AN EVALUATION OF CLUSTERING AND DISTRICTING MODELS FOR HOUSEHOLD SOCIO-ECONOMIC INDICATORS IN ADDRESS-BASED POPULATION REGISTER SYSTEM

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ŞEYMA ÖZCAN YAVUZOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
GEODETIC AND GEOGRAPHICAL INFORMATION TECHNOLOGIES

DECEMBER 2009

Approval of the thesis:

AN EVALUATION OF CLUSTERING AND DISTRICTING MODELS FOR HOUSEHOLD SOCIO-ECONOMIC INDICATORS IN ADDRESS-BASED POPULATION REGISTER SYSTEM

submitted by **ŞEYMA ÖZCAN YAVUZOĞLU** in partial fulfillment of the requirements for the degree of Master **of Science in Geodetic and Geographical Information Technologies, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of Natural and Applied Sciences

Assoc. Prof. Dr. Mahmut Onur Karslıoğlu _____
Head of Department, Geodetic and Geographical Information Technologies, METU

Assoc. Prof. Dr. H. Şebnem Düzgün _____
Supervisor, Mining Engineering Dept., METU

Examining Committee Members:

Prof. Dr. Vedat Toprak _____
Geological Engineering Dept., METU

Assoc. Prof. Dr. H. Şebnem Düzgün _____
Supervisor, Mining Engineering Dept., METU

Assoc. Prof. Dr. Oğuz Işık _____
City and Regional Planning Dept., METU

Assoc. Prof. Dr. Ayşegül Aksoy _____
Environmental Engineering Dept., METU

Dr. B. Burçak Başbuğ Erkan _____
Department of Statistics, METU

Date: 11 December 2009

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Şeyma ÖZCAN YAVUZOĞLU

Signature :

ABSTRACT

AN EVALUATION OF CLUSTERING AND DISTRICTING MODELS
FOR HOUSEHOLD SOCIO-ECONOMIC INDICATORS
IN ADDRESS-BASED POPULATION REGISTER SYSTEM

Yavuzoğlu, Şeyma Özcan
M.Sc., Geodetic and Geographical Information Technologies
Supervisor: Assoc. Prof. Dr. H. Şebnem Düzgün

December 2009, 297 pages

Census operations are very important events in the history of a nation. These operations cover every bit of land and property of the country and its citizens. Census data is also known as demographic data providing valuable information to various users, particularly planners to know the trends in the key areas. Since 2006, Turkey aims to produce this census data not as “de-facto” (static) but as “de-jure” (real-time) by the new Address Based Register Information System (ABPRS). Besides, by this new register based census, personal information is matched with their address information and censuses gained a spatial dimension. Data obtained from this kind of a system can be a great input for the creation of “small statistical areas (SSAs)” which can compose of street blocks or any other small geographical unit to which social data can be referenced and to establish a complete census geography for Turkey. Because, statistics on large administrative units are only necessary for policy design only at an extremely abstracted level of analysis which is far from "real" problems as experienced by individuals.

In this thesis, it is aimed to employ some spatial clustering and districting methodologies to automatically produce SSAs which are basically built upon the

ABPRS data that is geo-referenced with the aid of geographical information systems (GIS) and thus help improving the census geography concept which is limited with only higher level administrative boundaries in Turkey. In order to have a clear idea of what strategy to choose for its realization, small area identification criteria and methodologies are searched by looking into the United Nations' recommendations and by taking some national and international applications into consideration. In addition, spatial clustering methods are examined for obtaining SSAs which fulfills these criteria in an automated fashion. Simulated annealing on k-means clustering, only k-means clustering and simulated annealing on k-means clustering of Self-Organizing Map (SOM) unified distances are deemed as suitable methods. Then these methods are implemented on parcel and block datasets having either raw data or socio-economic status (SES) indices in nine neighborhoods of Keçiören whose graphical and non-graphical raw data are manipulated, geo-referenced and combined in common basemaps. Consequently, simulated annealing refinement on k-means clustering of SOM u-distances is selected as the optimum method for constructing SSAs for all datasets after making a comparative quality assessment study which allows us to see how much each method obeyed the basic criteria of small area identification while creating SSA layers.

Keywords: Census Geography, Small Statistical Areas (SSAs), Geographical Information Systems (GIS), Spatial Clustering, Districting, K-means, Self-Organizing Map (SOM), Simulated Annealing.

ÖZ

ADRESE DAYALI NÜFUS KAYIT SİSTEMİ
HANEHALKI SOSYO-EKONOMİK İNDİKATÖRLERİNİN KÜMELEME VE BÖLGE TASARIMI MODELLERİ İLE DEĞERLENDİRİLMESİ

Yavuzoğlu, Şeyma Özcan
Y. Lisans, Jeodezi ve Coğrafi Bilgi Teknolojileri
Tez Yöneticisi: Doç. Dr. H. Şebnem Düzgün

Aralık 2009, 297 sayfa

Nüfus sayımları bir ulusun tarihindeki en önemli olaylardandır. Bu sayımlar bir ülkeye ve vatandaşlarına ait bütün mülkü ve araziyi kapsamaktadır. Sayım verisi, özellikle anahtar bölgelerdeki eğilimleri öğrenmeyi amaçlayan planlamacılar olmak üzere, daha birçok kullanıcıya demografik veri olarak da bilinen çok değerli bir bilgiyi sağlamaktadır. Türkiye, 2006 yılından itibaren yeni Adrese Dayalı Nüfus Kayıt Sistemi (ADNKS) çalışması sayesinde sayım verisini “de-facto” (statik) değil “de-jure” (gerçek zamanlı) olarak üretmeyi amaçlamıştır. Ayrıca yeni kayıt bazlı sayım yöntemi ile kişi bilgileri ikamet ettikleri adres verisiyle eşleştirilmiş ve sayımlar mekansal bir boyut kazanmıştır. Böyle bir sistemden elde edilen bilgi, soysal verinin coğrafi olarak eşlendiği adalardan ya da herhangi başka küçük coğrafi ünitelerden oluşabilecek “küçük istatistiki alanların” oluşturulmasında ve tam bir sayım coğrafyası oluşturulmasında önemli bir girdi oluşturacaktır. Çünkü büyük idari alanlara ait istatistiksel veri, bireylerin tecrübe ettiği “gerçek” sorunlara değinmeyen, oldukça soyut analizler sonucunda geliştirilebilecek politikalar için gereklidir.

Bu tezde, Coğrafi Bilgi Sistemleri (CBS) yardımıyla coğrafi olarak ilişkilendirilmiş ADNKS verisi üzerine kurulu, küçük istatistiki alanların otomatik olarak

üretilmesini sağlayan bazı kümeleme ve bölgeleme metodolojilerinin kullanılması ve böylelikle, Türkiye'de büyük kademe idari alanlar ile sınırlı kalmış sayım coğrafyası kavramının iyileştirilmesine yardımcı olunması amaçlanmıştır. Bu amacın gerçekleştirilmesi yolunda hangi stratejinin seçilmesi gerektiğine ilişkin belirli bir fikir sahibi olmak amacıyla, Birleşmiş Milletler – İstatistik Bölümü tavsiyelerine bakılarak ve bazı ulusal ve uluslar arası uygulamalar dikkate alınarak, küçük alanların belirlenmesi ile ilgili kurallar ve yöntemler araştırılmıştır. Ek olarak, kurallara uyumlu küçük alanların otomatik olarak üretilmesini sağlayabilecek mekansal kümeleme yöntemleri araştırılmıştır. K-ortalamaları üzerine benzetilmiş tavlama, sadece k-ortalamaları ve Öz Düzenleyici Haritaların (ÖDH) sağladığı birleşik uzaklık (benzerlik) değerleri üzerine uygulanan k-ortalamalarının benzetilmiş tavlama ile iyileştirilmesi yöntemleri uygun bulunmuştur. Daha sonra bu yöntemler, Keçiören ilçesinde bulunan 9 mahallenin parsel ve adalarına ait grafik ve öznitelik verilerin işlenmesi, coğrafi olarak ilişkilendirilmesi ve ortak altlıklarda birleştirilmesi sayesinde elde edilen haritalar üzerinde uygulanmıştır. Sonuç olarak, ÖDH kümeleme çıktısı üzerinde uygulanan k-ortalamalarının bileşik tavlama yöntemi ile iyileştirilmesi, küçük alan belirlenmesine ilişkin temel kurallara her yöntemde ne kadar uyulduğunu görmemizi sağlayan karşılaştırmalı bir kalite değerlendirme çalışması sonrasında küçük alanların oluşturulmasında kullanılabilecek en ideal yöntem olarak seçilmiştir.

Anahtar kelimeler: Sayım Coğrafyası, Küçük İstatistiki Alanlar (KİA), Coğrafi Bilgi Sistemleri (CBS), Mekansal Kümeleme, Bölgeleme, K-Ortalamaları, Öz Düzenleyici Haritalar (ÖDH), Benzetilmiş Tavlama.

To My Beloved Mom, Dad and Dear Sisters…

To My One and Only...

ACKNOWLEDGMENTS

Firstly, I want to thank my advisor Assoc. Prof. Dr Şebnem Düzgün for her guidance, advice, criticism, encouragement and insight throughout the study. She was always there to listen and to give advice even if sometimes I was shy to ask for. She taught me how to ask questions and express my ideas. She showed me different ways to approach a research problem and she has reminded me to stop running and first to think. She is a complete role model for me. Besides, I would like to thank Prof. Dr. Vedat Toprak, Assoc. Prof. Dr. Oğuz Işık, Assoc. Prof. Dr. Ayşegül Aksoy and Dr. B. Burçak Başbuğ Erkan in my thesis committee, who gave a very detailed critique of my draft at my first jury and help shaping my thesis for an aim.

My special thanks goes to Mr. Hasan Aztopal - head of GIS Team, Dr. Levent Akçay – head of Address Frame Group and Muharrem Gürleyen Gök from Population and Migration Statistics in TURKSTAT for their never-ending supervision and patience throughout this thesis. I am most obliged to their kind and helpful behavior at all times for providing me such a valuable non-graphical data and having trust in me. Additionally, I am grateful to the staff of Development and City Planning Department at Keçiören Municipality for giving me valuable graphical datasets about Keçiören.

I would like to express my sincere gratitude to my friends, İlksen URGANCI, Gökçe Türkmendağ, Funda Arıkan, Oya Yarkınoğlu Gücük, Ulaş Canatalı, Sibel Sarı and all GGIT assistants for their warm friendships, suggestions, encouragement and joy throughout my graduate program. Those long hours in GGIT lab and meals at Zeynel will not be forgotten. Also, I would like to thank all my friends in TURKSTAT, especially Güneş İnan for happiness, friendship, fun, intelligence and great support. They never gave up on me even if I behaved like an unsocial lunatic.

I am greatly indebted to my mother Fahriye Özcan, my father Ahmet Özcan, and my elder sisters Hümeyra, Süheyla and Süreyya Özcan who provided me every opportunity and gave endless love, sincerity, friendship, unconditional support,

patience, kindness and compassion, happiness and trust to me for all of my life.

I would also like to thank my cat Gürbüz for sleeping beside me and providing me with peaceful moments while I was working.

Final thanks go to the God, who has granted me such a magnificent, hilarious, understanding and patient husband, Ayhan Yavuzoğlu.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AGNES: AGlometarive NESting
AMMOWI: Ankara Metropolitan Municipality Office of Water and Infrastructure.
ANNs: Artificial Neural Networks
AYBIS: Infrastructure Information System of Ankara
AZM: Automated Zone Matching
AZP: Automated Zoning Procedure
AZTool: Automated Zone Tool
BARD: Better Automated ReDistricting
BG: Block Groups
BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies
BMU: Best Matching Unit
BNA: Block numbering areas
CAD: Computer Aided Design
CF tree: Clustering Feature Tree
CF: Clustering Feature
CI: Computational intelligence
CLARANS: Clustering Large Applications Based on Randomized Search
CLIQUE: Clustering In QUEst
CSR: Complete Spatial Randomness
CURE:Clustering Using Representatives
DBSCAN: A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density
DEM: Digital Elevation Model
DENCLUE: Clustering Based on Density Distribution Functions
DF: Dwelling Frame
DIANA: DIvisive ANAlysis
DM: Data Mining ()
EA: Enumeration Area

EA: Enumeration Area

EA: Evolutionary Algorithm

ECW: Enhanced Wavelet Compression

ED: Enumeration district

EM: Expectation-Maximization

EP: Evolutionary Programming

ESDA: Exploratory Spatial Data Analysis

ESs: Evolution Strategies

EU: European Union

Eurostat: European Union Statistical Office

FA: Factor analysis

FCM: Fuzzy c-means

GAM: Geographical Analysis Machine

GAs: Genetic Algorithms

GIS: Geographical Information Systems

GISCO: Geographic Information System of the European Commission

GPS: Global Positioning System

INSPIRE: Infrastructure for Spatial Information in the European Community

KD: Knowledge Discovery

LISA: Local Indicators of Spatial Autocorrelation

MAUP: Modifiable areal unit problem

MAUP: Modifiable Areal Unit Problem

MinPts: Minimum number of objects.

MMA: Metropolitan Municipality of Ankara-

NEM: Neighborhood EM

NSO: National Statistical Office

NUTS: Nomenclature Territorial Units for Statistics

OA: Output area

ONS: Office for National Statistics

OPTICS: Ordering Points to identify the Clustering Structure

PAM: Partitioning Around Medoids

PCA: Principal Component Analysis

PCA: Principal components analysis

SA: Simulated annealing

SAL: Small Area Layer

SES: Socio-Economic Status

SEU: Social Exclusion Unit

SOFM: Self-organizing feature maps

SOM: Self-organizing maps

SSA: Small Statistical Areas

SSE: Sum of the squared error

StatsSA: Statistics South Africa

STING: STatistical INformation Grid approach

TIGER: Topologically Integrated Geographic Encoding and Referencing

TURKSTAT: Turkish Statistical Institute

TURKSTAT: Turkish Statistical Institute

UK: United Kingdom

UN: United Nations

US: United States

ZDES: Zone Design System

CHAPTER 1

INTRODUCTION

According to United Nations (2008), “the most important capital a society can have is the human capital. Assessing the quantity and quality of this capital at small area, regional and national levels is an essential component of modern government”. So, besides searching answer to question “How many are we?”, there is also a need to provide an answer to “Who are we?” in terms of age, sex, education, occupation, income levels and other crucial characteristics, as well as to “Where do we live?” in terms of the location that a person occupies. Answers to these questions can be given by population census that involves a spatial dimension.

Turkey with around 71.5 million population is the 17th populated country in the world and relatively large country with total surface area of 769 604 km^2. Proportion of population living in provinces and district centers (957 units) is 75% and the rest of the population living in towns and villages (36127 units). Total fertility rate is 2.14 children and life expectancy at birth is around 73.6 years. Population size is gradually increasing with the growth rate of 1.3 % per year (TURKSTAT, 2008).

In Turkey, the first population census was carried out in 1927 and the next population censuses were carried out between 1935 and 1990 regularly, in years ending with 0 and 5 in Turkey. After 1990, population censuses have been decided to be carried out in years ending with 0. However, according to the Official Statistical Program of Turkey, next census is going to be carried out in 2011, in order to keep up with the population census calendar of European countries which will also carry out population census in year 2011.

Until year 2006, all population censuses were carried out with traditional method in one day by application of a curfew with “de facto” definition. Information on usual

residence (de jure) population and household structure are not available from past censuses because of listing only present persons on the census day. Also, the mobility of population from one location to another couldn't be tracked by traditional method. Therefore, Turkish Statistical Institute (TURKSTAT) aimed to change the method of population census method into register based census in order to produce more reliable and up to date information on population size and characteristics by the Address Based Population Register System (ABPRS). For this purpose, population registration system has been improved to cover usual residence address of all people living in the country. As a result, the population census in Turkey now gained a spatial dimension which was not available before 2006.

However, in order to evaluate and include this new high resolution spatial dimension of censuses into policy-making effectively, the act of describing the state of Turkey has to transform into a new perspective from just aggregating the statistics on large administrative areas. Guiblin et al (2001-b) denotes that the "…recent increasing importance of information and knowledge systems in the new process oriented methods adopted in the production, distribution and consumption of goods and services in the societies caused the demand for a new geographical base for statistics". However, a complete census geography to meet these information and knowledge demands from society does not exist in Turkey except high level large administrative levels (Nomenclature Units for Territorial Statistics -NUTS level 1-2-3) which are also used for statistical data dissemination in Turkey.

In addition, administrative division is not enough for a complete census geography establishment for many reasons. First of all, administrative boundaries are always subject to change because of political or population trends. As a matter of fact, administrative divisions of Turkey have been changed so many times; and these alterations cause serious problems for statistical data users by making time-series comparisons almost impossible. Researchers may not have accurate results from their studies; formulation of regional policies and development plans may cause unsatisfactory results; and goods and services may not be conveyed to the needy areas. Therefore, a flexible census geography is needed to improve the current

statistical system of Turkey which is independent from administrative hierarchy to an extent.

Secondly, statistical units must have a standardized structure in terms of population. However, it is obvious that almost none of the administrative units are comparable to each other in terms of population numbers in Turkey. For example, the population of a neighborhood may be many times larger than another in the same district. This condition prevents making good judgments among administrative units, which are also statistical areas, and makes them incomparable to each other. Therefore, Turkish statistical system needs standardized and comparable statistical units.

The third considerable disadvantage of current structure is that without a high resolution data the policy judgments on a particular area may have misleading effects. Yetik (2003) mentions that “the majority of statistics, being developed, are on the basis of the cities, which are the biggest geographical unit of administrative classifications and very few of them are based on districts.” The provinces and districts have relatively big sizes in respect to population; and when statistics are produced for these divisions, the specific characteristics of local inhabitants are ignored. As a result, data users see only the average. Therefore, Turkish statistical system needs homogeneous and smaller statistical units to the possible extent to reduce such kinds of errors.

The number of drawbacks may be increased, but these three reasons are adequate to show the disadvantages of a lacking geographic base of the current statistical system. It is clear that, administrative areas have been designed for administrative purposes only, not for statistical purposes. Thus, starting out from the drawbacks of the current system, the main aim of this thesis is to create a new small statistical area layer that would constitute a basis for a complete census geography in Turkey. But it must be clarified that this new system would not completely substitute the current statistical constitution and administrative division; rather it must be thought as a complementary to the present system.

There are many steps throughout the thesis to reach to the main aim. First step includes a literature review mainly about the concept of “census geography”. This literature review also showed that there are few studies concerning with the census geography needs of Turkey (Kırlangıçoğlu, 2005). There are some studies (Aydinoğlu & Yomralıoğlu, 2009; Mataracı et al, 2009) that take the first step about those subjects and demonstrate some kind of GIS usage in such kind of studies, but none of them propose concrete answers to the question of how census geography should be in Turkey. The literature survey also helped to understand the present census-based statistical system, administrative system and already in use census geography of Turkey. Of course, it would be very hard to interpret the current system without knowing the others in the world. Therefore international applications and census geographies were examined. Then, the countries with the most developed census geography, the USA, UK and South Africa and TANDEM study as an international application were taken as the main models because of their reliable, respected, settled and well-working statistical systems in Chapter 2. So, the rules and methodologies for a small statistical area layer are proposed for Turkey in order to build a census geography upon.

The main point of those rules is to establish and identify small statistical areas which requires basically four fundamental requirements for the census geography units; homogeneity, compactness, population equity and contiguity. These concepts are also explained in detail in Chapter 2. Besides census geography and small area identification concepts, statistical activities in Turkey are explained in Chapter 2 on behalf of geography. These studies include upgrading the regional statistical system of Turkey for the compliance with the European Union and the new register based census methodology adopted by TURKSTAT mentioned earlier as ABPRS.

In Chapter 3, a second base literature survey is done to inspect spatial clustering and districting algorithms that would allow us to automatically define the small statistical areas that would conform to small area identification rules which was elaborately explained in Chapter 2. As a result of this examination, Simulated annealing on K-means clustering, only K-means clustering and Simulated annealing on K-means clustering of Self-Organizing Maps’ unified distances are selected as

methods for several reasons and logic behind these algorithms is extensively explained. In addition, the software packages that are used constructing small areas are briefly introduced at the end of Chapter 3.

In Chapter 4, some neighborhoods are selected from Keçiören district as a case study area for several reasons. Study area's census data related to ABPRS is obtained from Turkish Statistical Institute (TURKSTAT) and examined in an attempt to decide how to use the data. Then, most of the digital graphical data related to study area is obtained from Keçiören municipality. Some methodologies are developed in order to geocode the non-geographical census data with its geographical cover. Of course, all the gathered data, some of which were in different formats, different characteristics etc. needed some manipulations like data conversion, editing, integration and also data generation. As a result of all these processes, necessary basemaps composing of parcels and blocks are obtained on which small statistical areas can be established.

After constructing fundamental input, in Chapter 5, principal component analysis is implemented in order to get involved with the data and to get a composite socio-economic index from separate indicators contained in census data. This index and raw data are used separately as deterministic inputs for all the analyses carried out for parcels and blocks. Then, selected clustering and districting algorithms are applied in three different methods in order to derive small statistical area layers both for parcel and block resolutions. At the end of Chapter 5, the small statistical area output layers are compared visually and with quality assessment measures such as compactness, equal population, homogeneity and contiguity and the optimum method is selected.

Finally, the last chapter is the "conclusion" in which, problems, solutions and difficulties of the analysis in the previous chapter are summarized and interpreted. In addition, recommendations for further studies are introduced as a result deficiencies experienced in this study.

CHAPTER 2

LITERATURE REVIEW ON CENSUS GEOGRAPHIES AND SMALL STATISTICAL AREAS

This chapter includes concepts, rules and world-wide applications related to census geographies and foundation of small statistical areas (SSAs) which serve the basis for census geographies. Firstly, some concepts related to population censuses are explained, because population censuses and census geographies are two terms that have close interference. Then, the role of Geographical Information Systems (GIS) is introduced as a tool for the development of census geographies, those which are used as a template for the collection, analysis and dissemination of census information. The chapter continues with the explanation of the necessity and the basic rules for establishing SSAs and gives examples from the national and international applications in order to develop an appropriate methodology for the study. Then, the current situation of statistics and census activities in Turkey is discussed on behalf of the necessity of a complete census geography and hence, SSAs. Finally, SSA system is proposed for a new census geography hierarchy and dissemination of census data in Turkey.

2.1. Population Censuses

According to United Nations (2008), a population census is the total process of collecting, compiling, evaluating, analyzing and publishing or otherwise disseminating demographic, economic and social data pertaining, at a specified time, to all persons in a country or in a well delimited part of a country.

In order to plan for, and implement, economic and social development, administrative activity or scientific research, it is necessary to have reliable and detailed data on the size, distribution and composition of population. The population

census is a primary source of statistics, covering not only the settled population but also homeless persons and nomadic groups. Data from population censuses should allow presentation and analysis in terms of statistics on persons and households and for a wide variety of geographical units, ranging from the country as a whole to individual small localities or city blocks. In order to implement a country-wide population census there are four essential features to obey (UN, 2008);

1) Individual enumeration: Each individual and each set of living quarters is enumerated separately and that the characteristics thereof are separately recorded.
2) Universality within a defined territory: The census should cover a precisely defined territory (for example, the entire country or a well-delimited part of it).
3) Simultaneity: Each person and each set of living quarters should be enumerated as of the same well-defined point in time and the data collected should refer to a well-defined reference period.
4) Defined periodicity: Censuses should be taken at regular intervals so that comparable information is made available in a fixed sequence. It is conventional that a national census be taken at least every 10 years. But, some countries may find it necessary to carry out censuses more frequently because of the rapidity of major changes in their population and/or its housing circumstances.

As part of their preparation for the 2010 global round of population and housing censuses, some countries are developing, testing, and implementing alternative methods for collecting, processing and disseminating key statistics that used to be generated by the traditional approach to population censuses. Even so, the crucial principle of providing detailed statistics at the lowest geographical level remains of paramount importance. There are 2 approaches in basic for conducting a population census that are currently in use (UN, 2008);

- The traditional approach: This approach comprises a complex operation of actively collecting information from individuals and households on a range of topics at a specified time. Members of the public respond to a census questionnaire, or interviewers are deployed to collect information from

respondents. For interviewer-based censuses, enumerators assigned to different enumeration areas cover all households and persons in the enumeration area during a specified and usually short period of time in order to meet the requirements of universality and simultaneity. The traditional census is a matchless way in providing a snapshot of the entire population at a specified period and the availability of data for small geographic domains. Because of the complexity and expense of such censuses, they are usually implemented only once every 5 or 10 years.

- The register-based approach: The concept of producing census-like results based on registers emerged in the 2000 round of censuses. The philosophy underlying this concept is to take advantage of the existing administrative sources, namely, different kinds of registers, of which the following are of primary importance: households, dwelling addresses and individuals. In the next iteration these are linked at the individual level with information on business, tax, education, employment and other relevant registers. While it is theoretically possible to link the records on the basis of the name of the individuals, the existence of a unique identification number for each individual, household and dwelling is of crucial importance, as it allows much more effective and reliable linking of records from different registers. Register-based approach is also implemented in Turkey in 2006-2007 periods under name Address Based Population Register System (ABPRS) and will be explained in forthcoming sections detailed. The census records obtained from this study was an important source of information and constitutes a backbone of this thesis.

2.2. The role of Maps and GIS in Population Censuses

For censuses, there are certain major elements that must be taken into account. In general, census operations can be divided into six phases: (a) preparatory work, (b) enumeration, (c) data processing, (d) building of needed databases and dissemination of the results, (e) evaluation of the results, and (f) analysis of the results. In all these phases, mapping has a vital role. Traditionally, the role of maps in the census process has been to support enumeration and to present aggregate

census results in cartographic form. In general terms, mapping serves several purposes in the census process as depicted in Figure 2.1 and explained as follows (UN, 2009);

(*a*) *Maps ensure coverage and facilitate census operations (pre-enumeration)*. The census office needs to ensure that every household and person in the country is counted and that no households or individuals are counted twice. For this purpose, census geographers partition the national territory into small data-collection units. Maps showing enumeration areas thus provide an essential control device that guarantees coverage of the census.

(*b*) *Maps support data collection and can help monitor census activities (during enumeration)*. During the census, maps ensure that enumerators can easily identify their assigned geographic areas, in which they will enumerate households. Maps are also given to the census supervisors assigned to enumerators to support planning and control tasks. Maps can thus also play a role in monitoring the progress of census operations. This allows supervisors to strategically plan, make assignments, identify problem areas and implement remedial action quickly.

It is still the case that in many countries there are only a limited range of maps available and these often do not show sufficient detail to enable the boundaries of small areas to be clearly defined. This is particularly likely to apply in areas of unplanned settlement. It is thus common to supplement the maps with other material, such as (*a*) lists of households and/or (*b*) a textual description of the boundary including roads, railway lines, power lines, rivers and other physical features. This description may also include obvious landmarks on the boundary (school buildings, water points and others). In NPE (Nation-wide Population-Enumeration Studies), it is obligatory to define the boundaries of registry regions which are defined over maps. But in Turkey, because of the non-existence of up-to-date and trusted maps to base the enumeration, the registry region definitions is being done over lists which addresses take place.

However, it is not appropriate for field staff to rely entirely on a list of households, written or verbal descriptions and directions, or on local knowledge of the area boundaries. Reliance on verbal descriptions or local knowledge very often leads to confusion and error because people tend to have mental images (or mental maps) of places and these images may not coincide with the area as it really is reflected in the design of the enumeration area. For the same kind of reason, the supervisor's mental map of an enumeration area may differ markedly from that of an enumerator. To overcome such problems, it is important that the best possible quality maps be the basis for census enumeration operations and that the collection staff receive comprehensive training in the correct use of the maps and associated textual material if that is provided.

(c) *Maps make it easier to present, analyze and disseminate census results (post-enumeration)*. The cartographic presentation of census results provides a powerful means for visualizing the results of a census. This supports the identification of local patterns of important demographic and social indicators. Maps are thus an integral part of policy analysis in the public and private sectors.

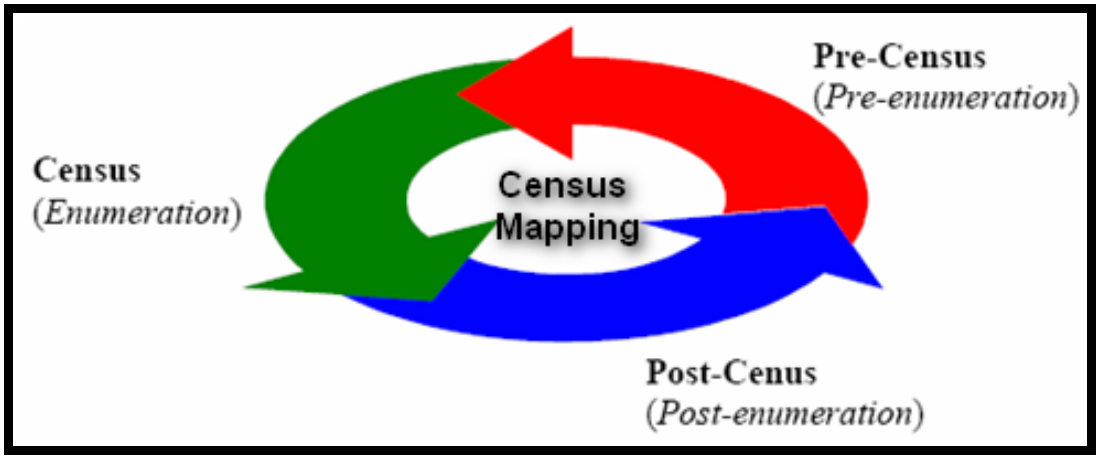


Figure 2.1 Mapping activities in a census cycle (UN, 2004).

Today, maps are one form of information display categorized under the broader term geographic information, and the form this geographic information most often takes is the geographically referenced database (or geodatabase). Before census mapping commences, the census agency needs to determine the appropriate technology for mapping. An agency should assess existing maps, available in any format, that are known to be accurate, and use them with new maps prepared as required. The new maps can either be produced as hand-drawn maps of enumeration areas, or they can incorporate overlays or other technological assistance. Alternatively, a GIS could be implemented. There are many definitions of GIS used by different people from different organizations and backgrounds. Generally;

> *"A GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data: 1. input; 2. data management (data storage and retrieval); 3. manipulation and analysis; and 4. output"* (Aronoff, 1993).

The most important feature of GIS is that it links tabular data to geographic locations. If this feature is handled in census case, for instance, GIS can link tabular data such as labor force, education degrees, average household size, population number etc. to geometric representations of statistical areas. GIS also, as different from a simple computer database program, can be used for representations, queries

and analyses based on a spatial perspective. For instance, the answers of questions such as: Which districts are next to District C? How many schools are inside of Province B? What types of businesses are located between the roads A and B? It is impossible to answer these spatial questions without using a spatial database program such as GIS.

It is clear that, "…the GIS technology is necessary to enable more sophisticated use of statistical data and the geographical information, and it offers a better system in terms of standards, methodology, collection, data processing and its dissemination." (UN, 2004)

In recent years, many countries have adopted the use of GIS to facilitate census mapping in the production of both enumeration maps and dissemination products. As the cost is declining and the basic technology is now well established, it is expected that this will continue. It is likely that the census could be a useful catalyst for increasing capacity within the statistical office (or the country as a whole). So, it is the GIS technology, which has the answer for the Turkish Statistical Institute (TURKSTAT) to improve the efficiency and utility of the department's role in all phases of a census to have more successful results. TURKSTAT uses GIS technology since 1992 to improve publications by thematic maps of demographic, population, regional, agricultural and environmental statistics. However, it is needed to use GIS technology in all phases of a census instead of using GIS for only dissemination purposes by thematic maps.

Where accurate and current maps at relevant scales are not available for a country or part of a country like in the case of Turkey, the technological alternatives depicted in Figure 2.2 and described in the following paragraphs could be employed in a GIS in data acquisition, data management and data dissemination phases for a census (UN, 2009);

(a) *Satellite images.* Although currently relatively expensive to acquire, the price of satellite imagery is declining in real terms. A satellite image typically covers a large area and can be cost-effective compared to other sources. Imagery should be pre-processed by the supplier so that it is rectified and

georeferenced (a known scale and orientation, with some latitudes and longitudes, is printed on the face of the image);

(*b*) *Aerial photography.* Acquisition of aerial photographs for large tracts of a country may be expensive. However, existing archives of photographs can be an excellent resource for preliminary counts of dwellings and as a base for basic maps. In some cases digital aerial photographs can be a cost-effective way of initiating some components of a geographic information system (GIS);

(*c*) *Global positioning systems (GPS).* Making hand-drawn maps or digital maps from a GIS for use by enumerators in the field can be greatly assisted by GPS. A simple, hand-held GPS receiver will give latitude and longitude coordinates with reasonable accuracy of key points. Depending upon the system selected, a GPS may also track linear features and thus be useful for mapping boundaries. Maps printed from a GIS or hand-drawn map can be enhanced by the addition of latitudes and longitudes recorded at key points to provide orientation, scale and absolute position. Such information will be particularly valuable for dissemination purposes or if the work is a component of developing a GIS for later use.

In this thesis, satellite images are used in order to get an up-to-date basemap for the study area.

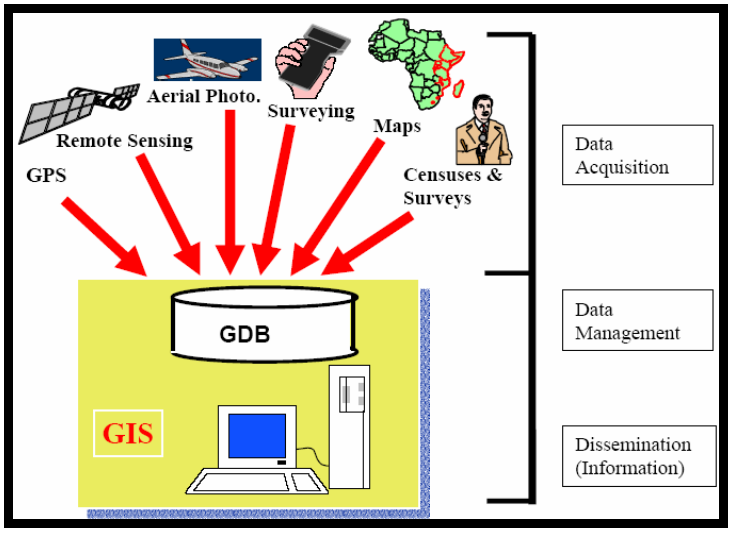


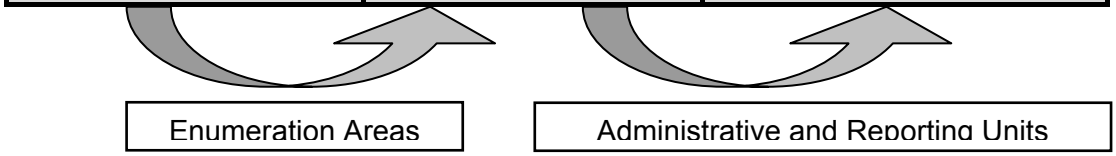
Figure 2.2 Technological alternatives and sources for GIS (UN, 2004).

Some statistical offices were early adopters of GIS. Population, social and economic statistics are the foundation of public planning and management. The spatial distribution of socioeconomic indicators guides policy decisions on regional development, service provision and many other areas. Digital techniques allow better management, faster retrieval and improved presentation of such data. There has therefore always been a close linkage between geography and statistics — as reflected, for instance, by the fact that in many countries the national statistical and mapping agencies are housed under the same roof. This close integration of GIS in statistical applications yields large benefits to national statistical offices as it reduces the cost and time required to collect, compile and distribute information. In TURKSTAT, GIS applications are mainly used by "GIS Team", but these applications covers only presenting the aggregate results of censuses or surveys as thematic maps for publications using large administrative units.

Census mapping process is very similar to a standard geographic information process. Therefore, it is easy to integrate GIS with census mapping operations in all three stages of the process which are pre-census, census and post-census (Table 2.1). In this way, census mapping activities go beyond a technical approach, become a united whole with technology and better reflect the population structure in local, regional and nation-wide scopes.

Table 2.1 GIS with census mapping: Stages of integration (UN, 2004)

PRE - CENSUS	CENSUS	POST-CENSUS
GIS GPS Photo / Video Digital Mapping Satellite Imagery	GPS GIS Digital Mapping	Databases Maintenance GIS Internet Statistical Database



Cartographic automation, GIS and other geospatial tools have enabled more efficient production of both enumerator maps and thematic maps of census results. In addition, advances in technology and new tasks for GIS using new data sources, such as remote sensing and GPS-enabled location recording, have expanded the power of geographic representation within a national statistical office (NSO).

Adoption of GIS should thus be seen as a major strategic decision with impacts beyond the census operation, and many issues need to be considered, since GIS has some costs as it has benefits (UN, 2008). These costs and benefits are listed below;

(a) Benefits;

(i) A closer linkage between maps for enumerators and map-based products for users;

(ii) The cost of intercensal updating of the base map will be less with a digital base map;

(iii) Producing duplicate maps may be less expensive with a GIS solution;

(vi) GIS will have increased ability to undertake quality assurance of geographic boundaries;

(v) The census agency will have a greater ability to perform spatial queries under GIS;

(vi) Space needed to store input maps for digital purposes will be far less;

(b) Costs;

(i) GIS requires additional technical expertise;

(ii) GIS will require a higher level of computing infrastructure;

(iii) A census system can proceed on the basis of basic maps. However, use of GIS in this task requires that a digital map base exists. If it is necessary to create the digital map base, significant lead times are required as well as significant funding. In both cases, more experienced technical staff is required;

(iv) In most cases, the preparation of maps and/or GIS will not be the core business of a statistical agency.

Prior to developing the mapping program for the census, consideration needs to be given to the geographic classification to be used and the mapping infrastructure available to carry out the mapping tasks. The census geographic work schema is illustrated in Figure 2.3. As the geography on which the census is collected will determine the geography on which the census data can be disseminated, a geographic classification should be devised in parallel with the development of census mapping which is mentioned with the red box on Figure 2.3.

In addition, the benefits of geographic data automation and classification in statistics are shared by the users of census and survey data. The data integration functions provided by geographic information systems, which allow the linking of information from many different subject areas, have led to a much wider use of statistical information. This in turn has increased the pressure on statistical offices to produce high-quality spatially referenced information for smaller geographic units. If carefully planned — that is, if the NSO can collect the information in small units and then aggregate it appropriately — then it should be able to satisfy the needs of many new data customers.

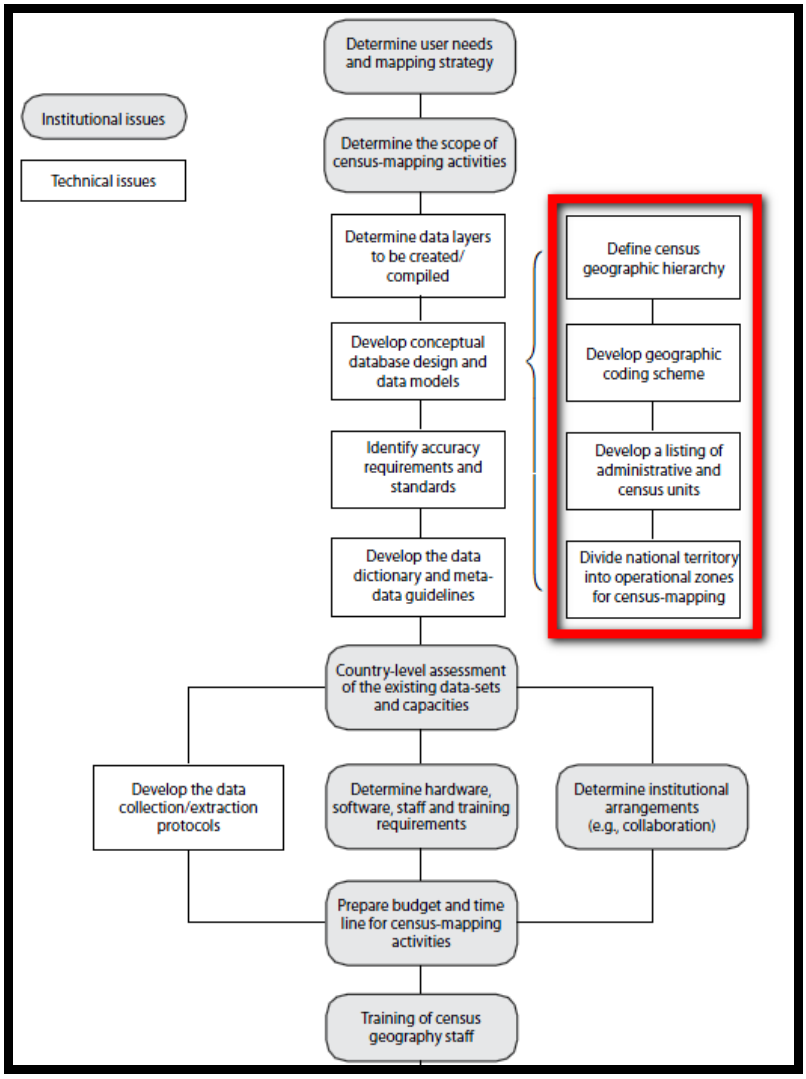


Figure 2.3 Stages in planning census geographic work (UN, 2009).

Moreover, the proposal behind use of new methods of spatial analysis is the availability of population data with a higher level of "granularity" (or spatial specificity) than previously, at the enumeration area (EA), population cluster or other small-area levels. If analysts or other GIS users wish to analyse the spatial distribution of population or map demographic or other variables in relation to others, they can now make use of a variety of techniques that range from simple queries to measurements to transformations, descriptive summaries and models.

The details of designing a general geographic classification, including the definition of the various areas of the geographic classification and their relationship to one

another, are more complex than those involved in census mapping. The design of (EAs), in other words small areas and other census management areas are of crucial importance for the census and are explained in following sections.

2.3. Census Geographic Classification and Constructing an EA-level Database for the Census

Reorganizing the NSO around a geographic information core means embracing the relationship between a country's geography and the various sets of information that the NSO uses and produces. The relationship between geography and databases occurs through the mechanism of coding. The first step is to link the management material (maps) to technical content (address lists).

The United Nations (2009) definition of "geocoding" is broader. It represents the connection between statistical observations and real-world locations expressed in terms of latitude and longitude or other locational attributes. Simply put, geographic coding is a way to ensure that the data know where they are.

Geocoding for a census is designed to cover a continuum of spatial scales, from individual housing units through enumeration area-level up to higher administrative or national levels. Its successful use depends on a country establishing a set of administrative areas with known territories and digital representation in the form of computer-coding.

Administrative hierarchies are based on the idea that inside the territory of the country there are boundaries that serve to demarcate the actual land extent at the state or provincial and district levels, or for the purposes of voting or health monitoring or postal delivery. Together, these various geographies can be stored in a database with the administrative-level code and number of units. For example, units at administrative level two (ADM 2s) are provinces, while units at administrative level three (ADM 3s) are districts. Ideally, any geospatial operation would have access to these units in GIS format for use in its various projects. The NSO role in administrative boundary delineations will vary by country.

One of the earliest decisions in census-planning pertains to the administrative areas for which census data will be reported. Administrative areas can be any special geographic unit, but mainly they are units of administration, i.e., some governmental authority has jurisdiction over the territory. Census preparation involves creating a list of all administrative and statistical reporting units in the country. The relationships among all types of administrative and reporting unit boundaries should be defined. Every country has its own specific administrative hierarchy, that is, a system by which the country and each lower level set of administrative units (except the lowest) are subdivided to form the next lower level. For example, for the purposes of the census a country may have been divided into seven hierarchical levels in urban areas and six in rural areas (Figure 2.4) (UN, 2009).

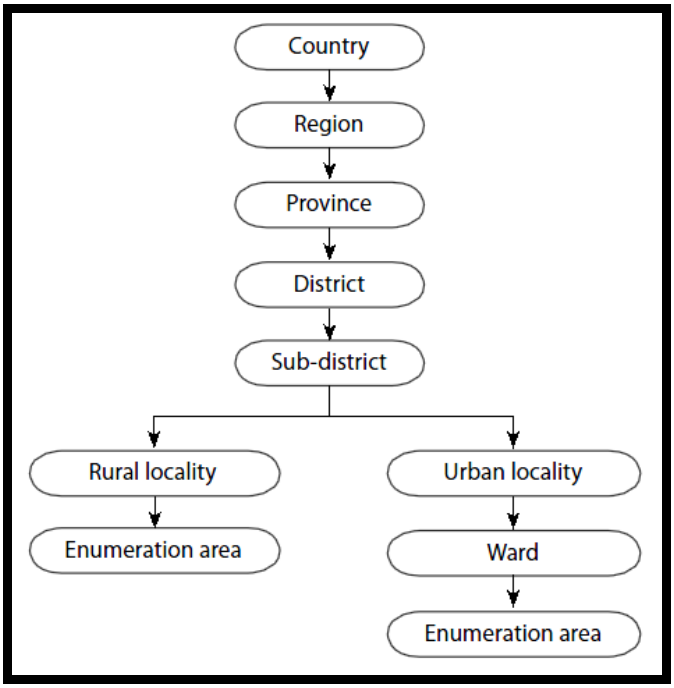


Figure 2.4. A simple census geographic hierarchy (UN, 2009).

Only some of these hierarchical levels may have actual administrative roles; for example, the province, district and locality levels may have capitals with local government offices that are responsible for those regions. Other units may have statistical roles alone; that is, they are designed for the display of data and not for administering territory. Figure 2.5 illustrates the nesting of administrative and census

units, using a simple example with only four hierarchical levels. In some instances, however, administrative units may not be completely nested. Especially when considering both administrative and other statistical reporting units, the census office may need to deal with a very complex system of geographic regions.

