy of classification methods. Among these factors are data set characteristics such as size, skewness and bimodality, map purpose, and the knowledge of the intended map users. In addition, the relationship between data classification methods and the visual complexity of the maps which they produce has not been investigated in any great depth. Although these issues have been recognized in geographic literature, very few generally accepted guidelines exist for the application of classification techniques in the preparation of data for choropleth mapping.

Recent advances in computer technology have made possible the development of another sub-discipline of geography -- geographic information systems. These systems perform some or all of the following functions by automated means: data collection, storage, manipulation, analysis, generalization and display. The proliferation of geographic information systems has made the mass-production of maps possible by both geographers and non-geographers, with the result that

> "Graphic crimes of the most basic kind that would make the most junior cartographer wince are perpetrated, and worse, not recognized by users of the system" (Jupe, 1987, p. 346).

In an evaluation of four of the most popular commercial mapping packages for personal computers, Noronha (1987) concluded that none had a satisfactory classing procedure. These conclusions and those of other researchers (Dobson, 1983; Morrison, 1983; Abler, 1987) support the premise that it is critical for geographers to stand alongside system architects and developers in order to ensure that geographic information systems are based on geographic knowledge. Furthermore, with rapid technological advances, expert systems with a geographical component will soon be commonplace. It is essential for geographers to be the researchers that formalize geographic theory for input into these systems (Robinson et al., 1986). Geographers must ensure that the results of decades of geographic inquiry are correctly incorporated into the new systems produced through the ongoing evolution of geographical analysis and display.

1.3 Objectives of the Study

The purpose of this study is to evaluate the classification methods incorporated in six computer programs in terms of the mathematical accuracy of the classed data sets, the visual complexity of the maps they produce, and their operational advantages and disadvantages. Five of the programs are designed to classify data for choropleth mapping. These are the programs prepared by K. Chang, G. Jenks, M. Monmonier, H.

Moellering and M. Wasilenko and C. Youngmann. The sixth is a general purpose statistics program called the Statistical Analysis Software (SAS) package.

By combining the mathematical accuracy and visual complexity aspects of choropleth mapping with the opeerational attributes of the programs, the most thorough evaluation of this subject to date is presented.

1.4 The Classing Programs

For the purpose of this study, the programs are referred to by the name of the person(s) who proposed them. The CHANG program (Chang, 1974) is based on the work of Jenks and Coulson (1963) and provides the user with eight options for data classification. Seven of these options are "traditional" classing methods (explained in detail in Chapter II), while the eighth option allows the user to define his or her own class breakpoints.

The other five programs are all based on newer techniques which select initial class breakpoints and then "reiterate" the class memberships until an optimal solution based on a mathematical accuracy measure is achieved. Each of them defines the "optimal" solution as the one that maximizes the homogeneity of data values within classes. These five so-called "iterative" techniques differ from one another in the way that they initially select class breakpoints, and the accuracy measures that they employ.

The JENKS program (Jenks, 1977) is theoretically based on the "grouping for maximum homogeneity" principle proposed by Fisher (1958). The MONMONIER and YOUNGMANN programs are both based on the work of Jenks and Caspall (1971) who first proposed that iterative techniques be used for choropleth data classification. YOUNGMANN replicates Jenks and Caspall's proposal of class breakpoint determination by absolute deviations from class means. MONMONIER adds another option whereby breakpoints are determined by squared deviations from class means. The MOELLERING program groups observations into classes on the basis of information theory. The SAS program performs a disjoint cluster analysis.

1.5 The Data

The data used in this study are from the 1981 Census of Canada, and include socioeconomic characteristics of the population, households and dwellings such as participation rates in the labour force, average household income and average value of dwellings respectively.

In order to test the possible effect of data set size on the classing programs, three different sizes of data set were used.

Past research (Evans, 1977; Chang, 1979) suggests that the accuracy of a classing method is affected by different data distributions. Therefore, this study includes a wide selection of these distributions to provide rigorous testing of the classing methods. Seven different data sets are tested for each area, representing right- and left-skewed, L- and J-shaped, rectangular, bell-shaped, and bimodal distributions.

In addition to the twenty-one data sets mentioned above (seven data distributions for each of three census metropolitan areas), one data set which contains extreme outliers is tested.

1.6 Overview of Methodology

Each data set is input into each computer program, and the resulting class breakpoints and measures of accuracy for a five-class map are tabulated. Four measures of mathematical accuracy are applied to each of the classed data sets. The methodology pertaining to the application of these particular accuracy measures is explained in detail in Chapter III.

The resulting class intervals for all of the programs and all of the data sets are mapped using GIMMS

(Waugh, 1980) software and the Versatec plotter. The visual complexity of each map is measured using the fragmentation index, calculated using ARC/Info software (Environmental Systems Research Institute, 1987).

The maps are then visually inspected by the author in order to investigate the relationship between complexity measures and map appearance on the basis of adjacency of class members and clusters of units in the same class. The final evaluation combines the results of the mathematical accuracy tests, the visual complexity tests and the operational attributes of the programs.

1.7 Thesis Structure

The thesis is structured in five parts, including this introductory chapter, which introduces the problem and gives a brief overview of the study. Chapter II provides a detailed literature review of the choropleth map classification problem. Areas covered include classification techniques, mathematical accuracy measures, visual characteristics of choropleth maps, and other related issues.

The third chapter discusses the methodology of the study beginning with the selection of the data sets, and includes a description of the computer programs and the mode of testing for mathematical accuracy. It then explains the procedure for the visual complexity testing and the comprehensive evaluation of each of the programs.

The results of the study are found in chapter IV. They are presented in tabular form, and are then discussed in terms of mathematical accuracy, visual complexity, additional testing on program options, and finally in terms of the comprehensive evaluation.

Chapter V gives a brief summary of the results, and emphasizes the major findings. Areas for future research are delineated, and the practical applications of the findings are presented. In conclusion, the study is positioned in the realm of geographic inquiry.

Appendices of calculations, computer printouts, plots of data distributions and sample map output are found at the end of the thesis.

II. LITERATURE REVIEW

2.1 Introduction

Identification of the problem of data classification for choropleth mapping can be traced back as far as the 1950s in geographic literature. Robinson (1954) noted that different interpretations of the same data can occur if different sets of class intervals are used. Shortly thereafter, several articles (Mackay, 1955; Schultz, 1961; Jenks and Coulson, 1963) appeared. Results of different classing techniques were illustrated and an attempt (Jenks and Coulson, 1963) was made to evaluate them (Figure 2.1).

Although each of these papers contained a few guidelines regarding the selection of class intervals, the prevailing conclusion was that cartographers must be both cautious and rigorous when applying classification techniques for two reasons: there is no one universally accepted method, and the class selection issue is

> "...probably the most important problem which the cartographer must face in using isopleth and choropleth maps" (Mackay, 1955, p. 71).

More than twenty years later, Gilmartin (1987, p. x) reaffirmed this belief:

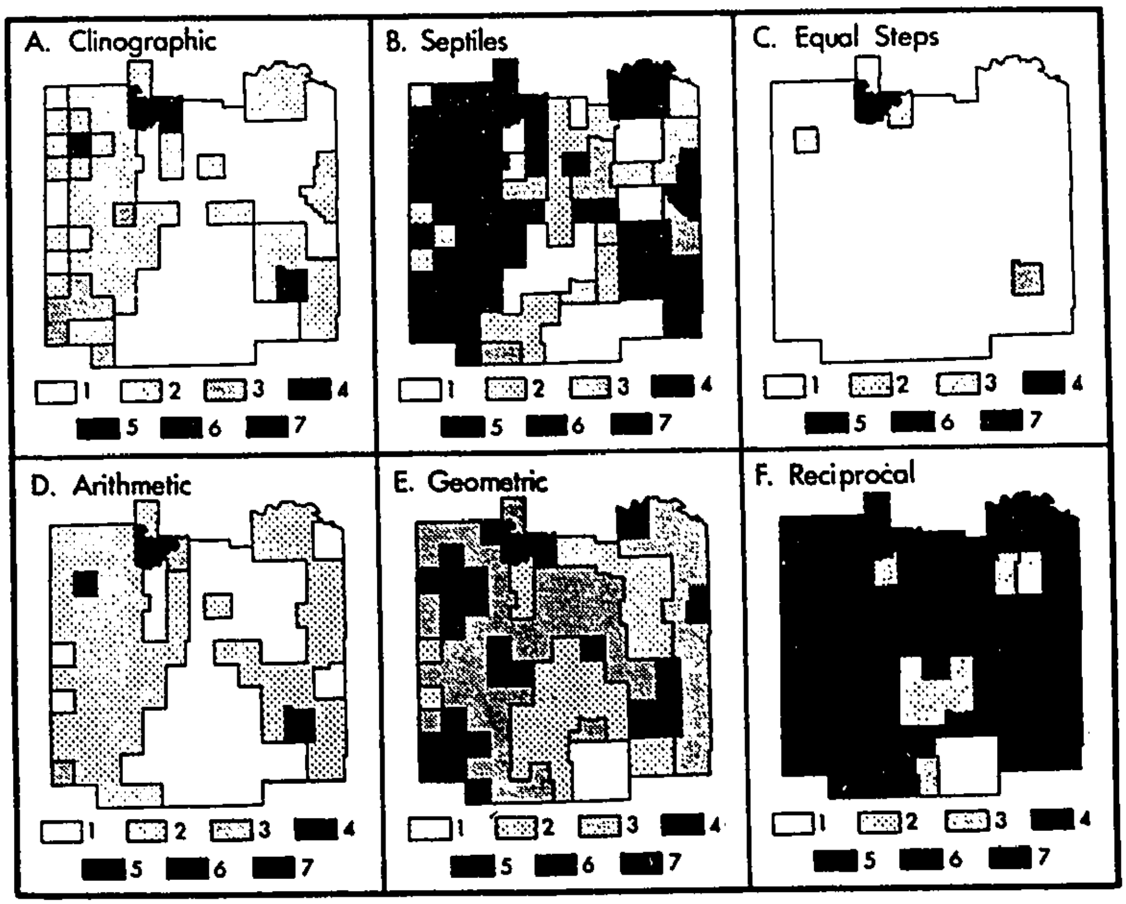


Figure 2.1 Different types of generalization can be achieved by varying the method used in classing data for statistical maps. All of the above maps were constructed with seven classes and the same shadings were used throughout. Therefore, differences can be attributed to the different mathematical systems used in classing. (From Jenks and Coulson, 1963, p. 22)

> "Although classification procedures are a standard topic in cartographic education and are usually a necessary step in the compilation of choropleth maps, there has been a notable lack of guidelines to assist the cartographer in selecting the most suitable set of class intervals for a given map".

Despite the vast body of literature on classing techniques that has accumulated between the 1950s and the 1980s, many authors of recent articles reach similar conclusions to those of the earlier authors. For example, Paslawski (1984, p. 159) stated that:

> "...we have not yet been presented with a general idea of class selection, which would introduce some order to the extensive knowledge on this subject."

The importance of the problem is thought by some authors, for example MacEachren (1983), to be even greater with the advent of computer technology, because of the ease in which class intervals can be calculated, and the "aura" of accuracy associated with the resulting computer-drawn maps.

2.2 Data Classing Techniques

Several methods of categorizing data classing techniques have been outlined in the literature (Evans 1977, Paslawski 1984, Robinson et al., 1984). In this section, data classing techniques are explained within the three categories described by Robinson et al. (1984) as regular series, systematically unequal series, and irregular series. In addition, a fourth category has been added in which other techniques are described which do not strictly fit into the categories described above.

2.2.1 Regular Series

There are five types of classification method which can be considered to be regular series. One of the most commonly used and easily understood methods divides the data range into equal steps, thus creating class ranges of equal value. This is performed by dividing the range of the data set by the number of classes desired; the quctient is added to the lowest value in the ordered data set, and subsequently added to the upper limit of each class to define the class breakpoints (Figure 2.2, column 2). The advantages of the equal step technique are that it is easy to perform, and that the ranges as presented in the legend are usually comprehended well by the user, especially if the breakpoint values are rounded for presentation.

The major drawback to this method is that for skewed data it may derive empty classes, thus reducing the number of classes represented on the map. In a similar manner, this method may also create classes that are "over-filled" with values thus causing a large portion of the map area

Class	Septiles[1]	Equal Step[2]	Arithmetic[3]	Geometric[4]	Reciprocal[5]
1	1.6– 3.1	1.6– 16.0	1.6– 5.1	1.6– 2.8	1.6– 1.8
2	3.2– 4.3	16.1– 30.6	5.2– 12.4	2.9– 5.2	1.9– 2.1
3	4.4– 5.2	30.7– 45.1	12.5– 23.3	5.3– 9.5	2.2– 2.7
4	5.3– 6.6	45.2– 59.7	23.4– 37.9	9.6– 17.2	2.8– 3.6
5	6.7– 7.6	59.8– 74.2	38.0– 56.1	17.3– 31.3	3.7– 5.3
6	7.7– 8.9	74.3– 88.7	56.2– 77.9	31.4– 56.9	5.4– 10.2
7	9.0–103.4	88.8–103.4	78.0–103.4	57.0–103.4	10.3–103.4

[1] The one hundred and five unit areas are divided into seven groups of fifteen unit areas each. The first unit area value is 1.6, the sixteenth is 3.2, the thirty-first is 4.4, etc.

[2] Obtain range of data, 103.4–1.6 = 101.8; divide this by number of classes desired, 101.8÷7 = 14.54; start with lowest value and add to obtain class limits, 1.6 + 14.54 = 16.14 + 14.54 = 30.68 etc.

[3] Use A + X + 2X + ... + NX = B, where A is the smallest value and B is the largest value and N is the number of classes. Since A = 1.6 and 1 + 2 + 3 + 4 + 5 + 6 + 7 = 28 then 1.6 + 28X = 103.4, X = 3.64, thus 1.6 + 3.64 = 5.24 + 2 (3.64) = 12.52 + 3 (3.64) = 23.44 etc.

[4] Log 103.4 = 2.01452, Log 1.6 = 0.20412. Log difference 1.81040÷7 = .25863. Thus 2.01452–.25863 = 1.75589–.25863 = 1.49726 etc. Anti-log 2.01452 = 103.4, 1.75589 = 57.0, 1.49726 = 31.4 etc.

[5] Reciprocal 1.6 = .625,000,000, and 103.4 = .009,671,180, reciprocal difference = .615,328,820÷7 = .087,904,117. Thus .625,000,000–.087,904,117 = .537,095,883–.087,904,117 = .449,191,766 etc. Converting these reciprocals to density values .625,000,000 = 1.6, .537,095,883 = 1.86, .449,191,766 = 2.22 etc.

Figure 2.2 Example calculations of the equal frequency, equal step, arithmetic, geometric and reciprocal methods of data classification. (From Jenks and Coulson, 1963, p. 128) to have one shade.

The equal frequency method places the same number of data observations into each class. It is also generically called the quantile method, or more specifically the quartile, quintile, etc. methods depending on the number of classes that are used. This technique is simple to perform: the desired number of classes is divided into the total number of observations in the data set in order to determine the number of observations that will be in each class; this number of observations of the ordered data set is then placed into the first class, and then into the second, etc. recording the class breakpoints along the way (Figure 2.2, column 1). It is impossible to obtain empty or over-filled classes with this technique.

A drawback associated with the equal frequency method is that "ties" between data values may occur where breakpoints have been assigned. Frequently the quotient resulting from the division of the number of the observations by the number of classes is not an integer, thus necessitating unequal numbers of observations amongst classes. In addition, it has been suggested by Chang (1979) that the equal frequency technique often produces a map of high complexity, or a "quilt-like pattern".

The equal membership method classifies the data into equal portions of some secondary variable (Figures 2.3 and 2.4). The most commonly employed variable is the area of the map units themselves; by employing a cumulative frequency graph, class limits can be determined such that each class will cover an equal portion of the map area (Dutton, 1983; Robinson et al., 1984).

A fourth technique defines class limits using the mean-and-standard-deviation method. For this technique, the ordered data set is divided into classes at breakpoints based on the data set mean, and standard deviations on either side of the mean (Figure 2.5). This method often provides good results when the data set is normally-distributed, but if the data set is skewed it may result in classes with few observations and/or over-filled classes. When performed on a rectangular data set, it replicates the equal frequency method.

The nested-means method proposed by Scripter (1970) is similar to the mean-and-standard-deviation method. The overall mean of the data set is calculated as being the first break, and the means of the upper "half" and lower "half" are calculated as secondary breakpoints; the means of the resulting quarters can be determined as breakpoints, if desired (Figure 2.6).

The main advantages of this technique are that it

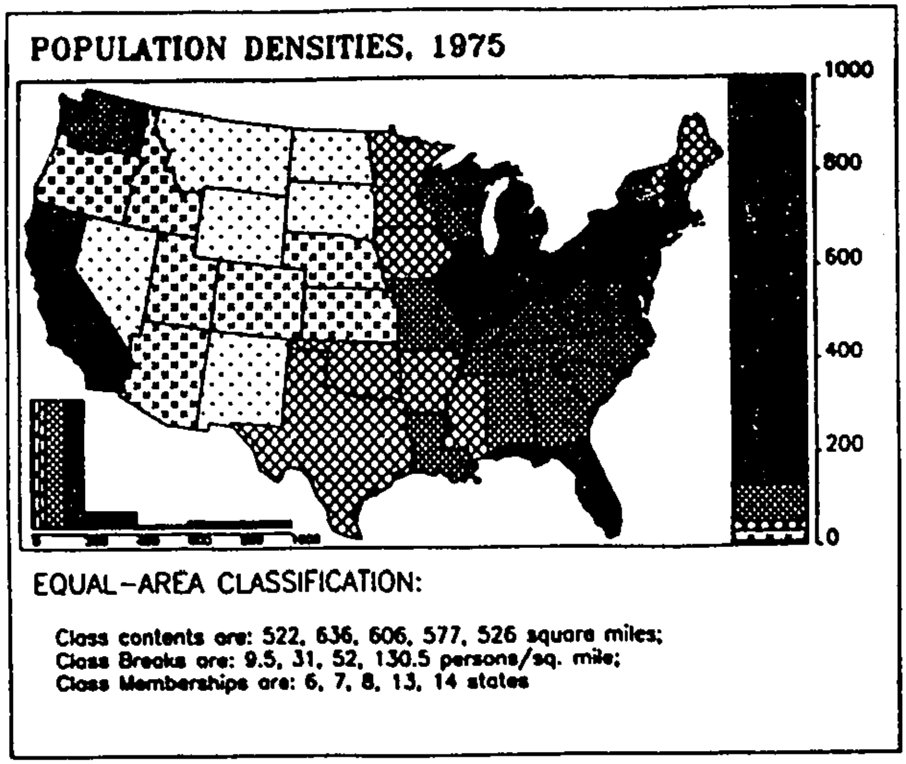


Figure 2.3 An equal membership classification that divides the land area equally into five classes. (From Dutton, 1983, p. 294)

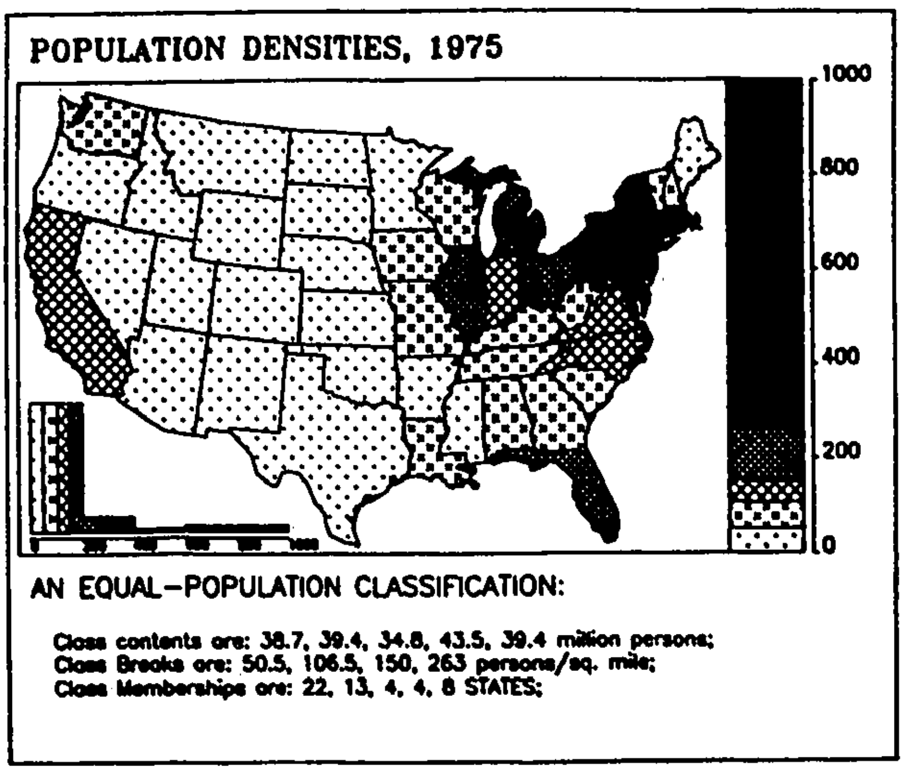


Figure 2.4 An equal membership classification that divides the population counts equally into five classes. (From Dutton, 1983, p. 294)

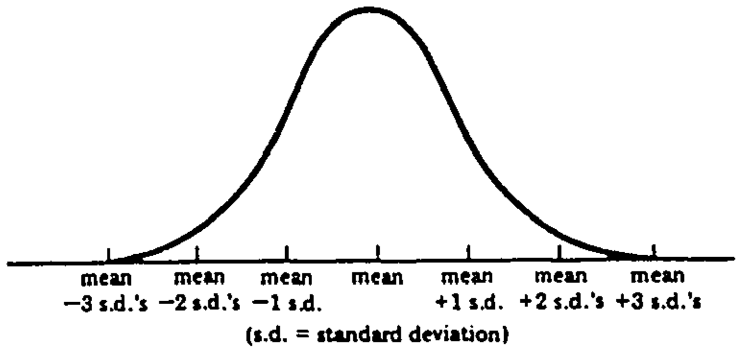


Figure 2.5 Illustration of the mean-and-standard deviation method of data classification.

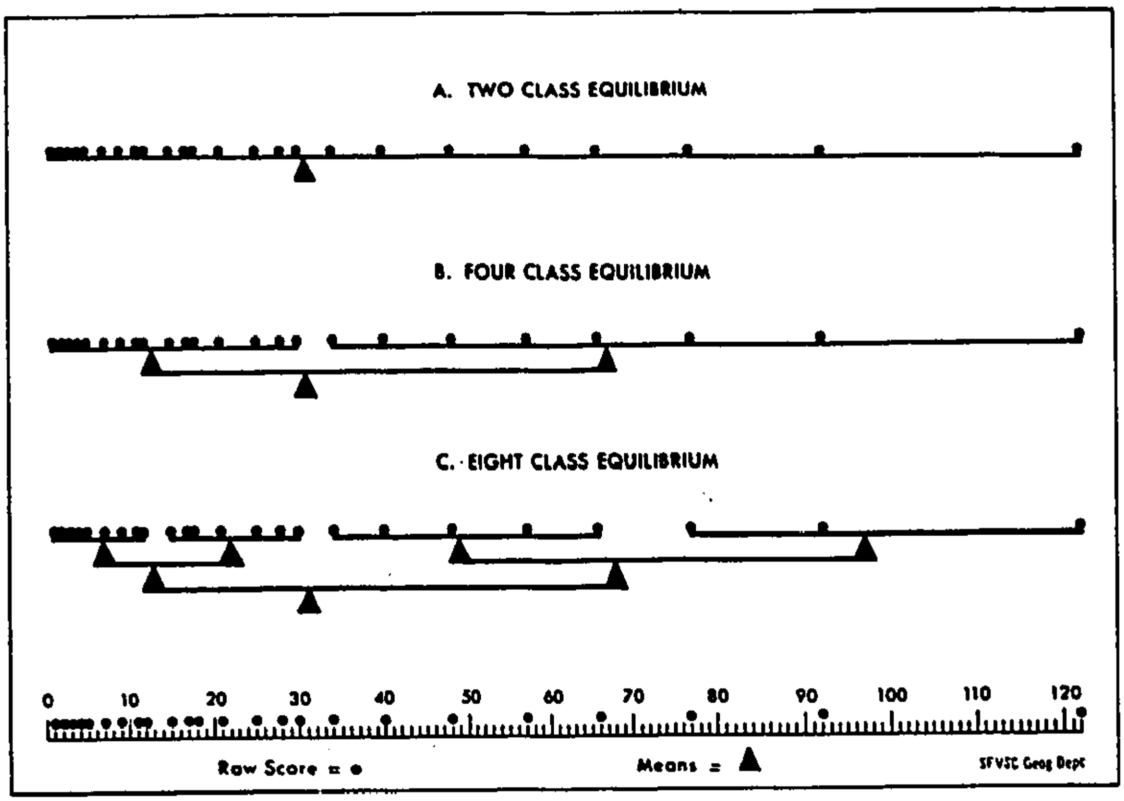


Figure 2.6 Successive levels of the nested-means hierarchy. (From Scripter, 1970, p. 391)

creates classes that can be considered to be in a state of "equilibrium", and that it is not possible to derive empty classes. A major disadvantage is that it is limited to 2, 4, 8, etc. classes.

2.2.2 Systematically Unequal Series

Three basic types of unequal series are used as classification techniques: the arithmetic, geometric, and reciprocal methods. Each of these methods involves the computation of the rate of increase of the values in the ordered data set. The arithmetic method defines class limits according to a numerical difference; the geometric method employs a numerical relation, and the reciprocal method uses a reciprocal function to define class limits. In each case, the classes derived are unequal, but in a systematic manner, since they represent a stated mathematical function (Figure 2.2, columns 3, 4, and 5).

All three of these methods can be easily calculated with the use of a computer. There is evidence that these methods portray L- and J-shaped data sets well (Evans, 1977), which is the major benefit of using these techniques. On the other hand, because the geometric and reciprocal methods utilise logarithms in their calculations of class breaks, neither are able to handle zeros or negative values in the raw data set.

One very important attribute of both the regular series and systematically-unequal series is that they do not group data values according to their similarity. For all of these techniques, it is the range of the ordered data sets that is the basis for creating class breakpoints. Data values are then grouped together simply by their position in the ordered data set.

2.2.3 Irregular Series

Techniques in the irregular series category can be further classified into three sub-categories: graphic techniques, iterative techniques, and "miscellaneous" techniques.

2.2.3.1 Graphic Techniques

Four types of graphs can be employed in the determination of "natural breaks" in a data set. The most commonly used graph is a simple array of the raw (ordered) data (Figure 2.7). Breakpoints are determined either as gaps in the array, or abrupt changes in the slope of the curve.

Another type of graph often employed is the frequency histogram, in which data values are ordered along the x-axis, and their frequency of occurrence recorded on the y-axis (Figure 2.8). Interpretation of

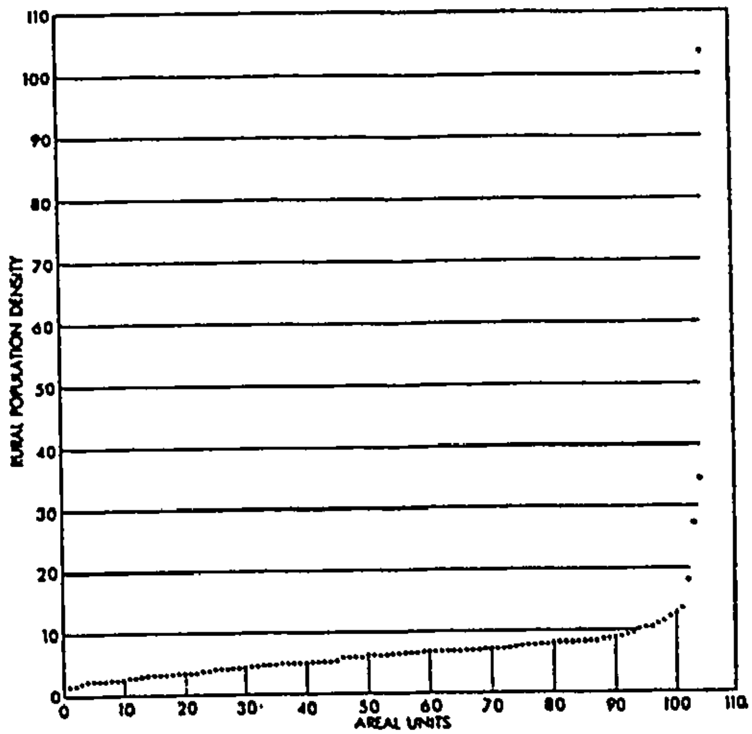


Figure 2.7 A graphic array of a raw (ordered) data set. (From Jenks and Coulson, 1963, p. 127)

natural breaks on the frequency histogram is done by identifying gaps or clumps along the x-axis. It should be noted that the arrangement of any gaps or clumps is completely dependent upon the number of chosen intervals and on the width of these intervals along the x-axis.

The two other graphic techniques both require that the size of areal units be known. The clinographic curve portrays the data values ordered on the y-axis against the cumulative areas (in percent of total area) to which they refer on the logarithmically-scaled x-axis (Figure 2.9).

The cumulative frequency curve shares the same y-axis as the clinographic curve, and also shows cumulative areas on the x-axis, but unlike the clinographic curve, the areas are displayed in square units on an arithmetic scale (Figure 2.10).

For both the clinographic and cumulative frequency curves, significant changes in slope are considered to describe natural breaks (Robinson et al., 1984). Theoretically, these methods are sound, but in reality undisputable "natural breaks" rarely occur and the graph interpretation procedure is highly subjective. Furthermore, there is often a discrepancy between the number of "natural breaks" that occur and the number of classes desired.

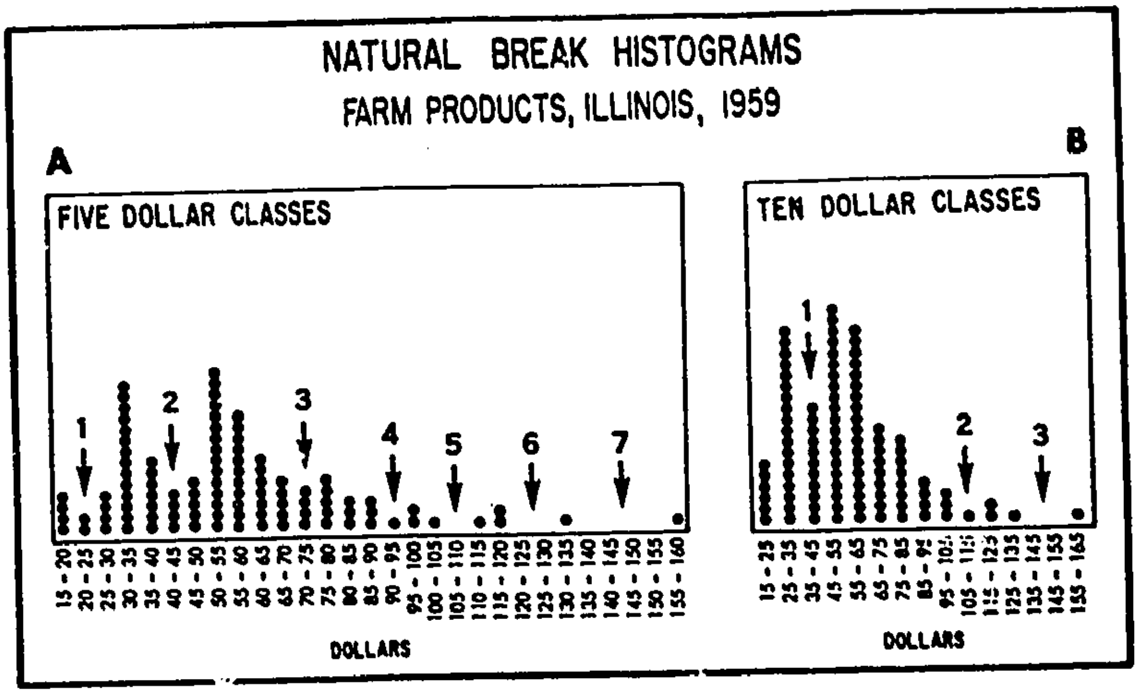


Figure 2.8 Mappable data are frequently plotted on histograms so that "natural breaks" in the statistical distribution can be identified and used as class limits for a choroplethic map. These two histograms illustrate one serious problem with this technique. The number of breaks, and to a lesser degree, their location, varies with the size of the interval used to plot the data. (From Jenks and Caspall, 1971, p. 223)

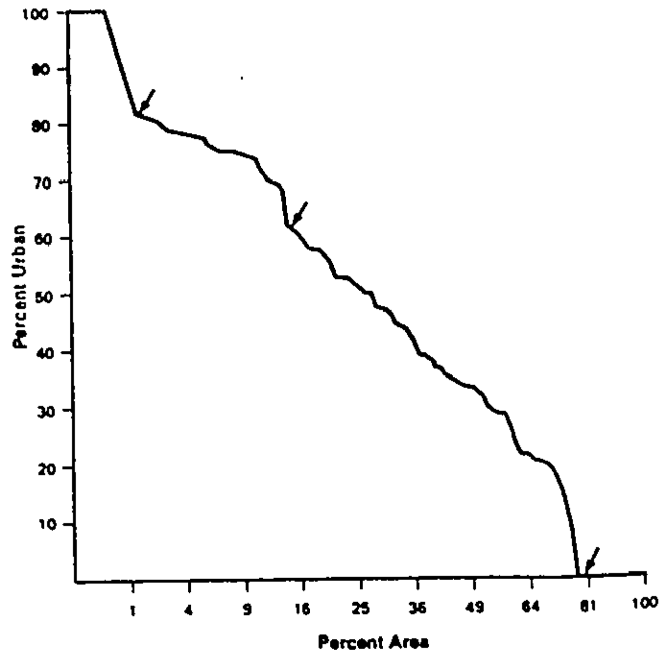


Figure 2.9 Illustration of a clinographic curve. The arrows denote points on the curve which might be selected for class limits. (From Robinson et al, 1984, p. 362)

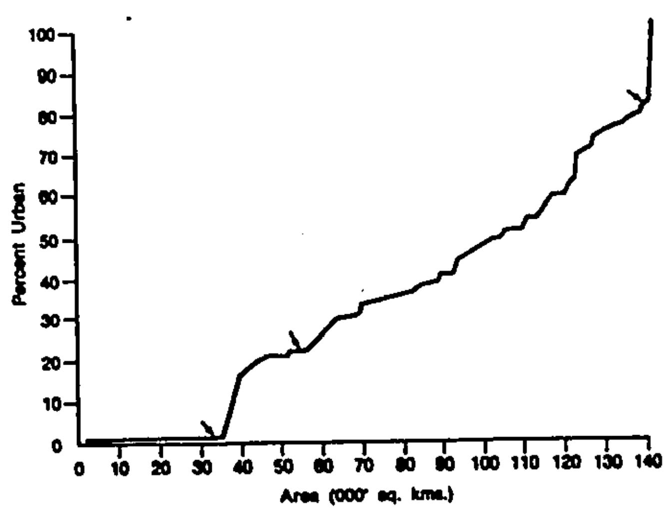


Figure 2.10 Illustration of a cumulative frequency curve. The arrows indicate points on the curve which might be used as class limits. (From Robinson et al, 1984, p. 364)

2.2.3.2 Iterative Techniques

Iterative techniques are made feasible through the use of computers. These techniques define "optimal" classes on the basis of some statistical criterion. Various combinations of class intervals of the ordered data set is attempted in order to find the "optimal" set, as measured by an accuracy statistic. The ways in which iterative techniques differ from one another are in the way they initially define breakpoints, and the accuracy statistic that they attempt to optimize. This section describes the theory behind the iterative techniques.

Jenks and Caspall (1971) were the first to introduce the iterative concept to data classification for choropleth maps. The first part of this procedure, referred to as "reiterative cycling" (Jenks and Caspall, 1971), arbitrarily divides the ordered data set into the desired number of classes, and then calculates the tabular accuracy index in the following manner:

1. Calculate the mean of the total data set (grand mean);
2. Calculate the mean of each class;
3. Calculate the worst-case tabular error, i.e. the sum of the absolute differences between each observation and the grand mean;

4. Calculate the sum of the tabular error for each class, i.e. the sum of the absolute differences between each observation in a class and the class mean;
5. Divide the sum of the class tabular errors (4) by the tabular error of the worst-case (3), and subtract the quotient from one. The result is the Tabular Accuracy Index (TAI) which can be considered to represent the accuracy of a set of class intervals as a percentage, if multiplied by 100. For example, a set of class intervals with a TAI of 0.82603 can be considered to be 82.603% accurate (Appendix 1).

The TAI is stored and class intervals are rearranged so that each value is grouped with the nearest class mean. The TAI is calculated again, and compared with the TAI of the first set of class intervals. If the second TAI is higher, the new set of class intervals replaces the first set; if it is lower, the first set of intervals is saved. This procedure is repeated over and over again (hence the term "reiterative cycling") until the class intervals and TAI repeat themselves.

The next stage of the process is referred to as "forced cycling". Using the class intervals resulting from the reiterative cycling described above, a new set of intervals is derived by forcing the first observation in the second class into the first class. Again, the TAI is calculated and compared to the previous TAI. If it is higher, another member of the second class is transferred to the first class, and the TAI is re-calculated. If the TAI decreases, the observation which was moved is placed back into the second class, and the same procedure attempted for shifting an observation from the third class into the second class. This "downwards" forcing is repeated until the TAI has been maximized, and then "upwards" forcing performed, i.e. moving the largest observation in the first class into the second class, calculating the TAI again, and so on.

The method can also be applied to maximizing the overview accuracy index (OAI), boundary accuracy index (BAI), or the composite accuracy index (CAI) also described by Jenks and Caspall (1971).

Other iterative techniques proceed in a similar manner as the reiterative cycling and forced cycling procedure of Jenks and Caspall, defining classes according to some statistical criterion. For example, Jenks (1977) devised two methods of class limit determination based on the "grouping for maximum homogeneity" principle of Fisher (1958). One of the methods minimizes the squared deviations about class means using the "Variance" measure. The second method minimizes the absolute deviations about class medians. In (Robinson et al., 1978) Jenks proposed a modification of the optimization statistics. Variance is replaced by a standardized measure called the "goodness of variance fit (GVF)", and the sum of absolute deviations is replaced by another standardized measure, the "goodness of absolute deviation fit (GADF).

Wasilenko and Moellering (1977) proposed an iterative technique that bases its grouping criterion on information theory. In this method, data values are grouped together by an "uncertainty statistic" which measures the probabilities of equality of the group members against a theoretical maximum probability of equality. The authors stated that a desirable grouping of data for choropleth mapping is achieved by maximizing within-class equality and minimizing between-class equality (Appendix 2).

The major drawback in using information statistics to classify data is that zero and negative data values cannot be processed due to the employment of a logarithm in the calculation of the probabilities.

Another iterative technique which may be used to classify data for choropleth maps is disjoint cluster analysis. In this type of analysis, a set of points called cluster "seeds" are selected as the first attempt to define means for classes. The other observations in the data set are grouped together with these on the basis of Euclidean distance; class means are re-calculated each time an observation is added to a class, and observations re-sorted as necessary.

This technique tends to be effective in portraying outliers, but unless a minimum distance between cluster seeds is defined, it may create classes of very few observations <u>and</u> over-filled classes in the same set of class intervals. In addition, for data sets with less than one hundred observations, results may not be satisfactory because the selection of cluster seeds is order-dependent (SAS User's Guide, 1982).

2.2.4 Other Techniques

Several other classing techniques have been proposed, yet rarely implemented. Principal components analysis has been used as a means of determining natural breaks. The first step in this technique is to calculate a matrix of similarity coefficients or "eigenvalues" between all pairs of data units. Similar pairs are placed in the same group in an attempt to find homogeneous natural clusters in the data set. The disadvantages of this method are that as with the "natural breaks" classing methods discussed in 2.2.3.1, the determination of natural clusters is subjective, and the desired number of "natural" clusters may not be obtained. In addition, Monmonier (1973) found that this method is limited in the number of observations that it can handle due to large amounts of computer storage required to process large matrices.

Several types of linkage analysis have been tested for the classification of data for multi-dimensional choropleth maps (Youngmann, 1972). Single, complete, average, and centroid linkage analyses are examples of these hierarchical grouping techniques, all of which employ similarity matrices based either on correlation coefficients or interobservation distances to group together "most-similar" pairs of observations. The result of any type of linkage analysis is a dendritic structure where, at separate points, all observations belong each to their individual class <u>and</u> all belong to the same class (Figure 2.11).

