
by

Yoshimasa Ito

B.S., Civil Engineering
Massachusetts Institute of Technology, 1993

SUBMITTED TO THE DEPARTMENT OF URBAN STUDIES AND PLANNING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER IN CITY PLANNING AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 1996

© 1996 Yoshimasa Ito. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part.

Signature of Author: ____________________________________________

Department of Urban Studies and Planning
April 17, 1996

Certified by: ________________________________

Qing Shen
Assistant Professor of Urban Studies and Planning
Thesis Supervisor

Accepted by: ________________________________

J. Mark Schuster
Associate Professor of Urban Studies and Planning
Chairman, MCP Committee
Areal Representation Issues in Using Census Data for Urban Research:
A Comparative Analysis of Japanese and U.S. Formats

by

Yoshimasa Ito

Submitted to the Department of Urban Studies and Planning on April 17, 1996 in partial fulfillment of the requirements for the Degree of Master in City Planning

ABSTRACT

Spatial data are methodical representations of physical reality. Because there are many possible representations for any reality, spatial data users will encounter different representations. Therefore, it is important that spatial data users develop abilities to explore the various strengths and limitations that are associated with the formats of the spatial data that they use for research.

This thesis explores the effects of different census data formats on spatial analyses within the context of urban studies and planning, through a comparative analysis of the Japanese and United States census data. These data formats represent extremely different methods of representing spatial reality. The Japanese “mesh” data format is an example of ‘raster’ representation, while the U.S. approach is a ‘vector’ representation of geographic areas. Four urban-related problems that utilize census data are examined. Each problem contains elements that expose the characteristics of the census data formats being applied, while representing a broad category of urban problems.

The methodology for this analysis is three-tiered. First, a review of literature that discusses the general issues related to spatial data use, as well as issues that are specific to the urban-related problems, is conducted. Then, the methodologies for defining the geographic entities in each census data format are described. Third, the issues and complexities associated with the application of census data to these urban problems, as well as possible procedures or tools that may facilitate the analyses, are explored.

The findings indicate that the degrees of suitability for utilizing these census data formats depend on the particular needs of both the urban problems and their broad categories. Generally, the strength of the U.S. census format is evident in enumeration, analyses of areas pre-defined by political or physical boundaries, and urban analyses where high spatial resolution is desired. The Japanese mesh squares facilitate temporal analyses of larger regions, definition of the analysis area by some urban models, and analyses where areal unit consistency is desired. Land-use information in the form of maps or remote sensing data is described as a potential tool for facilitating some of the analyses.

Thesis Supervisor: Qing Shen
Title: Assistant Professor of Urban Studies and Planning
Acknowledgments

Thank you, Professors Qing Shen and Joe Ferreira, for allowing me to selfishly complete my thesis under a very unusual time constraint, and for enabling me to start developing some ideas about spatial data and their roles in urban studies and planning. It is hard to believe that it’s been almost five years since I started hanging around at the Computer Resources Laboratory as a UROPer.

Professor Atsuyuki Okabe of the Faculty of Urban Engineering, University of Tokyo, offered me valuable advice and readings when I began to examine the Japanese “mesh” census data format. During my ten-month stay as a research student at the University, I also received valuable help from Professor Yasushi Asami, Dr. Yukio Sadahiro, Keiichi Okunuki and Yutaka Gotoh, all of Okabe/Asami Laboratory.

My grandfather passed away due to a sudden illness, as this project was getting under way. Although physical distance kept us from seeing each other frequently, he offered me much needed support and encouragement during the last several years.

This thesis is dedicated to Jinpachi Ito.
# Table of Contents

Abstract .................................................. 2

Acknowledgments ........................................... 3

Chapter 1: 
Introduction ............................................. 6
  1.1 Statement of the Problem
  1.2 Research Methodology

Chapter 2: 
Literature Review ........................................ 11
  2.1 About the Data
  2.2 Use of Spatial Data
  2.3 The Proposed Urban Studies and Planning Problems
  2.4 Discussion

Chapter 3: 
The Census Data Formats Examined .................... 19
  3.1 Japanese Census Data Format
    3.1.1 Early Versions of Japanese Census Data
    3.1.2 Creation of “Mesh” Census Data in Japan
    3.1.3 Basic Processing of Mesh Data
  3.2 U.S. Census Data Format
  3.3 Comparing Inherent Characteristics of the Two Representations

Chapter 4: 
Counting of Populations Using Census Data ........... 37
  4.1 Methods for Counting People
  4.2 Relating Census Data and Spatial Representation
    4.2.1 Assigning Attribute Data to Japanese Census Mesh Squares
    4.2.2 Assigning Attribute Data to U.S. Census Geographic Entities
  4.3 Counts of Specified Regions Using Census Data
    4.3.1 Enumeration with U.S. Census Data
    4.3.2 Enumeration with Japanese Census Data
  4.4 Discussion
Chapter 5:
Measuring Population and Employment Changes Over Time

5.1 Non-static Geographic Entities Over Time
5.2 Forecasting and Estimations for Non-census Years
5.3 Meaningful Areal Units for Urban Research
5.4 (Mis) representation Using Maps
5.5 Discussion

Chapter 6:
Inter-city Transit Demand Models and Census Data

6.1 Description of Example: Suburban Rail Station
6.2 Definition of Gravity Model
6.3 Critique of Model and Mesh Representation
   6.3.1 Spatial Resolution and Population Homogeneity
6.4 Discussion

Chapter 7:
Evaluating Usage of Local Transit Stops

7.1 Different Types of Ridership Forecast Models
7.2 The Period Route Segment Model
7.3 Description of Example: Hakodate, Japan
7.4 Extracting of Data
7.5 Critique of Model and Census Formats
7.6 Discussion

Chapter 8:
Conclusion

8.1 Urban Problems Revisited
8.2 Evaluation of Data Formats
8.3 Japan's Future Spatial Data Scheme

Bibliography
Chapter 1: Introduction

1.1 Statement of the Problem

The ongoing improvements in geographic information systems (GIS) technology and related product development have enabled a larger number of people to access more useful spatial data than ever before. Spatial data is some representation of a physical reality, and there are many different ways to represent one reality; spatial data users will face an increasing number of such representations, in the form of new data products. Therefore, it is now important that spatial data users develop abilities to explore the various pro’s and con’s as well as strengths and limitations that are associated with the spatial data format(s) that they are constrained to utilize for research.

This paper first examines and compares existing and known spatial data formats, in this case the census data formats of the United States and Japan. As with many other census data formats being implemented in industrialized nations, those in the U.S. and in Japan are produced in timely fashion, with allocated budgets that are sufficiently large. The geographic entities that partition the respective nations for census tabulation purposes are related in a rigid hierarchy in order to take differing scales into account. In both countries, the census is conducted by well respected national agencies; the data products are generally considered to be of reasonably legitimate quality.

However, the census formats in U.S. and Japan also represent two extremely different representations of physical reality, in a spatial sense. In short, the Japanese census format represents the raster data structure approach. The nation is partitioned by congruent squares that form a grid, and attribute data is tabulated for each of these squares. No particular attention is given to physical details such as roads and concentrations of development. In comparison, the U.S. census partitions the nation with census areas that are incongruent and irregularly shaped. The boundaries are drawn with the goal of
identifying communities that are internally homogeneous to some extent and, distinct in some way from the surrounding areas. A significant component of the U.S. census data product illustrates man made and natural objects such as roads and rivers, among other details. While the Japanese format answers the question "What is happening in this area?" for the data user, the U.S. format provides answers to the "Where is this object (such as street address or community) located?" question.

Due to these significant differences, it is reasonable to expect that the two data formats often exhibit different strengths and weaknesses, as well as pose different complexities when the data are utilized for urban studies and planning related research. In order to understand these differences as well as similarities, the two census formats will be compared within the context of several proposed urban-related problems. The proposed problems will be diverse, and will be designed to bring many of each format's characteristics to light. The information gained from this process will be useful to persons engaged in spatial data evaluation processes for their research.

1.2 Research Methodology

The urban-related problems are proposed in order to gauge the suitability of the two data sets being examined, in light of the aforementioned differences. The problems are strategically designed in order to facilitate a broad gauging of the dataset's suitability, as potential users of spatial data will undoubtedly face a diverse set of needs. Urban analyses can be generally sorted into the following categories, although real-life problems often have several components which require separate analyses:

1. **Time-constant:** Analyses that seek information about some condition at one particular time. Calculating the January, 1996 median income of a neighborhood is a time-constant analysis.

---

1National Center for Geographic Information and Analysis, p. 44.
2. **Time-series**: Analysis of change over a period of time. An example is the calculation of change in median income of a neighborhood between January, 1995 and January, 1996.

3. **Macro-area**: Analysis where the extent of the analysis area is significantly larger than the areal unit that contains the object or objects in question. Example: analyzing the effects of a chemical spill when winds can carry the fumes for 10 to 20 miles.

4. **Micro-area**: Analysis of an object or objects which affect the immediate area only, with the analysis area usually determined by a radius or some other well defined threshold. Example: analyzing the effects of a fire siren which is audible within a 1 mile radius.

Below, four urban-related problems that may utilize census data are proposed. Each corresponds to the four analysis categories described above, in the same order. Described for each problem are 1) the problem setting; 2) the nature of the problem that necessitate particular capabilities of the applied spatial data; and 3) a brief mention if the problem is an actual census data application in the U.S. and/or Japan.

1. **Performing a population count for user-defined areas.** Clearly, the desired spatial data characteristic is the ability to provide accurate counts of objects within client specified boundaries, such as municipalities or neighborhoods. Enumeration for defining electoral districts is an actual U.S. census data application.

2. **Examining population changes in traditional downtown areas versus adjacent neighborhoods over time.** Population distributions of some cities exhibit a so-called donutization effect, where the night population in the traditional city center migrate to suburbs, clearing the way for commercial development. The required capabilities include comparisons of the same areas over a period of time, and to draw distinctions between the downtown and suburban areas.

3. **Forecasting ridership along an urban/suburban rail corridor.** A gravity model that incorporates production, attractiveness and impedance factors is defined, relating the
areal units throughout a relatively vast region. The comparison focuses on the effects of changing spatial resolution and the differing methodologies of the two census formats when forecasting transit ridership throughout a large area.

4. Determining the overall usage of local trolley stops along a route. This information is useful when studying the feasibility of the transit stops’ existence, or selecting heavily used facilities for upgrading. A discussion of a demand model incorporating U.S. census data is taken from the literature. Accurate re-creation of the analysis zone is a principal goal. Here again, spatial resolution and format methodologies are important issues.

The subsequent chapters of this thesis are organized in the following manner. In Chapter 2, previous scholarship that is relevant to the stated objectives will be introduced, along with their reasons for relevance. Chapter 3 describes the geographic entities that comprise the U.S. and Japanese census data, with detailed definitions for particular areal units that are relevant in later chapters. The next four chapters are dedicated to discussions of the four analysis categories and urban problems.

In analyzing each of the proposed urban problems, actual census data is extracted and analyzed to present much of the Japanese case. Previous works of U.S. census data applications are cited in order to perform the comparison as well as to supplement the general discussion. The primary reasons for this approach were data availability and the constraint of time. It was also assumed that a comparative discussion weighing more on the Japanese data side would be more beneficial to the English-speaking audience, who may be relatively more familiar with the United States census data format.

The discussions for each of the proposed problems will attempt to answer the following questions:

- How do the characteristics and complexities associated with the formats' methodology affect the analysis of this problem, and more generally, the category to which the problem belongs?
• What procedures, if any, would facilitate the analysis using these census data formats?
• Is there a decidedly more appropriate data format to use for this category or problem?

Finally, the findings in these chapters will be summarized and, in conclusion, statements regarding each of the census data formats in general will be made.
Chapter 2: Literature Review

The course of formulating the proposal of this thesis can be generally described as a four-tiered process outlined here:

1. Acquisition of Japanese census data from 1970 through 1990 in its space delimited form; understanding its method and hierarchy of areal representation, past stated objectives and expectations that shaped its design, and simple analytical techniques normally associated with raster data models, through readings.

2. Examination of the U.S. census data format in a similar fashion as step 1, via a number of data products offered by the U.S. Bureau of the Census and literature. Contemplation of a comparative study involving the two aforementioned census data formats, focusing on areal representation differences and their consequences in typical urban studies and planning related problems.

3. Examination of previous works that address areal representation, or spatial data issues, particularly works that discuss common problems associated with spatial data, manipulations of areal representations that illustrate departures from the spatial 'reality', and error estimation techniques.

4. Examination of literature that apply census or other spatial data with attributes to typical urban related problems, such as population counts and projections, population and employment changes over a period of time, and transit ridership forecast models.

Broadly speaking, the purpose of this chapter is to introduce and discuss the literature which serve as a) descriptive guides which explain in detail the workings of each census data formats, historical precedents, and typical applications (steps 1 and 2 above); b) a foundation of past and recent works related to the thesis topic upon which to build (step 3 above); and d) examples of urban problems which will be the starting points of the comparative analysis (step 4 above).
2.1 About the Data

It is important to fully understand the methodologies of both census data formats before embarking on any applications. In comparing the situations in Japan and the U.S., the significant difference in the sheer volume and content of literature and other information that serve to educate users and potential users of census data in the two countries mainly stems from one reality: A complete U.S. census data file is readily available to the person seeking it, and the Japanese census data file is not. The array of U.S. census related data products available to the public is even overwhelming at first. In contrast, Kubo (1987) points out that the government had always been wary of releasing census data since the inception of the Japanese "mesh" grid system. Permission to use the data was granted for national and local government uses and full-time university research only; requests from other organizations had been refused. Redistribution of obtained census data was prohibited without permission, and transfer of data required fees to be paid. This policy resulted in, among other things, a small census data user population and a literature base of narrow scope and limited applications in Japan.

With respect to urban studies and planning applications of Japanese census data, it is generally agreed that Okudaira (1982) is the definitive source of information. This book addresses the broader topic of urban and regional analysis, not just census data. The dearth of books on this matter in Japan reflects the fact that the use of this data is largely limited to ministry policy makers and university researchers in national planning and economics. He starts by devoting a section to the history of census data format development in Japan, then describes the geographical entities and their hierarchy. Also described are some basic data processing procedures (spatial query, distance calculation) and examples of cluster and factor analysis using census data. Okudaira claims that grid square census data holds distinct advantages for certain macro level analyses. Yet, he
ends the discussions with a chapter on an possibly emerging data structure for local area analysis, which is not unlike the block groups of the U.S. census data.