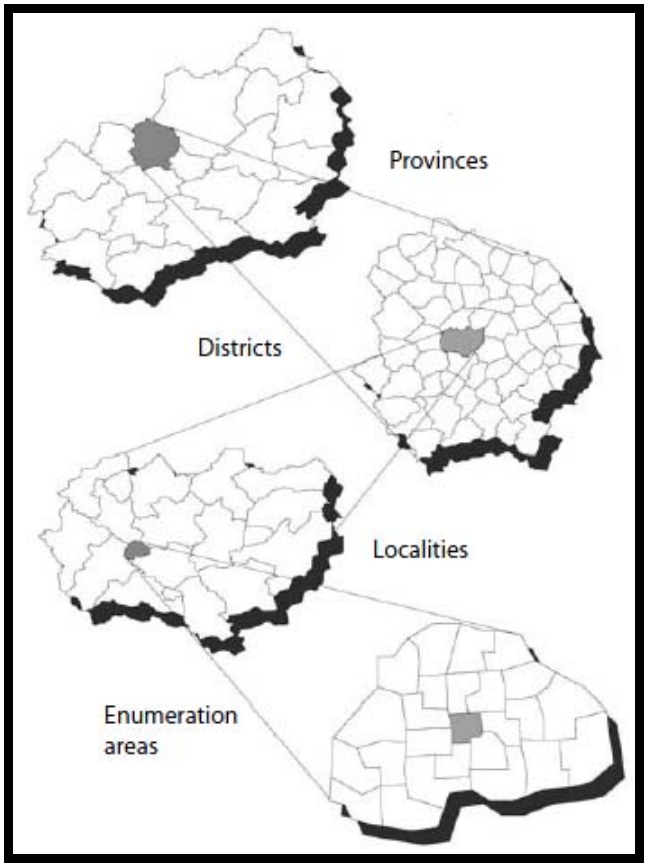


Figure 2.5 Illustration of a nested administrative hierarchy (UN, 2009).

Enumeration areas are the operational small geographic units for the collection or output of census data and are defined early in the census process. But in some countries these areas are solely used for data collection purposes in order to facilitate the enumerator's work. In these countries, the dissemination of census data via small output geographies is designed at the post-census phase. Whether manual or digital cartographic techniques are used, some of the delineation rules of enumeration areas or post-census small areas identification are similar. The design of enumeration areas should take various criteria into account. Among these criteria

are some that facilitate census data collection, while others pertain to the usefulness of EAs in producing output products. According to Coombes (1995) and UN (2009), correctly delineated, census EAs or small areas will accommodate to following rules;

- Be mutually exclusive (non-overlapping) and exhaustive (cover the entire country).
- Have boundaries that are easily identifiable on the ground.
- Be consistent with the administrative hierarchy.
- Be compact and have no pockets or disjoined sections.
- Have populations of approximately equally size.
- Be small and accessible enough to be covered by an enumerator within the census period.
- Be small and flexible enough to allow the widest range of tabulations for different statistical reporting units.
- Address the needs of government departments and other data users.
- Be useful for other types of censuses and data-collection activities as well.
- Be large enough to guarantee data privacy.
- Be socially homogenous at the possible extent.

In order to understand census geography classification and small area identification rules thoroughly, some national and international applications are examined in following parts of this thesis.

2.4. National Applications

In this section, United States, United Kingdom and South Africa census geography policies are closely examined with the aim to develop an appropriate methodology to create small area geography policy for Turkey according to rules accepted in these examples.

2.4.1. United States Census Geography Policy

According U.S. Census Bureau (2005-a), the success of a census or sample survey depends not only on how well the Census Bureau designs the questionnaire,

collects the data, and processes the results, but also on how well it links the collected data to geographic areas. In defining the geographic area framework for each specific census or survey, the geographic requirements consist of identifying the legal, administrative, and statistical entities to be used; publishing official standards for those entities, where appropriate; determining the names, numeric codes, and boundaries for the entities selected; entering the required information about these entities into the geographical data base; preparing the maps necessary to support the data collection and data dissemination functions; linking the address appearing on each census or survey questionnaire to its proper geographic location; and providing the reference files and technology needed to assign the data collected to the full set of geographic entities used to report the results of that census or survey. For these reasons, U.S. Census Bureau needed a stable structure of census collection geography at the census block level after the nationwide census at 1980s (U.S Census Bureau, 2005-a). The development of the Topologically Integrated Geographic Encoding and Referencing (TIGER) System, an automated geographic database, permitted the Census Bureau to delineate census blocks on a nationwide basis for the 1990 census (U. S. Census Bureau, 2005-e).

A modern society has vast informational needs, and a Nation as large as the United States and its territories contains many different kinds of geographic situations and settlement patterns. To respond to these needs and provide statistical data for these diverse situations and patterns, the geographic programs at the U.S. Census Bureau include several kinds of entities.

In its data collection operations, the Census Bureau must assign each person, household, housing unit, institution, farm, business establishment, or other responding entity to a specific location, and then assign that location to the tabulation units appropriate to the particular census or sample survey. This geographic coding (geocoding) process assures that the Census Bureau can provide correct counts for small geographic entities, and that both the Census Bureau and data users can accumulate the data for small entities to provide totals for larger geographic entities. Geography, then, is a basic element of the Census

Bureau's system for organizing and presenting statistical data to the public (U.S. Census Bureau, 2005-b).

The many geographic entities the Census Bureau recognizes and delineates often result in a geographic pattern that is quite complex. But to put simply, the U.S. Census Bureau classifies all geographic entities into two broad categories:

- Legal and administrative entities, and
- Statistical entities.

Legal and administrative entities generally originate from legal actions, treaties, statutes, ordinances, resolutions, court decisions, and the like. Local officials and others require data for governmental entities to fulfill a variety of needs. They require the boundaries of legal and administrative entities to manage a wide variety of programs and to conduct elections. The Census Bureau generally accepts, according to documentation by the appropriate authorities, the boundaries of these entities as they exist. Although the Census Bureau's data tabulations for legal and administrative entities are sufficient to satisfy the needs of many data users, information for these jurisdictions alone does not meet all data needs. Therefore, the Census Bureau also presents data for a second geographic category, statistical entities.

To provide the data tabulations needed by a majority of users, U.S. Census Bureau entangles the legal/administrative and statistical entities within a common framework, the geographic hierarchy (U.S. Census Bureau, 2005-b). There are numerous types of statistical entities, both large and small in population or land area. They include the groupings of States into regions and divisions, the metropolitan areas, the urbanized areas, some types of county subdivisions, and the small-area sub-hierarchy of census tracts and their subdivisions (Figure 2.6 and 2.7). Table 2.2 provides brief definitions of the most familiar types of statistical areas for U.S. Census Geography from top-to-bottom.

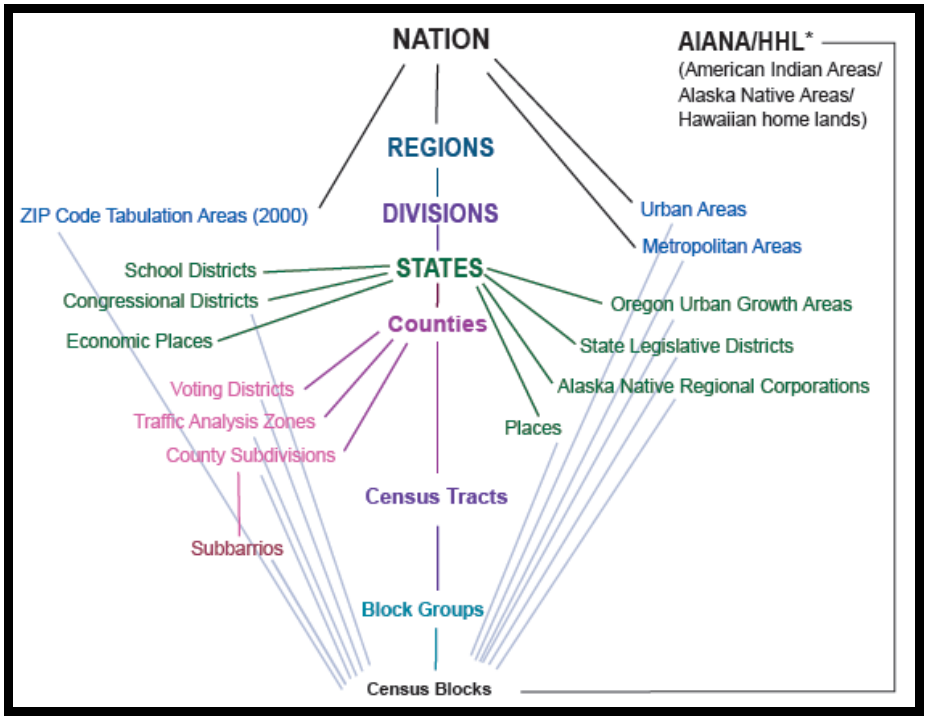


Figure 2.6 Standard hierarchy of legal and statistical census geographic entities (U.S. Census Bureau, 2009).

The Census Bureau recognizes numerous legally defined geographic entities for data presentation purposes, entities that generally are well known, such as States, counties, cities, and townships, whose governments function to provide services to the people living and working within their borders. These governmental units, however, usually do not provide sufficient geographic coverage to give a comprehensive, detailed picture of the distribution of the population on the lanrge extent, especially in highly populated counties. Moreover, in U.S., many of these governmental units have frequently changing boundaries, vastly differing population densities, extensive variation in population characteristics, and wide-ranging area sizes. These situations made it difficult for data users to summarize and analyze census statistics. To meet the need for geographic areas that would effectively supplement and complement the legally established areas, the Census Bureau, in association with data users across the Nation, has devised several types of geographic entities that generally define small, relatively permanent geographic

areas for which the Census Bureau can present statistics (U.S. Census Bureau, 2005-c).

Census blocks, enumeration districts (EDs), and block numbering areas (BNAs) were first used as operational units for taking and tabulating the census. As data users needed more small-area statistics, these operational units came into use as official entities for the tabulation and dissemination of decennial census statistics.

The delineation of census blocks and BGs could not begin until the TIGER data base contained an updated system of physical features and geographic boundaries. But now, perhaps nowhere within the framework of Census Bureau geography is the effect greater than at the small-area unit level (census tracts/BNAs, BGs, and census blocks) (U.S. Census Bureau, 2005-c). This has meant a vast expansion in the number of geographic entities in the data products of the Census Bureau, with the resulting increased opportunities for detailed data analysis. The availability of these low-level geographic entities provides extensive flexibility for data users to obtain counts for geographic units of specific interest to them.

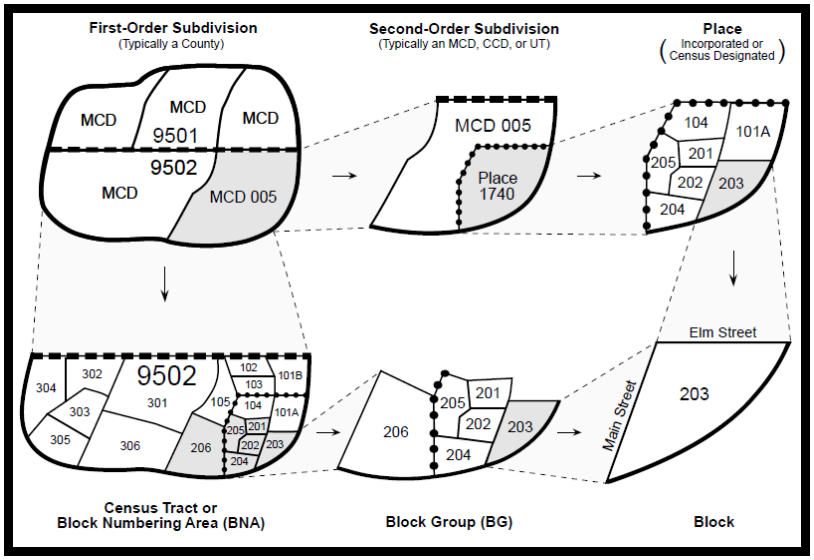


Figure 2.7 Small Area Geography in U.S. (U.S. Census Bureau, 2005-b)

Table 2.2 Main Legal and statistical entities in U.S. Census Geography (U.S. Census Bureau, 2008)

Legal and Statistical Entities	Definition
Census Regions (Both legal and statistical)	Census Regions are groupings of states and the District of Columbia that subdivide the United States for the presentation of census data. There are four census regions—Northeast, Midwest, South, and West. Each of the four census regions is divided into two or more census divisions.
Census Divisions (Both legal and statistical)	Census Divisions are groupings of states and the District of Columbia that are subdivisions of the four census regions (see Census Region). There are nine census divisions, which the Census Bureau established in 1910 for the presentation of census data.
States or Equivalent entities (Both legal and statistical)	States and equivalent entities are the primary governmental divisions of the United States. In addition to the 50 states, the Census Bureau treats some areas as the statistical equivalents of states for the purpose of data presentation where there are no states.
Counties or equivalent entities (Both legal and statistical)	The primary legal divisions of most states are termed counties. The Census Bureau continues to present data for these historical entities in order to provide comparable geographic units at the county level of the geographic hierarchy for these states and represents them as statistical entities in data products, but in reality they are legal entities. The Census Bureau treats some entities on other states as equivalents of counties for purposes of data presentation where there are no counties.
Census Tracts or Block Numbering Areas (BNAs) (Statistical entities)	Census Tracts are small, relatively permanent statistical subdivisions of a county or equivalent entity and are updated by local participants prior to each decennial census. Census tracts generally have a population size between 1,200 and 8,000 people with an optimum size of 4,000 people. The spatial size of census tracts varies widely depending on the density of settlement. Census tract boundaries are delineated with the intention of being maintained over a long time so that statistical comparisons can be made. Census tracts are split due to population growth or merged as a result of population decline. Census tract boundaries generally follow visible and identifiable features. State and county boundaries always are census tract boundaries in the standard census geographic hierarchy. When first established, census tracts are to be as homogeneous as possible with respect to population characteristics, economic status, and living conditions. Block numbering areas (BNAs) are geographic entities similar to census tracts, and delineated in counties (or the statistical equivalents of counties) without census tracts.

Table 2.2 (Continued) Main Legal and statistical entities in U.S. Census Geography (U.S. Census Bureau, 2008)

Block Groups (BGs) (Statistical entities)	Block Groups (BGs) are clusters of blocks within the same census tracts. Block groups generally contain between 600 and 3,000 people. A BG usually covers a contiguous area. Each census tract contains at least one BG and BGs are uniquely numbered within census tract. Within the standard census geographic hierarchy, BGs never cross county or census tract boundaries but may cross other boundaries on hierarchy.
Census Blocks (Statistical entities)	Census blocks are statistical areas bounded by visible features, such as streets, roads, streams, and railroad tracks, and by non-visible boundaries, such as city, town, township, and county limits, and short line-of-sight extensions of streets and roads. Generally, census blocks are small in area; for example, a block in a city bounded on all sides by streets. Census blocks in suburban and rural areas may be large, irregular, and bounded by a variety of features, such as roads, streams, and/or transmission line rights-of-way. In remote areas, census blocks may encompass hundreds of square miles.

While both categories of geographic units, legal/administrative entities and statistical entities, serve the common purpose of presenting Census Bureau data, the concepts, principles, and criteria for recognizing the entities in each category of legal and statistical involve different preparations by the Census Bureau. For both categories, it is critical that the Census Bureau establish and implement standards, guidelines, and criteria for defining, identifying, and delineating the geographic entities.

Delineation process for legal/administrative entities generally is well defined. Legislation or administrative measures create them, specify their governmental or administrative functions, and contain provisions for establishing and changing their names and boundaries.

On the other hand, a similar set of operations applies to statistical entities, but there is an important difference. Once the Census Bureau justifies the need for a new type of geographic area in terms of various principles, it must establish generally accepted criteria and guidelines for the identification and delineation of the new entity which are explained below (U.S. Census Bureau, 2005-b).

- Consistency is especially relevant to statistical areas, where the Census Bureau is largely responsible for establishing and implementing the criteria, standards, and guidelines that define these areas.
- It is desirable to maintain historical comparability of geographic entities from one census or sample survey to the next.
- The U.S. Census Bureau uses two basic principles in establishing and revising statistical entities. One recognizes a statistical entity by the similarity of its component parts, or the homogeneity principle. The functional integration principle views a statistical entity as a nucleus with its surrounding zone of influence. The homogeneity principle involves combining a group of people, housing units, or business establishments with similar characteristics into a single geographic area. The functional integration principle involves the grouping together, into a single statistical area, the people, housing, or business establishments that share a central nucleus along with the surrounding, functionally related entities, such as a large city and its suburbs. Because sources of these data generally involve looking at relationships among smaller entities, statistical entities based on functional integration often are more extensive in size than those based on homogeneity.
- The need for appropriate boundaries is a longstanding concern of census geography. Census tract and BNA boundaries generally follow permanent, visible features, such as streets, roads, highways, rivers, canals, railroads, and high-tension power lines. Pipelines and ridge lines may be acceptable when no other choice is available.
- Easy identification is a greater concern for most statistical entities. In establishing names for statistical entities, U.S. Census Bureau encourages the use of descriptive terms such as names that are known and already in local use. This requirement originally stemmed from the need for enumerators to know the exact limits of the areas they were assigned to enumerate. The use of such definite, easily recognized boundaries also makes it possible for data users to relate information from local records or other sources to the appropriate statistical entity.

- The Census Bureau provides population size guidelines in its criteria for most types of statistical entities. The size criterion generally determines the maximum number of such entities that someone can establish within a given county or other jurisdiction. In subdividing larger areas such as counties into smaller entities (for example, census tracts or BNAs), it is important to keep in mind their minimum desirable population size because of the many data items that are based only on sample responses and confidentiality. For such reasons, the Census Bureau recommends that a census tract contain at least 2,500 people.
- The geographic coverage of each type of statistical area varies according to their purpose. To be of use, major regions and subregions usually must cover the entire Nation; that is, they must provide complete geographic coverage. On the local level, census tracts or BNAs must cover an entire county and so do BGs and blocks.
- Compactness of shape is a desirable quality in a statistical entity, particularly for functionally defined ones; thus, it usually makes sense for their peripheries to be approximately equidistant from the centers. Twisted or elongated areas present the possibility that the statistical characteristics of the extremities will differ from those of the center or each other. If there are irregularities of shape, they should reflect geographic peculiarities related to the population, housing units or establishments the area contains, and there should be a justification in terms of major criteria such as integration or homogeneity.

Census tracts, BNAs, census block groups which conform to the guidelines above are the smallest geographic areas for which the U.S. Census Bureau collects, tabulates and disseminates decennial census data. Census data for these areas serve as a valuable source for small-area geographic studies. The BG is the smallest geographic entity for which the decennial census tabulates and publishes sample data. It has now largely replaced the earlier enumeration district (ED) as a small-area geographic unit for purposes of data presentation (U.S. Census Bureau, 2005-d, 2005-e).

2.4.2. United Kingdom Census Geography Policy

The UK Census, undertaken every ten years, collects population and other statistics essential to those who have to plan and allocate resources. Although the Census occurs simultaneously in all parts of the UK, the responsible body in England and Wales is the Office for National Statistics (ONS). The most recent Census took place on 2001 (ONS, 2009-a).

According to ONS (2009-b), geography is the key to virtually all National Statistics. It provides the structure for collecting, processing, storing and aggregating the data. Because in UK, the framework provided by geography is often the only factor for different datasets have in common. Usually, data collection is undertaken by enumerators visiting every identifiable address and leaving a census form for completion. At this point, the entire address-level geographical distribution of the population is to some degree known by enumerators. However, the census is undertaken in the context of confidentiality constraints designed to prevent the disclosure of information relating to identifiable individuals, and eventually geographical aggregation is one of the keys to dissemination of census results.

Before 2001, a feature of previous UK censuses was that the small area geographies used for the output of data have been the same as those for data collection, namely the enumeration districts (EDs) (Figure 2.8-a). ED design was mainly done according to three principles;

- EDs typically contained 200 households and 400 persons.
- ED design was undertaken by manually drawing boundaries in such a way as to equalize the workloads of enumerators as far as possible.
- EDs were constructed to nest neatly within the larger administrative boundaries of UK which were wards.

But, there were wide variations in ED shapes and sizes which caused the enumerated populations of some EDs falling below the minimum thresholds for the publication of census results (Martin, 1997).

Besides problems related to EDs, wards were also problematic for collecting and publishing statistics which are stated below (ONS, 2005);

- Wards are especially drawn up for electoral purposes, not census.
- Wards have an average population of around 6000 people and the data produced can mask local variations and hide patterns.
- Wards also vary greatly in size – from less than 1000 resident population to over 30000. This variation in size makes wards less preferable for national comparison or applying policy.
- Wards are subject to regular boundary change. Boundary changes make it difficult to compare data over time and integrate data from separate datasets produced at slightly different times. Consequently, the changes in boundaries affect the stability of EDs, because EDs nests within wards.

In addition to problems related to EDs and wards, in UK, there are many different geographic unit types (administrative, health, electoral, postcode, etc.) and their boundaries frequently do not align. A range of geographies is liable to frequent revisions. The UK's inconsistent geography has made it extremely challenging to produce and compare meaningful statistics over time (ONS, 2009-b). Especially, the ED population outputs of 1991 were criticized because there was no integration with widely used postcode geography (Martin, 2000).

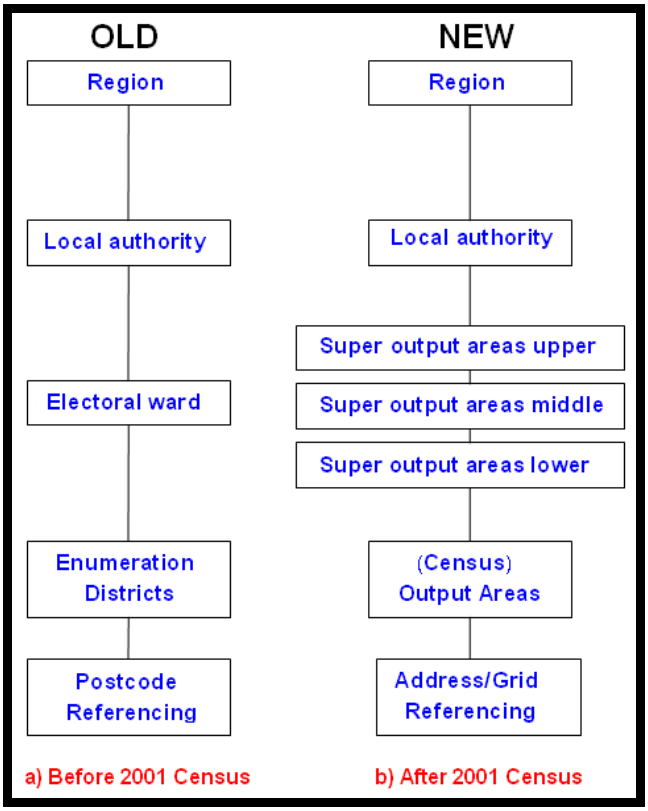


Figure 2.8 UK Census Geography hierarchy: a) Before 2001 census and b) After 2001 census (Frosztega & Estibals, 2004).

In the light of these problems, ONS decided to adopt a new, more flexible and future-proof approach. Part of the Neighborhood Renewal Agenda called "Report of Policy Action Team 18: Better Information" (SEU, 2000) highlighted a critical need for better information about local areas to overcome problems related to geographical referencing identified below;

- Provision of flexible outputs for small areas.
- Enabling national comparability.
- Reduction on the impact of boundary change on information about small areas.

In order to overcome these deficiencies, an entirely new methodology was adopted for 2001 Census, based on Openshaw and Rao's (1995) Automated Zoning Procedure (AZP), with Output Area (OA) creation taking place at the post-

enumeration phase and making use of population totals and social characteristics actually collected by the census. The OA geography for the 2001 Census in UK has been created by ONS and Prof. David J. Martin from Southampton University independently from the ED geography used for data collection on a GIS environment. By this way, ED remained only as a collection geography and OAs became the main statistical “building blocks” for census data release.

In order to design OAs which addresses the problems mentioned before, some principles and rules have been developed as explained below (Martin, 2002);

1) Firstly, OAs should be above the minimum and maximum population and household thresholds.
2) Secodly, OAs must be created from the unit postcode areas and individual addresses which are populated with the necessary individual and aggregate census information.
3) Thirdly, OAs should be standardized to the extent possible according to their population sizes, internal maximum social homogeneity (based on tenure of household and dwelling type), and the shape of areas (especially more compact geographical shapes are preferred).
4) Lastly, Oas must be bounded by obvious barriers such as major roads, ward boundaries where possible, rivers, streams etc. But in time, ward boundary change will progressively break the relationship with wards. In addition, urban/rural mixes should be avoided where possible.

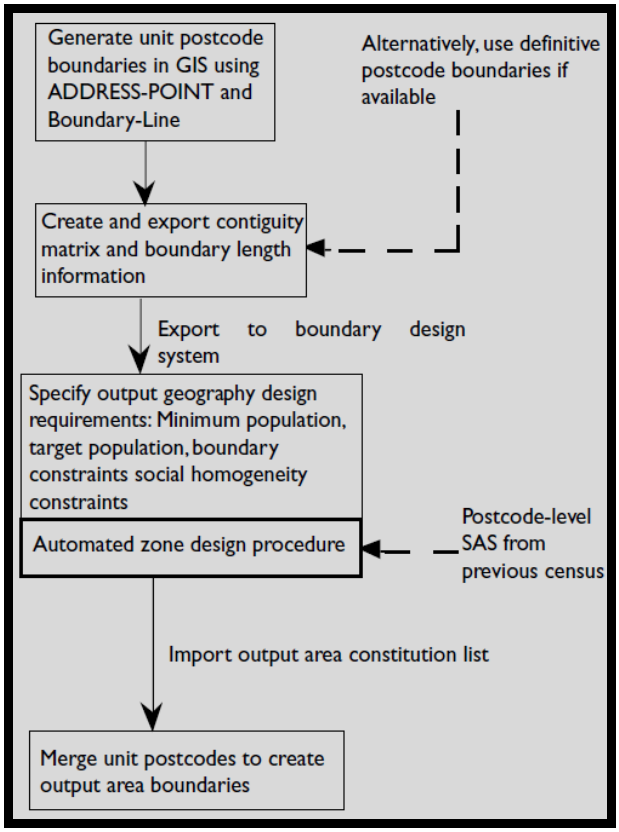


Figure 2.9 Basic structure of output geography design system (Martin, 1997)

Output areas in England are created using fully automated systems with common and consistent criteria across the country. The production system has developed recent years because of the increased power in computing and automatic zoning methods, and also through the availability of co-ordinate referenced and post-coded data. Martin (2003) says that the automatic zoning methods "…make use of the contiguity information available from the GIS containing the unit postcode polygons." A tool is used to apply this method called "Automated Zone Matching (AZM)" tool (Web 1). Through the procedure, first step is to estimate the number of probable OAs that should fall within a constraining polygon; secondly, according to the given population thresholds, adjacent postcodes are randomly aggregated to form above-threshold OAs; thirdly, some kind of statistical measurements are done to test the results according to design principles defined at the beginning; then the deviations from the target population sizes are measured by the sum of the squared differences between OA populations and the target population size. Here the

measurements of social homogeneity and geometric shape are considered separately. Finally, postcode polygons are swapped between adjacent OAs terms of their effects on these statistical measures (Martin, 2002).

In England system, two of the most effective variables on measuring the homogeneity of the regions are, as mentioned above, dwelling type and tenure; because the structure of the built environment and property ownership patterns may reflect the characteristics of that area. There are, according to the current system, seven dwelling and four tenure categories (Table 2.3). These two are combined with each other, by equal weights, to determine the almost homogeneous areas. Sometimes 'ethnic groups' data are used while defining the homogeneity but it has little effect on the final decisions.

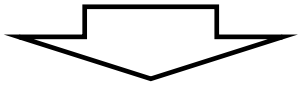
Table 2.3 Dwelling type and tenure categories in England.

Dwelling type	**Tenure**
Owner-occupied	Detached
Rented privately	Semi-detached
LA/HA	Terraced
Other	Flat
	Part-house
	Commercial
	Non-permanent

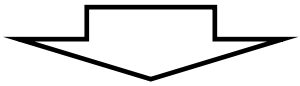
LA/HA: Rented from Local Authority (LA) or Health Authority (HA) (Martin, 2002).

Application of this methodology has resulted in 175.434 OAs. Based on these OAs, ONS created a census geography hierarchy composing of layers with areas increasing in size as one moves up (Figure 2.8-b). The characteristics of these geographical layers are explained in diagram below from bottom-to-top (Figure 2.10) (ONS, 2005).

Source Records: Individual addresses, postcode boundaries and census information



	Output Area Hierarchy	Characteristics
Wards (8800 units average population: 6000)	**Output Areas (OAs)** (175.434 units)	The lowest level unit to be used for aggregating and publishing census information. OAs are built to be as homogenous as possible and with a target size of 125 households (300 persons) after 2001 census.
	Lower Layer Super Output Areas (SOAs) (34.378 units)	Minimum population 1000; mean 1500. Built from groups of OAs (typically 5) The Lower Layer SOAs in England and Wales were generated again by AZP algorithm which merged OAs by taking into account measures of population size, mutual proximity and social homogeneity.
	Middle Layer Super Output Areas (SOAs) (7.193 units)	Minimum population 5000; mean 7200. Built from groups of Lower Layer SOAs. The Middle Layer SOAs, were generated via a two-stage process: (1) A draft set was generated by computer, in the same manner as the Lower Layer SOAs. (2) Local authorities and other local agencies were invited to propose changes to the draft boundaries in order to establish SOAs that better met local needs.
	Upper Layer Super Output Areas (SOAs)	Minimum population size 25,000. The nature of Upper Layer SOAs has yet to be determined.



Local Authority Districts

Figure 2.10 Geographical Layers of Output Areas in England and Wales (ONS, 2005).

UK census geography is one of the good examples of prototype automated output geography designs in the world which uses GIS for enumeration planning and management. ONS uses AZM (Automated Zone Matching) tool (which is developed by Prof. David Martin from Southampton University (Martin, 2002) to accomplish the preferred constraints for small areas. This tool's algorithm is based on the Openshaw's Automated Zoning Procedure (AZP). However, disadvantage of the tool is that it works with old ESRI Arc/Info GIS command prompts which are available in versions before 8.1 and used for preparing coverage type files. So, the

coverage files before and after version 8.1 differs by their attribute table structure. The tool does not accept the newer version coverage files to work with. Even the tool is freeware, it was impossible to use the tool for this particular thesis. But, after getting in touch with Prof. David Martin by e-mail, it is learned that a new tool called "AZTool" (Automated Zone Tool) is being prepared which uses the same procedure to produce small areas and works on newer versions of ESRI ArcGIS Desktop software and MS Windows operating systems. For further studies, this tool can also be employed to produce small areas for Turkey.

2.4.3. South African Census Geography

Every GIS professional is familiar with the notion that location can integrate disparate layers of information. This is no different in Statistics South Africa (Stats SA). The different layers of information benefit both the divisions of geography and social statistics where key national demographic and economic data are gathered (Lombaard, 2005).

Good statistics requires good frames. These frames are governed by quality methodologies and standards. Stats SA has over the past few years, elevated the importance of the geographic frame, geographic methods and geographic standards, in essence, the inclusion of geographical Knowledge in the production and dissemination of statistics. Thus, Stats SA formalized this strategic positioning of geo-information in the system of statistics (Lehohla, 2005).

According to Lehohla (2005), "…central to the change that occurred in Stats SA from 1996 was that Stats SA did not only recognize the importance of utilizing standard geographies for collecting, disseminating and comparing data but implemented a focused investment in geography because Stats SA knew that it constituted a relevance glue". Initiated as standard census geographies, and expanding as the geographical framework within which various statistics can operate, the geographical frame maintains the relationships between geographical layers, and it enables proper geographical usage of statistical data.

The data collected during the national population census of 1996 was published and disseminated at the EA spatial level. This has allowed all users of the data to aggregate the data based on their specific spatial entities of choice.

For census 2001, the division of the country into EAs was done according to certain rules. The fundamental rule of demarcating an EA was that an EA should not cross an administrative or social boundary (Statistics South Africa, 2007). Census mapping is important for execution, collection and dissemination of census data and the process must ensure that EA polygons are contained within a geographical frame i.e. when EAs are aggregated, they constitute the boundaries for higher geographical areas like municipalities and provinces and EAs cannot cross such boundaries. It was further intended to demarcate as near as possible to the 1996 EA, magisterial districts and tribal authority boundary as possible. It was also a requirement for an EA to be categorized by type and either fall under urban, farms or traditional areas and not to mix any land uses types.

But the dissemination of census data was changed with the 2001 census, because, as Grobbelaar, (2005) explains; there was a differencing problem which comprised cross tabulation of 2 or more variables at the EA geographical level. This resulted in such small totals that an individual's anonymity was manifested.

Due to concerns in regards to confidentiality, the lowest level on which the Census 2001 data was released was the second tier of the spatial hierarchy namely the sub place, which relates to suburbs, wards, villages, farms or informal settlements. In an effort to address user concerns in regards to the above mentioned, Stats SA undertook the responsibility of supplying users with custom aggregated data sets, based on the users spatial entity preferences, as long as confidentiality was kept in tact.

Subsequently a project was initiated in 2004, with the objective of developing a spatial layer for the purposes of disseminating data for certain census variables at a level lower than the sub place and as spatially similar to the EA, as permitted by confidentiality.

In order to address a continuous demand from users and balance the confidentiality requirement, Stats SA undertook a study to identify a geographical layer that comprises of units that contain a large enough population to reduce the risk of the possible identification of individuals when cross tabulation of variables is done. The fundamental finding of this study was that a minimum population total of 500 persons per EA is required to ensure confidentiality. Inseparably linked to this was the restricting of census variables to be published with the layer. Based on this, the development of a spatial layer for dissemination purposes was initiated which corresponds as much as possible to the EA layer, but with optimal confidentiality. This Small Area spatial layer is located between the EA and Sub Place in the geographical frame hierarchy (Figure 2.11).

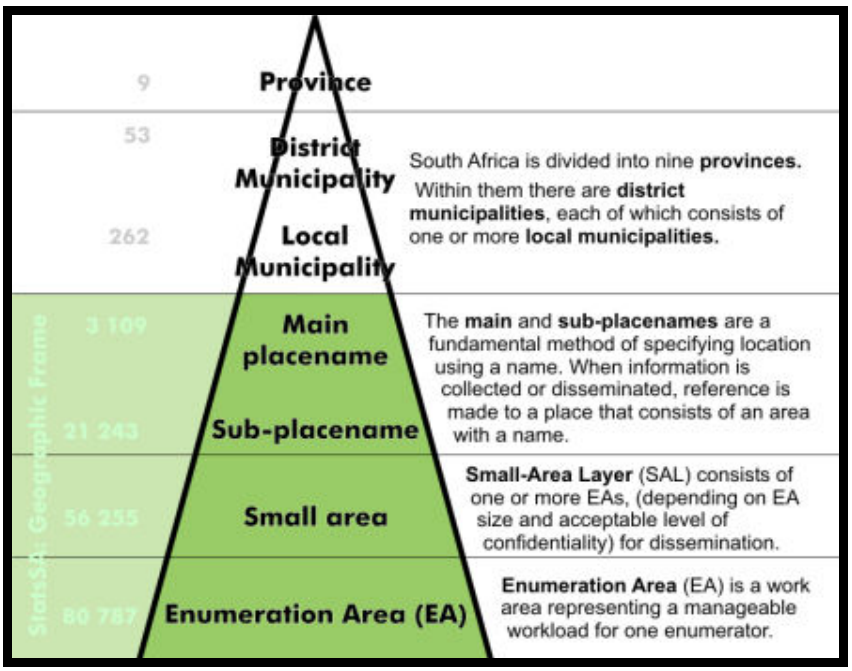


Figure 2.11 South African census geography hierarchy in year 2004 (Khumalo, 2009).

The automated spatial creation of the Small Area Layer (SAL) was based on the principle of merging individual Enumeration Areas within the Enumeration Area spatial layer. Merging was based upon a unique code allocated to the Enumeration

area if it adhered to a certain attribute and spatial requirements rule set. The rule set is as follows;

- The Enumeration Areas can only be merged if they are within same Sub Place (Consistency).
- The Enumeration Areas can only be merged if they have the same EA geography attribute type (Urban Formal, Urban Informal, Farms and Traditional (Tribal) Areas) (Homogeneity).
- An Enumeration Area can only be merged if its population is less than 500 (Population threshold).
- The resulting Small Area Layer (SAL) polygons must have a population total of 500 and more (Equal population).

The Small Area Layer therefore cannot be seen as the optimal solution for data dissemination not even to mention the possibility of using it for data collection - It is merely the result of an effort to bridge the gap between user needs and methodological and legislative constraints.

During the process of geographical classification, traditional authorities were worried about the demarcation of their areas. The board assured them that their land would not be divided between two different municipalities during determination of boundaries. But, the tribal authorities perceived the creation of new bodies in their areas as a way of diminishing their importance as actors of development to mere objects under local municipalities. Also, traditional leaders in most parts of the province were not adequately consulted for verification of boundaries during the municipal demarcation exercise, resulting in administrative boundaries that do not always suits or fit the ones that tribal authorities would have prescribed in their areas. Further to this, the process created significant animosity and have resulted in mistrust and they have they are of the opinion that their boundaries are not being recognized as valid ones (Khumalo, 2009). Using the GIS, different layers of information were overlaid on aerial photographs, providing an overhead view. This technology provided several advantages, but also had drawbacks as the use of these digital tools can never compensate for field visits to physically verify boundaries, especially in areas with undocumented boundaries like those in tribal areas (Statistics South Africa, 2007).

A quality census starts with knowledge of the whereabouts of all dwellings in the country, which enables every household to be visited, thus ensuring that every person in all parts of the country is counted by obeying the social boundaries like tribal boundaries. The Dwelling Frame (DF) project was therefore initiated in 2004 to capture the exact location and characteristics of every dwelling unit in the country by capturing the lines of latitude and lines of longitude through the use of GPS (Figure 2.12). The process of geo-referencing the dwellings will enable one to tell what exactly exists on the ground and will inform the process of demarcation or delineation of EAs to assign workloads to enumerators, assist in locating dwellings and managing fieldwork during enumeration, provide a register of dwellings against which census data is collected, and can be used for matching and cross-checking processed census records in the census post-enumeration survey (Statistics South Africa, 2007; Laldaparsad, 2007).

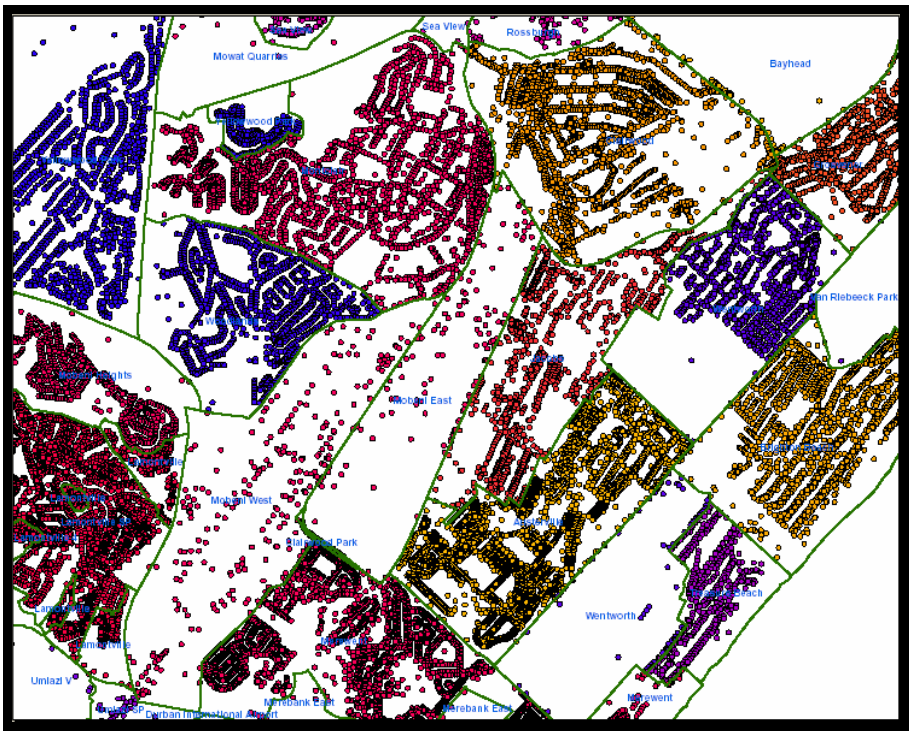


Figure 2.12 A snapshot from the dwelling frame project which shows the geographical positions of every dwelling (Lehohla, 2005).

After the dwelling frame project, Statistics South Africa has declared that for census 2011 EA boundaries that are crossing social boundaries are no longer needed. Figure 2.13 shows the geographical frame to be used for census 2011, with the register of dwellings forming the most basic unit of the hierarchy. The dwelling information will be used in the demarcation of EAs and to inform correct place-name boundaries. It is important to note that in census 2001 the smallest working unit was the enumeration area, but for census 2011 the smallest working unit will be dwelling frame which will provide a register of the spatial location of each dwelling unit in the country.

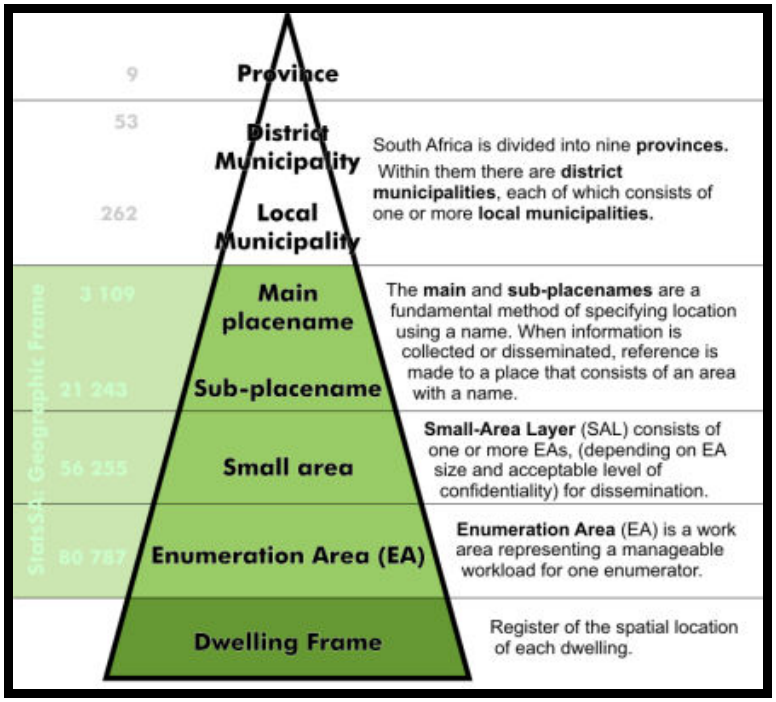


Figure 2.13 South African census geography hierarchy for 2011 census (Khumalo, 2009).

South Africa expects that this detailed level of geography can re-define or correct all layers of geography in the frame, making elements of the frame more geographically reliable, for the production of reliable statistics.

2.5. International Applications

Besides national census geography applications, there are also international applications which aim comparability across nations by means of small area statistics rather than by large area statistics.

2.5.1. TANDEM Project: Towards a Common Geographical Base for Statistics across Europe

Over the recent years most NSI's (National Statistical Institutes) have noticed an increasing demand for high quality statistics, with higher resolution, disseminated with increasingly higher frequencies and harmonized over ever larger areas. Eurostat is the statistical office of the European communities whose task is to provide the European Union (EU) with statistics at European level that enable comparisons between countries and regions. GISCO, on the other hand, is the permanent service of Eurostat which promotes and stimulates the use of GIS within the European statistical system and the commission. Eurostat/GISCO has for some time argued for the need on basic statistical areas that could be used to improve the resolution and comparability of area-based statistics within the EU. Also, developments both in the field of remote sensing and efforts on the part of many NSIs to collect information with point-based strategies, have led to a growing need to agree on a common system of grids and grid methods to increase the comparability of both types of spatial statistics. In response to papers submitted by United Kingdom, Sweden and Finland at the meeting of the Working Party on Geographical Information Systems for Statistics held in Luxembourg on 20 and 21 October 1999 (Guiblin et al, 2001-a), it was suggested that a combined grid and region-based approach would be needed to tackle the limitations inherent in the Nomenclature of Territorial Units for Statistics (NUTS) system. As a result of these and other developments, the Tandem consortium, consisting of GIS groups from the Office of National Statistics (UK), Statistics Finland and Statistics Sweden, was invited to apply for a commission grant to study these questions further.

NUTS were established by Eurostat more than 30 years ago in order to provide a single uniform breakdown of territorial units for the production of regional statistics for the EU. The NUTS nomenclature (classification) was created and developed according to the following principles (Web 12):

a) The NUTS favours institutional breakdowns: Different criteria may be used in subdividing national territory into regions. These are normally split between normative and analytic criteria;

- *Normative regions* are the expression of a political will; their limits are fixed according to the tasks allocated to the territorial communities, according to the sizes of population necessary to carry out these tasks efficiently and economically, and according to historical, cultural and other factors;
- *Analytical (or functional) regions* are defined according to analytical requirements; they group together zones using geographical criteria (e.g., altitude or type of soil) or using socio-economic criteria (e.g., homogeneity, complementarities or polarity of regional economies).

For practical reasons to do with data availability and the implementation of regional policies, the NUTS nomenclature is based primarily on the institutional divisions (administrative boundaries) currently in force in the Member States (normative criteria).

b) The NUTS favours regional units of a general character: Territorial units specific to certain fields of activity (mining regions, rail traffic regions, farming regions, labour-market regions, etc.) may sometimes be used in certain Member States.

c) The NUTS is a three-level hierarchical classification: Since this is a hierarchical classification, the NUTS subdivides each Member State into a whole number of NUTS 1 regions, each of which is in turn subdivided into a whole number of NUTS 2 regions and so on.

The NUTS Regulation lays down the following minimum and maximum thresholds for the average size of the NUTS regions (Table 2.4).

Table 2.4 Minimum and maximum thresholds for the average size of the NUTS.

Level	Minimum	Maximum
NUTS 1	3 million	7 million
NUTS 2	800 000	3 million
NUTS 3	150 000	800 000

At a more detailed level, there are the districts and municipalities. These are called "Local Administrative Units" (LAU) and are not subject of the NUTS Regulation. So, detailed levels of breakdown were not considered.

The NUTS nomenclature serves as a reference for;

- The collection, development and harmonization of EU regional statistics,
- The socio-economic analyses of the regions,
- The framing of Community regional policies.

In the light of above explanations about NUTS, it is evident that NUTS has limitations, because"…projects contributing to the development of societies are no longer limited to the development of infrastructures within administrative borders, but has been forced to shift their focus to the development of networks whose output patterns are not satisfactorily captured by crude systems of "large area statistics" (Guiblin et al, 2001-b). In addition, the system of the NUTS areas or administrative areas in general, is far from ideal for flexible and comparable regional statistics. The system faces the problem of what has been termed the modifiable area unit problem (MAUP). However, little effort has been made internationally to achieve better spatial comparability in terms of regional statistics. So, Tandem consortium is not concerned with "large area statistics" as those aggregated on the NUTS hierarchy of administrative units, but focuses mainly on the feasibility of providing a system of "small area statistics (SAS)" for the EU. But, this system should not be regarded as substitute for the classical system of "large area statistics" such as NUTS, instead should rather be seen as a valuable extension to, or upgrading of, existing practices.