Because of the hierarchical nature of these

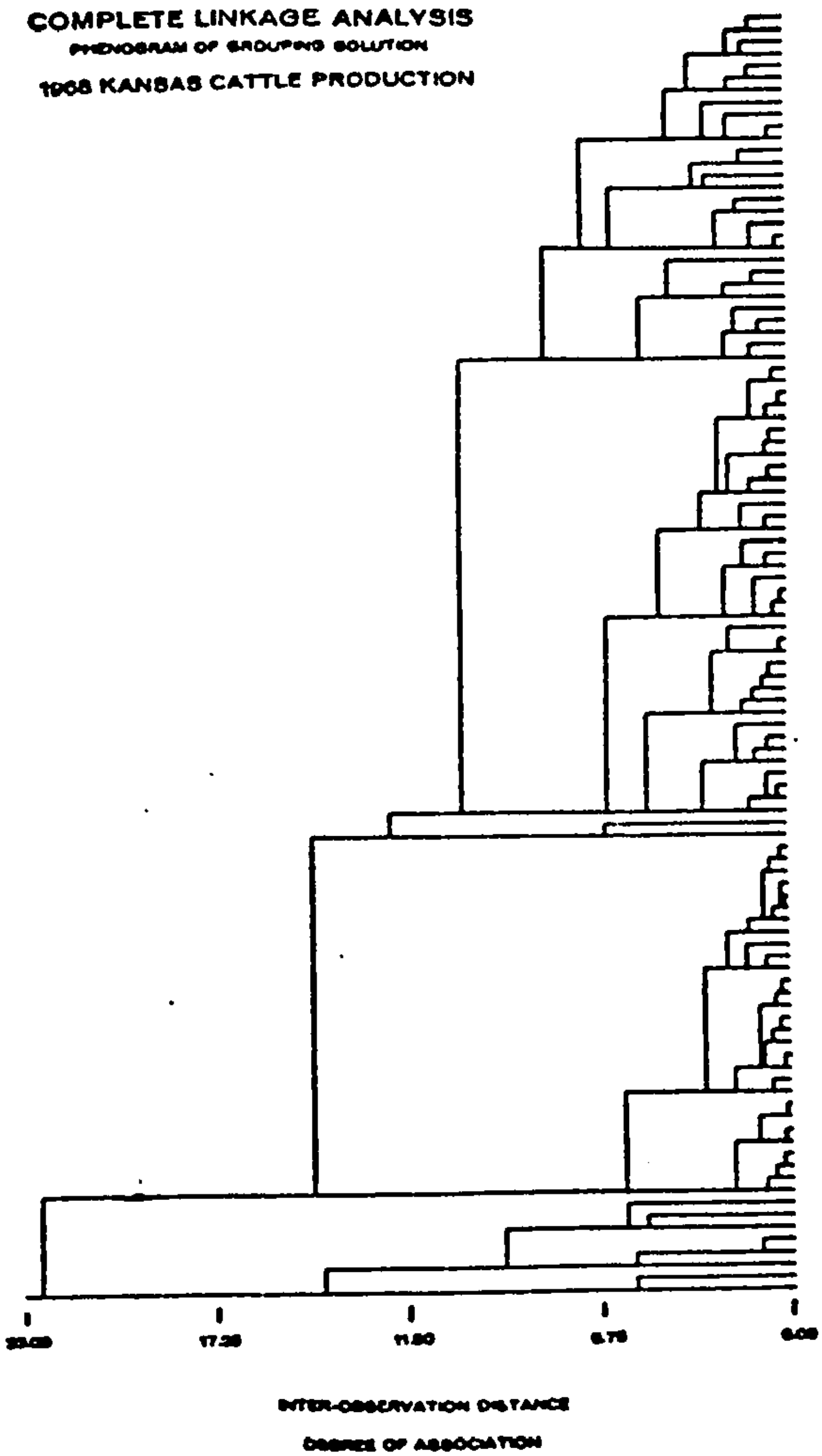


Figure 2.11 Example of hierarchical structure resulting from linkage analyses. (From Youngmann, 1972, p. 151)

structures, once observations have been paired together they cannot be separated further along in the analysis even if it appears that they are not as similar to one another as they are to other groups of observations. For this reason, the success of using linkage analysis to derive optimal class intervals has been quite limited (Youngmann, 1972).

In (1972) Monmonier proposed a method of classing that takes the spatial relationships of data values into account. In this method, the values are grouped together according to their adjacency, as determined by a similarity matrix describing the adjacency of all pairs of values.

Stegena and Csillag (1987, p. 145) suggested a "novel method" for the statistical determination of classes. First, iterative t-tests are performed on the data to determine the number of classes, then the equal frequency technique is applied. In this manner, the authors claim that the maximum information content is achieved.

2.3 Evaluating the Mathematical Accuracy of Classed Data

2.3.1 Mathematical Accuracy Measures

Initially, the evaluation of classed data sets took place in the form of simple visual comparison of the maps produced from the different classing techniques. This method was supplemented by the application of statistical measures of the mathematical accuracy of the classed data beginning with the work of Jenks and Coulson (1963). In this study, they propose a procedure for data classification which begins by utilizing graphic techniques and ends with the application of the "D-value", a statistic which measures the similarity of values within class ranges (Appendix 3).

In the four-part procedure proposed by Jenks and Coulson (1963), the data are first displayed on a frequency graph and a clinograph in an attempt to locate "natural breaks". Secondly, a graphic array of the data is prepared by plotting the ordered data set, and the shape of this curve compared to curves of known mathematical functions such as straight line, arithmetic, geometric, and reciprocal progressions. The third step in the procedure is to calculate the class breakpoints according to the function or functions chosen. The final step involves testing of all of the classed data sets derived from the previous steps through the application of the D-value, which compares the data classes to theoretical data classes. Low D-values are analogous to a low coefficient of variation, and therefore the classing technique which produces the lowest D-value is selected.

Following the introduction of the D-value, great emphasis was placed on the mathematical accuracy of class intervals, and other statistical measures were developed. Some methods of evaluating class intervals were introduced in conjunction with new classification methods, while others appeared once the literature on classing methods became voluminous yet remained inconclusive.

As mentioned in section 2.2.3.2, Jenks and Caspall (1971) introduced an iterative technique for the classification of data on the bases of tabular, overview, or boundary error. As part of this technique they defined statistics to measure the magnitude of each of these types of errors. These statistics are applied to classed data as measures of mathematical accuracy, where minimum error equates with maximum accuracy. Overview error is the absolute sum of the differences in volume between the areal units of the generalized data set (choropleth maps) and the raw data set. Tabular error is based on the same principle, but measures the difference in values between the classed data and the data model, not the difference in volume. Boundary error refers to the human tendency to overemphasize the importance of boundaries between areal units of differing data values. It is measured as the absolute sum of the differences between the values of the classed boundaries and the boundaries of the raw data set. Boundaries are assigned a value equal to the average of the two values that they separate. The overview accuracy index (OAI), the tabular accuracy index (TAI), and the boundary accuracy index (BAI) employ maximum possible error amounts as limits in their calculations (Appendix 1). In addition, the authors recommended the use of an index which combines the TAI, OAI, and BAI as the composite accuracy index (CAI). The CAI, calculated as (OAI + TAI + BAI), when divided by its algebraically-defined maximum, derives the map accuracy index, or MAI (Jenks and Caspall, 1971).

Two pairs of accuracy measures which are closely related are the "Variance" (Jenks, 1977) and "goodness of variance fit" (GVF) (Jenks, in Robinson et al., 1978), the "sum of absolute deviations about class medians", and the "goodness of absolute deviation fit" (GADF) (ibid.). Variance measures the sum of squared deviations of observations in a particular class to the class mean. Variance is also the numerator in the GVF equation, where the denominator is the sum of squared deviations from the total mean. In a similar fashion, the "sum of absolute deviations about class medians" is the numerator of the GADF formula. Both the GVF and the GADF measure the accuracy of a given set of class intervals (numerator) to the accuracy of the worst possible case, or the mean model (denominator).

Another statistic used to measure accuracy is the "percent of equality" measure (Wasilenko and Moellering, 1977). The derivation of this statistic is from information theory, and it refers to the within-class equality of observations. As within-class equality is maximized, the equality of observations <u>between</u> classes is minimized.

Other statistics used to measure the accuracy of classed data relate to the "between- and within-class variance" of analysis of variance. Examples of this include the F-ratio (Monmonier, 1972) and the R-squared measure employed as a by-product of the SAS (1982) disjoint cluster analysis algorithm.

Numerous statistics to measure the mathematical accuracy of classed data are found in the literature. Some concentrate on maximizing within-class homogeneity of data values, while others create hierarchical grouping of data. No one measure has gained universal acceptance or has been supported by a majority of reviewers as the appropriate method of data classification for choropleth mapping.

2.3.2 Experimentation using Mathematical Accuracy Measures

Some of the mathematical accuracy measures described in section 2.3.1 have been applied in various other research studies in addition to the ones in which they were proposed by their author(s).

Youngmann (1972) used the TAI proposed by Jenks and Caspall (1971) to evaluate four different types of linkage analysis for the determination of optimal classes for multi-dimensional choropleth maps. He found that the TAIs of all of the hierarchical techniques were low relative to iterative techniques, and that their TAIs could be improved by reiterating observations about the class means.

Monmonier (1972) tested the effectiveness of his proposed contiguity-biased classing approach by comparing natural breaks and equal steps classifications with versions of each modified by the contiguity matrix. Evaluation using visual inspection and the F-ratio led Monmonier to conclude that the contiguity-biased method succeeded in aggregating regions of the maps, but at the expense of mathematical accuracy.

Monmonier (1973) tested his proposed method of defining "natural breaks" through principle components analysis with the iterative technique proposed by Jenks and Caspall (1971) using the TAI. The principle components method produced TAI values that were comparable to those of the Jenks-Caspall algorithm. Thus, he concluded that despite the relatively extensive computer costs of the principle components method, it is a promising alternative to existing classing methods because it effectively portrays natural clusters, and determines the optimum number of classes for a choropleth map.

Chang (1979) used the GVF and GADF measures proposed by Jenks (in Robinson et al., 1978) to test the accuracy of classed data derived from the application of the equal step, equal frequency, arithmetic, mean and standard deviation, and natural breaks classing techniques to three sets of data. He found that the equal frequency technique produced the least accurate classes, while the equal step method provided reasonably good classifications.

Smith (1986) employed the GVF to evaluate the equal frequency, equal step, standard deviation, natural breaks, and Jenks and Caspall (1971) iterative technique on one-hundred-seventeen data sets. The GVF indicated that the iterative technique was the only one which consistently produced accurate classed data. For the "traditional" classing techniques (for example the standard deviation and equal steps methods) the GVF varied by as much as thirty-eight percent.

Coulson (1987) used the GVF to compare the equal step, equal frequency, standard deviation, arithmetic and geometric techniques, and Jenks (1977) iterative technique. He found that the iterative solution provided the most accurate results, while the performances of the "traditional" techniques were much more variable.

2.4 Visual Characteristics of Choropleth Maps

2.4.1 Map Communication Studies

Cartographic and geographic literature on classification for choropleth mapping focussed on mathematical accuracy until the 1970s. During that decade, attention turned towards the visual characteristics of choropleth maps. This change in focus was undoubtedly fueled by the research in map communication that became increasingly popular during that time.

The increased interest in map communication led to the development of numerous models and descriptions of cartographic communication (see (Board, 1981) for a summary) ranging from relatively straightforward descriptions of the cartographic process (Figure 2.12) to more complex schematics of the communication of cartographic information (Figure 2.13).

The findings and conclusions of map communication studies strongly suggested that cartographers cannot focus solely on the mathematical accuracy of maps, but that they must also consider the amount of information that their maps actually communicate to users. In discussing map communication, Olson (1976, p. 151) stated that:

> "Surely one of the major reasons that we present information on a map is to organize it into a comprehensible visual form".

After lengthy experimentation on choropleth map accuracy, Youngmann (1972, p. 186) concluded that:

> "...there are many other factors that affect the communication between map maker and map reader that merit investigation."

These findings were not lost on geographers studying the choropleth map data classification problem, as they shifted their focus to the visual characteristics of choropleth maps.

2.4.2 Map Complexity

Numerous experiments (Muehrcke, 1969; Monmonier, 1974; Muller, 1976b; Chang, 1978, 1979) have indicated that map communication and perception are enhanced by decreasing map complexity. Not surprisingly, this finding

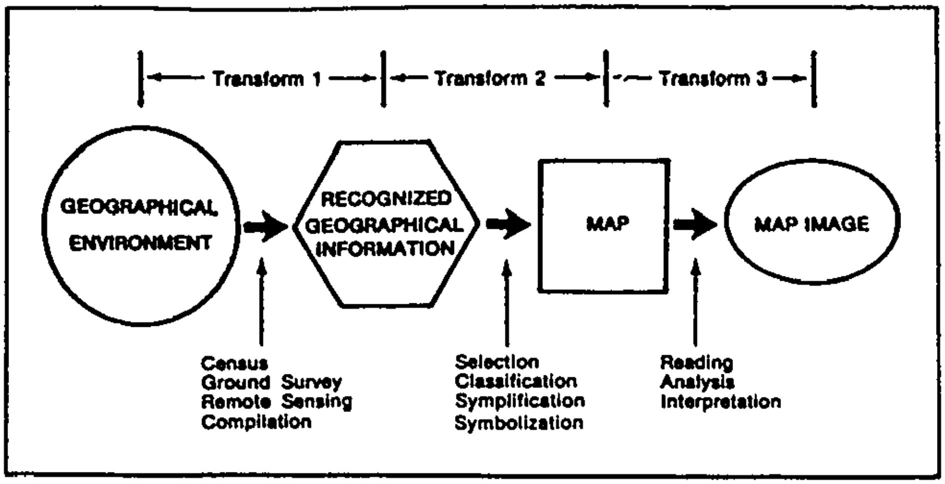


Figure 2.12 Model of the Cartographic Process. (From Muehrcke, 1981, p.16)

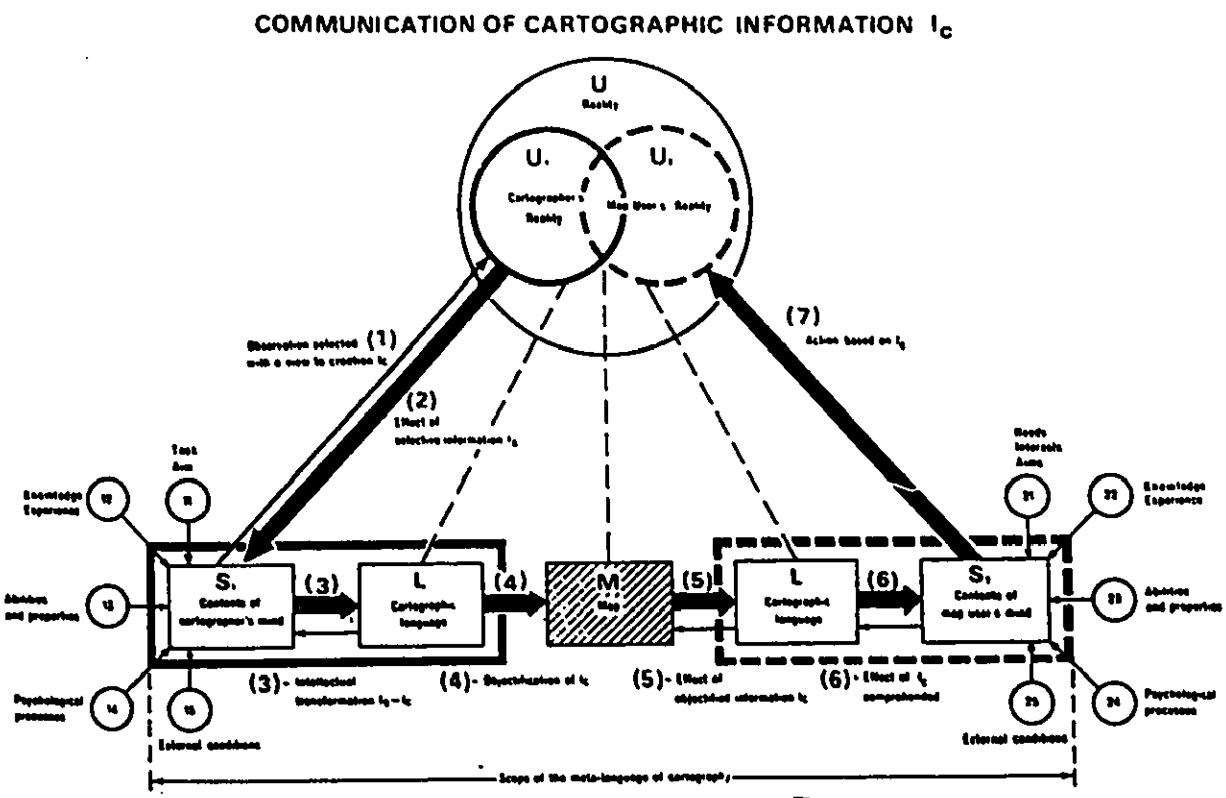


Figure 2.13 Model of Cartographic Communication. (From Kolacny, 1969, p. 48)

is supported by the Gestalt view of perception in which the "whole" is thought to be "greater than the sum of the parts". In cartographic terms, it indicates that a map user is more likely to gather and understand information presented on a map where areal units are grouped into contiguous regions rather than fragmented.

One of the first studies of visual characteristics of choropleth maps utilized trend surface analysis, under the assumption that the power of the polynomial equation which is fitted to the map surface reflects the visual complexity of the map (Muehrcke, 1969). Another visual characteristic is described as spatial autocorrelation, which refers to the correlation of shades of adjacent areal units. The degree of spatial autocorrelation has been measured using Kendall's t-statistic, a non-parametric correlation coefficient for ordered, classed data (Olson, 1972b).

Monmonier (1974, p. 159) stated that:

> "cartographic and psychological literature both identify complexity as an important influence of visual pattern recognition."

In the same paper he proceeded to define three new indices for measuring complexity based on the number of contiguous areal units assigned to the same class: the fragmentation index, the size disparity index, and relative entropy.

The fragmentation index is calculated as the number of regions (i.e. number of groups of adjacent areal units assigned to the same class) divided by the number of areal units on the map (therefore ranging in value from zero to one). The size-disparity index utilizes a Lorenz curve to measure the difference in size between areal units. Relative entropy measures the degree of equality of size of the areal units; high inequality reflects high map complexity.

Muller (1975b) defined six visual characteristics of choropleth maps and statistics to measure them: Blackness, Redundancy, Aggregation, Compactness, Complexity, and Contrast. Blackness simply refers to the amount of dark shading on a map, and is calculated as a function of the number, size, and shade of the areal units on a map. Redundancy reflects the proportion of areal units that belong to the same class. Aggregation is ostensibly the opposite of fragmentation, described in (Monmonier 1974), but is calculated on a different principle; it is based on the class membership of each pair of adjacent areal units. Compactness is a function of the adjacency of all areal units to one similar unit, and the subsequent adjacency of all other units. Complexity is calculated by recording the class memberships of three mutually contiguous areal units meeting at a single point. Contrast refers to the cumulative difference in tones between each pair of adjacent cells.

Ultimately, map complexity is the characteristic which is measured in most of the measures described above. Fragmentation, aggregation, trend surface analysis, entropy, spatial autocorrelation and the number of regions indices are all measures of map complexity.

Map complexity is a critical issue in data classification because of the suggested trade-off between complexity and accuracy (Jenks and Caspall, 1971; Monmonier, 1972; Muller, 1976b), i.e. as a map becomes more mathematically accurate through decreasing generalization, it gradually becomes more complex (Figure 2.14).

2.4.3 The Evaluation of Visual Characteristics

Although experimentation on the visual characteristics of choropleth maps was widespread during the 1970s, many researchers found that their experiments were inconclusive, and that further research in the field of perception was necessary in order for them to continue their studies.

Monmonier (1974) compared fragmentation, the size disparity index, relative entropy, autocorrelation (using

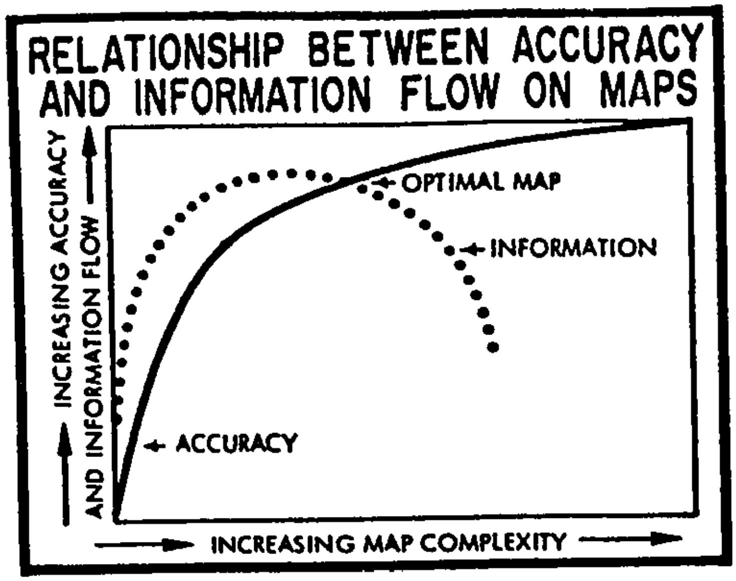


Figure 2.14 Graphic display of the relationship between map accuracy and information flow. The optimal representation of a distribution will be achieved when both of these map functions are maximized. (From Jenks and Caspall, 1971, p. 243)

Kendall's tau) and trend surface analysis on data for which the means, standard deviations, and minimum and maximum values were calculated. Intercorrelations of the measures were calculated to determine similarities in behaviours amongst them. The results showed a close inverse relationship between fragmentation and autocorrelation; similarly, relative entropy and size disparity were strongly negatively correlated. The order of the trend surface most closely representing the surface did not appear to be related to any of the other measures. On the basis of these findings, Monmonier suggests that there are three components of map complexity applicable to choropleth maps: fragmentation and autocorrelation, size inequality amongst areas, and spatial trends.

Olson (1975) tested map complexity (as autocorrelation) by presenting geography students with a series of maps of varying complexity and asking them to put them in order of decreasing complexity. Autocorrelation was calculated using Kendall's tau statistic, the proportion of identical neighbours, the average differences between neighbours, the number of clusters, and a weighted proportion statistic that takes into account both clusters and the irregularity of values within clusters. Results of the study found that subjects ordered the maps not only on the basis of complexity but also by pattern sequences. No comparisons amongst the methods of calculating autocorrelation were made, and therefore the testing was inconclusive in this regard.

Map blackness, aggregation, complexity and contrast were employed in an experiment by Muller (1976b) to determine the effect of varying numbers of class intervals on choropleth maps. These four measures were calculated for map sets with the number of classes ranging from three to nine. Map blackness and contrast were found to be nearly identical for all the map groups, while aggregation decreased and complexity increased with increasing numbers of classes. Muller concluded that the findings illustrate the contradictory relationship between map enhancement (decreased generalization through increased number of classes) and map consistency (low complexity).

In (1978) Chang tested five of the "traditional" classing techniques methods using test subjects to measure visual characteristics. Results of the study indicated that map users prefer maps of low complexity, but are concerned also with the distribution of values in each class. The other major finding was that there was a high degree of consistency amongst the complexity measures, causing Chang to suggest that any one of the measures on its own is a reliable indicator of map complexity.

Measures of map complexity and map accuracy were tested simultaneously in an experiment by Chang (1979). The classification techniques employed were the equal-step, equal frequency, arithmetic, mean-and-standard-deviation, and natural break methods; map accuracy was measured using Jenks' GVF and GADF statistics. Visual measures included the number of regions, fragmentation, aggregation, contrast, and the size disparity index. Chang tested the hypothesis that given the same number of classes, maps with an uneven distribution of class frequencies are more likely to have simpler pattern characteristics than maps with an even allocation of values in classes. He found that the measures of visual characteristics were highly consistent with one another, and that map complexity increased with more even distributions of class frequencies. However, map complexity did not necessarily increase with map accuracy. The equal frequency classing method produced maps that were the least accurate and the most complex, whereas the equal step classing method produced the most desirable maps in terms of both accuracy and complexity.

2.4.3 Map Comparison Experiments

A large portion of the experiments involving visual characteristics of choropleth maps were experiments that do not directly test the visual characteristics themselves, but instead incorporated them in other experiments in an attempt to gain a better understanding of how these characteristics are perceived. In most cases this testing was in the form of map comparison experiments.

Monmonier (1975) approached the selection of class intervals with an algorithm that attempted to maximize the visual correspondence between the variable to be categorized and a referent variable with which it is associated. In (Monmonier, 1976) he modified this algorithm to include weighting factors for the internal homogeneity of mapping categories, land area of the enumeration tracts, and other associated factors.

Muller (1976a) used map coadjacency, complexity, correlation and blackness to test subjects' abilities to judge spatial correspondence. In a similar experiment (1976b) he used correlation, complexity, and blackness to try to identify the elements of visual recognition that map readers use to associate or differentiate spatial distributions on choropleth maps. His experimentation with map comparison was based on the premise that visual map comparison can supplement statistical analysis in providing information as to where and how spatial correspondence occurs.

Muller (1976a, 1976b) concluded from his experiments that both blackness and complexity have more of an impact on map comparison than contrast. In addition, he found that positive correlations were perceived much better than negative correlations.

Lloyd and Steinke (1976, 1977, Steinke and Lloyd, 1981) performed a series of map comparison experiments involving visual characteristics. Several of these experiments involved the use of pairs of choropleth maps produced using different classing techniques. Subjects were asked to rate the similarity of pairs of maps, and measures including blackness, complexity, and correlation were calculated. In (1983) they performed a similar experiment, except that the map pairs were shown consecutively rather than simultaneously; Steinke and Lloyd believed that in this manner they could try to determine whether or not people can form mental images of choropleth maps, and if they make comparisons of images in the same way as comparisons of actual map patterns.

As a result of their initial map comparison experiments, Lloyd and Steinke (1976) found that subjects used both blackness and similarity of pattern to judge map similarity, and that the classing technique did not have a significant effect on their ability to judge map similarity. Their later experiments added to their earlier finding regarding map blackness, in that of all the visual characteristics, map blackness is used most for map comparison. Complexity and correlation were also used by subjects, but not to the same extent as map blackness, and not necessarily in that order. Contrary to their first experiment, the 1981 testing showed that the use of visual characteristics was affected by the classing technique employed.

2.5 Data Distributions

As a prelude to statistical analysis of any kind, data sets are often examined in order to determine the presence of characteristics which may have an effect on the analysis. The most commonly referred to characteristics include skewness and kurtosis, both of which refer to the distribution of the values within the data set. These data distributions are described by the shape of the curve of the ordered values of the data set as plotted on a graph. Skewness refers to the degree to which the bulk of values in an ordered data set occur on one side or the other of the 'middle' of the data set.

Kurtosis refers to the 'peakiness' of the data set, or the degree to which the values of the ordered data rise from the x-axis (Figure 2.15).

Probably the most common data distribution is that of the bell-curve, also referred to as a "normal distribution", where the bulk of the data values occur at the middle of the ordered data set. Other common distributions include J- and L-shaped, rectangular, and bimodal or multimodal (Figure 2.16).

The number of choropleth data classification experiments which have taken the distribution of data into account is limited. Their results, however, have indicated that different data distributions affect the "success" of class intervals.

Monmonier (1972) stated that the equal intervals and equal steps techniques are acceptable methods for determining class breaks in rectangular data distributions. In the same article, he found that skewed data were better represented by the application of the systematically unequal series techniques (described in section 2.2.2) that use arithmetic or geometric functions to determine class breakpoints.

Evans (1977) found that the geometric technique for the selection of class intervals gave a good

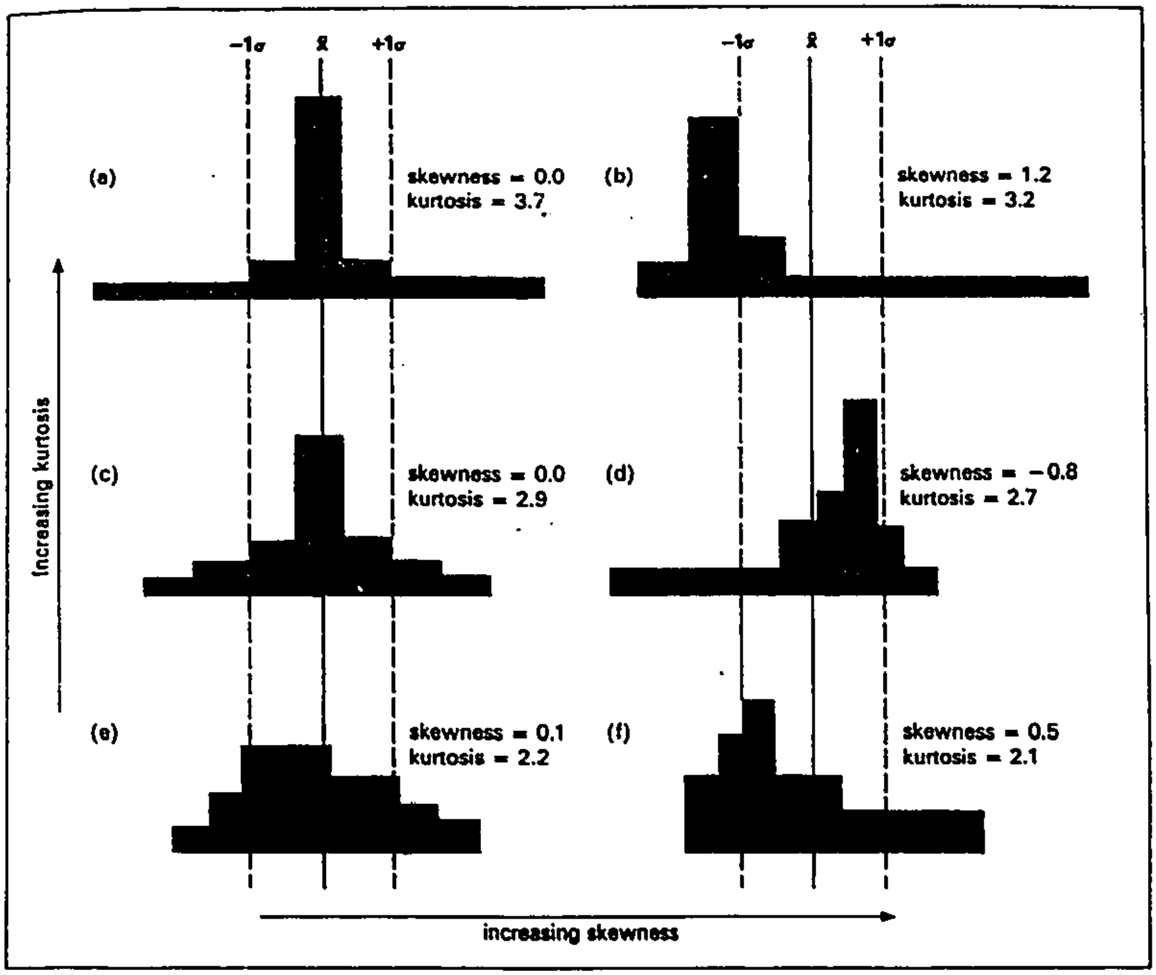


Figure 2.15 Skewness and kurtosis (from Ebdon, 1977, p. 27).

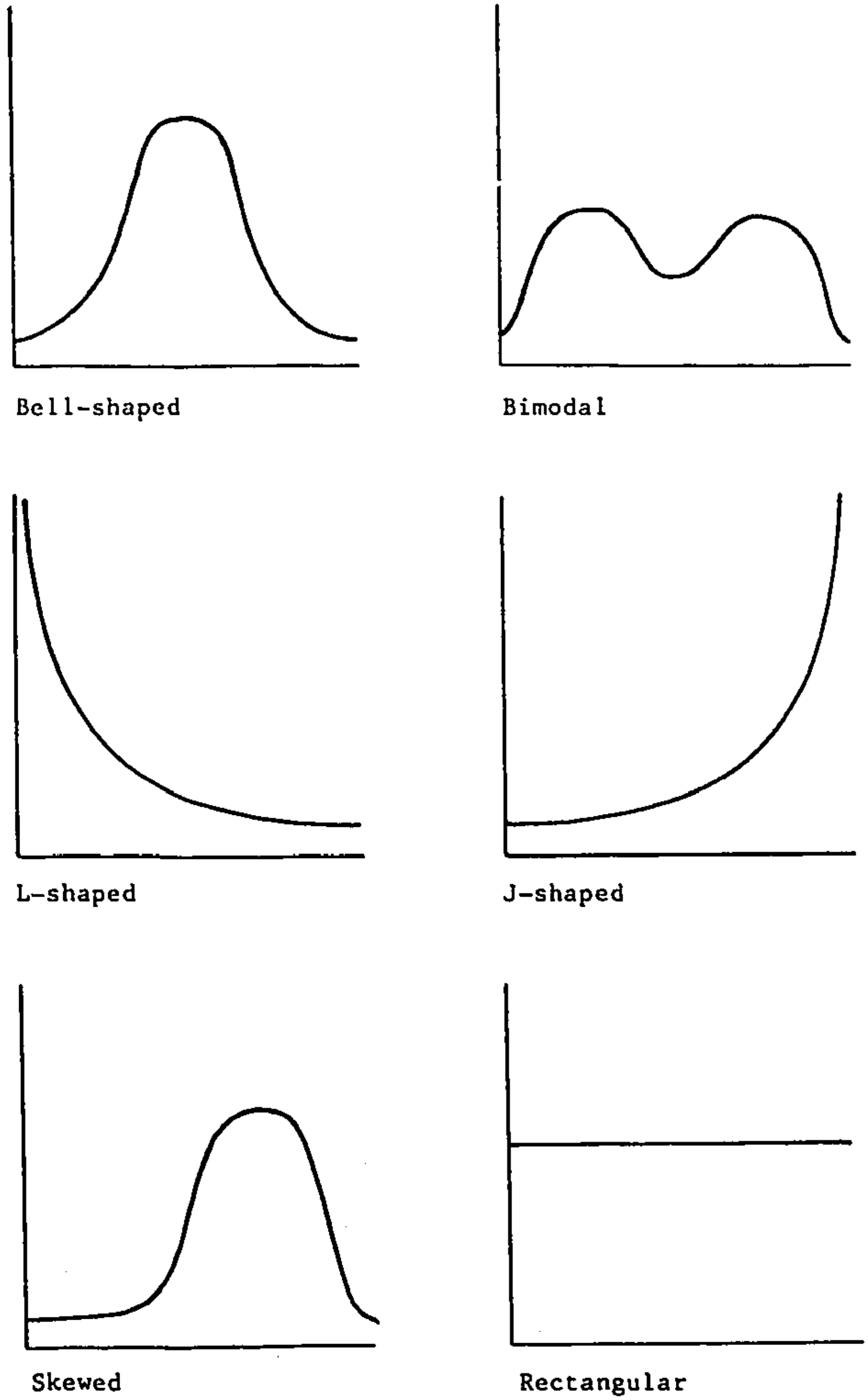


Figure 2.16 Data distributions

representation of J-shaped distributions, but did not provide good results when applied to rectangular or normally-distributed data sets. In the same experiment, he found that the mean-and-standard-deviation method, which operated well on normally-distributed data, did not provide good results when applied to rectangular or J-shaped distributions. The equal step technique performed extremely well when applied to rectangular data distributions, but under-represented the middle classes of normal distributions and over-represented the lower observations in J-shaped distributions.

Chang (1979) concluded that the equal frequency method provided a highly accurate representation of rectangular distributions, but when applied to normal distributions produced maps which were highly complex and highly inaccurate.

The first experiment to test the effect of data distributions on iterative classing techniques was performed by Wasilenko and Moellering (1976). In this study, they compared their information theoretic classing technique (see 2.3.1) with the iterative technique based on tabular accuracy proposed by Jenks and Caspall (1971). For this experiment they generated five data sets of differing distributions, and measured the accuracy of the classed data with the TAI and the percent of equality measures. They found that regardless of data distribution, the Jenks-Caspall technique produced better TAI values, as expected. For the percent of equality measure, the Wasilenko and Moellering technique produced slightly better results for U-shaped and rectangular distributions, tied with the Jenks-Caspall technique for the normal, bimodal, and left-skewed distributions, and was worse for the L-shaped distribution.

Smith (1986) tested the effects of skewness and kurtosis on five classing techniques. The study used the GVF proposed by Jenks (Robinson et al., 1978) to evaluate the accuracy of data classed using the equal frequency, equal step, standard deviation, natural breaks techniques as well as the Jenks and Caspall (1971) iterative technique. Results of his study indicated that the equal frequency and standard deviation methods perform adequately on normal distributions but poorly on skewed or kurtosed distributions. The natural breaks and equal step techniques were both unreliable in that they produced accurate and inaccurate classes regardless of data distribution. The only classing method that consistently provided accurate results was the Jenks-Caspall iterative technique.

Coulson (1987) tested the effects of rectangular and skewed distributions on the standard deviation, arithmetic, geometric, equal frequency, equal step, and Jenks (1977) iterative techniques. He found that the iterative technique produced the most accurate classes in both cases. For the rectangular distribution, the equal step, equal frequency, and arithmetic techniques provided acceptable solutions. For the skewed distribution, the arithmetic technique produced accurate classes, but all of the other "traditional" techniques created classes of very low accuracy, or empty classes.

To date, there has not been a clear indication of whether or not data distributions effect the accuracy of iterative classing techniques.

2.6 Other Issues

2.6.1 Introduction

This section outlines some additional issues which have appeared in choropleth map literature over the past few decades.

2.6.2 Guidelines for Choropleth Mapping

While the number of classes chosen is dependent upon the number of observations in the data set and the purpose of the map, there are several universally accepted guidelines regarding the number of classes to use in choropleth mapping. Although the exact numbers differ according to the author or experiment cited, between three and eight classes are normally used. Employing fewer than three classes is thought to create an overgeneralized map, sacrificing the accuracy of the data. On the other hand, research has indicated that the human eye cannot distinguish between more than eight different tonal values (black and white), and therefore the number of classes on a choropleth map should not exceed eight (Jenks and Coulson, 1963; Johnston, 1968; Jenks, 1967; Robinson et al., 1984). Other problems associated with the use of more than eight classes are that the map begins to lose its generalizing function, and that a greater degree of accuracy may be implied than actually exists.

Although there is a lack of consensus regarding an optimal classification method, research in the field has yielded several guidelines regarding choropleth classes. For example, it is generally accepted that choropleth class intervals should:

1. Encompass the full range of the data;
2. Not have any empty classes or any overlapping classes (Jenks and Coulson, 1963; Monmonier 1975);
3. Maximize homogeneity within classes; and

4. Reflect critical values (if any) in the data set (Robinson et al., 1984).

Other recommendations regarding choropleth class intervals have been made, but are more contentious than those described above. Jenks and Coulson (1963, p. 120) state that class intervals should divide the data into "reasonably equal groups", and should "have a logical mathematical relationship if possible". These recommendations have both been criticized: if equal groups are used, outliers and other critical data values may disappear into groups, and "logical" mathematical relationships are not necessarily "logical" in terms of map user perception (Stephanovic and Vries-Baayen, 1984).

Other researchers (Evans, 1977; Chang, 1979) have recommended that class intervals be selected according to the distribution of the data set, while numerous others (Fisher, 1958; Wasilenko and Moellering, 1977; Coulson, 1987, among others) emphasize the importance of minimizing between-group homogeneity as well as maximizing within-group homogeneity.

2.6.3 Error in Choropleth Mapping

The literature on choropleth mapping that has accumulated over the past few decades includes many articles focussing on the identification, measurement, and possible minimization of the different types of map error (see (Muller, 1987) for a summary). Choropleth map error, as a subset of map error, can be viewed as either inherent or induced error, or a combination of the two.

Inherent error is a direct result of the generalization process, i.e. because each individual value in the data set is not displayed as its own value, but instead is displayed as the representative value for the class to which it belongs, a certain amount of error of representation is inherent. This type of error, which cannot be eradicated, varies widely depending upon the classification method employed (Jenks and Caspall, 1971). Research performed and statistical measures developed regarding the mathematical accuracy of class intervals are, for the most part, addressing inherent error.

Induced error occurs when the user misinterprets the data presented on a choropleth map, either because of his/her cartographic background, or because the map is poorly designed. Induced errors can be eliminated by taking into consideration the needs of the user, through careful map design, and most importantly for choropleth mapping, through the selection of class intervals that best portray the geographic surface.

The difference in the size of areal units on choropleth maps creates an error that can be considered as inherent or induced, or a combination of the two. When areal units differ greatly in size, high values assigned to large areal units can be perceived as disproportionately large values. It has been suggested (Williams, 1976), that this type of error must be "corrected" by taking the sizes of areal units into consideration when assigning values to units.

2.6.4 Unclassed Choropleth Maps

In the early 1970s, Tobler (1973) proposed the use of unclassed choropleth maps in an answer to the choropleth data classification problem. On the unclassed choropleth map, each areal unit is shaded uniquely in proportion to its data value, thereby eliminating the classification procedure. Since it's introduction, the unclassed choropleth map has created a good deal of controversy in cartographic literature. Dobson (1973, p. 358) argued that the unclassed choropleth map "...renders the choroplethic form of map vacuous". His opinion stemmed from the theory that as the amount of information on a map increases, map complexity also increases; if complexity becomes great enough, the flow of information actually decreases, and the map user cannot comprehend all of the information presented on the map.

Dobson's view is supported by psychophysical research (Mackworth and Morandi, 1967; Gould and Dill, 1969) on how people collect information visually. In contrast however, Peterson (1978) tested subjects on their ability to perceive values in unclassed choropleth maps, and found that people are much better at perceiving values than previously thought. He concluded (Peterson, 1978, p. 6) that because unclassed maps are more accurate (i.e. less generalized) than classed maps, and because they do not seem to be too complex to be understood, "Cartographers, it would seem, must now justify their reasons for using the class interval method..." .

Dobson (1980) criticized Peterson's research on the basis of his methodology, and in turn, Peterson (1980) responded by attempting to justify the validity of his experiments. Clearly, the issue is far from decided.

2.7 Current Research

Recently, several researchers (Paslawski, 1984; Stephanovic and Vries-Baayens, 1984; Gilmartin, 1987) have lamented the fact that data classification research is still inconclusive. Stephanovic and Vries-Baayen (p. 52) stated that:

"There is still no agreement on the criteria for an optimal classification system."

However, Coulson (1987) asserted that the variance option of the JENKS program (Jenks 1977) achieved the objective of minimizing variance within classes while maximizing variance between classes. This statement is tested in this examination, as well as whether or not the objective produces the "best" choropleth maps.

Presently, map communication research continues to support the consideration of the visual characteristics of choropleth maps as well as their accuracy. As Eastman (1987, p. 108) points out:

"The map is the meeting ground for two organisational processes - those of the cartographer and those of the reader. It is therefore essential that we understand the latter if we are ever to take full advantage of the creative medium we employ."

Research on data classification is not as popular today as it was throughout the 1960s and 1970s. In all probability, this is due to the lack of generally-applicable conclusions regarding classification methods in spite of more than twenty years of research.

2.8 Summary

Although geographers (Monmonier, 1974; Muller 1976a; Chang, 1979) conducting research on measures of visual characteristics of choropleth maps have performed diverse studies with different results and conclusions, all of them agree on two major points: further research into map perception is desirable, and development of a classification method that takes visual characteristics into consideration is necessary to optimize class interval selection for choropleth mapping.