The other major source of descriptive material on Japanese census data is the data source itself, the Statistics Bureau of the Management and Coordination Agency (1990). While the data files are off-limits to the general public, the Bureau publishes volumes containing tabulations and thematic maps of selected question items, and produces guidebooks that contain grid descriptions and some statistics, such as average mesh population, total number of mesh squares, etc. Kubo (1987) provides insight into developments of spatial data in other Japanese ministries, agencies, local governments and industry, from a historical perspective. Pointing out the sluggish GIS development in Japan compared to other data-rich countries, Kubo blames the lack of data availability and the unwillingness of the various ministries to share their data with each other, as well as the general apathy of government officials towards new GIS developments elsewhere, particularly in the United States. Specifically, this alludes to the lack of a standardized digitizing effort for all areas in Japan, which has resulted in a general lack of good digitized maps.

As expected, the definitive description for the U.S. census data and its data products comes from the data source, the Bureau of the Census (1990). It is actually quite difficult to learn the dozens of geographical entities defined by the U.S. census, and the several hierarchy systems to which the geographical entities belong. The 1990 census introduced TIGER (Topologically Integrated Geographic Encoding and Referencing) system, which enables the assigning of an address to its proper block group and higher entities, generation of maps, updating of boundary information, and address matching of sites onto digital representations of road and other physical features, among other tasks. Since this guide is published dicennially, little historical perspective other than the most recent changes are offered in each edition.
There seems to be an abundance of literature that discusses all facets of the U.S. census, from the data collecting procedures and analytical methods, to the politicking involving the use of census data, and such. Literature was examined with the thesis scope in mind. Kaplan (1980) provides all of the above, in general terms that enable the newcomer to understand census data and the census process. Myers (1992) provides an introductory, how-to handbook for analysis using U.S. census data. Myers brings to light certain inherent complexities associated with the data, such as the splitting and changing boundaries of census tracts, although these complexities are not illustrated with actual examples, and possible solutions are not explored. The census bureau acknowledges the fact but does not stress the characteristic as a complexity. Myers contributes to the discussion by using simplified but specific urban analysis examples to illustrate census data use at local levels (analyses using block group and census tract aggregate levels).

Finally, the National Center for Geographic Information and Analysis (1990) provides textbook style explanations of GIS within the context of census data applications, which is helpful in guiding the thesis through the stage of visualizing census data through GIS to illustrate some assertions in the example urban problems.

### 2.2 Use of Spatial Data

The U.S. and Japanese data formats contain inherent characteristics that are particular to each format, but much of the past scholarship has focused on issues and difficulties faced by spatial data analysts everywhere. A most central foci of this thesis are the effects of different areal representations of reality on urban research. Census data formats are “shrink-wrapped” for use; that is, the areal units for the data are pre-defined. Of course, most research involves the analyses of areas that do not coincide with these pre-defined areas. When using census data for urban research, areal units do not usually coincide perfectly with areas that is the focus of the analysis. Neither the Japanese nor U.S. census areas coincide with most transit analysis zones or municipal boundaries, for
example. Furthermore, what if the analysis took place at the sub-municipality level, such as neighborhoods? There are no known and rigid boundaries for these most local areas. For these reasons, the urban researcher must be aware of the consequences that are brought about by the spatial representation method of the spatial data that is to be used for his analysis.

In the past, geographers led the way in scholarship regarding different areal representations; urban researchers were often constrained by the data format available to them, and thus many chose not to investigate other options. Among these advances in geography, an often-quoted framework for this issue is presented by Openshaw (1984), who coined the term "Modifiable Areal Unit Problem" (MAUP). Openshaw describes the problem as twofold: a given region can be partitioned into smaller sub-regions in many different combinations; there are also many ways to aggregate smaller sub-regions to create larger regions. Respectively, the scale effect and the zoning/aggregation effect causes spatial analyses of the same area using different partitions or levels of aggregation to produce results that potentially are quite different.

Openshaw concludes in a somewhat controversial manner. The MAUP is not solvable when one looks for the "correct" level of aggregation or areal units. It can be utilized as a powerful tool by analysts, however, who may manipulate areal units to their liking, in order to maximize the legitimacy of their theories. In urban analysis, this recommendation can be interpreted in two ways. First, most urban studies and planning research seeks information of defined areas such as municipalities, transit corridors, and neighborhoods. Thus, the areal units of study are already given, and using them will produce meaningful information.¹ The second interpretation is more sinister; the MAUP extends to possibilities where the inherent characteristics of the problems are unintentionally and intentionally applied to produce misleading maps and manipulate audiences. Monmonier (1977, 1991) humorously but strikingly presents examples of how

¹ This statement was shaped during a discussion involving Professor Shen and the author.
to "lie with maps", selecting aggregate levels and drawing areal boundaries so that the analysis would suggest the desired messages. As aggregate levels and boundaries are changed, the implied message changes as well.

The MAUP zone/aggregation effect is illustrated in Chen (1994)'s comparison of census tract and block group levels of aggregation in modeling housing and demographic diversity; results are shown to be significantly different from each other. It is important to point out the tradeoffs that always occur when analyses using different aggregate levels are compared. When lower aggregate level data is used, the result can show what is happening at a very local level. If a local area is distinct from the surrounding areas, this analysis will expose this condition. However, analysis at this level is sometimes deceptive. For example, a small area which is statistically an outlier among the surrounding area will achieve “prominence”, thus creating statistical noise that detracts from illustrating the overall trend of the larger area. It is also known that smaller areal units are more prone to higher enumeration and other representation errors, when the analysis is conducted in percentage figures.

2.3 The Proposed Urban Studies and Planning Problems

a. Population count. Counting of individuals within some specified boundary is performed using various forms of data, including census data, depending on what is the most dependable source for the location. The objective is to perform an accurate count of a client specified area, and these works explore the endeavor of enumeration. Rhind (1991) provides a comprehensive description of counting procedures in relatively data rich European countries. Many European countries as well as Japan maintain a population registry at the municipality level. Rhind's discussion outlines the strengths and limitations of the registry; while the registry can be highly accurate, it does not reflect changes over time in intervals, like a census can.
It is widely recognized that a primary source of enumeration error found in the Japanese census format is due to the aggregation of block group-like areal units in forming the rigorous grid squares which are represented as having straight edges. It is obvious that the aggregation of block groups cannot form straight lines, and an inherent counting difference exists between the sum of the aggregated sub-regions and the actual number that truly exists within the rigorously defined square. Brusegard and Menger (1989) explains that errors are man-produced as a result of a series of disaggregations and aggregations to satisfy some client specified boundary. Koshizuka (1984) and Ihara, Iwasa and Yamaguchi (1995) evaluate the extent of error in grid-square statistics, with the latter research concluding that a 1 km\(^2\) must contain more than 3,000 records for the figure to be less than 10\% erroneous with 95\% confidence. Also described in detail are the rules for assignment of block groups that lie over the boundaries of grid squares. These rules have changed virtually every time a new census was conducted, and Ihara et. al. (1995) suggest that the changes were for the better.

b. Temporal changes in population/employment / Projection. The strengths and weaknesses of Japanese and U.S. census data formats when applied to temporal analyses, as well as the feasibility of population projections using local census areas for non-census years is discussed in these works. Okudaira (1982) emphasizes that a major advantage of grid squares is the stability of their boundaries over time. Smith and Shahidullah (1995) as well as Tayman (1992) explore various methods of projecting populations for census tracts. Projections become more difficult as census areas become more disaggregated, because small areas are subject to higher enumeration errors. Howenstine (1993) addresses the problems caused by changing physical characteristics of U.S. census tracts over time. Census tracts are supposed to have visible boundaries such as roads and contain between 2,500 and 8,000 people for an average of 4,000. Their boundaries may shift, and tracts may be merged or split in order to account for changes in infrastructure and population. Since split tracts are more frequent than the other two changes, methods
to reconcile the different representations are discussed. It is concluded that leaving the recently split tracts intact and splitting the previously aggregated corresponding area by proportion of populations in the new tracts performs the reconciliation with the smallest deviation from the actual breakdown.

c. *Demand forecast along a transit route.* Foot (1981) provides a textbook-like explanation for urban models, including gravity models. In Chapter 6, the gravity model is described in detail, and is adapted for the implementation of census data.

d. *Usage of local transit stops.* Azar and Ferreira (1995) outline the Period Route Segment (PRS) model which forecasts transit ridership for each defined traffic analysis zone (TAZ). Here, U.S. census data block groups are implemented to create the TAZs. Chapter 7 discusses the prospects of applying Japanese mesh data to a local transit model such as the PRS.

2.4 Discussion

The examination of the literature enabled an understanding of 1) U.S. and Japanese census data formats’ methodology; 2) common issues and problems faced by spatial data users in general; 3) previously tried approaches to the four proposed urban problems, within the U.S. census data context; and 4) tools, such as procedures and techniques, that may facilitate the analyses of the proposed problems. These materials will be restated more thoroughly in later chapters.

This thesis contributes to previous works as a comparative discussion of two census data formats within the context of the proposed urban problems and, more generally, the broader categories of problems. As conclusions regarding the relative suitability of both census formats and the feasibility of tools are drawn for categories of problems, it is hoped that such conclusions can be helpful in approaching different problems that fall into these categories.
Chapter 3:  
The Census Data Formats Examined

This chapter is an overview of the two areal representations, the Japanese and United States census data formats. Historical forces that led to the current formats is described. Then, the definitions of geographic entities which are relevant in this paper, as well as the hierarchies that relate the geographic entities are presented. Finally, a general discussion of data analysis issues that arise from these definitions is presented. These issues will be examined closely within the context of specific urban studies and planning problems later on.

3.1 Japanese Census Data Format

3.1.1 Early Versions of Japanese Census Data

The statement below by social scientist Toshio Sanuki in 1969 is a representative opinion of the times that advocated the creation of what is today called "mesh data" in Japan.

Creative strategic maneuvers are necessary in order to respond to the challenges of the twenty first century and emerge victorious. Formulation of strategy begins with a close look at today's reality. A reorganization of the nation's regional statistics and data, the fundamental material needed to judge this reality, is essential in order to realize the future goals on the Japanese archipelago. Previous regional statistics have been compiled and tallied for a political jurisdiction, primarily for use by that district. It can be said that such data served its purpose of sufficiently fulfilling municipal needs. However, due to rapid changes taking place in our nation, there exists now a need to analyze large regions that transcend municipalities, as well as to examine small parcels within municipalities. Prefectural and municipal data is thus insufficient for projecting future national land use, implementing planning and development strategies. Furthermore, if national land development of the future must incorporate the fourth dimensional element called "time", the data itself must be of a consecutive and consistent nature over a long period of time.¹

Regional statistics incorporate various information of a particular region and tally the information by tracts, or unit areas. Political jurisdiction were common unit areas in pre-

¹Okudaira, pp. 151-152.
war Japan. Japan was a nation of many villages, and their boundaries were often clear and unquestioned, often demarcated by nature as well. The numbers of towns and villages decreased with modernization, but the biggest dip occurred in 1953 as a result of the Town-Village Annexation Ordinance being implemented. These events proved very inconvenient for people utilizing regional statistical data, as the statistical unit areas became too large, and time series analyses became impossible.

It had become evident that division of regions according to historical, natural and political perspectives had its limits in terms of consistency and continuity. If such traits are important, regions must be mechanically divided by equidistant horizontal and vertical lines. The congruent squares that result from such a procedure would form the basis for this new grid, or "mesh" system.

It is generally agreed that the Finnish geographer J.G. Granö was the first to utilize the mesh system in the scholastic arena. In 1929, he presented a paper on the regional analyses of various natural and anthropological trends using a 1km² mesh system. Since then, mesh users have gradually increased, and its merits are being recognized.

3.1.2 Creation of "Mesh" Census Data in Japan

In Japan, census data is commonly referred to as "mesh data". Other names include 'grid', 'cell', 'grid square system', 'block grid system', and 'grid coordinate system. One can imagine a grid created by vertical and horizontal lines, all 1 kilometer apart. The 1km² squares created by this "mesh" system serve as the fundamental regional unit. Thus, any point in the country belong to one of the 386,522 mesh squares.

This is not to say that actual census surveys are conducted with the mesh as a research unit. Each enumerator conducts census research in an area of approximately 50 households and physically resembling a U.S. census block group. The mesh squares are an aggregation of these research units which are "fitted" into the rigorous mesh square. To resolve disputes involving the research units that lie over mesh boundaries, various
techniques have been implemented in different census years, and it is generally conceded that all such techniques are imperfect and that errors will result.

Since the mesh system was introduced for census data in 1970, the census has been conducted every five years (1970, 1975, 1980, 1985, 1990). In previous years, the bureau had conducted census surveys, but the data had been tallied with fundamental regions falling along municipal boundaries and natural barriers. In the 1960's, with the nation experiencing rapid economic growth, demand rose for a new type of data that would allow forecasting and optimizing of infrastructure projects. It quickly became clear that existing forms of census data was not very useful. Among the reasons given were:

(1) since boundaries of municipalities changed with time (including revamping of municipal hierarchies, mergers, etc..), a time series analysis of a given region could not be performed,

(2) data in which the fundamental regions are mechanically and equally created are easier to handle, when performing comparisons of regions, as well as when calculating for the cumulative totals of several meshes,

(3) the mesh nature facilitates numerical labeling and orderly mapping of meshes.

Once the creation of the mesh system was justified, the actual procedure for creating the mesh squares was discussed. It was eventually agreed that a grid using lines parallel to longitudinal and latitudinal lines would be the best procedure. It should be noted here, however, that while latitudinal lines are equidistant everywhere, spacing between longitudinal lines becomes smaller as one travels north in the northern hemisphere (thus these lines are actually not parallel). Therefore, mesh areas will not be constant as one travels northward and southward. Comparing the northern and southern extremes, there is a 16% area difference between meshes in the metropolitan Sapporo and metropolitan Kagoshima areas. Between Tokyo and Kita-Kyushu, where the core of Japanese economic activities lies, the error is merely 2.5%. Nevertheless, this longitude-latitude
system was selected over other coordinate systems (UTM, 17-coordinate systems) due to our general familiarity with longitudinal and latitudinal lines.