According to (Guiblin et al, 2001-b), apart from the reason that large areas are inefficient for projects contributing to the society's development, small statistical areas (SSAs) are essential according to Tandem consortium due to subjects summarized below;

- Necessity for alternative comparable building blocks/territorial divisions for European system for SAS,
- Visualization of data more effectively,
- Combination and comparison of data on different spatial units,
- Making better statistical/spatial analysis of data (to test spatial patterns and trends),
- International work promoting the use of spatial data such as INSPIRE (Infrastructure for Spatial Information in the European Community),
- Aggregation, disaggregation, re-aggregation,
- Delineation of functional areas (urban-rural delineation)
- Departure from static and descriptive spatial analysis to dynamic and analytical spatial analysis.

In Europe, there are fundamentally two practices used for data collection and analysis: point and area-based methods. So, definition of statistics which is useful for describing and analyzing the spatial distribution of phenomena in space is mainly based on a system of regions that are either equal in terms of area (regular tessellation) or population (irregular tessellation). Regular tessellation system is mainly based on small gridded delineation of the region into equal areas. On the other hand, irregular tessellation (blobs) considers the delineation of an area roughly equal in terms of population. For both of these systems, Tandem formulated four fundamental requirements for small statistical areas (SSAs);

1. SSAs should be as small as possible: Regarding the relative size of the areas in an integrated system of regular and irregular tessellations it is important that they are as small as possible in relation to size of the territory that must be analyzed. This seems, in terms of irregular tessellations, to indicate that they should ideally be much smaller than the smallest NUTS level.

2. SSAs should be comparable in terms of population and/or area: A key issue for flexible spatial analysis is the concept of scalability. A properly defined system of geographical building blocks based on high resolution base data would form the basis for a hierarchy of comparable territorial divisions. This would facilitate the combination of data given different spatial units or different spatial scales and provide more flexibility than the generation of combinations using existing administrative sources. A standardized System of Small Area Statistics should provide scalability of data for different scales of spatial analysis. Different scales may be thought of as different sizes of study "windows", which could be used to recommend the resolution of data needed.
3. SSAs should be homogenous: High resolution territorial data are usually necessary when one performs spatial analysis using GIS software. Statistics grouped by administrative areas are often not applicable for spatial analysis because they differ widely in size (or population). Also, the ecological fallacy and the modifiable area problem (MAUP) can be minimized when using high resolution data, preferably data optimized for size, shape and homogeneity of key characteristics, as building blocks for analysis.
4. SSAs should be linked to an adequate amount of statistics: The need for statistics is obvious, because without statistics there would be no analysis. Generally, statistics are available on relatively large administrative areas only. But, in order to develop small area statistics there are many methods available to either aggregate data from registers with coordinates on micro level, or providing by disaggregating statistics from larger to smaller units (small area estimation techniques etc.)

Since, the scope of this thesis does not cover the production of SSAs by using regular tessellation (grid data model), the method proposed by Tandem study for producing SSAs using irregular tessellation (areas) are further analyzed and found that the work package for irregular tesellation proposes the evaluation of zoning methods able to provide different output areas for geographical statistics. The tool that TANDEM utilizes in order to create SSAs is the same with the tool that United

Kingdom utilizes for creating output areas, namely AZM tool (Tammilehto-Luode, 2003).

2.6. Statistics and Geographic Information in Turkey

The Turkish Statistical Institute (TURKSTAT) is the only authorized technical and scientific institute which produces publications to fulfill Turkey's information needs on social, economic, and cultural subjects. The main function of TURKSTAT is to comprehensively determine information needs, collect and compile data, and finally, to present information to its users according to the highest international standards. TURKSTAT has 26 regional offices and 1 central office in Turkey. The statistics and products generated by TURKSTAT are currently used as a guide by governmental institutions and foundations, universities, private organizations, decision makers and researchers.

However, Demir & Toprak (2004) say that "…as the world entered into a phase, which is commonly described as the process of globalization, statistical offices of many countries come to cope with the challenges of the new demands from decision-makers and researchers." Therefore, statistics in Turkey has to be examined to see general picture and necessities in terms of census geography.

2.6.1. Steps towards compliance with European Union

The declaration of Turkey as the formal candidate country to the European Union (EU) in December 1999 caused to think about the adoption process to EU in the area of statistics. European Commission has '*The Accession Partnership*' rules, which declare short and long term priorities to the candidate countries. The priorities about statistics for Turkey include the following:

For the short term;

> *"…adopting a strategy for the further development of statistics, in particular demographic and social statistics, regional statistics, business statistics,*

external trade and agricultural statistics; bring the business register up to EU standards."

For the long term,;

> *"…adopting EU compatible statistical methodologies and practices, in particular as regards GDP (Gross Domestic Production) estimation, harmonized consumer price indices, short-term indicators, social statistics, business register and balance of payments; aligning macro-economic statistics further with the statistical acquis; and ensuring adequate training of staff and improve the administrative capacity"* (Demir & Toprak, 2004)

After the Accession Partnership, Turkey began setting itself as a member state of EU; and it was clear that, statistics would play a vital role in attaining the goals. This problem had a major priority among all the others. TURKSTAT was aware of that and began the harmonization studies in statistics immediately. Studies about this subject are still underway, and today TURKSTAT still has to realize many projects to reach the international standards (mainly European Union), and to reach the aim of improving the statistical system of Turkey. In this framework, a High Level Committee has been established to evaluate the situation and to identify overall and key objectives for the adoption studies of the SIS. Some steps taken on this way are described in the 'Country Paper: Republic Of Turkey', prepared by TURKSTAT (2002). There are mainly eight steps defined in this paper, and one of them is the project of "Upgrading the Turkish Statistical System".

TURKSTAT has prepared the proposal of this project in November 2001 with the assistance of two EU consultants in order to accomplish the short and long term priorities of Accession Partnership and National Plan. Its total budget is 15.3 million Euro for 36 months period (2002-2004). The project includes many components to upgrade the Turkish statistical system, but one of them, '*Upgrading of the regional statistical system, introduction of NUTS classification, data collection and dissemination system, and a regional indicator database*' is the focus point of this thesis. This component will be examined, and a new geographic statistical system

will be proposed to develop the existing NUTS classifications of Turkey, and to make a contribution to the adoption process.

2.6.2. NUTS in Turkey

As explained in section 2.5.1, Eurostat (2003) defines NUTS as the regions established by Eurostat to provide a single uniform breakdown of territorial units for the production of regional statistics for the European Union (EU). One of the most important aims of this regulation is to take under control the inevitable change in the administrative boundaries through years in the Member States. They use NUTS regions to minimize the effects of these changes on the availability and comparability of regional statistics. It has the benefit of being well established, considerably stable, hierarchical and well organized to the national statistical regions.

In Turkey, until recently, statistical classifications of regions were made according to administrative divisions (local statistics are still based on administrative boundaries). Turkey, consisting of 81 provinces, partly adapted to the European statistical classification in September 2002. Turkey was divided into 12 NUTS-1 units, 26 NUTS- 2 units and 81 NUTS-3 (cities) (Figure 2.14) (Appendix A). All regional planning efforts are carried and incentives extended on the basis of these NUTS regions. These NUTS regions were created and developed according to the principles of Eurostat which was mentioned in section 2.5.1. The next lower level administrative unit is composed of the districts, however the current boundary structure of some existent or new districts in Adana, Ankara, Antalya, Diyarbakır, Erzurum, Eskişehir, İstanbul, İzmir, Kocaeli, Mersin, Sakarya and Samsun cities is not precise and definitive.

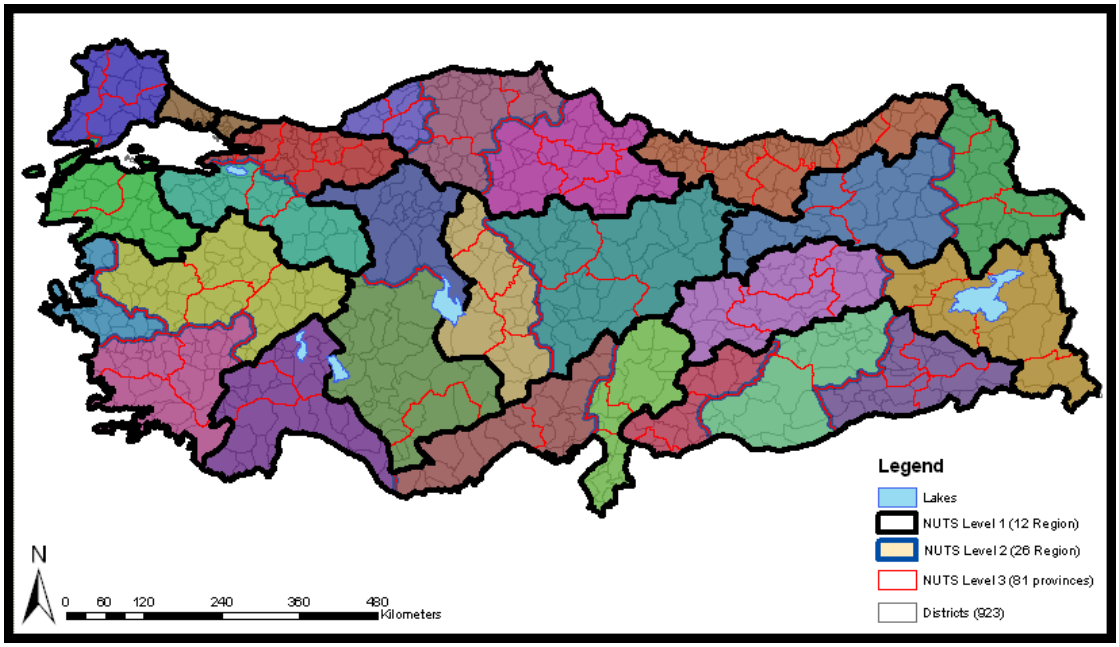


Figure 2.14 Current NUTS 1, NUTS 2, and NUTS 3 regions of Turkey

In addition to the problems with district boundaries, the NUTS regional breakdown in Turkey is also problematic, because, contrary to some minimum and maximum population thresholds for the average size of the NUTS regions determined by Eurostat, in Turkey these population requirements have been partly ignored. For instance, population number of a NUTS 3 region can not exceed 800.000 according to Eurostat's rules; but Istanbul, which is also a NUTS 3 region, has more than 12 million inhabitants according to 2008 Address Based Population Register System (ABPRS) results which can be checked on the website of TURKSTAT (Web 11).

Table 2.5 Population thresholds of NUTS regions in Turkey according to the results of 2008 ABPRS study (165).

Level	Minimum	Maximum
NUTS 1	**2.201.862**	**12.697.164**
NUTS 2	**737.308**	**12.697.164**
NUTS 3	**75.675**	**12.697.164**

In addition, NUTS regions are also the main units for data dissemination in TURKSAT. So, majority of statistics, being disseminated, are on the basis of the cities, which are the geographical unit of administrative classifications and very few of them are produced and disseminated based on non-precise districts. The cities and districts have relatively big sizes in respect to population; and when statistics are produced for these divisions, the specific characteristics of local inhabitants are ignored. As a result, data users see only the average, not the real. Therefore, NUTS regional breakdown and district level is not adequate for data dissemination and production.

2.6.3. Census Activities in Turkey

The main census activities maintained by TURKSTAT are population censuses, 'General Agriculture Census', 'Building Census', and 'General Industry and Business Census'. In parallel to the aims of the thesis, only the population censuses will be examined. Until year 2006 population censuses were named as "General Population Census" and were carried out by traditional method.

The aim of the "General Population Censuses" in Turkey were to accurately determine the total number of population, and social and economic characteristics of people according to the administrative boundaries. Governmental functions like defense, taxation, and justice require administrative units which are the lower level units of nation. These units are sometimes 'natural' or 'historical' regions, and they are more or less arbitrary units.

"The success of a census or sample survey depends not only on how well the authorized institution designs the questionnaire, collects the data, and processes the results, but also on how well it links the collected data to geographic areas" (US Census Bureau, 2005-a). As it was mentioned just before, the administrative geography is also used for statistical purposes in Turkey; but it must not be forgotten that; administrative areas are designed mainly for administrative purposes and in most cases not appropriate for statistical purposes.

Comparability of geographic areas from one census to the next is often a major concern for data users, and some users require a stable set of boundaries that permit historical comparisons. This, sometimes, becomes impossible because of the changed, divided, or combined administrative boundaries. For instance, the total number of provinces in Turkey was 63 in 1950, 67 in 1960, 73 in 1990, 80 in 1997; and now, since 1999, this number is 81. That is, 18 new provinces have been added to the original ones in 49 years. The similar changes occur in local levels, too; and these changes make it impossible to generate, analyze and use the time-series data and cause loss of information about related parts of the country.

TURKSTAT produced statistics mainly for provinces, and in some cases, for district level by “General Population Censuses”. The areas of some provinces are too large for specific regional studies such as environmental impact assessment or market analyses and for many other scientific analyses. Thus, it is essential to produce statistics for smaller areas. Furthermore, municipal and metropolitan boundaries are not uniformly regulated across the country, so they may not closely follow urban population boundaries.

After year 2006, the population census method has been changed into register based method and censuses are named as “Address Based Population Register System” which is explained in the next section.

2.6.4. Address Based Population Register System (ABPRS)

The first population census was carried out in 1927 and the next population censuses were carried out between 1935 and 1990 regularly, in years ending with 0 and 5 in Turkey. After 1990, population censuses have been decided to be carried out in years ending with 0. According to the Official Statistical Programme of Turkey, next census is going to be carried out in 2011, in order to keep up with the population census calendar of European countries which will also carry out population census in year 2011.

All population censuses were carried out with traditional method in one day by application of a curfew with “de facto” definition. Information on usual residence (de jure) population and household structure are not available from past censuses because of listing only present persons on the census day. However main problem in the census was over counting of population (imaginary population) and out of date because of ten years interval. Therefore, TURKSTAT aimed to change the method of population census in order to produce more reliable and up to date information on population size and characteristics. For this purpose, population registration system has been improved to cover usual residence address of all people living in the country. This system is considered as a base source for the next census planed in 2011.

MERNIS (Central Population Registration System) database implemented by the General Directorate of Population and Citizenship Affairs (GDPCA) of the Ministry of Interior keeps population registers of Turkish Citizens with unique 11 digits identification number in the form of family ledgers. The system provides information on vital events and relationship between members of generations. However, no information on address of usual residence of Turkish Citizen is available. Therefore, the system could not be used for census purposes. The project of Address Based Population Registration System (ABPRS) has been implemented for the period between 2006 and 2007. Main purposes of the project are to establish National Address Database that covers all addresses with unique ID within the boundaries of the country and to improve registration system by collecting information on place of usual residence for Turkish Citizens and foreigners living in Turkey with a field work pairing the citizen’s ID with address ID. In addition, administrative purpose of the project is to provide standardized and updated information on residence address of persons living in the country particularly for public services. Thus, bureaucracy issues will be decreased by sharing this information with governmental bodies (Demirci & Taştı, 2009).

Legal base for the project is the Population Services Law which was approved in April 2006. Turkish Statistical Institute was charged to establish two databases and General Directorate of Population and Citizenship Affairs, Ministry of Interior was

charged for updating and maintenance of the system. The project was realized by application of five stages given in the Population Services Law: i) establishment of the National Address Database, ii) field application for collecting information on usual residence addresses, iii) data processing, iv) checking the usual residence addresses, and v) updating the system.

According to the Address and Numbering Regulation, name of streets and building number are given by municipalities in localities having municipal organizations, and by special provincial administrations in villages. In order to eliminate the problems related with changes of name of streets, unique code is given to each street. This study constituted the source for the National Address Database (NAD). During this stage, around 40 millions addresses were entered in the NAD from all over the country by secure web based applications.

The field application was carried out by TURKSTAT in order to collect information on the unique ID, place of residence, age, sex, relationship between household members and completed level of education of Turkish Citizens and foreigners living in Turkey. Two different forms were used to collect information for households and institutional places. Also, separate form was used for persons who were not registered in MERNIS and did not have ID number.

These studies have been finished at the end of July 2007. Under the Official Statistical Program, statistics on population size, births, deaths and other vital events, and migration is panned to be produced according to this improved population registers (Demirci & Taştı, 2009).

It is obvious that geography is a basic element of the census system for organizing and presenting statistical data to the public; but the geography term in the current statistical system needs to be revised and now, with the new register based census methodology, it is possible to construct more detailed levels for Turkey's statistical system because the census data is now available with a spatial dimension like addresses. Although the TURKSTAT's data tabulations for legal and administrative entities are sufficient to satisfy the needs of many data users, information for these

jurisdictions alone does not meet all data needs. Therefore, the TURKSTAT must present data for a second geographic category, small statistical areas and this thesis aims to contruct them by utilizing ABPRS data.

2.6.5. Small Statistical Areas

Backer et al. (2002) defend that improving the general quality of the classical systems of official statistics is not the only reason to create a new geographical base for statistics in Europe. The main reason is to answer the existing and potential importance of information and knowledge systems in production, distribution and consumption of goods and services in the societies. These activities are the main concerns of almost all institutions and organizations taking part through hierarchies of especially public projects. According to them, "…projects contributing to the development of societies are no longer limited to the development of infrastructures within administrative borders, but have been forced to shift their focus to the development of networks whose output patterns are not satisfactorily captured by crude systems of 'large area statistics'." Turkey strongly needs a new system of "small area statistics", too. This system should not completely substitute the current statistical constitution, rather it must be thought as a complementary to the present system.

A 'system of small area statistics' (SSAs) can be defined both formally and functionally. Formally, it is a system of knowledge based on statistical micro data. It includes statistics, regular and irregular geographical features, and methods. "A SSAS consists of, and is in turn itself a part of and designed to fit into, a constantly changing hierarchical network of processes dedicated to the production and analysis of qualified spatial information" (Backer et al., 2002). Functionally, SSAs is a collection of processes such as aggregations, benchmarks, and data, spatial, and temporal analysis. It is used "…to improve the results of overriding projects to counter threats and exploit opportunities in view of private and collective efforts to improve the human condition" (Backer et al., 2002).

The aim of the first “system of small areas” (SSA) idea was to meet the demand for better information for projects to investigate the opportunities and withstand to present and potential problems. It would also highly improve the comparability of statistics across the EU and make aggregations, dis-aggregations, and re-aggregations of data easier in a network of systems of statistical areas. SSA also raised the quality of data sets by focusing on them and creating higher resolutions when used in combination with methods like sampling, small area estimation etc.

There are standards, guidelines, and criteria for defining, identifying, and delineating the small areas to be used in a population census. Small statistical areas must specify precise criteria for establishing the new component entities. After examining the international applications, basically four fundamental requirements have been formulated for small statistical areas (SSAs); (1) Homogeneity, (2) Compactness, (3) Population Equity within same level of statistical areas and (4) Easy Identification. The explanations of these requirements were given elaboratley throughout Chapter 2. As a result, this thesis aims to contruct SSAs for Turkish Statistical System by considering these criteria and by utilizing statsitical indicators from ABPRS.

CHAPTER 3

SPATIAL CLUSTERING AS A KNOWLEDGE DISCOVERY METHOD FOR SSAs

Census boundaries can be modified in several different ways. The simplest way is to aggregate the areas (i.e. EAs) to create a new set of spatial areas for a specific level in a census geography hierarchy based on attributes associated with each census area. Clustering techniques can also be used to group the census areas into new spatial features based on a set of census attributes and constraints. For this reason, this chapter includes concepts and previous applications related to spatial clustering techniques used in knowledge discovery and data mining domain. Since, clustering techniques are important in terms of SSA layer creation, stages of and concepts related to knowledge discovery and data mining terms are summarized in the beginning of the chapter. The chapter continues with explaining the difference between classic clustering and spatial clustering concepts and applications particularly used for mining knowledge on very large databases. Also, spatial clustering is discussed on the behalf of GIS. Then, some of these methods, developed by former studies and utilized in this thesis, are more elaborately explained. Finally, main software packages which are used for creating SSAs by utilizing selected spatial clustering techniques are explained by means of their capabilities.

3.1. Knowledge Discovery (KD) and Data Mining (DM)

Nowadays, we are face to face with a reality of life: data overload. In many fields, digital data collection and storage capacities resulted with the emergence of very large databases which exceeded the human ability to analyze the knowledge contained implicitly in them. As a result, there is a growing demand for computerized tools which automate this process and assist humans in extracting useful knowledge

from rapidly extending datasets. These tools are the subject of the emerging field of knowledge discovery (KD).

Fayyad (1996) defines KD as "…the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in the data". According to him, *data* are set of facts, i.e. incidents of crime, and *pattern* is an expression in a language or representation describing a subset of the data. By *nontrivial,* it is meant that some sort of search or inference is involved. The term *process* implies that KD is composed of many iterative steps explained as follows and are depicted in Figure 3.1;

1) **Domain knowledge:** Developing an understanding of the application domain and the relevant prior knowledge.
2) **Creation of target dataset:** Selecting a dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed.
3) **Data cleaning and preprocessing:** Removing noise, collecting necessary information about noise, deciding on strategies for handling missing data fields.
4) **Data reduction and projection:** Finding useful features to represent the data depending on the goal of the task.
5) **Consolidation:** Finding a particular data-mining method such as summarization, classification, regression or clustering which meets our goal.
6) **Exploratory analysis:** Choosing data mining algorithms or selection methods to be used for searching data patterns.
7) **Data mining:** Searching for patterns of interest in a particular representational form.
8) **Interpretation:** Visualization of the extracted patterns.
9) **Use of extracted knowledge:** Using the knowledge directly, incorporating the knowledge into another system for further action, or simply reporting it to interested parties.

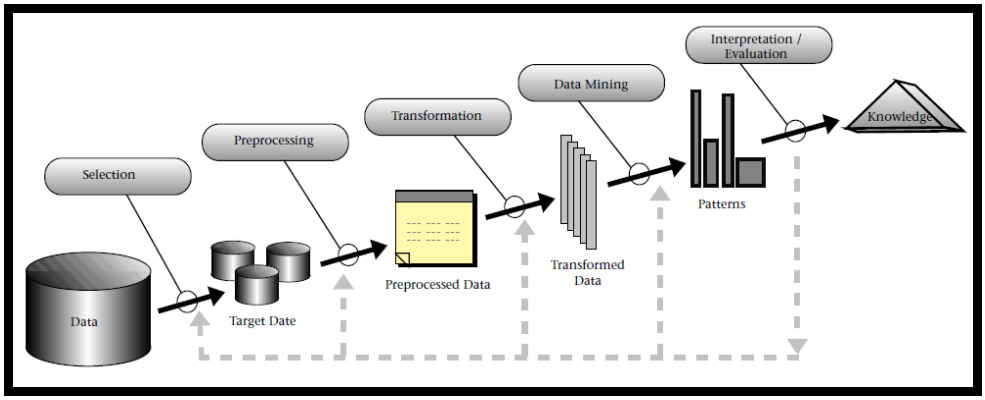


Figure 3.1 An overview of steps that compose the KD process (Fayyad, 1996).

KD has an interdisciplinary nature because it lies on the intersection of research fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence (AI), knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large data sets. But, knowledge discovery from data is fundamentally a statistical research. Because, statistics provide a conventional language and a framework for quantifying the implicit knowledge that we can infer from a particular sample that represents population.

KD refers to the overall process of discovering useful knowledge from data. So, there is a distinction between KD and data mining (DM), because, DM refers to a particular step in this process. Data mining is the application of specific prediction or description algorithms for extracting patterns from data. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing data (Fayyad, 1996). The goals of prediction and description can be achieved using a variety of particular data mining methods as below;

- Classification,
- Regression,
- Cluster Analysis,
- Summarization,
- Dependency modeling,
- Change and deviation detection.

3.2. Cluster Analysis

The knowledge discovery process employed in this study is a modified version of the general steps presented in Fayyad (1996) and follows the approach of Qi & Zhu's (2003) which consists of four major steps; data preparation, data cleaning and preprocessing and finally knowledge examination and interpretation (Figure 3.2).

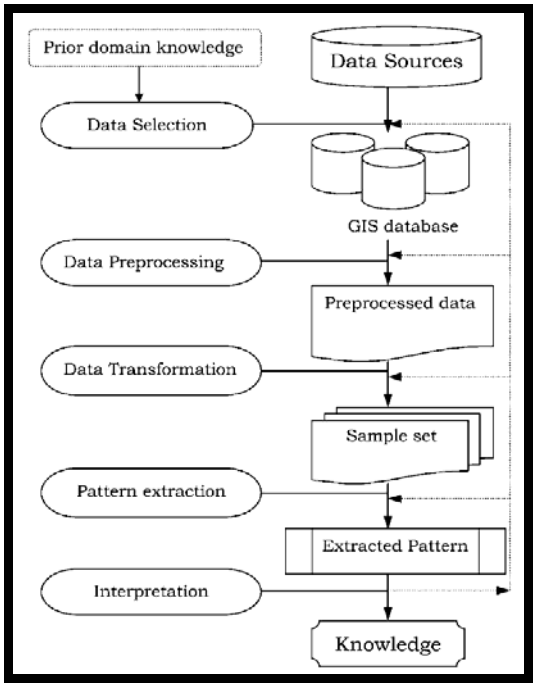


Figure 3.2 The KD process employed adapted from Qi & Zhu (2003).

The central step of KD process is the extraction of patterns from data. To achieve this various data mining algorithms have been developed and clustering algorithms are the most important tasks in data mining and KD literature (Qi & Zhu, 2003).

There are many different definitions about the clustering concept. Aldenderfer & Blashfield (1989) denotes that "…cluster analysis is the generic name of a wide variety of procedures that can be used to create classification" and defines clustering as a multivariate statistical procedure that starts with a dataset containing information about a sample of entities and attempts to reorganize these entities into relatively homogenous groups. Guo (2002), on the other hand, describes clustering

as the task of organizing a set of objects into groups such that the objects in the same group are similar to each other and different from those in the other groups. Jain et al (1999) give a more technical description of clustering that “cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity”. Clustering is also an unsupervised process because there are no predefined classes and no examples that would show what kind of relationships there are in the underlying data or there may be no experts who would declare their opinions about the domain (Delavar, 2007; Hatzichristos, 2004).

Briefly, we can say that the main concern in the clustering process is to reveal the organization of patterns into sensible groups in an unsupervised manner, which allows us to discover similarities and differences, as well as to derive useful inferences about them. This idea has been applied in many areas including astronomy, archeology, medicine, chemistry, education, psychology, linguistics and sociology, but particularly has a greater impact on data mining, document retrieval, image segmentation and pattern classification.

Despite the differences of goals, data types, methods used, Aldenderfer and Blashfield (1989), generalizes the clustering process as;

1) Selection of a sample to be clustered,
2) Definition of a set of variables on which to measure the entities in the sample,
3) Computation of the similarities among entities,
4) Use of cluster analysis method to create groups of similar entities,
5) Validation of the resulting cluster solution.

The variety of techniques for representing data, measuring similarity between data elements and grouping entities has produced a rich and often confusing assortment of clustering methods. The selection of a suitable algorithm for a specific task is determined by several factors, including the nature of data source, appropriate knowledge representation and desired accuracy. So, different algorithms are suitable for different problem configurations and there is no clustering technique that

is universally applicable in uncovering the variety of structures present in multidimensional datasets (Qi et al, 2003). This fact results from the implicit assumptions of clustering algorithms about based on similarity (distance) measures and grouping (linkage rules) criteria.

3.2.1. Similarity Measures

As mentioned earlier, clustering is a process of grouping data items based on a measure of similarity. Similarity measures how alike two cases are. So, in clustering methods use the dissimilarities, in other word distances, between objects. Actually, this is the first rule of Geography as Tobler (1979) denotes that “everything is related to everything else, but nearby things are more related than distant things”. These distances can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects. The most conventional way of computing this distance in a multidimensional space is to compute Euclidean distances, but of course, the selection of the right method can change related to application. These distance measures can be summarized as below (Statsoft, 2008);

- **Euclidean distance:** This is the most common chosen type of distance. It is the geometric distance in multidimensional space.
- **Squared Euclidean distance:** Standard Euclidean distance is squared in order to put greater weight on objects that are further apart.
- **City-Block (Manhattan) distance:** This distance is the average distance across dimensions. Results are usually similar to what Euclidean distance yields.
- **Chebychev distance:** This distance measure may be appropriate in cases when one wants to define two objects as "different" if they are different on any one of the dimensions.
- **Power distance:** This measure increases or decreases the weight that is placed on dimensions on which the respective objects are very different.
- **Percent disagreement:** This measure is particularly useful if the data for the dimensions included in the analysis are categorical in nature.

3.2.2. Grouping Criteria

The similarity measures represent the intra-cluster distance, but grouping criteria stands for inter-cluster distance (Wu & Chow, 2003). In other words, a linkage rule is needed when two clusters are sufficiently similar to be linked together. The linkage rules are usually used in hierarchical clustering algorithms. There are numerous linkage rules and some of them are summarized below (Statsoft, 2008);

- **Single Linkage (nearest neighbor):** In this method, the distance between two clusters is determined by distance of the two closest objects (nearest neighbors) in different clusters.
- **Complete Linkage (furthest neighbor):** In this method, the distances between clusters are determined by the greatest distance between any two objects in different clusters.
- **Unweighted pair-group average:** In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in two different clusters.
- **Weighted pair-group average:** This method is identical to the *unweighted* pair-group average method, except that in the computations, the size of the respective clusters is used as a weight.
- **Unweighted pair-group centroid:** The centroid of a cluster is the average point in the multidimensional space defined by the dimensions. In a sense, it is the center of gravity for the respective cluster. In this method, the distance between two clusters is determined as the difference between centroids.
- **Weighted pair-group centroid:** This method is identical to the previous one, except that weighting is introduced into the computations to take into consideration differences in cluster sizes.
- **Ward's method:** This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the sum of squares of any two clusters that can be formed at each step. In general, this method is regarded as very efficient; however, it tends to create clusters of small size.

3.3. Spatial Knowledge Discovery and Spatial Data Mining

Koperski et al (1996) defines spatial data as objects that occupy space. Nowadays, there is an exponential rise in the size of databases. The phenomenon is more serious in geospatial sciences. Birimcombe (2003) refers to 1990's as a period of transition from data-poverty to data-richness, since as digital spatial datasets have grown rapidly in scope, coverage and volume. He put forwards this state to the developments below;

- Improved technology and wider use of GPS, remote sensing and digital photogrammetry for collecting physical data,
- The introduction of new approaches to obtaining lifestyle and preference data such as through loyalty cards,
- Increased computing power to process raw data coupled with the falling cost of data storage,
- The advent of data warehousing technologies,
- Increasingly efficient ways of accessing and delivering data on-line.

Numerous applications, e.g. geographic information systems and computer aided design (CAD) systems, require the management of spatial data such as points, lines and polygons. The space of interest may either be an abstraction of a real two dimensional or three dimensional spaces, such as a part of the surface of the earth or some high dimensional space of feature vectors. These complex structures are kept in spatial databases that have unique characteristics. Li & Wang (2005) describe a spatial database as a storage where spatial objects are represented by spatial types and spatial relationships among these types. He also mentions the topologic structure of spatial objects which is often recognized by spatial indexing structures and accessed by spatial access methods.

Thus the exponential rises in the size of spatial databases, their increasingly complex structures and the rate at which they can accumulate on even a daily basis are leading to an urgent need for techniques that can mine very large spatial databases for the knowledge they contain. In order to understand and make full use of these data repositories, a few techniques have been tried, e.g. expert system,

database management system, spatial data analysis, machine learning, and artificial intelligence (Li et al, 2005). These and advances in spatial data structures, spatial reasoning and computational geometry paved the way for the study of spatial data mining (Koperski, 1996).

Many excellent studies on data mining have been conducted, however; most of these studies are concerned with knowledge discovery on non-spatial data (Han and Ng, 1994). But, efficient tools for extracting Information from geospatial data are crucial to organizations which make decisions based on large spatial datasets. These organizations are spread across many domains including ecology and environment management, public safety, transportation, public health, business, travel and tourism (Chawla et al, 2001).

According to Lİ et al (2005) spatial data mining and knowledge discovery (SDMKD) refers to the efficient extraction of hidden, implicit, interesting, previously unknown, potentially useful, ultimately understandable, spatial or non-spatial knowledge (rules, regularities, patterns, constraints) from incomplete, noisy, fuzzy, random and practical data in large spatial databases. Spatial data mining aims to automate such a knowledge discovery process. Thus, it plays an important role in;

- Extracting interesting spatial patterns and features,
- Capturing intrinsic relationships between spatial and non-spatial data;
- Presenting data regularity concisely and at higher conceptual levels;
- Helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance (Han et al, 1994).

In the presence of spatial data, the standard approach in the mining community is to materialize the spatial attributes and rebuild the model with these new spatial attributes (Chawla et al, 2001). But, the distinct features of a spatial database pose challenges and bring opportunities for mining information from spatial data. The difference between classical and spatial data mining parallels the difference between classical and spatial statistics (Chawla et al, 2001). Spatial statistics aims to model the special properties of spatial data with classic statistics. In brief, the major challenges towards spatial data mining result from the huge amount of spatial

data, the complexity of spatial data types and spatial accessing methods. In order to understand these, we have to look at the special nature of spatial data (Bacao, et al, 2005-a; Anselin, 1990);

- Firstly, there is a necessity of good efficiency on large spatial databases, because databases contain more than just a few thousand objects (Ester et al, 1996). Contrary to data used in statistics, spatial datasets are very large, since, they are large in terms of records (n) but also in terms of variables (p). But, most data processing algorithms have memory and time processing requirements that grow more than proportionally with the number of instances (n) (Bacao et al, 2005-a). Many analysis techniques scale somewhere between $O(n^3)$ and $O(n(\log(n)))$ in terms of computational complexity, with the majority falling somewhere around $O(n^2)$ (Buttenfield et al, 2000). But this complexity can be tremendous for spatial databases for that they not only have attribute information but also spatial features and the topological relationship between them.
- Secondly, classic statistics and classic data mining algorithms often make assumptions which violate Tobler's first law of geography (1979). In fact, spatial data is characterized by the existence of spatial dependency which is directly taken into account spatial statistical techniques under the heading of spatial autocorrelation. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife and temperature vary gradually over space.
- Thridly, spatial data is heterogenous which results from the unique nature of each place. Thus, spatial data very rarely presents stationary characteristics. This means that the functional forms and parameters may vary and not homogenous in different areas of the map. It is also related to isotropy which assumes that pattern is similar in all directions. So, most spatial processes are nonstationary and anistropic (Bailey and Gattrell, 1995).
- Fourth, spatial data are complex. Spatial dimension means each item of data has a spatial reference where each entity occurs on the continuous surface, or where the spatial referenced relationship exists between two neighbor entities. Spatial data includes not only positional data and attribute

data, but also spatial relationships among spatial entities. Moreover, spatial data structure is more complex than the tables in ordinary relational database. Besides tabular data, there are vector and raster graphic data in a spatial database. And the features of graphic data are not explicitly stored in the database (Li et al, 2005).

Azimi & Delavar (2007) divide spatial data mining techniques into four general groups as described below;

- **Spatial Association Rules:** Spatial association rules use spatial and non-spatial predicates in order to describe spatial objects using relations with other objects.
- **Spatial Clustering:** Clustering is the task of grouping the objects of a database into meaningful subclasses so that the members of subclasses are as similar as possible whereas the members of different clusters differ as much as possible from each other.
- **Spatial Trend Detection:** Spatial trend is defined as a regular change of one or more non-spatial attributes when moving away from a given object.
- **Spatial Classification:** The task of classification is to assign an object to a class from a given set of classes based on the attribute values of the object.

Spatial data mining is a demanding field since huge amounts of spatial data have been collected in various applications such as real-estate marketing, traffic accident analysis, environmental assessment, disaster management and crime analysis. A spatial dataset consists of a set of cases and each case has a spatial location and a set of variables. Such a data matrix can be decomposed into two parts: The attribute space X and the geographic space S (consisting of spatial locations), which are shown in Figure 3.3 with two rectangles. The number of cases is referred to as the dataset size (n) and the number of variables is referred to as the dataset dimensionality (d). When we say a dataset is large, it means that dataset has a large number of cases. When we say a dataset is high dimensional, it means that dataset has a large number of variables (Guo et al, 2005).

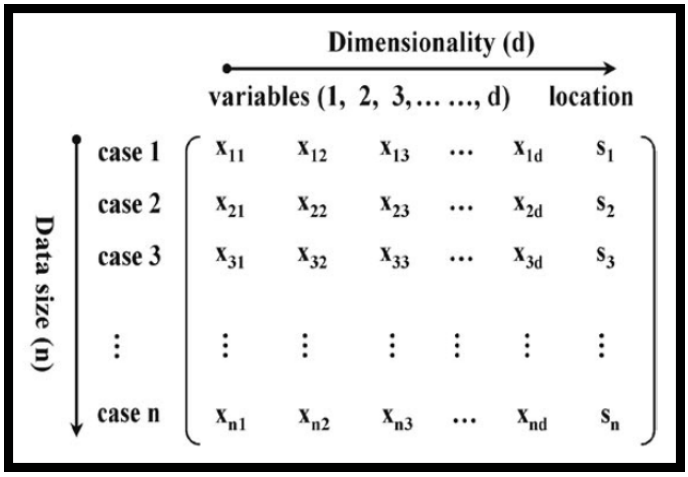


Figure 3.3 A spatial dataset matrix (Guo et al, 2005)

Thus, there is a need to discover knowledge from these spatial databases, but, because of the lack of primary knowledge, in other words domain knowledge, about the data, spatial clustering is one of the most valuable methods in spatial data mining, which is also our focus of view in this thesis.

3.3.1. Spatial Clustering

“The spatial cluster detection lies in the heart of spatial data mining” (Birimcombe, 2003). Spatial cluster analysis seeks to form a segmentation into regions which minimize within-cluster variation but maximize between-cluster variation by means of spatial statistical methods adapted from classic statistics in order to handle special nature of spatial data Spatial clustering can be based on combinations of non-spatial attributes, spatial attributes, and proximity of the objects or events in space, time, and space–time depending on different applications (Han & Miller, 2009; Jiang, 2004).

In order to define a spatial cluster we first must consider the kinds of data that are being studied. The information to be clustered may be event-based, population-based, field-based, or feature-based as described below;

- **Event-based data** include point locations (such as the places of residence and cases of disease in people, or the locations of a species of tree in a forest) and counts (accidents at particular road intersections).
- **Population-based data** incorporate information on the population from which the events arose, and include disease rates with case counts in the numerator and size of the at-risk population in the denominator. Gahegan (2005) explaines this kind of data more detailly. He denotes that, in this kid of datasets, it is not sufficient to simply find where the individual cases of the disease are clustered, but to find regions that have high rates of the disease. This requires that the disease rate be compared with the background population that is susceptible to the disease, so that regions with higher-than expected disease rates can be found .
- **Field-based data** are observations that are continuously distributed over space, and include concentrations and temperatures.
- **Feature-based data** include boundaries and polygons that may be derived from field-based data, such as zones of rapid change in an attribute's value.

In the light of this classification Jacquez (2008) defines a spatial cluster as an excess of events (for event- and population based data, such as a cancer cluster) or of values (for field-based data, such as a grouping of excessively high concentrations of cadmium in soils) in geographic space. For feature-based data, a cluster might be a spatial aggregation of boundaries.

Therefore, there are various forms of geographic data such as point, line and polygon which can be used in spatial clustering. Among them, the most popular form is point data. The location of a spatial object or an event is represented as a point by means of its coordinates. Examples of point data include the locations of buildings, the centers of roads and lands and the locations of outbreaks in epidemics or crime events. On the other hand, Han et al (2009-b) denotes that recent improvements in satellites and tracking facilities have made it possible to collect a large amount of trajectory data of moving objects. Examples include vehicle position data, hurricane track data and animal movement data. The first studies of spatial clustering method are developed for points, maybe, because of their simplicity as a feature.

Spatial cluster analysis plays an important role in quantifying geographic variation patterns. It is commonly used in disease surveillance, spatial epidemiology, population genetics, landscape ecology, crime analysis and many other fields. Some classical applications of spatial data clustering in fields of crime analysis and epidemiology are summarized at below;

- **John Snow's Analysis**: In 1850's, Dr. John Snow performed spatial data analysis to prove his hypothesis that cholera was spread by the drinking of water coming from a water pump in Broad Street affected with cholera bacteria in London. Because, the cholera cases were forming a tight cluster around the water pump when he plotted them on the map with paper and pen (Bailey and Gatrell, 1995).
- **Crime Hot-Spot Analysis:** This analysis helps police to identify high crime areas, types of crimes being committed and the best way to respond. In many cases crime hotspot is defined as an area where crime incidents are geographically concentrated (Levine, 2009-a).

These examples were some famous applications of spatial clustering in various areas. But from a more methodological point of view, the spatial clustering algorithms vary among themselves when we try to categorize them. It is difficult to provide a rigid categorization of clustering methods because these categories may overlap, so that a method may have features from several categories. These methods are broadly named as 'hybrid methods' (Han et al, 2009-b). In Jain et al (1999), taxonomy of spatial clustering methods is described based on their contrary properties as summarized below;

- **Agglomerative vs. divisive:** This aspect changes up to the algorithmic structure and operation. An agglomerative approach begins with each pattern in a distinct cluster, and successively merges clusters together until a stopping criterion is satisfied. A divisive method begins with all patterns in a single cluster and performs splitting until a stopping criterion is met.
- **Monothetic vs. polythetic:** This aspect relates to the ordered or simultaneous use of features in the clustering process. Most algorithms are polythetic; that is, all features enter into the computation of distances

between patterns, and decisions are based on those distances. Each of these clusters is further divided independently.

- **Hard vs. fuzzy:** A hard clustering algorithm allocates each pattern to a single cluster during its operation and in its output. A fuzzy clustering method assigns degrees of membership in several clusters to each input pattern. A fuzzy clustering can be converted to a hard clustering by assigning each pattern to the cluster with the largest measure of membership.
- **Deterministic vs. stochastic:** This issue is most relevant to partitional approaches designed to optimize a squared error function. This optimization can be accomplished using traditional techniques or through a random search of the state space consisting of all possible labelings.
- **Incremental vs. non-incremental:** This issue arises when the pattern set to be clustered is large, and constraints on execution time or memory space affect the architecture of the algorithm. Earlier, clustering methodology was not containing many examples of clustering algorithms designed to work with large data sets, but the advent of data mining has fostered the development of clustering algorithms that minimize the number of scans through the pattern set, reduce the number of patterns examined during execution, or reduce the size of data structures used in the algorithm's operations.

In this study, spatial clustering methods are explored in eight groups in a more detailed fashion. These are;

- Clustering techniques in exploratory spatial data analysis (ESDA),
- Partitioning methods,
- Hierarchical methods,
- Density-based methods,
- Grid-based methods,
- Fuzzy clustering,
- Artificial neural networks for spatial clustering, and
- Methaheuristical clustering methods.

3.3.2. Clustering techniques in Exploratory Spatial Data Analysis (ESDA)

ESDA is built on a solid statistical background. Its main aim is to identify data properties for purpose of pattern detection in data, hypothesis formulation from data and some aspects of model assessment such as goodness of fit and identification of data effects on model fit. It is based on the use of graphical and visual methods and the use of numerical techniques that are statistically robust (Bailey & Gatrell, 1995). Clustering plays an important role in ESDA, because it involves the identification and description of spatial patterns such as outliers, clusters, hotspots, coldspots, trends and boundaries. It has two primary objectives;

- **Objective 1:** Pattern recognition using visualization, spatial statistics and geostatistics to identify the locations, magnitudes and shapes of statistically significant pattern descriptors.
- **Objective 2:** Hypothesis generation to specify realistic and testable explanations for the geographic patterns found under Objective 1 against the null hypothesis which is called “complete spatial randomness” (CSR). CSR assumes that , there is no clustering in the spatial data and it is completely random (Jacquez, 2008).

There are many cluster statistics, but Jacquez (2008) generalizes them under three headings as global, local and focused;

- **Global cluster statistics** are sensitive to spatial clustering or departures from the null hypothesis, that occur anywhere in the study area. Many early tests for spatial pattern, such as Moran’s I were global in nature, and provided one statistic, such as a global autocorrelation coefficient, that summarized spatial pattern over the entire study area. While global statistics can identify whether spatial structure, such as clustering, autocorrelation, uniformity, exists, they do not identify where the clusters are, nor do they quantify how spatial dependency varies from one place to another. But there are some visualization techniques.
- **Local statistics** such as Local Indicators of Spatial Autocorrelation (LISA), G_i^* statistics and geographically weighted regression quantify spatial

autocorrelation and clustering within the small areas that together comprise the study area. Many local statistics have global counterparts that often are calculated as functions of local statistics. For example, Moran's global autocorrelation statistic is the scaled sum of the LISA statistics.

- **Focused statistics** quantify clustering around a specific location called a focus. These tests are particularly useful for applications like exploring possible clusters of disease near potential sources of environmental pollutants.

Some famous examples of spatial clustering algorithms based on ESDA are Geographical Analysis Machine (GAM), Spatial Scan Statistic, Besag-Newell's method and Fortheringham-Zhan's method (Conley et al, 2005).

3.4. Partitioning Methods

Partitioning algorithms construct a partition of a database D of n objects into a set of k clusters, so, organizes the objects into k partitions (k<=n), where each partition represents a cluster. The algorithm satisfies two requirements;

1) Each group must contain at least one object, and

2) Each object must belong to exactly one group.

Thus, the partitioning algorithm typically starts with an initial partition of D and then uses an iterative relocation technique, in other words optimization function that attempts to improve the partitioning by moving objects from one group to another. The type of optimization functions are usually proximity or dissimilarity measures which simplifies to assign a data item to the 'closest' cluster. Examples of partitioning algorithms are described below;

- **K-means:** K-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the intra-cluster similarity is high, but the inter-cluster similarity is low. Cluster similarity is measured with respect to the sum of squared distances of objects to the mean value of a cluster, which is the centroid or center of gravity of a cluster. This criterion tries to make the resulting k clusters as compact and as separate as possible.