Monmonier (1974, p.168) emphasized the importance of considering visual characteristics of maps in the classification of data in his statement:

> "Further experimental research with map readers is needed in order to assess the relative importance of these metrics in visual map pattern recognition. On the basis of this type of research, class-interval selection can be made more responsive to user needs."

Further to Monmonier's (1974) statement, Robinson and Petchenik (1975, p.120) stated the importance of taking the needs of the map user into consideration when creating a map:

> "A good map is not simply concerned with transmitting information but with enhancing the map user's understanding of reality".

In (1979, p. 22) Chang restated this belief:

> "There is an urgent need for a computer program which can incorporate objective functions of both map accuracy and map complexity".

The dilemma facing the cartographer is how to determine the point where both mathematical accuracy and information flow can be maximized.

III. METHODOLOGY

3.1 Introduction

The methodology for proceeding with this study is outlined and justified in this chapter. Elements of the methodology include the selection of data sets to represent a cross-section of data distributions, and the selection and application of measures of both accuracy and complexity in order to ensure a comprehensive evaluation.

In addition, the options and restrictions of each of the six computer programs selected for this study are described.

3.2 Selection of the Data Sets

In order to ensure rigorous testing of the six computer programs, data sets are selected according to two factors: size, and distribution. Three different sizes of data sets were chosen for this study. The "small" data sets are from the St. John's census metropolitan area, which has thirty-three census tracts and therefore thirty-three observations. The one-hundred-and-forty-seven census tracts in the Edmonton census metropolitan area provide "medium" data sets. The "large" data sets are from the Toronto census metropolitan area, which has six-hundred-eight census tracts.

Of the twelve cities for which data were available, St. John's and Toronto were chosen as the smallest and largest respectively. Edmonton was chosen for the medium-sized data sets because it is representative of the size of all the medium-sized cities, and does not contain physiographic features which would inhibit the visual inspection of the maps. It is noted that the selection of these particular data sets may affect the success of the classing techniques, as cities with different spatial patterns of census tracts may affect the visual complexity of the maps produced by each program.

The second factor for selecting data sets is data distribution. Within each of the three areas producing different sizes of data sets, data sets are selected which display a cross-section of common data distributions: bell-shaped, right-skewed, left-skewed, J-shaped, L-shaped, rectangular, and bimodal.

Although frequency histograms are often used to display data sets because they are easy to construct and comprehend, they are not suitable for this study. It is impractical to display each value of the Toronto data sets (608 observations) on a frequency histogram, as the histogram would be extremely difficult to interpret. If the data set is generalized and a frequency histogram constructed such that each bar represents for example ten or twenty values, the shape of the resulting histogram is entirely dependent upon the number of values represented by each bar (Figure 2.8). In addition, the interpretation of frequency histograms can be highly subjective depending upon the viewer's background knowledge, interpretive skills, and preconceptions regarding the distribution of the data set.

A theoretically sound approach to determining distributions of data sets is chosen -- a method which displays all observations of a data set rather than a generalized version, yet can display large data sets in a manner that is concise, objective and relatively easy to comprehend. This method is termed the empirical cumulative distribution function, and has been described by Wilk and Gnanadesikan (1968, p. 1) as playing:

> "...a key role in the statistical treatment of one-dimensional samples, being of relevance for summarization and palatable description as well as for exposure and inference".

In order to calculate points for empirical cumulative distribution functions, one-half is subtracted from each data observation in the ordered data set, and the result divided by the number of observations in the data set. The results of these calculations are used as

Y-values, which paired with X-values that are simply the original data values, provide the points for the plot (Appendix 5). The resulting graph is one in which the curve is always increasing, thus eliminating the artificial peaks and troughs often created in frequency histograms.

A computer program for calculating pairs of values for empirical cumulative distribution frequencies was written, and graphs of one-hundred data sets generated on the Versatec plotter. Of these one-hundred data sets, seven displaying right- and left-skewed, L- and J-shaped, rectangular, bell-shaped and bimodal distributions are chosen for each of the three geographic areas. In addition, one Toronto data set which has extreme outliers is chosen as a further test of the abilities of the programs. Therefore, the number of data sets used in the study is twenty-two, i.e. three census metropolitan areas multiplied by seven data distributions, plus one data set with outliers.

3.3 The Computer Programs

The following discussion of the six computer programs used in this study makes reference to many of the classification methods and accuracy measures explained in detail in Chapter II. These methods and measures are not dealt with in any detail in this chapter, but their usage in the various programs is delineated. For each program, the classification and evaluation techniques, and other options and restrictions are discussed. Sample output from each program can be found in Appendix 4.

3.3.1 CHANG

3.3.1.1 Classification and Evaluation Techniques

The CHANG program, based on the work of Jenks and Coulson (1963), offers a choice of seven of the "traditional" classification methods: equal step, equal frequency, arithmetic, geometric, reciprocal, mean-and-standard-deviation, and the nested-mean methods. It should be noted that CHANG uses a different mean-and-standard-deviation method than the conventional method described in Chapter II. The CHANG method is based on the assumption that the data set is normally distributed; the area under the normal curve is divided equally into the given number of classes, and z-values are computed. Class limits are calculated according to the formula [mean + z (standard deviation)] (Chang, 1974). For each calculation of class intervals in CHANG, the D-value is printed out for each individual class and for the total of all classes so that the cartographer can compare the D-values derived from different classing techniques (Appendix 4a).

3.3.1.2 Other Options and Restrictions

Initially, the maximum number of observations in a data set processed by the CHANG program was five hundred. This was increased to twelve thousand observations by the Geocartographics Subdivision for one specific application. The maximum number of classes that can be derived is ten, which does not represent a limitation since more than eight classes are rarely used.

3.3.2 JENKS

3.3.2.1 Classification and Evaluation Techniques

The JENKS program, based on Fisher's (1958) principle of "grouping for maximum homogeneity", classifies data using one of two iterative techniques: by minimizing the absolute deviations about class medians, or by minimizing the squared deviations about class means. Although the mean and median are both measures of central tendency, they operate in different manners. The mean, commonly referred to as the "average" of a data set, can be likened to the "center of gravity" of that data set. It is calculated by dividing the sum of the data values by the number of observations. The median, on the other hand, refers to the middle value of the ordered data set; half of the values in the data set are lower than the median, and half of the values are higher. In general, the median is less influenced by outliers in a data set than the mean, and perhaps it is for this reason that Jenks chose the median for his first classification method, even though it is more common to use the mean for this type of calculation.

Whether the user chooses the absolute deviations method or the squared deviations method (or both), JENKS outputs only the optimal solution for each in the form of class intervals, unlike the MONMONIER and YOUNGMANN programs which both show some of the calculations leading up to the optimal solution. The accuracy measure(s) (either the sum of absolute deviations and/or the variance) are printed as well as the frequencies and median or mean for each class interval (Appendix 4b).

3.3.2.2 Other Options and Restrictions

The JENKS program, which can handle up to ten classes, was expanded for this study from its original limit of two-hundred data values to six-hundred-fifty data values. In addition, a "user-defined" option (for calculating variance only) was added to the program, thus enabling the user to implement his/her own breakpoints and determine the resulting variance.

3.3.3 MOELLERING

3.3.3.1 Classification and Evaluation Techniques

The MOELLERING program utilizes information statistics in an iterative framework to define class breakpoints. It offers only one classing method, for which it displays the upper limits of classes, frequencies, within- and between-class statistics, and the percent of accuracy according to the equality measure derived from the classification procedure (Appendix 4c).

3.3.3.2 Other Options and Restrictions

The maximum number of observations in a data set that can be processed by MOELLERING is one-thousand, expanded from two-hundred-fifty for this study. The maximum number of classes that can be calculated is fifteen, although for each data set MOELLERING automatically determines the minimum and maximum "desirable" number of classes, and prints solutions for these numbers only, regardless of the number(s) of classes that the user requests (Appendix 2).

The MOELLERING program has one major restriction not shared with any of the other programs included in the study: data sets with zeros and/or negative values cannot be processed because of the logarithms employed in the calculation of the equality measure.

3.3.4 MONMONIER

3.3.4.1 Classification and Evaluation Techniques

The MONMONIER program, based on the work of Jenks and Caspall (1971), offers two choices of iterative classing techniques. The first is based on minimization of the absolute deviations about class means. For each set of class intervals, the applicable accuracy measure (TAI absolute or squared), class frequencies, class breakpoints and class intervals are displayed. Only the optimal solution is printed, but a history of the number of iterations and improvements is presented at the top of the optimal solution as "jiggles" and "improves" respectively. Output includes the mean, sum of deviations, the low and high values for the total data set and for each class, and frequencies. Breakpoints are derived as the mid-point between the upper and lower class limits (Appendix 4d).

3.3.4.2 Other Options and Restrictions

MONMONIER is limited to one-thousand data observations (expanded from five-hundred originally) and fifteen classes; it was also modified to enable the user to run both the absolute and squared deviations methods in the same batch submission. This program gives the user several options for fine-tuning the breakpoints themselves. Aside from the choice of absolute or squared deviations, the user may opt to have "rounded" breakpoints, thus creating values which will likely be more meaningful and memorable to the user than values with numerous decimal places (Monmonier, 1982). The base for rounding is also user-specified, so that for data sets with small ranges, a smaller rounding factor can be chosen than for data sets with large ranges.

Another option that MONMONIER offers is "gapped" intervals, where the minimum and maximum of each class are actual data values rather than numbers representing the beginning and ending values of classes which may or may not be actual data values. This does not affect the class membership, but merely indicates specifically the range of actual data values in each class. It should be noted that this option is only available when the rounding option has already been selected. Like CHANG and JENKS, this program has an option for user-defined or "mandated" breakpoints.

3.3.5 YOUNGMANN

3.3.5.1 Classification and Evaluation Techniques

Based on the work of Jenks and Caspall (1971), the classification method used in this program is almost the same as the first option in MONMONIER, i.e. it attempts to minimize the absolute deviations about class means. Unlike MONMONIER, for each set of class intervals, the history of iterations is shown in a four-step procedure: the initial configuration, reiteration of the original configuration, forcing, and reiteration of the forcing. The Tabular Accuracy Index does not necessarily increase from step one to step four, and hence the user must peruse all stages of the calculations in order to determine the set of intervals with the highest Tabular Accuracy Index. The results are presented in the form of class intervals, with frequencies and the mean for each class displayed (Appendix 4e).

3.3.5.2 Other Options and Limitations

The YOUNGMANN program was expanded by the Geocartographics Subdivision in order to accept up to one-thousand observations from the original maximum of five-hundred. YOUNGMANN offers a unique option that takes the sizes of the areal units into consideration when calculating the Tabular Accuracy Index, thereby attempting to compensate for the disproportionate visual influence that large areas hold in comparison to smaller areas. The methodology for this option is based on the Overview Accuracy Index presented by Jenks and Caspall (1971). Of course, the use of this option necessitates that areas of the units be known and entered into the computer. The resulting output displays all of the iterations resulting from the attempt to maximize the Tabular Accuracy Index as well as the four-step procedure performed in trying to maximize the Overview Accuracy Index. This option is likely more valuable in cases where there is great disparity amongst the sizes of areal units, and conversely will not have as much effect on classification when the areas of the units are more uniform.

3.3.6 SAS

3.3.6.1 Classification and Evaluation Techniques

The SAS FASTCLUS procedure performs a disjoint cluster analysis on the basis of "nearest centroid sorting". Initial cluster seeds are chosen, and then the data values are grouped to the closest seed. This process is reiterated until the most homogeneous classes are identified by the application of the sum of squared distances from cluster means.

Although the FASTCLUS procedure measures accuracy in a similar manner to the JENKS program, it differs from the other iterative techniques in this study in that the iterations do not proceed on the basis of maximizing an accuracy statistic. Instead, once the values have been assigned to the nearest seed, the distances of each cluster member from the mean are summed and squared, and the initial seeds are recalculated. This process reiterates until the recalculation of the seeds does not produce new seeds.

There is no single accuracy measure emphasized by the FASTCLUS procedure. Several statistics associated with accuracy are outputted, such as the between-cluster variance (R-squared), the ratio of between- to within-cluster variance (variance ratio), and the root-mean-square distance between observations in the cluster, among others (SAS User's Guide, 1982). Cluster means, standard deviations, and other descriptions of the data are printed optionally (Appendix 4f).

3.3.6.2 Other Options and Restrictions

The SAS program can handle data sets with up to one-hundred-thousand observations, and up to two-hundred-fifty classes. These limits differ greatly from the other programs because the FASTCLUS program was not designed for choropleth mapping; it is more strictly related to cluster analyses where it is much more common to be working with data sets of such magnitude.

The FASTCLUS procedure is intended as a classification method for large data sets. It has been noted (SAS User's Guide, 1982) that for data sets with fewer than one-hundred observations, the results may be sensitive to the order of the values in the data set.

Options in the FASTCLUS procedure that may be of particular interest for choropleth mapping include the RADIUS option, which allows the user to input a minimum distance between initial cluster seeds, and the DRIFT option, which replaces the cluster seeds with the means of current cluster members. The effect of both of these options would be to make the algorithm less sensitive to the ordering of the values in the data set by increasing the distances amongst the cluster seeds.

3.4 Testing Procedures and Measures

3.4.1 Introduction

The programs are evaluated both in terms of mathematical accuracy and in terms of the visual characteristics of the maps they produce.

For the evaluation of mathematical accuracy, four different accuracy statistics are chosen and applied to the classed data sets produced by each program. The decision to use four measures rather than just one was made because of the different theoretical bases of the measures used in the programs themselves and because of the fact that the literature does not identify one measure as being superior to the others.

The evaluation of the visual characteristics of the maps produced by each program is conducted using one statistical measure, the fragmentation index, and a visual inspection of the maps. As noted in section 2.4 the majority of statistical measures used for evaluating maps are based on the complexity of the map, and are correlated with one another. Therefore, the fragmentation index is considered to be representative of all of the complexity measures.

Once both the mathematical accuracy and visual characteristics are evaluated separately, the results are combined. In this manner, each program could be evaluated on both of the dimensions important for the optimal display of spatial phenomena in choropleth map form.

3.4.2 Mathematical Accuracy

The first step in the testing procedure is to run each data set through each computer program and tabulate class breakpoints or intervals, class frequencies and accuracy measures for a five-class map. The limit of five classes was chosen because:

a) it is approximately the midpoint in the commonly accepted range of three to eight classes;

b) it is compatible with several other documented studies (Youngmann, 1972; Robinson et al., 1984; Smith 1986);

c) an informal survey of the choropleth maps produced by the Geocartographics Subdivision revealed that five was the most commonly used number of classes.

It should be noted, however, that it is possible that the success of the programs relative to one another may improve or deteriorate if a smaller or greater number of classes is used. Within each program, it is expected that increasing the number of classes would improve the accuracy of the classed data due to the decreased generalization of the data (Jenks and Caspall, 1971; Youngmann, 1972; Muller, 1975b).

The result of the first set of computer tests is that each program produces a classed data set for each of the twenty-two raw data sets. All of the classing techniques offered in the programs are used to create classed data sets. However, for the CHANG program, only the "best" technique (as measured by the D-value) out of the six offered is chosen for comparison with the other programs, since the CHANG program has already indicated that the other five are mathematically inferior. The seventh option in the CHANG program, i.e. the nested-means method, cannot be tested since it cannot be used to produce five classes (see section 2.2).

Both the JENKS and MONMONIER programs offer a choice of two classing techniques. In both cases, both options are tested and retained for the inter-program comparison since they have independent accuracy measures and the programs do not give an indication of one technique being superior to the other.

Therefore, eight sets of classed data are compared for each of the raw data sets, i.e. one classed data set each from CHANG, MOELLERING, YOUNGMANN, and SAS, and two each from JENKS and MONMONIER. The accuracy of each of these classed data sets is then calculated using four different measures: the D-value (Jenks and Coulson, 1963), the TAI based on absolute deviations (Jenks and Caspall, 1971), the TAI based on squared deviations (Jenks, 1977), and the Variance (Jenks, 1977).

The four measures are applied using the "user-defined class breakpoints" options in the CHANG, JENKS, and MONMONIER programs. For example, the right-skewed St. John's data set is fed into each of the computer programs, producing eight sets of classed data. From the first set of computer tests, the D-value is known for the classed data produced by CHANG. The breakpoints of the other seven classed data sets are fed into the CHANG "user-defined class breakpoints" option, and the D-value calculated for each. Thus, the mathematical accuracy of the eight classed data sets can be compared using the D-value. In a similar procedure, the "user-defined class breakpoints" option in JENKS is used to calculate the Variance measure for all classed data sets, and MONMONIER's "user-defined" option calculated the TAI based on absolute deviations and TAI based on squared deviations.

Once the results of the application of the four accuracy measures are tabulated for each data set, one is able to make comparisons based on, for example, the individual accuracy measures, accuracy compared to size of data set, and accuracy compared to data distribution.

3.4.3 Additional Testing with Accuracy Measures

Both the MONMONIER and YOUNGMANN programs offer options which attempt to increase the interpretability and therefore information content of maps produced by their classing algorithms. These options are tested on a few data sets in order to determine the effect of map enhancement on map accuracy.

The option tested in the MONMONIER program rounds off class breaks so that they are more easily comprehended by the map user. Because the rounded class breaks may change the class memberships, a decrease in map accuracy is anticipated. The YOUNGMANN option that is tested is the "area" option, where weighting factors are utilized to compensate for the differences in size amongst areal units. This option is evaluated by comparing maps produced with the regular YOUNGMANN program with maps of the same data produced using the "area" option.

3.4.4 Visual Complexity

Once the mathematical accuracy of the classed data sets has been evaluated, each of the classed data sets is mapped using GIMMS software and a Versatec plotter. The appearance of the maps is evaluated in two ways: through the application of the fragmentation index, a measure of complexity formulated by Monmonier (1974), and through visual inspection by the author.

The fragmentation index measures map complexity as the number of map "regions", or contiguous areas of equal value, divided by the total number of areas. Because map complexity is considered to be lower when mapped patterns are aggregated rather than being in a "quilt-like" pattern, maps with relatively few "regions" compared to their number of areas will be less complex.

The fragmentation index varies between zero and one, with a value closer to zero describing a map of extremely low complexity (very few regions) and a value tending towards one depicting a highly complex map with many regions. The value one is subtracted from both the numerator and the denominator of the fragmentation index in order to ensure that the range of the index is always between zero and one.

The fragmentation index was calculated using ARC/Info. The results of all of the classing programs were used as input together with a file describing the adjacency of the polygons (areas on the map) associated with the data. ARC/Info contains a "dissolve" program which simply erases boundary lines between areas associated with equal class value, thus creating the "regions" which are the numerator in the equation of the fragmentation index. The denominator is the total number of areas on the map before the dissolve program is run, that is, the number of values in the data set.

The purpose of the visual inspection is to corroborate the results of the calculation of the fragmentation index.

3.5 Summary and Objectives of Study

The methodology of this study is designed to provide a thorough evaluation of the choropleth map data classing problem. Specifically, the objectives of the study are to:

1) evaluate the mathematical accuracy of classed data sets produced by various classing techniques;
2) evaluate the visual complexity of maps produced from the classed data sets;
3) combine the evaluation of mathematical accuracy and visual complexity for each computer program;
4) test the effect of data distribution on classing techniques;
5) test the effect of data set size on classing accuracy;

6) test additional options in the computer programs;
7) document the operational advantages and disadvantages of each of the programs.

The manner of selection of the data sets, and the design of the methodology to consider both the mathematical accuracy of classed data and the visual complexity of the maps they produce are oriented towards achieving the objectives of the study.

IV. RESULTS

4.1 Introduction

In this chapter, the results of the testing are presented according to the objectives stated in section 3.5. In order to carry out the evaluation, it was necessary to analyze the results in several formats.

4.2 Analysis of the ECDFs

The graphs (see Appendix 5) produced by the ECDF algorithm were interpreted first by the author, then certified by others. Although there were ninety-nine data sets to choose from (thirty-three for each area), four of the desired distributions could not be found: left-skewed and bell-shaped for St. John's, bimodal for Edmonton, and bimodal for Toronto. This necessitated the creation of four artificial data sets, which was accomplished by inputting data into the ECDF program, producing graphs, and adjusting the inputted data until the desired graph shapes were generated.

The values within each of the "real" data sets obtained from the 1981 Census of Canada are spatial data, in that they are interrelated in planar space. As discussed in section 1.1.2, maps are a powerful tool for the display of spatial data, and therefore maps are one of the elements under examination in this study. However, because the artificial data lack a spatial dimension, mapping is not a suitable display medium for these data. Therefore, maps were not produced for the artificial data sets, and map complexity was not evaluated.

4.3 Mathematical Accuracy

4.3.1 Presentation of the Results

Once the accuracy measures for each program for each data set were tabulated (Table 4.1), the results were ranked from "best" to "worst" by program and data set for each measure (Table 4.2). The "best" value is represented by the lowest value in the cases of the D-value and the Variance, and by the highest value for the TAI (absolute deviations) and TAI (squared deviations). Values calculated by the accuracy measures were assigned ranks from 1 to 7 as "best" to "worst", with tied values being assigned the same rank. In addition, on the same table, the mean ranks for each accuracy measure and the overall mean ranks were calculated.

The next step in the evaluation required compiling the number of times that each program ranked first and last for each of the data sets by each accuracy measure, and for all measures combined (Table 4.3). Another table

Table 4.1 St. John's: Tabulated Accuracy Measures

D-VALUE	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	0.09399	0.05469	0.11600	0.00003	0.00569	0.19804	0.02641
JENKS (abs.)	0.27460	1.04366	0.22951	0.00006	0.01417	0.27017	0.04219
JENKS (var.)	0.31500	0.80086	0.22951	0.00006	0.01002	0.26518	0.04367
MOEL.	0.08370	N/A	0.10412	0.00007	0.01921	0.27465	0.04445
MON. (abs.)	0.27460	1.04216	0.22951	0.00007	0.01921	0.26873	0.04219
MON. (sq.)	0.08388	1.03977	0.22951	0.00006	0.01002	0.26518	0.05008
YOUNG.	0.27460	0.80223	0.22951	0.00007	0.01921	0.26873	0.04219
SAS	0.35867	1.00460	0.20622	0.00008	0.01225	0.26251	0.05283

TAI (absolute deviations)	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	0.73670	0.65020	0.71964	0.75262	0.68756	0.63052	0.80265
JENKS (abs.)	0.77999	0.69713	0.78420	0.81100	0.82895	0.69869	0.84522
JENKS (var.)	0.77625	0.69862	0.78420	0.81100	0.81763	0.69827	0.84502
MOEL.	0.67015	N/A	0.73231	0.80299	0.83163	0.62775	0.82223
MON. (abs.)	0.77999	0.69850	0.78420	0.81229	0.83163	0.69977	0.84522
MON. (sq.)	0.75630	0.69850	0.78420	0.81100	0.81763	0.69827	0.81777
YOUNG.	0.77999	0.70047	0.78420	0.81229	0.83163	0.69977	0.84522
SAS	0.75697	0.60923	0.75126	0.76886	0.81403	0.64516	0.78500

TAI (squared deviations)	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	0.93582	0.84098	0.89276	0.94242	0.90092	0.84690	0.95112
JENKS (abs.)	0.94443	0.93005	0.95693	0.96485	0.96393	0.93036	0.96422
JENKS (var.)	0.94736	0.94246	0.95693	0.96485	0.96448	0.93223	0.96464
MOEL.	0.88509	N/A	0.90124	0.96090	0.96326	0.89575	0.95874
MON. (abs.)	0.94443	0.93114	0.95693	0.96422	0.96326	0.93088	0.96422
MON. (sq.)	0.94692	0.93380	0.95693	0.96485	0.96448	0.93223	0.96016
YOUNG.	0.94443	0.93935	0.95693	0.96422	0.96326	0.93088	0.96422
SAS	0.94315	0.89995	0.94076	0.94844	0.96428	0.91149	0.92817

VARIANCE	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	114.22	1553.69	44.15	1.53	636.90	1414.75	9.49
JENKS (abs.)	98.90	683.43	17.73	0.91	231.84	643.57	6.95
JENKS (var.)	93.67	562.20	17.73	0.95	228.30	626.28	6.87
MOEL.	204.50	N/A	40.66	1.02	236.10	963.33	8.01
MON. (abs.)	98.90	672.80	17.73	0.93	236.14	638.70	6.95
MON. (sq.)	94.46	646.80	17.73	0.91	228.32	626.28	7.73
YOUNG.	98.90	592.56	17.73	0.93	236.14	638.70	6.95
SAS	101.17	977.50	24.39	1.37	229.62	817.92	13.95

Table 4.1 Edmonton: Tabulated Accuracy Measures

D-VALUE	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	0.14587	0.00759	N/A	0.02285	0.05107	0.06843	0.32484
JENKS (abs.)	0.76434	0.00982	4.00625	0.04646	0.12492	0.82474	0.79407
JENKS (var.)	0.18106	0.00769	1.23630	0.02259	0.18452	0.62070	0.71499
MOEL.	0.21801	0.00936	N/A	0.04298	0.21151	0.48709	0.99758
MON. (abs.)	0.22216	0.00982	1.68488	0.02980	0.21429	0.74167	0.31105
MON. (sq.)	0.18106	0.00769	1.20734	0.02259	0.21332	0.62070	0.71499
YOUNG.	0.22216	0.00982	1.68488	0.02980	0.11505	0.50392	0.83806
SAS	0.54999	0.00919	1.23630	0.03655	0.81186	0.79431	0.71499

TAI (absolute deviations)	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	0.62700	0.64607	N/A	0.48924	0.76206	0.69313	0.85802
JENKS (abs.)	0.73449	0.75132	0.76167	0.71017	0.77650	0.73424	0.86321
JENKS (var.)	0.72758	0.73603	0.73723	0.70388	0.77360	0.70712	0.84556
MOEL.	0.72873	0.73095	0.62285	0.70010	0.76952	0.72050	0.81231
MON. (abs.)	0.73292	0.75132	0.76453	0.71014	0.76499	0.74140	0.85530
MON. (sq.)	0.72758	0.73603	0.73723	0.70388	0.76296	0.70712	0.84556
YOUNG.	0.73292	0.75132	0.76453	0.71014	0.77695	0.73570	0.86718
SAS	0.68450	0.58641	0.73723	0.59162	0.56503	0.56080	0.84556

TAI (squared deviations)	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	0.87357	0.88804	N/A	0.63057	0.94920	0.91283	0.96118
JENKS (abs.)	0.91757	0.91953	0.95194	0.88114	0.95065	0.90052	0.96192
JENKS (var.)	0.93183	0.92202	0.97228	0.91845	0.95264	0.92137	0.96895
MOEL.	0.91106	0.91514	0.74345	0.88990	0.94690	0.91457	0.94091
MON. (abs.)	0.92679	0.91953	0.95371	0.91337	0.94936	0.91594	0.96158
MON. (sq.)	0.93183	0.92202	0.97228	0.91845	0.94984	0.92137	0.96895
YOUNG.	0.92679	0.91953	0.95371	0.91337	0.95022	0.90967	0.95269
SAS	0.90046	0.84580	0.97228	0.86985	0.82318	0.82769	0.96895

VARIANCE	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectangular	Bell-shaped	Bimodal
CHANG	219.93	595.50	N/A	1648.22	3940.21	589.15	49.51
JENKS (abs.)	143.39	428.03	9.77	900.72	3827.76	672.31	48.56
JENKS (var.)	118.57	414.64	5.64	618.19	3673.50	531.46	39.60
MOEL.	154.71	451.53	52.16	834.34	4118.21	577.38	75.36
MON. (abs.)	127.35	428.03	9.41	656.61	3927.44	568.11	48.99
MON. (sq.)	118.57	414.76	5.64	618.11	3890.30	531.46	39.60
YOUNG.	127.35	428.03	9.41	656.61	3860.53	610.49	60.33
SAS	173.16	820.15	5.64	986.28	14437.61	1164.56	39.60

Table 4.1 Toronto: Tabulated Accuracy Measures

D-VALUE	Right-skewed	Left-skewed	L-shaped	J-shaped	Rectan-gular	Bell-shaped	Bimodal	Bell-shaped with outliers
CHANG	0.49499	0.30621	0.53507	0.04043	0.25755	0.01127	0.31649	0.91781
JENKS (abs.)	0.76138	0.26193	0.23323	0.08897	0.31001	0.04553	0.66031	0.91949
JENKS (var.)	0.58882	0.29136	0.79265	0.06284	0.29974	0.03250	0.67888	0.94180
MOEL.	0.72944	0.27812	0.73112	0.09325	0.17313	0.04452	0.99399	N/A
MON. (abs.)	0.60205	0.27797	0.27573	0.08116	0.30960	0.04085	0.31918	0.91453
MON. (sq.)	0.51565	0.29136	1.02717	0.05361	0.33985	0.03250	0.57548	0.94180
YOUNG.	0.73672	0.27797	0.20206	0.08116	0.30960	0.03960	0.71313	0.91947
SAS	0.19667	1.13324	0.48253	0.07666	0.84756	0.01419	0.54331	0.90965
TAI (absolute deviations)								
CHANG	0.47659	0.60804	0.78711	0.36071	0.75535	0.62283	0.85220	0.67653
JENKS (abs.)	0.73122	0.74306	0.80333	0.75552	0.77250	0.89060	0.85786	0.67734
JENKS (var.)	0.71656	0.74241	0.63962	0.74867	0.77213	0.69998	0.82229	0.64636
MOEL.	0.71347	0.72357	0.73429	0.74171	0.75793	0.69948	0.81112	N/A
MON. (abs.)	0.72291	0.74322	0.80455	0.75568	0.77249	0.71123	0.84793	0.68123
MON. (sq.)	0.67014	0.74241	0.77076	0.74744	0.77233	0.69998	0.83830	0.64636
YOUNG.	0.72220	0.74322	0.79782	0.75568	0.77249	0.71098	0.86375	0.66981
SAS	0.50950	0.48276	0.75649	0.50295	0.57659	0.61178	0.81979	0.50885
TAI (squared deviations)								
CHANG	0.71888	0.85818	0.95724	0.60088	0.92989	0.88217	0.95994	0.77331
JENKS (abs.)	0.86928	0.93925	0.94922	0.93936	0.95183	0.89060	0.96075	0.78150
JENKS (var.)	0.89968	0.93925	0.90168	0.94210	0.95194	0.90938	0.96753	0.89003
MOEL.	0.86873	0.92565	0.85989	0.92592	0.93986	0.88948	0.94082	N/A
MON. (abs.)	0.89555	0.93848	0.94321	0.94041	0.95177	0.90182	0.95901	0.78948
MON. (sq.)	0.89459	0.93925	0.95729	0.94204	0.95189	0.90938	0.96690	0.89003
YOUNG.	0.87232	0.93876	0.94736	0.94041	0.95177	0.90391	0.95192	0.77253
SAS	0.80011	0.88684	0.93368	0.77357	0.82178	0.86785	0.96583	0.81182
VARIANCE								
CHANG	2629628.80	21420.93	2666.87	54238.21	27173.86	3685.84	209.98	7989.02
JENKS (abs.)	1222779.00	9215.96	3166.62	8241.39	18670.81	3064.14	205.71	7706.81
JENKS (var.)	938470.00	9176.36	2454.06	7869.75	18627.33	2538.70	170.22	3879.77
MOEL.	1227925.00	11230.00	8738.00	10066.97	23311.95	3095.86	310.19	N/A
MON. (abs.)	977116.00	9292.28	3541.70	8097.68	18695.01	2750.05	214.84	7425.59
MON. (sq.)	986022.00	9175.83	2663.84	7878.03	18645.95	2538.13	173.47	3880.02
YOUNG.	1194330.00	9249.70	3282.81	8097.68	18695.01	2691.80	252.02	8023.75
SAS	1869670.90	37352.00	4136.13	30770.46	69079.01	3702.95	179.10	6639.19

Table 4.2 Program Rankings for Each Accuracy Measure

	ST. JOHN'S							EDMONTON							TORONTO								
D-VALUE	R.S.	L.S.	L.	J.	Rect.	Bell.	Bi.	R.S.	L.S.	L.	J.	Rect.	Bell.	Bi.	R.S.	L.S.	L.	J.	Rect.	Bell.	Bi.	Out.	Mean Rank
CHANG	3	1	2	1	1	1	1	1	1	N/A	3	1	1	2	2	7	5	1	2	1	1	3	1.86
JENKS (abs.)	5	7	6	3	5	7	3	8	7	6	8	3	8	6	8	1	2	7	6	8	5	5	5.64
JENKS (var.)	7	2	6	3	2.5	3.5	5	2.5	2.5	2.5	1.5	4	4.5	4	4	5.5	7	3	3	3.5	6	6.5	4.04
MOEL.	1	N/A	1	6	7	8	6	4	5	N/A	7	5	2	8	6	4	6	8	1	7	8	N/A	5.27
MON. (abs)	5	6	6	6	7	5.5	3	5.5	7	4.5	4.5	7	6	1	5	2.5	3	5.5	4.5	6	2	2	4.75
MON. (sq.)	2	5	6	3	2.5	3.5	7	2.5	2.5	1	1.5	6	4.5	4	3	5.5	8	2	7	3.5	4	6.5	4.11
YOUNG.	5	3	6	6	7	5.5	3	5.5	7	4.5	4.5	2	3	7	7	2.5	1	5.5	4.5	5	7	4	4.79
SAS	8	4	3	8	4	2	8	7	4	2.5	6	8	7	4	1	8	4	4	8	2	3	1	4.84
TAI (absolute deviations)																							
CHANG	7	6	8	8	8	7	7	8	7	N/A	8	7	7	3	8	7	4	8	7	7	3	3	6.52
JENKS (abs.)	2	5	3	4	4	3	2	1	2	6	1	2	3	2	1	3	2	3	1	1	2	2	2.50
JENKS (var.)	4	2	3	4	5.5	4.5	4	5.5	4.5	4	4.5	3	5.5	6	4	4.5	8	4	5	4.5	6	5.5	4.61
MOEL.	8	N/A	7	6	2	8	5	4	6	7	6	4	4	8	5	6	7	6	6	6	8	N/A	5.95
MON. (abs.)	2	3.5	3	1.5	2	1.5	2	2.5	2	1.5	2.5	5	1	4	2	1.5	1	1.5	2.5	2	4	1	2.25
MON. (sq.)	6	3.5	3	4	5.5	4.5	6	5.5	4.5	4	4.5	6	5.5	6	6	4.5	5	5	4	4.5	5	5.5	4.90
YOUNG.	2	1	3	1.5	2	1.5	2	2.5	2	1.5	2.5	1	2	1	3	1.5	3	1.5	2.5	3	1	4	2.04
SAS	5	7	6	7	7	6	8	7	8	4	7	8	8	6	7	8	6	7	8	8	7	7	6.90
TAI (squared deviations)																							
CHANG	6	7	8	8	8	8	7	8	7	N/A	8	6	5	6	8	8	2	8	7	7	5	6	6.81
JENKS (abs.)	4	5	3	2	4	5	3	5	4	6	6	2	7	4	5	2	3	5	3	5	4	5	4.18
JENKS (var.)	1	1	3	2	1.5	1.5	1	1.5	1.5	2	1.5	1	1.5	2	1	2	7	1	1	1.5	1	1.5	1.72
MOEL.	8	N/A	7	6	6	7	6	6	6	7	5	7	4	8	6	6	8	6	6	6	8	N/A	6.45
MON. (abs.)	4	4	3	4.5	6	3.5	3	3.5	4	4.5	3.5	5	3	5	2	5	5	3.5	4.5	4	6	4	4.11
MON. (sq.)	2	3	3	2	1.5	1.5	5	1.5	1.5	2	1.5	4	1.5	2	3	2	1	2	2	1.5	2	1.5	2.13
YOUNG.	4	2	3	4.5	6	3.5	3	3.5	4	4.5	3.5	3	6	7	4	4	4	3.5	4.5	3	7	7	4.29
SAS	7	6	6	7	3	6	8	7	8	2	7	8	8	2	7	7	6	7	8	8	3	3	6.09
VARIANCE																							
CHANG	7	7	8	8	8	8	7	8	7	N/A	8	6	5	6	8	7	3	7	7	7	5	6	6.81
JENKS (abs.)	4	5	3	1.5	4	5	3	5	4	7	6	2	7	4	5.5	3	4	5	3	5	4	5	4.31
JENKS (var.)	1	1	3	5	1	1.5	1	1.5	1	2	2	1	1.5	2	1	2	1	1	1	2	1	1	1.56
MOEL.	8	N/A	7	6	5	6	6	6	6	6	5	7	4	8	5.5	6	8	6	6	6	8	N/A	6.28
MON. (abs.)	4	4	3	3.5	6.5	3.5	3	3.5	4	4.5	3.5	5	3	5	2	5	6	3.5	4.5	4	6	4	4.04
MON. (sq.)	2	3	3	1.5	2	1.5	5	1.5	2	2	1	4	1.5	2	3	1	2	2	2	1	2	2	2.13
YOUNG.	4	2	3	3.5	6.5	3.5	3	3.5	4	4.5	3.5	3	6	7	4	4	5	3.5	4.5	3	7	7	4.31
SAS	6	6	6	7	3	7	8	7	8	2	7	8	8	2	7	8	7	8	8	8	3	3	6.22

Overall Mean Ranks	
CHANG	5.50
JENKS (abs.)	4.16
JENKS (var.)	2.98
MOELLERING	6.00
MON. (abs.)	3.78
MON. (sq.)	3.32
YOUNGMANN	3.86
SAS	6.01

Table 4.3 Number of Occurrences of "best" and "worst" Accuracy Measures for each Classing Program

NUMBER OF OCCURRENCES OF "BEST" ACCURACY MEASURE FOR EACH PROGRAM

PROGRAM	D-VALUE				TAI (ABS)				TAI (SQ)				VARIANCE				TOTAL (all measures combined)
	S	E	T	TOT	S	E	T	TOT	S	E	T	TOT	S	E	T	TOT	
CHANG	5	4	3	12	0	0	0	0	0	0	0	0	0	0	0	0	12
JENKS (abs.)	0	0	1	1	3	3	3	9	2	0	1	3	2	0	0	2	15
JENKS (var.)	0	1	0	1	1	0	0	1	7	7	7	21	6	6	6	18	41
MOELLERING	2	0	1	3	1	0	0	1	0	0	0	0	0	0	0	0	4
MONMONIER (abs.)	0	1	0	1	6	3	4	13	1	0	0	1	1	0	0	1	16
MONMONIER (sq.)	0	2	0	2	1	0	0	1	4	6	4	14	3	5	2	10	27
YOUNGMANN	0	0	1	1	7	4	3	14	1	0	0	1	1	0	0	1	17
SAS	0	0	2	2	0	0	0	0	0	2	0	2	0	2	0	2	6

NUMBER OF OCCURRENCES OF "WORST" ACCURACY MEASURE FOR EACH PROGRAM

PROGRAM	D-VALUE				TAI (ABS)				TAI (SQ)				VARIANCE				TOTAL (all measures combined)
	S	E	T	TOT	S	E	T	TOT	S	E	T	TOT	S	E	T	TOT	
CHANG	0	0	0	0	3	2	2	7	5	2	3	10	5	2	2	9	26
JENKS (abs.)	2	5	2	9	0	0	0	0	0	0	0	0	0	0	0	0	9
JENKS (var.)	1	0	1	2	0	0	1	1	0	0	0	0	0	0	0	0	3
MOELLERING	2	0	2	4	2	2	1	5	1	2	2	5	2	2	2	6	20
MONMONIER (abs.)	2	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3
MONMONIER (sq.)	1	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	3
YOUNGMANN	2	2	0	4	0	0	0	0	0	0	1	1	0	0	1	1	6
SAS	3	1	2	6	2	3	4	9	1	3	2	6	1	3	3	7	28

S - ST. JOHN'S DATA SET (33 OBSERVATIONS)

E - EDMONTON DATA SET (147 OBSERVATIONS)

T - TORONTO DATA SET (602 OBSERVATIONS)

TOT - TOTAL FOR ALL DATA SETS COMBINED

(Table 4.4) shows the number of times that the programs produced the same classed data sets.

4.3.2 Program Accuracy for Each Measure

The following account of the results refers to Table 4.4. It should be explained in advance of the discussion that when recording the "first" and "last" rankings for each of the accuracy measures, the occurrences of each may exceed the number of data sets, i.e. twenty-two, due to tied results amongst the programs. For example, for the TAI based on absolute deviations, the JENKS (absolute deviations) program ranked first nine times, the MONMONIER (absolute deviations) program ranked first thirteen times, and the YOUNGMANN program ranked first in fourteen cases. In some of these cases these three programs shared the first ranking, and therefore when the number of first rankings is summed it exceeds the number of data sets.

The CHANG program produced the best D-value in twelve out of the twenty-two data sets. All of the other programs rated much lower, ranging from having the best D-value three times (MOELLERING) to only one time (JENKS (absolute deviations and variance), MONMONIER (absolute deviations) and YOUNGMANN). JENKS (absolute deviations) and SAS ranked last nine and six times respectively, with the other programs falling into the two to four range

Table 4.4 Occurrences of identical classed data resulting from different classing programs

	St. John's							Edmonton							Toronto							
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8
CHANG																						
JENKS (abs.)	*		#	#			*		*													
JENKS (var.)			#	#	#	#		#	#	#	#		#	#		#						#
MOEL.					*																	
MON. (abs.)	*		#	*	*	*	*	*	*	*						*		*	*			
MON. (sq.)			#	#	#	#		#	#	#	#		#	#		#						#
YOUNG.	*		#	*	*	*	*	*	*	*						*		*	*			
SAS										#				#								

__Note:__ The # symbol refers to identical results, usually between MONMONIER (squared deviation , and JENKS (variance), while the * symbol refers to like pairs most often occurring between JENKS (absolute deviations), MONMONIER (absolute deviations), and YOUNGMANN. When # and * occur in the same data set, they refer to two different similar pairs of class intervals.