Mesh hierarchy and mesh labels (called mesh codes) were created as follows. Each 1 km² mesh has a corresponding 8 digit mesh code. The first 4 digits denote the level I mesh group (each level I mesh is approximately 80 km by 80 km, thus consisting of exactly 64 level II meshes, each approximately 10 km by 10 km), the next 2 digits denote the level II mesh group (1 through 64), and the last 2 digits denote the level III, or fundamental mesh (there are 100 fundamental mesh squares in one level II mesh).

The longitudinal lines are drawn at each degree (123°E through 149°E). In central Japan, these lines are approximately 80 km apart. The latitudinal lines are drawn at 40 minute intervals (36°, 36°40', 37°20',...), and these lines are approximately 80 km apart. Thus, a level I mesh group is created by adjacent latitudinal and longitudinal lines. As stated earlier, the first 4 digits correspond to the level I mesh. Of these, the first 2 digits are calculated as follows:

\[ \text{LATITUDE} \times 1.5 \]

The next 2 digits are derived as follows:

\[ \text{LAST 2 DIGITS OF LONGITUDE} \]

In Figure 3-1, the 4 digit code is given as 5 4 3 8. The lines that intersect at the lower left corner of each level I mesh is used for the calculations.

Next, each level I mesh is subdivided into 64 squares of 10 km by 10 km (level II meshes). Each is labeled, from 00-07, 10-17, 20-27,... to 70-77 as illustrated. The level II mesh in Figure 3-2 would have the code 5 4 3 8 - 2 3. Finally, each level II mesh is subdivided into 100 squares of 1 km² each (fundamental meshes, or level III meshes). Each is labeled from 00 to 99 as illustrated. The level III mesh in Figure 3-3 would have the meshcode 5 4 3 8 - 2 3 - 5 2.
Figure 3-1 through 3-3 The Mesh System

3.1

Region A is a Level I Mesh Square # 5 4 3 8
5 4 = 36° x 1.5; 3 8 = lower two digits of longitude on the west side

3.2

Region B is a Level II Mesh Square # 5 4 3 8 2 3
Each Level II Region corresponds to one 1/25000 scale topographical map
Region C is a Level III Mesh Square #5 4 3 8 2 3 5 2

(National Center for Geographic Information and Analysis, 1993)
Furthermore, the Bureau of Statistics has compiled a level IV mesh system (500 m by 500 m meshes) for selected metropolitan areas, which simply involves a subdivision of the fundamental meshes into 4 squares.

In Japan, the national census data and the national business establishment data are compiled in the mesh system every five years. The national business establishment data lags by one year, so the latest business establishment data set was compiled in 1991. All of the fields that appear in the data sets are summarized here (Table 3-2).

3.1.3 Basic Processing of Mesh Data

The characteristics of mesh data make it suitable for the following procedures, among others, to be performed.

SIMPLE CALCULATIONS. Since the mesh squares are assumed to be congruent, many types of calculations within a mesh or encompassing multiple meshes are possible. In an example of mesh census data, the population which is not part of the labor force (NLP) is calculated by adding the values of two fields:

\[
\text{NLP} = [\text{night population under 15}] + [\text{night population over 65}]
\]

It is easy to obtain this figure for each mesh. Furthermore, the density of NLP over a particular area can be calculated by adding each NLP and then dividing that figure by the number of cells. Other calculations can be made using the values of other fields, or categories, in the census.

GEOMETRIC QUERY. Occasionally, there arises a need to select a sub-region of a certain shape (square, circle) from some region. For example, a circular region is cut to select the areas which are affected by some poisonous spillage that emits fumes equally in all directions. Although the square shape of each mesh is somewhat limiting, the ideal shape can be closely imitated.

CALCULATING DISTANCE. If two mesh of coordinates \((x_i, y_i)\) and \((x_j, y_j)\) are given, the distance between them can be calculated as the following:
Table 3.2

Complete List of Tabulations: 1985 Census Data

| Name of survey          | Survey year | Mesh code | Mesh level (I thru IV) | 1/25000 Map of the Level II mesh corresponding to mesh | Number of cities, wards, towns and villages within mesh | Prefectural and city codes for each municipality within mesh | Night population (Total, Male, Female; the same for all fields below) | Labor force | Employed workers | Totally unemployed | Non labor force | Hired personnel (including management) | Business owner | Member of family business | Type 1 sector personnel | Agricultural personnel | Forestry personnel | Fishery personnel | Type 2 sector personnel | Mining personnel |
|------------------------|-------------|-----------|-----------------------|-----------------------------------------------------|------------------------------------------------------|--------------------------------------------------------|-------------------------------------------------------------|-------------|-------------------|------------------------|-------------------|---------------------------|------------------|-----------------------------|----------------------|------------------|----------------------|----------------------|
|                        |             |           |                       |                                                     |                                                      |                                                        | 0-4 years old                                               |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 5-9 years old                                               |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 10-14 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 15-19 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 20-24 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 25-29 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 30-34 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 35-39 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 40-44 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 45-49 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 50-54 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 55-59 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 60-64 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 65-69 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 70-74 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 75-79 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 80-84 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 85 years old and over                                       |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 0-2 years old                                               |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 0-5 years old                                               |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 3-5 years old                                               |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 6-11 years old                                              |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 12-14 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 15-17 years old                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 18 years old                                                |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | 19 years old                                                |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Labor force                                                 |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Employed workers                                            |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Totally unemployed                                          |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Non labor force                                             |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Hired personnel (including management)                    |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Business owner                                              |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Member of family business                                   |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Type 1 sector personnel                                    |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Agricultural personnel                                     |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Forestry personnel                                          |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Fishery personnel                                           |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Type 2 sector personnel                                    |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
|                        |             |           |                       |                                                     |                                                      |                                                        | Mining personnel                                            |             |                   |                        |                   |                           |                  |                             |                      |                  |                      |                      |
Construction personnel
Manufacturing personnel
Type 3 sector personnel
Electric/Gas/Heat/Water personnel
Transport/Communication personnel
Small distributor/Restaurant-Bar personnel
Financial/Insurance personnel
Real Estate personnel
Service industry personnel
Civil Service personnel
Professional/Technical personnel
Management personnel
Office Administrative personnel
Sales personnel
Agri/Forestry/Fishery worker
Construction worker
Transport/Communication worker
Technician/Manufacturing worker/Other worker
Security Service worker
Service industry worker

(No more tabs by sex)
Commuters 15 years old and over (work, school)
Persons working at home
Commute within city, ward, town or village (work, school)
Commute within prefecture but different municipality (w, s)
Commute outside prefecture (w, s)

Total households
General household
Normal household
Sub-household

One person household (general, normal)
2,3,4 person household (g, n)
5,6,7+ person household (g, n)

Type of household
Extended household
Nuclear household
Other
Household with relative under 6 years of age
Household with relative over 65 years of age

Condition of household
All commuters, with school commuter under 12 years of age
All non-commuters are senior citizens
All non-commuters are either senior citizens or small children

Household supported by Agri/Forestry/Fishery
Household supported by Agri/Forestry/Fishery and other sector
Household not supported by Agri/Forestry/Fishery
Household without employed persons
Type of abode
Single home
Nagaya (tenement)
Multiple family (1-2 floors, 3-5 floors, 6+ floors)
Ownership
Households living in an abode
Households living in an owned home
Households living in public housing
Households living in rented housing
Households living in company housing
Households renting within a home
Households renting within a home, 1 person household

Male/Female ratio
Average age
Percentage of pre-adolescents
Percentage of persons at a 'productive' age
Percentage of senior citizens
Percentage of labor force
Percentage of working people
Percentage of working females
Percentage of hired people
Percentage of people who own businesses
Percentage of working people in Type 1 sector
Percentage of working people in agriculture
Percentage of working people in Type 2 sector
Percentage of working people in construction
Percentage of working people in manufacturing
Percentage of working people in Type 3 sector
Percentage of working people in small distributor/small store/restaurant/bar
Percentage of working people in service industry
Percentage of working people in professional/technical/managerial/administrative position
Percentage of workers as technicians, manufacturing and other labor positions
Percentage of workers in sales and service
Percentage of commuters
Percentage of commuters to other municipalities
Percentage of nuclear households
Percentage of households with children under 6 years of age
Percentage of households with people 65 years of age and over
Percentage of single homes
Percentage of nagaya
Percentage of purchased homes
Percentage of public housing
Percentage of rented home

Number of rooms per household
Number of tatami mats per household
Average member per household
Number of rooms per person
Number of tatami mats per person

Complete List of Tabulations: 1981 Business Establishment Data

Name of survey
Survey year
Mesh code
Mesh level (I thru IV)
1/25000 Map of the Level II mesh corresponding to mesh
Number of cities, wards, towns and villages within mesh
Prefectural and city codes for each municipality within mesh

(All categories have two fields: Number of facilities, Number of employees, unless otherwise noted)

All industries
Type 2 sector industries
Mining
Construction
Manufacturing
Light industry-materials
Light industry-processing
Heavy industry-materials
Heavy industry-processing
Chemical industry
Other industry
Type 3 sector industries
Small distributor-retail
Small distributor
Small retail
Textile/Clothing distributor
Restaurant-Bar distributor
Restaurant-Bar
Financial/Insurance
Real Estate
Transport/Communication
Electricity/Gas/Water/Heat
Service
Laundry/Cosmetic
Medical
Education
Social insurance/welfare
Life services
Medical/Sanitary/Welfare related services
Entertainment related service
Administrative related service
Civil Service

Manufacturing 1-9 employees
Manufacturing 10-29 employees
Manufacturing 30-99 employees
Manufacturing 100-299 employees
Manufacturing 300-499 employees
Manufacturing 500-999 employees
Manufacturing 1000+ employees
Distributors/Retail 1-9 employees
Distributors/Retail 10-29 employees
Distributors/Retail 30-49 employees
Distributors/Retail 50-99 employees
Distributors/Retail 100-299 employees
Distributors/Retail 300+ employees
<table>
<thead>
<tr>
<th>Service</th>
<th>1-9 employees</th>
<th>10-29 employees</th>
<th>30-49 employees</th>
<th>50-99 employees</th>
<th>100-299 employees</th>
<th>300+ employees</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>All industries</th>
<th>1-4 employees</th>
<th>5-19 employees</th>
<th>20-29 employees</th>
<th>30-49 employees</th>
<th>50-99 employees</th>
<th>100-299 employees</th>
<th>300+ employees</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Management practice</th>
<th>Self</th>
<th>Corporate</th>
<th>Public</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Business Format</th>
<th>Store/Restaurant-Bar</th>
<th>Office</th>
<th>Factory/Plant/Works</th>
</tr>
</thead>
</table>

|------------------|-------------|-------------|-------------|-------------|-------------|---------------------|

<table>
<thead>
<tr>
<th>Average number of employees</th>
<th>All industry</th>
<th>Manufacturing</th>
<th>Supplier/Retail</th>
<th>Service</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Proportion by industry</th>
<th>Manufacturing</th>
<th>Distributor/Retail</th>
<th>Service</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Proportion by time of founding</th>
<th>before 1954</th>
<th>1955 - 1972</th>
<th>1973 and thereafter</th>
</tr>
</thead>
</table>

(Japan Agency of Management and Coordination, Bureau of Statistics, 1985)
\[ d_{ij} = \left\{ (x_i - x_j)^n + (y_i - y_j)^n \right\}^{1/n} \]

*n*=2 when direct travel is possible, and *n*=1 when only horizontal and vertical movements are possible.

### 3.2 U.S. Census Data Format

A long tradition of census taking exists in the United States, beginning from its colonial period. One-time enumeration in the colonies of Virginia (1624), New York (1712), Connecticut (1756), Massachusetts (1764) and Rhode Island (1774) were conducted at the request of the British for administrative purposes. However, guidelines that defined the purpose of the modern U.S. census was outlined in the U.S. Constitution, adopted in 1787. The cost of fighting the Revolutionary War had been high, and each State was to contribute to the common defense of the new nation on the basis of their populations. The new constitution also defined an electoral system where States would be represented in Congress on the basis of their populations. The Constitution required the President of the Union to conduct a popular census every ten years.

The U.S. census is still conducted once every ten years, and has a stated mission of acting as a tool to redefine electoral districts. The data structure hierarchy is illustrated in Figure 3-4. It is evident that the hierarchy contains numerous geographic entities, for various purposes. In later chapters, the census tracts, block groups and blocks will be the entities that will receive much discussion. The Japanese Level III mesh squares have an average population of 749 throughout Japan (1985) although the range is very wide. This geographic entity is the least aggregated level where data for the entire country is available. Since U.S. block groups average 1,000 people, this entity arguably is a sound choice for comparative purposes. Census tracts are important, as they are the entities of choice for many spatial analyses, because of MAUP, and for relative simplicity. These entities are briefly described here (U.S. Bureau of the Census, 1990). Their spatial relationships are illustrated in Figure 3-5.
Figure 3-4 Hierarchy of U.S. Census Geographic Entities
(Number inside parentheses denote mean population)

Metropolitan

Central City

Census Tract (4,000)

Block (85)

Block Group (1,000)

Nonmetropolitan

County

Block Group (1,000)

Block Numbering Area (4,000)

Block (30)

(National Center for Geographic Information and Analysis, 1993)
**Census Tracts** - Census tracts are small areas with generally stable boundaries, defined within counties and statistically equivalent entities, usually in metropolitan areas and other highly populated counties. They are designed by local committees of data users to be relatively homogeneous with respect to population characteristics, economic status, and living conditions at the time they are established. Census tracts average 4,000 persons, but the number of inhabitants generally ranges from 2,500 to 8,000 persons.

**Block Groups (BG's)** - BG's are combinations of census blocks within census tracts and Block Numbering Areas (BNA's are the non-metropolitan equivalent of census tracts). All the blocks in a BG have the same first digit in their identifying numbers; e.g., BG 4 contains all blocks numbered from 401 to 499 in a census tract or BNA. The entire nation and its territories are subdivided into blocks for the first time in the 1990 census.