- **K-medoids:** The k-means algorithm is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data. This sensitivity is particularly diminished by taking the most centrally located object within a cluster, instead of taking the mean value of the objects. The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities by using an absolute-error criterion. PAM (partitioning around medoids) was one of the first k-medoids algorithms introduced.
- **Expectation-Maximization (EM):** The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given dataset when the data is incomplete. It can be viewed as an extension to the k-means, but in EM, instead of assigning each object to a cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. In other words, there are no strict boundaries between clusters. In order to take spatial data into account, neighborhood EM (NEM) is proposed which penalizes the objects which are not neighbors to each other.
- **(Clustering Large Applications Based on Randomized Search) CLARANS:** A typical k-medoids like PAM works efficiently for small datasets. To deal with larger datasets, a sampling based method, called CLARA (Clustering LARge Applications) is proposed. Instead of finding representative objects for the entire data set, CLARA draws a sample of the data set, applies PAM on the sample, and finds the medoids of the sample. The point is that, if the sample is drawn in a sufficiently random way, the medoids of the sample would approximate the medoids of the entire data set. To come up with better approximations, CLARA draws multiple samples and gives the best clustering as the output. Here, for accuracy, the quality of a clustering is measured based on the average dissimilarity of all objects in the entire dataset, not only for those objects in the samples. However, CLARA cannot find the best clustering if any of the best k-medoids are not selected during sampling. To improve the quality and scalability of CLARA, another algorithm called CLARANS was proposed which combines the sampling technique with PAM. Whereas, CLARA draws a sample of nodes at

the beginning of the search from entire dataset, CLARANS dynamically draws a random sample of neighbors in each step of a search (Han et al, 2009-b). CLARANS has been experimentally shown to be much more efficient than PAM. In addition, CLARANS is able to find clusterings of better quality than CLARA (Ester et al, 1997).

3.5. Hierarchical Methods

Hierarchical methods create a hierarchical decomposition of dataset D by grouping data objects n into a tree of clusters. Hierarchical algorithms can be further divided in two subclasses;

1) Agglomerative: Bottom-up (Merging),

2) Divisive: Top-Down (Splitting).

Hierarchical decomposition is presented by a dendogram, a tree that iteratively splits D into smaller subsets consists of only one object. In such a hierarchy, each node of the tree represents a cluster of D. A dendogram can either be created from the leaves up to the root (agglomerative approach) or from the root down to the leaves (divisive approach) (Figure 3.4). In contrast to partitioning algorithms, hierarchical algorithms do not need k as an input. However, a termination condition has to be defined indicating when the merge or division process should be terminated. One example of a termination condition in the agglomerative approach is the minimum distance between all clusters (Ester et al, 1997).

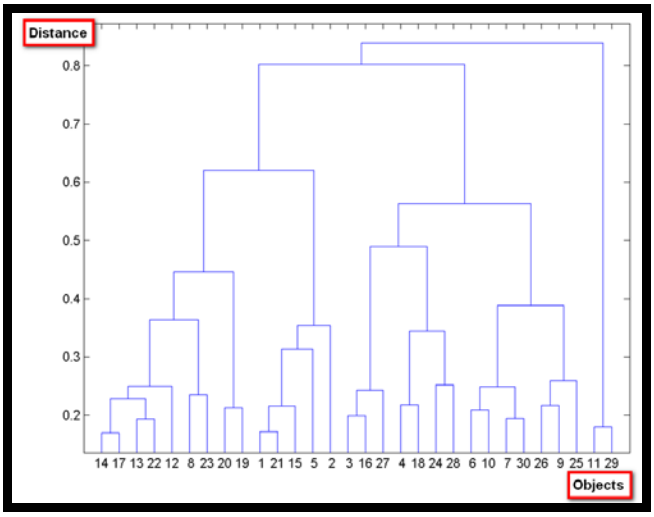


Figure 3.4 A sample dendogram (Web 6).

The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. That is, if we decide that merge or split gave the right results, the method cannot go back and correct it. This results from the grouping rules like single linkage or complete linkage which is widely used in hierarchical algorithms and explained at section 3.2.2. But in recent studies in literature, hierarchical methods are used with partitioning methods hybridly, thus hierarchical methods gain the capability of iterative relocation of partitioning algorithms. Examples of hierarchical methods used in spatial clustering are summarized below;

- **Agglomerative and Divisive Hierarchical Clustering:** AGNES (AGlometarive NESting) and DIANA (DIvisive ANAlysis) are two earlier hierarchical clustering algorithms. AGNES is an agglomerative (bottom-up) algorithm which starts by placing each object in its own cluster and then merging these atomic clusters into larger clusters until all of the objects are in one cluster or until a certain termination condition is satisfied. On the other hand, DIANA is a divisive (to-down) algorithm that does the reverse of AGNES by starting with all objects in one cluster (Figure 3.5). It subdivides the cluster into smaller and smaller pieces until a certain termination condition is satisfied.

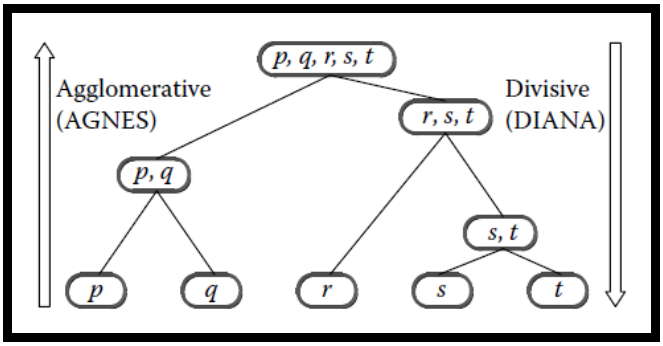


Figure 3.5 Agglomerative and divisive hierarchical clustering on a set of data objects {p,q,r,s,t} (Han et al, 2009-b).

In either AGNES or DIANA, one can specify a desired number of clusters as a termination condition. The drawback of these two algorithms is the selection of merge or split points. Such a decision is critical because once a group of objects are merged or split, the process continues to iterate on the newly generated clusters. It cannot undo what was done previously or cannot perform object swapping between clusters. One of the solutions to this problem which would improve the clustering quality is to integrate hierarchical methods with other clustering techniques. Two such methods are introduced in the following subsections.

- **BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies):** BIRCH is designed for clustering large amount of numerical data by integration of two steps:
  1) Hierarchical clustering (at the initial micro-clustering stage)
  2) Other clustering algorithms such as iterative partitioning (at the later macro-clustering stage).

  It overcomes the two difficulties of agglomerative clustering methods: (1) scalability and (2) the inability to undo what was done in the previous step. BIRCH introduces two concepts, clustering feature (CF) and clustering feature tree (CF tree), which are used to summarize cluster representations (Figure 3.6). These structures help the clustering method achieve a good speed and scalability in large databases (Zhang, Rmakrishnan and Livny, 1996).

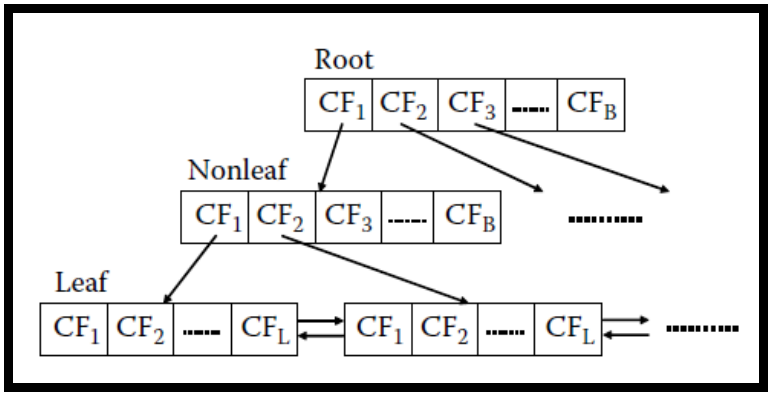


Figure 3.6 A CF tree structure (Han et al, 2009-b).

However BIRCH also has some disadvantages. A CF tree node can only hold a limited number of entries due to its size, so the resulting cluster is not always considered as a natural cluster. Moreover, if the clusters are not spherical in shape, BIRCH does not perform well because it uses the notion of radius or diameter to control the boundary of a cluster (Han et al, 2009-b).

- **CURE (Clustering Using Representatives):** Cure starts with partitioning the dataset into p partitions then by random sampling it eliminates the outliers in order to get unbiased clusters. Then, it randomly selects well scattered points from the clusters and shrinks them towards the center of cluster by a specified fraction (Figure 3.7). Unlike BIRCH, the algorithm adjusts well to arbitrary shaped clusters and avoids the single-link effect of pure hierarchical method. It also tries to speed up the processing time by eliminating the clusters which grows too slow.

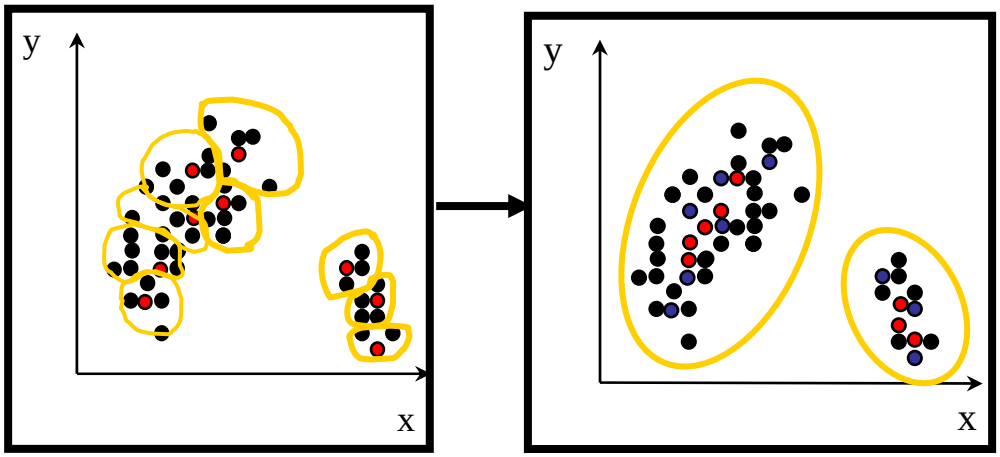


Figure 3.7 Shrinking the representatives (Han & Kamber, 2000).

- **Chameleon – A Hierarchical Clustering Algorithm Using Dynamic Modeling:** Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between two clusters. It tries to overcome the weakness of agglomerative algorithms, because they ignore the information about the interconnectivity of objects in two separate clusters and also the other set of schemes ignores the information about the closeness of two clusters as defined by the similarity of the closest objects across two clusters. Chameleon determines the pair of most similar sub-clusters by taking into accounts both the interconnectivity and closeness of the clusters. Chameleon models the data using a *k*-nearest neighbor graph, where each vertex of the graph represents a data object, and there exists an edge between two vertices (objects) if one object is among the *k*-most similar objects of the other. The edges are weighted to reflect the similarity between objects. Chameleon uses an algorithm that consists of two distinct phases. In the first phase, it uses a graph partitioning algorithm to partition the *k*-nearest neighbor graph into a large number of relatively small sub clusters. In the second phase, it uses an agglomerative hierarchical clustering algorithm that repeatedly merges subclusters based on their similarity (Figure 3.8) (Han et al, 1999). Chameleon shows greater power at discovering arbitrarily shaped clusters than several weel known algorithms such as BIRCH and density-

based algorithm DBSCAN. However, the processing cost of high-dimensional data is worse than others (Han et al, 2009-b).

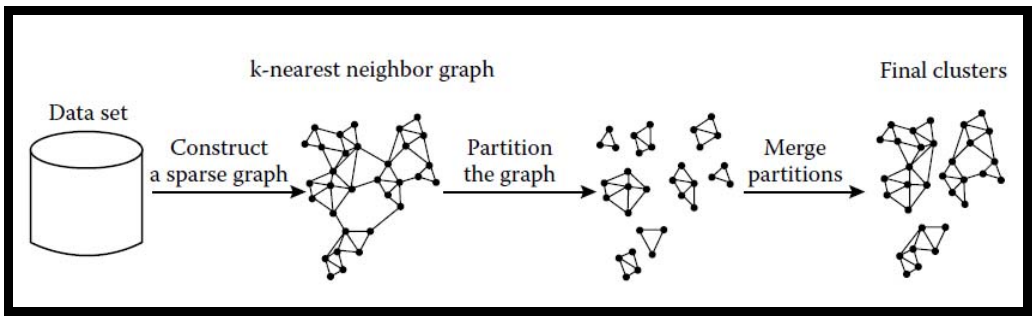


Figure 3.8 Chameleon: Hierarchical custering based on k-nearest neighbors and dynamic modeling (Han et al, 1999).

3.6. Density-based Methods

Most partitioning methods cluster objects based on the distance between objects and most hierarchical algorithms cluster objects according to grouping criteria or both. But such methods can only find spherical-shaped clusters and it is difficult for them to find clusters of arbitrary shape. Also they are weak at handling noise. Density-based clustering methods have been developed on the notion of density. Their general idea is to continue growing a given cluster as long as the density (the number of objects or data points) in the 'neighborhood' exceeds a threshold. Moreover, they take into account dense clusters in the data space that are separated by regions of low density which represents noise. They are also faster than other algorithms because they scan the database one time (Han et al, 2009-a, Han et al, 2000). Representative algorithms are explained below;

- **DBSCAN – A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density**: The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with outlier objects. It defines a cluster as a maximal set of density-connected points. DBSCAN searches for clusters by checking the ε-neighborhood of each point in the database. This process

starts with an arbitrary point p. If the ε-neighborhood of a point p contains more than minimum number of objects (MinPts), a new cluster with p as a core object is created. DBSCAN then iteratively collects directly density reachable objects from these core objects. This process terminates when no new point can be added to any cluster (Figure 3.9) (Ester et al, 1996).

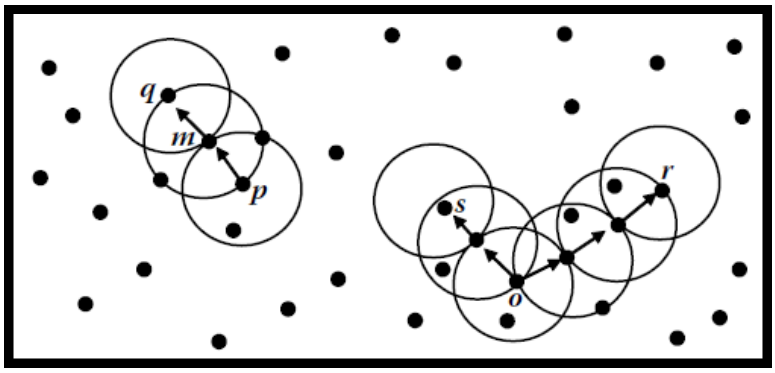


Figure 3.9 Density accessibility and density connectivity in density-based clustering (Ester et al, 1996)

The algorithm is effective at finding arbitrary shaped clusters. However, the user is responsible of selecting parameter values such as ε-neighborhood or MinPts. But, parameter setting happens to be in real-world examples or high-dimensional datasets. GDBSCAN is a modified version of DBSCAN. It is developed because databases may also contain extended objects such as polygons other than points. In GDBSCAN, neighborhood notion is indicated by the assumption that two polygons have non-empty intersection. Also, non-spatial attributes such as the average income of a city can be used to define cardinality of the neighborhood. The equivalent of the parameters of ε-neighborhood and MinPts are NPred and MinWeight (Han et al, 2009-b).

- **OPTICS – Ordering Points to identify the Clustering Structure:** DBSCAN fails to find the optimal clustering when the data space has both dense and sparse regions. It is favorable in local densities rather than global ones. However, when the scale extends, the parameters equivalent to find clusters also increases. To overcome this difficulty a cluster analysis called OPTICS

was proposed by Ankerst et al (1999). This algorithm extends DBSCAN with a set of distance parameters in order to create orders of clusters in large areas. It extends the core distance p of DBSCAN cluster with a reachability distance (q) (Figure 3.10).

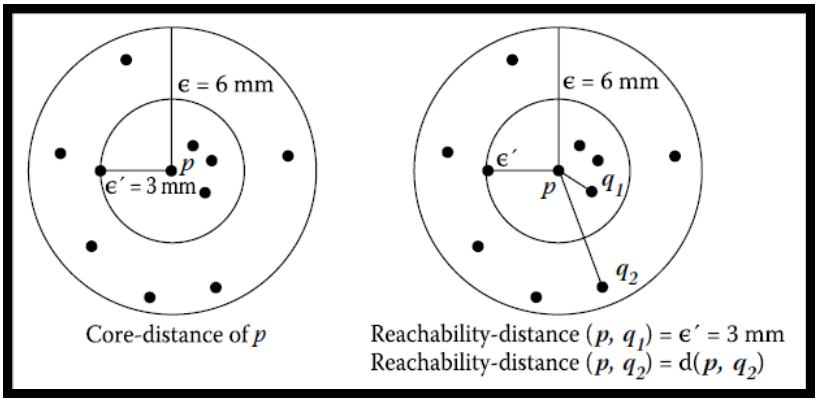


Figure 3.10 Parameter distances in OPTICS (Han et al, 2009-b).

- **DENCLUE – Clustering Based on Density Distribution Functions:** DENCLUE was proposed by Hinneburg and Keim (1998). This algorithm has a solid mathematical background based on a set of distribution functions. It is especially good at handling datasets with large amount of outliers and allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional datasets. Also, it is significantly faster existing algorithms. But, it needs large number of parameters which is not preferable by novice user (Han et al, 2000). The steps of the method is built on the following ideas;
    1) It uses a grid map that is laid over the objects. Then, it keeps information about the grid cells that actually contain objects and manages these cells in a tree-based structure,
    2) It models the influence of each object using a mathematical function called influence function, which describes the impact of object within its neighborhood,
    3) The overall density of the data space is calculated as the sum of the influence function applied to all objects,

4) Finally, clusters can be determined mathematically by identifying density attractors, where density attractors are local maximum of the overall density function (Hinneburg et al, 1998).

A sample application applied to an arbitrary empirical shape with different thresholds can be seen in Figure 3.11 below;

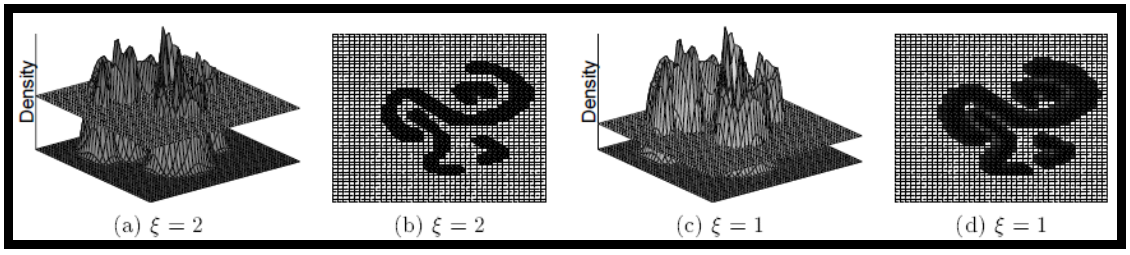


Figure 3.11 Sample applications with different thresholds (Hinneburg et al, 1998).

3.7. Grid-Based Methods

Grid-based methods divide the data space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the multi-resolution grid data structure for example raster data or simply images. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension. Representative studies are explained below (Han et al, 2009-b);

- **STING (STatistical INformation Grid approach):** STING starts with dividing the spatial area into rectangular cells, corresponding to different levels of resolution. As can be seen in Figure 3.12, each cell at a high level is partitioned into a number of smaller cells in the next lower level.

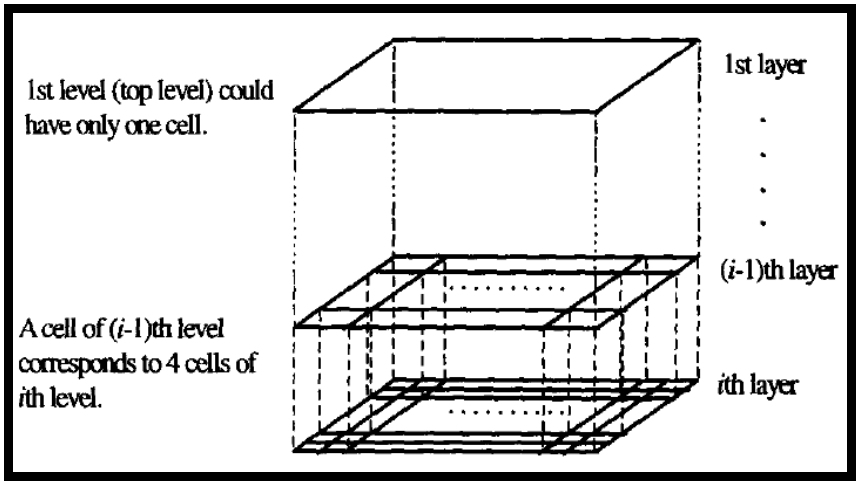


Figure 3.12 Hierarchical partitioning of cells in STING (Wang, Yang and Muntz, 1997)

The statistical information of each cell is calculated and stored previously and is used to answer queries. Parameters such as count, mean, standard deviation can be easily calculated from parameters of a lower level cell. Thus, it uses a to-down approach to answer data queries. Its main advantage is its speed compared to other spatial clustering algorithms. Its disadvantage is that all cluster boundaries are either horizontal or vertical and no diagonal boundary can be detected (Wang et al, 1997).

- **WaveCluster:** This algorithm is based on the notion of wavelet transform. A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band (Figure 3.13).

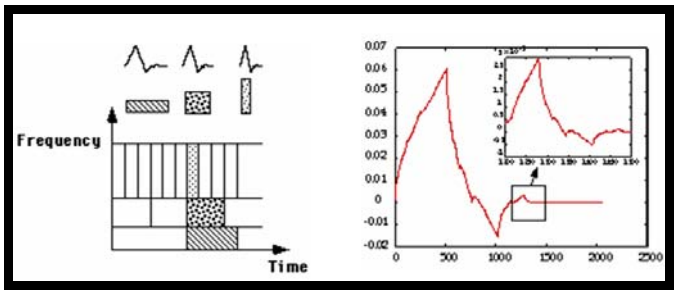


Figure 3.13 Wavelet transformation of different features (Sheikholeslami, et al, 1998).

The steps of algorithm can be summarized as below;

1) It summarizes the data by imposing a multidimensional grid structure onto data space,
2) The multidimensional spatial data objects are represented in a n-dimensional feature space,
3) Algorithm applies wavelet transform on feature space to find the dense regions in the feature space,
4) The wavelet transform is applied iteratively which result in clusters at different scales from fine to coarse.

The major advantages of wave cluster are that it is completely unsupervised, effective at removal of outliers and applicable to multi-resolution grids. But, it is complex to handle, detects arbitrary shaped clusters at different scales and is not sensitive, is not sensitive to noise and only applicable to low-dimensional data (Sheikholeslami et al, 1998). These advantages and disadvantages can be seen in Figure 3.14 when different wavelet frequency is applied to an empirical data.

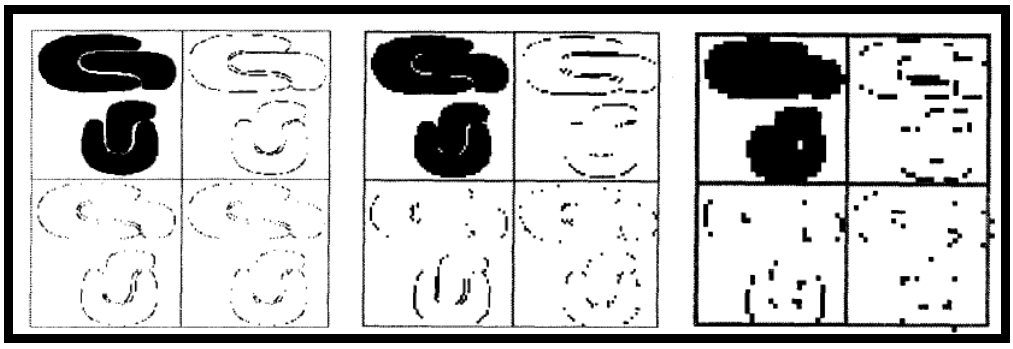


Figure 3.14 Multi-resolution wavelet representations with different frequencies (Sheikholeslami, 1998).

- **CLIQUE (Clustering In QUEst):** CLIQUE aims to automatically identify subspaces of a high-dimensional data space that allows better clustering than original space. CLIQUE can be considered as both density-based and grid-based, because it starts with partitioning each dimension a same number of equal length intervals, then if the total data points contained in

preliminary clusters exceeds the input parameter, these clusters are connected within a subspace . It can automatically find subspaces of high-dimensionality, but it fails at getting the accuracy of the clustering result which may be degraded because of the simplicity of the method (Agrawal et al, 1998).

3.8. Fuzzy Clustering Methods

Traditional clustering approaches generate clusters and in a group, each object belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering extends this notion to link each object with every cluster using a membership function. The outputs of such algorithms are clusters, but not disjoint groups. In fuzzy clustering, each cluster is a fuzzy set of all the objects. For example, in Figure 3.15, the rectangles enclose two “hard” clusters in the data: H_1={1,2,3,4,5} and H_2={6,7,8,9}. But, a fuzzy clustering might produce two fuzzy clusters F_1 and F_2 depicted by ellipses in the figure. The objects in these fuzzy clusters will then have membership values in [1,0] for each cluster. For example, object numbered 1 would have an ordered pair as (1,0.9) for cluster F_1 in which “1” stands for the number of the object and “0.9” stands for its membership value to the cluster F_1. The most popular fuzzy clustering algorithm is the fuzzy c-means (FCM) algorithm. FCM is better than the k-means algorithm at avoiding local minimum, but if one uses squared error criterion as in k-means, FCM can still converge to local minimum. But, there are different membership functions to avoid this problem such as those based on similarity decomposition and centroids of clusters (Jain et al, 1999).

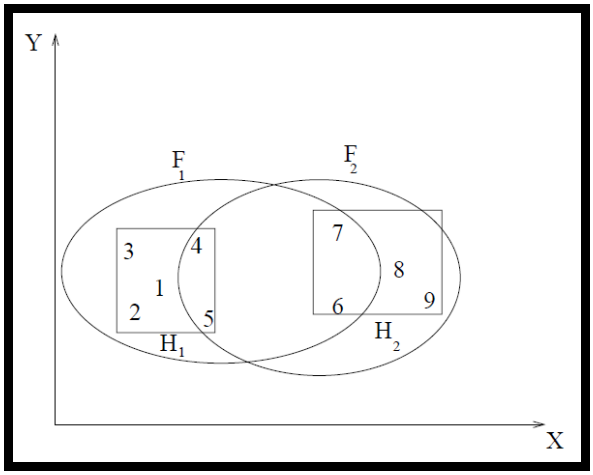


Figure 3.15 Hard vs. Fuzzy clusters (Jain et al, 1999).

3.9. Artificial Neural Networks (ANNs) for Spatial Clustering

An artificial neural network (ANN), usually called "neural network" (NN), is a mathematical model or computational model that tries to simulate the structure and functional aspects of biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase (Figure 3.16).

In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data (Wikipedia, 2009-a).

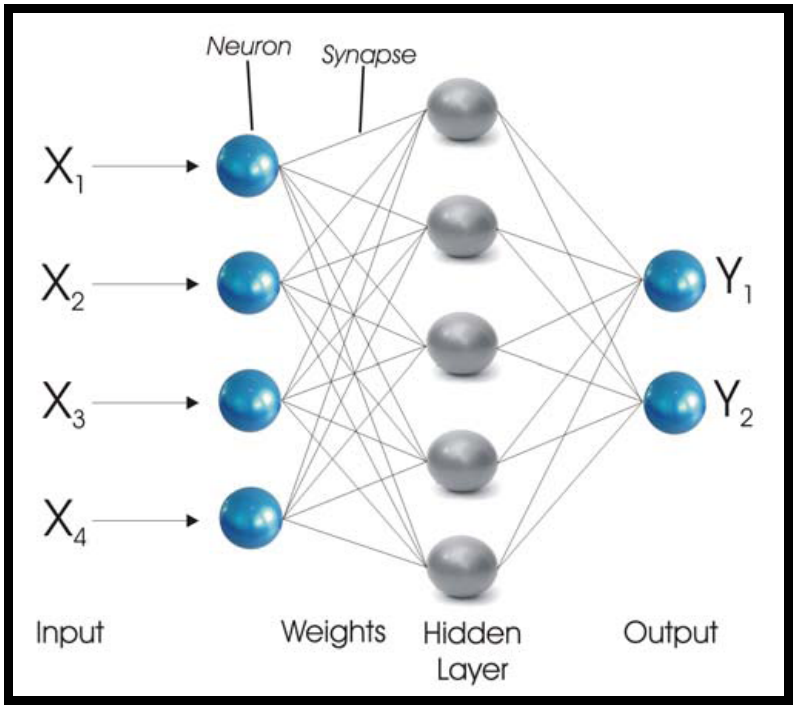


Figure 3.16 An Artificial neural network (Coleman, 2008).

ANNs have been used extensively over the past three decades. Competitive or in other words winner-take-all neural networks are often used to cluster input data (Jain & Mao, 1996). In competitive learning, similar patterns are grouped by the network and represented by a single unit called neuron. This grouping is done based on data correlations. Well known examples of ANNs used for clustering include Kohonen's learning vector quantization (LVQ) and self-organizing map (SOM). The architectures of ANNs are quite simple. The weights between the input nodes and output nodes are iteratively changed. This process is called learning process and continues until a termination criterion is met. The learning procedures of ANNs are quite similar to those in some classical clustering approaches. For example, the relationship between the k-means algorithm and SOM is explained in Bacao et al (2005-b).

3.10. Metaheuristical Methods for Spatial Clustering

One of the most widely used automated metaheuristics method of spatial clustering is called “simulated annealing (SA)”. The simulated SA is a sequential search technique. Simulated annealing procedures are designed to avoid (or recover from) solutions which correspond to local optima of the objective functions. This is accomplished by accepting with some probability a new solution for the next iteration of lower quality (as measured by the criterion function). This probability measure is controlled by a critical parameter called the temperature, which is typically specified in terms of a starting (first iteration) and final temperature value (Jain et al, 1999). “Simulated Annealing exploits an analogy between the way in which molten metal freezes into a minimum energy crystalline structure (the annealing process) and the search for a function optimum” (Altman et al, 2005). At each iteration, simulated annealing randomly generates a candidate point (or set of points) within a local neighborhood of the current solution. The probability of moving from the current solution to one of the candidate points is a function of both the difference in the value of the objective function at each point, and a temperature parameter. At high temperatures, candidate points that are “worse" than the current solution can be selected as the solution in the next iterate. This helps the heuristic to avoid a local optimum. In each iteration, the temperature is reduced gradually, so that the probability of getting the optimal solution becomes vanishingly small which makes simulated annealing computationally slow.

Another metaheuristical method for clustering is the evolutionary approaches. An evolutionary algorithm (EA) can generally be regarded as a computer program that simulates evolutionary processes in which the Darwinian notion of natural selection is maintained. An EA typically starts by randomly initializing a population of individuals to obtain the globally optimal partition of the data (Jain et al, 1999). Each individual is a candidate solution and are encoded as chromosomes. The most commonly used evolutionary operators are: selection, recombination and mutation. Each step transforms one or more input chromosomes into one or more output chromosomes. A fitness function evaluated on a chromosome determines a

chromosome's likelihood of surviving into the next generation. A typical evolutionary algorithm for spatial clustering can be explained in Xiao (2008) with the below steps:

1) Firstly, random populations of individuals are selected. Each solution here corresponds to a valid cluster of the data. Then, a fitness value is associated with each individual which shows how appropriate an individual is to be a cluster.

2) Secondly, evolutionary operators are used such as selection, recombination and mutation in order to generate next population of solutions. Again, the fitness values of newly generated individuals are calculated. The individulas with a small error survives to the next generation.

3) Then, step 2 iteratively repeats until some termination condition is satisfied.

The best known evolutionary techniques are Genetic Algorithms (GAs), evolution strategies (ESs) and evolutionary programming (EP). Out of these three approaches, GAs are most frequently used in clustering.

3.11. Geographical Information Systems (GIS), Spatial Data Analysis and Spatial Data Mining

As a result of developments in technology and %80 of existent data being spatially referenced (Li et al, 2005), "Geographical Information System (GIS)" has emerged as powerful tool which has potential to organize complex spatial environment with tabular relationships. The emphasis is on developing digital spatial database, using the data sets derived from precise navigation and imaging satellites, aircrafts, digitization of maps and relational databases. There are many different definitions of GIS in literature. A couple of them are cited below;

> Burrough (1986) defined GIS as *"…set of tools for collecting, storing, retrieving at will, transforming, and displaying spatial data from the real world for a particular set of purposes".*
>
> Aronoff (1991) defines GIS as *"a computer based system that provides the following four sets of capabilities to handle georeferenced data: 1) input; 2)*

data management (data storage and retrieval); 3) manipulation and analysis; 4) output".

A more detailed definition of GIS is noted by Chrisman (1996). He indicates that GIS is the *"…organized activity to;*

- *Measure aspects of geographic phenomena and processes;*
- *Represent these measurements usually in the form of a computer database to emphasize spatial themes, entities and relationships;*
- *Operate upon these representations to produce measurements and to discover new relationship by integrating disparate sources; and*
- *Transform these representations to conform to other frameworks of entities and relationships."*

In 1991, Openshaw wrote his thoughts about the inadequacy of spatial analysis toolbox that is served with GIS. His main concern was that among 1000 commands, few were concerned with spatial data analysis rather than data visualization and manipulation such as creating thematic maps, buffering, overlay and query. At the same year, Ding and Fortheringham indicated that spatial data analysis and GIS were developed independently from one another. He was sharing the same opinion with Openshaw (1991) that GISs became display devices without claiming any spatial analytical capabilities and perhaps some basic descriptive statistics related to area and distance computations. In these years, there was an increasing need to develop new, largely automated exploratory spatial data techniques integrated with GIS. But, still, GIS is an aid in spatial data analysis by its geo-relational database structure which combines tabular and locational information. The link between these two allows fast computation of various characteristics of data. Also, it provides tools for appropriate visualization.

In 1992, Anselin indicates that the need of spatial data analysis and GIS integration is partly accomplished by augmenting these simple mapping techniques with new methods such as detection of local and global patterns and associations as part of an inductive approach to exploring data (exploratory spatial data analysis or ESDA). This gap has been filled by close or loose coupled spatial analysis modules with GIS

softwares. But, even if they are partially integrated, Leung (2000) denotes that their capabilities in spatial analysis and decision making are still far from satisfactory.

As with the widespread growth in the availability in spatial data and emergence of very large data bases, in 2000, Anselin states that classic spatial data analyses such as standard error or significance tests lost their meaning. He adds that computations for traditional hypothesis testing and statistical modeling became unfeasible with datasets with hundreds of observations involving matrices.

The databases with unimaginable size emerged a new demand to convert all this data into knowledge leading to better decisions and robust tools that demand little in terms of assumptions about the distributions underlying in the data and processing capacities of computers (Bacao et al, 2005-c). Thus, knowledge discovery and spatial data mining algorithms, which automatically discover implicit knowledge from huge amounts of spatial data, has recently received wide attentions. Also, this necessity gave birth to a new concept called "geocomputation". Macmillan (1997) determines geocomputation as a postmodern turn in the science of geography. He states that;

> *"But the key feature of geocomputation as far as I am concerned is the domain it belongs to – the domain of scientific research. Just as astronomy emerged with extraordinary vitality in the post-Galilean world, so geography can emerge from its slumbers in a geocomputational world."*

According to Openshaw (2001), "Geocomputation is the application of computationally intensive approaches to the problems of doing physical and human geography in particular and the geosciences in general". Geocomputation is based on computational intelligence which is also called artificial intelligence. The four leading technology of geocomputation are;

- Data created by GIS,
- Tools provided by computational intelligence,
- Power provided by high performance computers, and
- Philosophy provided by science (Openshaw, 2001).

Computational intelligence (CI), as an instrument for geocomputational tools, combines three main technologies aimed at the development of intelligent systems. These are granular computing, neural networks and evolutionary approaches. They are complementary to each other (Bacao et al 2005-c) (Figure 3.17).

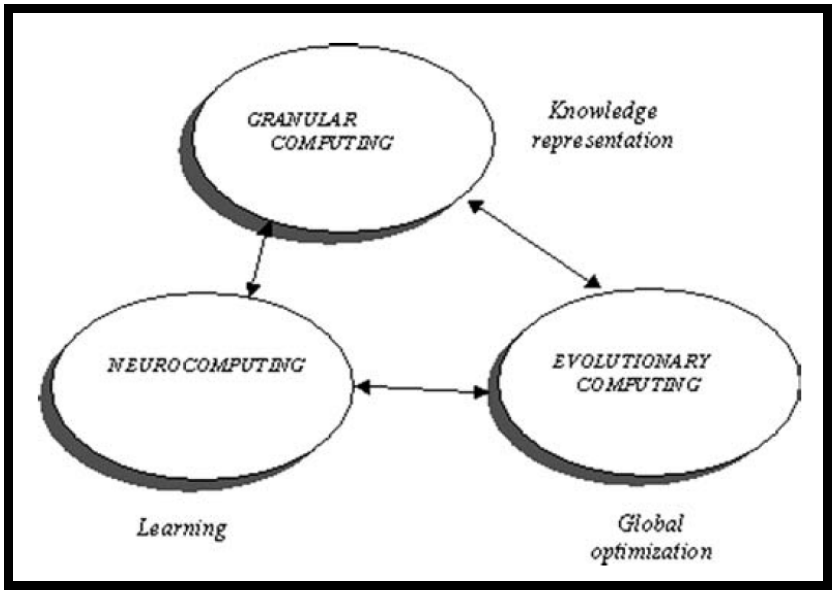


Figure 3.17 Main links in Computational Intelligence (Bacao et al, 2005-c).

Bacao et al (2005-c) justifies the interest of GIS researchers in CI tools with their properties that includes the ability to handle very large numbers of spatial objects, sensitivity to the special nature of spatial data, flexibility, freedom from distributional assumptions and ability to generate mappable results.

3.12. Clustering Methods Implemented in Study

In this thesis, traditional and novel spatial clustering methods are used as an extension of spatial data mining techniques that is based on computational intelligence integrated with GIS techniques for socio-economic data. Clustering techniques listed below are implemented to fulfill some of the principles such as

compactness, contiguity, homogeneity and equal population to achieve small statistical areas for the study area. These are;

- K-means as a classic representative of partitioning methods,
- Self-organizing maps (SOM) from artificial neural networks, and
- Simulated annealing from metaheuristic clustering methods.

3.12.1. K-means Clustering

Spatially, cluster analysis will seek to form segmentation into regions which minimize within-cluster variation but maximize between-cluster variation (Birimcombe, 2003). Clustering has a long and rich history in a variety of scientific fields (Jain, 2008). One of the most popular and simple partitioning clustering algorithms, k-means, was first published in 1955 by Hugo Steinhaus methodologically. But, the term “k-means” was first used by James MacQueen in 1967. Then, the standard algorithm was first proposed by Stuart Lloyd in 1957 which is also known as “Lloyd’s algorithm” (Wikipedia, 2009-b). Even though K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering because of its ease of implementation, simplicity, efficiency, and empirical success.

K-means clustering algorithm is extensively used in epidemiology, criminology, geodemographics and archeology for exploratory spatial data analysis and spatial data mining. As for computer vision, k-means is commonly used as a form of image segmentation. The results of the segmentation are used to aid border detection and object recognition (Wikipedia, 2009-b).

K-means clustering is an algorithm to classify or to group objects based on attributes into K number of group where K is a positive integer number. First, K initial centroids are chosen according to a user specified parameter. Each point is assigned to the closest centroid and each collection of points assigned to a centroid is a cluster. To assign a point to the closest centroid a proximity measure is needed which quantifies the notion “closest”. Euclidean distance is often used for data points. If we use Euclidean distance as a proximity measure, quality assessment of clustering is done using the sum of the squared error (SSE) which is also known as scatter or

square distortion. To compute total SSE, error of each data point, i.e., its Euclidean distance to the closest centroid is calculated. For example, if we want to compare different sets of clusters that are produced by two different runs of k-means, one with the smallest squared error (SSE) is preferred. The grouping is done by minimizing the sum of squares of distances between data and the centroid. For this reason, the centroid of each cluster is then updated based on the points assigned to the cluster. Thus, the partition optimizes a statistical homogeneity criterion – namely the total expected squared dissimilarity is minimized. This optimization iterates till no point changes clusters, or equivalently, until the centroids remain the same (Estivill-Castro & Houle, 2000; Kumar et al, 2006; 121). The whole basic process can be seen in Figure 3.18.

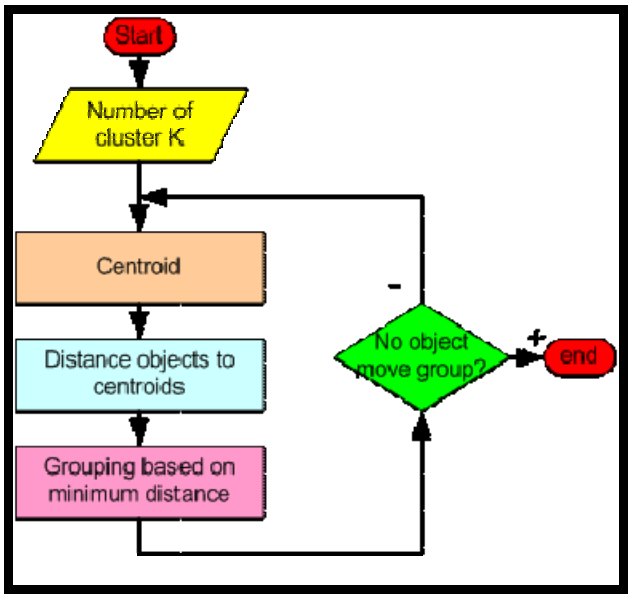


Figure 3.18 K-means clustering process (Web 7).

Pena et al (1999) defines k-means process in a reasonable mathematical formula. According to this formulation, k-means algorithm finds locally optimal solutions using as clustering criterion F, the sum of squared Euclidean distances between each element and its nearest cluster center (centroid). This objective function minimizes the SSE, the square-error criterion. Therefore, it follows that;

$$F(\{C_1,...,C_K\}) = \sum_{i=1}^{K} \sum_{j=1}^{K_i} \left\| w_{ij} - \overline{w_i} \right\| \tag{3.1}$$

Where;

F = Optimization function F as clustering criterion (SSE),

$\{C_1,...,C_K\}$ = Current partition of the database,

K = Number of clusters,

K_i = Number of objects in cluster i,

w_{ij} = j[th] object of the i[th] cluster,

$\overline{w_i}$ = the centroid of the i[th] cluster which is defined as;

$$\overline{w_i} = \frac{1}{K_i} \sum_{j=1}^{K_i} w_{ij}, i = 1,..., K \tag{3.2}$$

Considering Equation 3.1 and 3.2, k-means pseudo-code can be written as in Figure 3.19 below;

Step 1. Select the initial partition of the database in K clusters $\{C_1,...,C_K\}$,

Step 2. Calculate the cluster centroids $\overline{w_i} = \frac{1}{K_i} \sum_{j=1}^{K_i} w_{ij}, i = 1,..., K$,

Step 3. FOR every w_i in the database and following the instance order DO,

Step 3.1. Reassign instance w_i to its closest cluster centroid, $w_i \in C_s$ is moved from C_s to C_t if $\left\| w_i - \overline{w_t} \right\| \leq \left\| w_i - \overline{w_j} \right\|$ for all $j = 1,..., K, j \neq s$,

Step 3.2. Recalculate the centroids for clusters C_s and C_t ,

Step 4. If the cluster membership is stabilized THEN stop ELSE go to Step 3.

Figure 3.19 The pseudo-code of k-means algorithm (Pena et al, 1999).

The process of k-means pseudo-code is illustrated in Figure 3.20. Starting from three centroids, the final clusters are found in four assignment-update steps. In this figure, each subfigure shows (1) the centroids at the start of the operation, (2) the assignment of the points to those centroids. The centroids are indicated by "+" symbol and all points belonging to the same cluster have the same marker shape.

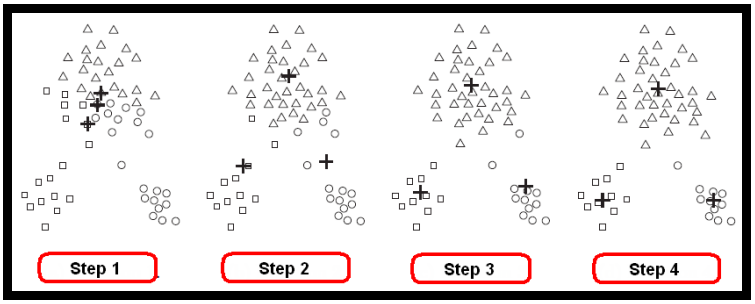


Figure 3.20 Using the k-means algorithm to find three clusters in sample data (Kumar et al, 2006).

In the first step shown in Figure 3.20, points are assigned to the initial centroids, which are all in the larger group of points. After points are assigned to a centroid, the centroid is updated. In the second step, points are assigned to the updated centroids, and the centroids are updated again in step 3. The k-means converges to a solution in step 4 because no more changes occur.

Huge size of data files are involved in clustering for data mining. Most clustering algorithms handles data by using complex similarity measures making their computational cost $O(n^2)$ unacceptable for clustering large datasets. For this reason, data-mining researchers adapt k-means or its variants for efficient processing of large sets with both numeric and categorical attributes. K-means only requires O(t*m*k*n) time, where t is the number of iterations over the entire dataset, m is the dimension, k is the number of clusters, and n is the number of data items. Number of iterations, t, is typically small (5-10) because most changes occur in the first few iterations. Moreover, t, m and k is significantly small than n, so one may simply describe k-means as requiring O(n) time (Estivill et al, 2000; Kumar, 2000).

The procedure always converges to a solution, although the solution is typically a local optimum. Global optimization measures the distance from every object to every other objects and takes the minimal sum of all distances as the best solution. But, solving this is computationally almost impossible, particularly when the number of objects is large. For example, with 6000 incidents grouped to 20 partitions, one cannot solve this with any normal computer since there are 6000!/20!*5980!=1456*1057 combinations. Therefore, k-means algorithm makes initial guesses about the centroids and optimizes these locations in relation to nearby points. This is called local optimization (Levine, 2009-b).

Despite the efficiency, simplicity and success of k-means algorithm, it has also many drawbacks. In literature, one can see that it has been proposed by several scientists in different forms and under different assumptions. But later on, as the disadvantages of the algorithm are emerged, many researchers investigated variations of the method to overcome these drawbacks (Bock, 2007). Below, the major drawbacks of the k-means algorithm are explained;

a) The most critical choice is in a k-means algorithm is the number of clusters, k. While no mathematical criterion exists, a number of heuristics are available for choosing k. Typically, k-means run independently for different values of k and partition with the smallest SSE or appears most meaningful to the domain expert is selected (Jain, 2008). An inappropriate choice of k may yield to poor results. For example, k-means algorithm can correctly find out the clustering centres as shown in Figure (3.21-b) when k is equal to the natural cluster number. Otherwise, it will lead to an incorrect clustering depicted in Figure (3.21-a) and (3.21-c) where some objects do not locate at the centers of the corresponding clusters. Instead, they are at some boundary points among different clusters or at points biased from some cluster centers (Cheung, 2003).