1 = right-skewed
2 = left-skewed
3 = L-shaped
4 = J-shaped
5 = rectangular
6 = bell-shaped
7 = bi-modal
8 = bell-shaped with outliers

except CHANG which did not rank last for any data set.

Theoretically, the D-value measures the similarity of values in classes by the breakpoints they create in the data set rather than by the ranges of the classes themselves. In this manner, the D-value measures within-group homogeneity, but does not measure between-group heterogeneity. Since the iterative techniques all attempt to find distinct "clusters" of data by measuring the within-class homogeneity of classes using the ranges of the classes, it is not surprising that collectively they did not produce as good D-values as the CHANG program.

For the TAI calculated by absolute deviations from class means, both the YOUNGMANN and MONMONIER (absolute deviations) programs produced good results, with fourteen and thirteen "first" rankings respectively. The JENKS (absolute deviations) program produced lower results (nine "first" rankings), and the other programs ranked first once (JENKS (variance), MOELLERING, MONMONIER (squared deviations)) or did not ever rank first (CHANG, SAS).

These results illustrate the success of the iterative programs designed with the TAI (absolute deviations) as their optimization measure, i.e. MONMONIER (absolute deviations) and YOUNGMANN. The JENKS (absolute deviations) technique bases its classes on clustering about class medians rather than class means, and this undoubtedly accounts for it's lower accuracy according to the TAI (abs.).

The SAS, CHANG and MOELLERING programs ranked last for the TAI (abs.) nine, seven and five times respectively. The JENKS (absolute deviations), YOUNGMANN, and MONMONIER (both absolute and squared deviations) programs did not rank last for any of th data sets, and the JENKS (variance) program ranked last only once. This finding illustrates that although the iterative techniques based on squared deviations from class means (JENKS (variance) and MONMONIER (squared deviations) did not produce the best TAIs (abs.) overall, they still performed well compared to CHANG, MOELLERING and SAS.

The JENKS program (variance option) obtained "first" rankings for the TAI based on squared deviations in twenty-one of the twenty-two data sets, compared with fourteen for the MONMONIER (squared deviations) option. All of the other programs ranked first from zero to three times. The CHANG, MOELLERING and SAS programs all ranked last a comparatively high number of times, with ten, five and six respectively.

Similar to the results of the TAI (abs.) testing, these results illustrate that the programs designed to maximize accuracy according to an optimization function are successful in doing so. In this case, the JENKS variance option performed even better than the MONMONIER squared deviations option that is designed with the TAI (sq.) as it's optimization measure.

Because the Variance measure is closely related to the TAI based on squared deviations, it was expected that the programs which performed well with one would also perform well with the other. The results corroborate this expectation, with JENKS (variance) and MONMONIER (squared deviations) both producing good results, i.e. ranking first eighteen and ten times respectively. The accuracy of the other programs as measured by the Variance mirrored their performance with the TAI based on squared deviations.

The remaining six techniques ranked first two or less times, but again CHANG, MOELLERING and SAS ranked last comparatively frequently. These three programs had nine, six, and seven "worst" Variance measures respectively, compared to the iterative techniques based on absolute deviations (JENKS (absolute deviations), MONMONIER (absolute deviations) and YOUNGMANN) which ranked last once or not at all.

The TAI (abs.), TAI (sq.) and Variance are all different from the D-value in that they measure the within-class homogeneity of values using the lower and upper class limits in their calculation rather than class ranges based on breakpoints which may or may not be actual data values. Because of this methodology, they are better measures for identifying clusters in the data, since between-group heterogeneity is naturally taken into account as well as within-group homogeneity.

4.3.3 Program Accuracy for All Measures Combined

Aggregation of all of the accuracy measures indicates that the JENKS program variance option is the most accurate, with a mean rank of 2.98. Out of the eighty-eight (twenty-two data sets multiplied by four accuracy measures) measures resulting from the testing, it ranked first forty-one times, and ranked last only three times. The MONMONIER (squared deviations) program rated comparably well, with a mean rank of 3.32, twenty-seven first rankings, and only three last rankings.

The next best overall results were achieved by the MONMONIER (absolute deviations) option, the YOUNGMANN program and the JENKS (absolute deviations) option with mean ranks of 3.78, 3.86, and 4.16 respectively. The number of first and last rankings for each are MONMONIER (absolute deviations) sixteen and three, YOUNGMANN seventeen and six, and JENKS (absolute deviations) fifteen and nine.

The CHANG program ranked sixth overall, with a mean rank of 5.50, twelve first rankings and twenty-six last rankings. In terms of mean rank, MOELLERING ranked slightly higher than SAS (6.00 and 6.01 respectively). MOELLERING had twenty last rankings and four first rankings, while SAS had twenty-eight last rankings and six first rankings.

The results of the testing indicate that all of the iterative techniques based on deviations from class means or medians (JENKS both options, MONMONIER both options, YOUNGMANN) are successful in creating the data classes that they are designed to produce, as determined by the high measures of accuracy accorded them. In addition, the application of the accuracy measures indicates that, regardless of the accuracy measure applied, these techniques provide consistently accurate classes in comparison to the other techniques. CHANG, MOELLERING and SAS all produce classes with a wide range of accuracy regardless of the accuracy measure applied.

As noted in section 4.3.1, the TAI (sq.) and Variance measures are very similar conceptually, and therefore the programs that are designed to produce optimal solutions based on these measures are at an advantage in this study. Specifically, the MONMONIER squared deviations option will produce classed data which is rated as being accurate not only by it's own accuracy measure (TAI (sq.)), but by the Variance measure as well provided that the algorithm functions as it is designed to. The same is true in reverse for the application of the TAI (sq.) and Variance to the JENKS variance option.

4.4 Visual Characteristics

4.4.1 The Fragmentation Index

The application of the fragmentation index also warranted the compilation of results to aid in interpretation. The fragmentation indices are shown in Table 4.5, and then the results for each program are ranked by data set (Table 4.6). Rankings are assigned from least complex to most complex as 1 to 7, and the mean ranks calculated.

Results of the application of the fragmentation index are recorded on Table 4.6. Overall, the SAS program produced the least complex maps, with a mean rank of 1.60. The CHANG program had a mean rank of 3.46, and the MONMONIER (both options), YOUNGMANN and JENKS (variance option) had ranks ranging from 4.06 to 5.03. The JENKS

Table 4.5

Tabulated Complexity Measure - Fragmentation Index

Distribution	R.S.	L.S.	L-	J-	REC.	BELL	BI-	OUT.
ST. JOHN'S	STJ1	STJ2	STJ3	STJ4	STJ5	STJ6	STJ7	
CHANG	.438	--	.500	.219	.407	--	.375	
JENKS (ABS.)	.438	--	.407	.313	.407	--	.344	
JENKS (VAR.)	.438	--	.407	.313	.438	--	.407	
MOELLERING	.500	--	.500	.282	.375	--	.375	
MON. (ABS.)	.438	--	.407	.313	.375	--	.344	
MON. (SQ.)	.469	--	.407	.313	.438	--	.500	
YOUNGMANN	.438	--	.407	.313	.375	--	.344	
SAS	.407	--	.344	.250	.375	--	.344	
EDMONTON	EDM1	EDM2	EDM3	EDM4	EDM5	EDM6	EDM7	
CHANG	.137	.185	n/a	.151	.336	.240	--	
JENKS (ABS.)	.213	.261	.185	.322	.405	.274	--	
JENKS (VAR.)	.185	.151	.089	.288	.357	.261	--	
MOELLERING	.206	.261	.295	.336	.377	.281	--	
MON. (ABS.)	.199	.261	.178	.274	.357	.309	--	
MON. (SQ.)	.185	.178	.089	.295	.336	.261	--	
YOUNGMANN	.199	.261	.178	.274	.411	.343	--	
SAS	.158	.137	.089	.199	.117	.172	--	
TORONTO	TOR1	TOR2	TOR3	TOR4	TOR5	TOR6	TOR7	TOR8
CHANG	.084	.144	.135	.043	.417	.223	--	.386
JENKS (ABS.)	.313	.276	.195	.200	.368	.378	--	.374
JENKS (VAR.)	.249	.274	.132	.175	.358	.332	--	.279
MOELLERING	.302	.267	.276	.193	.368	.379	--	n/a
MON. (ABS.)	.269	.280	.203	.192	.366	.379	--	.371
MON. (SQ.)	.198	.274	.112	.177	.368	.216	--	.279
YOUNGMANN	.299	.280	.195	.192	.366	.369	--	.360
SAS	.101	.093	.132	.045	.162	.216	--	.127

Note: Raw data sets STJ2, STJ6, EDM7 and TOR7 were not mapped because they are artificial and have no spatial interrelationships. Therefore, the fragmentation index is not available for these data.

Note: "n/a" denotes program inability to produce classed data for that data set.

Table 4.6

PROGRAM RANKINGS FOR THE FRAGMENTATION INDEX

ST. JOHN'S	R.S. STJ1	L.S. STJ2	L- STJ3	J- STJ4	REC. STJ5	BELL STJ6	BI- STJ7	OUT.
CHANG	4	--	7.5	1	5.5	--	5.5	
JENKS (ABS.)	4	--	4	6	5.5	--	2.5	
JENKS (VAR.)	4	--	6	6	7.5	--	7	
MOELLERING	8	--	7.5	3	2.5	--	5.5	
MON.(ABS.)	4	--	4	6	2.5	--	2.5	
MON.(SQ.)	7	--	4	6	7.5	--	8	
YOUNGMANN	4	--	4	6	2.5	--	2.5	
SAS	1	--	1	2	2.5	--	2.5	

EDMONTON	R.S. EDM1	L.S. EDM2	L- EDM3	J- EDM4	REC. EDM5	BELL EDM6	BI- EDM7	OUT.
CHANG	1	4	n/a	1	2.5	2	--	
JENKS (ABS.)	8	6.5	6	7	7	5	--	
JENKS (VAR.)	3.5	2	2	5	4.5	3.5	--	
MOELLERING	7	6.5	7	8	6	6	--	
MON.(ABS.)	5.5	6.5	4.5	3.5	4.5	7	--	
MON.(SQ.)	3.5	3	2	6	2.5	3.5	--	
YOUNGMANN	5.5	6.5	4.5	3.5	8	8	--	
SAS	2	1	2	2	1	1	--	

TORONTO	R.S. TOR1	L.S. TOR2	L- TOR3	J- TOR4	REC. TOR5	BELL TOR6	BI- TOR7	OUT. TOR8
CHANG	1	2	4	1	8	2	--	7
JENKS (ABS.)	8	6	5.5	8	6	6	--	6
JENKS (VAR.)	4	4.5	2.5	3	2	3.5	--	2.5
MOELLERING	7	3	8	7	6	7.5	--	n/a
MON.(ABS.)	5	7.5	7	5.5	3.5	7.5	--	5
MON.(SQ.)	3	4.5	1	4	6	3.5	--	2.5
YOUNGMANN	6	7.5	5.5	5.5	3.5	5	--	4
SAS	2	1	2.5	2	1	1	--	1

MEAN RANKS FOR EACH PROGRAM

CHANG	3.46
JENKS (ABS)	5.83
JENKS (VAR)	4.05
MOELLERING	6.16
MON. (ABS)	5.03
MON. (SQ)	4.31
YOUNGMANN	5.11
SAS	1.60

Note: Raw data sets for STJ2, STJ6, EDM7 and TOR7 were not mapped because they are artifical and have no spatial interrelationships. Therefore, rankings for the fragmentation index are not available for these data.

Note: "n/a" denotes program inability to produce classed data for that data set.

(absolute deviations option) and MOELLERING programs produced the most complex maps, with mean ranks of 5.83 and 6.16 respectively.

4.4.2 Visual Inspection

The visual inspection of the maps was done by data set, so that the maps created from the classed data produced by the eight classing techniques could be compared with one another (see Appendix 6 for examples). This testing proved to be an important facet of this evaluation, as it revealed information that could not be discovered by the application of the accuracy measures or the complexity measure.

As discussed in section 2.4, maps which are of low complexity are more easily interpreted by the user. Thus, it stands to reason that it is preferable to present users with maps of low complexity. This premise is true providing that the low degree of complexity does not inhibit the information content of the map to the point where there is very little information portrayed. From the application of the fragmentation index, it is known that the SAS program produces maps of low complexity. However, the visual inspection of the maps revealed that many of the SAS maps have large aggregations of areas of the same class, thus their information content is extremely low. For much of the map area, the spatial relationships of the data are unknown (Figure 4.1).

4.5 Data Distributions

The evaluation of the effects of different data distributions on the accuracy of the programs was carried out by averaging the three rankings (one from each area) for each program for each distribution. Results are presented in Table 4.7. Data distributions did not appear to affect program accuracy greatly, i.e. programs ranked fairly consistently regardless of distribution, although the ranges of rankings for each program varied marginally.

The CHANG program ranked between 4.41 for the bimodal distributions to 6.33 for the J-shaped distributions. JENKS (absolute deviations) rankings ranged from 3.25 for rectangular distributions to 5.33 for bell-shaped. JENKS (variance) had the highest rankings overall, ranging from 2.46 (left-skewed and rectangular distributions) to 4.04 for L-shaped. MOELLERING rankings were consistently low, from 5.17 for rectangular distributions to 7.25 for bimodal. Both options of the MONMONIER program ranked well consistently. The absolute deviations option ranged from 3.42 for the right-skewed distributions to 4.96 for the rectangular distributions.

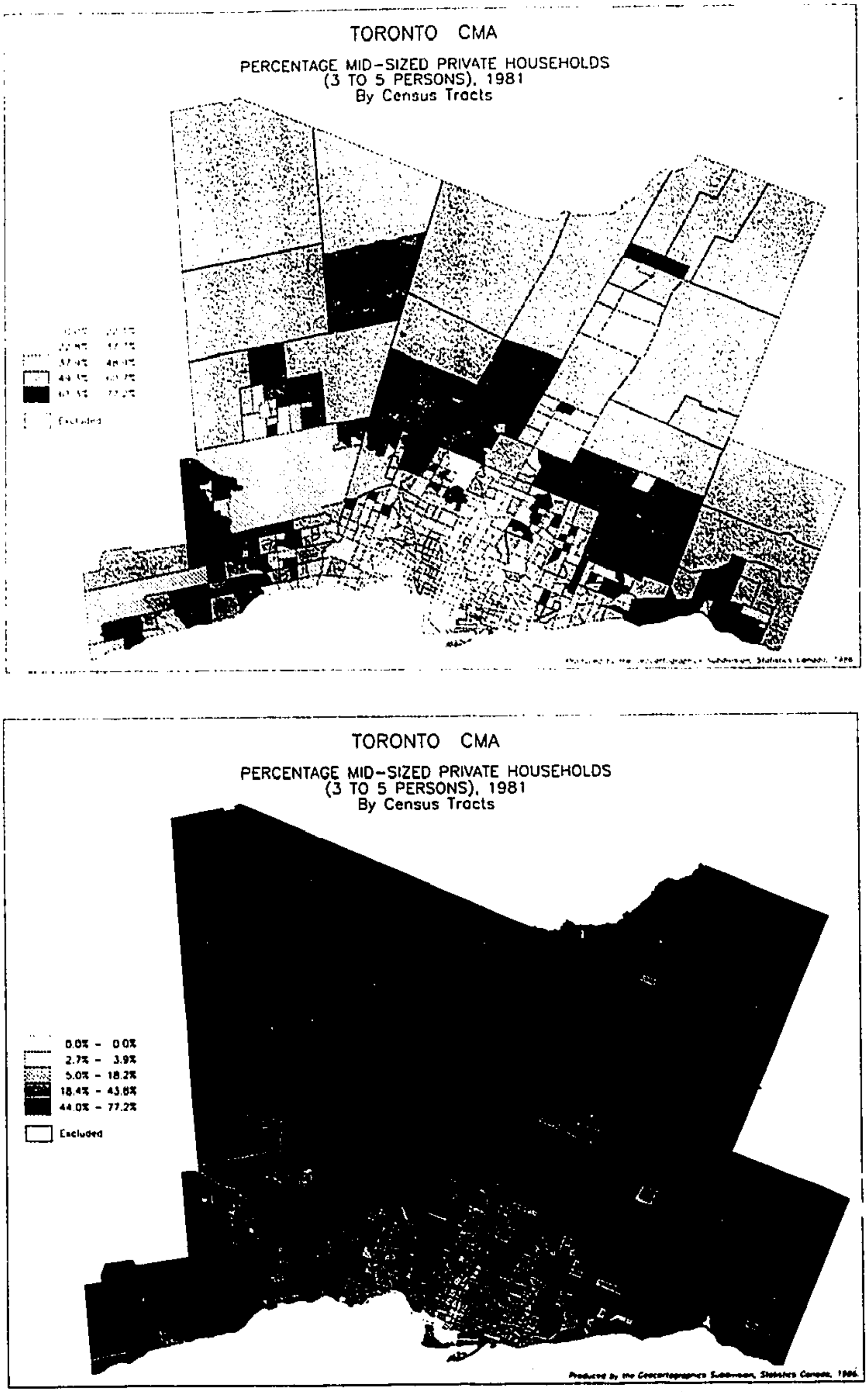


Figure 4.1 Most and Least Complex Maps - a) YOUNGMANN and b) SAS

Table 4.7

Program Rankings by Data Distribution

	R.S.	L.S.	L-	J-	REC.	BELL	BI-	OUT.
CHANG	6.17	6.00	5.00	6.33	5.58	5.33	4.41	4.46
JENKS (ABS.)	4.46	4.00	4.25	4.29	3.25	5.33	3.50	4.25
JENKS (VAR.)	2.83	2.46	4.04	2.71	2.46	2.96	3.25	3.63
MOELLERING	5.63	5.63	5.58	6.08	5.17	5.67	7.25	n/a
MON. (ABS.)	3.42	4.04	3.75	3.63	4.96	3.58	3.67	2.75
MON. (SQ.)	3.17	3.17	3.33	2.50	3.88	2.88	4.17	3.88
YOUNGMANN	4.00	3.08	3.58	3.63	3.88	3.75	4.58	5.50
SAS	6.33	6.83	4.54	6.83	6.75	6.50	5.17	3.50

Table 4.8

Program Rankings by Size of Data Set

	St. John's	Edmonton	Toronto
CHANG	5.92	5.38	5.31
JENKS (ABS.)	3.95	4.71	3.86
JENKS (VAR.)	2.88	2.71	3.31
MOELLERING	5.96	5.80	6.27
MON. (ABS.)	3.91	3.96	3.59
MON. (SQ.)	3.48	3.07	3.41
YOUNGMANN	3.57	3.89	4.09
SAS	6.04	6.09	5.94

Note: "n/a" denotes program inability to produce classed data for that data set.

The squared deviations option ranged from 2.50 for bell-shaped distributions to 4.17 for bimodal. The YOUNGMANN program also ranked well, from 3.08 for left-skewed distributions to 4.58 for bimodal. Overall, the SAS program ranked the lowest, from 4.54 for L-shaped distributions to 6.83 for left-skewed data.

Average ranks for each program were also calculated for the data set with outliers. All of the programs based on squared deviations of values from class means (JENKS (variance option), MONMONIER (squared deviations option) and SAS) isolated the outliers into their own class. This means that when mapped, the outliers can be identified separately rather than included in a class to which they are not homogeneous. However, the classed data sets that contain a separate class for outliers are not necessarily more accurate. This is illustrated in the results of this test, in that it was the absolute deviations option of the MONMONIER program that ranked first in terms of accuracy (2.75). The SAS (3.50), JENKS (variance option -- 3.63), and MONMONIER (squared deviations option -- 3.88) produced somewhat less accurate results. The MOELLERING program could not be tested using the data set with outliers because the outliers were zero values, and the MOELLERING program cannot process data sets with zeros.

4.6 Sizes of Data Sets

The methodology for evaluating the effects of distributions on the programs was also used for evaluating the effects of the sizes of data sets. Average ranks were calculated for each program for each size of data set. Results are shown on Table 4.8.

The size of data sets does not appear to be a contributing factor to the accuracy of the programs. The rankings for each program varied marginally amongst the different sizes of data set, and no trends appeared either with increasing or decreasing size.

4.7 Additional Testing on Selected Options

One data set from each area was randomly chosen to test options in the MONMONIER and YOUNGMANN programs. Extensive experimentation with these options is not within the scope of this study. However, even small tests aid in evaluating the programs. Results are tabulated on Tables 4.9 and 4.10.

4.7.1 The MONMONIER "Rounding" Option

The three data sets were inputted into the regular MONMONIER (absolute deviations) option and the rounding option. For the St. John's data set, the rounding option was not able to produce acceptable results, likely because

Table 4.9

The MONMONIER "Rounding" Option

	STJ3	EDM2	TOR5
MONMONIER regular			
TAI (abs.)	.78420	.75132	.77249
MONMONIER rounded			
TAI (abs.)	n/a	.72404	.77140

Table 4.10

The YOUNGMANN "Area" Option

	STJ3	EDM6	TOR5
YOUNGMANN regular			
TAI (abs.)	.78420	.73570	.77249
YOUNGMANN "Area" option			
TAI (abs.)	.74181	.64053	.75123

of the relatively few number of data values. For the Edmonton data, the rounding option had a TAI of .72404 compared to the TAI of the regular option of .75132. The difference between the TAIs for the Toronto data set was marginal -- .77249 for the regular option, and .77140 for the rounding option (Table 4.9).

This test was too small to make any conclusions about the rounding option. However, data set size could have an effect on the rounding option, as large data sets provide more "suboptimal" solutions to choose from. If, as stated by Monmonier (1982), the rounded class breaks enhance the map user's understanding of the map, this test corroborates his experimentation with the rounding option (1982), and illustrates the contradictory relationship between map accuracy and map enhancement.

4.7.2 The YOUNGMANN "Area" Option

The TAIs dropped considerably using the area option for each of the three areas. St. John's TAI decreased from .78420 to .74181, Edmonton decreased from .75132 to .64053, and Toronto dropped from .77249 to .75123 (Table 4.10).

The classed data sets produced by the area option were also mapped, and then the maps compared to the maps of the regular YOUNGMANN program. A visual comparison of the maps shows that the area option is successful in downplaying the perceived importance of large areas on the map (Figure 4.2). As with MONMONIER's rounding option, the inverse relationship between map accuracy and map enhancement is illustrated by this test.

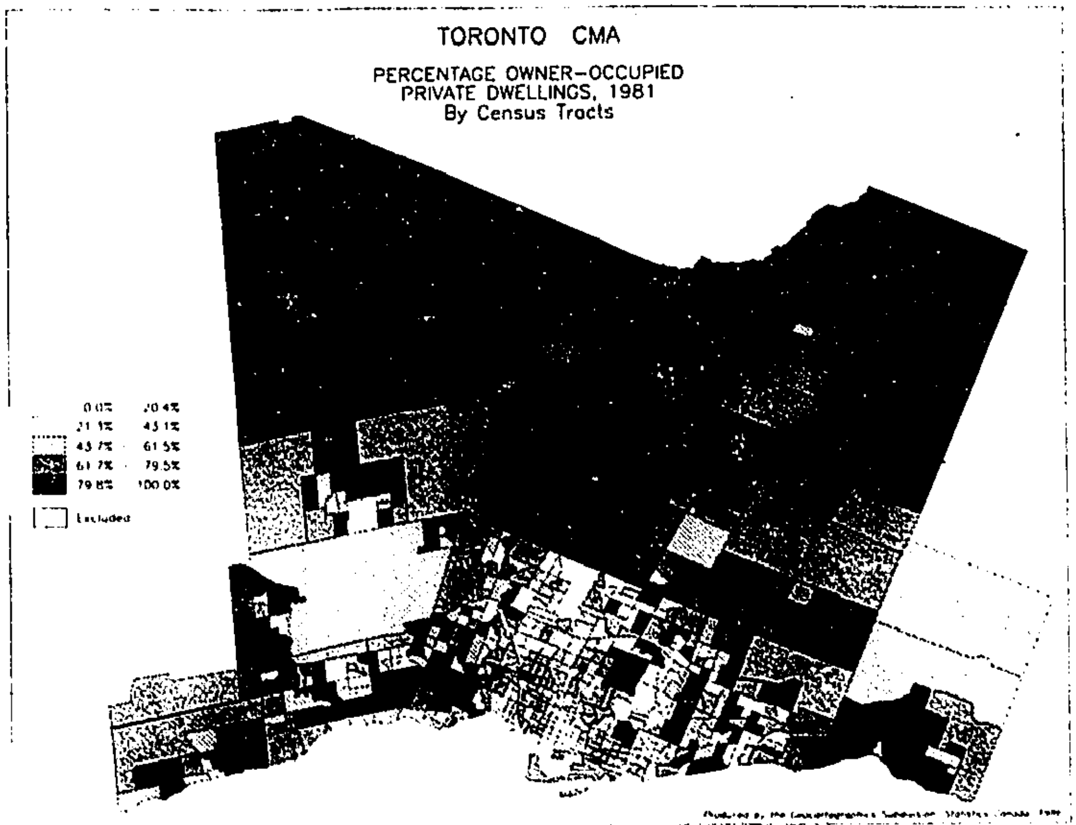
4.8 Comprehensive Results for Each Program

Results of the mathematical accuracy and visual characteristics testing are combined in this section, and discussed along with the operational advantages and disadvantages of each of the programs.

4.8.1 CHANG

The CHANG program ranked sixth overall in terms of mathematical accuracy. The maps produced by CHANG's classed data sets were of low complexity, but similar to the SAS maps, some of them had low information content. Within these overall rankings, CHANG varied comparatively widely, i.e. it produced accurate classes for some data sets and inaccurate ones for others. In addition, it varied in it's ability to provide accurate results based on data distribution; it ranked very poorly for skewed data and J-shaped data.

The main advantage of the CHANG program is that it has a user-defined class breakpoints option, and thus a)



b)

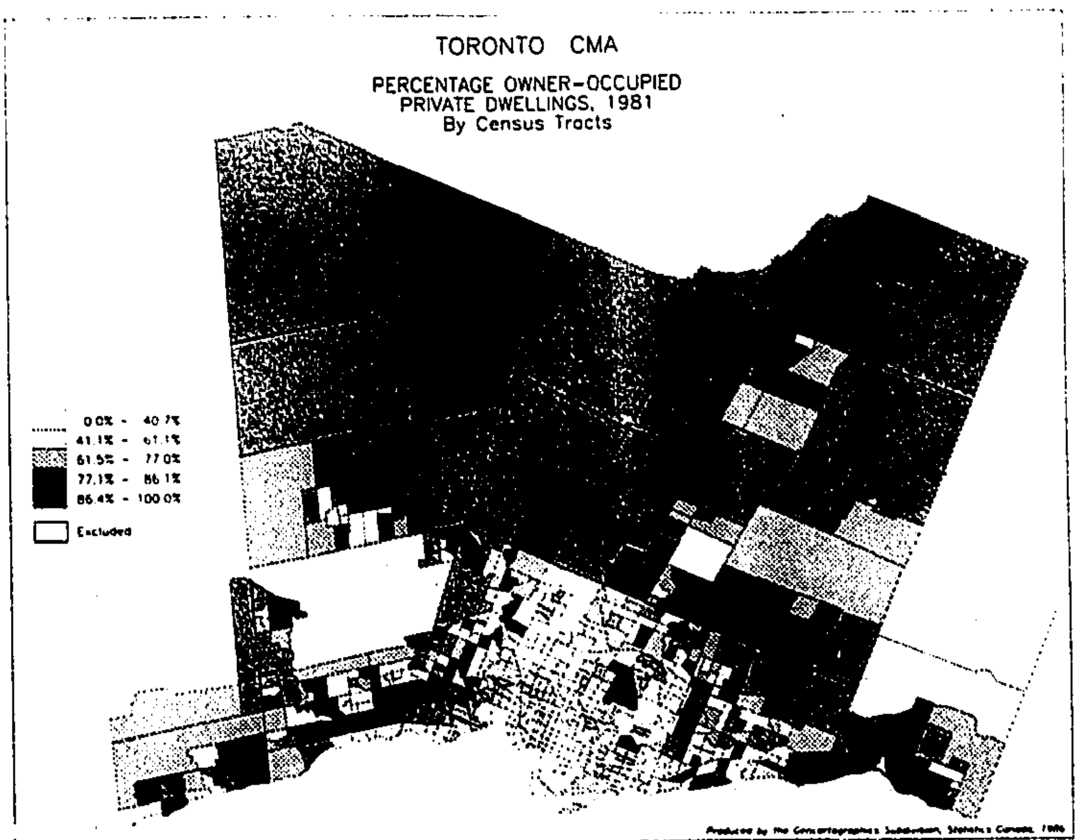


Figure 4.2 **Rectangular data distribution mapped by a) regular Youngmann classing program, and b) Youngmann "area" option.**

breakpoints can be set at "meaningful" values such as zero (for gain/loss or increase/decrease data), averages, or some other criterion significant to the data.

There are several drawbacks in the operation of CHANG. One is that it requires FORTRAN format statements for it's execution, and therefore is not as user-friendly as most of the other programs. Another disadvantage of the CHANG program is that the geometric and reciprocal techniques cannot operate on data sets with zeros because of the logarithms used in their calculations.

Several problems with the CHANG program were identified during the testing. One is that the equal frequency option may result in equal data values straddling two classes, or may calculate class frequencies incorrectly. Another problem is that when using the mean-and-standard-deviation option, negative values are outputted as negative class breaks, even when there are no negative values in the data. A third problem occurs when the data distribution is such that the first breakpoint is calculated between zero and one. In cases like this, the zero will be used as a denominator later in the program calculations, and an error message ensues. This occurred for the L-shaped Edmonton data set, and thus CHANG was not able to provide results for this data.

4.8.2 JENKS

Although the JENKS absolute deviations option ranked in the lower half of the techniques regarding mathematical accuracy, it still provided acceptable results regardless of data distribution. The maps produced by this technique were comparatively complex, ranking seventh overall.

The variance option produced excellent results. It ranked first overall in terms of mathematical accuracy, and third in terms of map complexity. The results did not vary significantly according to data distribution. This option is useful for isolating outliers in data sets.

The JENKS program is easily executed and results easily interpreted. The user-defined option allows for the inputting of breakpoints at "meaningful" values such as zero (for gain/loss or increase/decrease data), averages, or some other criterion significant to the data set. The only disadvantages associated with the JENKS program is that the user-defined option operates only when all breakpoints to acquire the requested number of classes are inputted. For example, it will not run if five classes are requested and only one breakpoint specified.

4.8.3 MOELLERING

Overall, the performance of the MOELLERING program was poor. It ranked seventh in terms of mathematical accuracy and produced the most complex maps of all of the techniques.

Although MOELLERING is easily executed, the output is more difficult to interpret than some of the other programs, as it is not obvious what accuracy measure is being maximized. The operational disadvantages of MOELLERING are numerous. It is the only one of the six programs tested that could not be modified to exclude data values without extensive restructuring of the entire algorithm. This operation was necessary because the raw data sets included "dummy" values which occur when data values are not available, or, particularly in the case of the census data used, when revealing particular data values would conflict with confidentiality and/or validity guidelines. Because it was not feasible to re-program the MOELLERING algorithm, a special "pre-processor" had to be programmed and added as an extra step before executing MOELLERING.

The major disadvantage of MOELLERING is that zeros and negative values cannot be classed. This hurdle was overcome, but not solved, by removing these numbers in the pre-processor stage. However, spatial data can include both zeros and negative values, and for these data the MOELLERING program cannot be used.

MOELLERING automatically determines the number of classes that is appropriate for the inputted data set, and calculates classes only for the number of classes within the range it specifies. Although the stated range is generally broad enough that this is not a problem, there may be situations when the cartographer wishes to have a number of classes outside the range specified by MOELLERING, and therefore this is another limiting factor of this program.

4.8.4 MONMONIER

Both the absolute deviations option and the squared deviations option in MONMONIER produced consistently good results both mathematically and in terms of visual characteristics. The absolute deviations option ranked fourth overall for accuracy, and the variance option ranked second overall. As with the variance option in the JENKS program, the MONMONIER squared deviations option is good for isolating outliers.

The MONMONIER program is simple to execute, and interpretation of the output is straightforward. It offers several options of interest to the cartographer: the rounding option, which yields more memorable class breakpoints with a minor sacrifice of accuracy, a user-defined breakpoints option, and an option which produces "gapped" intervals between class maxima and minima rather than breakpoints.

One operational disadvantage of the MONMONIER program is that it requires approximately 110% of the units of central processing time of the JENKS program, which is the most comparable program. In addition, independent testing of MONMONIER revealed a problem in the algorithm, i.e. when two consecutive classes contain only one value each, illogical class breakpoints are outputted.

4.8.5 YOUNGMANN

Although the theory underlying the YOUNGMANN program is the same as MONMONIER (absolute deviations option), the results were different due to a slight difference in the algorithm. YOUNGMANN ranked fourth overall in terms of accuracy, but still produced consistently good results regardless of data distribution. The maps it produced ranked fourth in complexity.

The YOUNGMANN program is easy to run, but the output requires more effort to interpret than those of JENKS and MONMONIER. The results of all steps of the forcing cycles are outputted, and the final solution is not necessarily the most accurate solution. Therefore, careful observation is required to identify the optimal solution.

The best reason for using the YOUNGMANN program is to experiment with the "area" option, in order to compensate for the perceived exaggerated importance of large areal units.

4.8.6 SAS

The SAS program ranked last in terms of mathematical accuracy and first in map complexity. However, the first ranking for map complexity does not mean that the SAS maps are superior; in fact, the visual inspection revealed that they were largely inferior to the maps produced by the other techniques because of their low information content.

SAS is the least user-friendly program in this study. It is possible that it would produce better results if some of the options outlined in section 3.3.6.2 were used, so that it would not over-fill classes. In all likelihood, this would increase the information content of the maps.

4.9 Summary

This chapter has presented the results of the testing in detail. Chapter Five puts these results into perspective by summarizing and drawing conclusions from them. The conclusions are then placed in the context of their contribution to the advancement of knowledge regarding the display of geographic data.

V. CONCLUSIONS

5.1 Summary

In this study, the classing techniques contained in six computer programs were evaluated in several ways. Firstly, the mathematical accuracy of the classed data sets outputted from each program was measured using four accuracy measures. Secondly, the visual complexity of the maps produced from the classed data sets was evaluated by the application of a complexity measure and by visual inspection. The third component of the evaluation documented the operational attributes of each of the programs.

Within the evaluations of mathematical accuracy and visual complexity, two factors were introduced and their effects examined. The first of these factors was the distribution of data values within data sets, and the second was the size of data sets.

5.2 Major Findings

The combined results of the mathematical accuracy testing and the evaluation of the visual complexity of the maps show that JENKS, MONMONIER and YOUNGMANN are superior to CHANG, MOELLERING and SAS. In this study, the JENKS, MONMONIER and YOUNGMANN programs provided consistently good classes regardless of data distribution. This finding corroborates earlier experimentation on the effects of data distributions on iterative classing techniques (Wasilenko and Moellering, 1977; Smith, 1986; Coulson, 1987).

In addition to their good results in terms of mathematical accuracy and map complexity, JENKS, MONMONIER, and YOUNGMANN were easy to execute, and outputs were readily interpreted. Options found in the MONMONIER and YOUNGMANN programs can ease user interpretation of the data with minimal sacrifices in map accuracy. The MONMONIER "rounding" option provides class breakpoints that are rounded, and thus more easily comprehended by the user. YOUNGMANN's "area" option compensates for differences in the size of the areal units, thereby reducing the exaggerated importance of large areas.

CHANG produced classed data of comparatively low accuracy, and was not able to provide a solution for one of the data sets used (L-shaped Edmonton data). The CHANG maps were generally of low complexity, and visual inspection showed that for some of the maps the information content was low.

MOELLERING ranked poorly both in accuracy and in visual complexity, and was unable to produce results for one of the input data sets (left-skewed St. John's data). In addition, its usefulness is severely limited because of its inability to process data sets with zero and negative data values. Spatial data are by no means limited to positive values only, and thus the MOELLERING program cannot be considered as a generally applicable classing program.

The SAS program produced consistently poor results in terms of mathematical accuracy. Maps produced from the SAS classed data were of very low complexity, and did not provide much information about the spatial inter-relationships of the data due to the over-filling of classes.

The comparison of map accuracy vs. map complexity found that the programs that produced the least complex maps (SAS and CHANG) also ranked poorly in terms of mathematical accuracy (last and fifth respectively). This finding supports the concern of researchers regarding the trade-off between accuracy and complexity (see section 3.4.2). However, the programs producing the most complex maps (MOELLERING and JENKS absolute deviations option respectively) were not ranked highest in terms of mathematical accuracy, as would be expected following the accuracy vs. complexity argument.

Although the JENKS (both options), MONMONIER (both options) and YOUNGMANN programs all produced good results, the variance option of the JENKS program produced the best results overall. It ranked highest in terms of mathematical accuracy, and produced maps of medium complexity regardless of data distribution. These results, combined with its ease of use and interpretation of output, identify it as the best method of classifying data for choropleth mapping. This finding affirms Coulson's (1987, p. 19) belief that the JENKS algorithm (variance option) achieves the:

> "geographical objective of seeking to group the data units such that the resulting map replicates the original data as closely as possible, given the generalization to a limited number of classes."

The consensus amongst many geographers (Jenks, 1977; Wasilenko and Moellering, 1977; Monmonier, 1982; Coulson, 1987; among others) is that the classification techniques that maximize between-group heterogeneity as well as maximizing within-group homogeneity provide the best representation of choropleth map data. This belief lends further support for the JENKS variance algorithm as the superior classification method, as it is the one that best achieves these relationships.

5.3 Future Research

The areas for future research can be described in three categories. The first of these is comprised of the research questions that result from the findings of the study itself. Secondly, research that should be done as an extension of the work is identified, followed by that which is related to the original research problem but is outside the scope of the study.

The selection of grey scales used in the area shadings for choropleth maps has been studied in the past, but during the visual inspection of the maps in this study a new area of experimentation was identified. Specifically, this research raised the question of whether grey scales for choropleth mapping should be regularly spaced to maximize user perception of differences in the grey scale, or whether they should reflect the relative proximity of the classes to one another. This question bears some investigation into map user perception of grey scales and shaded areas on choropleth maps.

Another research question resulting directly from the findings of this study is whether or not the "optimal" point of balance of mathematical accuracy and map complexity can be quantified. This study identified the JENKS program as being superior in this regard by comparing it to the other programs. However, quantification of the relationship between accuracy and complexity, and determining an acceptable range of values for this "optimal" solution would be of great benefit towards the formalization of geographic theory for data classification.

There are numerous areas for future research which are extensions of this investigation. In this study, the size of the data sets did not have a noticeable effect on the classing programs. While it may be expected that this finding would not change with different sizes of data sets other than those used for the study, this should be a factor in further experimentation. As an extension of this factor, the effect of data sets with different spatial organizations should be tested to see whether or not they have an effect on the complexity of the maps resulting from the classing programs. This could be tested, for example, by using other cities to provide small, medium and large data sets. The pattern of census tracts in the other cities may cause the classing programs to behave differently.

Although complexity is one of the major factors affecting the transmission of information on choropleth maps, it is not the only factor, and the broader question of the relationship between map accuracy and information flow should be further researched. Both the MONMONIER and YOUNGMANN programs attempt to increase information flow not by reducing complexity, but through the use of more easily interpretable class breakpoints (MONMONIER) and through correction of the disproportionate importance accorded to large areas (YOUNGMANN). Results of the small tests done in this study indicate that both programs succeed in achieving their objectives. However, further testing is necessary before conclusions can be drawn regarding their general applicability. Another method of downplaying the importance of large areas is through the use of an ecumene (Statistics Canada, 1981). This should be investigated in conjunction with the testing of the YOUNGMANN area option.

Further extensions of this research could also include testing with different numbers of classes, different accuracy measures and/or classing programs that were not available at the time of the initiation of the study. In addition, complexity measures other than the fragmentation index could be employed in the evaluation of map complexity.

As stated in section 1.7, this study focussed on evaluating the visual characteristics of the maps produced by the application of the fragmentation index, which measures map complexity. While complexity has historically been the most prominently featured visual characteristic in choropleth map classification experiments, many other visual characteristics affect map interpretation and perception, for example symbolization, scaling and use of colour. All of these factors, and any others which might be discovered through a map user survey, could be refined in the effort to formalize geographic theory.

Researching the broader issue of the relationship between map accuracy and information flow requires experimentation with test subjects, as the investigation relies upon human perception. Although clearly outside the scope of this study, a map user study would be of great value in corroborating the findings of this research, and in furthering their applicability. A map user study would assist in the determination of all the ways in which the information flow of choropleth maps could be improved.

5.4 Practical Applications of the Findings

Findings and conclusions of this research have been applied in the production of the 1986 Metropolitan Atlas series by Statistics Canada. Choropleth maps in the earlier 1981 Metropolitan Atlas series had classifications determined by the traditional classing methods found in the CHANG program. However, the MONMONIER and JENKS programs were adopted for the 1986 series based on the results of this study.

Several geographers, who recognize the limitations and inadequacies of the equal step and equal frequency techniques that are so often the only options offered in commercial programs, have expressed interest in the findings. Indeed, some developers of commercial mapping packages have expressed direct interest in these findings. In light of the increasing criticism of this aspect of commercial mapping packages (Anderson and Child, 1987; Noronha, 1987) interest in and implementation of iterative classing techniques will become more widespread once the results of this research and other studies like it are known. Once the first iterative classing technique is incorporated into a commercial mapping system, interest will blossom further.

5.5 Conclusions

The magnitude of the choropleth map data classification problem has not decreased over the period of time that it has been studied. In fact, today it is even more critical, with the advent of computer technology that enables choropleth maps to be generated almost instantly.

Dobson (1983a) proposed that computer technology, and in particular the field of computer graphics, has revolutionized the discipline of geography such that a new form of "automated geography" is possible. In another article on the same topic he stated that (1983b, p. 349):

> "...geographers should assume leadership to assure that the new spatial analysis techniques are used correctly..."