**Blocks** - These are the smallest geographic units for which the Census Bureau tabulates data. For the 1990 census, the Census Bureau numbered blocks throughout the nation and its territories for the first time. Many 1980 census blocks were revised and renumbered to meet the requirement that their boundaries follow visible features such as streets, streams, and railroad tracks, as well as to reflect new and corrected street patterns. Unlike the 1980 census, blocks are not split between geographic entities; rather, a unique three-digit block number, sometimes with an alphabetic suffix, applies to each entity. For example, the 1980 census reported data for the place and nonplace portions of block 101; in the 1990 census, there are data for two specific block numbers: 101A inside the place and 101B outside. For the 1990 census, the entire United States and its territories were divided into more than 7 million blocks. For the 1980 census, data were tabulated for only 2.5 million blocks; in nonblock-numbered areas, the ED, usually covering a much larger area than a block, was the smallest area for which the Census Bureau tabulated data.
Census Tracts are Defined Using Physical Boundaries

(U.S. Bureau of the Census, 1990)
An important feature of the U.S. census to be pointed out at this stage is that an effort is made to keep the average populations at census tract and block group levels more or less constant. This means that boundaries at those aggregation levels can and will change. This is one feature which differentiates the U.S. data set considerably from that of Japan.

3.3 Comparing Inherent Characteristics of the Two Representations

It can be argued that many of the differences between the census processes in the two countries stem from the different motives and conditions for taking the census. As stated earlier in this paper, the mesh format of the census data in Japan was introduced as an analysis tool to estimate and optimize social and economic growth. In comparison, the U.S. census tracts and block groups are created with a degree of population homogeneity and enumeration ranges as goals, so their areas may change over a period of time.

The mechanical nature of the mesh system does not take into account the various "exceptions" that occur in any landscape, which can prevent one mesh square from reflecting the characteristics of the greater area of its vicinity. An example of such a situation is the park situated amidst a downtown business district. If a mesh happened to be situated directly over that park (assuming that the park is approximately 1km²), that mesh would not possess many of the characteristics of the surrounding cells. While this aberration would be treated at face value for certain analyses, the characteristic of the larger surrounding city blocks may be desired for other analyses. Further, an artificial smoothing of the aberrational and surrounding mesh squares to dampen the contrast may or may not be feasible, depending on the nature of the problem.

Another issue particular to the Japanese mesh system is the formula for assigning each census research unit of approximately 50 households to each mesh square. Over the years, various formulas have been implemented to assign those research units that were split by mesh boundaries. With the majority of these formulas, it can't be helped that the
resulting mesh squares really aren't true squares; the rigorous nature of the square will always be compromised by the shape of the smaller research units themselves. Nevertheless, the mesh are assumed to be perfect squares, resulting in a difference between the TRUE nature of the area within the perfect square, and the REPRESENTATION of that same area using mesh. This represents a problem when a high degree of accuracy in individual mesh representation is of the researcher's highest priority.
Chapter 4:
Counting of Populations Using Census Data

This chapter discusses the procedures and the issues that arise when Japanese and U.S. census data are applied to perform population counts of a specified area such as municipalities, neighborhoods, or electoral districts. Chapter 3 described the concept of dividing the U.S. and Japan into geographic entities that spatially represent all areas, but the process of assigning attribute data, such as population, to these geographic entities is not a trivial task. This assignment process, as well as the nature of the geographic entities themselves, significantly affect the enumeration process for both census formats.

4.1 Methods for Counting People

An array of methods that exist throughout the world for a population count can be generalized into three types. In places where no on-ground tabulations are available or collected, the areal extent of an urbanized area gained from satellite images is used to estimate population. This is often used in cities of the developing world, and is associated with some obvious problems. The expected level of error becomes quite significant at local levels and it is virtually useless for counts of sparsely populated areas. Information regarding socio-economic and demographic breakdowns, as well as temporal changes cannot be obtained. However, this may be the only available method for estimating population in countries where neither a census is held nor a population register is maintained.

Residents of most western European countries and Japan, among others, submit information to local population registers. Such registers, normally administered by municipalities, hold information regarding individuals' postal address as well as typical identifiers such as name, birth date, sex, marital status and unique identification numbers. Therefore, a municipality always will refer to their own population registry for questions
regarding its population. Such registers can maintain high accuracy only if residents dutifully report changes, typically changes of address, marriage, births and deaths. Many countries use the tactic of forced encouragement in order to update registry information. For example, some countries use the registry to identify and send out social security payment information, and others may use the registry to identify women of certain ages who need to be informed of cancer screening procedures. The extreme case of absolute reliance on registry data for population statistics is Denmark, where traditional census activities have been abandoned in favor of 37 registries that hold statistics for the national population. The high accuracy level of their registries is reflected by the generally satisfactory level of public service delivery. Accuracy of these registers differ among countries, however. The Swedish register and census data from 1980 revealed that only 0.3 percent of people recorded in the registers were recorded incorrectly. On the other hand, Italians and Spanish have been known to be slow to notify changes. In Japan, where registry statistics are known to be highly accurate (updating is considered very important, as this information is necessary for school/job application, social services, marriage, and other various activities), address-linked population counts can theoretically be performed using these figures, although the data is completely closed from public access.

The third method of discussion is that of counting using census data. In the case of Japan, it has already been established that its registry is highly accurate, and queries such as population figures for incorporated municipalities can be satisfied with very high accuracy; the limitations of the rigorously defined census mesh squares to handle natural and political boundaries has been discussed. However, when performing population counts of municipal sub-areas or areas that are specified by other boundaries, the use of the registry becomes quite a complex task. First, the registry data must be related spatially in order to query only the records within the specified boundary. This may be manageable for small areas at a time, but considering that each Japanese resident is
supposedly recorded in all of the registries, the geocoding of the national population is far too time and resource consuming, not to mention the constant updating process that would be required. This reality makes the population count using spatially oriented census data very feasible. In the U.S., the census is the primary data source for population counts; in fact, counting is the primary mission of the U.S. census as prescribed by the country's constitution.

4.2 Relating Census Data and Spatial Representation

The census attribute data is initially collected and processed independently from associated census geographic entities. At the most basic level, enumerations are performed by individuals who cover their assigned enumeration areas. These enumeration areas, which normally are clearly defined, are then aggregated to form census geographic entities at the lowest aggregation level. Theoretically this is a trivial task; however, the reality is that the real world is not demarcated with and comprised of census geographic entities. Described in this section are some of the difficulties involved in assigning objects to their correct census areal units. Each of these complexities will in turn affect enumerations of client-specified areas in different ways.

4.2.1 Assigning Attribute Data to Japanese Census Mesh Squares

First, the decision making process for assigning basic census areas to the mesh squares must be discussed, as this is the primary source of inherent counting error. Although the mesh grid system had been instituted for the first time only in 1970 with a total of five census years from 1970 to 1990, the rules for census area assignment have changed four times, each an attempt to undo the wrongs of the previous. Particularly the most recent census in 1990 experienced a dramatic rule change, as the definition and the characteristic of the basic census areas were altered.
The basic census area was initially defined as building blocks for mesh squares in the 1970 census. Each area is enumerated by one enumerator from the Bureau, and each area contains an average of 50 households. Since this was the only requirement, these census areas possessed the following characteristics:

- Areas were relatively small in densely populated areas while relatively large in sparsely populated areas.
- Mesh Squares in densely populated areas were often composed of multiple census areas, while census areas included multiple mesh squares in sparsely populated areas.

In 1990, the basic census areas were redefined as the equivalent of U.S. blocks for areas that are urbanized and developed in block units, and other areas were defined using natural or artificial boundaries. This change occurred because the initial definition caused boundaries of census areas to change between census years to maintain the mean household number. Blocks are, for the most part, consistent over a period of time.

Below, the assignment rules for each census year are described.

**1970:** For mesh squares that completely enclose five or more basic census areas (i.e. urbanized regions), census areas that lie on mesh square boundaries were assigned to the mesh square in which the largest areal proportion lie. In non-urbanized regions, individual households were assigned to mesh squares depending on their locations. For the Tohoku, Hokkaido and Southwestern regions, all assignments were decided according to the latter, non-urbanized version.

**1975:** The 1970 assignment rules proved to be much too time-consuming. This time, the population centroids for each basic census area was calculated and, using a computer digitizer to read their coordinates, assigned them to mesh squares that enclosed the centroids.

**1980:** This time, officials were concerned with the little regard for accuracy that was shown in 1975. The population centroid was still used for census areas under 1 km², but larger areas that were divided by mesh square boundaries were separated where they were
crossed by the boundaries, with each partition being assigned to their respective mesh squares.

1985: Same as 1980.

1990: Same as 1980, but using the redefined basic census areas.

Regardless of the methods implemented, it is inevitable that the assigning of these areas to rigorously defined squares will result in some errors at all boundaries. It is likely that assignment decisions improved over time, however. Koshizuka (1984) proposed a model relating reported mesh population and its error relative to the actual mesh population, with 95% confidence, as follows:

$$\left| \frac{P_{\text{reported}} - P_{\text{actual}}}{P_{\text{actual}}} \right| \leq 8 \left\{ 0.252 (\alpha^2 + \sigma^2) + 0.13 \alpha \right\}^{0.5} / (P_{\text{actual}}^{3/4} \pi^{1/2} \alpha),$$

where $\alpha$ is the average population of all basic census areas associated with the mesh, and $\sigma^2$ is the variance of the basic census area populations. $P_{\text{actual}}$ is calculated by aggregating basic census areas according to the method used in 1970 for non-urbanized regions. It is shown here that the relative error is directly related to the $-3/4$ th power of the actual population. Koshizuka shows that a counting error of less than 10% with more than 95% confidence can only be achieved with mesh squares of more than 10,000 people, using the 1975 1km$^2$ mesh squares of Utsunomiya City, Japan in his model. However, Ihara et. al. (1994) puts the threshold at 3,000 people, using 1990 1km$^2$ mesh squares for the same city. This perceived improvement is significant, and Ihara et. al. attributes the change to improving methods for aggregating basic census areas over time. The function relating mesh population and relative percent error with 95% confidence is illustrated in Figure 4-1.

4.2.2 Assigning Attribute Data to U.S. Census Geographic Entities

The issue of spatially representing attributes for the U.S. census data requires much terse discussion. Aggregation and disaggregation within the geographic hierarchy requires no forced fitting of sub regions into rigorous aggregates or a partition of a region.
Figure 4-1 Mesh Enumeration Error Distributions

Nominal Error (# of people)

(a) Nominal Error for Each Mesh in Utsunomiya City

Percent Error

(b) Percent Error for Each Mesh in Utsunomiya City
(c) Percent Error Range with 95% Confidence for Each Mesh in Utsunomiya City

(Ihara, Iwasa, Yamaguchi, 1995)
into smaller disaggregated sub regions with estimated population shares. Each level change is merely spatial and attribute addition or subtraction. It is assumed that a highly accurate count can be achieved by summing the populations of the blocks that compose the more aggregated areal unit in question.

4.3 Counts of Specified Regions Using Census Data

Population counts of specified regions can be performed using census data. The immediate difficulty is that of census areas and their aggregations not matching exactly the specified regions. The Japanese and U.S. census data are not exempt from this problem, although the U.S. census tracts do not cross county boundaries in most situations. It can be said that the issue becomes acute when performing counts of communities and areas of the sub-county level in the U.S. because of the aforementioned characteristic of census tracts. Furthermore, counting for smaller areas and sparsely populated areas tend to result in larger "representation error" levels. Here, representation error is defined as an error caused when objects are assigned as attributes of some area, when in reality those objects are really attributes of a different area. Representation errors will form a larger percentage of the total actual enumeration in small and sparse areas. This can be said for both the U.S. and Japanese data formats. With this in mind, the issues of enumeration are presented here.

4.3.1 Enumeration with U.S. Census Data

As stated before, U.S. block groups or census tracts aggregate exactly to County areas, so enumeration issues become important for sub-county enumeration. The following generalizations can be made, and are illustrated using real census data.

- Block/Block group level enumeration are known to be highly accurate, and thus are the best building blocks for enumeration of incorporated and unincorporated areas. This process, while certainly time-consuming, is the best method to insure the highest
degree of enumeration accuracy. It is also very unlikely that a political boundary would split blocks in urban and suburban areas. This insures that enumeration of cities and incorporated towns can be achieved with accuracy using blocks. Areas that are not incorporated will also benefit from the use of blocks for enumeration. The small blocks will come the closest to providing boundaries that encloses the specified area.

- *Census tracts are generally bound by physically stable features, such as roads, rivers, and coastlines.* Since most unincorporated areas to be enumerated are bound by such features, the definition of census tracts proves to be convenient for this purpose.

### 4.3.2 Enumeration with Japanese Census Data

It has been discussed previously that enumeration errors result from the aggregation of basic census areas to form the 1km² mesh squares. This inherent error will affect all attempts to perform population counts in areas of all sizes. The enumeration complexities that arise from the spatial representation of grid squares are separate issues, and are described here.

- *Grid boundaries do not take into account any physical or political boundaries.* While this intended effect is justified for other purposes, this characteristic makes it difficult to conduct enumeration of any specified area using mesh squares. When a specified boundary splits a mesh square, one must estimate the proportion of the specified area that is contained within the grid, and apply the ratio for the attribute values. However, even this procedure is tricky, as the Japanese mesh squares do not possess the consistent nature of the U.S. census tracts.

- *While U.S. census tracts were relatively consistent in terms of population, and socio-economic homogeneity, the grids possess none of those characteristics.* This necessitates the referencing of a map with a grid overlay in order to get an idea of the actual characteristic of an area within a grid. In particular, grids that contain a
significant amount of uninhabited land, special facilities (such as an airport), or a portion of a body of water (such as a grid that is split by a coastline) need special attention. This is especially true when enumeration figures are used to calculate population density, or categorizing of settlements (urban, suburban, or rural).

4.4 Discussion

The degree of accuracy by which census data can present an enumeration of attributes (like population) undoubtedly affects all analyses and other tasks that utilize census data. If enumeration is a stated mission for census data (such as in the U.S.), then its ability to provide highly accurate counts is crucial. In other analyses that use census data, its ability to provide accurate counts validates the analyses; otherwise, it must be noted as a disclaimer.