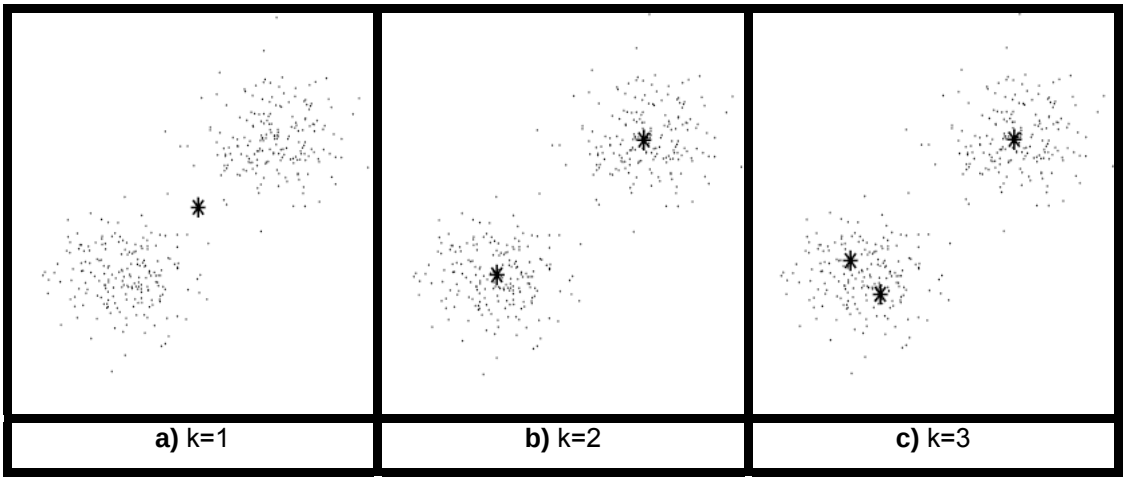


Figure 3.21 The results of k-means algorithm with two natural clusters where pre-determined number of clusters are a)k=1, b)k=2, c)k=3; "*" denotes the locations of centroids (Cheung, 2003).

b) When random initialization of centroids is used, different runs of k-means typically produce different total SSEs. For example, in Figure 3.20, even the initial cenroids are not from one natural cluster, minimum SSE clustering is still found. In figure 3.22, however, even though the initial centroids seem to be better distributed over data points, we obtain a clustering which is far from being optimal. This situation emerges from the fact that k-means algorithm is trapped in the local optimum which means that the algorithm only searches the nearby objects around centroids instead of looking at all choices. One way to overcome getting trapped in local optimum is to run k-means algorithm for a given k with several different initial centroids and choose the partition with the smallest SSE.

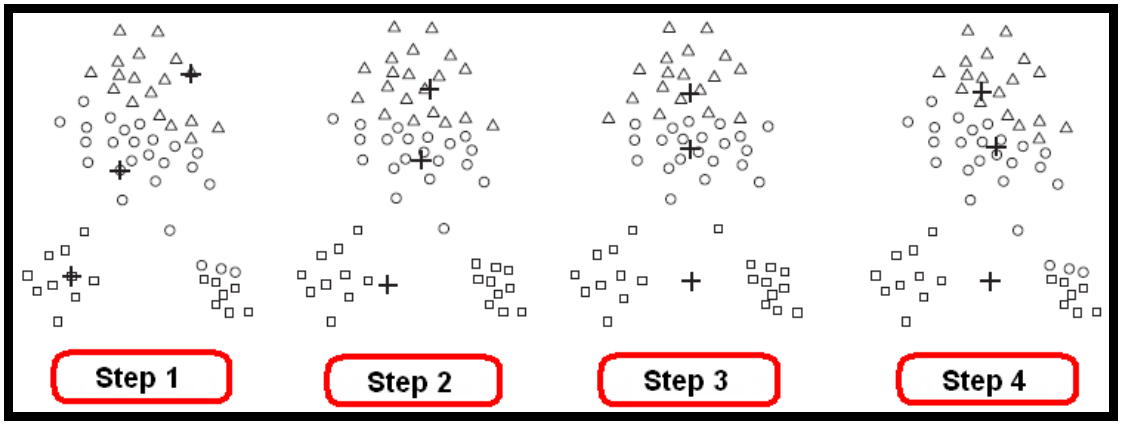
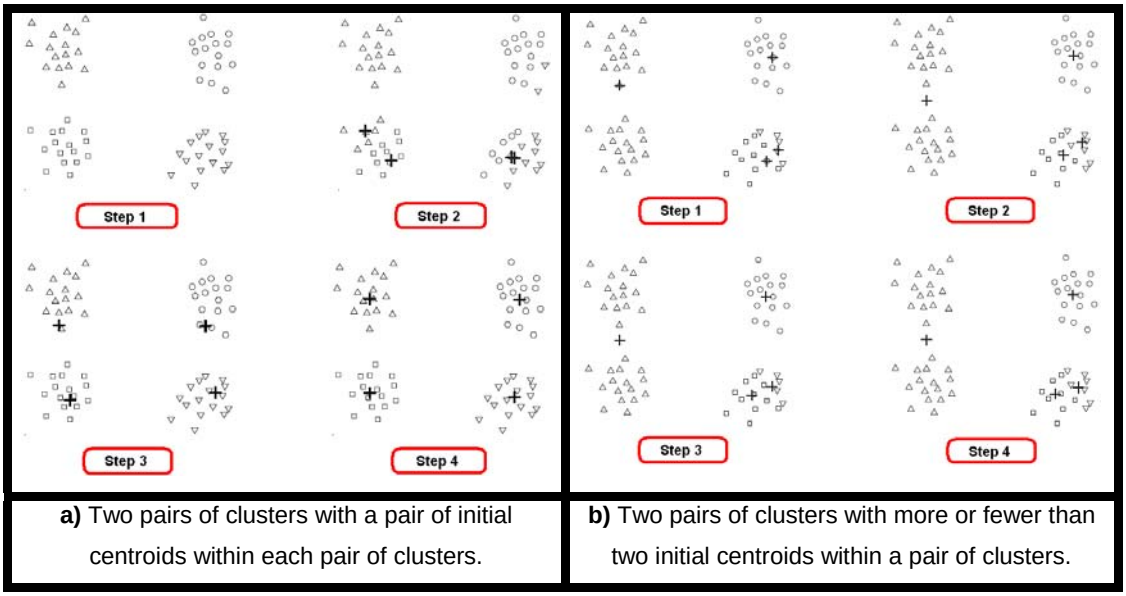


Figure 3.22 Poor initial centroids for k-means (Kumar et al, 2006)

c) The technique of performing multiple runs with different set of randomly chosen centroids and then selecting the set of clusters with the minimum SSE may not work well depending on the data set and the number of clusters sought. For example, in both Figure (3.23-a) and (3.23-b), the data consists of two pairs of clusters, where the clusters in each (top-bottom) pair are closer to each other than to the clusters in the other pair. In Figure (3.23-a), Step 2 shows that if we start with two initial centroids per pair of clusters, even when both centroids are in a single cluster, the centroids will redistribute themselves so that the true clusters are found. However, Figure (3.23-b) Step 2 shows that if a pair of clusters has only one initial centroid and the other pair has three, then two of the natural clusters will be combined and one true cluster will be split. In this case, because the pairs are farther apart than the clusters within a pair, the k-means algorithm will not redistribute the centroids between pairs of clusters, and thus, only a local minimum will be achieved.



a) Two pairs of clusters with a pair of initial centroids within each pair of clusters.

b) Two pairs of clusters with more or fewer than two initial centroids within a pair of clusters.

Figure 3.23 The two different initial centroids for k-means clustering within two pairs of natural clusters (Kumar et al, 2006).

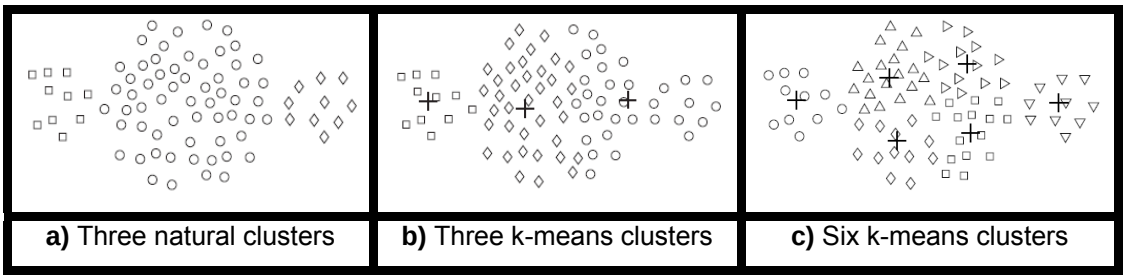
d) As k-means uses means as centroids, they are commonly adopted as representative of the data points of the cluster. However, it is possible for the average of the coordinates have no valid interpretation. For example, the average of the coordinates may indicate the school lies in the middle of a lake (Estivill et al, 2000). Also, arithmetic mean is not robust to outliers. Very far data instances from the centroid may pull the centroid away from the real one.

e) Another problem with k-means algorithm is that empty clusters can be obtained if no points are allocated to a cluster during the assignment step. This notion is also called “dead-unit” problem. That is, if some units are initialized far from the input dataset in comparison to other units, they then immediately become dead without getting a clustering chance in the whole clustering process (Web 7).

f) When applying k-means to a multivariate data, the clustering process will be done assuming that each attribute has the same weight. So which attribute contributes most to the clustering is not known.

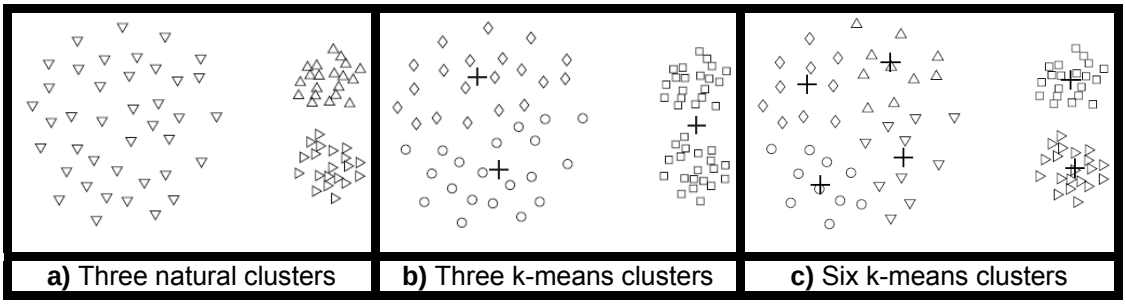
g) K-means and its variations have a number of limitations with respect to finding different types of clusters. In particular, k-means has difficulty detecting the natural clusters, when clusters have non-spherical shapes or widely different sizes or densities (Kumar et al, 2006). It implies that the data clusters are ball-shaped because it performs clustering is usually based on Euclidean distance as shown in Equation 3.1 (Cheung, 2003). This is illustrated by Figures 3.24, 3.25 and 3.26.In Figure 3.24 k-means cannot find the three natural clusters because one of the clusters is much larger than the than the other two, and hence, the larger cluster is broken, while one of the smaller clusters is combined with a portion of the larger cluster. In Figure 3.25, k-means fails to find the three natural clusters because the two smaller clusters are much denser than the larger cluster. Finally, in Figure 3.26, k-means finds two clusters that mix portions of the two natural clusters, because the shape of the natural clusters is not globular. However these limitations can be overcome, in some sense, if the user is willing to accept a clustering that breaks the natural clusters into a number of subclusters. Figure 3.26 shows what happens to the three previous datasets if six clusters are found instead of two or three. Each smaller cluster is pure in sense that it contains only points from one of the natural clusters.

K-means is restricted to data for which there is a notion of a center (centroid). A related technique, k-medoid clustering does not have this restriction, but is more expensive.



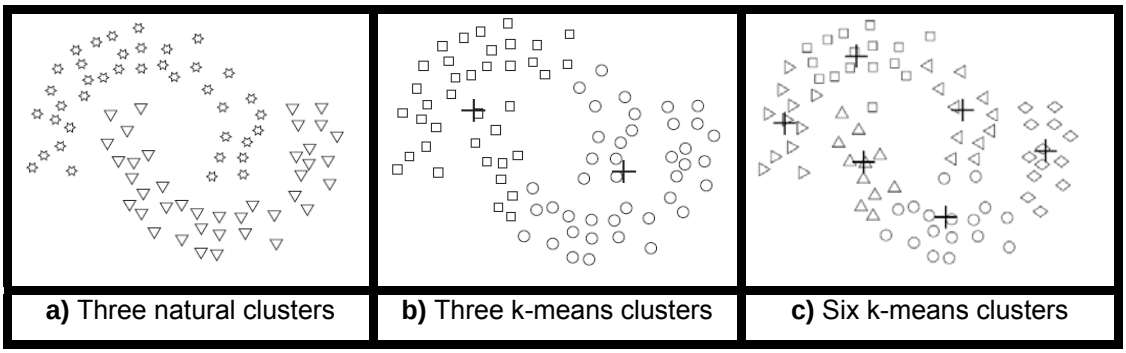
a) Three natural clusters	**b)** Three k-means clusters	**c)** Six k-means clusters

Figure 3.24 K-means with clusters of different size (Kumar et al, 2006).



a) Three natural clusters | **b)** Three k-means clusters | **c)** Six k-means clusters

Figure 3.25 K-means with clusters of different density (Kumar et al, 2006).



a) Three natural clusters | **b)** Three k-means clusters | **c)** Six k-means clusters

Figure 3.26 K-means with non-globular clusters (Kumar et al, 2006).

3.12.2. Self-Organizing Maps (SOM)

There is a long cartographic tradition of describing cities through a focus on the characteristics of their residents (Spielman & Thill, 2008). Visualization of population census data by cartographic means has for many decades have been an important in the hands of demographers, policy makers and community groups. The advent of GIS and related analytical approaches further increased the availability of map-based solutions for census-based data (Skupin & Hagelman, 2005). But, Basara & Yuan (2008) points out the complex, composite and mutually interacting nature of populations which end up with vast amount of interrelated and nonlinear multidimensional data. Even if map-based solutions are useful tools for describing the environment and presenting demographic data by preserving topological relations, they are limited in that they can only show one or two-dimensional picture of the social characteristics of an area. While maps are an efficient and familiar medium, they have limitations when it comes to displaying multiple pieces of

information about the same location (Spielman et al, 2008). So, for multidimensional data, maps are not a good type of data reduction utilities.

As another solution, today, most of the known systems use statistical classifiers. But statistical classification has many disadvantages which have been pointed out by Openshaw (1996). These disadvantages include;

1) Data reduction techniques, such as factor analysis (FA), reflect the choice of the original attributes. But, the method does not take geography into account, and the factor labeling process is subjective and difficult.
2) Normalizing techniques in statistical classification, such as principal components analysis (PCA), which assume linear relationships and are sensitive to non-normality; although there is no reason to assume that the important relationships that occur in the data are linear.
3) Classification algorithms often date back to the 1970s, when computers were much slower and computing time was very expensive. Shortcuts and algorithms designed to minimize computer costs were necessary. But there is no problem with the computational power today.

Thus, complexity arises when standard statistical modeling methods are applied to nonlinear and skewed spatial data sets with interactive variables.

Conventional multivariate statistical methods such as factor analysis (FA) and principle component analysis (PCA) serve the needs of data projection, while k-means and other clustering techniques deal with data quantization. Data projection and data quantization are commonly carried out sequentially, but any conclusions inferred from the second step are then conditional upon the outcome of the first step. To avoid this problem, data compression methods capable of both tasks simultaneously should be given preference. In addition, discovery of how to make the cluster assignment or allocation stage more sensitive to the spatial nature of the task is also important (Yan & Thill, 2009; Openshaw et al, 1997).

It has been suggested that artificial neural networks (Openshaw et al, 1997), may provide the power and flexibility to improve the overall quality of multidimensional geo-referenced data classifications. The method has many advantages, the major

one being that neural networks use data to discover patterns and geographical relationships. In addition, they can handle nonlinear relationships; they manage noise, and have a high degree of automation. They can achieve the same or even higher efficiency (Openshaw et al, 1997) in most of the functions of conventional methods. They do not contain any hypothesis about the nature or distribution of data, but are a lot more usable by the average user than conventional methods. Neural networks provide valuable help in the handling of problems of a geographical nature that have been impossible to solve so far (Hatzichristos, 2004).

Machine learning techniques of data mining, while still seldom used in urban analysis, have the potential to help analysts develop detailed differentiation of the urban landscape (Spielman et al, 2008). As described in Section 3.9, neural networks are a category of machine-learning methods that have recently become widely used in various problems involving prediction, classification, and pattern recognition. The basic principle of the neural network is the ability to `learn' from a complex set of input data through a training process, where a signal coming from outside the system is elaborated and transmitted by a set of `neurons' or `nodes' in such a manner that a stimulus response connection between input and output is formed (Kauko, 2005).

One of the best known and most efficient neural-network methods for achieving unsupervised classification is the “Self-Organizing Map (SOM)” method. The SOM algorithm, first introduced by Teuvo Kohonen was developed from the basic information processing modeling in the human brain’s cortical cells and much of the literature still uses the image of neurons in describing the building blocks of a SOM (Li & Shanmuganathan, 2007). The main problem in handling large amount of information is to find structures for memorizing, classifying and representing data as efficient as the human brain. The human thinking when processing perceptive and subconscious information tends to squeeze them via forming “reduced representations” of the most relevant facts, without loosing awareness of their interrelationships (Franzini et al, 2001).

Kohonen describes SOM as a “visualization and analysis tool for high-dimensional data” (Kohonen, 2001). SOMs have been applied to clustering, dimensionality

reduction, classification, sampling, vector quantization and data mining problems in fields as diverse as geographical information systems (GIS), bioinformatics, medical research, physical anthropology, natural language processing, document retrieval systems, and ecology (Izenman, 2008). The SOM method, also known as Kohonen neural networks or self-organizing feature maps (SOFM), is one of the methods increasingly adopted within geocomputational models. Because it provides a novel approach to the analysis of multi-temporal, multi-attribute, geographic data, in which the dimensionality reducing and clustering ability of the SOM method is combined with the integrated handling of two-dimensional geometry visualization and associated attributes provided by GIS (Skupin & Hagelman, 2005). This is accomplished by reducing high-dimensional spatial dataset to a lower dimensional nonlinear manifold, usually two or three dimensions, and in displaying graphically the results of such data reduction. Kohonen (1996) describes this technique as "…a mapping from a high-dimensional data space onto a (usually) two-dimensional lattice of points". In other words, disordered and vast amount of information is preprocessed and analyzed to give visual patterns, thus forming a landscape of the phenomenon described by the dataset.

Moreover, SOM is a topology preserving algorithm that data samples that are close (similar) to each other in the input space are also close to each other on the low dimensional space. In this sense, SOM resembles a geographic map concerning the distribution of phenomena, referring to the first law of geography: "everything is related to everything else, but near things are more related to each other" (Tobler, 1979).

The added value of the SOM is its ability to discover hidden data patterns, structures, and relationships in multivariate datasets. It also can conceptualize and map data in one-dimensional (1-D), two-dimensional (2-D), or three-dimensional (3-D) output space using a variety of topological structures (e.g., linear, rectangular, toroidal, spherical, cubic, etc.).

The input data for the SOM is the attribute space X with, for example, with “n” size and “d” dimension as shown in Figure 3.27.

$$X = \begin{Bmatrix} x_{1,1}, x_{1,2}, \ldots x_{1,i}, \ldots x_{1,\mathbf{d}} \\ x_{2,1}, x_{2,2}, \ldots x_{2,i}, \ldots x_{2,\mathbf{d}} \\ \ldots \\ \ldots \\ x_{n,1}, x_{n,2}, \ldots x_{n,i}, \ldots x_{n,\mathbf{d}} \end{Bmatrix}$$

Figure 3.27 Input vectors extracted from the raw attribute data.

Normalization (scaling) must be applied to raw data at the preprocessing step by using one of the normalization methods such as variance, minimum-maximum or range normalization in order to assure that all variables have equal weights and none of them dominates other variables. The user can also assign a weight to some variables so that they have a specified level of impact on the clustering (Guo et al, 2005; Kauko, 2009).

The SOM training algorithm involves essentially two processes namely;

- Vector quantization, and
- Vector projection (Vesanto & Alhoniemi, 2000).

Vector quantization is to create the representative set of vectors, in other words weights from the input vectors of the raw data. The output vectors can be denoted as $m = \{m_{i1}, m_{i2}, ..., m_{ik}\}$ with the same dimension as the input vectors but lesser in size than input vectors where (n>k). So in general, vector quantization reduces the number of vectors, in other words create clusters from the raw data with a reduced number of similar vectors (Jiang, 2004). We can say that SOM carries out a many-to-one projection, i.e., more than one data item can be represented to the same output vector if they are similar enough. The output vectors, representatives of

the input vectors, have also been called synaptic weight vectors, prototypes, codebook vectors, reference vectors and model vectors (Izenman, 2008).

The other process, vector projection aims at projecting the k output vectors (codebook vectors) onto a regular tessellation which is called SOM plot. The SOM plot is the end product of the SOM algorithm (after a large number of iteration steps) and is a graphical image. The SOM plot is displayed in the output space and consists of a grid (or network) of a large number of interconnected nodes (or artificial neurons) in two dimensions. The nodes are typically arranged on a square, rectangular or hexagonal grid as can be seen in Figure 3.28. Hexagonal grids are usually preferred because they provide equal distances between units in the output space and because of visualization concerns. In a two-dimensional SOM grid, for example, the set of rows are $K_1 = \{1,2,...,K_1\}$ and the set of columns are $K_2 = \{1,2,...,K_2\}$, where K_1 (the height) and K_2 (the width) are chosen by the user. Then, a node is defined by its coordinates, $(l_1, l_2) \varepsilon K_1 \times K_2$. The total number of nodes are; $k=K_1 \times K_2$ which is equal to the number of output vectors.

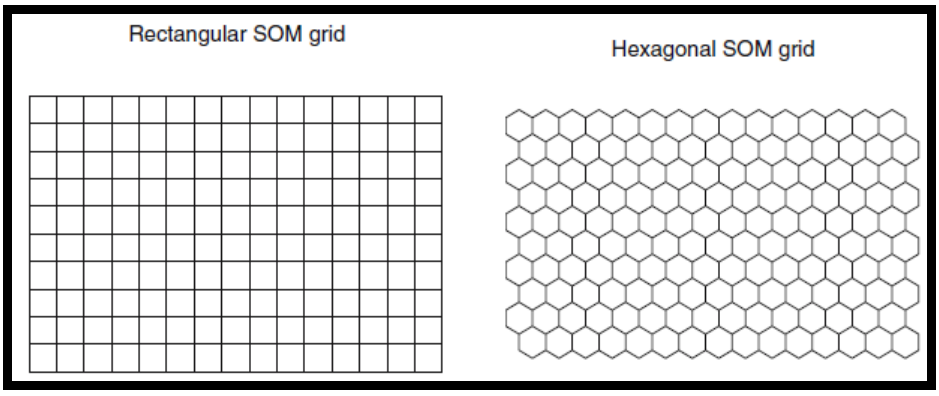


Figure 3.28 Displays of 10×15 rectangular and hexagonal SOM grids with 150 nodes (İzenman, 2008).

In the vector projection, each input vector is projected into a neuron by using output vectors (weights), where the projection is performed in a competitive fashion in order to ensure that “close” (similar) codebook vectors will be projected into neighboring

neurons on SOM plot. Both vector quantization and vector projection tasks are illustrated in Figure 3.29, where both input and output vectors are represented as a table format with columns as dimensions and rows as IDs of vectors. In figure, SOM is represented by a transitional color scale, which implies that similar neurons are being together. But it should be noted that, as an explanatory expression, the quantization and projection tasks are separated, which are actually combined together in SOM without being one after other (Jiang, 2004).

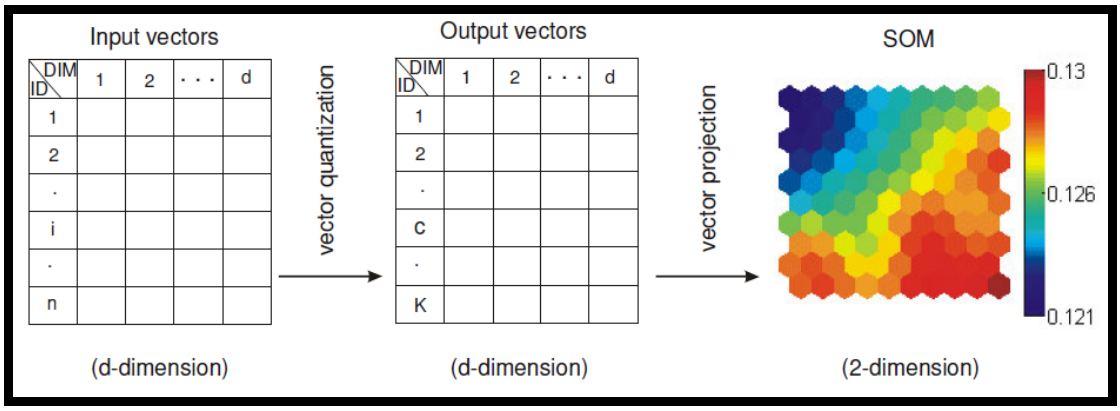


Figure 3.29 Illustration of SOM quantization and projection principle (Jiang, 2004).

In a more mathematical fashion, at the first step of the SOM algorithm, the map size is set up ($k=K_1xK_2$) and all the codebook vectors (m_k) are randomly initialized for each neuron so that they each consist of random weights (w_{ij}); i and j being their coordinates on SOM grid. The selection of map dimension is a matter that relates to the desired trade-off between the level of resolution (a larger map has a better resolution) and the level of generalization (a smaller map is more shallow). SOM feature maps of different sizes have different characteristics (Skupin et al, 2005). Small feature maps provide generalizations; large grids allow a unique location in geographic space to be mapped to a unique location in the synthetic attribute space. In a large feature map, where the number of buckets exceeds the number of observations, each bucket may hold few, if any observations; regions have very specific properties. On the contrary, in a small SOM feature map where the number of observations far exceeds the number of buckets, many observations will fall into

each bucket and regions of the map will represent general characteristics (Spielman et al, 2008) (Figure 3.30).

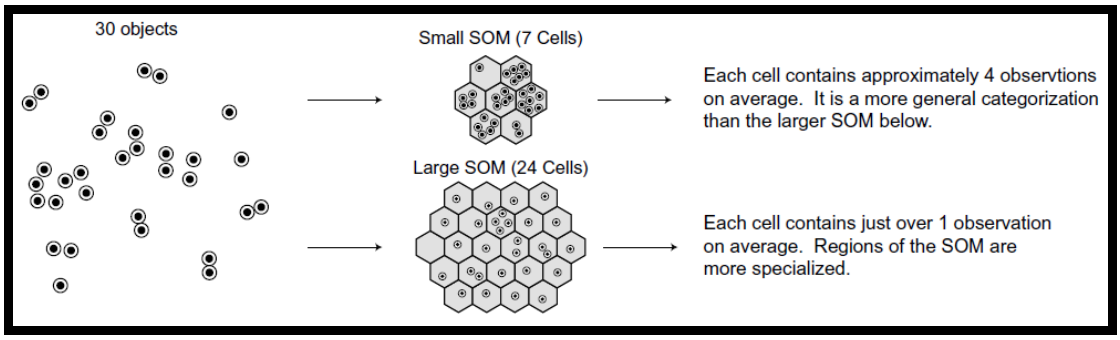


Figure 3.30 Self-organizing map size (Spielman et al, 2008)

However, the map size, in other words default number of neurons, is calculated using heuristics given by Vesanto et al (2000) below;

$$k = 5 * \sqrt{Number \quad of \quad Samples \ \ (n)} \tag{3.3}$$

Then, each normalized input vector (x) from raw data is randomly chosen and Euclidean distance between it and all the output (codebook) vectors is calculated (Equation 3.4). The output vector which is closest, in other words most similar, to that input vector is called the Best Matching Unit (BMU), denoted by (m_c);

$$\left\| x - m_c \right\| = \min_k \left\{ \left\| x - m_k \right\| \right\} \tag{3.4}$$

Next, the codebook vectors are updated by using a specific learning rate and neighborhood function as in Equation 3.5. The BMU and its topological neighbors are moved closer to the input vector in the input space. This adaptation procedure stretches the codebook vectors of the BMU and its topological neighbors towards the sample vector (BMU) which is usually called the unfolding phase of training. This is illustrated in Figure 3.31. The SOM update rule for the codebook vector of the unit (m_k) is:

$$m_k(t+1) = m_k(t) + \alpha(t) h_{ck}(t)[x - m_k(t)] \quad (3.5)$$

Where;

t = time,

$\alpha(t)$ =learning rate,

$h_{ck}(t)$ =The neighborhood kernel around the BMU.

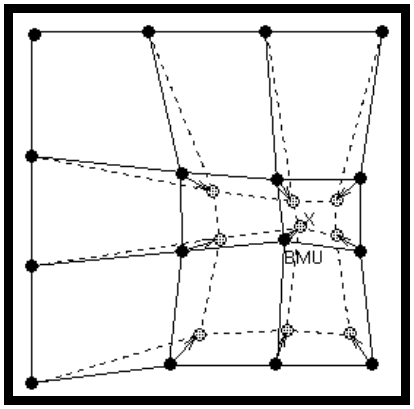


Figure 3.31 Updating the best matching unit (BMU) and its neighbors towards the input sample marked with x. The solid and dashed lines correspond to situation before and after updating, respectively (Web 8).

Neighborhood function has different formats, but Gaussian function is usually adopted and it is defined by Equation 3.6 below;

$$h_{ck}(t) = e^{-d_{ck}^2 / 2\sigma_t^2} \quad (3.6)$$

Where;

σ_t = The neighborhood radius at time t,

d_{ck} = The Euclidean distance between neurons c (BMU) and k on the SOM grid.

It should be noted that the size of the neighborhood reduces slowly as a function of time, i.e., it starts with fairly large neighborhoods and ends with small ones as can be seen in Figure 3.32 below.

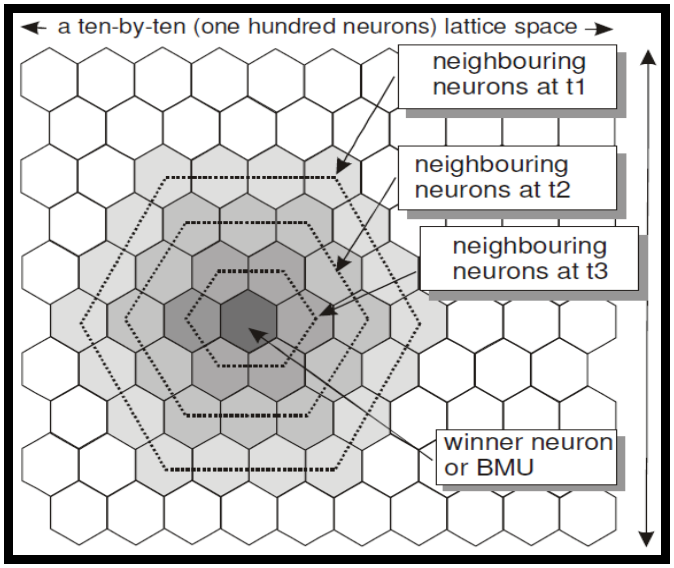


Figure 3.32 The characteristics of a 10x10 SOM with smaller neighborhoods at times t1<t2<t3 (Jiang, 2004).

The learning rate function ($\alpha(t)$) and neighborhood radius (r) can decrease linearly, exponentially or inversely proportionally to time t, but linear function is conventionally used. Usually the learning length is divided into two periods: t_1 for the initial coarse structuring period and t_2 for the fine structuring period. But in principle, learning rate begins from 0.5 (t_1) in the first phase, and from 0.05 (t_2) in the second phase until it reaches 0 in order to guarantee convergence and stability of the SOM. In addition, a useful “rule of thumb” is to run the algorithm steps for at least 10 times the neuron numbers divided by dimension length for rough phase and 40 times for finetuning phase (Kohonen, 2001, p. 112). Another “rule of thumb” is that neighborhood radius starts from k/4 and goes down to one fourth of that (unless this would be less than 1). On second phase, neighborhood radius starts from where it

stopped in first phase, and goes down to 1. The length of second phase is 4 times that of the first phase (122).

Considering Equation 3.3, 3.4, 3.5 and 3.6 k-means pseudo-code can be written as in Figure 3.33 below;

Let

X be the set of n input vectors in raw data having d dimensions as $x_{1d}, x_{2d},...,x_{nd}$

m_k be the codebook vectors on a $K_1 x K_2$ grid where they have weights as $w_{ij;}$ i and j are their coordinates on that grid

d_{ck} be the distance between neurons m_c (BMU) and m_k on the SOM grid.

$h_{ck}(t)$ be a neighborhood function having a neighborhood radius (r).

$\alpha(t)$ assuming values in $[0,1]$, initialized to a given initial learning rate

For each input vector in raw data;

Repeat;

Step 1. Calculate the distance between the input vectors and all neurons of the SOM ($\|x - m_c\| = \min_k \{\|x - m_k\|\}$) (this is called the calculation phase)

Step 2. Select the nearest unit as winner (BMU) that minimizes the distance (this is what is usually called the voting phase)

Step 3. Update each unit of the SOM according to the update function ($m_k(t+1) = m_k(t) + \alpha(t) h_{ck}(t)[x - m_k(t)]$) (this is what is usually called the updating phase)

Step 4. Decrease the $\alpha(t)$ and radius (r) in function $h_{ck}(t)$

Step 5. Until $\alpha(t)$ reaches 0 and radius (r) reaches 1.

Figure 3.33 The pseudo-code of SOM algorithm.

As mentioned earlier, in order SOM to converge to a stable solution, both the learning rate ($\alpha(t)$) and neighborhood radius (r) should converge to zero. In addition, the update of both parameters (Step 3 in psudo-code) can be done different from the procedure that involves presenting each individual data unit to the

network one-by-one which is called “iteration” or “sequential training”. This case is known as “batch training”. The difference of batch training from sequential training relies on the unit’s updating process and on the non-obligation to randomly present training input vectors to the network. Also sometimes the learning rate can be omitted. In this algorithm, neurons are updated only after an “epoch” which means to present input vectors to the network at once (Henriques, 2005). In each epoch the input space is divided using voronoi regions which are also known as thiessen polygons. These regions are polygons which include all points which are closer (similar) to a codebook vector than to any other (Figure 3.34).

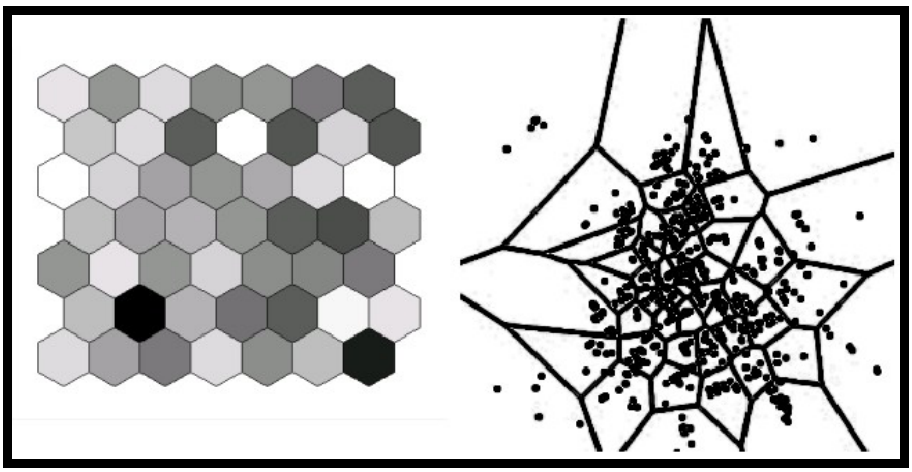


Figure 3.34 Voronoi regions for codebook vectors (Fincke et al, 2008).

In this case, the update processes of weight (codebook) vectors are calculated according to the Equation 2.7 below (Vesnato et al, 2000);

$$m_k(t+1) = \frac{\sum_{k=1}^{N} h_{ck}(t)\left[x - m_k(t)\right]}{\sum_{k=1}^{N} h_{ck}(t)} \tag{3.7}$$

While sequential algorithm is highly dependent on the order of data input, batch training overcomes this drawback by presenting the outputs at once. Also, batch

training is much faster than sequential algorithm (Izenman, 2008). The whole process of SOM algorithm can be summarized and visualized by the Figure 3.35 given below.

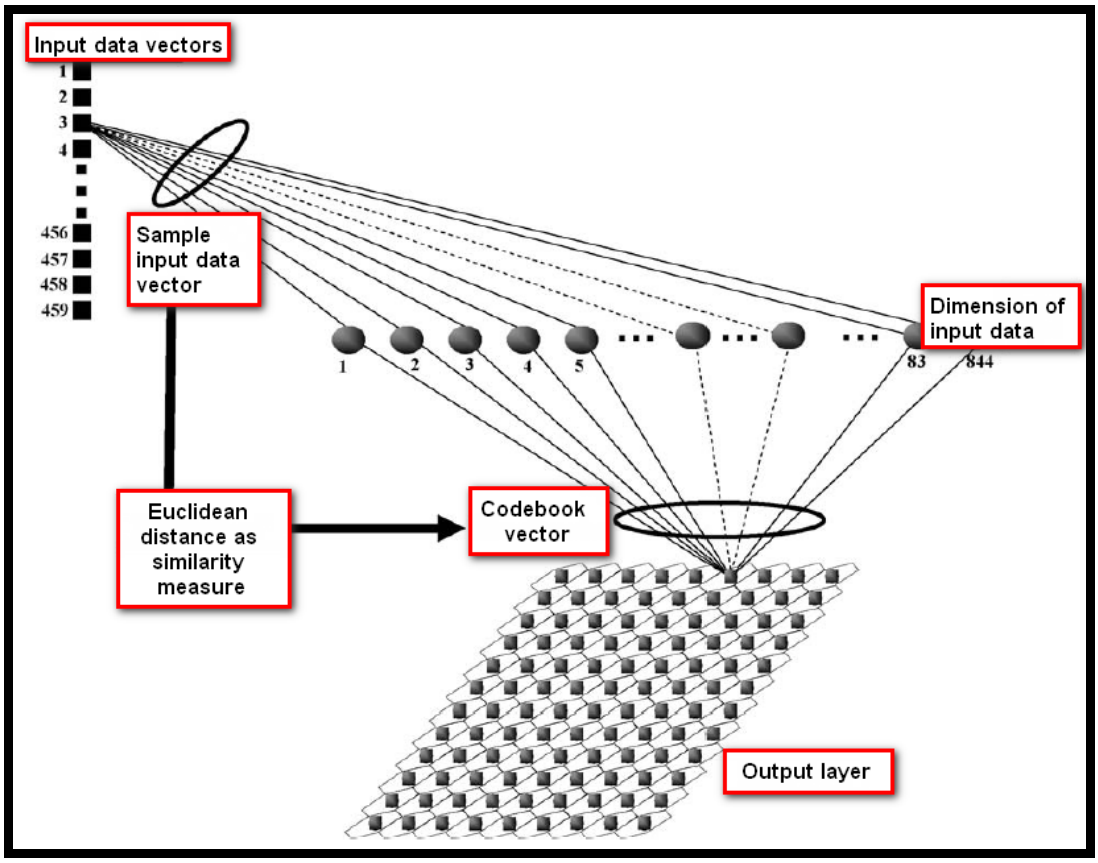


Figure 3.35 Self-organising map architecture: The input layer is linked to the cells of the output layer by connections called weights. The sampled vector composed of as much element as descriptors is compared to the virtual vectors associated to each neuron of the output using a distance measure.

For quality assessment of SOM training, in this thesis, we focused on quantization error (qe) and topographic error (te) (Kohonen, 2001). The quantization error is given by the average distance between a codebook vector and the input data vectors which share the given codebook vector as a BMU (Equation 3.8).

$$qe = \frac{\sum_{k=1}^{n} \|x - m_k\|}{n} \tag{3.8}$$

Where “x” is one of the input data vectors and “m_k” is BMU, in other words best codebook vector, for that particular input vector and “n” is the number of existing input data vectors.

The topographic error (te) evaluates the topology preservation of the SOM. Considering that, for each input data unit the closest codebook vector will be its BMU, the second closest unit will be called the second BMU or BMU2. Thus, topographic error measures the proportion of all input vectors for which the first and second BMUs are not adjacent units (Equation 3.9).

$$te = \frac{\sum_{k=1}^{n} u(x_k)}{n} \tag{3.9}$$

Where, “x_k” is one of the input data vectors, n is the number of data vectors and “$u(x_k)$” equals to one (1) if the BMU and the BMU2 are not adjacent and zero (0) otherwise. The SOM quality is better if both quality measurements are near or equal to zero (0).

The SOM algorithm has much in common with k-means clustering. In k-means clustering, items assigned to a particular cluster are averaged to obtain a “cluster centroid” (or “representative” of that cluster), which is subsequently updated. The crucial difference of SOM from the k-means classifier is the ‘neighborhood' concept that is, the concept of a ‘winner' node with adjacent neurons (at the beginning of the learning process, the exact extent of the hexagonal or rectangular neighborhood depends on the chosen network parameters such as radius of the neighborhood; at the end of the learning process, the Gaussian neighborhood comprises only the closest neighboring nodes and SOM begins to resemble k-means. In fact, if we omit the adaptation process (Step 4) from the pseudo-code, the algorithm becomes k-

means. This neighborhood concept is incorporated into the SOM algorithm but is missing from the k-means algorithm. In Figure 3.36 below, Gaussian squeeze can be seen during the update process of neurons in a particular neighborhood.

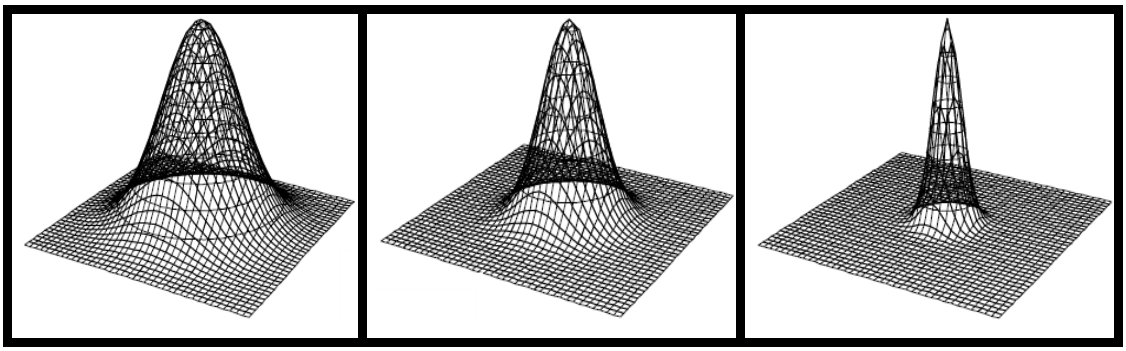


Figure 3.36 Gaussian squeeze during the adaptation phase (Franzini et al, 2001).

When dealing with spatial data, the common approach is to use only thematic attributes for the building of the SOM output (Koua & Kraak, 2004). This is often done when the output of the SOM is visualized together with a geographic map. But such a restriction is unnecessary and it has been proposed to include also the geometrical attributes such as latitude and longitude information of the cases into the SOM algorithm (Bacao et al, 2008). In such a way, similar data with close geometric distance are mapped onto the same node or to neighboring nodes in the map. So, SOM becomes a topology preserving data compressor and feature extractor (Silva et al, 2004).

In fact a SOM is not generated through completely mechanical procedures, but requires three kinds of external manipulation: first initialization, second selection of the networks parameters for the learning process and finally calibration of a stabilized map using labeled categories. So, the metaphor “unsupervised” is applied to the neural network by the fact that the nodes “compete for” and eventually “win” observations by some predefined equations. The “unsupervised” property refers to the results being strongly dependent on the input data only, but for the rest of the algorithm such as creating network as a SOM grid and deciding on the network parameters are compulsory prerequisites that is defined by the user (Kauko, 2009).

A different type of the cluster structure of a SOM is a u-matrix, where “u” stands for “unified distance” (Ultsch & Siemon, 1990). To visualize the results of a SOM, u-matrices may be used. The u-matrix is a representation of a SOM in which distances, in the input space, between neighboring neurons are represented, usually using a color or gray scale. If distances between neighboring neurons are small, then these neurons represent a cluster of pattern with similar characteristics. If the neurons are far apart, then they are located in a zone of the input space that has few patterns, and can be seen as a separation between clusters. Conventionally, the darker colors such as blue, represents lower values and these indicate how close the SOM neurons are to each other; whereas the brighter colors such as red represents higher values and these indicate how far apart the neurons are from each other (Fincke et al, 2008). . If we have K_1x K_2 nodes in a SOM grid than u-matrix will be a (K_1-1) x (K_2-1) matrix.

It is useful to think of the neurons on the feature map as buckets for data. For example, in order to define places with many wealthy householders, with high levels of education, high home-ownership rates, and low poverty rates would end up in buckets that are near each other and clustered in a region of the u-matrix. On the other hand, census tracts where poverty is abundant and residents typically have low levels of education would end up clustered in buckets in a different region of the u-matrix; probably quite far away from the well educated and wealthy people. Places that have both wealthy households and poor households would end up occupying a region of the map somewhere between the two extremes as fuzzy areas (Spielman et al, 2008). The u-matrix constitutes a particularly useful tool to analyze the results of a SOM, as it allows an appropriate interpretation of the clusters available in the data. (Fincke et al, 2008)

An additional visualization tool is a color map of the various component planes. In general, the “components” are the individual input variables that make up “X”, the raw data. The correlations and relationships in the input data space can be visualized using the component planes. The component planes show the values of the map elements for different attributes and how each input variable varies over the space of SOM grid (Koua & Kraak, 2004). Similar picture of variation means that two

input variables are closely correlated to each other. The component planes and the U-matrix are linked by position such as that the hexagon in a certain position in one SOM plot corresponds to the same hexagon in SOM plots.