Although his proposal elicited varied responses from several geographers (Cowen, 1983; Kellerman, 1983; Monmonier, 1983; among others), all agreed that geographers cannot ignore the implications of the misuse of computer technology in the analysis and display of spatial data. Because of the widespread and ever-increasing use of maps in disciplines outside of geography, geographers have a responsibility to share their expertise in spatial analysis and display.

The rapid pace of technological development is now more than ever calling for the increased participation of geographers in the development of geographic information systems (Morrison, 1983; Muller et al., 1986; Eastman, 1987). Geographically-based expert systems will become a reality in the near future, and geographers will lose their position as experts in the collection, analysis and display of spatial data unless efforts are made to share our geographic knowledge with the computer scientists and engineers that are the system architects. The contribution of this study to geographic knowledge is that a superior approach to the classification of data for choropleth mapping has been identified using the most thorough evaluation to date, corroborating the earlier less comprehensive work of Smith (1986) and Coulson (1987). This classification technique should be incorporated into the mapping systems used by geographers and non-geographers alike in order to ensure optimal classes for choropleth maps.

The "traditional" classing techniques that group data according to their position in an ordered data set, such as the equal steps, equal frequency, and arithmetic techniques, do not provide a good representation of the original spatial data. In contrast, techniques which group data according to their homogeneity reflect the natural clusters in the data, and therefore produce more accurate representations of the geographic surface.

The variance option of the JENKS program has been identified as the best of the techniques that group homogeneous data values. It produces consistently good representations of the geographic phenomena being mapped.

Appendix 1

YOUNGMANN'S TABULAR ACCURACY INDEX CALCULATION

Appendix 1

TABULAR ACCURACY INDEX CALCULATION
HYPOTHETICAL X DISTRIBUTION

Step 1. Calculate grand mean

	x_i	
A	42	N = 11
B	26	
C	32	
D	18	
E	98	
F	120	
G	96	
H	118	
I	104	
J	114	
K	66	

$$\sum_{i}^{N} x_i = 834$$

$$\bar{x} = \frac{\sum_{i}^{N} x_i}{N} = \frac{834}{11} = 75.82$$

Step 2. Calculate class means

Class 1

	x_{ij}	
A	42	$n_1 = 4$
B	26	
C	32	
D	18	

$$\sum_{i}^{n_1} x_{i1} = 118$$

$$\bar{x}_1 = \frac{\sum_{i}^{n_1} x_{i1}}{n_1} = \frac{118}{4} = 29.5$$

Class 2

K	66	$n_2 = 1$

$$\sum_{i}^{n_2} x_{i2} = 66$$

$$\bar{x}_2 = \frac{66}{1} = 66$$

Source: Youngmann, 1972, pp. 44-47.

Class 3

E	98	$n_3 = 6$
F	120	
G	96	
H	118	
I	104	
J	114	

$$\sum_{i}^{n_3} x_{i3} = 650$$

$$\bar{x}_3 = \frac{650}{6} = 108.33$$

Step 3. <u>Calculate worst-case tabular inaccuracy</u>

	x_i	-	$\bar{x}$				
A	| 42	-	75.82 |	=	| -33.82 |	=	33.82
B	| 26	-	75.82 |	=	| -49.82 |	=	49.82
C	| 32	-	75.82 |	=	| -43.82 |	=	43.82
D	| 18	-	75.82 |	=	| -57.82 |	=	57.82
E	| 98	-	75.82 |	=	| 22.18 |	=	22.18
F	| 120	-	75.82 |	=	| 44.18 |	=	44.18
G	| 96	-	75.82 |	=	| 20.18 |	=	20.18
H	| 118	-	75.82 |	=	| 42.18 |	=	42.18
I	| 104	-	75.82 |	=	| 28.18 |	=	28.18
J	| 114	-	75.82 |	=	| 38.18 |	=	38.18
K	| 66	-	75.82 |	=	| -9.82 |	=	9.82

$$\sum_{i}^{N} |x_i - \bar{x}| = 390.18$$

Step 4. <u>Calculate within-class tabular inaccuracy</u>

Class 1

	x_{ij}	-	$\bar{x}_i$				
A	| 42	-	29.5 |	=	| 12.5 |	=	12.5
B	| 26	-	29.5 |	=	| -3.5 |	=	3.5
C	| 32	-	29.5 |	=	| 2.5 |	=	2.5
D	| 18	-	29.5 |	=	| -11.5 |	=	11.5

$$\sum_{i}^{n_1} |x_{i1} - \bar{x}_1| = 30.0$$

Class 2

K	$\lvert 66 - 66 \rvert$	=	$\lvert 0.0 \rvert$	= 0.0

$$\sum_{i}^{n_2} \lvert x_{i2} - \bar{x}_2 \rvert = 0.0$$

Class 3

E	$\lvert 98 - 108.33 \rvert$	=	$\lvert 10.33 \rvert$	= 10.33
F	$\lvert 120 - 108.33 \rvert$	=	$\lvert 11.67 \rvert$	= 11.67
G	$\lvert 96 - 108.33 \rvert$	=	$\lvert -12.33 \rvert$	= 12.33
H	$\lvert 118 - 108.33 \rvert$	=	$\lvert 9.67 \rvert$	= 9.67
I	$\lvert 104 - 108.33 \rvert$	=	$\lvert -4.33 \rvert$	= 4.33
J	$\lvert 114 - 108.33 \rvert$	=	$\lvert 5.67 \rvert$	= 5.67

$$\sum_{i}^{n_3} \lvert x_{i3} - \bar{x}_3 \rvert = 54.0$$

$$m = 3 \qquad \sum_{j}^{m} \sum_{i}^{n_j} \lvert x_{ij} - \bar{x}_j \rvert = 84.0$$

Step 5. Tabular Accuracy Index

$$TAI = 1.0 - \frac{\sum_{j}^{m} \sum_{i}^{n_j} \lvert x_{ij} - \bar{x}_j \rvert}{\sum_{i}^{N} \lvert x_i - \bar{x} \rvert}$$

$$= 1.0 - \frac{84.0}{390.18}$$

$$= 1.0 - .2050$$

$$= .7950$$

Appendix 2

MOELLERING'S PROPOSED INFORMATION THEORETIC MODEL

Appendix 2

The Proposed Information Theoretic Model

For any given data with N number of observations, transform each observation into a proportion of the total, giving

$$Y_i = \frac{Z_i}{\sum_{i=1}^{N} Z_i} \tag{4}$$

where $\sum_{i=1}^{N} Y_i = 1.0$ and Z_i is the i^{th} datum. Through a modified version of Shannon's model (1951) in Semple (1973) and Griffin and Semple (1971), the information statistic H(Y) for the unclassed data set is

$$H(Y) = \sum_{i=1}^{N} Y_i \log_2 1/Y_i \quad . \tag{5}$$

The value of H(Y) is the actual statistic for a given data set. Regardless of further partitioning of the model, the sum of all partitions must equal H(Y). Therefore

$$H(Y) = H'(Y) + H''(Y) + H'''(Y) + \ldots . \tag{6}$$

The theoretic maximum statistic for the data will occur when all values are equal. Under this condition the maximum statistic for the data is

$$H(Y) = \log_2 N \quad . \tag{7}$$

In order to generalize, group the data into C classes such that $Y_c = \sum_{i\epsilon c} y_i$ and $\sum_{c=1}^{C} Y_c = 1.0$ and $c = 1,2,\ldots,C$.

The between class statistic $H^b(Y)$ may then be calculated

$$H^b(Y) = \sum_{c=1}^{C} Y_c \log_2 1/Y_c \tag{8}$$

Source: Wasilenko and Moellering, 1977, pp. 13-15, 21.

Consequently, the within class statistic $H^W(Y)$ may be defined as

$$H^W(Y) = \sum_{c=1}^{C} Y_c \left[\sum_{i\epsilon c} y_i/Y_c \log_2 Y_c/y_i \right] \quad . \tag{9}$$

By combining (8) and (9), a model for generalizing data for choropleth maps emerges

$$H(Y) = \sum_{c=1}^{C} Y_c \log_2 1/Y_c + \sum_{c=1}^{C} Y_c \left[\sum_{i\epsilon c} y_i/Y_c \log_2 Y_c/y_i \right] \quad . \tag{10}$$

The general relationship of this model may be expressed as;

$$H(Y) = H^b(Y) + H^W(Y) \tag{11}$$

and the sum of $H^b(Y)$ and $H^W(Y)$ must equal $H(Y)$ for any given data set.

It is also possible to calculate a percent of equality for each class and then the total percent of equality for the map. This is a simple relationship between the statistics actually calculated for each class and the maximum statistic for each class. The percent of equality for each class e_c is;

$$e_c = \frac{H^W(Y)}{H^m(Y)} \times 100 \tag{12}$$

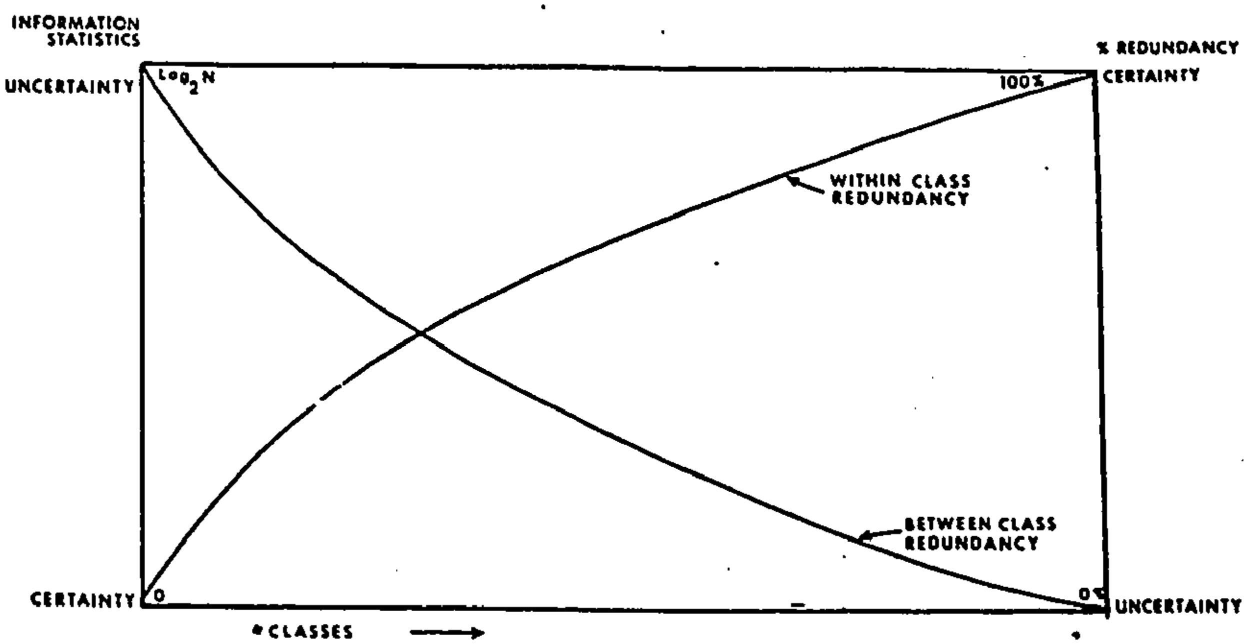
where $H^m(Y) = Y_c \log_2 I_c$, $H^m(Y)$ is the theoretic maximum statistic for a given class, and I_c is the number of observations in each class. Equation (12) may then be restated as

$$e_c = \frac{Y_c \left[\sum_{i\epsilon c} y_i/Y_c \log_2 Y_c/y_i \right]}{Y_c \log_2 I_c} \times 100 \quad . \tag{13}$$

The percent of equality for the map as a whole is

$$e = \frac{\sum_{c=1}^{C} Y_c \left[\sum_{i\epsilon c} y_i/Y_c \log_2 Y_c/y_i \right]}{\sum_{c=1}^{C} Y_c \log_2 I_c} \times 100 \tag{14}$$

The percent of equality measure developed here is analogous to the concept of relative entropy used by Monmonier (1974) as a measure of regional size disparity and the percent of inequality utilized by Semple, Youngmann and Zeller (1972) and Semple and Scorrar (1975). Also it is important to note that the search for a high within class statistic at this point is desirable, since by maximizing the within class statistic the between class statistic is minimized, thus producing a map with a low level of uncertainty while maintaining equality within the classes.



The relationship between Information Statistics, % of Redundancy and the concept of Uncertainty.

Appendix 3

CALCULATION OF THE D VALUE

Appendix 3

Type of Class Interval	Class Number	Class limits	1 Sum of N	2 N	3 Mean of class	4 Class range	5 Weighted Range	6 Th. Mean of class	7 Th. Weighted range	8 Diff.
A Clinographic	1	1.6– 5.9	168.0	45	3.75	4.4	1.179	3.8	1.158	.021
	2	6.0– 8.9	324.2	45	7.20	3.0	.417	7.5	.400	.017
	3	9.0– 13.9	119.9	11	10.9	5.0	.459	11.5	.435	.024
	4	14.0– 18.9	18.2	1	18.2	4.0	.220	16.5	.242	.022
	5	19.0– 27.9	27.4	1	27.4	9.0	.328	23.5	.383	.055
	6	28.0– 34.9	34.6	1	34.6	7.0	.202	31.5	.222	.020
	7	35.0–103.4	103.4	1	103.4	68.4	.662	69.2	.988	.326
									Sum of D	* .529
B Septile	1	1.6– 3.1	33.2	14	2.37	1.6	.675	2.4	.667	.008
	2	3.2– 4.3	53.9	15	3.75	1.2	.322	3.8	.316	.006
	3	4.4– 5.2	73.6	15	4.91	.9	.183	4.85	.186	.003
	4	5.3– 6.6	87.2	14	6.23	1.4	.225	6.0	.233	.008
	5	6.7– 7.6	111.7	16	6.98	1.0	.143	7.2	.139	.004
	6	7.7– 8.9	130.6	16	8.16	1.3	.159	8.35	.156	.003
	7	9.0–103.4	303.5	15	20.23	93.4	4.617	56.25	1.660	2.957
									Sum of D	* 2.989
C Equal Step	1	1.6– 16.0	612.1	.101	6.06	14.54	2.399	8.87	1.639	06.7
	2	16.1– 30.6	45.6	2	22.8	14.54	.637	23.41	.621	.610
	3	30.7– 45.1	34.6	1	34.6	14.54	.420	37.95	.383	0.37
	4	45.2– 59.7	0	0	0	14.54		52.49	.277	
	5	59.8– 74.2	0	0	0	14.54		67.03	.217	
	6	74.3– 88.7	0	0	0	14.54		81.57	.178	
	7	88.8–103.4	103.4	1	103.4	14.54	.140	96.11	.151	.011
									Sum of D	* .824

1 Sum of the population density values for all unit areas in the class.
2 Number of unit areas in the class.
3 Sum of density values divided by the number of unit areas in class.
4 It is assumed that the range of the class contains all values from the beginning of the class to the beginning of the next class.
5 Range of the class divided by the mean of the class.
6 Mean of each class determined by halving the range of the class and adding this amount to the beginning of the class.
7 Theoretical mean of the class divided into the range of the class.
8 The difference between the weighted range and the theoretical weighted range of each class. These values are totaled to obtain the value Sum of D. for each function.
* These are partial values since one or more classes are vacant. When class intervals in any series are vacant that set of intervals should be eliminated from further consideration. The Equal Step and Arithmetic functions are included in this table for illustrative purposes, but in practice parts C and D of this table would not have been calculated.

Source: Jenks and Coulson, 1963 pp. 130, 131

Appendix 4

SAMPLE COMPUTER PRINTOUTS

4a. CHANG Program

RECIPROCAL METHOD RESULTS******************

CLASS BREAKS ARE:

51.15
57.59
65.90
77.00

COMPILATION OF CLASS FREQUENCIES:

CLASS	FREQUENCY
1	1
2	1
3	8
4	48
5	81

COMPUTATION OF D VALUES

CLASS	D VALUE
1	0.00593
2	0.00547
3	0.00273
4	0.00286
5	0.00586

TOTAL D VALUE 0.02285

4b. JENKS Program

A 5 CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF 214.80002

CLASS	N	LARGEST	SMALLEST	MEDIAN	ABS.DEV.
5	27	92.60000	83.50000	85.80000	50.20000
4	35	83.40000	78.40000	81.10000	36.49998
3	27	79.10000	76.30000	77.30000	15.10001
2	27	75.70000	70.90000	74.40000	33.30000
1	23	70.10000	46.00000	66.70000	79.70000

A 4 CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF 257.80002

CLASS	N	LARGEST	SMALLEST	MEDIAN	ABS.DEV.
4	27	92.60000	83.50000	85.80000	50.20000
3	37	83.40000	78.90000	81.10000	40.69998
2	47	78.80000	72.50000	76.60000	66.70001
1	28	71.60000	46.00000	67.30000	100.20000

A 3 CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF 326.70004

CLASS	N	LARGEST	SMALLEST	MEDIAN	ABS.DEV.
3	48	92.60000	80.70000	84.05000	111.90000
2	63	80.50000	72.50000	77.20000	114.60002
1	28	71.60000	46.00000	67.30000	100.20000

A 2 CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF 492.00043

CLASS	N	LARGEST	SMALLEST	MEDIAN	ABS.DEV.
2	65	92.60000	78.80000	82.70000	172.10004
1	74	78.30000	46.00000	74.45000	319.90039

A 1 CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF 775.50049

CLASS	N	LARGEST	SMALLEST	MEDIAN	ABS.DEV.
1	139	92.60000	46.00000	77.80000	775.50049

```
THE CLASS BREAKS WITH  5 CLASSES OPTIMIZED

SUM1 =            4.87754
SUM2 =            4.87070
CHECK =           99.98263
CHECK1 =           99.98264

THE CLASSES FOR THE TOTAL SYSTEM

THE UPPER LIMIT FOR CLASS  1 =          67.90010071
THE UPPER LIMIT FOR CLASS  2 =          75.70009613
THE UPPER LIMIT FOR CLASS  3 =          79.60009766
THE UPPER LIMIT FOR CLASS  4 =          86.70009613
THE UPPER LIMIT FOR CLASS  5 =          92.60009766

THE NUMBER OF OBSERVATIONS IN CLASS  1 =  17
THE NUMBER OF OBSERVATIONS IN CLASS  2 =  33
THE NUMBER OF OBSERVATIONS IN CLASS  3 =  31
THE NUMBER OF OBSERVATIONS IN CLASS  4 =  38
THE NUMBER OF OBSERVATIONS IN CLASS  5 =  20

THE BREAK IS AT         0.00631 THE WITHIN CLASS STATISTIC IS        0.40900
THE BREAK IS AT         0.00704 THE WITHIN CLASS STATISTIC IS        1.12971
THE BREAK IS AT         0.00740 THE WITHIN CLASS STATISTIC IS        1.10959
THE BREAK IS AT         0.00787 THE WITHIN CLASS STATISTIC IS        1.51837
THE BREAK IS AT         0.00861 THE WITHIN CLASS STATISTIC IS        0.70227

THE BETWEEN CLASS STATISTIC =          2.24318

THE PERCENT OF ACCURACY OF CLASS  1 =        99.86742
THE PERCENT OF ACCURACY OF CLASS  2 =        99.98647
THE PERCENT OF ACCURACY OF CLASS  3 =        99.99787
THE PERCENT OF ACCURACY OF CLASS  4 =        99.99563
THE PERCENT OF ACCURACY OF CLASS  5 =        99.99146
```

4d. MONMONIER Program

```
#CLASSES =  5     JIGGLE  0     TAI =    0.710     IMPROVES =   0
#CLASSES = -5     JIGGLE  1     TAI =    0.710     IMPROVES =   0

***********************
*                     *
* ABSOLUTE DEVIATIONS *
*                     *
***********************
```

REPORT FOR 5 CATEGORIES WITH INDEX OF TABULAR ACCURACY = 0.71014

	ALL DATA	CLASS 1	CLASS 2	CLASS 3	CLASS 4	CLASS 5
MEAN:	77.434851	60.188889	68.847374	75.923912	81.074356	86.680771
SUM OF DEVIATIONS:	777.75604	34.755558	30.747375	65.252190	45.425621	49.261543
LOW VALUE:	46.000000	46.000000	65.500000	72.500000	78.800003	84.000000
HIGH VALUE:	92.599998	65.400002	71.599998	78.300003	83.500000	92.599998
HI/LO RANGE:	46.599998	19.400002	6.0999985	5.8000031	4.6999969	8.5999983
CATEGORY SIZE:	139	9	19	46	39	26

BREAK	GAP LOW	GAP HIGH	GAP RANGE	MID-POINT
1	65.400	65.500	0.99998E-01	65.450
2	71.600	72.500	0.90000	72.050
3	78.300	78.800	0.50000	78.550
4	83.500	84.000	0.50000	83.750

OPTIMAL CATEGORY LABELS

LOWER LIMIT:	46.000
BREAK 1:	65.450
BREAK 2:	72.050
BREAK 3:	78.550
BREAK 4:	83.750
UPPER LIMIT:	92.600

REITERATION OF FORCING

TABULAR ACCURACY ANALYSIS

CLASS	N	RANGE		MEAN	ELEVATION
1	7.	46.00000	63.40000	58.80000	58.80000
2	21.	64.70000	71.60000	68.48572	68.48572
3	46.	72.50000	78.30000	75.92392	75.92392
4	39.	78.80000	83.50000	81.07435	81.07435
5	26.	84.00000	92.60000	86.68078	86.68078

TABULAR ACCURACY INDEX = 0.70836

OVERVIEW ACCURACY INDEX = 0.70836

4f. SAS Program

PRINT OF INPUT BEFORE FASTCLUS

FASTCLUS PROCEDURE

CLUSTER SUMMARY

CLUSTER	MEMBERS	RMS ST DEV	MAX DISTANCE FROM SEED
1	1	.	0
2	22	2.097313	4.133333
3	30	2.936657	5.156667
4	80	2.794167	5.180435
5	6	2.327603	3.5

STATISTICS FOR VARIABLES

VARIABLE	TOTAL STD	WITHIN STD	R-SQUARED	VAR RATIO
EDM2A	7.41058119	2.71307390	0.86984993	6.68343779
OVER-ALL	7.41058119	2.71307390	0.86984993	6.68343779

APPROXIMATE EXPECTED OVER-ALL R-SQUARED = 0.9631 CUBIC CLUSTERING CRITERION =-10.9759
WARNING: THESE VALUES ARE INVALID IF VARIABLES ARE CORRELATED

CLUSTER MEANS

CLUSTER	EDM2A
1	46.00000
2	87.15909
3	69.98667
4	79.11500
5	60.93333

CLUSTER STANDARD DEVIATIONS

CLUSTER	EDM2A
1	.
2	2.097313
3	2.936657
4	2.794167
5	2.327603

4f.....cont'd

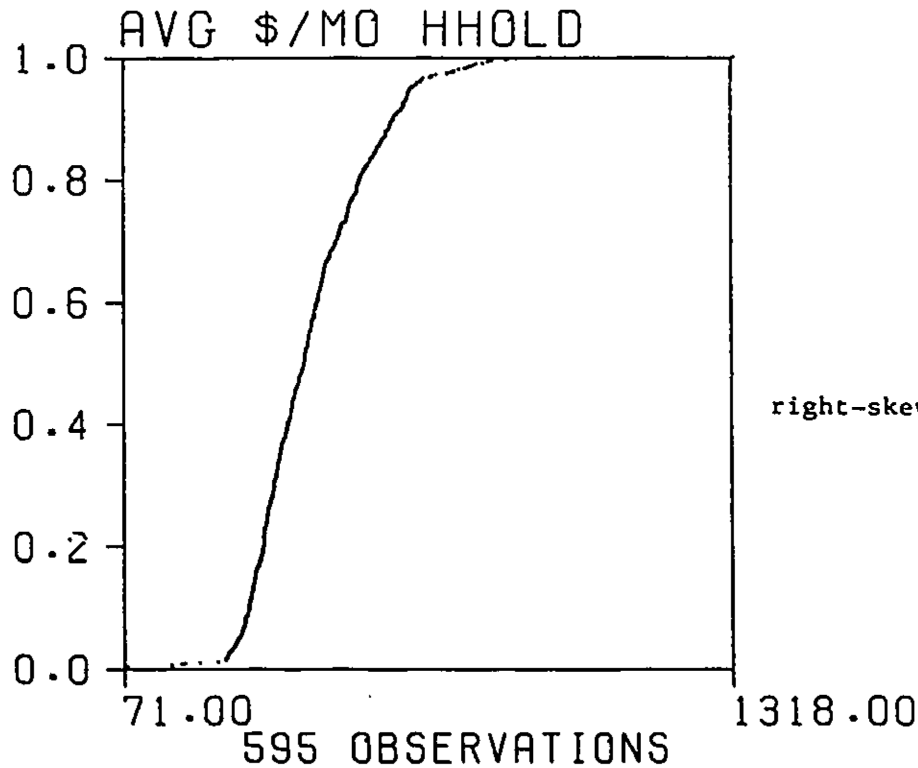
PRINT AFTER SORT

OBS	EDM2A	CLUSTER	DISTANCE
1	46.0000	1	0.00000
2	57.8000	5	2.80000
3	60.2000	5	0.40000
4	60.4000	5	0.60000
5	61.6000	5	1.80000
6	63.0000	5	3.20000
7	63.4000	5	3.60000
8	64.7000	3	4.54333
9	65.4000	3	3.84333
10	65.5000	3	3.74333
11	66.5000	3	2.74333
12	66.7000	3	2.54333
13	67.1000	3	2.14333
14	67.3000	3	1.94333
15	67.3000	3	1.94333
16	67.9000	3	1.34333
17	67.9000	3	1.34333
18	68.2000	3	1.04333
19	68.7000	3	0.54333
20	69.2000	3	0.04333
21	69.8000	3	0.55667
22	69.9000	3	0.65667
23	70.1000	3	0.85667
24	70.9000	3	1.65667
25	71.0000	3	1.75667
26	71.0000	3	1.75667
27	71.5000	3	2.25667
28	71.6000	3	2.35667
29	72.5000	3	3.25667
30	73.1000	3	3.85667
31	73.2000	3	3.95667
32	73.2000	3	3.95667
33	73.4000	3	4.15667
34	73.6000	3	4.35667
35	73.7000	3	4.45667
36	74.3000	3	5.05667
37	74.4000	3	5.15667
38	74.5000	4	5.10043
39	74.5000	4	5.10043
40	74.7000	4	4.90043
41	74.8000	4	4.80043
42	74.8000	4	4.80043
43	74.9000	4	4.70043
44	75.0000	4	4.60043
45	75.2000	4	4.40043
46	75.5000	4	4.10043
47	75.5000	4	4.10043
48	75.6000	4	4.00043
49	75.6000	4	4.00043
50	75.7000	4	3.90043
51	76.3000	4	3.30043
52	76.6000	4	3.00043
53	76.7000	4	2.90043
54	76.7000	4	2.90043
55	76.8000	4	2.80043
56	76.8000	4	2.80043