This chapter first discussed the assignment of basic census areas to create census geographic entities, and then the processing of these geographic entities to provide population counts of client specified areas. In both discussions, it became evident that the inherent characteristics of Japanese mesh data caused difficulties. The rigorously defined mesh squares provided difficulty both when aggregating basic census areas to create mesh squares, and when squares were applied to recreate client specified areas. A possible recommendation for alleviating these problems is the use of land use maps with mesh overlay; in order to identify aforementioned problems such as lack of homogeneity, and grids split by coastlines. This way, researchers will have more spatially fine attribute information; for example, instead of merely knowing that an attribute value corresponds to a mesh square, he would recognize that 53% of the 1km² area is sea water, and thus the attribute value corresponds to the remaining 47% of the mesh square. Accurately digitized land use maps would facilitate such a procedure, but consistent and comprehensive land use information for the entire country is still limited to paper maps produced by the Geographical Survey Institute of Japan.
The following three chapters will discuss census data use within the contexts of other urban related problems. Enumeration issues, although not be explicitly stated in subsequent chapters, play critical roles in these and other analyses.
Chapter 5: 
Measuring Population and Employment Changes Over Time

Census data is often used to gauge demographic changes over time. In general, the two common temporal analyses are the comparisons between data in different census years, and forecasts or projections based on existing figures. When performing these tasks within the context of the two census formats, distinct issues and complexities arise for each format. The modifiable areal unit problem (MAUP) is also an issue for both formats, although the degree by which each format is affected is not obvious.

5.1 Non-static Geographic Entities Over Time

Areal units such as the Japanese grid squares and the U.S. census tracts and block groups are examples of "ready-to-go" units, or areal units that have been defined by the census format and have attribute data tabulated along them. Therefore, without taking the MAUP or real boundaries of defined places into account, utilization of these units when performing spatial analyses, including temporal comparisons, requires the least amount of additional data processing. In this spirit, it is easy to see that the rigorous and unchanging boundaries of Japanese grid squares accommodates temporal comparisons.

However, an acknowledged problem with U.S. census tracts is that their boundaries change occasionally. Since census tracts strive for consistency amongst themselves, were meant to contain between 2,500 and 8,000 people each and have visible physical boundaries (such as a road), tracts often need to be added, removed, split, or have their boundaries shifted as population and physical characteristics change. According to the Bureau of the Census in 1980, 2,762 tracts (8%) throughout the United States were split (Howenstine, 1993).

In the case of split tracts, it becomes necessary for the researcher to adjust either the data that contains the split tract or the data from the census year that is being compared to
the data containing the split tract. One solution is to aggregate block or block group areas from the census before the split tract took place and match the new tract boundaries. This enables the data from the older year to replicate the split tracts of the recent year, but the sheer volume of calculations makes for a very tedious process. The more common solution is to simply aggregate the newly split tracts and remove the internal boundaries. The primary problem with this method is that the joined area becomes so large in population, well beyond the originally intended size. Therefore, it is useful to consider methods in which the older, aggregated tract is divided along the new boundaries, with the populations within each division estimated. Howenstine provides an error estimation analysis for two methods that reconcile split census tracts in Cook County, Illinois between 1970 and 1990. The following two methods were compared:

1. Splitting the 1970 tract along the 1990 split tract boundaries, then assigning populations based on the areal proportion of the divided areas.
2. Splitting the 1970 tract along the 1990 split tract boundaries, then assigning populations based on the population distribution of the split tracts in the later census year.

Cook County grew in population by 12% during the two decades, and the number of census tracts increased from 1,116 in 1970 to 1,352 in 1990. Figure 5-1 shows the ten 1970 tracts split into 31 1990 tracts. Table 5-1 summarizes the estimation errors for ten Cook County tracts that were split into 31 tracts between the two census years. For the first (1990 Area) and the second (1990 Population) methods, the 1970 census block data were used to aggregate for the actual data within each of the 1970 sub-areas, which was compared to the 1970 population that were distributed to the sub-areas according to proportion of land area and population distribution in 1990, respectively.

Also, since the selected Cook County tracts may not represent all possible physical circumstances, six scenarios which represent extreme situations of tract splitting necessity were developed. For each scenario, the two proposed methods were tested for estimation errors. The scenarios are described below (Howenstine, p. 427-8):
Figure 5-1  Split Census Tracts in Cook County, 1970-1990

(Howenstine, 1993)
Table 5-1. Population Estimation Errors (Percent of Population Misassigned) for 31 Suburban Cook County Census Tracts Using Two Methods of Estimation, 1970-1990

<table>
<thead>
<tr>
<th>Census Tract</th>
<th>FIPS Code</th>
<th>Number of Split Tracts</th>
<th>Growth Rate (%)</th>
<th>1990 Area</th>
<th>1990 Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8030</td>
<td>8</td>
<td>82</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>8043</td>
<td>6</td>
<td>35</td>
<td>58</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>8049</td>
<td>2</td>
<td>-27</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8080</td>
<td>2</td>
<td>-13</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>8171</td>
<td>2</td>
<td>-9</td>
<td>36</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>8157</td>
<td>2</td>
<td>-16</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>8231</td>
<td>2</td>
<td>-9</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8241</td>
<td>3</td>
<td>253</td>
<td>39</td>
<td>45</td>
</tr>
<tr>
<td>9</td>
<td>8262</td>
<td>2</td>
<td>10</td>
<td>62</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>8302</td>
<td>2</td>
<td>23</td>
<td>23</td>
<td>31</td>
</tr>
</tbody>
</table>

Average Error for Ten Tracts
Combined Error for Cumulative Area

Methods (% error)

(Howenstine, 1993)
1. There is no change in the population size but a recognition of internal dissimilarity; for example, an existing park may be partitioned from a dense residential area.

2. Population increase is equal numerically and proportionally in both parts of a split tract. This may occur if a uniform but moderately settled area becomes densely populated during this decade.

3. The population increase is equal in proportion but not in number; for example, when a village on one end of the tract and a sparse rural area on the other each double in population.

4. The population increase is equal in number but not in proportion. Using the example above, both the village and the rural area would gain the same number of people.

5. Population growth is uneven. There are many possible combinations, but here the sparse rural area develops quickly, overtaking the village in density. The village itself experiences only slow growth.

6. An entirely uninhabited half of a tract becomes densely settled, matching the density of the rest of the tract. This situation may occur if farmland adjacent to a subdivision becomes fully developed, or a park area shifts to dense residential use.

Table 5-2 summarizes the population estimation errors for the six hypothetical scenarios using the two methods of estimation. In each scenario, two parts of a tract are assigned hypothetical populations for two census years, 1970 and 1990. The 1970 population is reallocated using the two methods, Area and Population. For simplicity, one tract was split into two parts of equal area for all six scenarios. The error figures are the percent errors of people in 1970 who were misassigned.

Based on the results, it seems that an estimation according to area proportions is only useful when population is uniformly and proportionally changing. The Population method is most accurate when a tract is split merely to recognize a preexisting condition (such as a park) or split as a result of even growth throughout the study area.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1990 Area</td>
</tr>
<tr>
<td>1. No change in population</td>
<td>5000</td>
<td>5000</td>
<td>50</td>
</tr>
<tr>
<td>2. Uniform and proportional increase</td>
<td>2500</td>
<td>5000</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>5000</td>
<td></td>
</tr>
<tr>
<td>3. Proportional but not uniform</td>
<td>1000</td>
<td>2000</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>8000</td>
<td></td>
</tr>
<tr>
<td>4. Not proportional but equal increase</td>
<td>1000</td>
<td>3000</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>6000</td>
<td>13</td>
</tr>
<tr>
<td>5. Uneven increase</td>
<td>1000</td>
<td>6000</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>5000</td>
<td>35</td>
</tr>
<tr>
<td>6. Development of undeveloped area</td>
<td>0</td>
<td>5000</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>5000</td>
<td></td>
</tr>
</tbody>
</table>

Average misassignment (% error) across all scenarios

33 16

(Howenstine, 1993)
Howenstine's presentation of the six scenarios might lead some readers to frame the issue of population change within an areal unit as six unique and equally possible characteristics. However, are all six characteristics equally likely to occur in census areal units? The general lack of symmetry in the real world suggests that scenarios such as "Uniform and proportional increase" within areal units are quite rare. It is likely that, in most areas, population growth within census areal units are asymmetrical and not proportional. Researchers should not take for granted that Howenstine's six scenarios are equally distributed; rather, they need to determine the scenario or scenarios that are prevalent in their areas of analyses. Additional land-use information, in the form of land-use maps or remote sensing data between census years, sheds light on the development characteristics for census areal units.

Examination of census data over time is performed by a researcher who, for example, seeks rates of population and employment changes in a neighborhood in order to measure the effects of housing or economic policies over a period of five years. It is vital for such researchers to formulate a consistent and accurate method for representing the same area in multiple census years.

5.2 Forecasting and Estimations for Non-census Years

Projections of population, employment and some other figures are available at the state/prefecture or national levels in both the U.S. and in Japan. However, projections for small areas are necessary in order to perform local analysis involving non-census years. However, the very nature of the small, local areal units cause unique problems when the small areas are used as bases for projection. Local area data availability is limited (U.S. census is decennial, Japanese is every five years), and accuracy is questionable, as small areas are subject to greater enumeration errors than larger areas. Also, events such as construction of new buildings, or the loss of vacant land will have great impact on population growth rates for small areas. This sensitivity makes for a high degree of
uncertainty when projecting for small areas. Finally, both the U.S. and Japanese census data formats possess inherent characteristics which make small area projections difficult. The boundaries of U.S. census tracts change over time, and this makes for a tedious adjustment process (discussed in 5.1). The Japanese grid data contain enumeration errors that result from aggregating irregularly shaped areas to create mesh squares, thus calling the very idea of using mesh data for forecasting into question.

However, projections for small areas have been examined. Projections using U.S. census tracts seem much more feasible, especially when one may simply ignore tracts that change boundaries and given that individual enumerations are thought to be relatively accurate. Also, the need arises from the fact that the U.S. census is performed only once every ten years. Typical methods for census tract projection are outlined by Smith and Shahidullah (1995) as below, for projections using 1970 and 1980 populations as a base for forecasting 1990 figures for three Florida counties:

1. Linear extrapolation (LINE) - assumed that each tract will grow or decline by the same number of persons between 1980 and 1990 as between 1970 and 1980.
2. Exponential extrapolation (EXPO) - assumed that each tract will grow (decline) at the same percentage growth rate between 1980 and 1990 as between 1970 and 1980.
3. Shift share (SHIFT) - assumed that each tract's share of county population will change by the same numerical amount between 1980 and 1990 as between 1970 and 1980.
4. Share of growth (SHARE) - assumed that each tract's share of county population growth will be the same between 1980 and 1990 as between 1970 and 1980.
5. Average (AVG) - Average of projections from the above four techniques.
6. Composite (COMP) - An average of techniques that were selected according to the population growth characteristics of each tract.

Three different measures were used to compare the accuracy of each projection method. Mean absolute percent error (MAPE) is the average error when the direction of error is ignored. Mean algebraic percent error (MALPE) accounts for the different signs.
Since extreme errors with one sign can significantly affect the MALPE, the proportion of positive errors (%POS) was also used.

Judging from the tabulated results, some generalizations can be made. According to Table 5-3, MAPE's for all techniques decline as population size of the tracts increases. Also, as Table 5-4 illustrates, there is a concave-up U-shaped function that relates the population growth rate and the MAPE. Errors were higher for extreme growth and decline ends, and were smallest where there was moderate or no growth or decline. Finally, Table 5-5 illustrates that places losing population during the base period are likely to be underprojected and place growing rapidly are likely to be overprojected. This is because very high or very low growth rates usually do not persist for long periods; rather, they tend to move toward more moderate levels (Smith, p. 67, 70).

Projection and forecasting for small areal units is most feasible when figures for non-census years are absolutely necessary for whatever purpose, and when the situation can tolerate the extent of errors that were described above. It was shown here, however, that projection and forecasting is difficult and error-prone for small areas, regardless of the census format used. Any assertions referring to projections of small areas should be regarded as qualified statements, and the accuracy of projections should always be questioned.

5.3 Meaningful Areal Units for Urban Research

"What is happening where over a period of time?" is a typical question form that requires time-series analysis. For example, there is a possibility that the city of Hakodate experienced a "donutization" of its nighttime population in the densely developed part of the city, as residential areas there give way to industrial and office properties. A quick way to check for such movements is to calculate % changes in daytime and nighttime populations. Areas that show large decreases in their nighttime populations were likely to have been areas where donutizations occurred.
Table 5-3. Mean Absolute Percent Errors for Census Tracts by Population Size in 1970:
Three-County Sample

<table>
<thead>
<tr>
<th>Size</th>
<th>N</th>
<th>Linear</th>
<th>Exponential</th>
<th>Shift</th>
<th>Share</th>
<th>Average</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2500</td>
<td>37</td>
<td>34.5</td>
<td>44.4</td>
<td>44.1</td>
<td>32.4</td>
<td>36.5</td>
<td>32.2</td>
</tr>
<tr>
<td>2500-4999</td>
<td>165</td>
<td>19.3</td>
<td>26.6</td>
<td>27.6</td>
<td>16.7</td>
<td>21.5</td>
<td>16.9</td>
</tr>
<tr>
<td>5000-7499</td>
<td>147</td>
<td>17.6</td>
<td>25.5</td>
<td>24.3</td>
<td>15.2</td>
<td>19.6</td>
<td>15.7</td>
</tr>
<tr>
<td>7500+</td>
<td>72</td>
<td>17.2</td>
<td>25.4</td>
<td>20.7</td>
<td>15.5</td>
<td>18.9</td>
<td>15.5</td>
</tr>
<tr>
<td>Total</td>
<td>421</td>
<td>19.7</td>
<td>27.6</td>
<td>26.7</td>
<td>17.3</td>
<td>21.7</td>
<td>17.6</td>
</tr>
</tbody>
</table>

(Smith et. al., 1995)
Table 5-4: Mean Absolute Percent Errors for Census Tracts by Population Growth Rate 1970-1980: Three-County Sample

<table>
<thead>
<tr>
<th>Growth Rate</th>
<th>N</th>
<th>Linear</th>
<th>Exponential</th>
<th>Shift</th>
<th>Share</th>
<th>Average</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; -10%</td>
<td>81</td>
<td>27.5</td>
<td>16.4</td>
<td>49.2</td>
<td>22.1</td>
<td>27.7</td>
<td>21.8</td>
</tr>
<tr>
<td>-10% - 0%</td>
<td>68</td>
<td>10.5</td>
<td>10.3</td>
<td>22</td>
<td>10.2</td>
<td>12.5</td>
<td>10.3</td>
</tr>
<tr>
<td>0%-10%</td>
<td>75</td>
<td>9.9</td>
<td>10</td>
<td>14.1</td>
<td>9.5</td>
<td>9.4</td>
<td>9.8</td>
</tr>
<tr>
<td>10% - 25%</td>
<td>70</td>
<td>13.8</td>
<td>15.9</td>
<td>9.2</td>
<td>12.2</td>
<td>12.3</td>
<td>11.3</td>
</tr>
<tr>
<td>25% - 50%</td>
<td>43</td>
<td>23.1</td>
<td>27.3</td>
<td>21.1</td>
<td>21.5</td>
<td>23.1</td>
<td>21.8</td>
</tr>
<tr>
<td>50% +</td>
<td>84</td>
<td>31.3</td>
<td>77.9</td>
<td>37.6</td>
<td>27.6</td>
<td>41.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Total</td>
<td>421</td>
<td>19.7</td>
<td>27.6</td>
<td>26.7</td>
<td>17.3</td>
<td>21.7</td>
<td>17.6</td>
</tr>
</tbody>
</table>