In order to clarify the terms u-matrix and component planes, Figure 3.37 shows the visualization of SOM outputs regarding the simulated synthetic dataset of rare disease events as input and the respective identified spatial clusters. This visualization has been obtained from the study of Dai & Oyana (2009) which aims to introduce automatic cluster identifications for environmental applications. Figure 3.37 shows the variations and patterns of rare disease using u-matrix (top left) and four component planes with corresponding scale bars in the SOM. The U-matrix shows three obvious clusters of rare disease events. Comparing the four component planes, two clusters can be located on the north side, and the other cluster is located on the south side of the study area. The comparisons, between the four component planes (x coordinates of cases, y coordinates of cases, at risk population and kinds of rare disease cases), show very interesting results. First, rare disease hits harder in areas with small at risk populations, as can be seen in the component planes of population and case. In addition, the component planes show that most of the disease events occur on the southwest side and there is a large background population in the central west area.

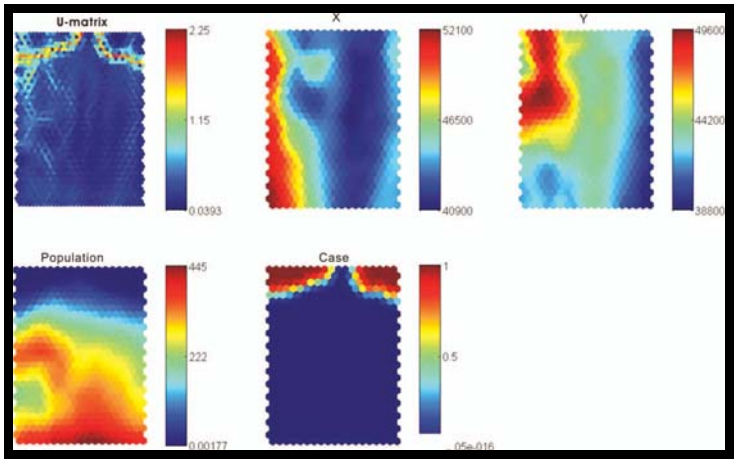


Figure 3.37 SOM visualization of the simulated synthetic dataset representing a rare disease (Dai & Oyana, 2009).

A common SOM (applied to geographic data) labels each node by the names of those geographic units that fall in a particular node (Guo et al, 2005). The neuron is labeled based on the majority of the labels it “won”. In this way, one may obtain a classification of similar types of observations that are significantly different from other types of observations. It should be noted that some nodes may fail to “win” any observations and thereby remain without a label (Kauko, 2005). Labeling function can be seen in Figure 3.38 which shows the output of Kaski & Kohonen’s (1996) study which aims to find out a homogenous structure about welfare and poverty among world countries. The order of the abbreviated country names indicates the similarity of the standard of living of the countries and the shades of gray indicate the degree of clustering. Light areas represent areas of a high degree of clustering and dark areas represent gaps in the degree of clustering. The countries labeled in lower case were not used in the computation of the map because too many indicators were missing from them. Dots denote map locations which did not correspond to any counties, in other words which did not win any observations. Subjectively, 6 clusters can be located on below SOM plot.

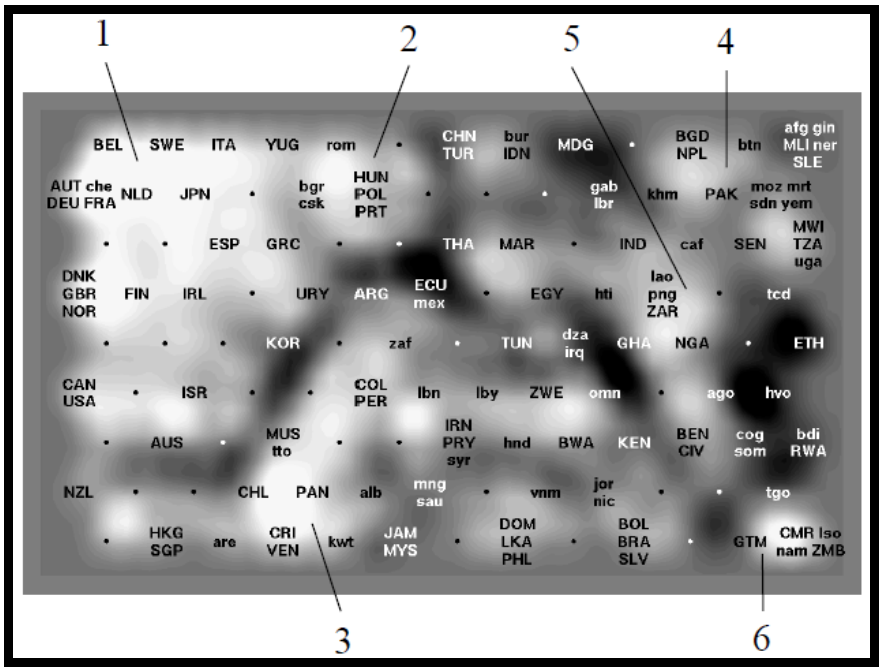


Figure 3.38 Labeled u-matrix to describe the standard of living in world’s countries (Kaski & Kohonen, 1996).

While the distance matrix (u-matrix) representation with labels is a good method for visualizing clusters, it does not provide a very clear picture of the overall shape of the data space because the visualization is tied to the SOM grid. If the user has very limited knowledge about the geographic locations of those places, labeling would not provide much helpful information in interpreting the spatial distribution of the discovered patterns. Also, one cannot see how much those nearby nodes are similar to each other since the SOM does not show the original data values, instead it color-codes them. But in order to make the output more understandable for an audience that is not familiar with SOMs, it is possible to code the resulting SOM classifications and export the results into a GIS. Coupling the SOM algorithm's pattern recognition capabilities with the spatial analysis capabilities of geographic information systems (GIS) provides mapping of the high-dimensional data.

Openshaw et al. (1995) summarizes the advantages and disadvantages of SOM algorithm as below. The advantages are;

1) Use of raw data removes the need for an orthonormalising linear filter.
2) The self-organising nature of a Kohonen map allows structure to emerge rather than be imposed from the top.
3) Incorporation of data uncertainty into the classification.
4) Simplicity and greatly reduced number of source code lines.
5) Possible to incorporate prior knowledge into the classification process making it more intelligent.
6) Fuzziness of the results is preserved in a particularly easy to use form.
7) Reduction in importance of knowing precisely how many clusters are needed.
8) Cluster interpretation is easier because the classification takes place in the data space rather than in some transform space.
9) Non-linear technology.
10) Less likely to be trapped in a local sub-optimum.

The major disadvantages are;

1) Extensive computer run times are needed requiring the use of parallel supercomputing to adequately train on large data sets.

2) A number of design aspects are entirely subjective, in particular; the number of training iterations, the architecture of the net, the updating process, and the choice of map the classification.

3) The current absence of experience interpreting the results.

4) Lack of experience with the technology.

The SOM algorithm has been used in the problem area of clustering and regionalization with promising results in literature. For example, Openshaw et al (1995) used SOM as an ability to apply a multivariate classification procedure to reduce the British census data in order to find out small statistical areas (SSAs). Similarly, Hatzichristos (2004) carried out a demographic classification of Athens, Greece, using the SOM algorithm in combination with fuzzy logic. Bacao et al. (2005-d) concentrated on the application of SOM for developing a variant of it which particularly interested in introducing the geographical knowledge into SOM algorithm. The study of Basara et al (2008) demonstrated a positive relationship between environmental conditions and health outcomes in communities using the SOM-GIS method to overcome vast amount of data and traditional methodological challenges. Kauko (2005; 2009) explored the variety of residential area types in three largest Dutch cities: Amsterdam, Rotterdam and The Hague. In his work, the SOM method and learning vector quantization is used to cluster and classify the multi-dimensional socio-demographic data and other objective indicators produced by official statistics. Koua et al (2004) investigated ways to integrate computational analysis based on SOM neural network for exploratory visualization to support visual data mining and knowledge discovery for health and demographic data. Li et al (2007) applied SOM to investigate city's social areas within Beppu City, Japan. Dai et al (2009) proposed a two-stage approach for automatic cluster identification for environmental health applications. In the first stage, they used SOM algorithm is used to accomplish two things; (1) to characterize and classify multivariate environmental datasets and (2) to visually explore the interactions between environmental variables and establish the number of clusters in each of the experimental datasets. In the second stage, they used a genetic algorithm to delineate final cluster boundaries. Silva et al (2004) applied SOM to discover urban social exclusion/inclusion in the city of Sao Jose dos Campos, Brasil. Lacayo &

Skupin (2007) developed a module for the conversion of SOM from its standard plot format (u-matrix) into a shapefile which facilitates the integration of a SOM into a GIS. Vesanto et al (2000) produced different approaches to clustering of the SOM in which they used SOM to create prototype vectors that are then clustered by k-means algorithm with the smallest Davies-Bouldin index. This approach is also adopted in this thesis.

3.12.3. Simulated Annealing

A very widely used method in combinatorial clustering optimization is simulated annealing. Simulated annealing sometimes called the Metropolis technique which was originally developed to solve a hard optimization in physics. It has subsequently developed into a global optimization method particularly suitable for problems with multiple local optimums (Openshaw & Rao, 1995). The basic running of the algorithm was told in section 3.10.

There are many different scientific branches in literature which utilize simulated annealing method. However, there are few applications which use simulated annealing for spatial analysis especially for census geography area. One of these studies is Openshaw & Rao's experiment (Openshaw & Rao, 1995) to find an optimum zone design methodology for UK's census geography called (Automated Zone Procedure) AZP. This procedure was also briefly explained in Chapter 2. Later, AZP procedure is used by Martin (2002) with the same aim, but this time the algorithm is enhanced by transforming it into a software called Automated Zone Matching (AZM) tool which uses a constrained simulated annealing methodology to derive the smallest dissemination unit for office of national statistics called Output Areas (OAs). Simulated annealing provides a robust method for optimization problems which are otherwise hard to solve. It was therefore attractive as a potential solution for finding optimum zone design for small areas.

In order to understand the basic simulated annealing AZP method (AZP-SA) a pseudo-code is given below;

Step 1. Randomly select an initial partition called P_0, and compute the squared error value, $E(P_0)$. Select values for the control parameters such as compactness, equal population and homogeneity, initial and final temperatures T_0 and T_f.

Step 2. Select a neighbor P_1 of P_0 and compute its squared error value, $E(P_1)$. If $E(P_1)$ is larger than $E(P_0)$, then assign P_1 to P_0 with a temperature-dependent probability. Else assign P_0 to P_1. Repeat this step for a fixed number of iterations.

Step 3. Reduce the value of T_0. If T_0 is greater than T_f, then go to step 2. Else stop.

Figure 3.39 The pseudo-code of simulated annealing algorithm (Jain et al, 1999).

The algorithm of simulated annealing can be best interpreted by referring to Martin's (2002) study. The aim in his study was to create SSAs in other words OAs from the smallest and widely used areal units in UK which are the postcode areas. Detailed information about census geography hierarchy of UK was given Chapter 2.

The AZP algorithm makes use of the contiguity information available from the GIS containing the postcode areas. It begins by estimating the approximate number of OAs that should nest within a constraining polygon such as ward boundaries by determining an input population target size, and then randomly aggregating adjacent postcode polygons to form equipopulous OAs. A number of statistical measures for this initial configuration are computed for each of the selected design constraints such as; Overall distance from target population is measured by the sum of the squared differences between OA populations and the target population size. The measurement of shape (compactness) and social homogeneity are also included in the constraints. This preliminary process is equal to step 1 in pseudo-code at Figure 3.39. Consideration is then given to the swapping of postcode polygons between adjacent OAs in terms of their impact on the statistical measures mentioned in step 1. For example, regarding population size, an improving swap will be one that reduces the total squared difference from target size by bringing an above-target and below-target OA closer to the desired size. This process also reduces or does

not make any change in the temperature of simulated annealing algorithm which brings the algorithm to optimized result closer or re-iterates. Such that, any swaps which serve to improve the overall solution in this way are accepted and incorporated into the emerging OA geography, while any that cause deterioration in the objective criteria are rejected.

The simulated annealing approach in Martin's (2002) study is illustrated in Figure 3.40. In Figure 3.40–a, an area is shown in which the postcode polygons have been grouped into three prototype OAs, indicated by the different shading. In Figure 3.40–b, one postcode polygon is selected for potential swapping into a neighbouring OA resulting in Figure 3.40–c. The overall quality of this configuration is assessed and found to be unsatisfactory, so the algorithm reverts to the original situation in Figure 3.40–d. Another postcode polygon is identified as a potential swap in Figure 3.40–e and this time results in an overall improvement (reduces the temperature), leading to its acceptance within the current best solution, as shown in Figure 3.40–f. No swaps are permitted which would produce sub-threshold OAs or break the internal contiguity of an OA. Once all available combinations have been tested in other words, once the final temperature is reached, the overall best solution is chosen.

This method can be thought as a constrained clustering which utilizes simulated annealing. A simulated annealing algorithm similar to the one in Martin's (2002) study is also implemented in this thesis to create small areas.

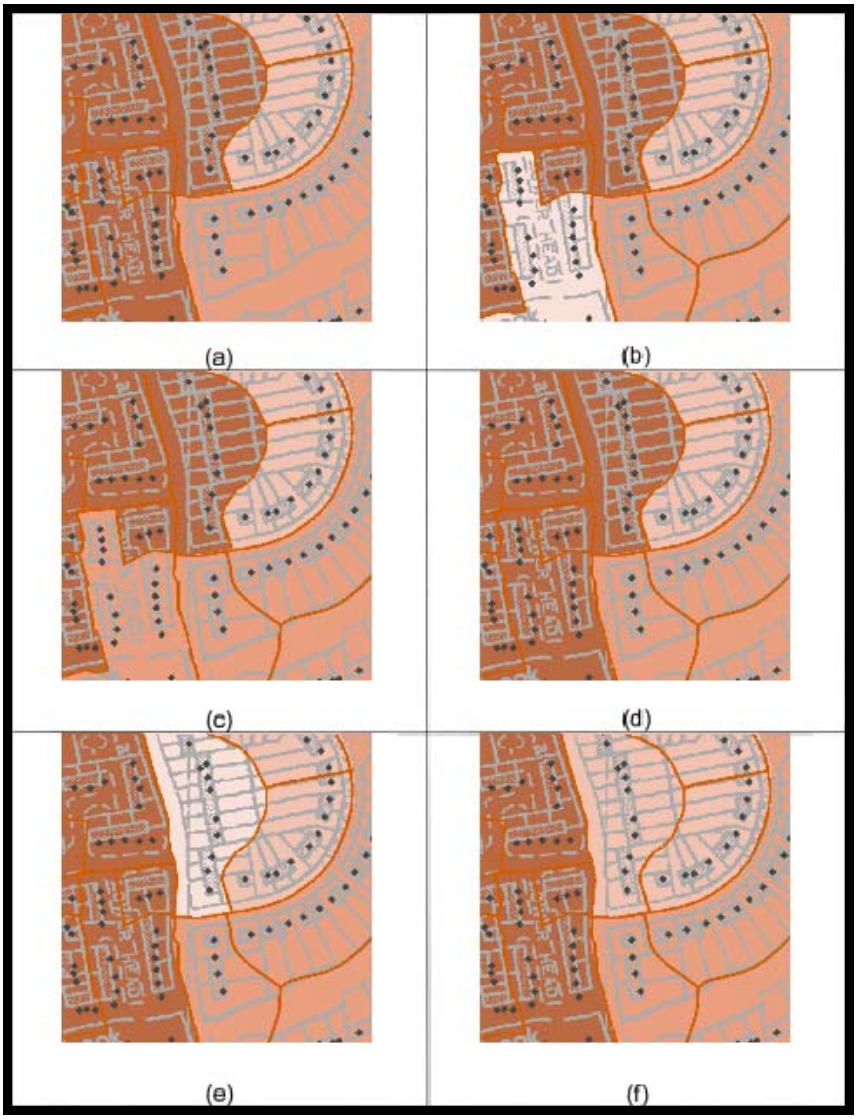


Figure 3.40 Illustration of the iterative swapping of postcode polygons between prototype output areas by the AZP algorithm (Martin, 2002)

3.13. Software Packages Employed for Developing SSA Clusters

There are mainly two software packages which are used to derive constrained clusters in this thesis. These clusters are accepted as small statistical areas that fulfill the small area identification principles which were clearly explained in Chapter 2. In following sections, these software and clustering logics underlying are clarified.

3.13.1. BARD: Better Automated ReDistricting

BARD was one of the two open source software packages for general districting and clustering analysis at the time of writing this thesis. The other software package is Martin's (2002) AZM tool. Unfortunately, as mentioned in Chapter 2, AZM tool only works with coverages that can be built with old Arc/INFO command prompts which are available with ESRI ArcGIS software having version older than 8.1. So, BARD package is utilized. BARD provides methods to create, display, compare, edit, automatically refine, evaluate, and profile political districting plans. BARD aims to provide a framework for scientific analysis of districting plans. BARD package is written by Micah Altman et al (2009) using R language and runs on R (Web 5, Web 9).

Since districting is a computationally complex partitioning problem not able to reach to an exact optimization solution, BARD implements a variety of selectable metaheuristics that can be used to refine existing or randomly-generated districting plans based on user-determined multi-criteria.

Furthermore, BARD supports automated generation of districting plans and profiling of plans by assigning different weights to various criteria, such as district compactness or equality of population. This functionality permits exploration of trade-offs among criteria. Districting is a computationally-intensive problem for even modest-sized states. Performance is thus an important consideration in BARD's design and implementation. The program implements performance enhancements such as evaluation caching and explicit memory management.

BARD fulfills jobs listed beow and shown in Figure 3.41;

1. First, BARD reads and processes districting data. BARD can read and write files representing districting plans in the standard ESRI shapefile format, permitting inter-operability with other GIS packages.
2. Second, BARD evaluates districting plans. BARD will generate textual and graphical reports for a single plan or comparison of multiple plans. Currently BARD shows areal differences between pairs of plans, counts 'holes'

(unclassified areas) in plans, computes common compactness scores (moment of inertia), calculates overall population deviation, checks for contiguity and calculates sum of squared error for homogeneity.

3. Third, BARD generates and refines districting plans. Plans can be automatically generated to use as starting points for further refinement, or evaluated in their own right. We provide a number of different procedures for automatically generating plans including plans for pure random generation of districts, random-walk based methods for generating contiguous equi-populous districts and both simple and weighted k-means based plan generation. Once generated (or provided), plans may be automatically refined using metaheuristics to meet chosen goals. The application of a metaheuristic to refine plans should yield a plan that is an improvement, given a chosen scoring formula. In this thesis, simulated annealing is used for metaheuristic districting.
4. Fourth, BARD compares multiple plans. BARD outputs the range of overall scores, the range of scores for each component, the differences among plans, and the correlations among score components.

Integration of BARD into an existing GIS system could significantly enhance ease of use for non-experts. For novice users, interface features such as wizards, accompanied by extensive help, and training examples would be valuable,

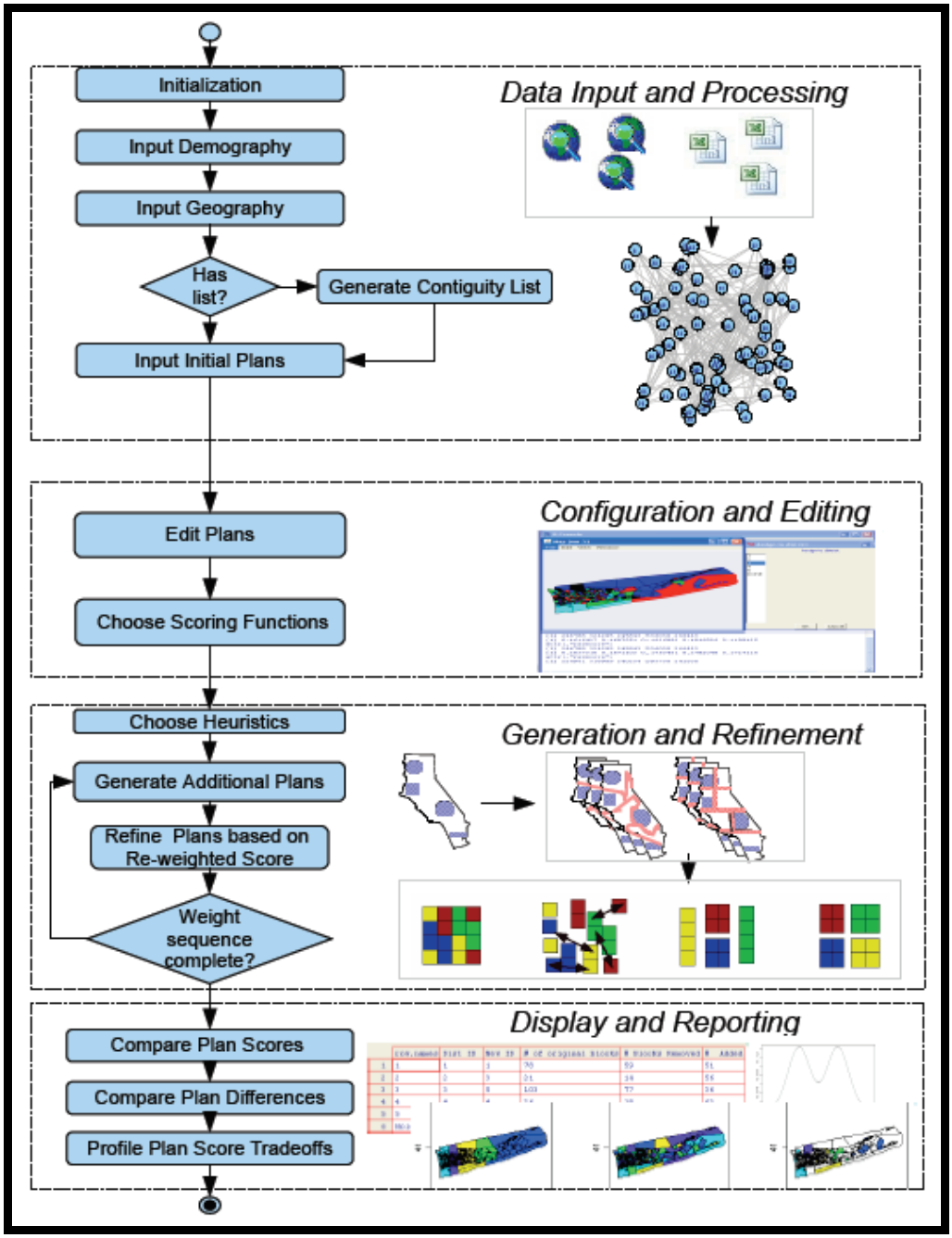


Figure 3.41 Phases of districting in BARD.

3.13.2. SOM Toolbox 2.0 on Matlab R2007b

SOM toolbox 2.0 contains freeware functions for the creation, visualization and the analysis of Teuvo Kohonen's Self-Organizing Maps which runs on Matlab environment. The first version of the toolbox was released in 1997.

The toolbox is closely related to SOM_PAK, a freeware software package implementing the SOM algorithm in C-code. The toolbox contains functions that uses SOM_PAK programs from Matlab and has functions for reading and writing data files in original SOM_PAK format. The freeware SOM program package SOM_PAK is sufficient, but it's not nearly as flexible as the MATLAB environment. So the Laboratory of Computer and Information Science at Helsinki University developed SOM Toolbox to offer a simple, well documented MATLAB function package which is easy to use and modify. The SOM_PAK files can also be accessed with the SOM toolbox, so it is possible to first train the map with the SOM_PAK and then use the SOM Toolbox for map visualization (Vesanto, Himberg et al, 2000).

Matlab environment is well-suited for SOM Toolbox because it has fast prototyping and customizing capabilities. Also, Matlab features a high programming language, powerful visualization, graphical user interface tools and a very efficient implementation of matrix calculus which are major advantages in the data mining research.

The SOM toolbox can be used for realizing jobs listed below;

1) Train SOM with different network topologies and learning parameters,
2) Compute different error, quality and measures for the SOM,
3) Visualize SOM using u-matrices, component planes, cluster color-coding and color linking between the SOM and other visualization methods,
4) Do correlation and cluster analysis with SOM.

The primary requirement of SOM toolbox 2.0 is to have Matlab software at least version 5.2. Second requirement is to have enough memory, because toolbox uses quite a lot of memory to speed iterations. The SOM toolbox 2.0 and related documents and reports can be downloaded freely from the web site of the Laboratory of Computer and Information Science cited at (Web 9).

3.14. Codes Developed in the Thesis

The codes developed in this thesis are related with retrieving the similarity values (unified distances) of each input data embedded on a SOM plane. As mentioned in section 3.12.2, SOM algorithm creates a plane that groups similar ones from input data in hexagons and each hexagon has an associated unified distance (similarity) value. These neurons (hexagons) have also a BMU number. This BMU number can be interpreted as the unique IDs of each hexagon on the u-matrix plane. In order to obtain the BMU of a particular hexagon and thus getting the unified distance value of similar input data in a particular neuron is important in order to visualize these similarity values on their actual geographical reference on the map. The functioning of the code which is developed by Matlab is explained in Chapter 5 en detail and codes are given in Appendix C.

CHAPTER 4

DESCRIPTION OF THE STUDY AREA AND DATA PROCESSING

Aronoff (1993) states that “A GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data: 1. input; 2. data management (data storage and retrieval); 3. manipulation and analysis; and 4. output”. Hence, data stands as a core component for GIS in order to assess our goal. For this reason, in this chapter, spatial and socio-economic structure of nine districts from Keçiören has been explored in order to present the current situation before conceptualizing clustering and districting methods based on exploratory heuristics and GIS. This chapter also aims to demonstrate how GIS can be a valuable tool for analyzing basic data and to construct a reference data and analysis set for the further clustering and districting studies relating to these nine districts.

4.1. Location of Case Study Area

Modern urban planning needs efficient descriptors of the distribution of socio-economic indicators spatially. Knowledge about residential patterns, distribution of services and socio-economic indicators can help the decision makers for future strategies of cities’ development. In order to facilitate this process, ”...clustering and redistricting of socio-economic indicators is a very efficient tool, because it allows the reduction of the information from a very high dimensional and complex input space to a low dimensional and visualizable output space” (Tuia et al, 2009). Thus, selection of the input space was the key for this study which reflected its outcomes on the input space. As an input space, 9 neighborhoods are chosen from Keçiören, Ankara and these neighborhoods can be seen in Figure 4.1 and Figure 4.2 in means of where they take place in Turkey (country), Ankara (city) and Keçiören (province) hierarchically and geographically. These neighborhoods are Ayvalı, Etlik, Aşağıeğlence, İncirli, Emrah, Basınevleri, Çiçekli, 19 Mayıs and Karargahtepe.

However, before justifying this selection, we need to examine their selection criteria which have been presented under the headings of urban development history, geographical and physical structure of study area and socio-economic characteristics of selected neighborhoods such as population, income and education.

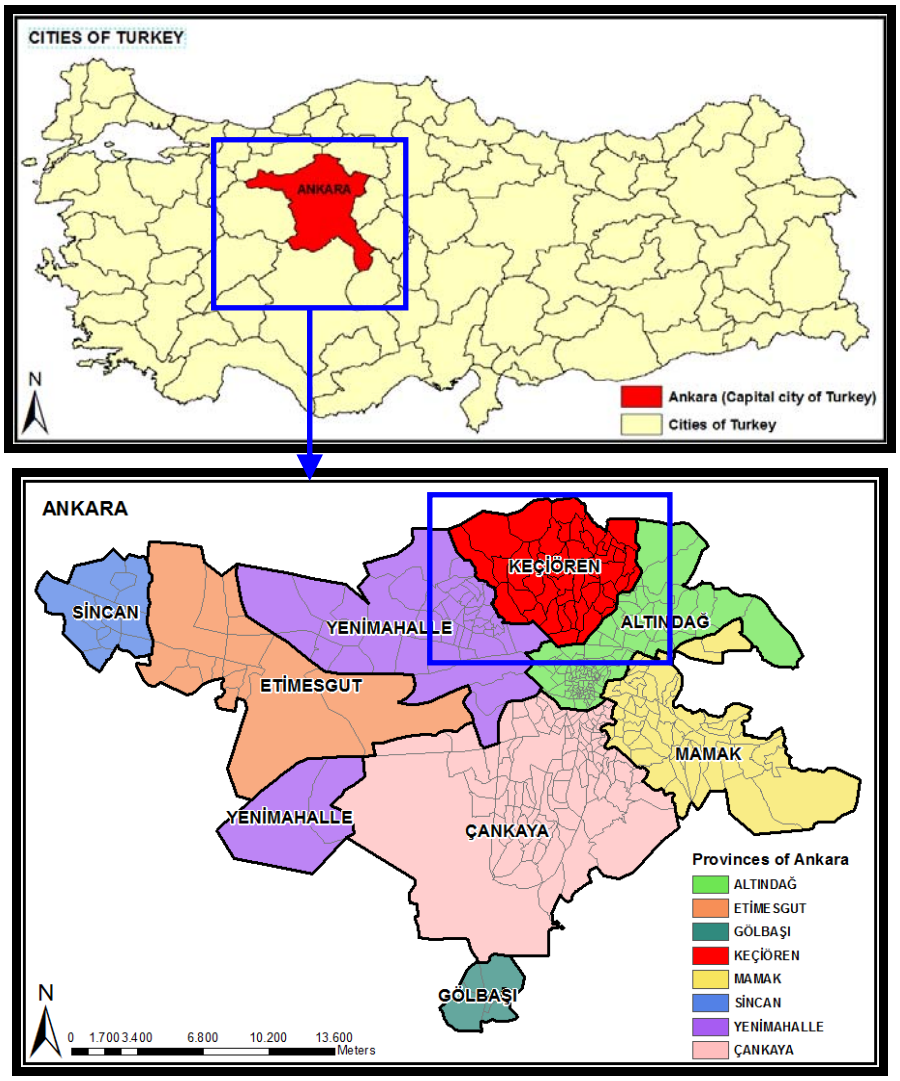


Figure 4.1 Location of case study neighborhoods in Turkey, Ankara

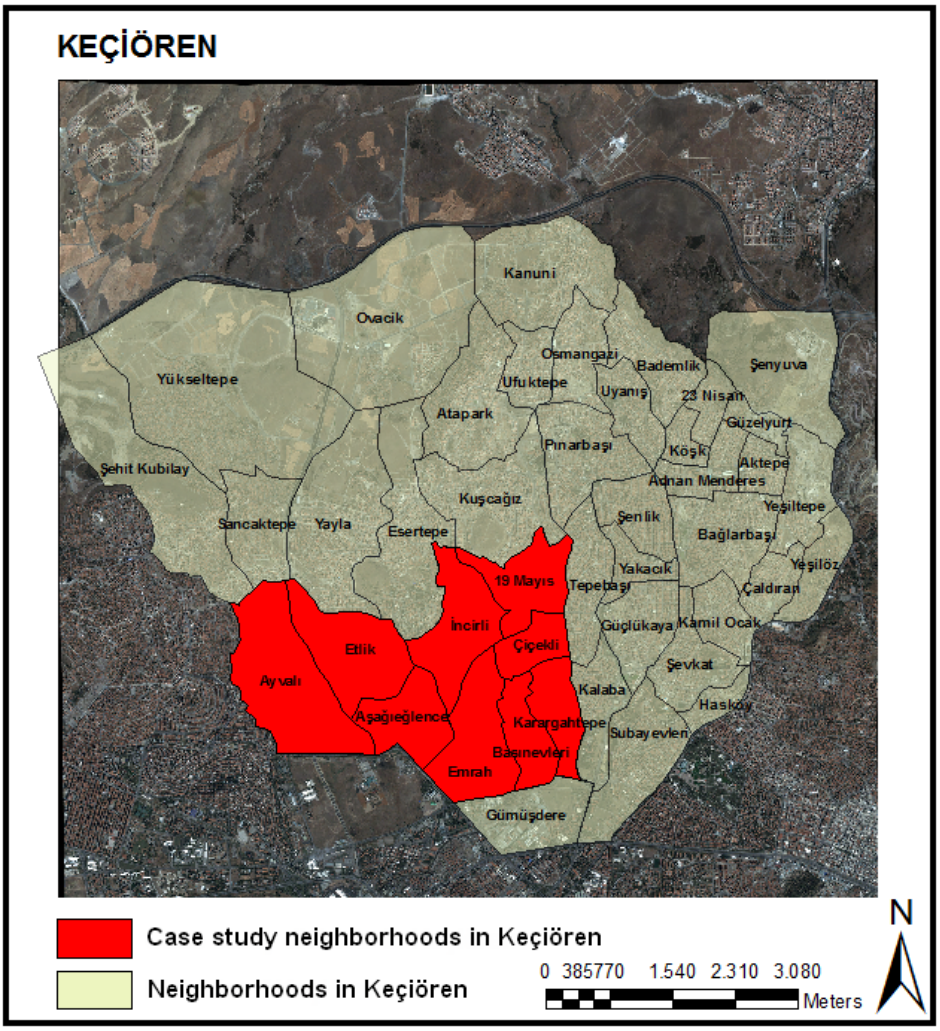


Figure 4.2 Location of case study neighborhoods in Keçiören.

4.2. Urban Development Stages of Study Area

The urban development stages of Keçiören province as a whole is studied by Kalaycıoğlu (2006) with referring to studies of Şenyapılı (2004 and 2005). According to Kalaycıoğlu's study, it is told that until 1940s, Keçiören was a recreation place in the North of Ankara where vineyards and orchards during summer were found. In Jansen Plan of 1932 for Ankara, it was proposed that this old vineyards and orchards and their typical design should be protected. Between years 1940-1950, Keçiören and Etlik were expanding as two separate areas on the sides of the city, and the transport to the city was provided by public city buses. Gradually the houses

in those gardens and vineyards were used both in winter and in summer. In those years frequent demands from Keçiören for construction were rejected because they did not fit into the Jansen Plan (Şenyapılı, 2004).

In 1940s, although there was an increase in the need for construction in order to provide housing need for immigrants from rural to urban area, no solutions to the problem of housing were found. So the squatter settlements have increased at the north fringes of Etlik by poor and unemployed people or by villagers who live at nearby villages who came city for working. In Ankara, urban development in places not covered by the plan took place in those squatter (gecekondu) areas. In 1950s, this situation worsened that there was also a serious rise in the number of squatters in Keçiören (Şenyapılı, 2004).

In 1950s, as a result of these developments in Keçiören's planned areas, the expansion of the city center and the squatter areas around the center, Keçiören became a settlement integrated to the city. In those years, seeing the inadequacy of Jansen Plan with regards to population increase in the city, a new plan was decided to be made. As a result, a new Ankara Plan was prepared in 1957 by Nihat Yücel – Rasit Uybadin, which proposed new settlement areas around Keçiören and a dense construction in the order of blocks and parcels in the region. In this plan; the density of net population proposed for Keçiören was 100 persons/ ha; for Etlik it was 245 persons/ha. On the other hand, no proposal was made for the increasing number of squatter settlements in Keçiören and Etlik areas.

In the year 1960s, the northern parts of Keçiören were full with squatter settlements. Especially, Emrah and Aşağı Eğlence neighbourhoods around Etlik were fast expanding places (Şenyapılı, 2004).

In the 1970s a new plan for Ankara, "1990 Ankara Structural Plan (1990 Ankara Nazım Planı)" was prepared. According to this plan, Etlik was considered to be a place where middle low income groups were settled (Ankara Metropoliten Alan Nazım Plan Bürosu, 1977). For some parts of İncirli improvement plans were proposed due to squatter settlements. On the other hand, in Aşağıeğlence, Etlik and

most part of the İncirli, there was a legal urban structure dominancy. Different from former plans, this plan proposed prevention zones in the squatter settlements in Keçiören. Later with the 1984 law, Keçiören was declared as one of the 8 neighborhood municipalities connected to the Ankara Metropolitan Municipality. From this year onwards district municipalities started to make 1/1000 scale Construction Improvement Plans (Islah İmar Planları) for the gecekondu areas. In Figure 4.3, northern part of Ayvalı, Etlik and İncirli, west part of the Ondokuz Mayıs neighborhoods and many parts in Basınevleri is proposed for construction improvement plans.

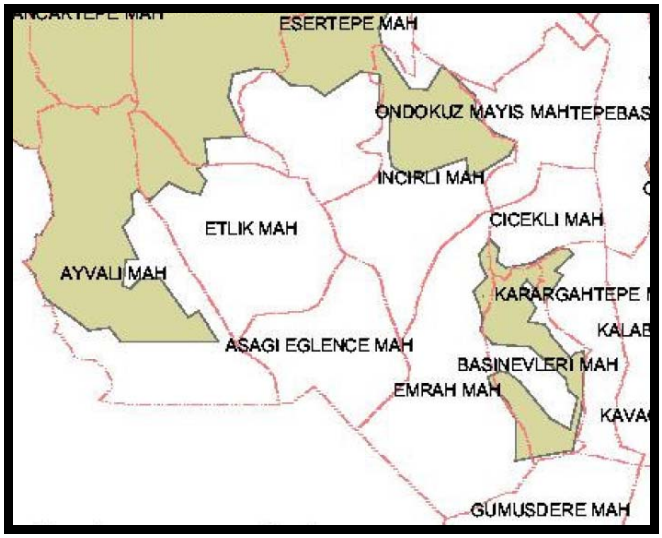


Figure 4.3 Places with construction improvement plans in study area at 1984 (Kalaycıoğlu, 2004).

These plans were completed for all of Keçiören in the year 2004 without any upper-scale plans of the district. However, in case study area with referring to the 2008 IKONOS satellite imagery, the west part of Ayvalı, north of İncirli and some parts of Basınevleri still needs to be improved. With improvement plans, building density of the settlements have been increased to 4 floor apartment buildings and social infrastructures like education and health facilities were proposed in the periphery where once squatter settlements took place. Present physical structure/form of

Keçiören is mostly a result of “Construction” and “Construction Improvement” plans which were made in the last 14 years.

4.2.1. Characteristics of the Study Area

Ankara is the capital city of Turkey and Keçiören is the most populated province of Ankara with 190 km^2 area. Keçiören is located on the northern part of Ankara. Districts neighbouring Keçiören are Altındağ from the east and southeast, Yenimahalle from the South and the West, Kazan from the north-east and Çubuk from the North. The distance from Keçiören to the Ankara city center is approximately 5 km. At present, Keçiören has 43 neighborhoods in municipal border. Study area is located on the south-west part of Keçiören (Figure 4.1).

In Keçiören, there are 136 public and 27 private elementary / high schools. 67 of these schools reside in the study area. In terms of higher education within the study area, Gülhane Military Medicine Academy covers a large area on the southern part in Emrah district. Gülhane Military Medicine Academy is also one of the two significant hospitals in Keçiören (Figure 4.3).

Apart from the health and education centers other outstanding units in study area are commercial centers. Two of these centers are: Anteras and Metro market on south of Ayvalı. Also, Aşağıeğlence and south parts of Etlik are the commercial centers of Keçiören (Figure 4.4).

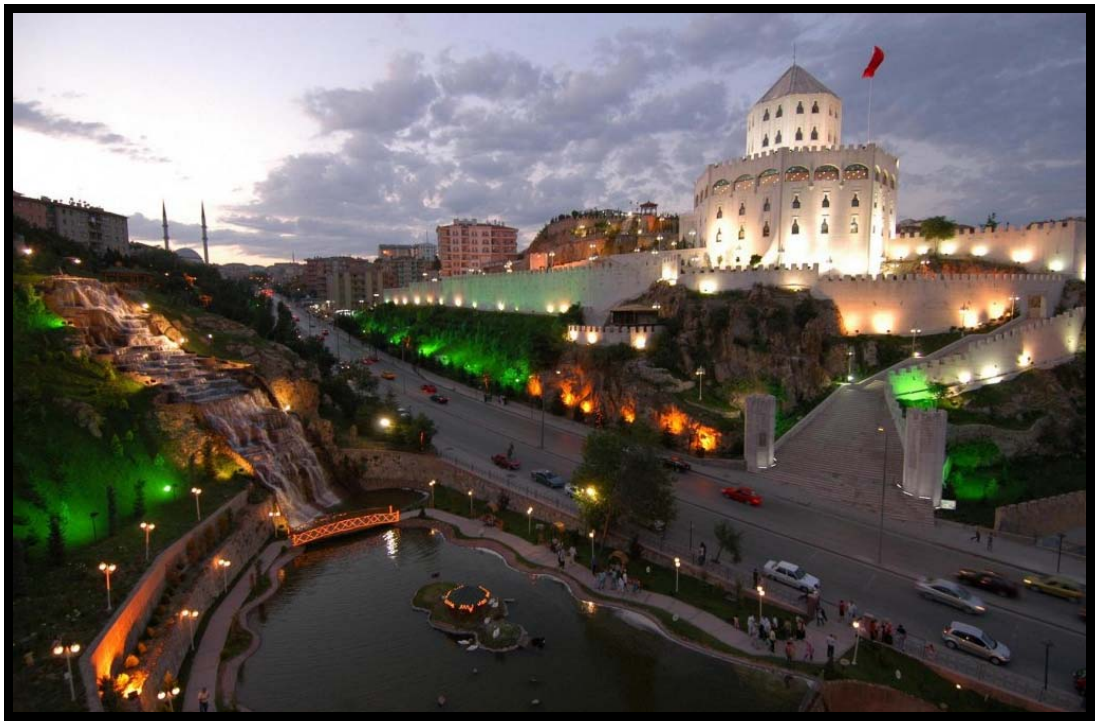


Figure 4.4 A View from Keçiören (Kalaba, Güçlükaya and Çiçekli Neighbourhoods in 2008) (Web 4)

The reasons for selecting neighborhoods from Keçiören as astudy site are based on three pillars;

1) Firstly, it is one of the oldest districts in Ankara where the population increase has risen much more than expected and at present it is the district with the highest population density in Ankara according to Address Based Population Register System (Web 11).

2) Hence, with respect to variables which are useful to understand socio-economic status in terms of level of education, level of income and income source there is great diversity within Keçiören. Besides these socio-economic characteristics, it can be said that Keçiören, especially the neighborhoods selected as study area offers variation in terms of housing and land-use pattern. It can be claimed that these characteristics of study area make it a suitable site for a representative study for constructing SSAs.

3) Thirdly, when urban development and its stages in Keçiören are considered, it can be said that the district became a residential place mainly for middle

income groups. However, poor groups also settled in this district since there are still squatter settlements, especially in the study area. Additionally, due to the Construction Improvement Plans in the last ten years in Keçiören, it is now difficult to assess the physical conditions of the settlement areas. Due to these reasons, Keçiören carries a priority rather than other districts of Ankara, for studies to be made in order to construct compact, homogenous and equipopulous SSAs based on both socio-economic and physical attributes.

4.3. Data Collection

Thesis database, storing information about these 9 neighborhoods is constructed by GIS. This study needs both graphical and non-graphical data about the study area. Graphical data are necessary to understand the physical features of the area like urban pattern, topography, transportation, etc. Table 4.1 gives information about the collected graphical raw data

Non-graphical data are as important as graphical data, and indispensable for understanding the population characteristics like demography and socio-economic situation. Sources non-graphical datasets, used in this thesis, are given Table 4.2.

Table 4.1 Graphical Raw Datasets

Class	Dataset	Data	Date	Type	Format	Source	Used Content
Graphical Raw Datasets	**Infrastructure Information System of Ankara (AYBIS)**	Buildings	2000	Polygon	Mapinfo (*.tab)	MMA-AMMOWI[1]	Building outdoor numbers
		Roads and Road names	2000	Polyline / Annotation Group	Mapinfo (*.tab)	MMA-AMMOWI	Roads and Road names
		Landmarks	2000	Point	Mapinfo (*.tab)	MMA-AMMOWI	Hospitals and clinics, post offices, police stations, bazaars, schools, mosques, government offices, parks
		Provinces	2000	Polygon	Mapinfo (*.tab)	MMA-AMMOWI	Province boundaries
		Districts	2000	Polygon	Mapinfo (*.tab)	MMA-AMMOWI	District boundaries
		Parcels	2000	Polygon	Mapinfo (*.tab)	MMA-AMMOWI	Current situation of parcelsand parcel IDs in 2000
		Blocks	2000	Polygon	Mapinfo (*.tab)	MMA-AMMOWI	Current situation of blocks and block IDs in 2000
		Digital Elevation Model (DEM)	2000	Image	(*.bmp)	MMA-AMMOWI	Nearest neighbor sampling of elevation points
		Hillshade	2000	Image	(*.tif)	MMA-AMMOWI	Relief
	1/1000 scaled basemaps of covering the study area	CAD based maps	2000	polygon	Netcad (*.ncz)	Keçiören Municipality[2]	Buildings
	Satellite Image	IKONOS Satellite Image of the study area	2008	ECW (Enhanced Wavelet Compression)	(*.ecw)	Keçiören Municipality	Pan-sharpened multispectral (RGB) imagery with 1meter/pixel
	1/1000 scaled development plans covering the study area	CAD based maps	2008	polygon	Netcad (*.ncz)	Keçiören Municipality	Very detailed development plans of the area covering all types of landuse components
	Center lines of roads	Roads	2008	Polyline	ESRI (*.shp)	Keçiören Municipality	Roads with the their latest names and constant presentation numbers in ABPRS

[1] Retrieved from MMA-AMMOWI: Metropolitan Municipality of Ankara- Ankara Metropolitan Municipality Office of Water and Infrastructure.

[2] Retrieved from Keçiören Municipality – Department of City Planning and Development Department.

Table 4.2 Non-graphical Raw Data

Class	Dataset	Data	Date	Type	Format	Source	Used Content
Non-graphical Raw Data	Address-Based Population Register System (ABPRS) Database	Adresses and socio-economic indicators	2008	Sheet	SAS files (*.egp)	Turkish Statistical Institute (TURKSTAT)	Addresses with road, street names, building names, type of building, outdoor and indoor numbers; Socio-economic indicators for the head of household.
		Number of buildings according to its type	2008	Sheet	SAS files (*.egp)	TURKSTAT	For Ankara and Keçiören
		Keçiören road/street names according to ABPRS	2008	Sheet	SAS files (*.egp)	TURKSTAT	New and old names of roads/streets with a constant presentation number.
	Population statistics	Population number and density	1990 2000 2008	Sheet	Excel (*.xls)	TURKSTAT	Population numbers and densities for provinces in Ankara

4.4. Data Preparation

In order to construct SSAs based on census information, necessary and appropriate data are collected, transferred and manipulated in the GIS environment. Thus, GIS consists of databases and maps which are linked to each other. This means that; the data for this study includes both tabular data as databases and graphical data as maps.

4.4.1. Establishment of an Address and Socio-Economic Indicators Attribute Database by ABPRS

In order to find SSAs in an area, firstly, one needs to have basic complete granular database which consists of addresses of each building for geocoding them on the map. Secondly, as each SSA must be homogenous and equipopulous in terms of socio-economic indicators and population, census information is needed. For these reasons, the ABPRS datasets form the bottleneck of this thesis.