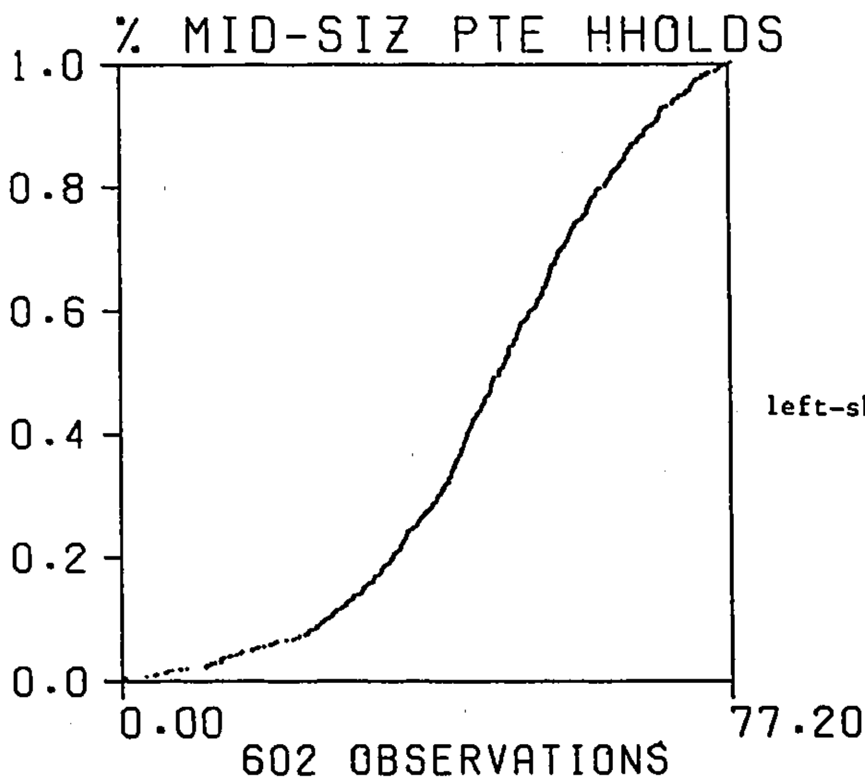
Appendix 5

SAMPLE ECDF PLOTS OF DATA CHARACTERISTICS

TORONTO

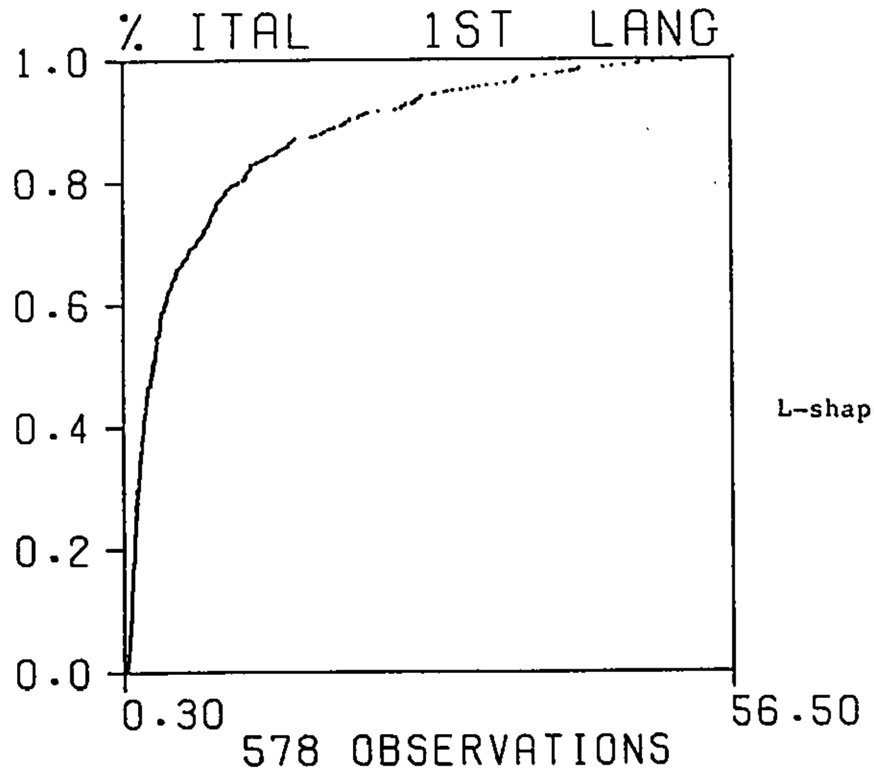


right-skewed

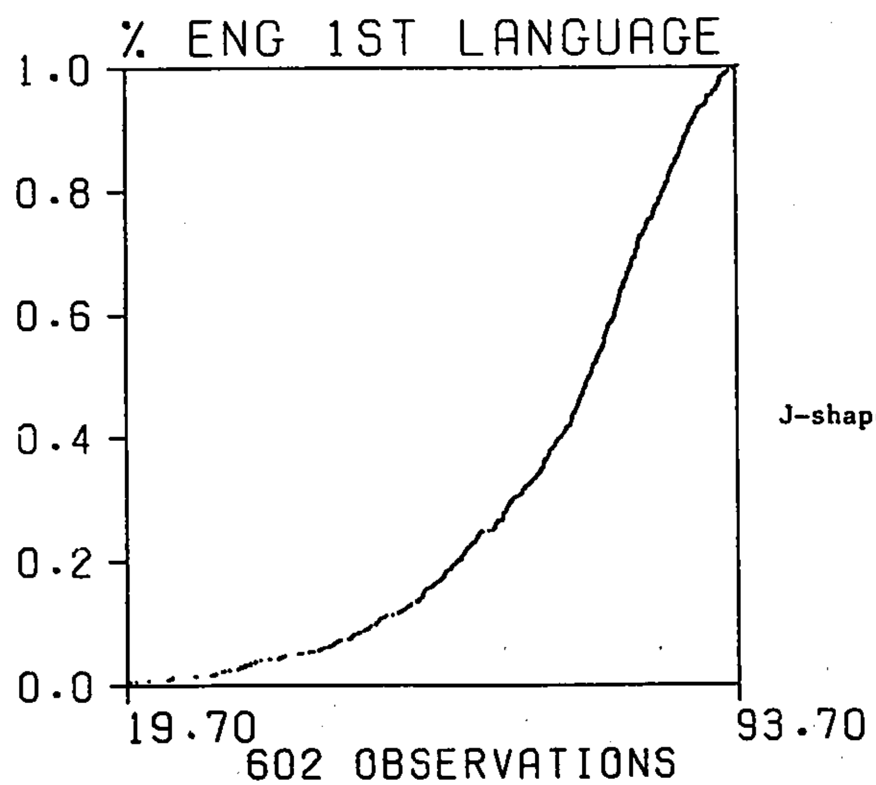


left-skewed

TORONTO

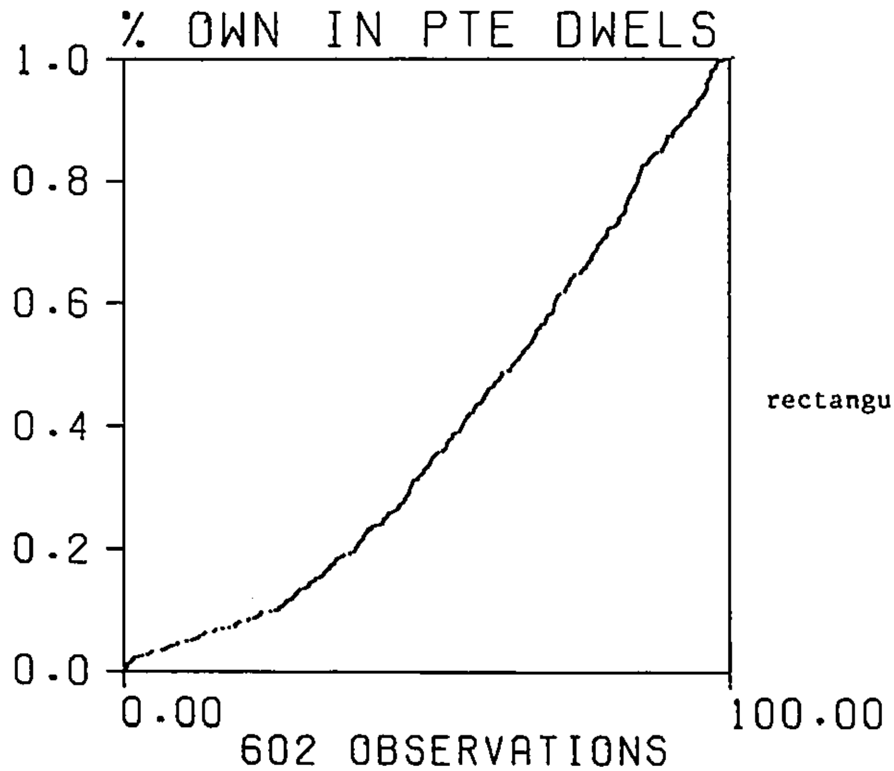


L-shaped



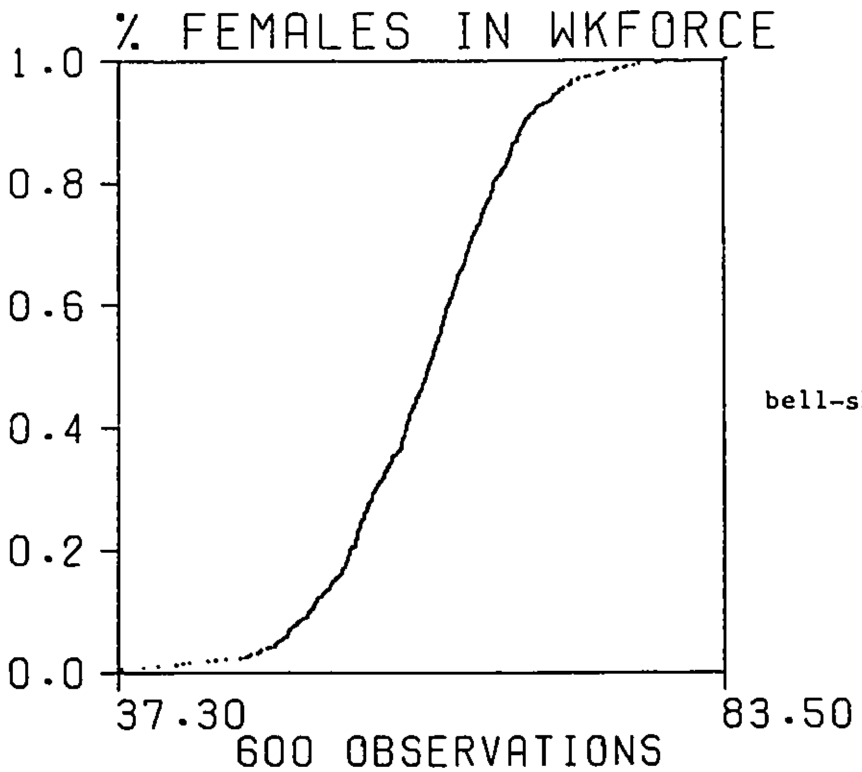
J-shaped

TORONTO

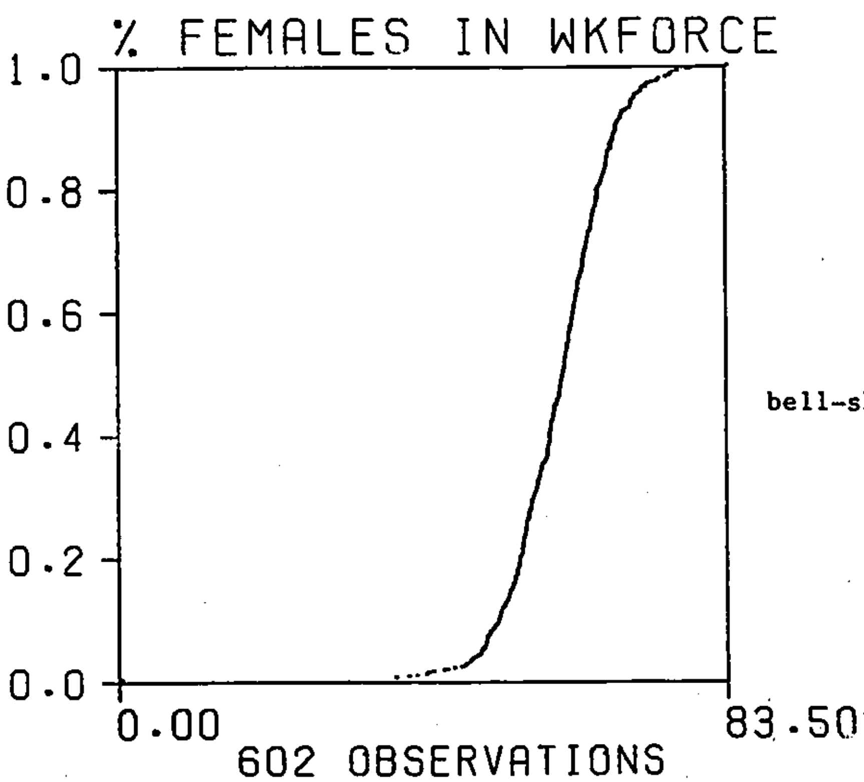


rectangular

TORONTO



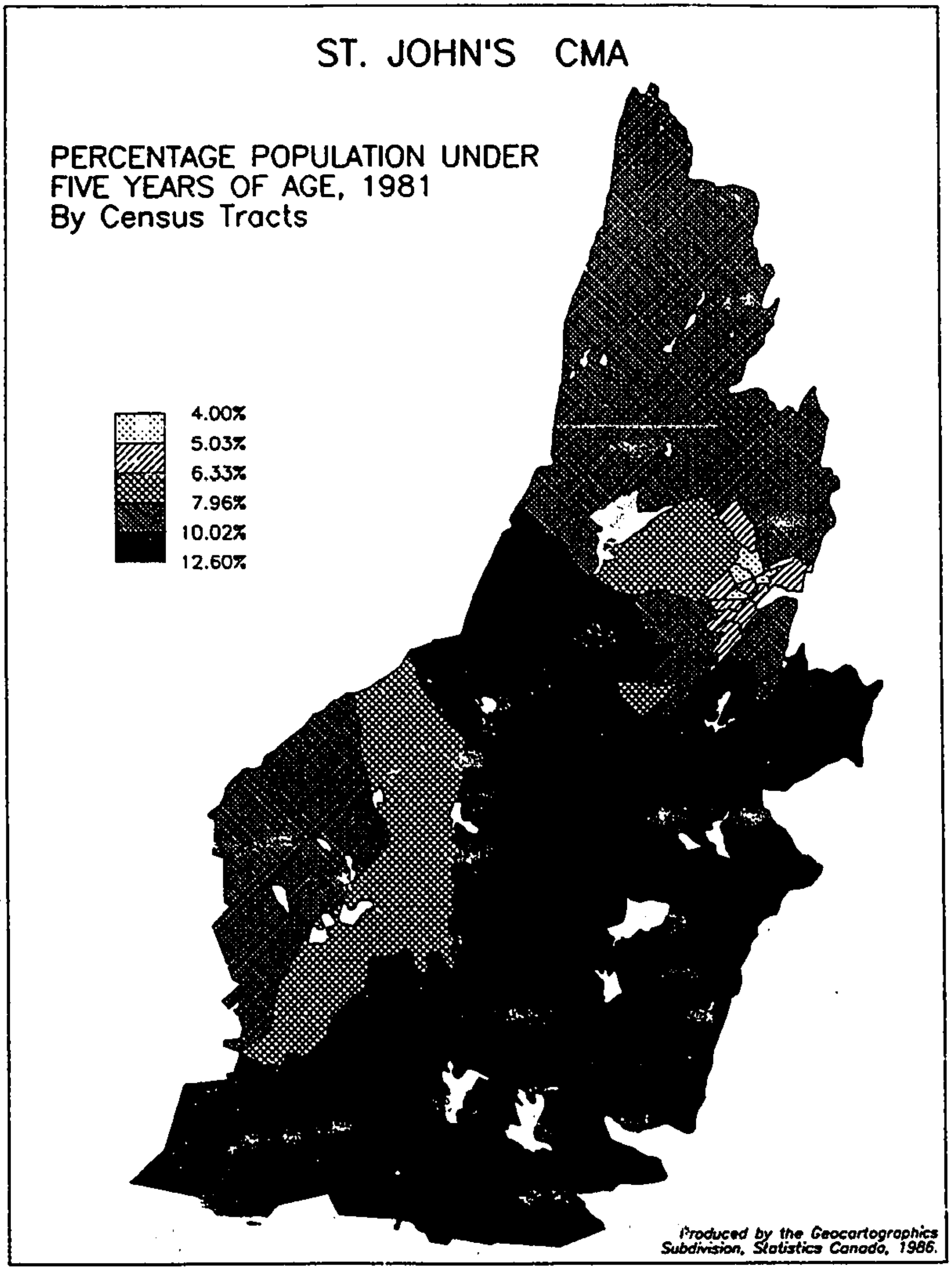
bell-shaped



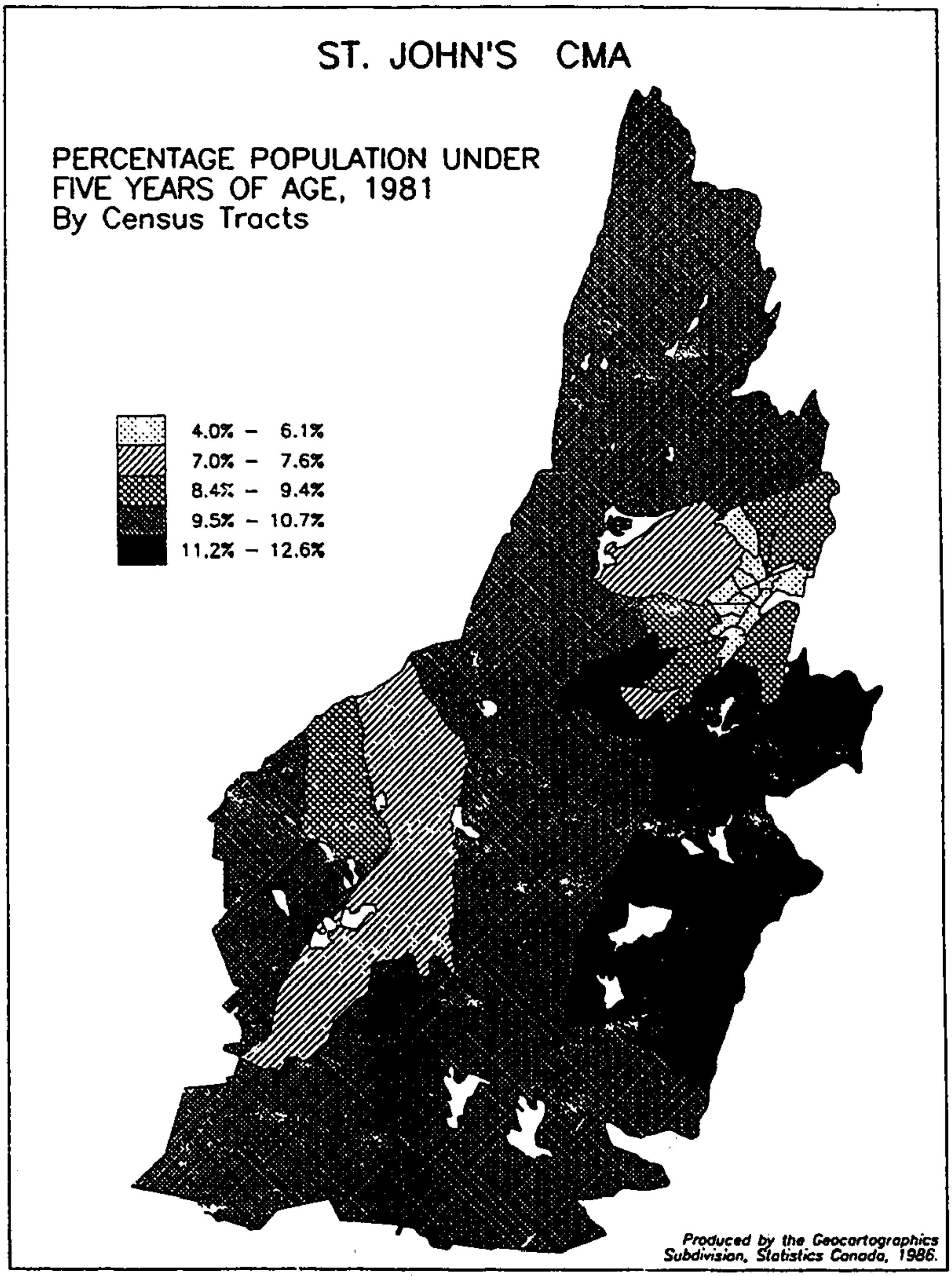
bell-shaped with outliers

Appendix 6

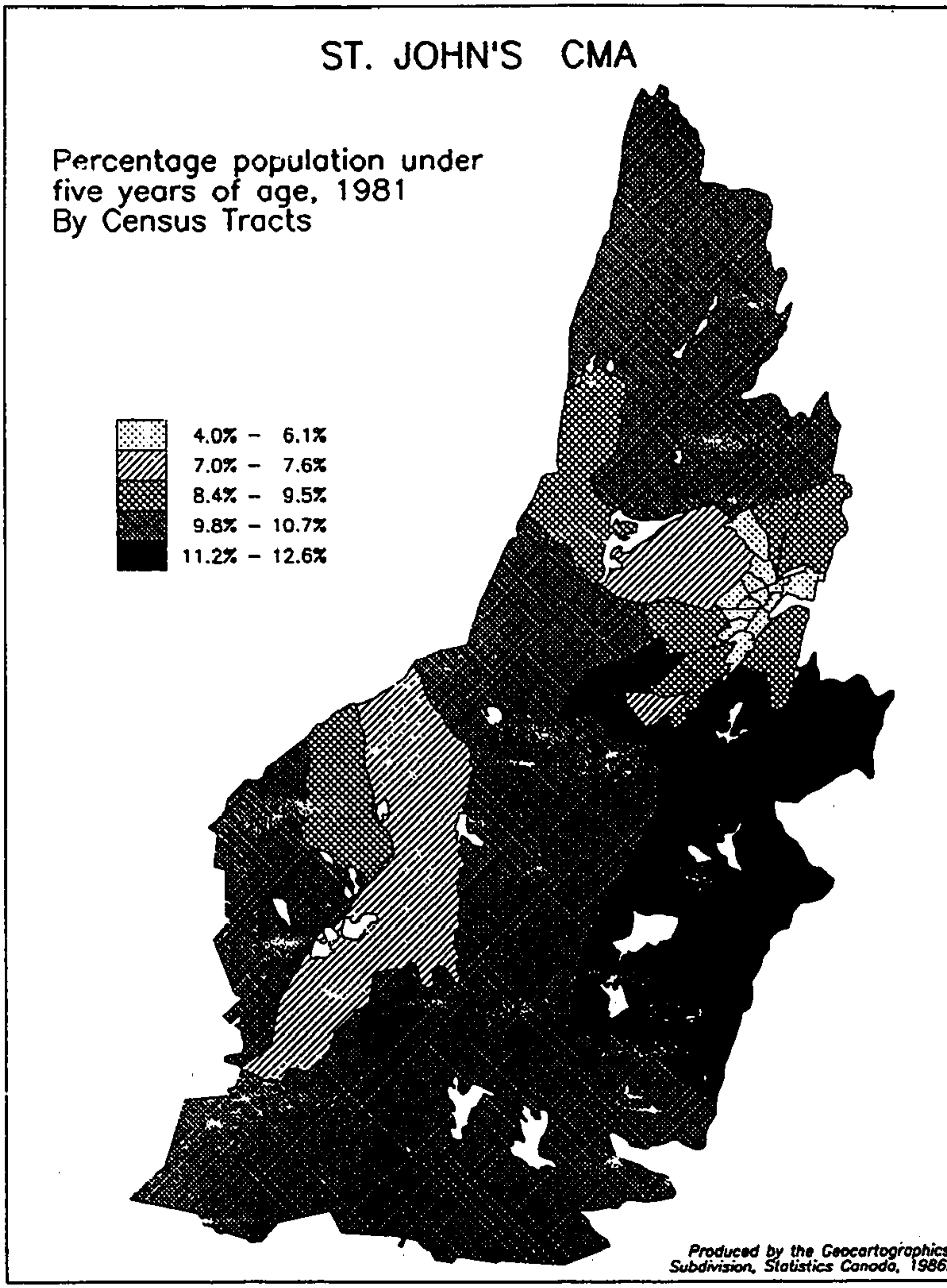
SAMPLE MAP OUTPUT



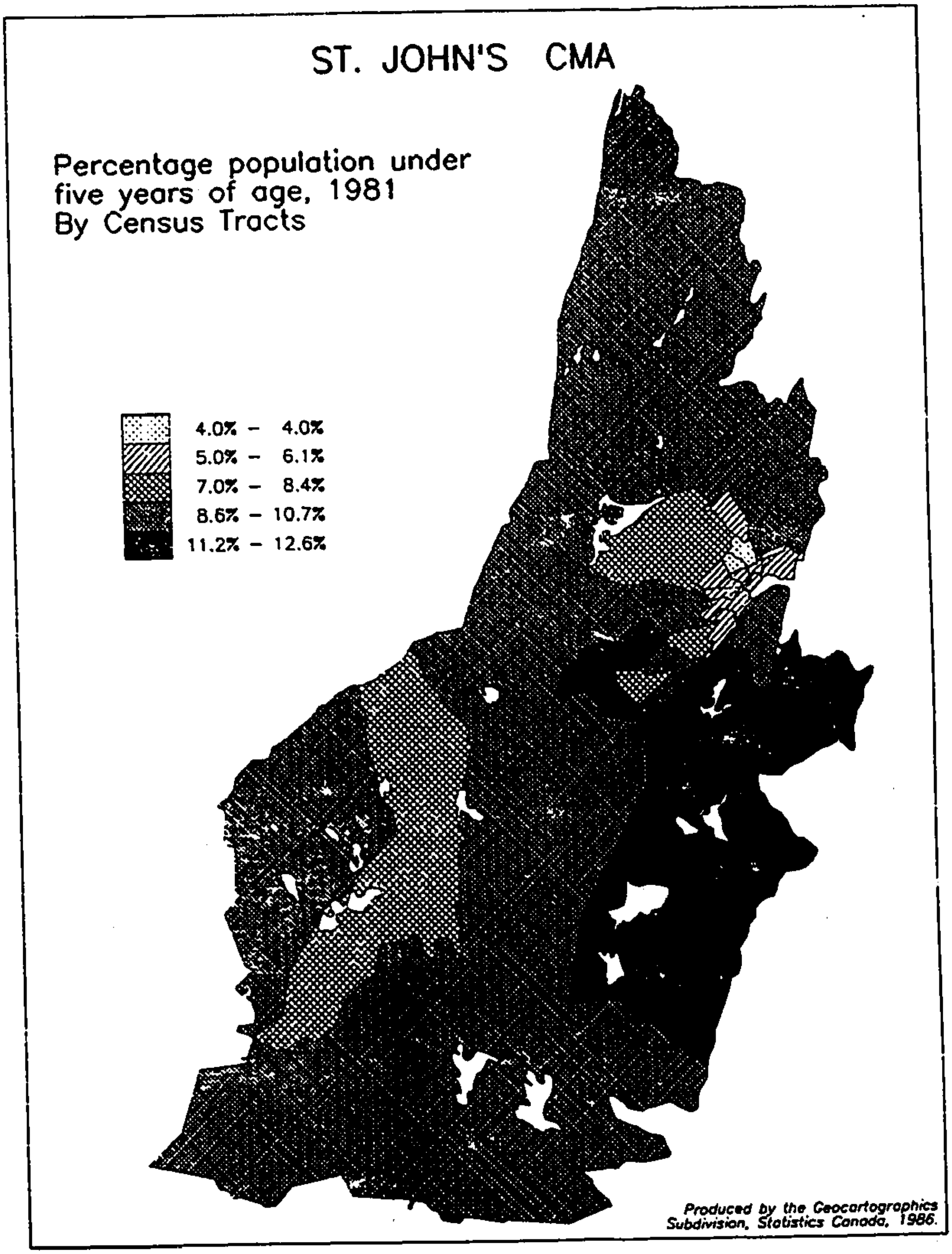
CHANG program results (geometric option) from bimodal data distribution.



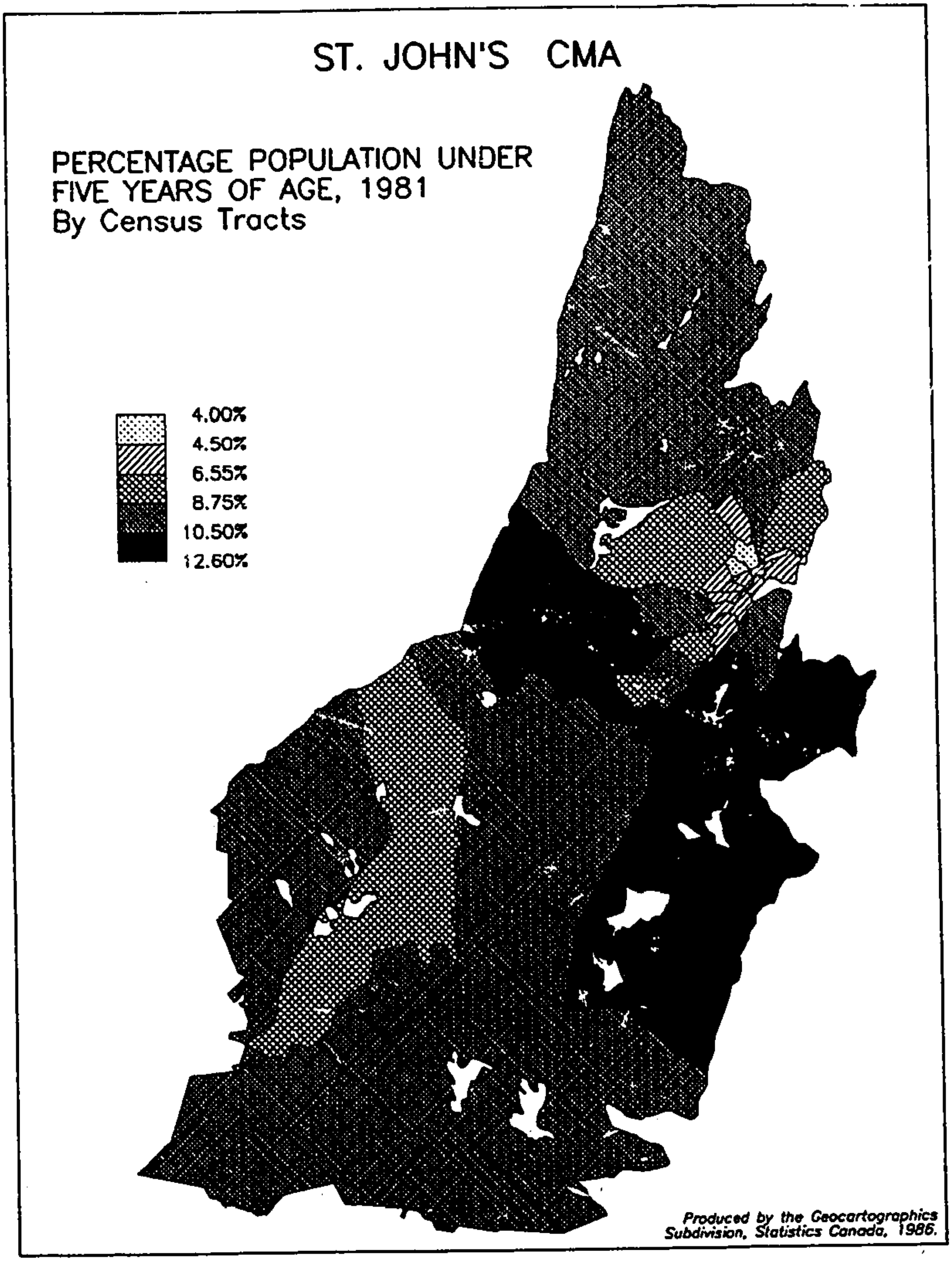
JENKS program results (absolute deviations option) from bimodal data distribution.



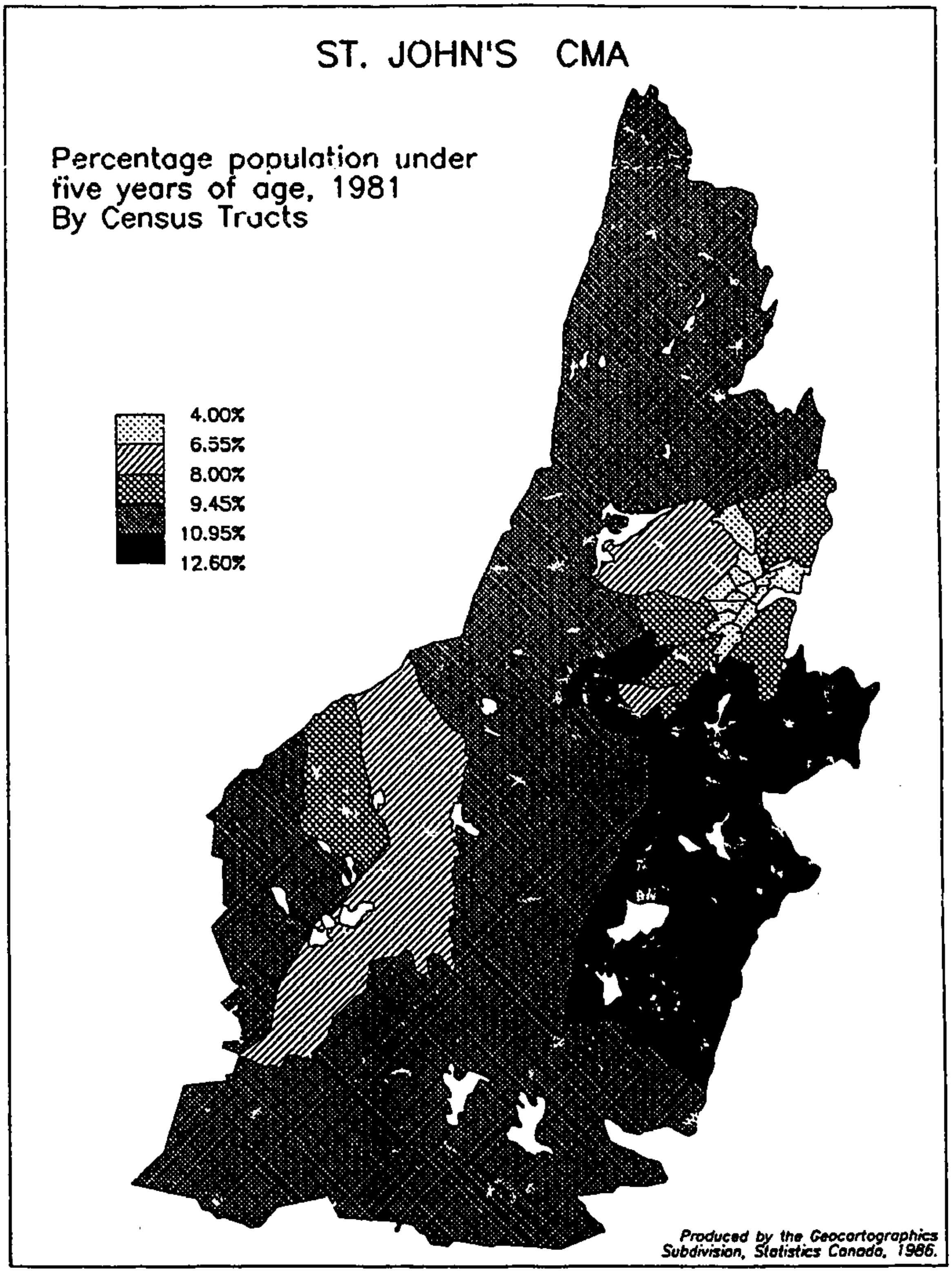
JENKS program results (variance option) from bimodal data distribution.



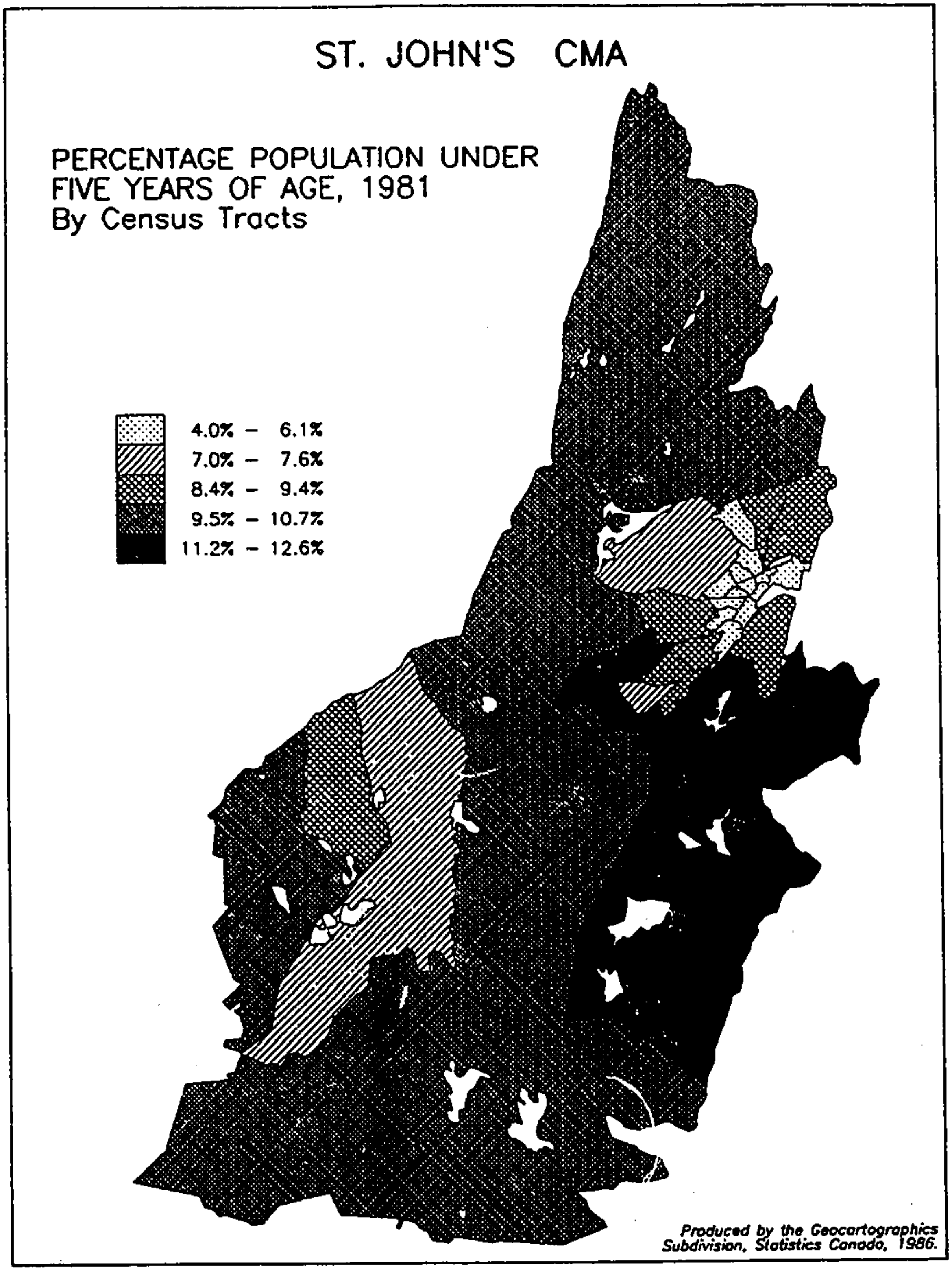
MOELLERING program results from bimodal data distribution.



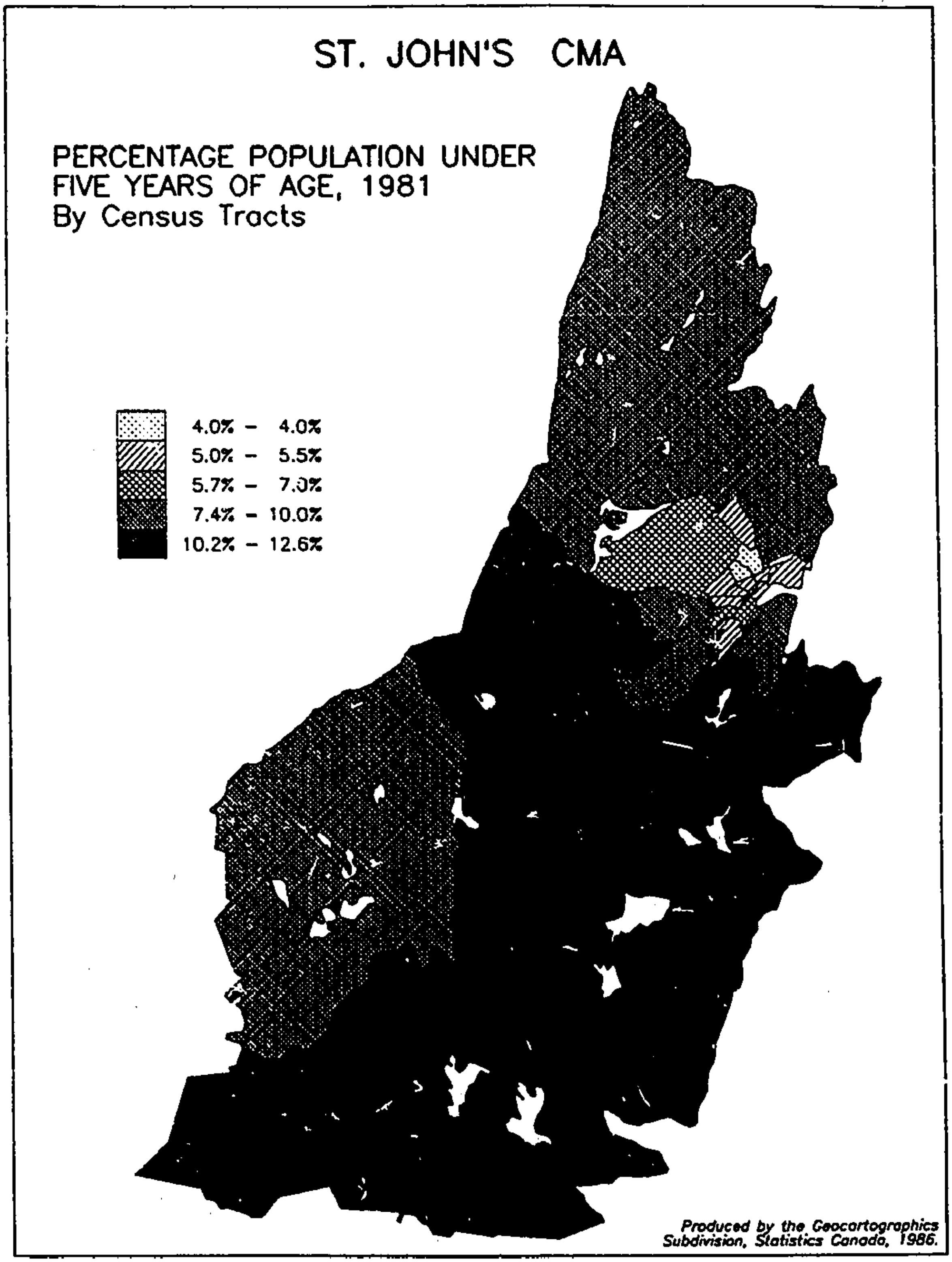
MONMONIER program results (absolute deviations option) from bimodal data distribution.



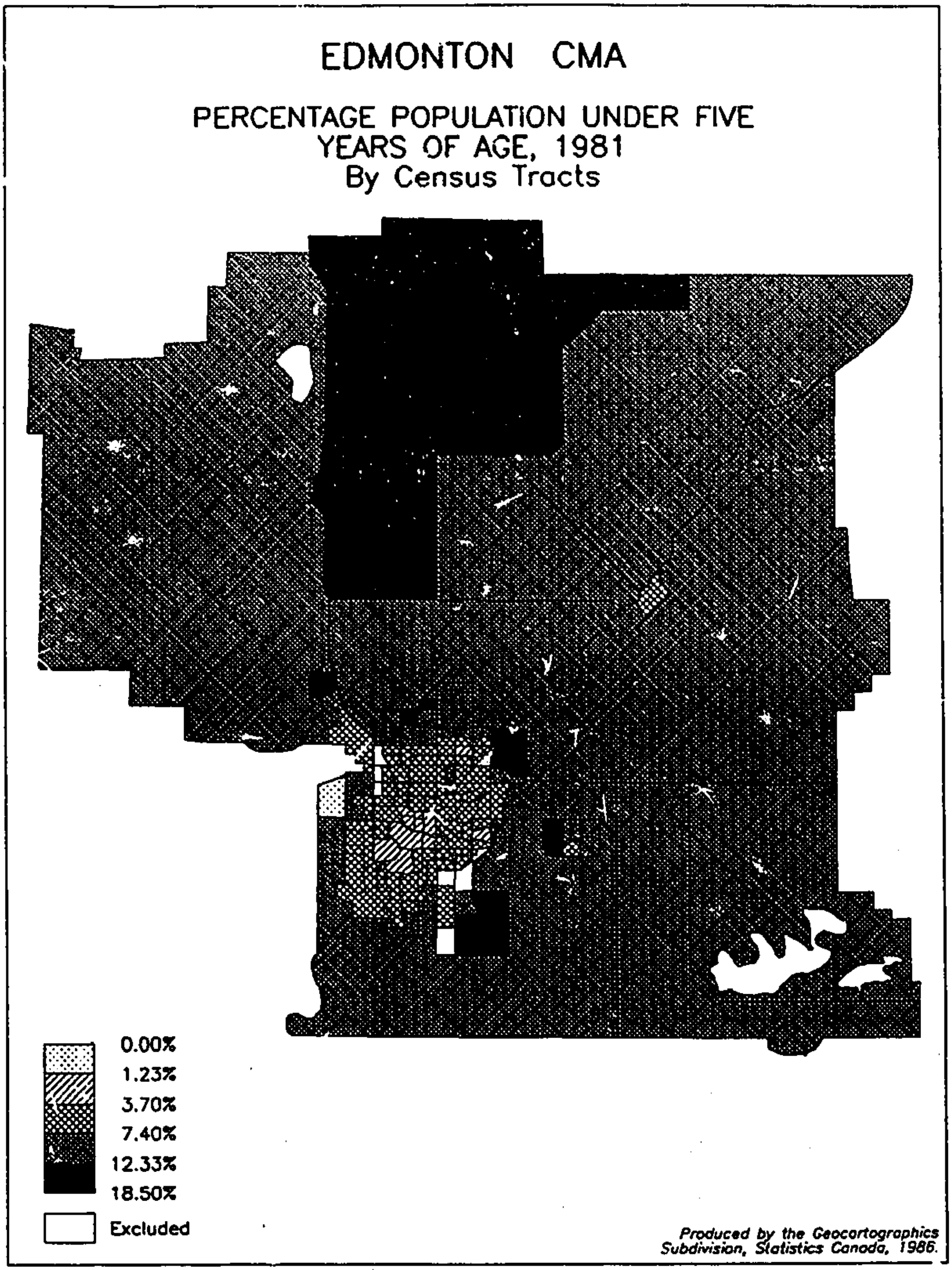
MONMONIER program results (squared deviations option) from bimodal data distribution.



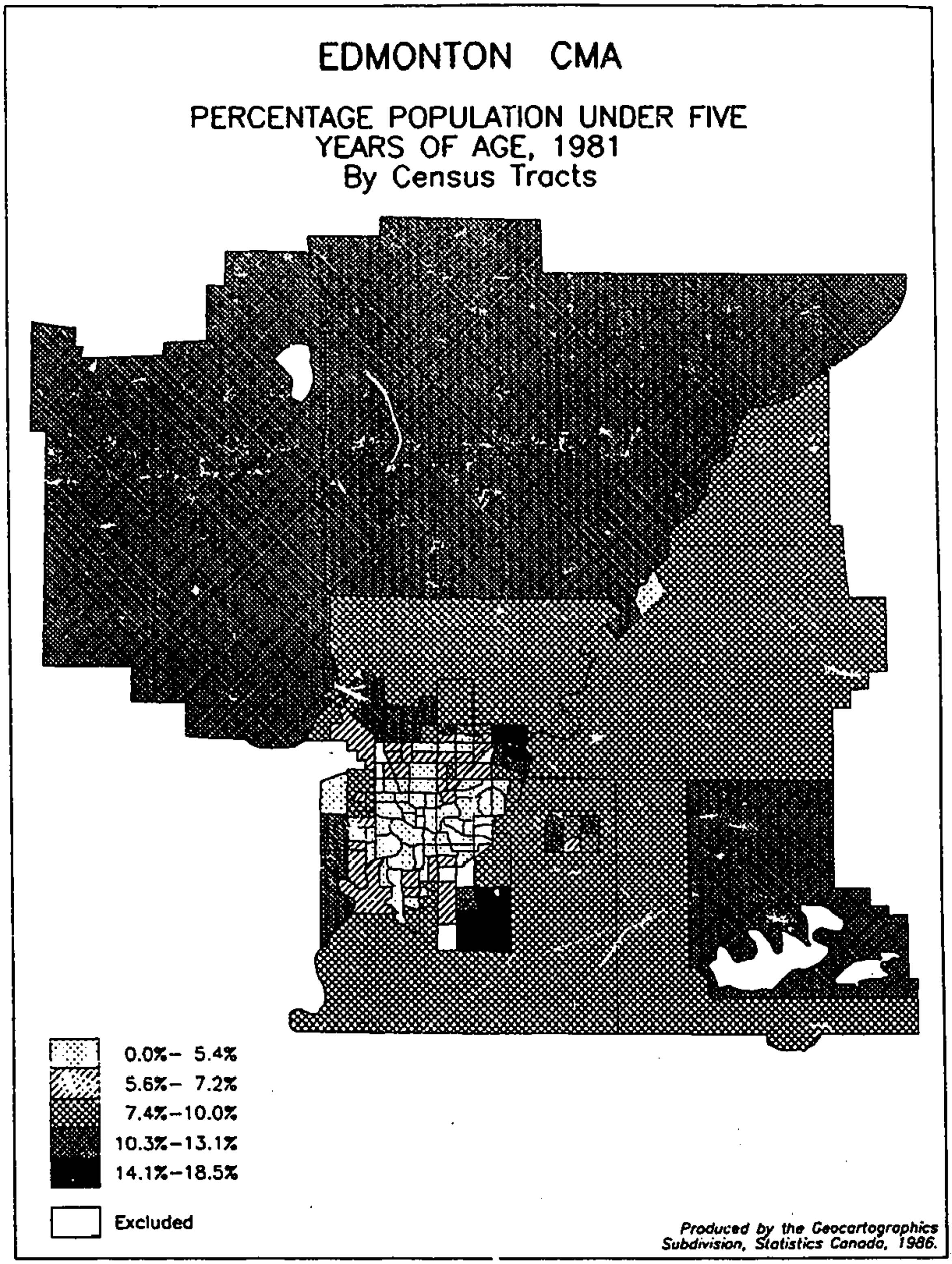
YOUNGMANN program results from bimodal data distribution.



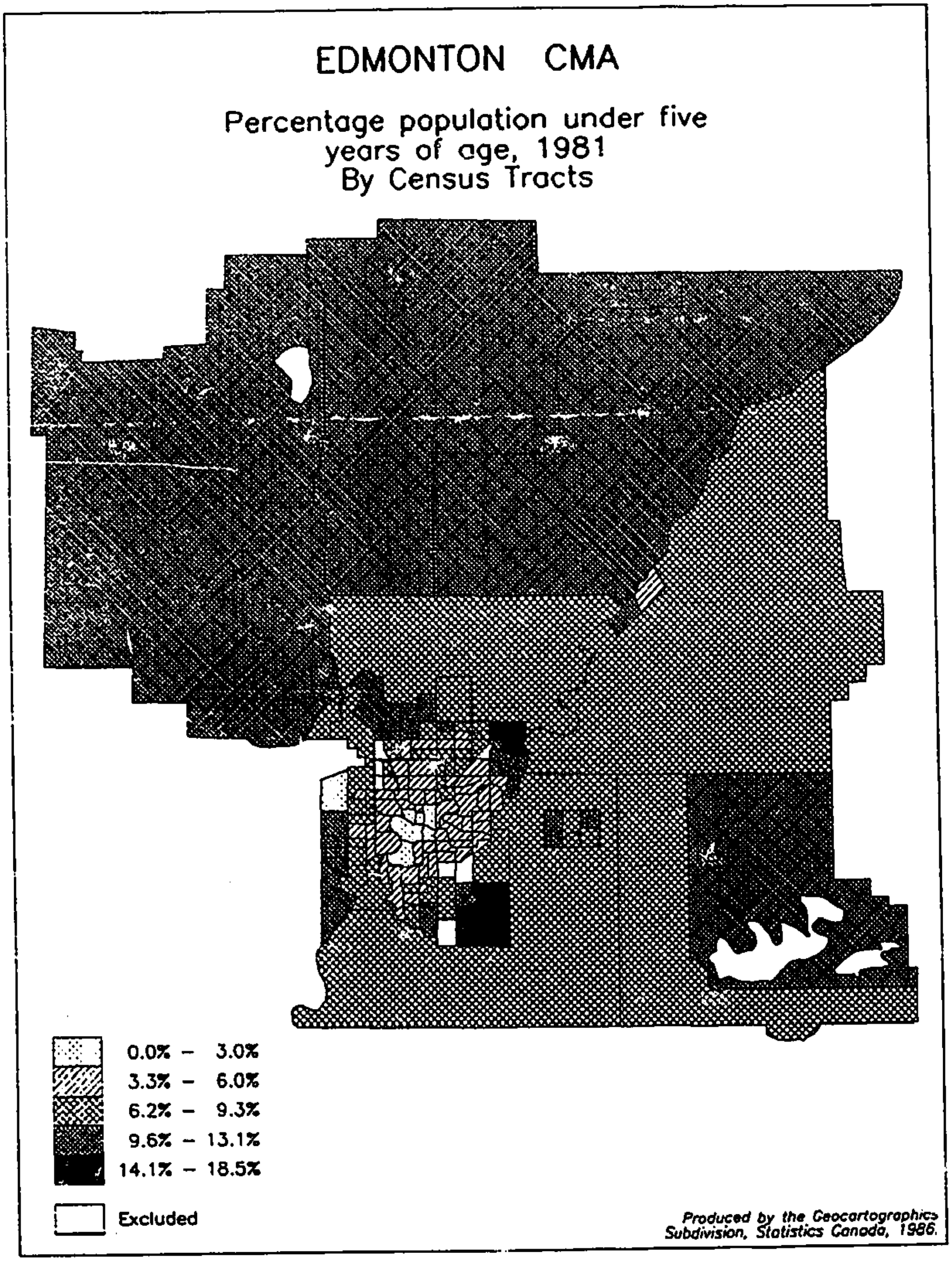
SAS program results from bimodal data distribution.



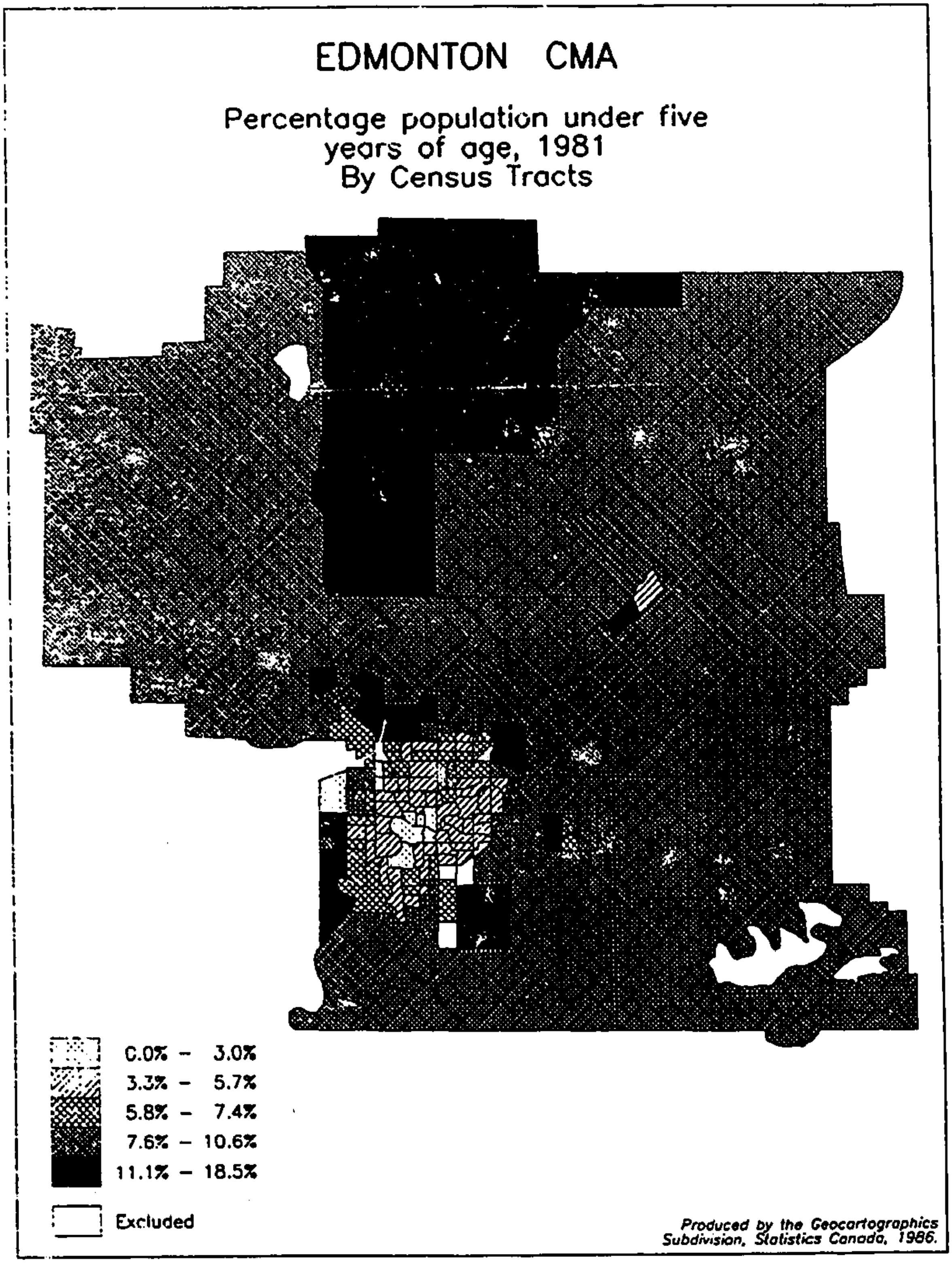
CHANG program results (arithmetic option) from right-skewed data distribution.