(Smith et. al., 1995)
Table 5-5. Mean Algebraic Percent Errors for Census Tracts by Population Growth Rate 1970-1980: Three-County Sample

<table>
<thead>
<tr>
<th>Growth Rate</th>
<th>N</th>
<th>Linear</th>
<th>Exponential</th>
<th>Shift</th>
<th>Share</th>
<th>Average</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; -10%</td>
<td>81</td>
<td>-26.7</td>
<td>-13.4</td>
<td>-49.2</td>
<td>-20.3</td>
<td>-27.4</td>
<td>-20.2</td>
</tr>
<tr>
<td>-10% - 0%</td>
<td>68</td>
<td>-0.9</td>
<td>-0.6</td>
<td>-17.7</td>
<td>0</td>
<td>-4.8</td>
<td>-0.5</td>
</tr>
<tr>
<td>0% - 10%</td>
<td>75</td>
<td>3.7</td>
<td>4</td>
<td>-10.3</td>
<td>3.1</td>
<td>0.1</td>
<td>3.6</td>
</tr>
<tr>
<td>10% - 25%</td>
<td>70</td>
<td>12.2</td>
<td>14.7</td>
<td>4</td>
<td>10.2</td>
<td>10.3</td>
<td>8.8</td>
</tr>
<tr>
<td>25% - 50%</td>
<td>43</td>
<td>13.1</td>
<td>21.1</td>
<td>10.9</td>
<td>9.6</td>
<td>13.7</td>
<td>11.2</td>
</tr>
<tr>
<td>50% +</td>
<td>84</td>
<td>16.5</td>
<td>73.3</td>
<td>25.7</td>
<td>10.6</td>
<td>31.5</td>
<td>13.6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>421</td>
<td>2</td>
<td>17.3</td>
<td>-7.2</td>
<td>1.4</td>
<td>3.4</td>
<td>2</td>
</tr>
</tbody>
</table>

(Smith et. al., 1995)
However, attention needs to be focused on the relationship between the nature of the population movement and the defined areal units. It is assumed that a trend of commercialization of a neighborhood, municipality or region is the result or the combinations of 1) personnel or locational decisions by local employers; 2) completion of project by a real estate developer; and 3) development policy decisions made by public officials at the municipal or regional level. Then, the donutization trend should appear in units of neighborhood, municipality, or region. Moreover, the typical urban researcher or concerned citizen who await the results of his analysis are interested in the trends that are taking place in those same units. Typical questions may be asked in ways such as: what is happening to the generations-old residential community comprised of these five blocks; or: how does the population migration in central Hakodate compare to that in Sapporo?

In light of the above, areal units to be used for this question should reflect the areal units that make the research meaningful. Does the Japanese mesh square representation satisfy these expectations? The census area for the central portion of Hakodate city and its adjacent areas are presented to illustrate. Figure 5-2 shows the mesh squares that contain the central Hakodate area. From the perspective of urban/municipal analysis, what meaningful conclusions can be made from these figures? Individual calculations for squares cannot pinpoint the population changes that are taking place in a particular neighborhood occupying a portion of a square, or a neighborhood split by a square boundary; it is almost impossible to replicate such neighborhoods by merely aggregating mesh squares. Due to this limitation, they are also not able to allow for the comparison between the traditionally densely populated region within the ring road and the outlying areas. Finally, squares by themselves are also not capable of producing figures for an entire municipality (Hakodate city) for comparison with other cities. This problem illustrates the need for fundamental census areal units that can be aggregated to form regions that match or reasonably proximate meaningful units from the urban research perspective. Some may argue that any declaration of "meaningful" units ignores the
Figure 5-2 Mesh Squares in Central Hakodate City, Japan

(Japan Geographical Survey Institute, 1989)
modifiable areal unit problem inherent in all areal representations. However, often in urban research, the usually elusive answer for the "which areal representation is meaningful?" question is already a given; that is, geographic entities whose characteristics are of interest to the urban researcher.

5.4 (Mis) representation Using Maps

The same data also illustrates a long recognized and inherent characteristic found in other spatial data representations as well. Often, results of spatial analyses are presented using thematic maps, with a legend providing numerical ranges and the corresponding symbols that denote numerical values. Depending on the design of the thematic maps, the spatial analyses can suggest a countless number of different findings.

Figure 5-3 and 5-4 are two thematic maps that suggest the notion that the region can be separated into three and two worlds, respectively. The first map separates the mesh squares by "population-losing", "population-gaining", and "insignificant change" (less than 0.5% change). The second map separates the mesh squares using the mean value as a boundary. Both approaches can be justified; while the first map is purely concerned with absolute signs, the second map takes a simple statistical approach. The findings suggested by these two maps differ significantly.

Similar to the MAUP, the so-called optimal representation for thematic maps is elusive, and efforts to seek it may be futile. If some pre-defined numerical range that is meaningful for some purpose exists, then those ranges should be used for these maps. For example, if a particular policy outlines different measures for areas with absolute gains and losses in population, then the approach taken in Figure 5-3 is appropriate. In the same spirit of Openshaw (1984)'s suggestion in dealing with the MAUP, this inherent characteristic of thematic mapping should not be viewed in dismay, but be utilized and even exploited as a powerful tool for analysts to illustrate their positions.
Figure 5-3  Thematic Map for Hakodate City
% Change in Night Population, 1980-1985

% Change in Population (1980 - 1985) > 0%
% Change in Population (1980 - 1985) < 0%
Insignificant Change in Population
No Data Available (Ocean)
Figure 5-4 Thematic Map for Hakodate City
% Change in Night Population, 1980-1985

% Change in Population (1980 - 1985) > Mean Change for All Squares
% Change in Population (1980 - 1985) < Mean Change for All Squares
No Data Available (Ocean)
5.5 Discussion

The ability of census formats to provide for comparative analyses between census years is important for various urban problems, such as projecting transit ridership for consecutive years including non-census years. Reconciliation of census areal units whose boundaries were altered over time, as well as forecasting and projection methods for census areas have been discussed, although the resulting accuracy is questionable and, the errors being too great for many situations.

The issues of meaningful areal units and thematic mapping methods are by no means confined to temporal analyses. However, the motivation for altering boundaries of U.S. census areas is for the enabling of meaningful areas to be created. U.S. census areas are designed to exhibit attribute values within consistent ranges among all areas, as well as to display homogeneity within each area. Alterations such as boundary shifts and split tracts occur when significant demographic change occurs (such as large population growth), or with change in the physical domain (such as the construction of a highway). It is far more likely that such census areas can be aggregated to create areal units that are most useful to the urban researcher. The resulting tradeoff is that while the dynamic nature of U.S. census areas facilitates the creation of meaningful analysis areas, temporal comparisons of static areas become difficult. The reverse is evident with Japanese mesh squares. The static mesh boundaries facilitate temporal comparisons, but the creation of useful areal units for urban research with mesh squares is quite a formidable task, requiring the exhaustive measures of using land use information and estimation to draw new boundaries and assign attribute values.
Chapter 6: Inter-city Transit Demand Models and Census Data

This chapter describes a gravity model that facilitates the selection of an inter-city or commuter rail station location within a large area, based on inter-zonal dynamics. The census areal units will serve as the zones to be used in the model. Previously discussed issues such as enumeration and forecasting complexities cannot be ignored, but these are peripheral to the model itself. Other issues, such as resolution (different levels of census area aggregation) and population homogeneity (maintaining population figures within some range for certain census geographic entities), are issues that directly affect gravity models; these will be discussed in detail here.

6.1 Description of Example: Suburban Rail Station

A commuter rail service that serves a suburban area is proposed. The general dearth of transit options in suburban areas results in any implemented transit to draw passengers from a wide area. This includes walkers, passengers utilizing a personal vehicle to reach a transit stop (park and ride), and passengers transferring from suburban buses. These characteristics place this problem in the category of Macro-area urban problems.

The large area in question is viewed as an aggregation of zones, demarcated by census areal units. It is known that each of these areal units holds a certain number of people who, for whatever reason, would choose to utilize a commuter rail facility if given the opportunity. Furthermore, each areal unit demonstrates different degrees of attractiveness for a commuter rail station within that area. This degree of attractiveness is separate from the number of potential users, and is measured in terms of its intrinsic characteristics such as proximity to a major road, perceived safety of the neighborhood, and availability of parking. Finally, these areal units are separated by known distances. This information
will be used to create a gravity model which suggests a rail station location from among the areal units in the large area.

6.2 Definition of Gravity Model

The suburban area in question is comprised of a total of $N$ areal units. Each areal unit is described by the three factors below.

PRODUCTION ($P_i$) - This variable corresponds to the number of commuters that would utilize a commuter rail service, if given the opportunity, in areal unit $i$.

ATTRACTIVENESS ($A_i$) - This value reflects the intrinsic qualities of areal unit $i$ which make it suitable as a commuter rail station location. This degree of attractiveness is separate from the number of potential users, and is measured in terms of its intrinsic characteristics such as proximity to a major road, perceived safety of the neighborhood, and availability of parking.

IMPEDANCE ($D_{ij}$) - This value measures distance between the centroids of any two areal units, $i$ and $j$.

In the simplest gravity-type model involving just one areal unit $j$, the degree of locational suitability for station location $i$ in relation to areal unit $j$ can be shown by:

$$S_i = f'(A_i) * f''(P_i) / f'''(D_{ij}),$$

where $f'(A)=A$, $f''(P)=P$, and $f'''(D)=D^2$ for the purposes of this discussion. Attractiveness ($A$) and production ($P$) are directly related to suitability, while distance ($D$) is inversely related to suitability. As there are $N$ areal units, however, the sum effects of all areal units on each areal unit in the large area is calculated separately. The summation version of the above equation is shown below:

$$S_i = \Sigma_{j=1}^{N} f'(A_i) * f''(P_j) / f'''(D_{ji}),$$

so that $S_i$ is now a function of production in, as well as distance to, all $N$ areal units, including itself.
To illustrate spatially, a sample area shown in Figure 6-1 is comprised of four areal units \( N = 4 \). \( N \) can be increased to fit real models that may be comprised of more areal units. For attractiveness (A) and production (P), four distinct values, \( A_1..A_4 \) and \( P_1..P_4 \) are counted or calculated beforehand. For \( D_{11} \) thru \( D_{44} \), a matrix of \( N \times N \), or 16 values contains the distances between each of the areal units. The distances are measured between the areal units' centroids (the distance to itself is assumed to be the distance from the centroid to the closest edge).

Finally, the suitability for each of the four areal units can be calculated as follows:

\[
S_1 = f'(A_1) \times \left[ \frac{f''(P_1)}{f'''(D_{11})} + \frac{f''(P_2)}{f'''(D_{21})} + \frac{f''(P_3)}{f'''(D_{31})} + \frac{f''(P_4)}{f'''(D_{41})} \right] =
300 \times \left\{ \frac{435}{0.25} + \frac{525}{1} + \frac{655}{1.96} + \frac{235}{1} \right\} = 911,969
\]

\[
S_2 = f'(A_2) \times \left[ \frac{f''(P_1)}{f'''(D_{12})} + \frac{f''(P_2)}{f'''(D_{22})} + \frac{f''(P_3)}{f'''(D_{32})} + \frac{f''(P_4)}{f'''(D_{42})} \right] =
200 \times \left\{ \frac{435}{1} + \frac{525}{0.25} + \frac{655}{1.96} + \frac{235}{1} \right\} = 620,837
\]

\[
S_3 = f'(A_3) \times \left[ \frac{f''(P_1)}{f'''(D_{13})} + \frac{f''(P_2)}{f'''(D_{23})} + \frac{f''(P_3)}{f'''(D_{33})} + \frac{f''(P_4)}{f'''(D_{43})} \right] =
250 \times \left\{ \frac{435}{1.96} + \frac{525}{1.96} + \frac{655}{0.25} + \frac{235}{1} \right\} = 889,464
\]

\[
S_4 = f'(A_4) \times \left[ \frac{f''(P_1)}{f'''(D_{14})} + \frac{f''(P_2)}{f'''(D_{24})} + \frac{f''(P_3)}{f'''(D_{34})} + \frac{f''(P_4)}{f'''(D_{44})} \right] =
300 \times \left\{ \frac{435}{1.96} + \frac{525}{1} + \frac{655}{1} + \frac{235}{0.25} \right\} = 702,582
\]

The calculations show that areal unit 1 is most suitable as a station location, although its production value is the second lowest among the four areal units.

6.3 Critique of Model and Mesh Representation

The model is now compared within the contexts of the Japanese and U.S. census formats, and the issues of spatial resolution and attribute homogeneity within the areal units.
Figure 6-1  $A_i$, $P_j$ and $D_{ji}$ Values for $N = 4$

($A_i$ and $P_j$ are Assigned Values)

Distance Matrix for $D_{ji}$ Values

\[
\begin{array}{c|cccc}
   & 1 & 2 & 3 & 4 \\
\hline
1  & 0.5 & 1 & 1 & 1.4 \\
2  & 1   & 0.5 & 1.4 & 1 \\
3  & 1   & 1.4 & 0.5 & 1 \\
4  & 1.4 & 1 & 1 & 0.5 \\
\end{array}
\]
6.3.1 Spatial Resolution and Population Homogeneity

Higher spatial resolution generally translates into 1) smaller areal units; and 2) a greater differentiation of the areal units based on distance from each other.