The purposes of the “Address Based Population Registration System” study were;

- To establish “National Address Database” that will store all address information in Turkey,
- To obtain personal information of Turkish citizens and foreigners residing in the standardized addresses that were defined in the “National Address Database”,
- To match usual residence addresses and the census registers of the General Directorate of Population and Citizenship Affairs (GDPCA) by using the Turkish Republic identification numbers.

The study finished at July 2007.

In NPE (Nation-wide Population-Enumeration Studies), it is obligatory to define the boundaries of registry regions which are defined over maps. But in Turkey, because of the non-existence of up-to-date and trusted maps to base the enumeration, the registry region definitions is being done over lists which addresses take place. In ABPRS study mainly two forms were used;

- Address forms where municipality exists,
- Address forms where municipality does not exist.

Since the study area is governed by Keçiören municipality, the address records are used which were gathered by the address forms where municipality exists. With Address Forms used where municipality exists, it was aimed to keep addresses (indoor, outdoor number) of all units (residences, mosques, barracks, garages, hospitals, workplaces, haylofts, barns, warehouses, parcels, constructions, etc..), to keep characteristics of all units (residence, personal workplace, public workplace, construction, parcel, etc...), to keep the number of person living or working at these units and if there is another opening door of the unit to a road and street, the number of this door’s number is taken under the name “outdoor number 2”.

In the first part of this form Table 4.3-a, the location of the dwelling unit is recorded in terms of city (province), district, sub-district, municipality, village, neighborhood (quarter) and locality (mevki) names. In second part of form (Table 4.3-b), the type and name of the road, street, public square, boulevard or cluster is checked and

written. Besides this information, the development level of the street, road, boulevard or Public Square was wanted to be evaluated by the surveyor in terms of developed, semi-developed and under-developed. But, this evaluation was up to the subjective opinion of the surveyor, so, this information is not accepted as an accurate indicator for this thesis. In the third section of form (Table 4.3-c), the address of the dwelling unit is recorded by taking the outdoor number of the building it belongs, indoor number and building or housing complex name into account. The third section is very important because in this section the characteristic (type) of the dwelling unit is recorded. For example, if the dwelling unit is an occupied resident, it is written down and code "2" is given. Or, if the dwelling unit is under construction, it is recorded and code "4" is given. The last section (Table 4.3-d) is especially added for the buildings which are located at the junction of two streets or roads. Because, it was observed that the systematic errors at numbering were usually resulting from these kinds of buildings. Sometimes, if there is an opening door to the street on both sides of the building, they would have been ignored or twice recorded by surveyors.

Table 4.3 Sections of address form that is used where municipality exists (Original of this form can be found in Appendix B).

a)

T.R.
PRIME MINISTRY
TURKSTAT

Province Name: .....

District Name: .....

Subdistrict Name: .....

Municipality Name: .....

Village Name: .....

Quarter Name: .....

CONSTANT PUBLICITY NUMBER: .....

Locality Name: .....

Table 4.3 (Continued) Sections of address form that is used where municipality exists (Original of this form can be found in Appendix B).

b)

2006

ADDRESS FORM

(Municipality Existent)

PLEASE WRITE DOWN THE NAME OF SQUARE/AVENUE/ROAD/STREET OR CLUSTER FOR WHICH THE INFORMATION FILLED BELOW BELONGS TO. MARK THE TYPE OF PLACE AND THE DEVELOPMENT LEVEL AS ONE OF THE MULTIPLE-CHOICE GIVEN. (DO NOT WRITE THE EXTENSIONS SUCH AS STR. FOR STREET, SQ. FOR SQUARE, RD. FOR ROAD ETC.)

TYPE: SQUARE ☐ 1 BOULEVARD ☐ 2 AVENUE ☐ 3 STREET ☐ 4 CLUSTER ☐ 5

DEVELOPMENT LEVEL DEVELOPED ☐ 1 SEMI-DEVELOPED ☐ 2 UNDER-DEVELOPED ☐ 3

SQUARE/ AVENUE/ ROAD/ STREET/CLUSTER CONSTANT PUBLICITY NUMBER ☐☐☐☐☐☐

c)

ROW NUMBER	ADDRESS'				CHARACTERISTIC OF NUMBERED SITE	CHARACTERISTIC CODE
	OUTDOOR NUMBER (IF IT IS A BUILDING PLOT WRITE DOWN THE ALLOTMENT NUMBER)	INDOOR NUMBER (DEPENDENT ON OUTDOOR NUMBER)	HOUSING DEVELOPMENT NAME	BLOCK NAME	Please write down the usage aim of the site clearly such as; residence, grocery, butcher, hairdresser, school, government, mosque, hospital, hayloft, shed, warehouse, garage, lot, allotment. If site is unemployed write unemployed residence or unemployed office.	1.Residence 2.Private workplace 3.Official workplace 4. Construction 5. Building Plot 6. Allotment 7. Cottage 8. Building plots reserved for public improvement 9. Other
	1	2	3	4	5	6

d)

IF BUILDING HAS ANOTHER DOOR WHICH OPENS OUT INTO SQUARE/AVENUE/ROAD/STREET, FILL IN THE BLANKS BELOW			CITY BLOCK/SECTION (ON MAP) /PARCEL			ADDITIONAL/ SUPPLEMENTARY NOTES		
Write down the name of Square, Avenue, Road or Street	Constant Publicity Number	Door Number	City Block No.	Section No.	Parcel No.	NOTE 1	NOTE 2	NOTE 3
7	8	9	10	11	12	13	14	15

The main purpose of ABPRS was to establish “National Address Database” that will store all address information of dwelling units in Turkey. Besides the address forms which were used for collecting addresses and types of dwelling units, personal information of Turkish citizens and foreigners residing in these dwelling units were also collected with the forms called “ABPRS Household Information Form”

(Appendix B). Both forms are filled simultaneously with the aim to collect compatible information related to dwelling unit addresses and the households residing in these dwelling units. Even information about all the members in one dwelling unit, in other words household, is registered in ABPRS, in this thesis the socio-economic information related to the head of household is taken into account.

The household forms and address forms had a common identity code which was given to each and every unique address. So, in the first part of household forms this identity code is pasted in order to easily join address and household information. The socio-economic indicators in this thesis are gathered from these household records. In household forms income level, income source and education level of each member of household is collected. The questions related to these indicators are written below;

a) What is the average monthly income of this household?

1) 0 - 150 YTL **2)** 151 - 350 YTL **3)** 351 - 500 YTL
4) 501 - 1000 YTL **5)** 1001 YTL and more **9)** Unknown

b) What is the income source of this household?

1) Steady income like salary yor payment
2) Income from commercial, labor intensive, industrial, agricultural sources or real estate and movable goods
3) Social welfare

c) What is your education level (head of household)?

1) Illiterate
2) Literate but no schooling
3) Primary school
4) Primary education
5) Secondary school
6) High school
7) University
8) Master's degree
9) Doctor's degree

A sample record of National Address Database filled with artificial data can be seen in Table 4.4. A sample record of household information related to the head of household in each dwelling unit can be seen in Table 4.5 again filled with artificial data.

Table 4.4 Structure of ABPRS database.

Columns in ABPRS Database	Possible Content of Column
Unique Address Code	12543981
City Name	ANKARA
City Plate Code	06
District Name	KEÇİÖREN
District Code	05
District Register Code	60
Subdistrict Name	CENTER
Subdistrict Code	0
Subdistrict Register Code	107
Village Name	CENTER
Village Code	0
Village Register Code	2442
Sub-level Settlement Name	CENTER
Sub-level Code	0
Sub-level Register Code	35271
Neighborhood Name	AŞAĞI EĞLENCE
Neighborhood Code	1686
Neighborhood Type Code	1
Neighborhood Type	Municipal Neighborhood
Neighborhood Identification Code	14
Road, Street, Boulevard or Square Name	Etlik
Road, Street, Boulevard or Square Code	366319
Road, Street, Boulevard or Square Identification Code	1
Road, Street, Boulevard or Square Type Code	3
Road, Street, Boulevard or Square Type	Road
Building or Housing Complex Name	Yeşilyurt

Table 4.4 (Continued) Structure of ABPRS database.

Outdoor Number	61
Outdoor Number 2	42
Building or Housing Complex Type	1
Indoor Number	12
Type Code of Dwelling Unit	2 (Occupied Resident)
Type of Housing Dwelling Unit	Occupied Resident

Table 4.5 A sample record from ABPRS Household Survey.

Socio-Economic Indicator	Possible Content
Unique Address Code	12543981
Proximity of Individual	Head of household
İncome Level	3 (351 – 500 YTL)
Income Source	1 (Steady income like salary yor payment)
Education level of Individual	6 (High School)

The records related to addresses and household socio-economic status which are retrieved from TURKSTAT were having "*.sas" estensions which is the file format of SAS statistical software. So, these files were converted to ".dbf" file format which is acceptable by many software today. The two files are joined one-to-one by using the unique address code as the common column in MS Office Access environment. By this way, address and indicator database is obtained for each head of household in each dwelling unit. According to world-wide applications explained in Chapter 2 and recommendations of UN (2009), the geocoding of addresses are done by using the addresses of buildings, not by pinpointing all addresses of independent dwelling units that an apartment or building have. There were 65563 independent dwelling unit addresses in 9 neighborhoods. For this reason, in each building, the frequencies of socio-economic indicators are counted and recorded on a separate database in MS Office Excel environment for each head of household. For example, in an apartment called Akka, illiterate head of households are counted and recorded under a column named "E1", or head of households which have a steady income are counted and recorded under a column named "S1".

The characteristic of each dwelling unit in terms of its usage is also taken into account and for each building the major characteristic of that building is recorded on the database. For example, if in a building there were 3 occupied residential units and 1 commercial unit, 3 is written to “L2” column and 1 is written to its “”L4” column. But the main characteristic code is given as 1 which indicates that this building is used majorly for residential purposes (Table 4.6).

Also, in this database a unique code has been given to each building by using the abbreviation of study neighborhoods by adding a new column to file. For example, if the building was in Ondokuz Mayıs neighborhood, the building is coded as “OD19”, or if the building was in Ayvalı, “A20” is given to that building. The coding of buildings was compulsory for the further parts of this study where the maps and the tabular data are related via these codes. ABPRS database as a whole lacks building codes which is a major deficiency.

As a result of this process, we came up with an attribute database with 6323 records, in other words buildings, whose table structure is depicted in Table 4.6.

Table 4.6 Physical and Socio-economic attributes database for each building.

Variable Construct		Variable Code	Variable Definition
Unique Building Code		CODE	Unique building code given to each building in this thesis.
Charasteristics Code of building		CHAR_C	Code giving the type of the building in general.
Physical Construct	Type of divisible units	L1	Number of empty residential units
		L2	Number of occupied residential units
		L4	Number of commercial units
		L11	Number of units that give public services like schools, religious centers, police stations, military establishments, municipalities, hospitals or official buildings etc.
		L20	Number of divisible units that is on construction process
		L22	Number of other units that is not classified

Table 4.6 (Continued) Physical and Socio-economic attributes database for each building.

Socio-economic Construct	Education	E1	Number of illiterate person
		E2	Number of literate person that did not go to any school
		E3	Number of person that is graduated from primary school
		E4	Number of person that is graduated from primary education
		E5	Number of person that is graduated from secondary school
		E6	Number of person that is graduated from high school
		E7	Number of person that is graduated from university
		E8	Number of person that have a masters degree
		E9	Number of person that have a doctors degree
		E10	Number of person whose education level
	Income Source	S1	Number of person who obtains a steady income like salary or payment
		S2	Number of person who obtains income from commercial, labor intensive, industrial, agricultural sources or real estate and movable goods
		S3	Number of person who lives with social welfare
		S12	Number of person who obtains both S1 and S2
		S123	Number of person who obtains both S1, S2 and S3
		S13	Number of person who obtains both S1and S3
		S23	Number of person who obtains both S2 and S3
		S9	Number of person whose source of income is unknown
	Income Level	A1	Number of person whose income is between 0 and 150 TL (Turkish Lira)
		A2	Number of person whose income is between 151 and 350 TL
		A3	Number of person whose income is between 351 and 500 TL
		A4	Number of person whose income is between 501 and 1000 TL
		A5	Number of person whose income is 1001 and more
		A9	Number of person whose amount of income is unknown

4.4.2. Establishing an Up-to-Date Geocoded Address Database for the Study Area

After establishing the socio-economic indicators table database for the buildings in study neighborhoods, in order to join each record/building in this database with its geographical equivalent, a building layer was constructed in ESRI ArcGIS 9.2 environment. The building layer which was obtained from Infrastructure Information System of Ankara (AYBIS) was constructed in year 2000, thus it was not up-to date. As being the most populated district in Ankara, Keçiören has a high urban development potential. Also, as a result of construction improvement plans many squatter areas in Keçiören are transformed into regular settlement formation. Thus, using up-to-date (2008) IKONOS satellite imagery (resolution: 1m/pixel) which was

retrieved from Keçiören municipality, the building polygon layer retrieved from AYBIS is updated by adding and digitizing non-existent buildings or by deleting the building polygons which were replaced by new ones (Figure 4.5). All the spatial layers used in this thesis have the same projection system which is Universal Transverse Mercator / European Datum 1950.

In order to match address information from ABPRS with the re-digitized building polygon layer, a digital road/street layer was needed which can be used as a reference. This necessity results from the fact that in a standard address, street/ road/ avenue/ boulevard name is compulsory to find a building. By ABPRS study, in order to prevent duplication of street or road name in a municipal border, some of the roads' and streets' name has been changed. Also, a unique identification code is also given to all of the streets or roads as to prevent the complexities between old and new names. For these reasons, the road layer which would be used in this study had to involve both the old and new names of streets or roads, because in ABPRS, the streets or road names in addresses are kept with their newer names. A polygon road layer in this format is provided from Keçiören Municipality which included the old name, new name and ABPRS identification code of the road and showed the centerlines of the roads (Figure 4.6).

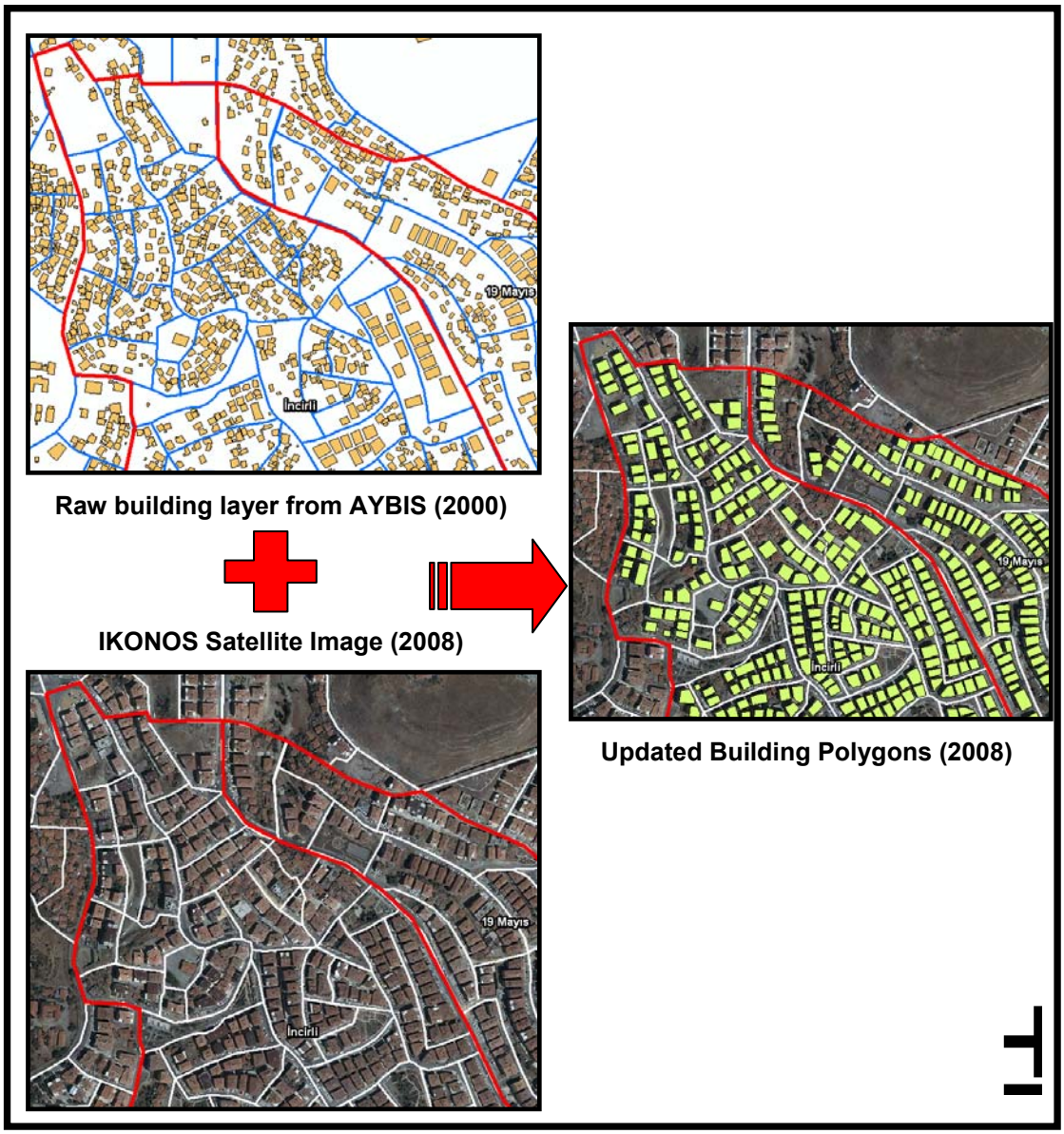


Figure 4.5 Creating an up-to-date building polygon layer by using satellite image of the study area.

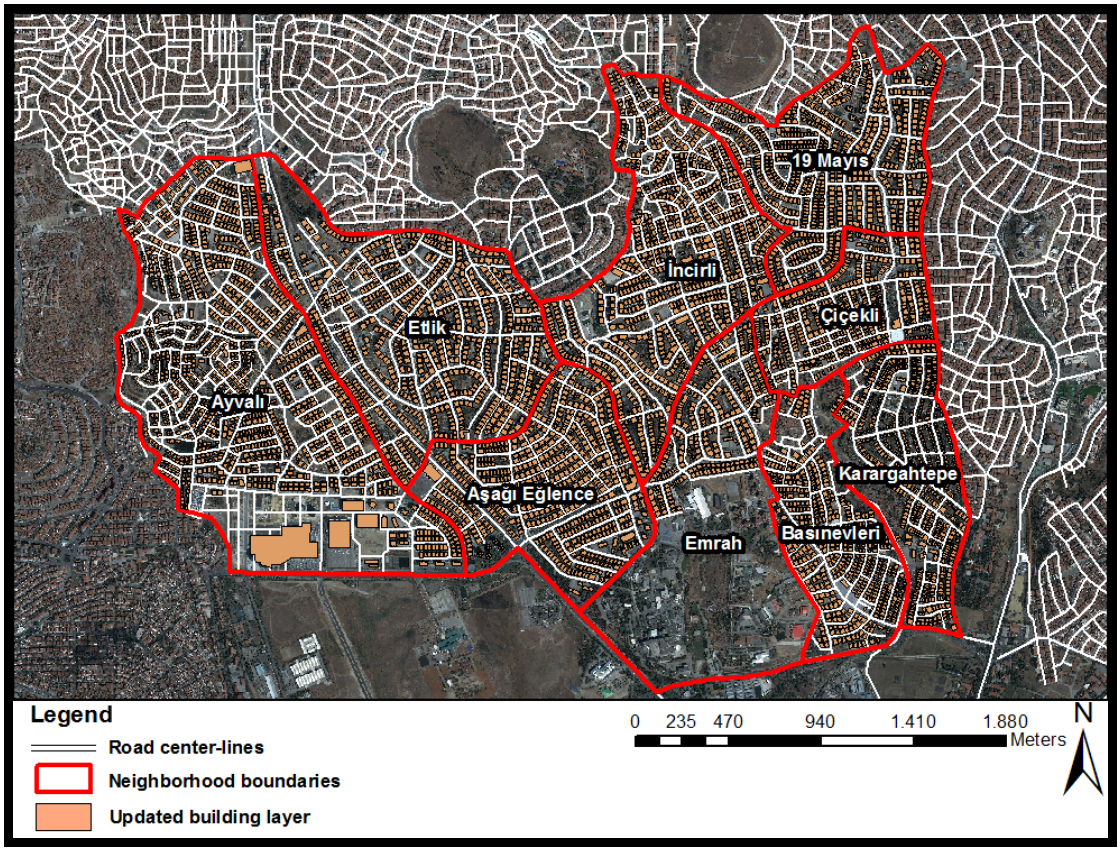


Figure 4.6 The road center-lines layer including the old and new names of streets or roads in ABPRS.

However, to geocode the addresses, taking the roads as a reference was not enough. One should also know the outdoor numbers of building polygons. Since, the study area was very big to make a field survey to collect each and every building's outdoor number and to mark them on the map, in this thesis, the outdoor numbers of the building polygons which was contained in the AYBIS building layer's attribute table (2000) are taken as a reference (Figure 4.7). But before doing this, for a small part of the study area, the accuracy of these outdoor numbers are checked and found that all of the outdoor numbers except for the areas that were newly built match 100%. For the newly built buildings, the outdoor numbers are continued from the last known outdoor number that matches.

Figure 4.7 Building layer from AYBIS database labeled with building outdoor numbers.

By using both the road centerlines and building outdoor numbers from AYBIS database, the 6323 address in ABPRS database are matched with the updated building polygon layer. In this process, the unique codes that were given to each building by using the abbreviations of neighborhoods, while establishing the address and socio-economic indicators database, were also recorded on a separate column. This column is used as the common column to join the geocoded buildings and the socio-economic indicators database that was established in section 4.4.1. So, as a result of this process, a new building layer is obtained in which every building has an address and each building is populated with the socio-economic and physical data which was including the type of the buildings and income level, income source and education level frequency counts. The completion of data manipulation phase that is explained in sections 4.4.1 and 4.4.2 took six and a half months.

4.4.3. Establishment of a Cadastral Infrastructure for SSA Layer Construction

The building layer geocoded and populated with socio-economic and physical attributes comprised of non-contiguous polygons, which was not an accepted data structure for SSA construction. Also, a building layer does not involve the areas where there is no building formation on them. So, in order to create a seamless and contiguous surface which can be used to create SSAs, parcel layer from AYBIS database (2000) and 1/1000 scaled development plans covering the study area which is retrieved from Keçiören municipality (2008) is used to create an updated parcel layer for the study area (Figure 4.8).

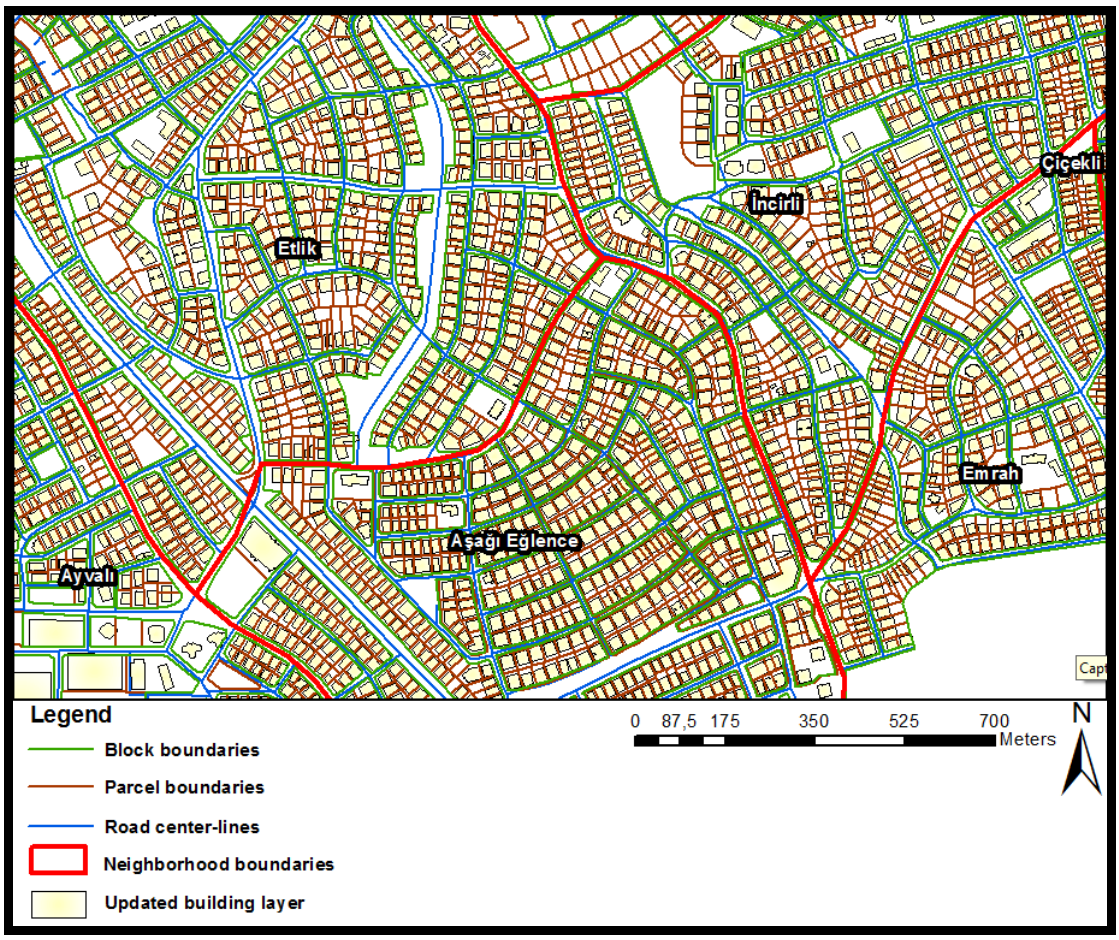


Figure 4.8 Parcel and Block boundaries retrieved from AYBIS (2000) and Keçiören Municipality

However, the parcel boundaries obtained from AYBIS and Keçiören municipality had gaps especially all over the roads. So, a parcel layer has been re-digitized by extending the parcel boundaries to the road center-lines (Figure 4.9). By this way, contiguous surfaces of parcels are generated. There are 6858 parcels in the study area.

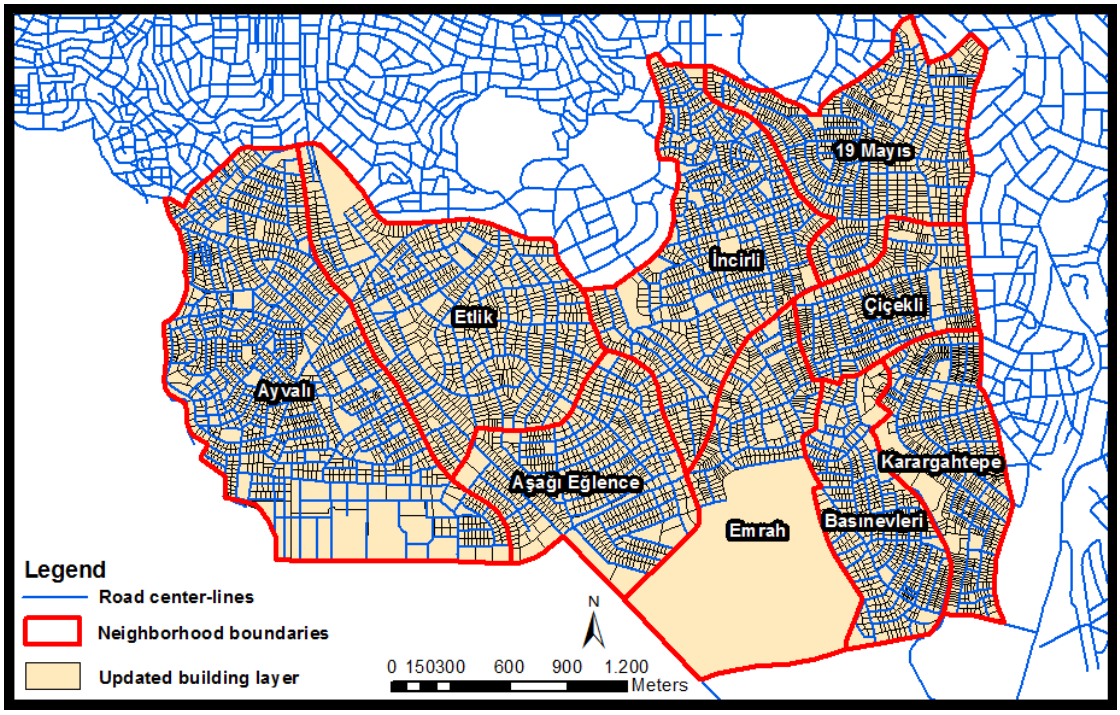


Figure 4.9 Re-digitized contiguous parcel boundaries.

In order to transfer the attribute information of buildings to these parcels a special method is used. Firstly, the building polygons are converted to points by using "Feature to Point" tool by ArcToolbox in ArcGIS 9.2. Then, "Spatial Join" is implemented between the building-point and parcel layer by using the latitude and longitude information that building points have. In this method, we use the location as common information, not a common column in the attribute tables of two layers. As a result of this process, a parcel layer is retrieved which has the same attributes such as addresses, physical attributes and socio-economic indicators that the building layer had. The problem with the parcel layer is that some parcels could not be spatially joined with building points because there was no building formation

upon these parcels, but still these parcels had some physical attributes. For example, some of them were public parks, empty parcels, construction sites or urban renewal areas. These parcels' physical attributes relating to their landuse characteristics are then updated by referring to AYBIS database and 1/1000 scaled development plans covering the study area. A land use map is created by taking the type characteristics of the parcels which can be seen in Figure 4.10.

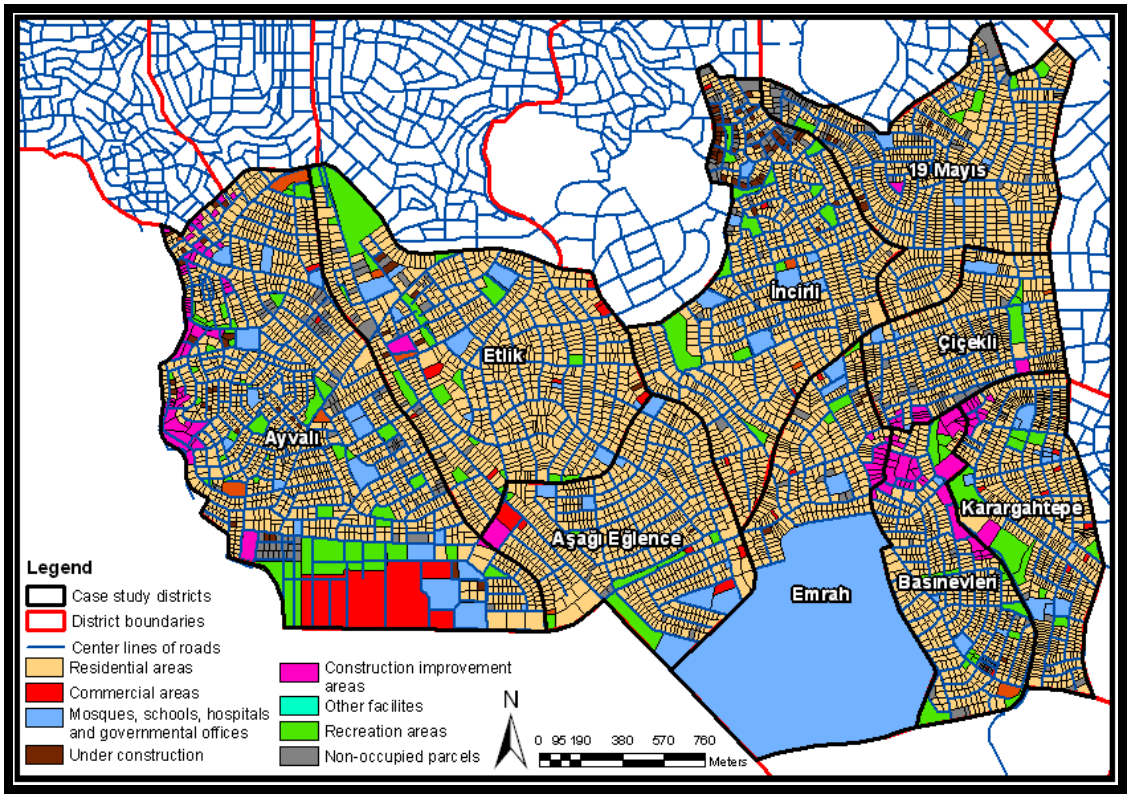


Figure 4.10 Land use map of case study area (digitized in this study).

Besides using parcels, blocks are also thought to be useful for creating SSAs. Because the BARD package which is used for redistricting process does not take road layer into account while clustering the parcels according to SSA creation rule which denotes that an SSA boundaries generally follow permanent, visible features, such as streets, roads, highways, rivers, canals, or railroads. So, blocks are also created by "Dissolve" tool in ArcToolbox in ArcGIS 9.2 (Figure 4.11). The important point in creating blocks was to take the sum of the frequencies of socio-economic or physical indicators that each parcel has. For example, in block "A" the number of

head of households is counted and summed, or the number of occupied residential dwelling units in each parcel are counted and summed. 711 blocks are created as a result of this process.

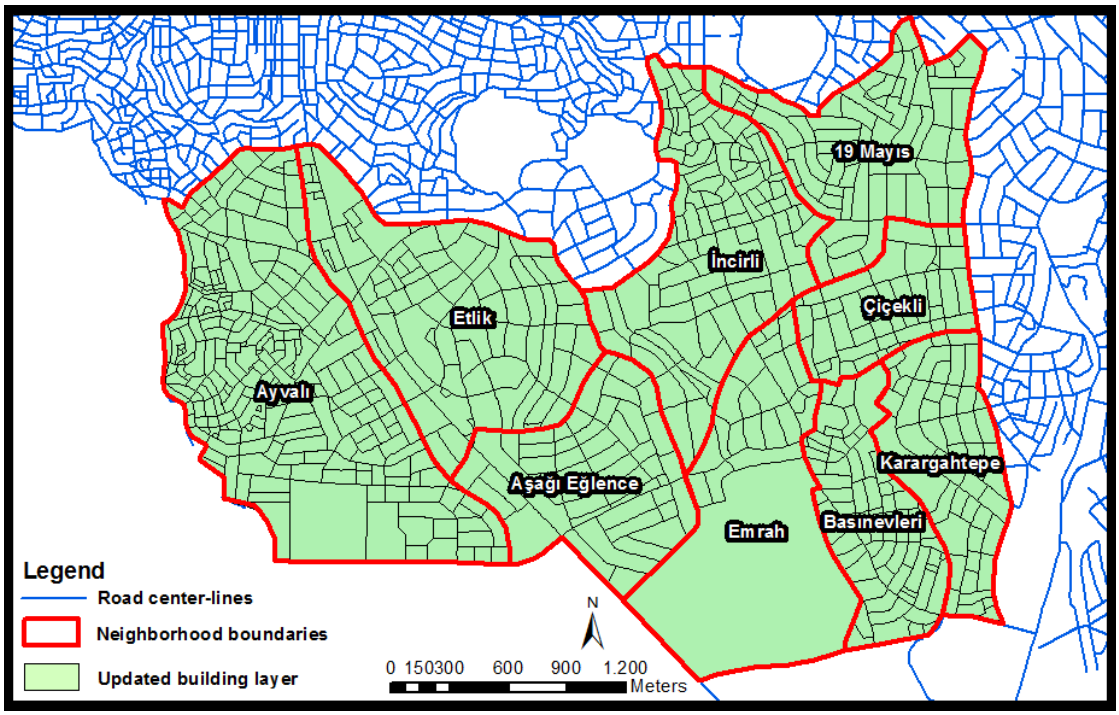


Figure 4.11 Block layer created for construction of SSAs.

There were some digitizing errors derived while creating blocks and parcels, like dangles, switchbacks, slivers, knots, loops, under and overshoots etc. in the ‘blocks’ and “pacels” map. They had to be corrected in order to use these blocks or parcels as the basic units of an SSA layer. Firstly, shapefiles of parcels and blocks was converted to geodatabase format to be able to make topology correction operations. Secondly, topology was built for the feature classes by using the ‘Create new topology’ tool of geodatabase format. Third step was to correct the errors in the coverage by topology validation. After validation, all of the errors were eliminated.

The parcel layer and block layer created in this chapter are used as the spatial and tabular inputs for the automated creation of SSAs in the study neighborhoods.

CHAPTER 5

EVALUATION OF CLUSTERING ALGORITHMS FOR THE CASE STUDY

This chapter deals with the main issue of the thesis, which is the creation of small statistical areas (SSAs) that builds up the whole census geography from bottom to top on the pilot study neighborhoods located in Keçiören. In order to achieve this aim, principal component analysis (PCA) is implemented to derive a single socio-economic status index for each parcel and block before migrating to the application of clustering algorithms. Then, as for the first method, raw material and SES index are separately used as an input for districting by BARD's simulated annealing clustering to derive SSAs at parcel and block granularity. Secondly, same inputs are used for k-means clustering in BARD for parcel and block datasets. K-means outputs are accepted as SSA layers for parcels and blocks. Thirdly, SOM clustering is done for block and parcel datasets using raw data and SES index separately. The SOM outputs are partitioned by BARD simulated annealing clustering to obtain SSA layers. Eventually, all SSA layers derived from these analyses are tested by some quality assessment measures and discussed at the end of this chapter to decide on the optimum (near-best) option.

5.1. Methodology

The aim in this thesis is to find a way to establish a SSA layer which can be used as a basis for the census geography hierarchy in Turkey. So, different analyses are carried out on a pre-selected study area which covers 9 neighborhoods located in Keçiören. The criteria for building SSAs was mentioned in Chapter 2 by giving examples form world-wide applications. These criteria for small area identification are summarized in the next page.

1) Homogeneity principle: The homogeneity principle involves combining a group of people, housing units, or business establishments with similar characteristics into a single geographic area.
2) Population size and equal population: The size criterion generally determines the maximum number of such entities mentioned above that someone can group within a given area. In subdividing larger areas such as neighborhoods into smaller entities, it is important to keep in mind their minimum desirable population size because of confidentiality. Also, it is desirable to obtain equal population areas because of comparability.
3) Compactness: For SSAs it usually makes sense if their peripheries are approximately equidistant from centers. Twisted or elongated areas present the possibility that the statistical characteristics of the extreme parts will differ from those of the center or from each other. Even if there are irregularities of shape, there should be a justification in terms of integration and homogeneity.
4) Appropriate boundaries: SSA boundaries generally follow permanent, visible features, such as streets, roads, highways, rivers, canals, railroads, and high-tension power lines.

Before starting to employ the clustering algorithms, principal component analysis (PCA) is implemented to derive composite socio-economic index (SES) for every parcel and block by using raw physical and household socio-economic indicators explained in Chapter 4. Because, in raw data, there are many indicators that are closely correlated and this correlation may dominate the clustering process. However, SES index is a summary index of all indicators in raw data and can be used as a unique parameter to produce homogenous SSAs. Anyway, for both raw data and SES index, three main methods are chosen and applied in this thesis to achieve SSAs that fulfill above criteria at parcel and block resolution separately. Therefore, the difference of raw data clustering and SES index clustering is seen clearly as the difference is seen for parcels and blocks. As a result, twelve SSA layer is obtained. These methods are explained below and flow of the study is shown in Figure 5.1;

- For the first method SES index and raw data for each parcel and block are used separately as an input for automated districting in BARD package which utilizes simulated annealing clustering method in order to produce contiguous, compact, equipopulous and homogenous SSAs. As a result of first method, two SSA layers are obtained using SES index and two layers are obtained using raw data for parcel and block datasets respectively.
- Second method involves application of k-means clustering with BARD package on parcel and block datasets, again using two separate inputs explained above. The outputs of k-means clustering is accepted as it is, because it already produces homogenous, compact, equipopulous and compact SSAs which results from the nature of the k-means algorithm explained in Chapter 3. As a result of second method, two SSA layers are obtained using SES index and two layers are obtained using raw physical and socio-economic data for parcel and block datasets respectively.
- Third method utilizes clustering with a neural network method called Self-Organizing Maps (SOM), by using same inputs mentioned in first method. The output planes of SOM is converted to geographical maps and consequently, similarity values (u-distances) for each parcel and block are used again as an input for BARD simulated annealing clustering to derive SSA layers. As a result of third method, two SSA layers are obtained using SES index and two layers are obtained using raw data for parcel and block datasets respectively.

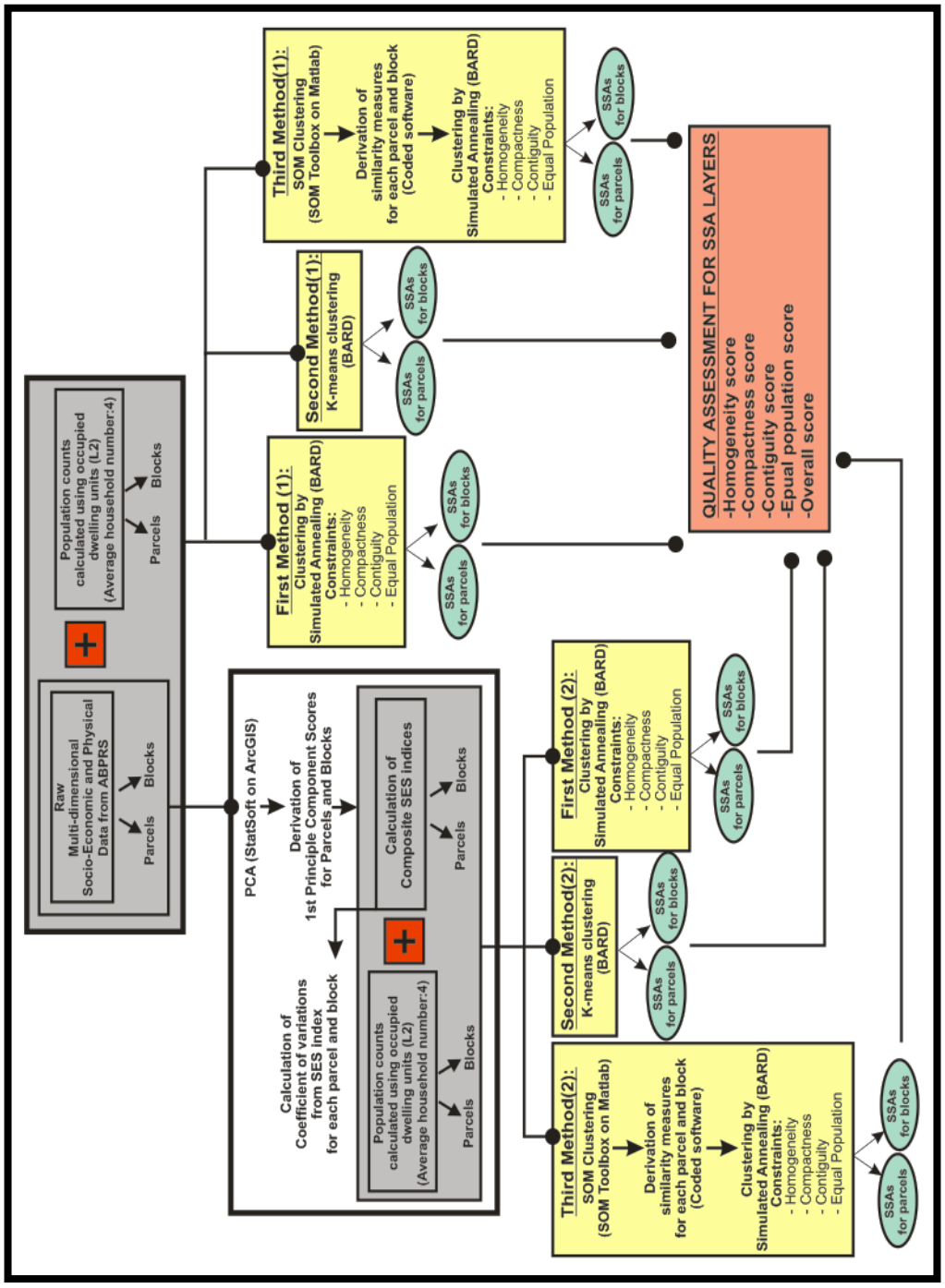


Figure 5.1 Schematic flow chart of methodology

The appropriate boundaries principle for all these three methods worked for blocks, but did not worked for parcels. Because, BARD package does not accept other layers involving road network, river boundaries, or any other layer that SSAs would nest in. On the other hand, blocks were grouped by referring to streets, roads or highways in Chapter 4. Also, in order to respect neighborhood boundaries, the

BARD simulation is run for parcels and blocks in each distinct neighborhood. So, in contrast to parcels, it was not very important to take the appropriate boundaries principle for blocks into account. Anyway, the analyses are done for both datasets to see the effect of districting geographically.

The output SSAs consisting of parcels and blocks derived separately from three methods by raw data and SES index for each neighborhood are eventually merged. But, there is an important point that must be mentioned here. The neighborhood boundaries have a changing nature in Turkey. For example, for this thesis the neighborhood boundaries are retrieved from 3 different sources and all these layers were different from each other. So, it is also very important for Turkey to stabilize the neighborhood boundaries in order to get permanent SSAs. Or, neighborhood boundaries must not be taken into consideration for building SSAs.

At this point, in order to run k-means clustering for the second method and run BARD simulations for the first and the third methods, the cluster number or in other words, number of SSAs has to be specified for the study area. As elaborately told in Chapter 2, for world-wide applications, it is observed that generally 125 household (almost 300-500 persons) is taken as a threshold for establishing the smallest level in a census geography hierarchy. This threshold can be accepted for parcels in study area. But, in the case of blocks, the number of households in many blocks generally changes between 150 and 300 households. So, in order to produce meaningful and comparable clusters for both parcels and blocks in study area, 500 household (1600-2000 persons) is selected as a threshold. Because it is again seen in world-wide applications that “500 household” is accepted as a threshold for the second level small areas. The small areas constructed with 500 households can be thought as the conjugates of census block groups in US, Lower layer Super Output Areas (LSOAs) in UK and “Small Area” level in South African census geography hierarchies. There are 63051 households (occupied dwelling units) in the study area. Consequently, “127” $\left(63051 \div 500 \cong 127\right)$ is selected as the number of small areas that will be searched in study area.