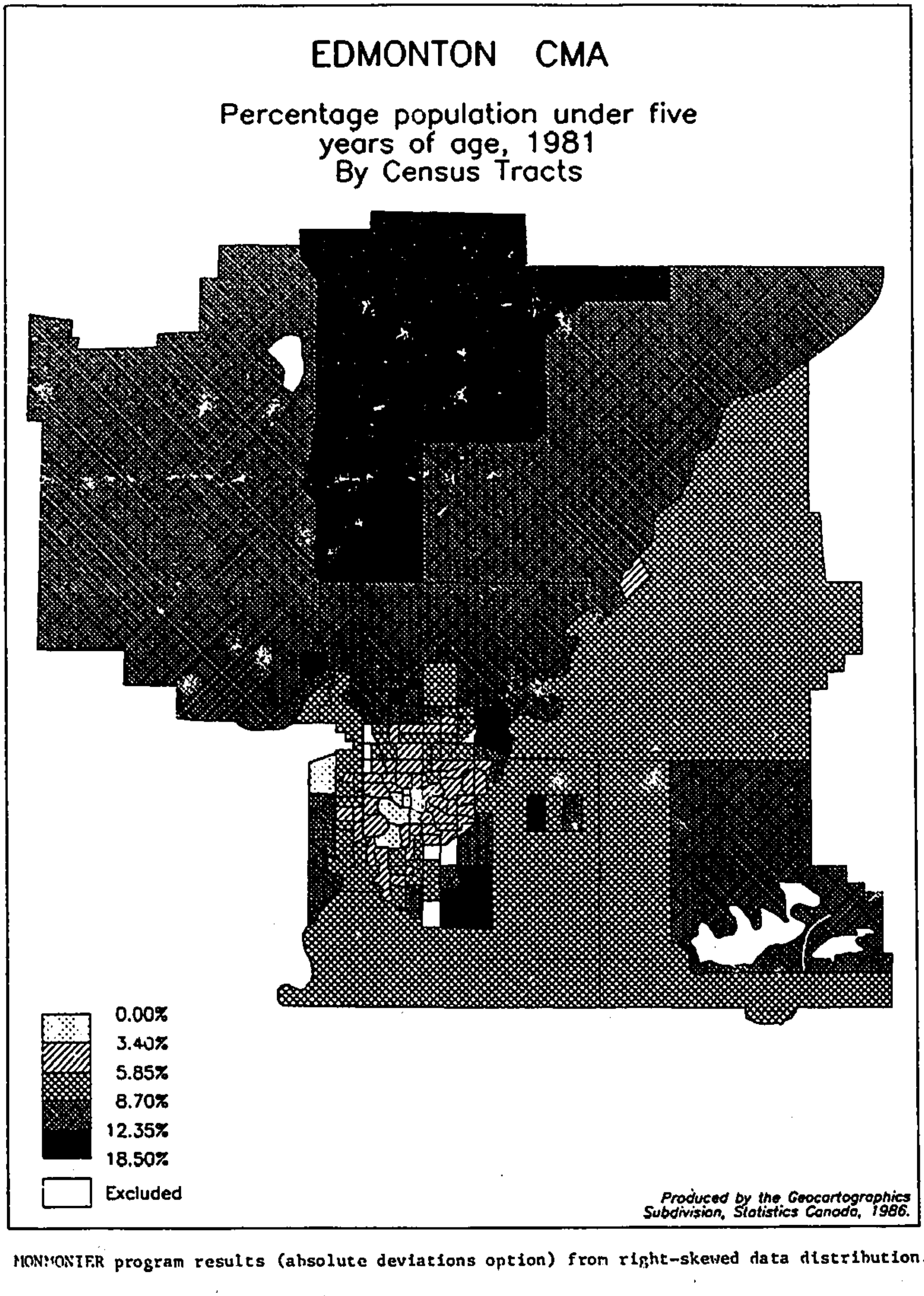
JENKS program results (absolute deviations option) from right-skewed data distribution.



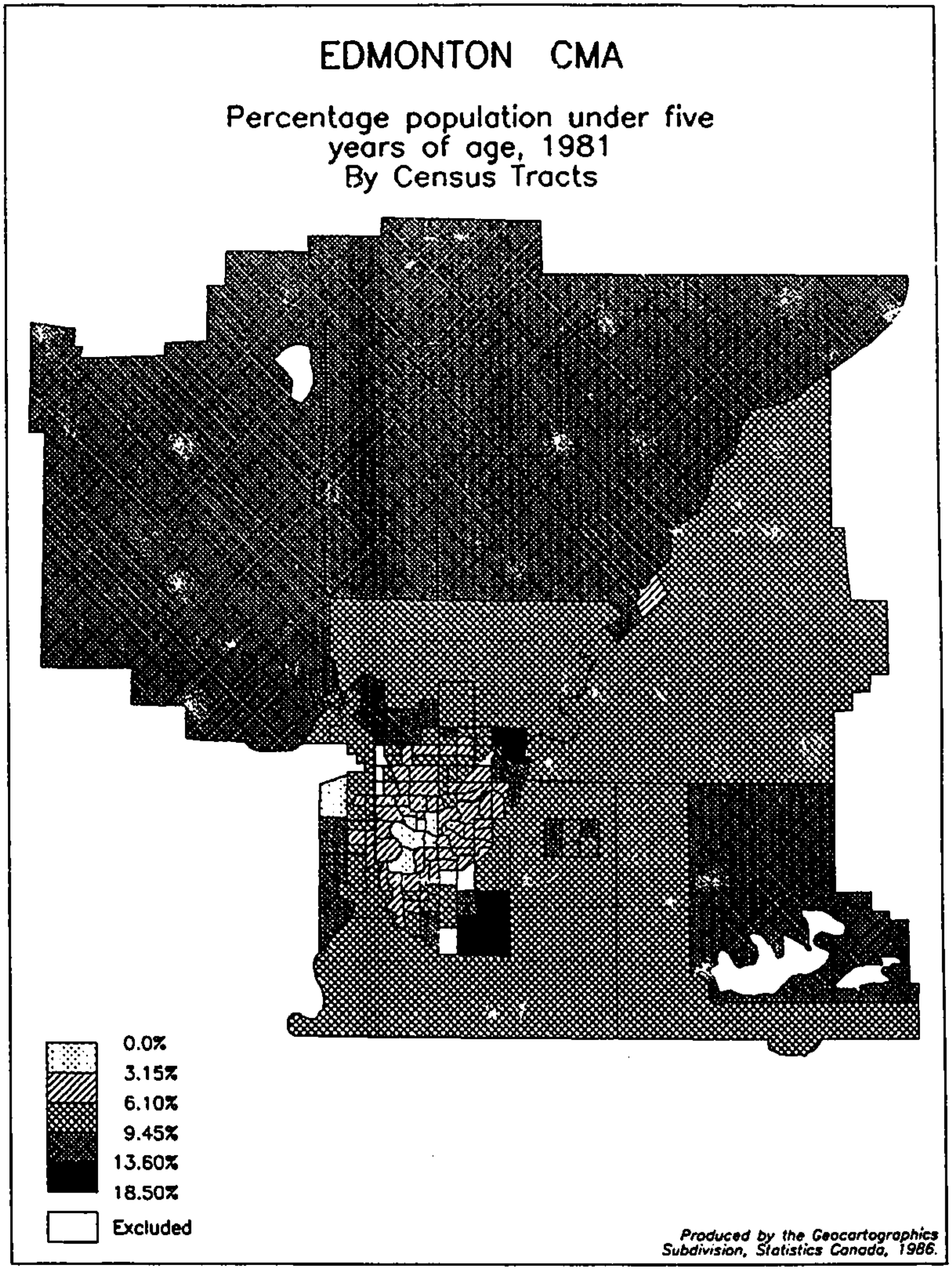
JENKS program results (variance option) from right-skewed data distribution.



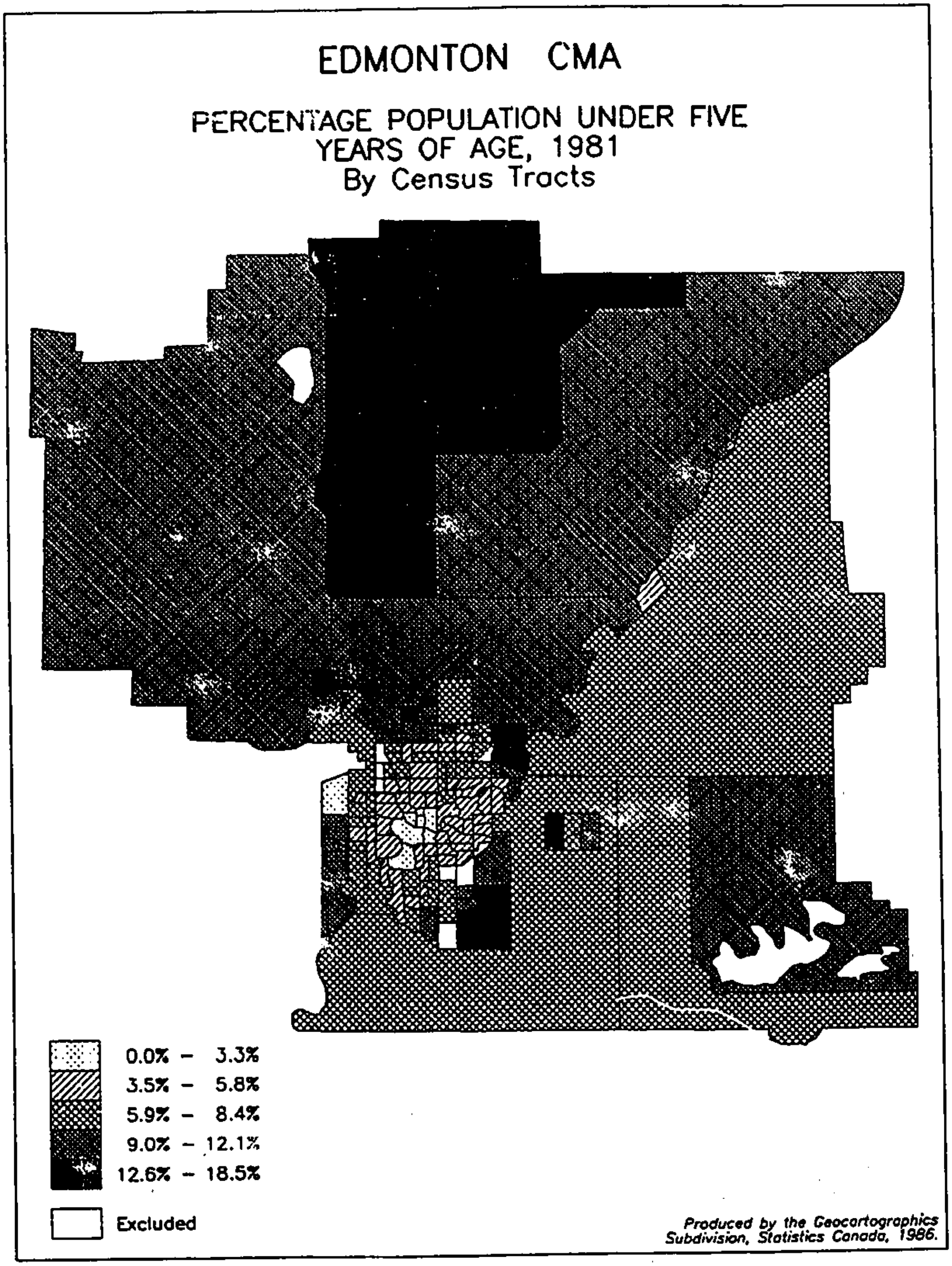
MOELLERING program results from right-skewed data distribution.



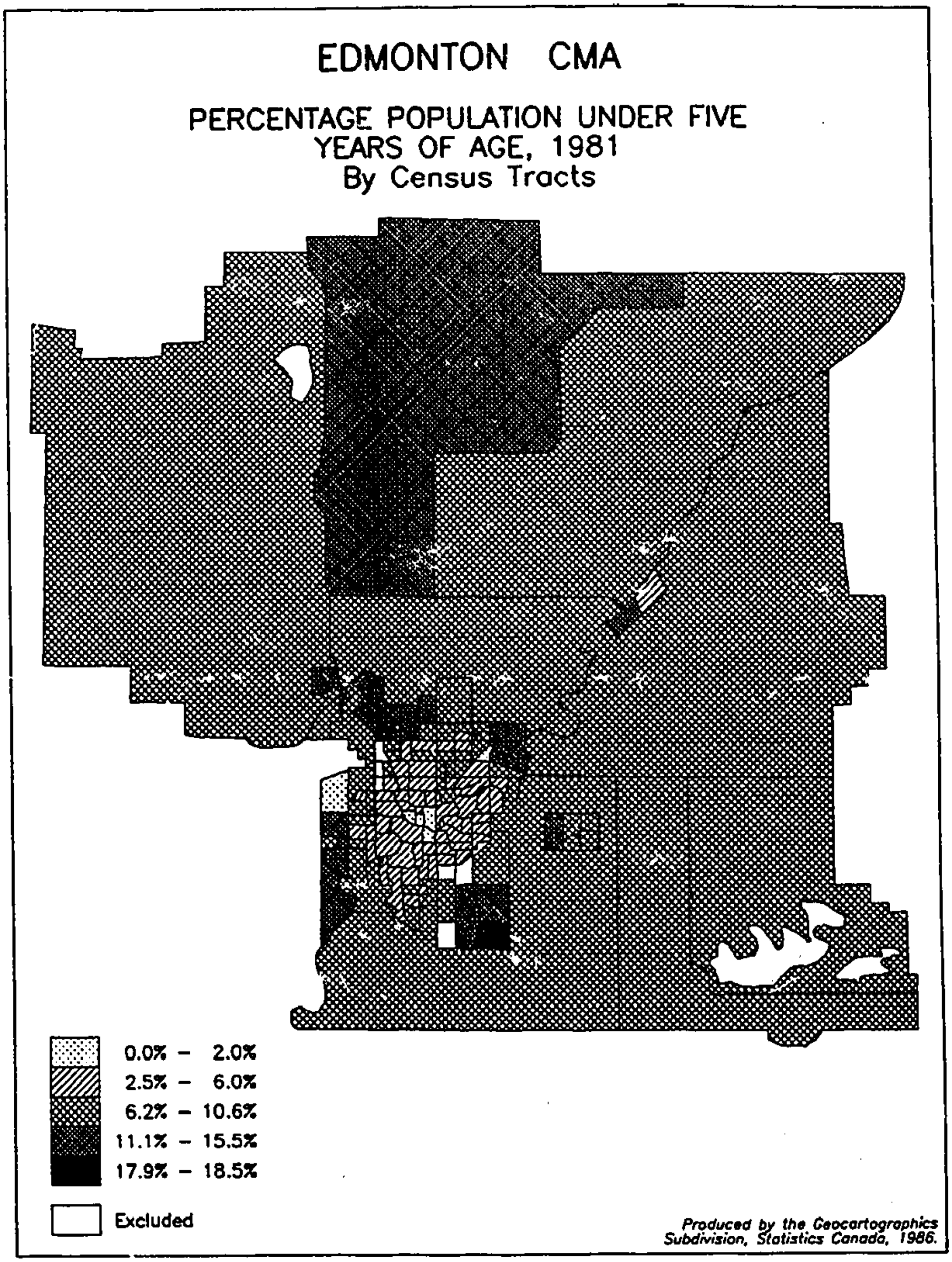
MONMONIER program results (absolute deviations option) from right-skewed data distribution.



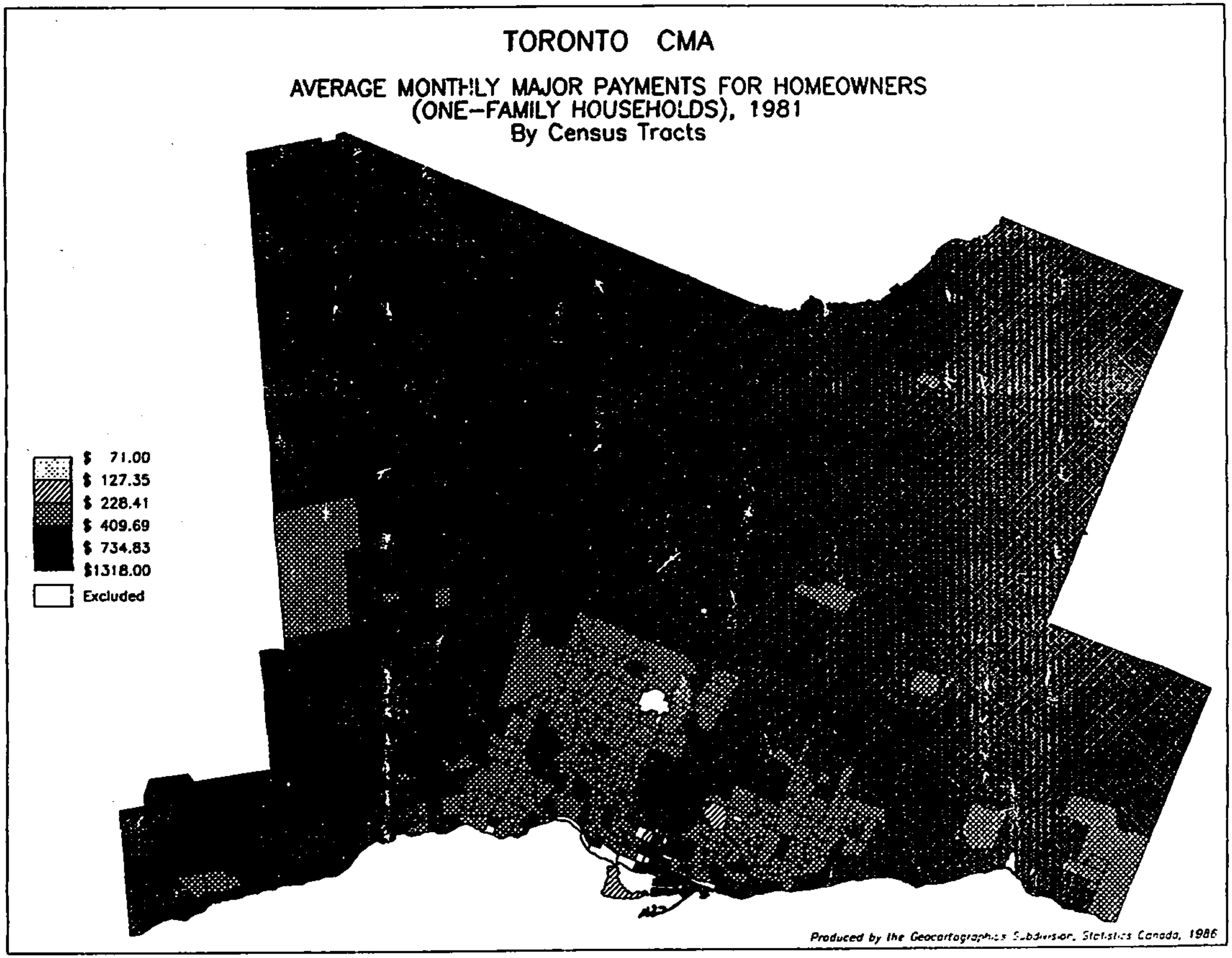
MONMONIER program results (squared deviations option) from right-skewed data distribution.



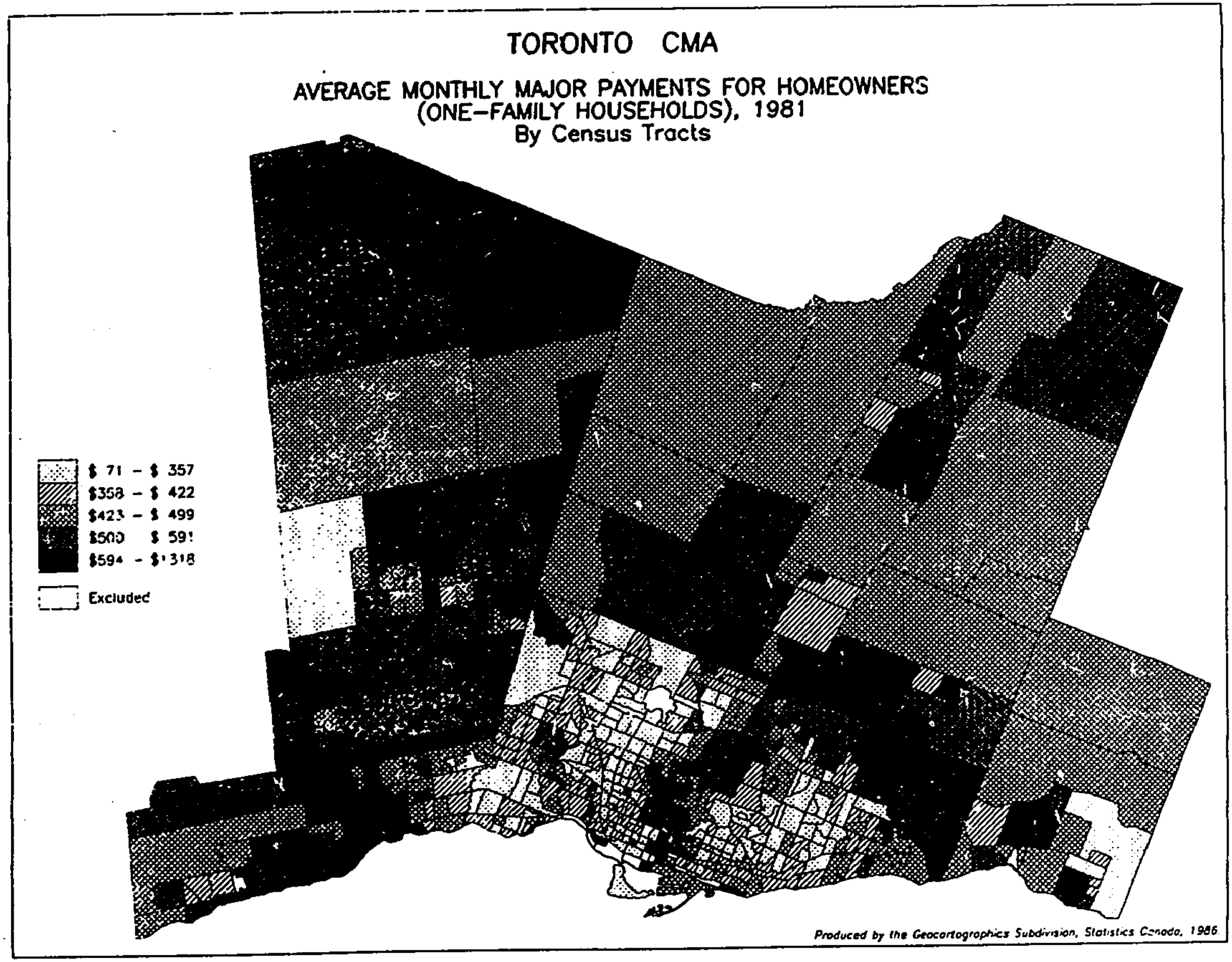
YOUNGMANN program results from right-skewed data distribution.



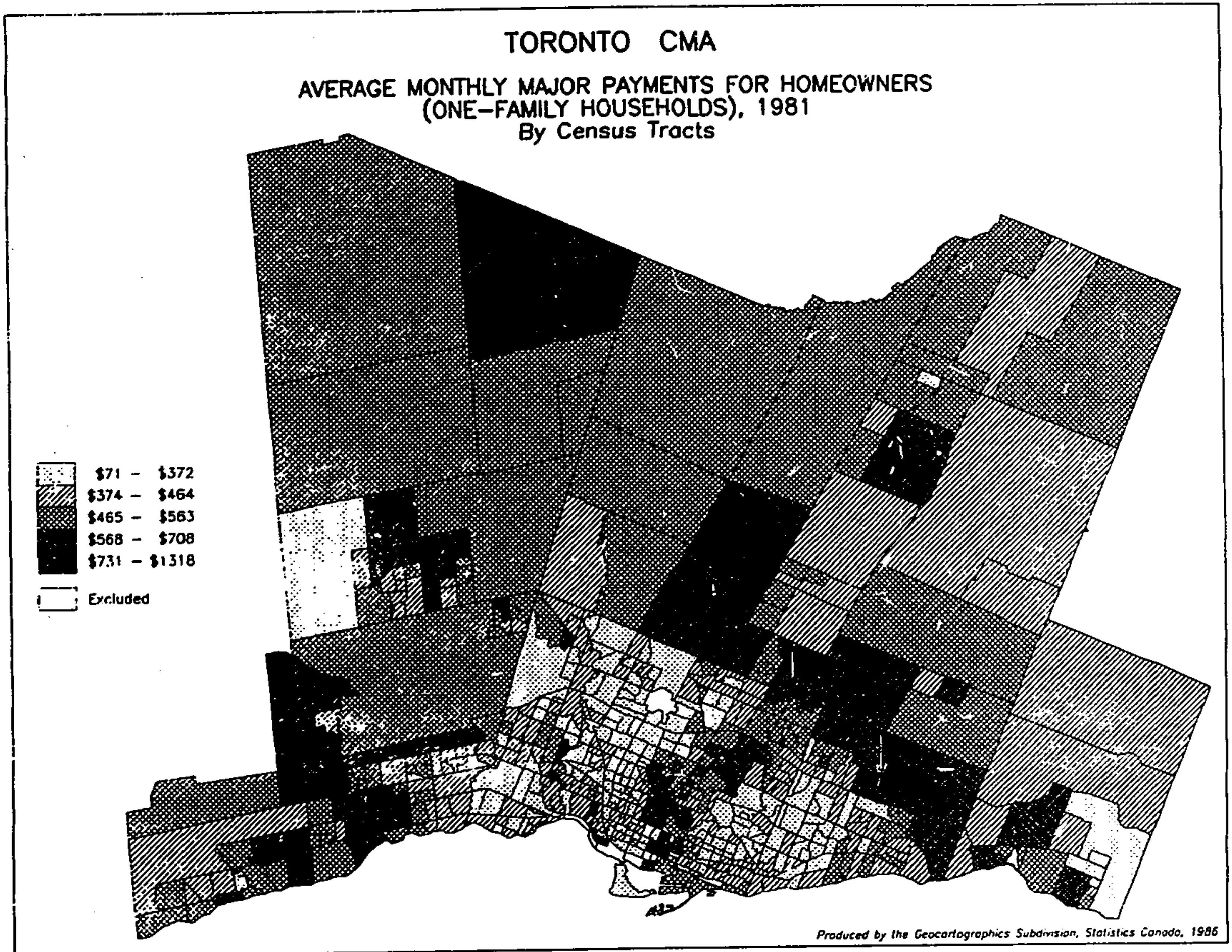
SAS program results from right-skewed data distribution.



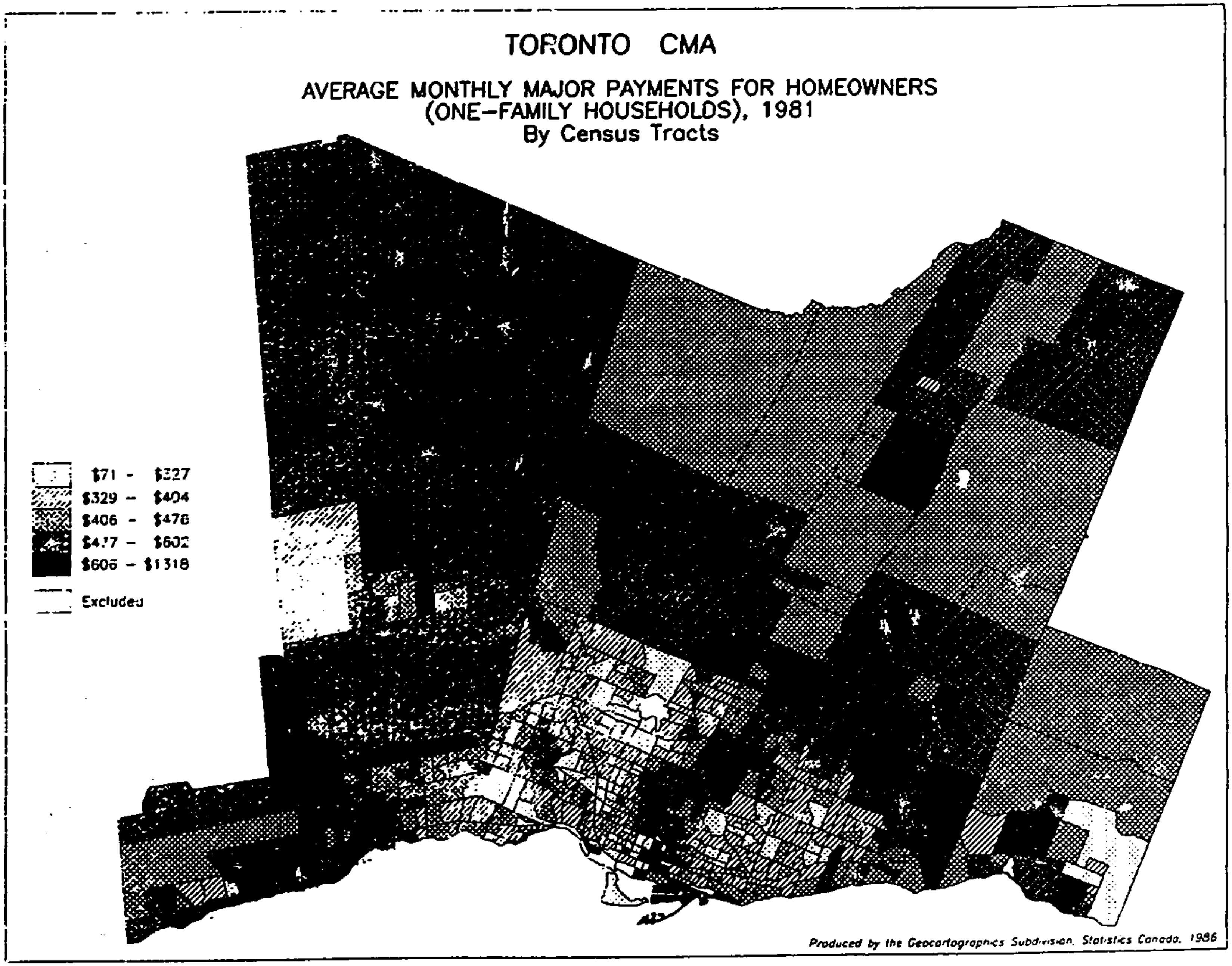
CHANG program results (geometric option) from right-skewed data distribution.



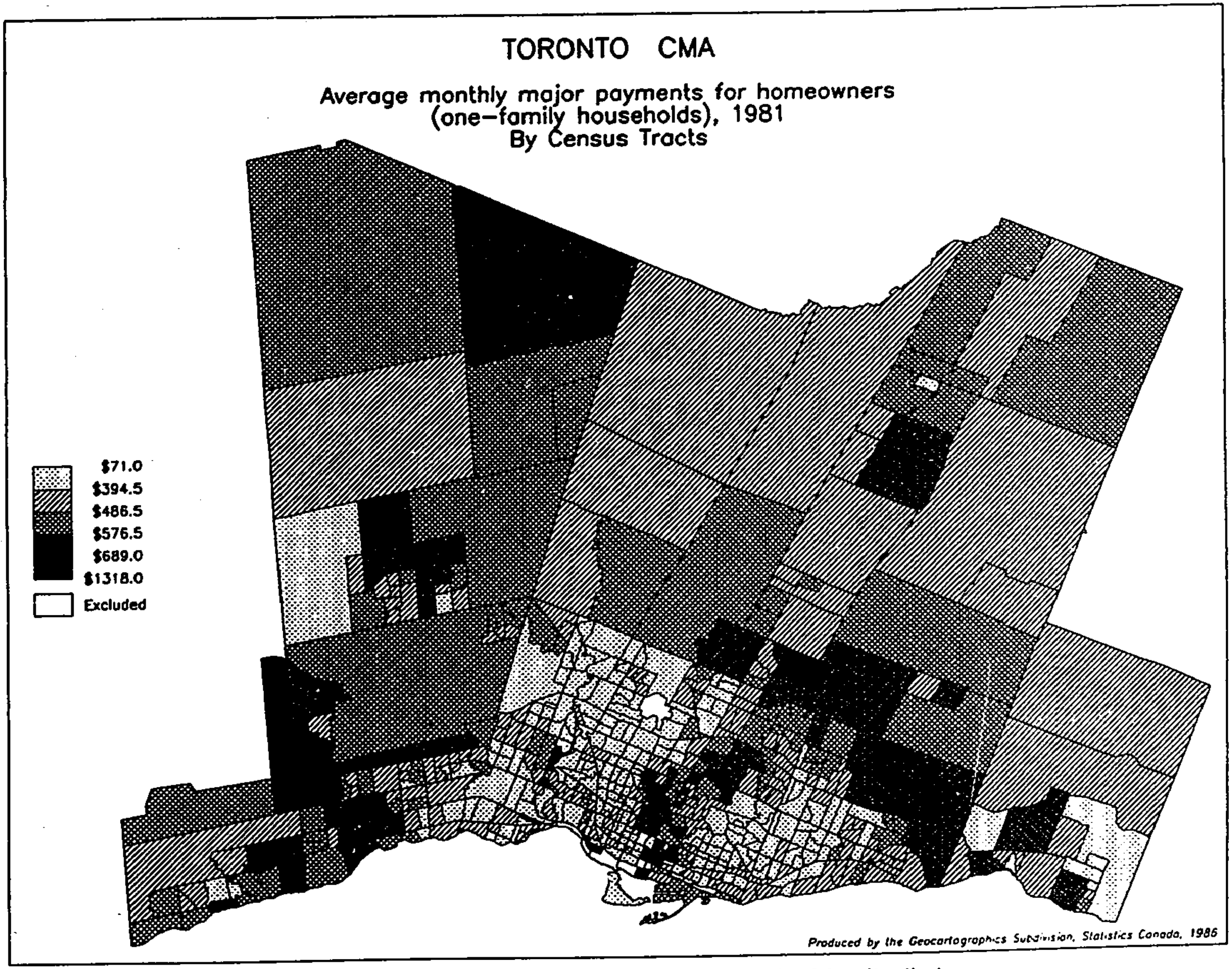
JENKS program results (absolute deviations option) from right-skewed data distribution.



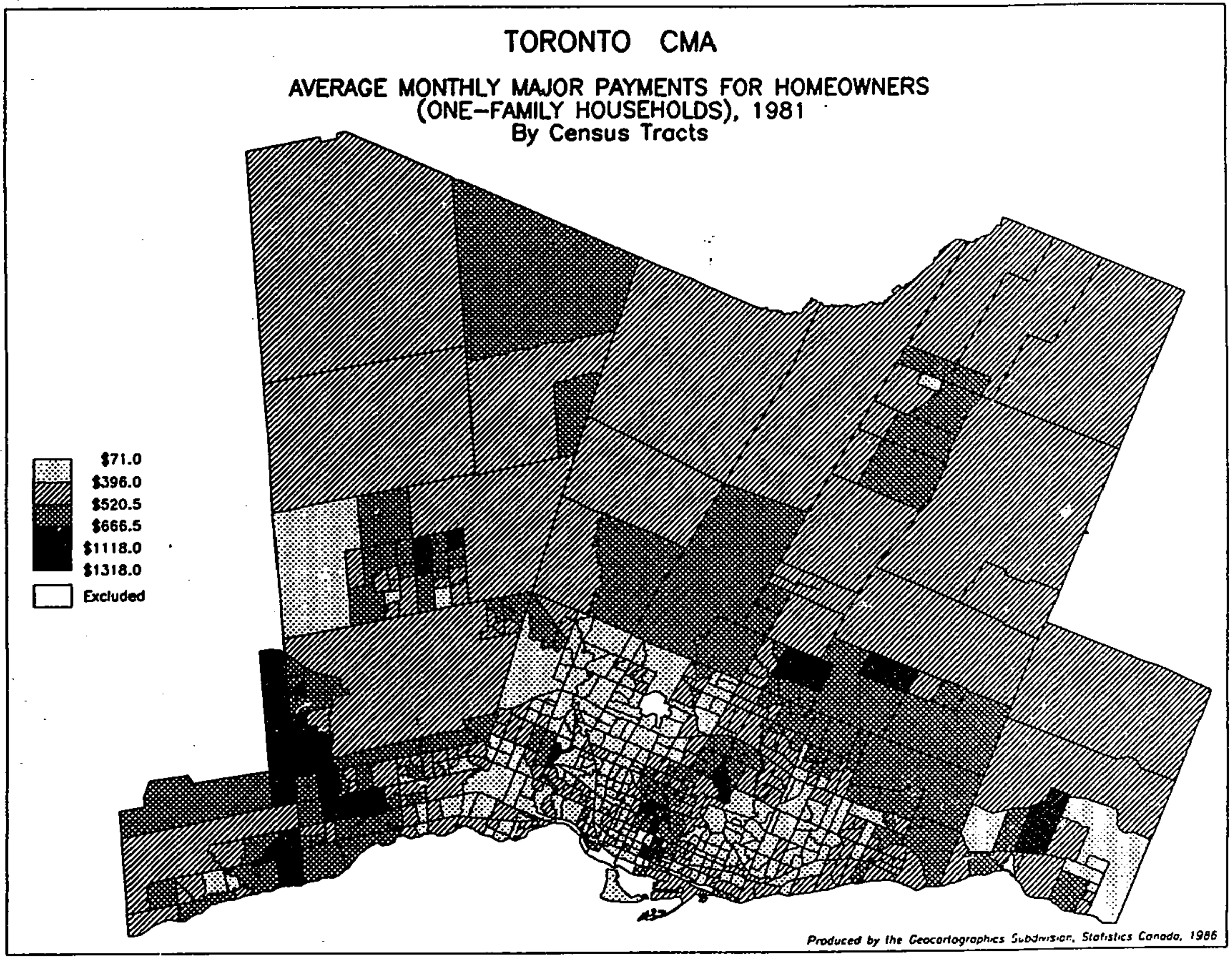
JENKS program results (variance option) from right-skewed data distribution.



MOELLERING program results from right-skewed data distribution.



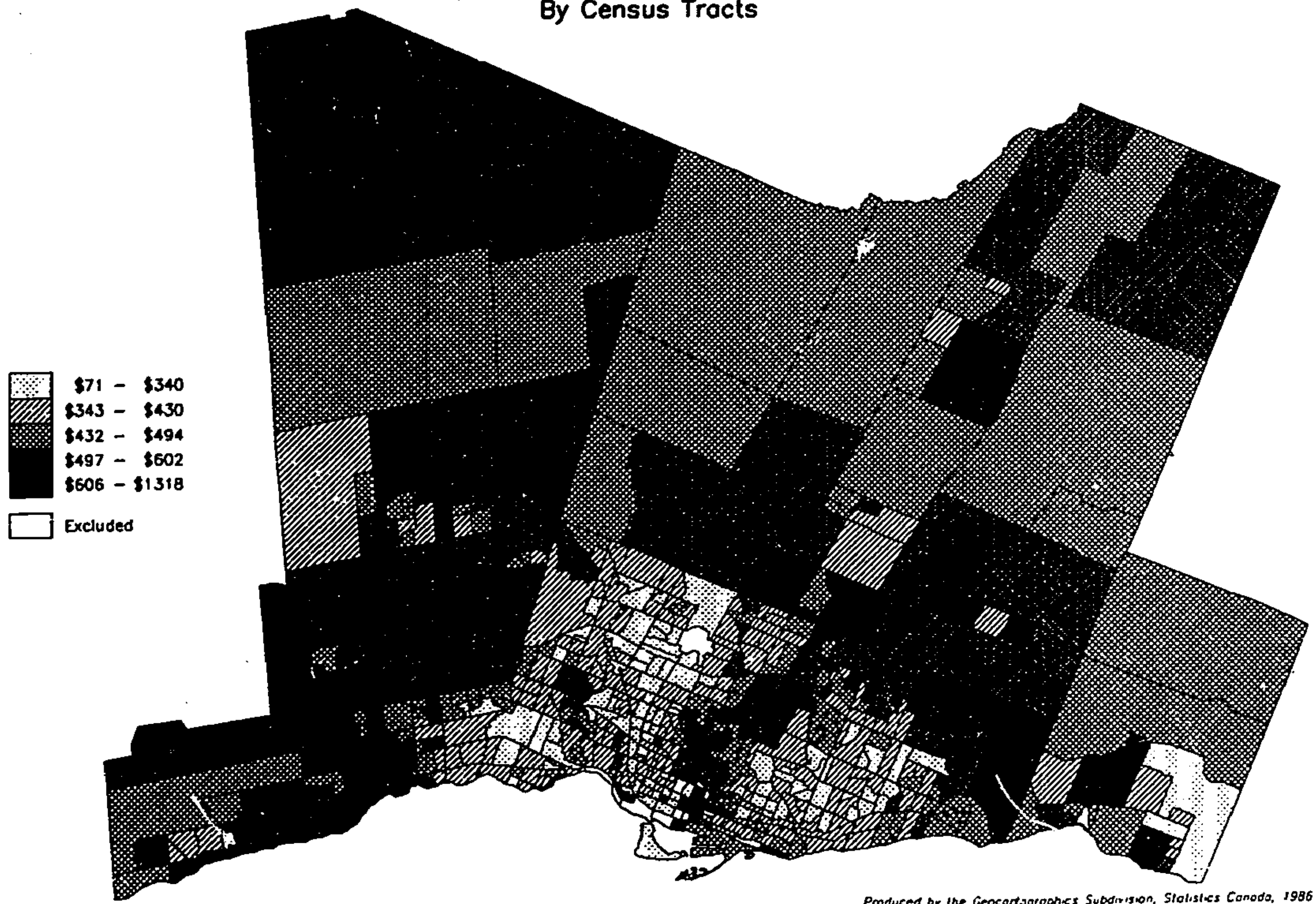
MONMONIER program results (absolute deviations option) from right-skewed data distribution.