Within the hierarchy of mesh squares, higher resolution facilitates the calculation of the attractiveness (A) factor, because each mesh square will contain smaller areas, and tend to become more homogenous (see Figure 6-2). The larger square contains a vacant lot, a densely developed neighborhood, and a pond. Each of these elements represents different levels of appropriateness for a new rail station; thus, a meaningful attractiveness factor is difficult to calculate. When this square is disaggregated mechanically, the new squares are more homogeneous, and the attractiveness factors reflect their respective areal units, instead of an average number for the larger area. While higher resolution mesh squares seem to facilitate attractiveness factor calculations, the 500m x 500m (highest resolution) squares are not available in most suburban locations; this fact hinders the present analysis.

In the U.S. census, block groups provide the highest resolution with socioeconomic data for suburban regions. In the metropolitan Boston, Massachusetts area¹, for example, block group areas range from very small to 6.15km². If these block groups were sorted by area into quintiles, the 1km² mesh squares, the highest available resolution for all areas in Japan, would belong in the second quintile from the larger extreme. The disaggregation process in the U.S. census is not mechanical as in mesh squares; areal units are disaggregated while population ranges are maintained. An important effect of the resulting population homogeneity among areal units is the higher number of block groups that still contain multiple elements such as a portion of a park and a densely developed neighborhood; block groups that contain park only are rare. Therefore, the most useful

¹ Includes all of Boston proper, as well as Brookline, Cambridge, Somerville, Watertown, Chelsea, Newton, parts of Belmont, and some adjacent areas. Mean block group area: 0.24km².
Figure 6-2 Higher Resolution Facilitates Attractiveness Factor Calculation

- Densely Populated Neighborhood
- Pond
- Original Square
- Squares After Disaggregation
Attractiveness (A) values can be calculated using block groups that have been separated according to different land-use patterns, using land-use data.

Attractiveness and production are intrinsic factors for each areal unit, but their influence upon other areal units is largely dependent on the distances that separate them. Figure 6-3a shows areal unit i, and three other areal units, a, b, and c. From the perspective of areal unit i, the other areal units are differentiated only by three distance values, so that

\[
S_i = f'(A_i) * \left\{ f''(P_i)/f'''(D_{ij}) + f''(P_a)/f'''(D_{ai}) + f''(P_b)/f'''(D_{bi}) + f''(P_c)/f'''(D_{ci}) \right\} = 100 * \left\{ (200/4) + (300/100) + (450/196) + (200/324) \right\} = 5591.3
\]

Figure 6-3b shows the same areas represented by areal units created by disaggregating areal units a, b, and c. Now, the areal units are differentiated by six distance values. Smaller areal units enable a finer differentiation of production (P) by distance, with the new suitability for areal unit i recalculated as below:

\[
S_i = f'(A_i) * \left\{ f''(P_i)/f'''(D_{ij}) + f''(P_a)/f'''(D_{ai}) + f''(P_b)/f'''(D_{bi}) + f''(P_c)/f'''(D_{ci}) \right\} + \left\{ f''(P_{a_1})/f'''(D_{a_1j}) + f''(P_{a_2})/f'''(D_{a_2j}) + f''(P_{b_1})/f'''(D_{b_1j}) + f''(P_{b_2})/f'''(D_{b_2j}) + f''(P_{c_1})/f'''(D_{c_1j}) + f''(P_{c_2})/f'''(D_{c_2j}) \right\} = 100 * \left\{ (200/4) + (100/81) + (200/121) + (150/169) + (300/225) + (50/289) + (150/361) \right\} = 5569.7
\]

Therefore, spatial resolution can significantly affect suitability calculations in gravity-type models.

Finally, the congruent mesh squares are all consistently represented by points. It is obvious that the use of mesh squares facilitates the calculations of the distances between the areal units, because of their square and congruent nature. Special algorithms are necessary in order to calculate distances between irregularly shaped U.S. block groups and census tracts. It is important to realize, however, that this trait of mesh squares becomes a true advantage only when the impedance factor is a function of Euclidean distances. This is not the case in many urban studies and planning related problems, where functions such as road distances and travel times are the appropriate impedance factors.
Figure 6-3 Increased Differentiation Based on Distance Due to Disaggregation

a.

\[ d_{ai} = 10 \]

\[ 4 \]

\[ i \]

\[ A_i = 100 \]

\[ P_i = 200 \]

\[ a \quad b \quad c \]

\[ P_a = 300 \quad P_b = 450 \quad P_c = 200 \]

\[ 4 \]

b.

\[ d_{a_1i} = 9 \]

\[ 2 \]

\[ i \]

\[ a_1 \quad a_2 \quad b_1 \quad b_2 \quad c_1 \quad c_2 \]

\[ P_{a_1} = 100 \quad P_{a_2} = 200 \quad P_{b_1} = 150 \quad P_{b_2} = 300 \quad P_{c_1} = 50 \quad P_{c_2} = 150 \]
6.4 Discussion

An important feature of the gravity model is that it can incorporate a high number of elements that surround the center point of analysis, and differentiate the effects of individual elements by calculating a statistical surface layer based on their distances from the center. The statistical surface can be expanded as desired, even to infinite areas. This feature makes the model appropriate for urban problems in the Macro-area category.

This chapter defined a particular gravity model that is influenced by three factors: production (P), attractiveness (A) and impedance (D). The characteristics of the different census formats affect the assignment of production and attractiveness factor values to particular areas, and the perception of distance between points in the total area. Smaller areal units, disaggregated from larger ones, probably will demonstrate more attractiveness-homogeneity than the larger square, facilitating the calculation of an attractiveness factor that is more meaningful to that particular local area. Disaggregation with the aid of accurate land-use information, in the form of maps or remote sensing data, would further enable the researcher to create small and attractiveness-homogeneous areal units. Smaller areal units also permit an increased differentiation of the areal units based on their distances from the point of comparison; this is advantageous when a particular analysis requires higher spatial resolution. A sampling of block groups in suburban Boston, Massachusetts revealed that typical suburban block groups provide higher resolution than level III (1km²) mesh squares, the most disaggregated level of spatial data in non-urban Japanese areas. When appropriate, resolution may be increased with further manual disaggregation using land-use information.

In many suburban and rural areas, the U.S. census tracts' and block groups' tendencies to fulfill population range requirements often result in larger areal units. It is difficult for single points to represent large areal units, as sub-areas within these areal units may possess rather contrasting attractiveness and production characteristics. This problem is
acute in places where pockets of densely populated developments exist within a sparsely populated region. Large and small areal units co-exist in the same area, causing this problem to be unpredictable and difficult to control. The congruent mesh squares, while not immune to this problem, demonstrate areal consistency within areal unit hierarchy levels.
Chapter 7: Evaluating Usage of Local Transit Stops

It is of interest to transit operators to see whether existing passenger facilities such as stations and depots are being overused or underused. In this chapter, analyses of local transit stops - the likes of buses, trolleys and subways - will be examined within the context of a local transit ridership forecast model. The aspect of the model to be focused upon is the creation of traffic analysis zones (TAZs) using census areal units. It is assumed that the TAZs corresponding to local transit stops will be the areas from which the users of the stops originate. Since local transit modes tend to be accessed by passengers in the immediate vicinity only, this problem is placed in the category of Micro-area urban problems.

7.1 Different Types of Ridership Forecast Models

In general, a study area or a transit corridor is broken up into numerous TAZs, and travel demand models are used to determine transit ridership between the TAZs. A transit trip table is generated as a result of a four-step process: trip generation, trip distribution, modal split and trip assignment models. There is also the final step of calibration which takes into account some errors associated with the particular model and the study area.

There have been a good deal of work done to improve the four-step models, with many new revised versions of the model being proposed. Here, one of these models, the "Period Route Segment" (PRS) Model is presented for discussion. The PRS model, first developed by Batchelder et. al. 1983 and later discussed within the context of GIS integration (Azar and Ferreira, 1995), is described as a relatively simple model which is appropriate for short-range, local routes which would cause little repercussions to the larger system in the event of alignment or scheduling changes, and "predicts ridership
based on socio-economic attributes, the physical characteristics of a bus route, and the attractiveness of down-route trip destinations" (Azar and Ferreira, 1995).

7.2 The Period Route Segment Model

The model estimates the AM peak and midday boarding estimates in each direction for each of the TAZs. The general form of the model is given below:

$$BOARDi^d = PRODi * OPPi^d * LOSi^d$$

- $BOARDi^d$ is the forecasted boarding count for zone $i$ in direction $d$.
- $PRODi$ is the trip production factor in the area surrounding zone $i$. This factor measures the ability of the area to generate trips. $PRODi$ is directly proportional to the number of adults, total population, total employment and the route distance in the zone (walking distance of a quarter mile), while inversely proportional to income level.
- $OPPi^d$ is the trip opportunity factor in direction $d$ from zone $i$. $OPPi^d$ measures the ability of the area around the zone to attract transit trips in the down-stream direction. The assumption made here is that a passenger will select a transit mode if 1) the mode is within a quarter mile distance, 2) the destination is at least a 6 minute ride from the departure point and less than 35 minutes, and 3) transferring passengers would remain on board a second mode for 3-15 minutes. Therefore, $OPPi^d$ is directly proportional to population and employment who fit into the above descriptions.
- $LOSi^d$, the level-of-service factor, is based on expected wait time for service and seat availability.

7.3 Description of Example: Hakodate, Japan

The old part of Hakodate city is served by an above-ground trolley system. In this scenario, the trolley system is studied in an effort to provide better service at the trolley stop facility level. Due to the recent population trends described in an earlier chapter,
certain older neighborhoods have experienced significant population losses, while transforming into more commercial areas. Other areas have gained population. This suggests that absolute usage levels as well as peak usage times for many trolley stops will change, resulting in shifts in passenger demand within the two trolley lines in operation.

A study of trolley stop utilization is important in order to gauge these changes and possibly upgrade stops that are experiencing a high level of usage.

7.4 Extracting of Data

Figure 7-1 is a reduced 1/25000 map that shows the locations of the two trolley lines. It is assumed here that a person within 500m of any trolley stop is likely to be capable of utilizing that stop. This distance is proximate to the quarter mile radius used by Azar and Ferreira for bus and subway access. Due to the relatively micro level of analysis that is thus required, the level IV mesh of 500m squares is applied. This level of data is available for densely inhabited districts (DIDs) in Japan, and thus is available for this particular location. As a rule, a mesh square is included if:

- Any portion of the mesh square is within a 500m radius of a trolley stop; AND
- More than half of the radius length connecting the trolley stop in question and the outer mesh square must lie within that outer mesh square, at least for one radius location.

This is shown in Figure 7-2.

7.5 Critique of Model and Census Formats

The defined constraints of the PRS model are evaluated, and the applications of U.S. and Japanese census formats are compared here.

1. The Assumed Width of the Zone. Does the rigorous definition of the zone satisfactorily take into account the riders? It is reasonable to assume that local transit usage levels will experience significant population losses, while transforming into more commercial areas. Other areas have gained population. This suggests that absolute usage levels as well as peak usage times for many trolley stops will change, resulting in shifts in passenger demand within the two trolley lines in operation.

A study of trolley stop utilization is important in order to gauge these changes and possibly upgrade stops that are experiencing a high level of usage.

7.4 Extracting of Data

Figure 7-1 is a reduced 1/25000 map that shows the locations of the two trolley lines. It is assumed here that a person within 500m of any trolley stop is likely to be capable of utilizing that stop. This distance is proximate to the quarter mile radius used by Azar and Ferreira for bus and subway access. Due to the relatively micro level of analysis that is thus required, the level IV mesh of 500m squares is applied. This level of data is available for densely inhabited districts (DIDs) in Japan, and thus is available for this particular location. As a rule, a mesh square is included if:

- Any portion of the mesh square is within a 500m radius of a trolley stop; AND
- More than half of the radius length connecting the trolley stop in question and the outer mesh square must lie within that outer mesh square, at least for one radius location.

This is shown in Figure 7-2.

7.5 Critique of Model and Census Formats

The defined constraints of the PRS model are evaluated, and the applications of U.S. and Japanese census formats are compared here.

1. The Assumed Width of the Zone. Does the rigorous definition of the zone satisfactorily take into account the riders? It is reasonable to assume that local transit usage levels will experience significant population losses, while transforming into more commercial areas. Other areas have gained population. This suggests that absolute usage levels as well as peak usage times for many trolley stops will change, resulting in shifts in passenger demand within the two trolley lines in operation.

A study of trolley stop utilization is important in order to gauge these changes and possibly upgrade stops that are experiencing a high level of usage.

7.4 Extracting of Data

Figure 7-1 is a reduced 1/25000 map that shows the locations of the two trolley lines. It is assumed here that a person within 500m of any trolley stop is likely to be capable of utilizing that stop. This distance is proximate to the quarter mile radius used by Azar and Ferreira for bus and subway access. Due to the relatively micro level of analysis that is thus required, the level IV mesh of 500m squares is applied. This level of data is available for densely inhabited districts (DIDs) in Japan, and thus is available for this particular location. As a rule, a mesh square is included if:

- Any portion of the mesh square is within a 500m radius of a trolley stop; AND
- More than half of the radius length connecting the trolley stop in question and the outer mesh square must lie within that outer mesh square, at least for one radius location.

This is shown in Figure 7-2.

7.5 Critique of Model and Census Formats

The defined constraints of the PRS model are evaluated, and the applications of U.S. and Japanese census formats are compared here.

1. The Assumed Width of the Zone. Does the rigorous definition of the zone satisfactorily take into account the riders? It is reasonable to assume that local transit usage levels will experience significant population losses, while transforming into more commercial areas. Other areas have gained population. This suggests that absolute usage levels as well as peak usage times for many trolley stops will change, resulting in shifts in passenger demand within the two trolley lines in operation.

A study of trolley stop utilization is important in order to gauge these changes and possibly upgrade stops that are experiencing a high level of usage.
Figure 7-1 Trolley Service; Hakodate, Japan

(Japan Geographical Survey Institute, 1989)
Figure 7-2  Selecting Mesh Squares for Forming Study Zone

Mesh Squares Included in Study Corridor
modes are accessed primarily by foot. Also, these modes are characterized by frequent stops that are located at close intervals. Bus service, furthermore, can be provided along many routes where transit demand exists. For these reasons, the 500m (or close to a quarter-mile) radius for local transit users is generally valid. Other transportation modes normally do not act as feeders for local transit modes.

2. Attributes' Failure to Illustrate Ridership. Figure 7-3 shows a mesh square which is dominated by a racetrack. This happens to be a major racetrack within the prefecture of Hokkaido, and can seat 20,000 spectators at a time. During the racing season, races occur frequently. However, the attribute data for this mesh square suggests a close-to-zero night time population, as well as a day time population which is far exceeded by the number of spectators during operating hours. A more extreme case of this flaw would be the case if the mesh square was dominated by a busy airport, which is utilized during all seasons and all hours of the day. While both Japanese and U.S. census areas share this problem, it is a more serious problem for Japanese mesh squares. Mesh squares, created without regard for actual land-use conditions, could easily contain only airport property; therefore its actual daytime population would be vastly understated by its reported population. Most U.S. census tracts fulfill the population range requirement, thus the degree of this representation problem is less.

3. Degrees of Areal Unit Resolution. For macro-area problems described in the previous chapter, higher resolution meant finer differentiation in terms of distance, as well as an increase in land-use homogeneity within the areal units. In micro-area analysis, the analysis area is rigorously defined, and smaller areal units are better able to re-create the analysis area. While the mesh squares provide the highest resolution at 0.25km$^2$, the mean area for block groups in the most urbanized areas of Boston, Massachusetts is about 0.12km$^2$.\(^1\) This mean varies according to the degree of

\(^1\)Calculated using Arcview 2 mile buffer on block groups in the Back Bay and South End neighborhoods of Boston, Massachusetts.
Figure 7-3 Instance of Attribute Data Not Forecasting Usage Adequately
urbanization and other factors, but this particular value indicates that U.S. block groups can provide higher resolution in dense urban analysis areas.

7.6 Discussion

In the trolley stop problem, the area of analysis is constrained by a prescribed distance from the trolley stop. This local area by itself contributes to the usage of the internal trolley stop. Therefore, this analysis is primarily concerned with what is happening in the immediate vicinity of the stop, and not in outlying areas. This characteristic differentiates the Micro-area analysis from the gravity model/Macro-area problem described in Chapter 6. Areal units far away from the analysis object are not weighed into the local TAZ, as the gravity model was designed to do.

U.S. census tracts and block groups are designed with attribute homogeneity and consistency in mind. Specifically, this means that the areal sizes of block groups vary in order to achieve an enumeration within the prescribed range; thus, "park only" and "racetrack only" areal units in U.S. census data are highly unlikely. As mentioned before, the block group sizes in urban areas provide a sufficiently high level of spatial resolution. In other areas that are not so densely developed, however, areal units whose areas are inflated in order to satisfy attribute requirements cause problems for Micro-area analysis. An inflated block group may be considerably larger than the analysis area that is prescribed. Furthermore, an inflated block group may contain, for example, an urban park and an adjacent, densely populated neighborhood (Figure 7-4). For this trolley stop in the park, the projected usage would be vastly overestimated by riders from the adjacent neighborhood. Common sense dictates that a local transit stop does not belong in an uninhabited area such as this, yet the analysis area created by this inflated block group contradicts that knowledge.

The local transit stop example illustrates that a Micro-area analysis benefits from the availability of high-resolution and areally uniform units. In the most urbanized areas, it
Figure 7-4 Inflated Block Group Overestimates Ridership

500m Buffer

Densely Populated Neighborhood

Block Group Boundary
can be generalized that U.S. block groups tend to be uniformly small, and provide higher resolution than the 500m x 500m Japanese mesh squares. It is also known that pedestrians perceive formidable physical barriers such as major roads and rivers as boundaries when selecting from multiple transit options. U.S. block groups, often separated by such boundaries, may facilitate the creation of analysis zones. In less dense areas, however, some block groups begin to increase in size. The inclusion of block groups that are too large for the buffer size often results in overestimation of the analysis area. Furthermore, these non-metropolitan block groups demonstrate a lack of uniformity in size, as pockets of denser areas are represented by block groups which may be considerably smaller than the surrounding block groups. This causes the overestimation of the local transit ridership to be random and unpredictable.

The inherent complexity shared by both data formats is that it is virtually impossible to re-create exactly a rigorously defined area such as a buffer; boundaries of areal units do not coincide with those of most buffers, and area-splitting and population reassignment methods discussed in Chapter 5 are largely ineffective when the nature of population distributions within areal units are unknown. The task of accurate re-creation may be simplified if individuals or households were spatially encoded, but this is usually not possible for large population groups. Here again, the use of land-use information can facilitate accurate re-allocation of populations. Particularly within areal units similar to the one in Figure 7-4, the populations are not homogeneously distributed. A land-use layer in a GIS, for example, would show that the population is concentrated in the neighborhood that occupies a portion of the areal unit. The population associated with that areal unit can then be re-allocated to the neighborhood portion. This now enables the buffering of the transit stop to yield realistic results.
Chapter 8: Conclusion

This thesis evaluated the applicability of existing census data formats to typical urban studies and planning related analyses. Comparisons of two given census spatial representations within the context of proposed urban problems and the broader problem categories were made; in the process, particular complexities associated with either or both of the formats were pointed out. The findings for each of the broad categories would facilitate census data application toward other problems that fall into these categories. The Japanese and U.S. census formats proved to provide stark contrast, as they respectively represent extremely different approaches to a two-dimensional representation of physical realities.

In order to illustrate some of the assertions in the thesis, actual Japanese census data was extracted and processed, while U.S. data examples were mainly drawn from other works. The decision for this time-saving approach was made in light of the fact that 1) an English-language audience is less likely to be familiar with Japanese census data issues than those of the U.S. census; and 2) a sufficient amount of literature containing U.S. data examples was found. Overall, the thesis had a secondary purpose of providing information about Japanese census data for unfamiliar users and potential users. The scope of the thesis was limited to the examination of the spatial aspects; issues such as questionnaires and the creation of attribute data fields were not discussed in detail.

8.1 Urban Problems Revisited

The previous discussions of each of the four categories of urban problems explored the desired characteristics of the census data to be applied, as well as the capabilities of the U.S. and Japanese census data to fulfill these needs. These findings are summarized below:
1. **Time-constant/Enumeration:** This category requires the ability to portray some characteristic of a client specified area at a particular instant in time. The ability of census formats to accurately and easily re-create a client specified area using census areal units is a key issue. For enumeration purposes, the population attribute must be accurately represented by each areal unit.

As the U.S. census fulfills a constitutional mission of enumerating the national population, it is significantly better suited for the task. U.S. census tracts often have visible boundaries such as rivers and roads; such areal units are better capable of replicating commonly specified areas for enumeration, such as towns and cities. Block groups are suitable for even smaller areas like neighborhoods. Enumeration with U.S. census areas are highly accurate, because blocks are enumerated and tabulated separately, with higher geographic entities being created by aggregating these blocks. It is difficult to replicate meaningful areas by aggregating the Japanese mesh squares which have rigid borders. Attribute accuracy is also questionable with mesh squares, due to the nature of the procedures involved in the aggregation of basic census areas to create the mesh squares.

2. **Time-series/Population change over time:** A temporal analysis of some specified area also requires accurate enumeration and area re-creation using census areal units. In addition, however, the re-creation must be possible at more than one instance, i.e., for more than one census year.

The comparison involving the re-creation of meaningful areas is also valid for this problem/category. However, a price is to be paid for this trait of U.S. census data. As the physical environment and development densities change over the years, the U.S. census areal units must accommodate the changes, and alter their boundaries. This causes temporal comparisons of areas affected by boundary changes to be problematic. Mesh squares maintain unchanging areal boundaries, thus enabling comparisons of mesh squares and their aggregations over time.
3. **Macro-area/Transit ridership forecast:** Smaller areal units tend to demonstrate more attractiveness-homogeneity than larger areal units, because sub-areas within large areal units may exhibit rather different attractiveness factors. Higher resolution also permits an increased differentiation of areal units based on distances that separate them; this is advantageous only when circumstances require high spatial resolution. Analysis of suburban areas using level III (1km$^2$) mesh generally does not provide the high resolution that suburban U.S. block groups can. The U.S. census tracts’ and block groups’ tendencies to fulfill population range requirements may result in areal inconsistencies among these areal units, with some areal units being much larger than other proximate areal units. This makes the aforementioned problem with large areal units difficult to identify and control.

4. **Micro-area/Local transit stop usage:** Higher resolution and areal uniformity among the census areal units would facilitate analysis. U.S. block groups and census tracts are created with population ranges as goals, so their areas will vary. In densely developed urban areas, block groups tend to be uniformly small in area, providing higher resolution over the 500m x 500m mesh squares. In non-metropolitan areas, however, these block groups become larger and vary significantly in area. If the buffer catches an oversized block group, then the resulting analysis area will be much larger than a realistic sphere for a local trolley stop. This results in an overestimation of the actual analysis area. Furthermore, the lack of size uniformity among non-metropolitan block groups make the overestimation difficult to be accounted for, as they may appear randomly. The congruent 500m x 500m mesh squares are immune from this problem.

8.2 Evaluation of Data Formats

The preparation of the Japanese census is much influenced by its frequency: the population and housing census is conducted every five years. Timely preparation of data products as well as cost efficiency, translating into the facility of data collection and
representation, are very important issues for the Japanese Bureau of Statistics. The other major influence of the Japanese format was the set of agendas back in the late 1960s that prompted the need for a new dataset: to enable accurate measurement of economic and demographic change during a period of miraculously fast growth, and to provide a system that is immune to changes in political boundaries due to reorganization, and facilitates numerical and spatial analysis.

It is argued by many that the current Japanese mesh formats fulfills these requirements quite well. Japan's economy, politics and society of the 1960s was amidst an intensive development phase, involving centralized economic planning based on national, rather than regional or local, goals; the pragmatic mesh squares may have served well as an indicator of change and development at the national, or even at the broad regional level. But then, how well do mesh data fulfill the needs of today's urban researchers? Societal interest and scholarship in urban and regional studies has taken a decidedly local turn since, focusing on trends, policies and goals of smaller and defined entities, like municipalities and even neighborhoods. This is where the use of mesh data becomes troublesome, as Chapters 4 and 5 in particular have shown. For local, defined areas, enumeration errors using mesh squares often become too significant, as the squares are unable to conform to the local boundaries. Increasingly, urban and regional research in the present context is becoming meaningful only if the results make statements regarding "meaningful" areas, like municipalities and neighborhoods, not aggregations of square areas.

The U.S. census data format is not completely immune to the problem of creating meaningful areas with accuracy; its geographic entities are not municipalities or neighborhoods, either. However, urban researchers often benefit from using this data format, as there is an attempt to create areal units that reflect the locations of communities and pockets of populations. Also, the U.S. census employs the approach of often using physical boundaries such as creeks and roads for areal unit boundaries. Since
neighborhoods and municipalities are often defined by such features, U.S. census areal units are more likely to be better suited for aggregation to form areas that are meaningful for urban researchers. Also, U.S. areal units prove to provide enumeration with extremely high accuracy.

In contrast, Chapters 6 and 7 discussed models that did not rely on existing boundaries to define analysis areas, but defined the areas within the models themselves. The gravity model (Macro-area analysis) calculated a statistical surface based on linear distance between the areal units, technically inviting all areal units to be included in the zone, but weighing heavily the proximate and influential areas. The circular buffer model (Micro-area analysis) defined the criteria for inclusion and exclusion of areal units to the analysis zone. In the most dense and some suburban areas, average block groups were smaller than the level IV mesh squares, and demonstrated areal uniformity to a certain degree. However, block group areas would become larger and demonstrate a lack of size and shape uniformity in non-metropolitan areas, as they remain constrained by defined population ranges within levels of aggregation. In short, these models and numerical analyses bring out some of the advantages associated with raster type areal representations such as mesh data, especially in non-metropolitan areas.

In discussions of the four categories of urban problems, the utilization of land-use information from various sources in order to facilitate the analyses was the common theme. The aim was to re-process the existing mesh data in order to produce areal units that 1) facilitate the re-creation of meaningful analysis areas or rigorously defined areas; 2) provide higher resolution; or 3) demonstrate increased land-use or other homogeneity. The discussion of these schemes dealing with the complexities presented by the census data formats suggests that future research and development could facilitate the use of these census data formats in urban related research. Here are some possible directions that might be pursued:
1) In a GIS which incorporates census data and land-use information, a mechanism which recognizes signs of population patterns from the land-use data and proportionally assigns population data from the census to specific areas where population is evident. In the reality of heterogeneous development, such a mechanism permits a more accurate representation of population.

2) A method of illustrating the degrees of homogeneity in population distribution within individual census areal units. This could be managed by one or more fields in the attributes table of a data set, where a statistical index could show the nature of distribution within the areal units. A user performing an area-splitting and population re-allocation procedure, for example, would find this information useful.

8.3 Japan’s Future Spatial Data Scheme

The analyses of the four presented urban problems suggest that, while the Japanese mesh census data fulfills its originally intended goals and, facilitates certain types of numeric analyses and modeling, it does not readily give useful information regarding specific local areas for which the boundaries are known. This is actually not considered a radical view in Japan. Block-level spatial data is now becoming available, but with only one attribute (night population), and its accuracy is known to be questionable because the figures are based on an address listing.

It would seem that the Japanese registry, if spatially coded frequently, could provide valuable spatial data that facilitates enumeration and the re-creation of meaningful areas with high accuracy. In addition to such tasks, it can be generalized that the registry provides a "home base" of sorts as well as a form of traceable identification to the Japanese citizens, as it contains information which can be considered relatively permanent; indeed, for many, name, sex, hometown, and even their address do not change over their lifetimes.
An important issue for urban researchers is the availability of all such data. As mentioned before, the census data is very expensive and difficult to obtain (although this is beginning to change); however, the registry information is \textit{completely} off-limits. To the outside observer, the reason for this may seem unclear, but the reason has more to do with historical remnants from the pre-World War II period when Japan's subjects were sorted into rigid and hereditary social classes. In those days, the registry included class information. Although this system was abolished, many Japanese today are conscious of their former class and, in many cases, their addresses may suggest affiliation with certain social classes. Therefore, municipalities are extremely protective about releasing any data derived from their registries, and this is not likely to change so soon.

In this climate, it is likely that the registry will continue to be used solely by municipalities for administrative purposes. Census data, projected to become more accessible in the near future, will probably be the most reliable source of spatial and tabular population information for urban researchers. And, as the long list of census data fields in Chapter 3 indicates, census data will also prevail as the richest source of population information. It is hoped that recent advances in GIS technology availability and use in Japan will serve as impetus for introducing innovative spatial data products that address the issues that were discussed in this paper, in the near future.
Bibliography