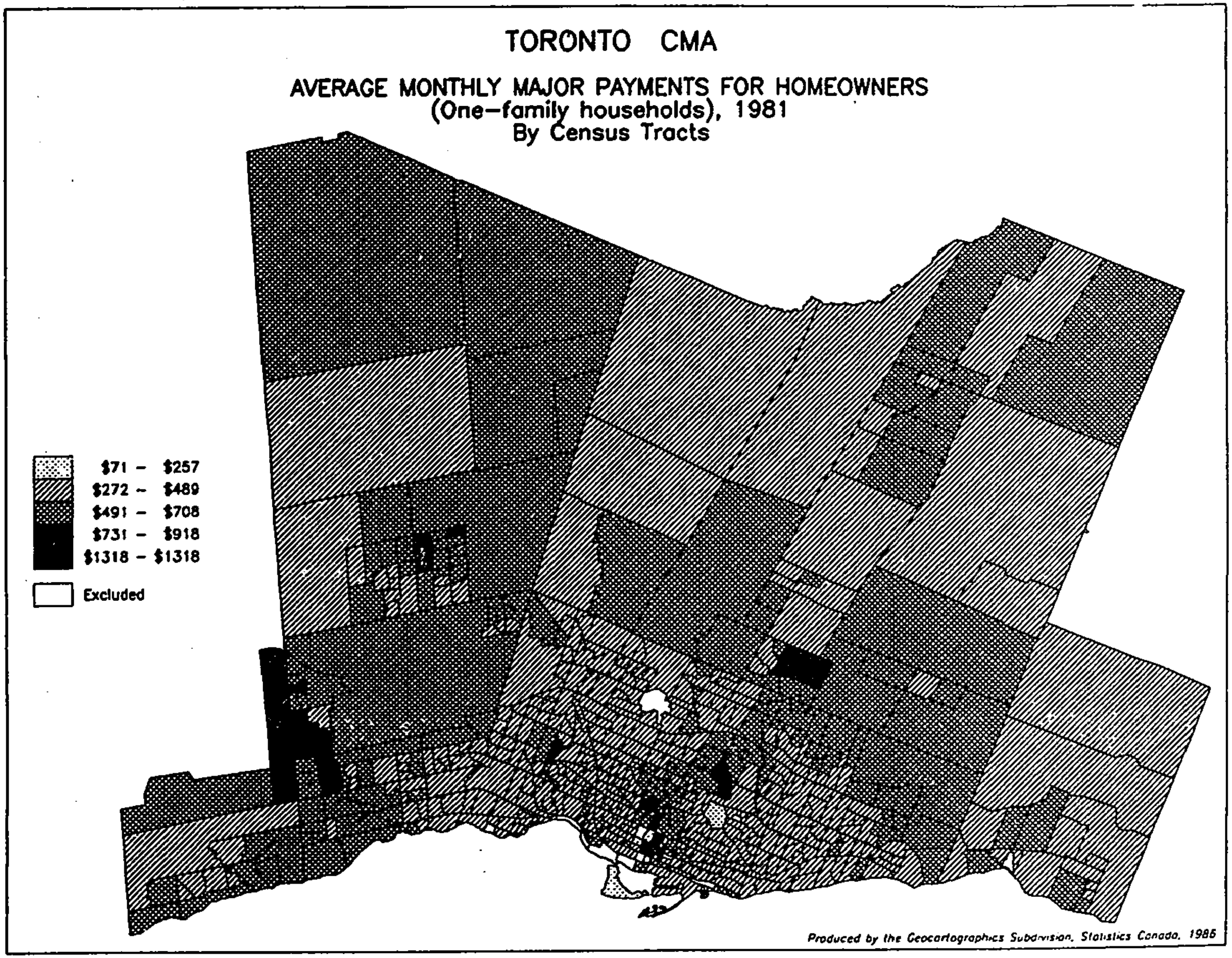
MONMONIER program results (squared deviations option) from right-skewed data distribution.

TORONTO CMA

AVERAGE MONTHLY MAJOR PAYMENTS FOR HOMEOWNERS
(One-family households), 1981
By Census Tracts



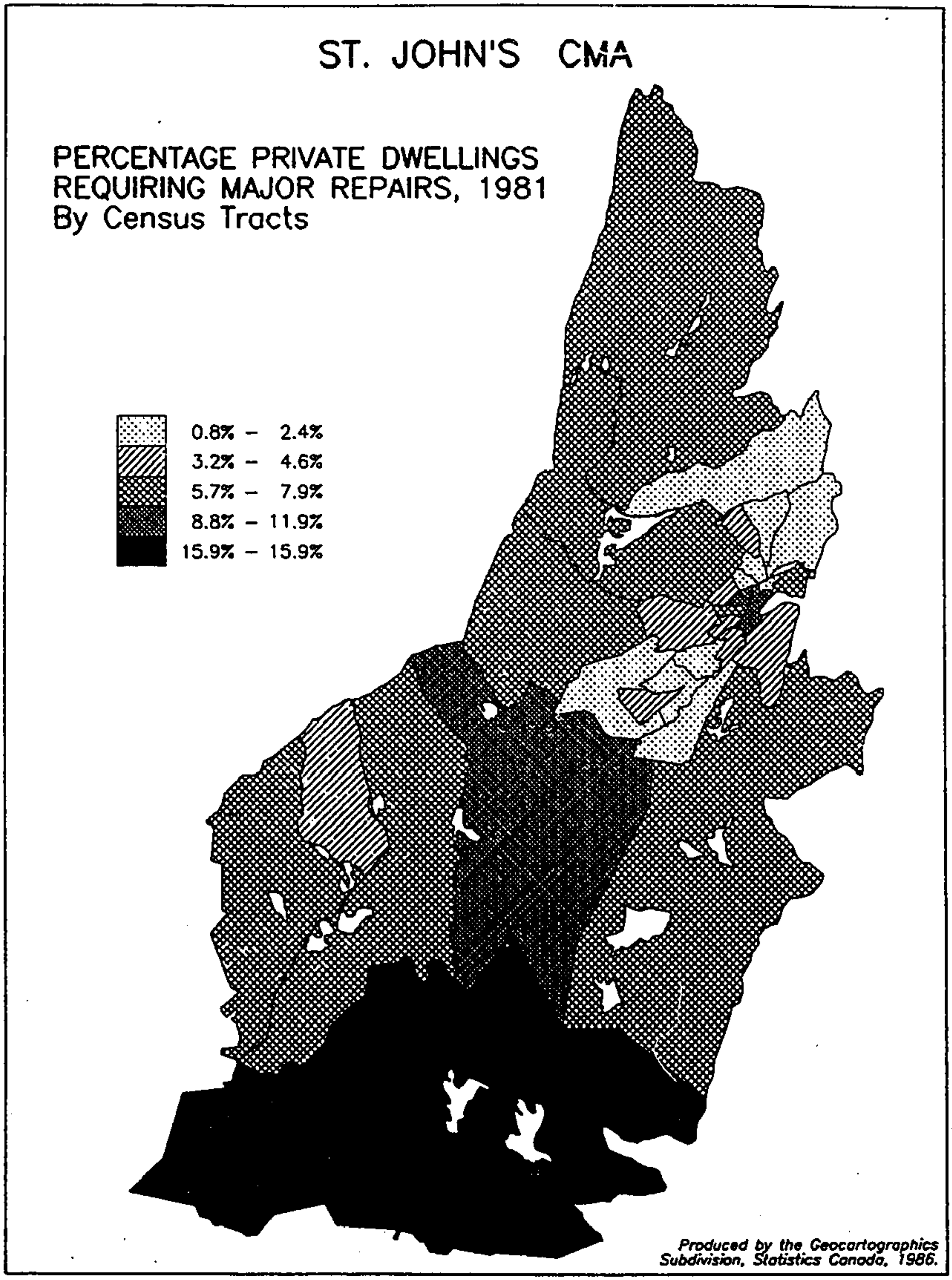
YOINGMANN program results from right-skewed data distribution.



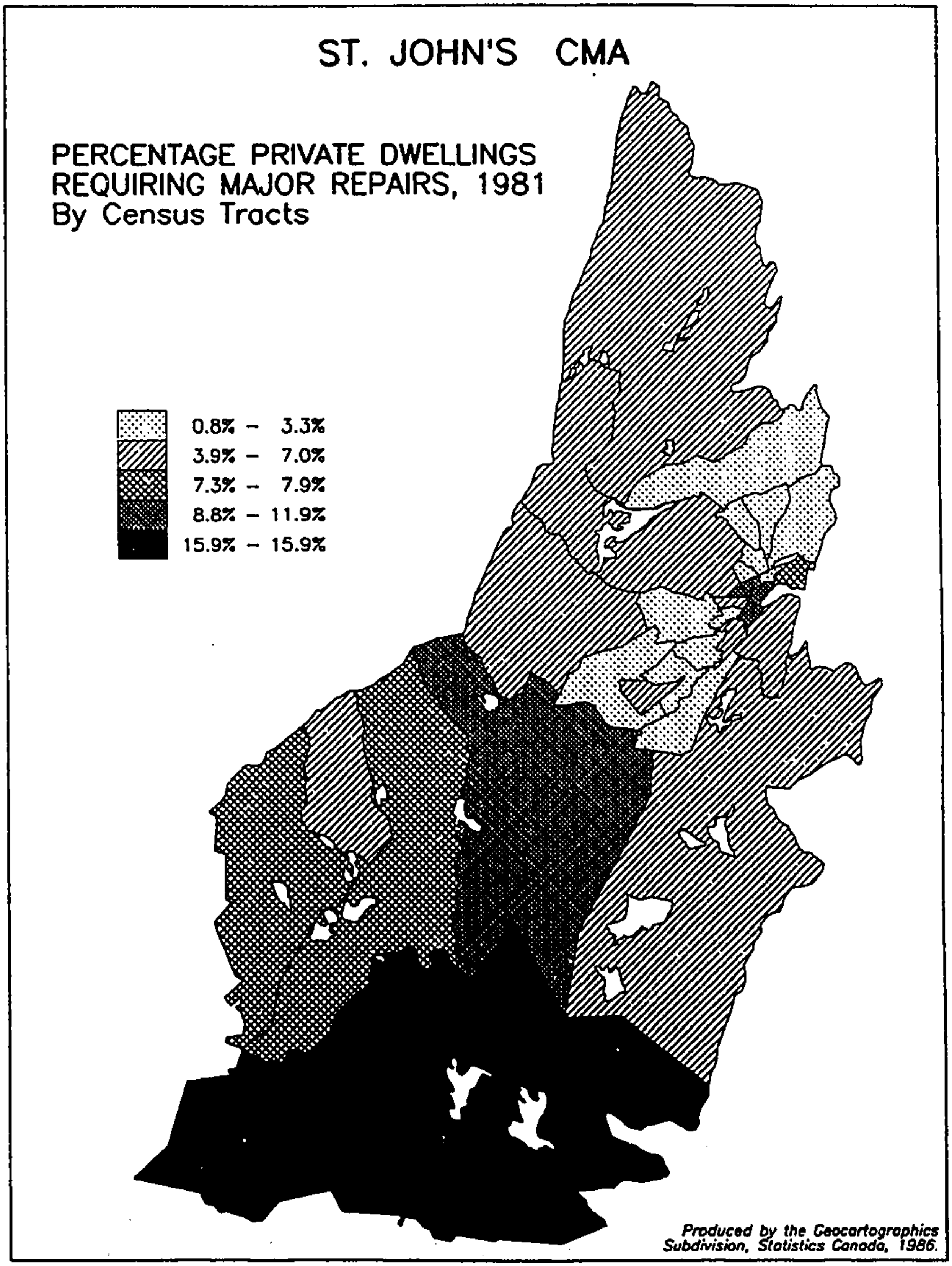
SAS program results from right-skewed data distribution.

Appendix 7

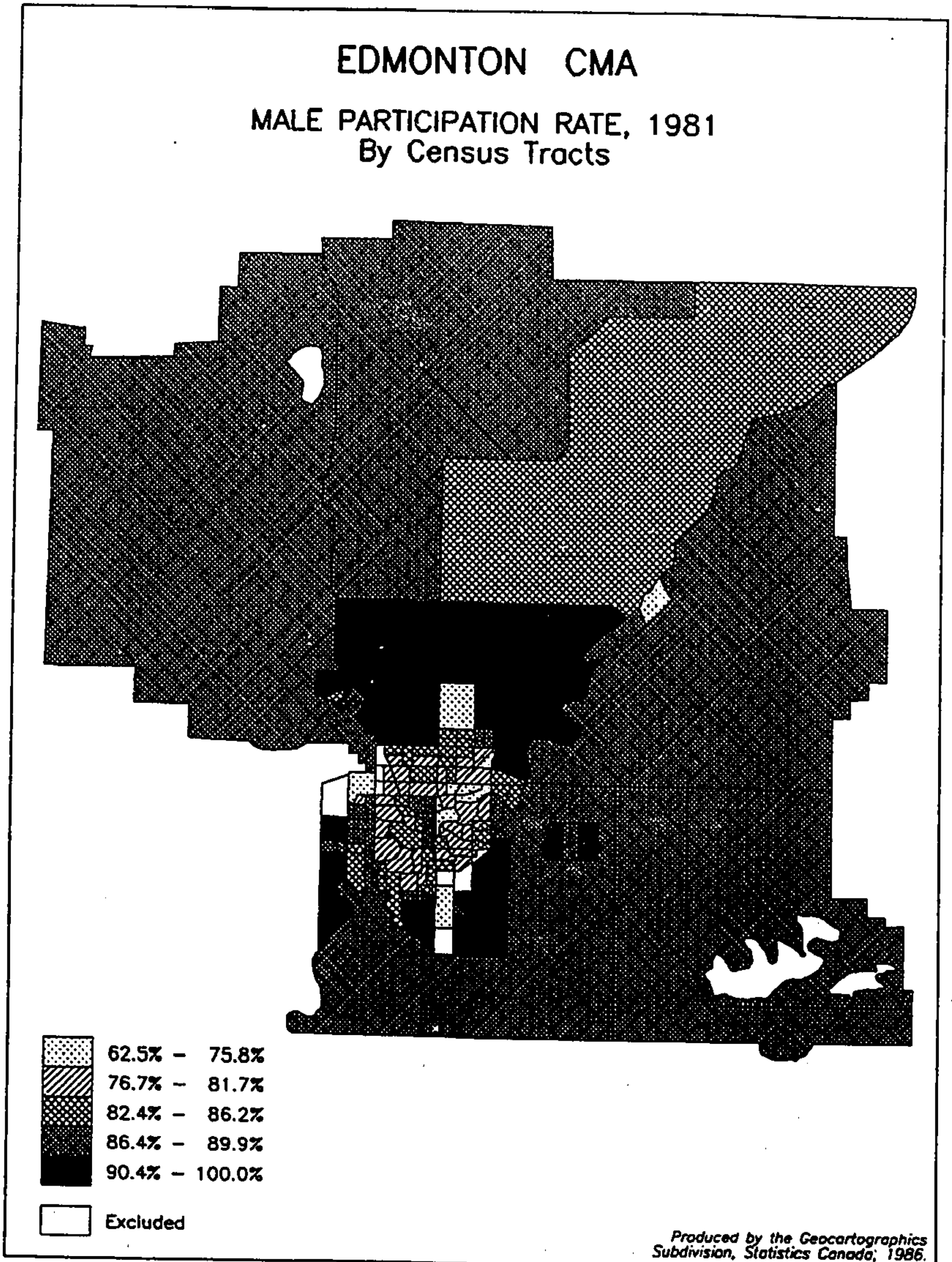
YOUNGMANN'S AREA OPTION



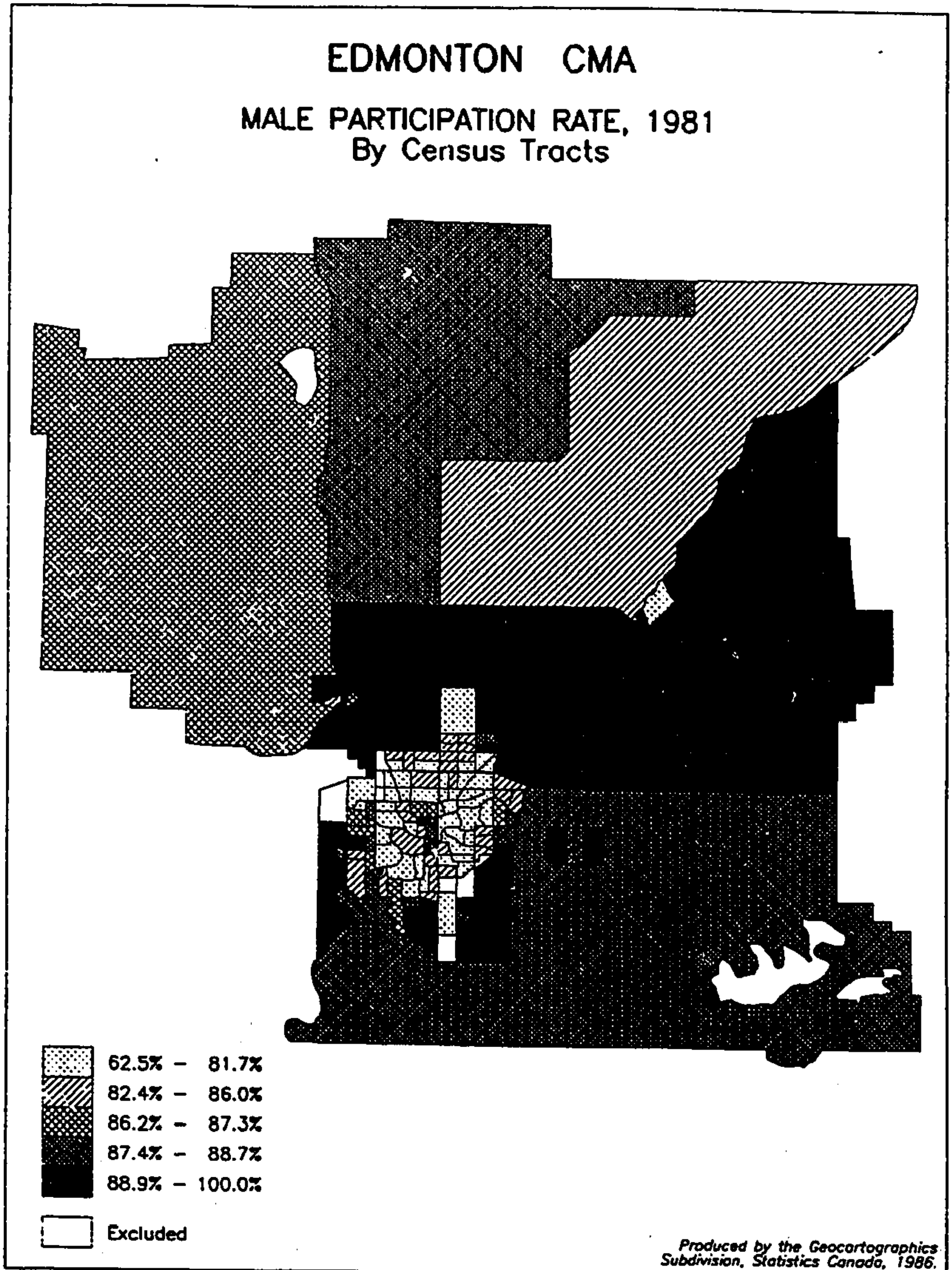
YOUNGMANN program results from L-shaped data distribution.



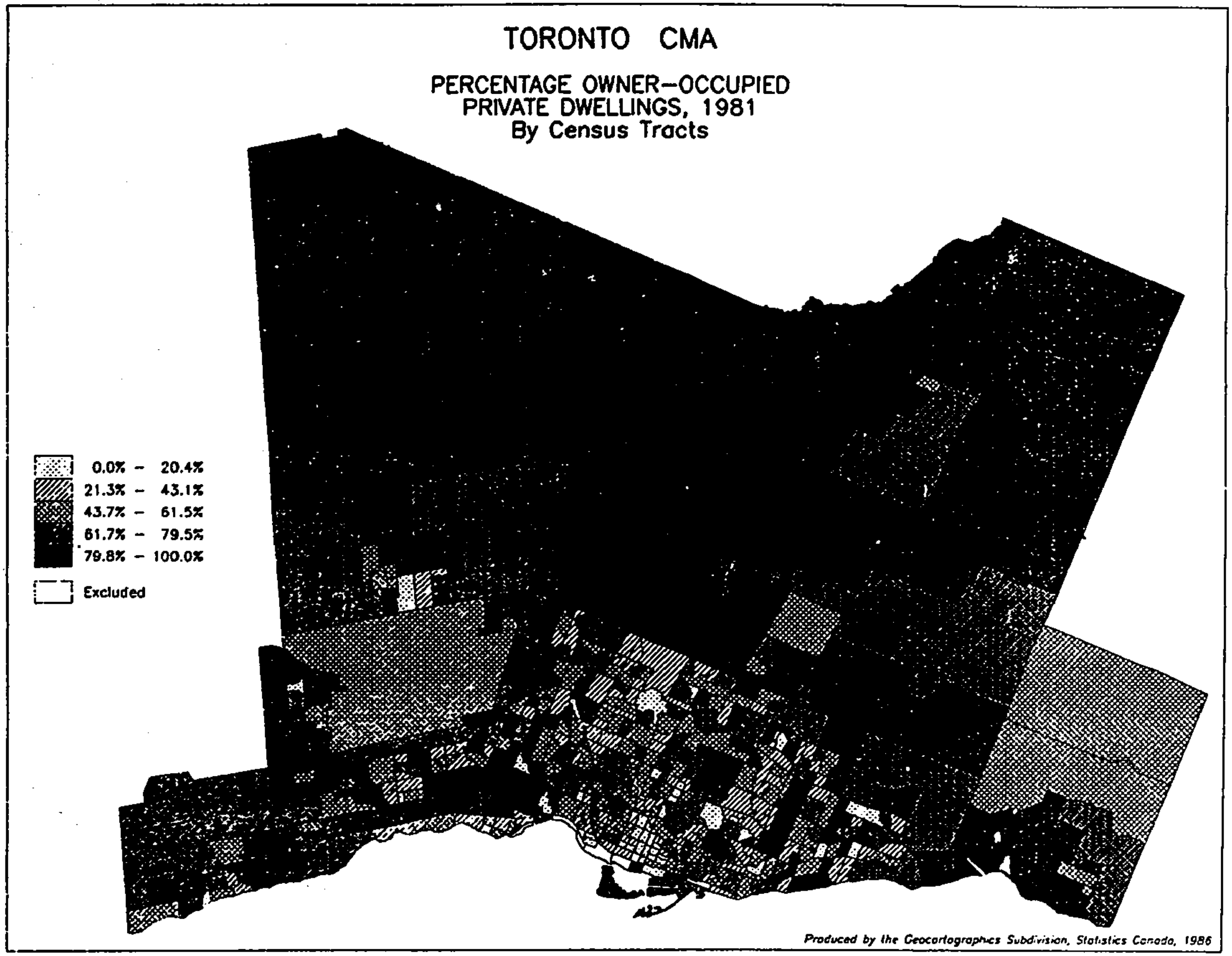
YOUNGMANN program results (area option) from L-shaped data distribution.



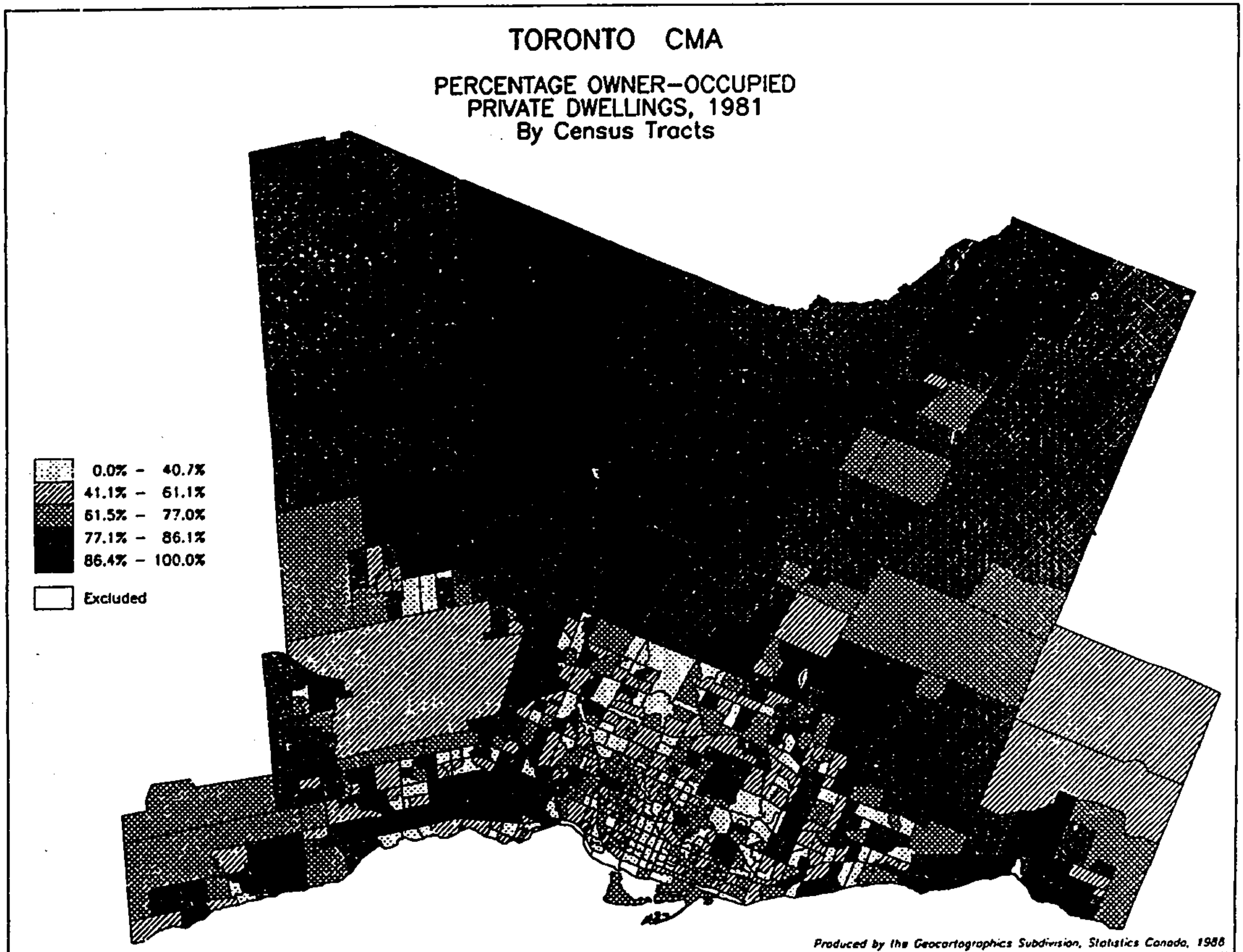
YOUNGMANN program results from left-skewed data distribution.



YOUNGMANN program results (area option) from left-skewed data distribution.



YOUNGHANN program results from rectangular data distribution.



YOUNGMANN program results (area option) from rectangular data distribution.

REFERENCES

Abler, R. F. (1987). "What Shall We Say? To Whom Shall We Speak?", Annals of the Association of American Geographers, Vol. 77, No. 4, pp. 511-524.

Anderson, J. and Child, J. (1987). "Choropleth Maps: Window Dressing or Workhorse?", GIS '87 - Second Annual Conference, Exhibits and Workshops on Geographic Information Systems, ACSM-ASPRS, Vol. 1, pp. 352-361.

Board, C. (1981). "Cartographic Communication", Cartographica, Vol. 18, No. 2, pp. 42-70.

Chang, K. (1974). "An Instructional Computer Program on Statistical Class Intervals", The Canadian Cartographer, Vol. 11, No. 1, pp. 69-77.

Chang, K. (1978). "Visual Aspects of Class Intervals in Choroplethic Mapping", The Cartographic Journal, Vol. 15, No. 1, pp. 42-48.

Chang, K. (1979). "Class Intervals and Measures of Map Complexity and Accuracy", paper presented at the AAG 75th Annual Meeting, Philadelphia, April 22-25, 23 pp.

Coulson, M.R.C. (1987). "In the Matter of Class Intervals for Choropleth Maps: With Particular Reference to the Work of George F. Jenks", Cartographica, Vol. 24, No. 2 , pp. 16-39.

Cowen, D. J. (1983). "Automated Geography and the DIDS (Decision Information Display System) Experiment" The Professional Geographer, Vol. 35, No. 3, pp. 339-340.

Dobson, J.E. (1983a). "Automated Geography", The Professional Geographer, Vol. 35, No. 2, pp. 135-143.

Dobson, J.E. (1983b). "Reply to Comments on "Automated Geography"", The Professional Geographer, Vol. 35, No. 3, pp. 349-353.

Dobson, M.W. (1973). "Choropleth Maps Without Class Intervals?: A Comment", Geographical Analysis, Vol. 5, No. 4, pp. 358-360.

Dobson, M.W. (1980). "Unclassed Choropleth Maps: A Comment", The American Cartographer, Vol. 7, No. 1, pp. 78-80.

Dutton, G. (1983). "Bivariate Construction of Equal-Class Thematic Maps", Proceedings, Auto-Carto 5, pp. 285-296.

Eastman, J.R. (1987). "Cognitive Models and Cartographic Design Research", The Cartographic Journal, Vol. 22, No. 2, pp. 95-101.

Eastman, J. R. (1987). "Graphic Syntax and Expert Systems for Map Design", 1987 ASPRS-ACSM Annual Convention - Technical Papers, Vol. 4, pp. 87-96.

Ebdon, D. (1977). "Statistics in Geography - A Practical Approach", Oxford England: Basil Blackwell, 192 pp.

Environmental Systems Research Institute (1987). ARC/Info User's Guide, Redlands California: ESRI.

Evans, I.S. (1977). "The Selection of Class Intervals", Transactions, Institute of British Geographers, Vol. 2, No. 1, pp. 98-124.

Fisher, W.D. (1958). "On Grouping for Maximum Homogeneity", Journal of the American Statistical Association, Vol. 53, pp. 789-798.

Gilmartin, P. (1987). Preface to "Studies in Cartography", Cartographica, Vol. 24, No. 2, pp. ix-xi.

Gould, J.D. and Dill, A.B. (1969). "Eye Movement Parameters and Pattern Discrimination", Perception and Psychphysics, Vol. 6, pp. 311-320.

Jenks, G.F. (1963). "Generalization in Statistical Mapping", Annals, Association of American Geographers, Vol. 53, No. 1, pp. 15-26.

Jenks, G.F. (1967). "The Data Model Concept in Statistical Mapping", International Yearbook of Cartography, Vol. 7, pp. 186-190.

Jenks, G.F. (1977). Optimal Data Classification for Choropleth Maps, Occasional Paper No. 2, Department of Geography, University of Kansas, 25 pp.

Jenks, G.F. and Caspall, F.C. (1971). "Error on Choroplethic Maps: Definition, Measurement, Reduction", Annals, Association of American Geographers, Vol. 61, No. 2, pp. 217-244.

Jenks, G.F. and Coulson, M.R.C. (1963). "Class Intervals for Statistical Maps", International Yearbook of Cartography, Vol. 3, pp. 119-134.

Johnston, R. J. (1968). "Choice in Classification: The Subjectivity of Objective Methods", Annals, Association of American Geographers, Vol. 58, pp. 575-589.

Jupe, D. (1987). "The New Technology: Will Cartography Need the Cartographer?", The Canadian Surveyor, Vol. 41, No. 3, pp. 341-346.

Kellerman, A. (1983). "Automated Geography: What Are the Real Challenges?" The Professional Geographer, Vol. 35, No. 3, pp. 342-343.

Kolacny, A. (1969). "Cartographic Information - a Fundamental Term in Modern Cartography", The Cartographic Journal, Vol. 6, pp. 47-49.

Lloyd, R. and Steinke, T. (1976). "The Decisionmaking Process for Judging the Similarity of Choropleth Maps", The American Cartographer, Vol. 3, No. 2, pp. 177-184.

Lloyd, R. and Steinke, T. (1977). "Visual and Statistical Comparison of Choropleth Maps", Annals, Association of American Geographers, Vol. 67, No. 3, pp. 429-436.

Lu, X.Z. (1987). "Regular Class Intervals With Maximum Homogeneity Within Each Class", Proceedings of the 13th International Cartographic Conference ICA, Vol. 1, pp. 337-343.

MacEachren, A.M. (1983). "Choropleth Map Accuracy: Characteristics of the Data", Proceedings, Auto-Carto 5, pp. 499-507.

Mackay, J.R. (1955). "An Analysis of Isopleth and Choropleth Class Intervals", Economic Geography, Vol. 31, No. 1, pp. 71-81.

Mackworth, N.H. and Morandi, A. (1967). "The Gaze Selects Informative Details Within Pictures", Perception and Psychophysics, Vol. 2, pp. 347-351.

Monmonier, M.S. (1972). "Contiguity-Biased Class-Interval Selection: A Method for Simplifying Patterns on Statistical Maps", Geographical Review, Vol. 62, No. 2, pp. 203-228.

Monmonier, M.S. (1973). "Eigenvalues and Principal Components: A Method for Detecting Natural Breaks for Choroplethic Maps", Proceedings ACSM Fall Convention, pp. 252-264.

Monmonier, M.S. (1974). "Measures of Pattern Complexity for Choroplethic Maps", The American Cartographer, Vol. 1, No. 2, pp. 159-169.

Monmonier, M.S. (1975). "Class Intervals to Enhance the Visual Correlation of Choroplethic Maps", The Canadian Cartographer, Vol. 12, No. 2, pp. 161-178.

Monmonier, M.S. (1976). "Modifying Objective Functions and Constraints for Maximizing Visual Correspondence of Choroplethic Maps", The Canadian Cartographer, Vol. 13, No. 1, pp. 21-34.

Monmonier, M.S. (1982). "Flat Laxity, Optimization, and Rounding in the Selection of Class Intervals", Cartographica, Vol. 19, No. 1, pp. 16-27.

Monmonier, M.S. (1983). "Comments on "Automated Geography"", The Professional Geographer, Vol. 35, No. 3, pp. 346-347.

Muehrcke, P.C. (1969). "Visual Pattern Analysis: A Look At Maps", unpublished Ph. D. dissertation. University of Michigan.

Muehrcke, P.C. (1981). "Maps in Geography", Cartographica, Vol. 18, No. 2, pp. 1-37.

Muller, J.C. (1975a). "Associations in Choropleth Map Comparison", Annals, Association of American Geographers, Vol. 65, No. 3, pp. 403-413.

Muller, J.C. (1975b). "Definition, Measurement, and Comparison of Map Attributes in Choroplethic Mapping", Proceedings of the AAG 71st Annual Meeting, Vol. 7, pp. 160-164.

Muller, J.C. (1976a). "Objective and Subjective Comparison in Choroplethic Mapping", The Cartographic Journal, Vol. 13, No. 2, pp. 156-166.

Muller, J.C. (1976b). "Numbers of Classes and Choropleth Pattern Characteristics", The American Cartographer, Vol. 3, No. 2, pp. 169-175.

Muller, J.C. (1987). "The Concept of Error in Cartography", Cartographica. Vol. 24, No. 2,pp. 1-15.

Muller, J.C., Johnson, R.D., and Vanzella, L.R. (1986). "A Knowledge-based Approach for Developing Cartographic Expertise", Proceedings, Second International Symposium on Spatial Data Handling, pp.557-571.

Noronha, V.T. (1987). "Choropleth Mapping in a Microcomputer Environment: A Critical Evaluation of Some Commercial Implementations", The American Cartographer, Vol. 14, No. 2, pp. 139-154.

Olson, J.M. (1972a). "The Effects of Class Interval Systems on Choropleth Map Correlation", The Canadian Cartographer, Vol. 9, No. 1, pp. 44-49.

Olson, J.M. (1972b). "Class Interval Systems on Maps of Observed Correlated Distributions", The Canadian Cartographer, Vol. 9, No. 2, pp. 122-131.

Olson, J.M. (1975). "Autocorrelation and Visual Map Complexity", Annals, Association of American Geographers, Vol. 65, No. 2, pp. 189-204.

Olson, J.M. (1976). "A Coordinated Approach to Map Communication Improvement", The American Cartographer, Vol. 3, No. 2, pp.151-159.

Paslawski, J. (1984). "In Search of a General Idea of Class Selection for Choropleth Maps", International Yearbook of Cartography, Vol. 24, pp. 159-171.

Peterson, M. (1978). "An Evaluation of No-Class Choropleth Mapping", paper presented at ICA 9th International Conference on Cartography, College Park, MD, July 26-August 2.

Peterson, M.P. (1980). "Unclassed Choropleth Maps: A Reply", The American Cartographer, Vol.7, No. 1, pp. 80-81.

Robinson, A. H. (1954). Elements of Cartography, New York: John Wiley & Sons, 173 pp.

Robinson, A.H., Sale, R.D. and Morrison, J.L. (1978). Elements of Cartography, 4th edition, New York: John Wiley & Sons.

Robinson, A.H., Sale, R.D., Morrison, J.L. and Muehrcke, P.C. (1984). Elements of Cartography, 5th edition, New York, John Wiley & Sons, 544 pp.

Robinson, A.H. and Petchenik, B.B. (1975). "The Map as a Communication System", The Cartographic Journal, Vol. 12, No. 1, pp. 7-15.

Robinson, V.B., Frank, A.U. and Blaze, M.A. (1986). "Expert Systems Applied to Problems in Geographic Information Systems: Introduction, Review and Prospects", Computer Environmental Urban Systems, Vol. 11, No. 4, pp. 161-173,

SAS Institute Inc. (1982). SAS User's Guide: Statistics, 1982 Edition, Cary, NC, 584 pp.

Schultz, G.M. (1961). "An Experiment in Selecting Value Scales for Statistical Distribution Maps", Surveying and Mapping, Vol. 21, No. 2, pp. 224-230.

Scripter, M.W. (1970). "Nested-Means Map Classes for Statistical Maps", Annals, Association of American Geographers, Vol. 60, No. 2, pp. 385-393.

Smith, R.M. (1986). "Comparing Traditional Methods for Selecting Class Intervals on Choropleth Maps", The Professional Geographer, Vol. 38, No. 1, pp. 62-67.

Statistics Canada (1984). Metropolitan Atlas Series - Edmonton, Dept. of Supply and Services, cat. 99-925, 65 pp.

Stefanovic, P. and Vries-Baayens, A. (1984). "Classification systems, Choropleth maps, and the Computer", The ITC Journal, 1984-1, pp. 52-57.

Stegena, L. and Csillag, F. (1987). "Statistical Determination of Class Intervals for Maps", The Cartographic Journal, Vol. 24, pp. 142-146.

Steinke, T.R. and Lloyd, R.E. (1981). "Cognitive Integration of Objective Choropleth Map Attribute Information", Cartographica, Vol. 18, No. 1, pp. 13-23.

Steinke, T. and Lloyd, R. (1983). "Judging the Similarity of Choropleth Map Images", Cartographica, Vol. 20, No. 4, pp. 35-42.

Tobler, W.R. (1973). "Choropleth Maps Without Class Intervals?", Geographical Analysis, Vol. 5, No.3, pp. 262-265.

Wasilenko, M. and Moellering, H. (1976). "Evaluating an Information Theoretic Model for Deriving Choropleth Classes", paper presented at the AAG 72nd Annual Meeting, New York, April 11-14, 17 pp.

Wasilenko, M. and Moellering, H. (1977). An Information Theoretic Model for the Derivation of Choropleth Classes, Discussion Paper No. 56, Department of Geography, Ohio State University, 40 pp.

Waugh, T. (1980). "The Development of the Gimms Computer Mapping System" in The Computer in Contemporary Cartography, D.R.F Taylor editor, Toronto: John Wiley and Sons, pp. 219-234.

Wilk, M.B. and Gnanadesikan, R. (1968). "Probability Plotting Methods for the Analysis of Data", Biometrika, Vol. 55, No. 1, pp. 1-17.

Williams, R.L. (1976). "The Misuse of Area in Mapping Census-Type Numbers", Historical Methods Newsletter, Vol. 9, No. 4, pp. 213-216.

Youngmann, C.E. (1972). "Representation Accuracy in Multi-dimensional Regionalization and Choropleth Mapping", Ph. D. dissertation, University of Kansas, 191 pp.