5.2. Constructing a Socio-Economic Status (SES) Index with PCA

Socio-economic analysis is inherently multi-dimensional. When measuring socio-economic variables such as population, income, employment or education one uses several response variables which must be considered together. These measurements are not mutually exclusive and are often inter-correlated. However, it may not be obvious how these measurements are spatially connected. The use of multivariate statistical analysis for database reduction techniques such as Principal Components Analysis (PCA) which is coupled with GIS is indicated in these conditions.

5.2.1 Fundamentals of PCA

PCA is a multivariate statistical technique used to reduce the number of variables in a dataset into a smaller number of dimensions. PCA analyzes individual measurements that are inter-correlated (McAdams & Demirci, 2006). PCA makes no assumption about the underlying statistical distribution of the data, thus the data is not labeled priory (Parinet et al., 2004). The primary interest in PCA is to try to understand any pattern or structure in the observations over p attributes or variables $(x_1, x_2, ..., x_p)$, a natural approach is to look for directions in the data space in which the n points are most spread out (Bailey & Gatrell, 1995).

Vyas & Kumaranayake (2006) define this notion in a more mathematical way. From an initial set of n correlated variables, PCA creates uncorrelated indices or components, where each component is a linear weighted combination of the initial variables. For example, in Equation 5.1, principal components are created by applying a linear transformation to our original variables:

$$\begin{aligned} PC_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ &\vdots \\ PC_m &= a_{11}X_1 + a_{12}X_2 + ... + a_{mp}X_p \end{aligned} \qquad (5.1)$$

In Equation 5.1, a_{mp} represents the weight of the m^{th} principal component on n^{th} variable. Diagrammatically , uncorrelated principal components can be visualized as orthogonal vectors, at right angles to each other, which means that the indices are measuring different dimensions in the data and the first major principal component is in the direction of the dominant data variability (Matejicek, 2006) (Figure 5.2).

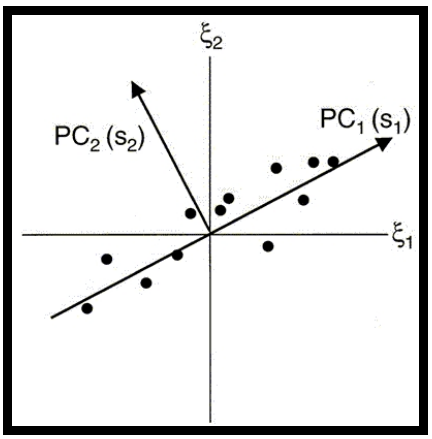


Figure 5.2 Principal components (PC) of a set of two-dimensional data. The original co-ordinate system is spanned by x_1 and x_2 . The orthogonal score vectors s_1 and s_2 are calculated according to the criterion of maximum variance.

The weights of each principal component on observations are given by the eigenvectors of the correlation matrix or if the original data were standardized, the co-variance matrix. The variance (λ) for each principal component is given by the eigenvalue of the corresponding eigenvector. The components are ordered so that the first component (PC_1) explains the largest possible amount of variation in the original data. The second component (PC_2) is completely uncorrelated with the first component and explains additional but less variation than the first component. Subsequent components are uncorrelated with the previous components; therefore, each component captures an additional dimension in the data, while explaining smaller and smaller proportions of the variation of the original variables. The higher

the degree of correlation among the original variables in the data, the fewer components required to capture common information (Vyas et al, 2006).

Over the last 20 years, PCA method is used in many fields. Some of these studies which are carried out in the last 10 years specialize on geodemographics, natural environment such as water quality or animal habitats, health studies and marketing. Voas & Williamson (2001) examined two demographic classification systems based on the analysis of 1991 census variables, for districts, wards and census enumeration districts in England and Wales. They tried to find out pair wise associations between variables using PCA in order to describe different localities on different scales. Filmer & Pritchett (2001) tried to estimate the relationship between household wealth and children's school enrollment by adapting asset ownership indicators where there is no data on income or household consumption expenditures. They built an asset ownership index and assigned households to a group on the basis of their value on the index as 'poor', 'middle' and 'rich'. Parinet et al., (2004) assessed the water quality of a tropical lake system consisting of 10 lakes by considering different chemical and physical variables existent in these lakes. In order to reveal euthrophication differences and relationships between these lakes, they gathered analytical variables from original variables by using PCA method. Essa & Nieuwoudt (2003) studied different dimensions of small-scale farmers at KawaZulu-Natal in South Africa by conducting PCA on data obtained from a sample survey of 160 households. Pettorelli et al., (2005) aimed to define the spatial pattern of major vegetation types available to deer in France. They used an integrated analysis of PCA and GIS to extract most of the variation in vegetation data. Vyas & Kumaranayeke (2006) applied PCA to asset ownership data to two different countries, Brazil and Ethiopia, in order to create socio-economic status (SES) indices in rural and urban. Specifically, they addressed issues like choice of variables, data preparation and data clustering which are also the goals of this thesis for PCA method. McAdams et al., (2006) explored the usefulness of PCA in examining the spatial attributes of water quality indicators from samples taken from Küçükçekmece Lake in Istanbul, Turkey. Yost et al., (2001) made their research on breast cancer risk and SES index conducted for blacks and whites. Their study

evaluates the relationship of SES index constructed by PCA and breast cancer incidence in California for four race/ethnic groups.

In most of the studies above, implementation of PCA on a dataset is done by using a common statistical software, since; PCA or other multivariate analysis methods are standard and routine functions for most of common statistical software. In GIS, one is able to map several variables and their distribution, but it is often difficult to determine the relationships of these variables clearly. To show the interrelationship of space and variables there is a need to integrate statistical analyses into the GIS (McAdams et al., 2006). PCA is a robust statistical technique to reduce data and develop composite variables. However, it is not linked spatially. The GIS can take the data and display spatial tendencies (Arslan, 2008). For this reason, in this thesis, PCA has been implemented directly from ESRI ArcGIS 9.2 by using a close coupled VBScript Macro called “Stat Tools”.

5.2.2. Constructing a Socio-economic Status (SES) index for the Case Study Area

The collection of accurate socio-economic indicators requires extensive resources of household surveys such as in the case of the survey conducted by TURKSTAT called Address Based Population Register System (ABPRS). In this part of the thesis, the issue is to aggregate over the range of different variables to derive a four-dimensional (landuse, education, income source and income amount) and a uni-dimensional SES index which differentiates socio-economic levels for parcels and blocks in nine districts in Keçiören. This is because each variable, used individually may not be sufficient to differentiate household SES. In addition, the SES indices can be valuable variable inputs to clustering and regionalization while trying to find out homogenous small areas, because SES index gives a unique value for parcels and blocks and it removes the domination of correlated variables in raw data. We divide this section into 3 parts to reflect the main steps in constructing a SES index: description of socio-economic indicators, application of PCA and interpretation of results and classification of households into socio-economic groups. Also, the steps

in the whole process of constructing a composite SES are summarized in Figure 5.3.

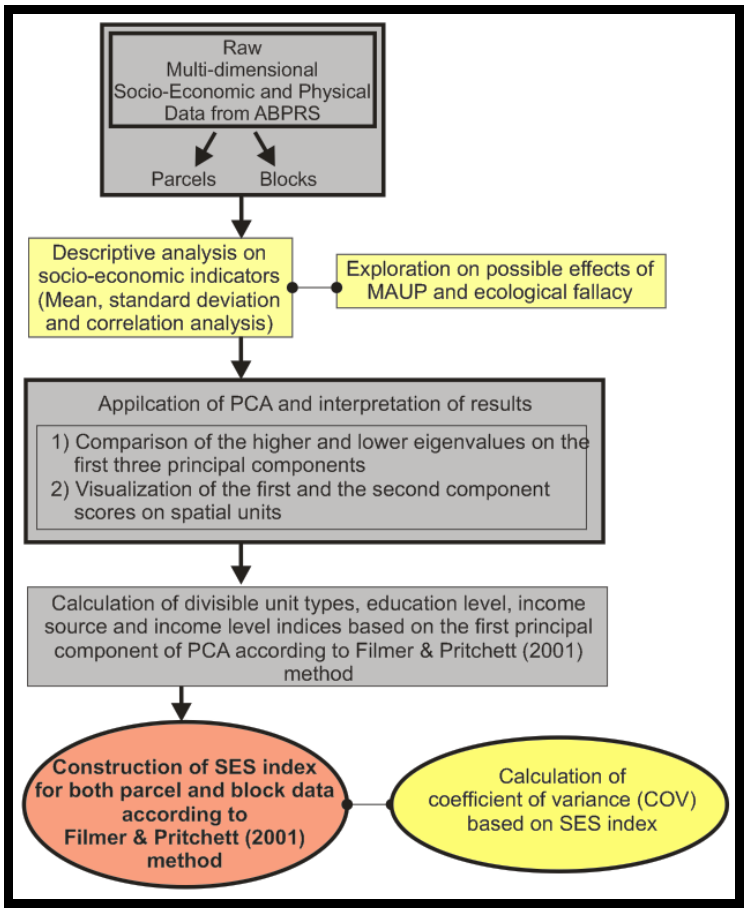


Figure 5.3 Steps followed for constructing a composite SES index.

5.2.3. Description of Socio-economic Indicators

According to Yost et al. (2000), a widely accepted definition of a SES is based on the combination of occupation, education and income which represent three main domains: class, status and power. However, in ABPRS (2008), information is collected mainly on two different fields as physical structure and socio-economic structure; the first one covers addresses and types of divisible units in a building and second one covers education level, source/sources of income and income level of each household. On the other hand, Vyas et al (2006) indicates that if there is an absence of a “best practice” approach, then substitute approach is to select

variables that proxy living standards, in this case 30 variable representing land use, education, income source and income level classes are used as an input for PCA conducted fro both parcels and blocks in nine districts of Keçiören. The full list of variable constructs, variable codes as used in database and variable definitions that is used in PCA analysis can be seen in Table 5.1.

Table 5.1 Variables used as an input for PCA

Variable Construct		Variable Code	Variable Definition
Physical Construct	Type of divisible units	L1	Number of empty residential units
		L2	Number of occupied residential units
		L4	Number of commercial units
		L11	Number of units that give public services like schools, religious centers, police stations, military establishments, municipalities, hospitals or official buildings etc.
		L20	Number of divisible units that is on construction process
		L22	Number of other units that is not classified
Socio-economic Construct	Education	E1	Number of illiterate person
		E2	Number of literate person that did not go to any school
		E3	Number of person that is graduated from primary school
		E4	Number of person that is graduated from primary education
		E5	Number of person that is graduated from secondary school
		E6	Number of person that is graduated from high school
		E7	Number of person that is graduated from university
		E8	Number of person that have a masters degree
		E9	Number of person that have a doctors degree
		E10	Number of person whose education level
	Income Source	S1	Number of person who obtains a steady income like salary or payment
		S2	Number of person who obtains income from commercial, labor intensive, industrial, agricultural sources or real estate and movable goods
		S3	Number of person who lives with social welfare
		S12	Number of person who obtains both S1 and S2
		S123	Number of person who obtains both S1, S2 and S3
		S13	Number of person who obtains both S1and S3
		S23	Number of person who obtains both S2 and S3
		S9	Number of person whose source of income is unknown
	Income Level	A1	Number of person whose income is between 0 and 150 TL (Turkish Lira)
		A2	Number of person whose income is between 151 and 350 TL
		A3	Number of person whose income is between 351 and 500 TL
		A4	Number of person whose income is between 501 and 1000 TL
		A5	Number of person whose income is 1001 and more
		A9	Number of person whose amount of income is unknown

(*) All variables are used for both parcels and blocks. For example, L1 indicates the total residential units in the parcel database. The same variable indicates total residential units in the block database.
(**) The socio-economic variables indicate head of household as a person.

PCA works best when distribution of variables varies across cases, in this instance parcels or blocks. Because, these are the variables that are more unequally distributed between households which are given more weight in PCA (McKenzie, 2003). Variables with low standard deviations would carry a low weight from the PCA. For example, if all households in an area are at the same education level, than this situation would exhibit no variation between households. Therefore, this variable would have little use in differentiating SES. Eventually, as a first step, descriptive analysis is carried out for all the variables for both parcel and block data, looking at means and standard deviations (Table 5.2).

Table 5.2 Descriptive statistics for variables at parcel and block scale.

Variable Construct		Variable Code	Parcel Data			Block Data		
			Mean	StdDev	cov	Mean	StdDev	cov
Physical Construct	Type of divisible units	L1	1,41	1,84	1,30	13,66	14,24	1,04
		L2	9,19	6,71	0,73	89,18	89,37	1,00
		L4	0,7	2,65	3,79	6,81	12,63	1,85
		L20	0,37	2,44	6,59	3,61	11,13	3,08
		L22	0,01	0,09	9,00	0,07	0,32	4,57
		L11	0,01	0,08	8,00	0,06	0,23	3,83
Socio-economic Construct	Education	E1	0,26	0,56	2,15	2,49	3,12	1,25
		E2	0,25	0,56	2,24	2,42	3,13	1,29
		E3	2,79	2,56	0,92	27,04	28,06	1,04
		E4	0,25	0,55	2,20	2,39	2,96	1,24
		E5	1,15	1,32	1,15	11,14	11,52	1,03
		E6	2,35	2,18	0,93	22,8	23,2	1,02
		E7	1,79	2,21	1,23	17,36	20,27	1,17
		E8	0,15	0,46	3,07	1,42	2,47	1,74
		E9	0,00	0,01	-	0	0,04	-
		E10	0,05	0,24	4,80	0,48	1,04	2,17
	Income Source	S1	7,42	5,68	0,77	71,99	72,5	1,01
		S2	0,97	1,34	1,38	9,42	10,27	1,09
		S3	0,11	0,36	3,27	1,03	1,64	1,59
		S12	0,35	0,75	2,14	3,37	5,05	1,50
		S123	0,00	0,07	-	0,04	0,23	-
		S13	0,05	0,23	4,60	0,44	0,91	2,07
		S23	0,01	0,11	11,00	0,11	0,36	3,27
		S9	0,18	0,48	2,67	1,74	2,38	1,37
	Income Level	A1	0,07	0,33	4,71	0,65	1,22	1,88
		A2	0,23	0,53	2,30	2,23	2,98	1,34
		A3	1,63	1,84	1,13	15,85	17,56	1,11
		A4	3,84	3,2	0,83	37,27	37,83	1,02
		A5	3,19	3,4	1,07	30,9	33,62	1,09
		A9	0,13	0,42	3,23	1,26	2,02	1,60

As can be seen in Table 5.2, the mean and standard deviation values increase as the aggregation level increases from parcel level to block. For example, this situation arises from the inherent property of spatial data called modifiable areal unit problem (MAUP). MAUP is a potential source of error that can affect spatial studies which utilize aggregate data sources (Unwin, 1996). Bailey et al (1995) indicates that especially in social sciences, data is in the form of aggregated measurements for households or individuals living in such zones which is also the case in this thesis. This kind of data is published in aggregated form because of confidentiality or lack of information for some individual data. Hence, individual data is presented in the form of aggregated zones. But this aggregation is only one of the possible configurations that can be drawn up. For example, Ratcliffe & McCullagh (1999) give an example for this situation. They state that enumeration districts containing comparable number of houses in England can be better sources of aggregation than police beats when displaying burglary rates in crime data. Therefore, the data spatially varies as the individual data is aggregated in a different way. In this respect, one should also be careful about ecological fallacy which assumes that the aggregated data variables represent individual people.

MAUP consists of two problems as statistical (aggregation effect) and geographical (scale effect);

- **Scale Effect:** The scale effect is the tendency, within a system of areal units, for different statistical results to be obtained from the same set of data when the information is grouped at different levels of resolution (135) (Figure 5.4-a). In the study of Voas et al (2001), scale effect is mentioned as the variability in enumeration districts changes or is lost when the data is aggregated to ward or county level.
- **Aggregation or Zoning Effect:** Zoning effect is the variability of statistical results obtained within a set of modifiable units as a function of the various ways these units can be grouped at a given scale and not as a result of the variation in the size of these units (135) (Figure 5.4-b). This effect can obviously be seen in our data. For example as the parcels are aggregated into blocks, the mean (1.15) and standard deviation (1.32) of variable E5 (number of persons that is graduated from high school) increases to 11.14

and 11.52 sequentially. But, one should be careful about not to derive any conclusion for another spatial resolution such as individual level households from this increase which is called ecological fallacy.

Therefore, MAUP follows the uncertainty, because different areal arrangements of the same data produce different results which also makes it difficult to obtain valid generalizations or comparable results.

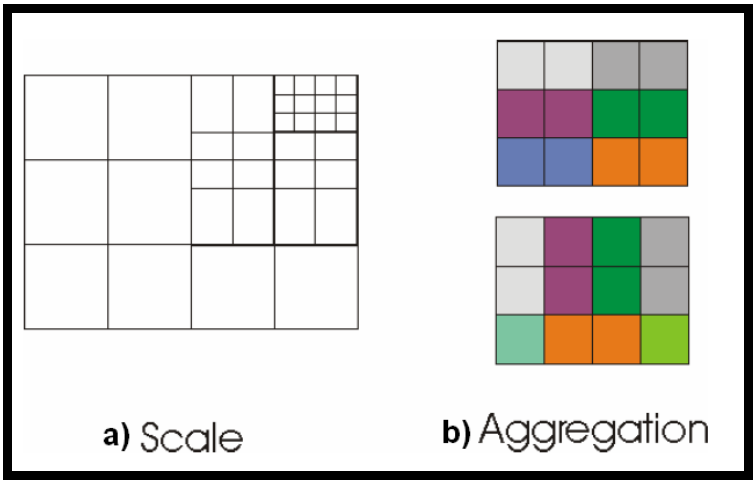


Figure 5.4 **a)** Scale effect and **b)** Aggregation effect of MAUP (Web 10).

As MAUP consists of two difficulties namely statistical and geographical, the solution to overcome these problems is also two-sided. For statistical aspect of the solution, Holt, Steel and Tranmer (1996) suggest to use well chosen variables as a result of regression analyses to adjust the area level results which may yield reliable estimates of individual relationships. From geographical aspect, Openshaw & Alvanides (1999) and Martin (2000) developed and implemented automated pattern-seeking algorithms within GIS like Zone Design System (ZDES) and Automated Zone Procedure (AZP) in order to find homogenous areas which have outcomes to overcome both difficulties of MAUP. In this thesis, we also apply clustering and zoning algorithms for both parcel and block data in order to find homogenous areas which partially solves MAUP problem; but effects of MAUP do not diminish completely, only decreases to a certain level and must not be ignored.

In Table 5.2, the results indicate that for parcel level data in nine district in Keçiören, L2 (occupied residence), E3 (primary education), E6 (high school), S1 (steady income), A4 (income between 501-1000 TL) and A5 (income level 1001+) variables have the most variation among others. For block level data, the standard deviation among modifiable units increases by MAUP effect and in addition to the variables mentioned for parcel level data, the most variation is also seen in variables L1 (empty residences), E7 (university), S12 (both steady income and temporary income) and A3 (income between 351-500 TL).

Correlation values between variables are also as much important as descriptive analysis for PCA because attribute values that are correlated with other have a predictive power among others (Vyas et al, 2006).The least correlated (<-0.5) and most correlated (>0.5) variables can be seen in Table 5.3 and Table 5.4 for both parcel and block level data respectively. In these tables, the most correlated variables are shown by dark color for coefficients greater than 0.5 and by darker color for coefficients greater than 0.8. For both tables there is no negatively correlated variable whose coefficient is lower than -0.5.

In Table 5.3 which shows the correlation coefficients for parcel level data, it can be seen that, L2 (occupied houses) is strongly correlated with E3 (primary school), E6 (high school) and A4 (income between 501-1000 TL). This means that most of the houses are occupied by people who have graduated from primary school or high school and earn money between 501 and 1000 TL. Variable E3 (primary school) is correlated with S1 (steady income), A3 (income between 351-500 TL) and A4 (income between 501-1000 TL). Thus, we can say that people graduated from primary schools usually have a steady salary which is between 351-1000 TL or vice versa. Variable E6 (high school) is correlated with S1 (steady income), A4 (income between 501-1000 TL) and A5 (income more than 1000 TL) which means that people graduated from high schools have a steady income and earns more money than 500 TL. S1 (steady income) is highly correlated with A4 (income between 501-1000 TL) and A5 (income more than 1000 TL). Thus, people who gain more than 500 TL usually have a steady income.

When we compare block level correlation coefficients in Table 5.4 with parcel level's in Table 5.3, correlations between variables increase remarkably. This issue also results from the aggregation effect of MAUP which was firstly mentioned by Gehlke & Biehl (1934). Their study was about searching grouping effects in census tract data, pure chance data and 1000 rural counties. Analogously, they found that correlation coefficients increase as the units of census tract areas increased in size from one tract to another. As mentioned earlier, this situation can be partially solved by aggregating the individual level data more homogeneously rather than arbitrarily and is also one of goals to be accomplished in this thesis.

The correlation of variables for block data in Table 5.4 is more or less the same with parcel level's (Table 5.3). The difference between them arises from the fact that the range expands for aggregated data. For example, L2 (occupied houses) was correlated with E3 (primary school), E6 (high school) and A4 (income between 501-1000 TL) for parcel data, whereas in block data L2 (occupied houses) variable seems to be strongly correlated with most of the variables from education construct, with S1 (steady salary), S2 (temporary salary), S12 (both S1 and S2) from source of income construct and again with most of the variables from income level construct. However, these results are rather deceptive when one tries to make comments about individual units considering the relationships of variables on a greater resolution which is defined also as ecological fallacy.

It is evident that examining the variables in pairs would not give us the localities and overall spatial distribution. So, a more comprehensive multivariate analysis is required like PCA.

Table 5.3 Correlation matrix for parcel level data.

Variable Construct		Variable Construct	Physical Construct						Socio-economic Construct																					
			Types of Divisible Units						Education										Income Source								Income Level					
Variable Construct		Variable Code	L1	L2	L4	L20	L22	L11	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	S1	S2	S3	S12	S123	S13	S23	S9	A1	A2	A3	A4	A5	A9
Physical Construct	Divisible Unit Types	L1	1	0,3	0	-0,1	-0	-0,1	0,1	0,1	0,2	0,1	0,2	0,2	0,2	0,1	-0	0,1	0,3	0,1	0,1	0,1	0	0,1	0	0,1	0	0,1	0,2	0,2	0,2	0,1
		L2	0,3	1	0,1	-0,2	-0	-0,1	0,3	0,3	0,8	0,3	0,6	0,8	0,7	0,3	-0	0,2	1	0,5	0,2	0,4	0,1	0,2	0,1	0,3	0,1	0,3	0,6	0,8	0,7	0,2
		L4	0	0,1	1	-0	0	0	0	0	0,1	0	0,1	0,1	0,1	0,1	0	0,1	0,1	0,1	0	0,1	0	0	0	0	0	0	0	0,1	0,1	0,1
		L20	-0,1	-0,2	-0	1	-0	-0	-0,1	-0,1	-0,2	-0,1	-0,1	-0,2	-0,1	-0,1	0	-0	-0,2	-0,1	-0	-0,1	-0	-0	-0	-0,1	-0	-0,1	-0,1	-0,2	-0,1	-0,1
		L22	-0	-0	0	-0	1	0	0	-0	-0	0	-0	-0	0	0	0	-0	-0	0	0	0	0	0	-0	0	0	0	-0	-0	0	0
		L11	-0,1	-0,1	0	-0	0	1	-0	-0	-0,1	-0	-0,1	-0,1	-0,1	-0	0,2	-0	-0,1	-0,1	-0	-0	0	-0	-0	-0	-0	-0	-0,1	-0,1	-0,1	-0
Socio-economic Construct	Education	E1	0,1	0,3	0	-0,1	0	-0	1	0,1	0,2	0,1	0,2	0,2	0,1	0	-0	0	0,3	0,1	0,1	0,1	0,1	0,1	0	0,1	0,1	0,2	0,3	0,3	0,1	0,1
		E2	0,1	0,3	0	-0,1	-0	-0	0,1	1	0,2	0,1	0,2	0,2	0,1	0	-0	0	0,3	0,1	0,1	0,1	0	0,1	0	0,1	0,1	0,2	0,3	0,3	0,2	0,1
		E3	0,2	0,8	0,1	-0,2	-0	-0,1	0,2	0,2	1	0,2	0,4	0,5	0,2	0,1	-0	0,1	0,7	0,4	0,2	0,2	0,1	0,2	0,1	0,2	0,1	0,3	0,7	0,7	0,4	0,2
		E4	0,1	0,3	0	-0,1	0	-0	0,1	0,1	0,2	1	0,2	0,2	0,1	0,1	-0	0,1	0,3	0,2	0,1	0,1	0,1	0,1	0	0,1	0,1	0,1	0,3	0,3	0,2	0,1
		E5	0,2	0,6	0,1	-0,1	-0	-0,1	0,2	0,2	0,4	0,2	1	0,4	0,2	0,1	-0	0,1	0,6	0,3	0,2	0,2	0,1	0,1	0,1	0,2	0,1	0,3	0,5	0,6	0,4	0,1
		E6	0,2	0,8	0,1	-0,2	-0	-0,1	0,2	0,2	0,5	0,2	0,4	1	0,4	0,2	-0	0,1	0,7	0,4	0,2	0,3	0,1	0,1	0,1	0,2	0,1	0,2	0,4	0,7	0,6	0,1
		E7	0,2	0,7	0,1	-0,1	0	-0,1	0,1	0,1	0,2	0,1	0,2	0,4	1	0,4	-0	0,1	0,6	0,3	0,1	0,4	0	0,1	0	0,1	0,1	0,1	0,2	0,4	0,8	0,1
		E8	0,1	0,3	0,1	-0,1	0	-0	0	0	0,1	0,1	0,1	0,2	0,4	1	0	0	0,3	0,1	0	0,2	0,1	0	0	0	0	0	0	0,1	0,5	0,1
		E9	-0	-0	0	0	0	0,2	-0	-0	-0	-0	-0	-0	-0	0	1	0	-0	-0	0	-0	0	0	0	0	0	-0	-0	-0	-0	0
		E10	0,1	0,2	0,1	-0	-0	-0	0	0	0,1	0,1	0,1	0,1	0,1	0	0	1	0,2	0,1	0	0,1	0	0	0	0,1	0	0,1	0,1	0,1	0,1	0,1
	Income Source	S1	0,3	1	0,1	-0,2	-0	-0,1	0,3	0,3	0,7	0,3	0,6	0,7	0,6	0,3	-0	0,2	1	0,3	0,2	0,3	0,1	0,1	0	0,2	0,1	0,3	0,6	0,8	0,7	0,2
		S2	0,1	0,5	0,1	-0,1	0	-0,1	0,1	0,1	0,4	0,2	0,3	0,4	0,3	0,1	-0	0,1	0,3	1	0,1	0,3	0	0,1	0,1	0,1	0,1	0,2	0,3	0,4	0,4	0,1
		S3	0,1	0,2	0	-0	0	-0	0,1	0,1	0,2	0,1	0,2	0,2	0,1	0	0	0	0,2	0,1	1	0,1	0,1	0,1	0	0,1	0,2	0,2	0,2	0,2	0,1	0,1
		S12	0,1	0,4	0,1	-0,1	0	-0	0,1	0,1	0,2	0,1	0,2	0,3	0,4	0,2	-0	0,1	0,3	0,3	0,1	1	0	0,1	0	0,1	0	0,1	0,1	0,2	0,5	0
		S123	0	0,1	0	-0	0	0	0,1	0	0,1	0,1	0,1	0,1	0	0,1	0	0	0,1	0	0,1	0	1	0	0	0	0	0	0,1	0,1	0,1	-0
		S13	0,1	0,2	0	-0	0	-0	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0	0	0	0,1	0,1	0,1	0,1	0	1	0	0	0	0,1	0,1	0,2	0,1	0
		S23	0	0,1	0	-0	-0	-0	0	0	0,1	0	0,1	0,1	0	0	0	0	0	0,1	0	0	0	0	1	0	-0	0	0,1	0,1	0,1	0
		S9	0,1	0,3	0	-0,1	0	-0	0,1	0,1	0,2	0,1	0,2	0,2	0,1	0	0	0,1	0,2	0,1	0,1	0,1	0	0	0	1	0,1	0,1	0,2	0,2	0,1	0,5
	Income Level	A1	0	0,1	0	-0	0	-0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0	0	0	0,1	0,1	0,2	0	0	0	-0	0,1	1	0,1	0,1	0,1	0	0
		A2	0,1	0,3	0	-0,1	0	-0	0,2	0,2	0,3	0,1	0,3	0,2	0,1	0	-0	0,1	0,3	0,2	0,2	0,1	0	0,1	0	0,1	0,1	1	0,3	0,2	0,1	0,1
		A3	0,2	0,6	0	-0,1	-0	-0,1	0,3	0,3	0,7	0,3	0,5	0,4	0,2	0	-0	0,1	0,6	0,3	0,2	0,1	0,1	0,1	0,1	0,2	0,1	0,3	1	0,5	0,2	0,1
		A4	0,2	0,8	0,1	-0,2	-0	-0,1	0,3	0,3	0,7	0,3	0,6	0,7	0,4	0,1	-0	0,1	0,8	0,4	0,2	0,2	0,1	0,2	0,1	0,2	0,1	0,2	0,5	1	0,4	0,1
		A5	0,2	0,7	0,1	-0,1	0	-0,1	0,1	0,2	0,4	0,2	0,4	0,6	0,8	0,5	-0	0,1	0,7	0,4	0,1	0,5	0,1	0,1	0,1	0,1	0	0,1	0,2	0,4	1	0,1
		A9	0,1	0,2	0,1	-0,1	0	-0	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0	0,1	0,2	0,1	0,1	0	-0	0	0	0,5	0	0,1	0,1	0,1	0,1	1

Table 5.4 Correlation matrix for block level data.

		Variable Construct	Physical Construct						Socio-economic Construct																					
			Types of Divisible Units						Education										Income Source								Income Level					
Variable Construct		Variable Code	L1	L2	L4	L20	L22	L11	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	S1	S2	S3	S12	S123	S13	S23	S9	A1	A2	A3	A4	A5	A9
Physical Construct	Divisible Unit Types	L1	1	0,8	0,5	0	0,2	0	0,7	0,6	0,8	0,7	0,8	0,8	0,7	0,5	0	0,5	0,8	0,7	0,6	0,7	0,1	0,5	0,2	0,7	0,5	0,6	0,8	0,8	0,8	0,6
		L2	0,8	1	0,6	-0,1	0,2	0	0,8	0,8	1	0,8	1	1	0,9	0,6	0	0,5	1	0,9	0,7	0,8	0,2	0,6	0,3	0,7	0,5	0,8	0,9	1	0,9	0,6
		L4	0,5	0,6	1	-0,1	0,2	0,1	0,5	0,5	0,5	0,5	0,5	0,5	0,6	0,6	0,1	0,4	0,6	0,5	0,4	0,5	0	0,4	0,2	0,5	0,3	0,4	0,5	0,5	0,6	0,5
		L20	0	-0,1	-0,1	1	-0	-0	-0,1	-0	-0,1	-0	-0,1	-0,1	-0	-0	-0	-0	-0,1	-0,1	0	-0	-0	-0	-0	-0	-0	-0	-0	-0,1	-0	-0
		L22	0,2	0,2	0,2	-0	1	0,1	0,1	0,1	0,1	0,2	0,1	0,2	0,2	0,2	-0	0,1	0,2	0,2	0,1	0,3	0	0,2	0,1	0,2	0	0,1	0,1	0,2	0,2	0,2
		L11	0	0	0,1	-0	0,1	1	0	0	0	0	0	0	0	0,1	0,2	0	0	0,1	0,1	0,1	0,1	0,1	0,1	0	-0,1	-0	-0	0	0	0
Socio-economic Construct	Education	E1	0,7	0,8	0,5	-0,1	0,1	0	1	0,7	0,8	0,7	0,8	0,8	0,6	0,4	0	0,4	0,8	0,7	0,6	0,6	0,1	0,5	0,2	0,7	0,4	0,7	0,8	0,8	0,7	0,5
		E2	0,6	0,8	0,5	-0	0,1	0	0,7	1	0,8	0,7	0,8	0,8	0,7	0,4	0	0,4	0,8	0,7	0,6	0,6	0,1	0,5	0,2	0,6	0,5	0,6	0,8	0,8	0,7	0,5
		E3	0,8	1	0,5	-0,1	0,1	0	0,8	0,8	1	0,8	0,9	0,9	0,7	0,5	0	0,5	1	0,8	0,7	0,7	0,1	0,6	0,3	0,7	0,6	0,8	1	1	0,8	0,6
		E4	0,7	0,8	0,5	-0	0,2	0	0,7	0,7	0,8	1	0,8	0,8	0,6	0,4	0	0,5	0,8	0,7	0,5	0,6	0,1	0,5	0,2	0,6	0,4	0,6	0,8	0,8	0,7	0,5
		E5	0,8	1	0,5	-0,1	0,1	0	0,8	0,8	0,9	0,8	1	0,9	0,8	0,5	0	0,5	0,9	0,8	0,7	0,7	0,2	0,6	0,3	0,7	0,5	0,8	0,9	1	0,8	0,6
		E6	0,8	1	0,5	-0,1	0,2	0	0,8	0,8	0,9	0,8	0,9	1	0,9	0,6	0	0,5	1	0,8	0,7	0,8	0,1	0,6	0,3	0,7	0,5	0,7	0,9	1	0,9	0,6
		E7	0,7	0,9	0,6	-0	0,2	0	0,6	0,7	0,7	0,6	0,8	0,9	1	0,8	0	0,5	0,9	0,7	0,6	0,8	0,2	0,5	0,2	0,6	0,4	0,6	0,7	0,8	1	0,6
		E8	0,5	0,6	0,6	-0	0,2	0,1	0,4	0,4	0,5	0,4	0,5	0,6	0,8	1	0	0,4	0,6	0,5	0,4	0,6	0,1	0,3	0,2	0,4	0,2	0,3	0,4	0,5	0,8	0,4
		E9	0	0	0,1	-0	-0	0,2	0	0	0	0	0	0	0	0	1	-0	0	0	0	0	-0	-0	-0	0	-0	-0	0	0	0	0,1
		E10	0,5	0,5	0,4	-0	0,1	0	0,4	0,4	0,5	0,5	0,5	0,5	0,5	0,4	-0	1	0,5	0,4	0,3	0,4	0,1	0,3	0,1	0,4	0,3	0,5	0,5	0,5	0,5	0,5
	Income Source	S1	0,8	1	0,6	-0,1	0,2	0	0,8	0,8	1	0,8	0,9	1	0,9	0,6	0	0,5	1	0,8	0,7	0,8	0,2	0,6	0,3	0,7	0,5	0,8	0,9	1	0,9	0,7
		S2	0,7	0,9	0,5	-0,1	0,2	0,1	0,7	0,7	0,8	0,7	0,8	0,8	0,7	0,5	0	0,4	0,8	1	0,6	0,8	0,1	0,5	0,3	0,6	0,5	0,7	0,8	0,9	0,8	0,5
		S3	0,6	0,7	0,4	0	0,1	0,1	0,6	0,6	0,7	0,5	0,7	0,7	0,6	0,4	0	0,3	0,7	0,6	1	0,5	0,1	0,4	0,2	0,5	0,5	0,7	0,7	0,7	0,6	0,4
		S12	0,7	0,8	0,5	-0	0,3	0,1	0,6	0,6	0,7	0,6	0,7	0,8	0,8	0,6	0	0,4	0,8	0,8	0,5	1	0,1	0,6	0,2	0,5	0,4	0,5	0,7	0,7	0,8	0,5
		S123	0,1	0,2	0	-0	0	0,1	0,1	0,1	0,1	0,1	0,2	0,1	0,2	0,1	-0	0,1	0,2	0,1	0,1	0,1	1	0,1	0,1	0,1	0,1	0,1	0,1	0,2	0,2	0,1
		S13	0,5	0,6	0,4	-0	0,2	0,1	0,5	0,5	0,6	0,5	0,6	0,6	0,5	0,3	-0	0,3	0,6	0,5	0,4	0,6	0,1	1	0,2	0,5	0,3	0,5	0,5	0,6	0,5	0,4
		S23	0,2	0,3	0,2	-0	0,1	0,1	0,2	0,2	0,3	0,2	0,3	0,3	0,2	0,2	-0	0,1	0,3	0,3	0,2	0,2	0,1	0,2	1	0,2	0,1	0,3	0,3	0,3	0,3	0,2
		S9	0,7	0,7	0,5	-0	0,2	0	0,7	0,6	0,7	0,6	0,7	0,7	0,6	0,4	0	0,4	0,7	0,6	0,5	0,5	0,1	0,5	0,2	1	0,4	0,6	0,7	0,7	0,6	0,7
	Income Level	A1	0,5	0,5	0,3	-0	0	-0,1	0,4	0,5	0,6	0,4	0,5	0,5	0,4	0,2	-0	0,3	0,5	0,5	0,5	0,4	0,1	0,3	0,1	0,4	1	0,5	0,5	0,6	0,4	0,4
		A2	0,6	0,8	0,4	-0	0,1	-0	0,7	0,6	0,8	0,6	0,8	0,7	0,6	0,3	-0	0,5	0,8	0,7	0,7	0,5	0,1	0,5	0,3	0,6	0,5	1	0,8	0,8	0,6	0,5
		A3	0,8	0,9	0,5	-0	0,1	-0	0,8	0,8	1	0,8	0,9	0,9	0,7	0,4	0	0,5	0,9	0,8	0,7	0,7	0,1	0,5	0,3	0,7	0,5	0,8	1	0,9	0,7	0,6
		A4	0,8	1	0,5	-0,1	0,2	0	0,8	0,8	1	0,8	1	1	0,8	0,5	0	0,5	1	0,9	0,7	0,7	0,2	0,6	0,3	0,7	0,6	0,8	0,9	1	0,8	0,6
		A5	0,8	0,9	0,6	-0	0,2	0	0,7	0,7	0,8	0,7	0,8	0,9	1	0,8	0	0,5	0,9	0,8	0,6	0,8	0,2	0,5	0,3	0,6	0,4	0,6	0,7	0,8	1	0,6
		A9	0,6	0,6	0,5	-0	0,2	0	0,5	0,5	0,6	0,5	0,6	0,6	0,6	0,4	0,1	0,5	0,7	0,5	0,4	0,5	0,1	0,4	0,2	0,7	0,4	0,5	0,6	0,6	0,6	1

5.2.4. Application of PCA and Interpretation of Results

Having examined the pair wise associations between variables, it is now considered how they combine to describe overall spatial distribution. PCA is carried out on all 30 variables both for all districts cumulatively and for each district one-by-one on parcel and block level data by using the ESRI ArcGIS 9.2 which is closely coupled with a VBScript macro called 'Stat Tools'. The aim of conducting PCA separately for all datasets was to understand whether the variance was changing drastically from one district to another or when compared to the overall frame. The results of this study showed that for all districts and each district separately, the percentage of total variances explained by the first components were changing between 19.8% and 24.83% for parcel data and between 48.39% and 56.81% for block level data. This shows that each district separately do not differentiate from the overall frame too much. So, in this part of thesis, PCA results are shown considering all 9 districts cumulatively for both parcel and block level data.

Following the PCA for parcel level data 8 components were retained whose eigenvalue is greater than one. The number of principal component extracted can also be defined by the user, but the common method used to select components is to consider where the associated eigenvalue is greater than one. The eigenvalue for each principal component indicates the percentage of variation in the total data explained. These components cumulatively accounted 52.26% of the variance and component 1 is the most important in the classification, explaining 23.19% of the variance in the parcel data (Table 5.5).

Different factors extracted represent different dimensions of households in 9 districts of Keçiören for parcel level data. The groupings of original variables in components can be seen by the magnitudes of factor loadings. The dominant loadings which have greater and lower factor loadings are presented by background coloring in Table 5.5. For a greater clarity, loadings of first few components are compared as higher and lower eigenvalues which are relatively easy to interpret in Table 5.6.

Table 5.5 Principle components and histograms of eigenvectors for all districts at parcel level (Number of parcels=6858).

			Principal Components							
			1	2	3	4	5	6	7	8
		Eigenvalues:	6,96	1,91	1,41	1,16	1,15	1,05	1,03	1,01
		% of Variance:	23,19	6,36	4,69	3,88	3,83	3,51	3,43	3,37
		Cumulative %:	23,19	29,54	34,23	38,11	41,94	45,45	48,88	52,26
		Eigenvectors:								
		Variable codes:	1	2	3	4	5	6	7	8
Physical Construct	Divisible Unit Types	L1	-0,12	-0,04	-0,01	0,1	-0,03	0,18	0,06	-0,27
		L2	-0,37	-0,03	0,03	0,01	-0,03	0,03	-0,03	0,01
		L4	-0,05	-0,09	-0,17	-0,07	-0,12	0,02	0,08	0,17
		L11	0,04	-0,01	-0,04	-0,58	-0,31	0,02	-0,03	-0,01
		L20	0,09	-0,01	0	0,05	0,16	-0,23	-0,17	0,21
		L22	0,01	-0,04	-0,03	-0,19	-0,01	-0,08	0,62	0,45
Socio-economic Construct	Education	E1	-0,13	0,22	0,03	-0,1	0,11	0,12	0,04	0,25
		E2	-0,13	0,16	0,06	-0,07	0,06	-0,09	-0,04	0,07
		E3	-0,29	0,23	0,05	0,06	-0,1	0,01	0,03	-0,02
		E4	-0,13	0,1	-0,05	0,04	-0,06	-0,06	-0,02	0,13
		E5	-0,24	0,11	0,06	0,05	-0,11	0	-0,01	0,01
		E6	-0,29	-0,04	0,04	0,05	-0,08	0,04	-0,01	-0,07
		E7	-0,24	-0,42	-0,03	-0,07	0,12	0,05	-0,04	0,01
		E8	-0,12	-0,4	-0,04	-0,12	0,2	0,08	-0,11	0,12
		E9	0,01	0	-0,02	-0,55	-0,38	0,05	-0,17	-0,14
		E10	-0,06	0	-0,22	-0,03	0,05	-0,06	-0,24	-0,22
	Income Source	S1	-0,36	-0,02	0,06	0,02	-0,03	0,17	-0,09	0,06
		S2	-0,2	-0,06	0,05	0,07	-0,14	-0,22	0,16	-0,13
		S3	-0,1	0,17	0,06	-0,34	0,49	-0,1	-0,01	-0,11
		S12	-0,16	-0,28	0,05	-0,06	0,03	-0,26	0,19	-0,03
		S123	-0,04	0	0,05	-0,06	0,05	-0,27	-0,52	0,53
		S13	-0,06	0,04	0,03	-0,05	0,01	-0,44	0,3	-0,13
		S23	-0,03	0,03	0,02	0,04	-0,09	-0,65	-0,15	-0,25
		S9	-0,11	0,13	-0,62	0,04	0,02	-0,02	0,01	0,05
	Income Level	A1	-0,06	0,12	-0,04	-0,3	0,5	0,11	0,02	-0,26
		A2	-0,13	0,24	0,07	-0,15	0,16	-0,06	0,08	0,07
		A3	-0,25	0,31	0,11	0,03	-0,09	0,07	-0,02	0,09
		A4	-0,31	0,13	0,08	0,08	-0,13	0,05	-0,03	-0,02
		A5	-0,28	-0,42	0	-0,03	0,06	-0,01	-0,02	0,01
		A9	-0,08	0,06	-0,69	0,04	0	-0,05	0,02	0,01
Physical Construct	Divisible Unit Types	L1								
		L2								
		L4								
		L11								
		L20								
		L22								
Socio-economic Construct	Education	E1								
		E2								
		E3								
		E4								
		E5								
		E6								
		E7								
		E8								
		E9								
		E10								
	Income Source	S1								
		S2								
		S3								
		S12								
		S123								
		S13								
		S23								
		S9								
	Income Level	A1								
		A2								
		A3								
		A4								
		A5								
		A9								

Table 5.6 Comparison of higher and lower eigenvalues on the first three principle components for parcel level data.

		Lower eigenvalues		Higher eigenvalues	
Component 1	Divisible Unit Types	L2: Occupied residences	-0,37	L4: Commercial	-0,05
				L11: Public services	0,09
				L20: Construction	0,01
				L22: Other facilities	0,04
	Education	E3: Primary school	-0,29	E9: Doctor's degree	0,01
		E5: Secondary School	-0,24	E10: Unknown education level	-0,06
		E6: High School	-0,29		
		E7: University	-0,24		
	Income Source	S1: Steady income	-0,36	S3: Social welfare	-0,1
		S2: Temporal income	-0,2	S13: Steady income+Social welfare	-0,08
		S12: Steady income+Temporal Income	-0,16	S23: Temporal income+Social welfare	-0,03
				S123: Steady income+Temporal Income+Social welfare	-0,04
	Income Level	A3: Income between 351-500 TL	-0,25	A1: Income between 0-150 TL	-0,06
		A4: Income between 501-1000 TL	-0,31	A9: Unknown income level	-0,08
Component 2	Education	E7: University	-0,42	E1: Illiterate	0,22
		E8: Master's degree	-0,4	E2: literate but no education	0,16
				E3: Primary school	0,23
	Income Source	S12: Steady income+Temporal Income	-0,28	S3: Social welfare	0,17
	Income Level	A5: Income more than 1000TL	-0,42	A2: Income between 151-350 TL	0,24
				A3: Income between 351-500 TL	0,31
Component 3	Education	E10: Unknown education level	-0,22		
	Income Source	S9: Unknown income source	-0,62		
	Income Level	A9: Unknown income level	-0,69		

It seems evident from Table 5.5 and 5.6 that first component generally has negative values but some of them are much lower than others in parcel level data. At the first stage, component 1 contrasts highly residential parcels (L2) with parcels that are reserved for commercial (L4), construction (L20), public service (L11) and other facilities (L22). For education, the conventionally adopted education levels such as primary (E3), secondary (E5), high school (E6) and university (E7) is contrasted with doctor's degree (E9) and households whose education level is unknown (E10) in the study area. As for income source, it is clear that the group getting social welfare with other types of income sources such as steady and temporal income (S3, S13, S23,

S123) opposes with the group of households getting only steady or temporal income (S1, S2). Lastly, component 1 contrasts the mid-income level gaining money between 351-1000 TL (A3, A4) with low-income level gaining 0-150 TL (A1) and the group whose income level is unknown (A9).

For further analysis, component 2 is considered. Firstly, component2 contrasts high education levels such as university (E7) and master's degree (E8) with low education levels such as illiteracy (E1), literacy with no education (E2) and primary school level (E3). For income source, component 2 opposes two polar; ones getting income from both steady and temporal sources (S12) and the ones depending only on social welfare (S3). Parallel with income source, for income level, component 2 contrasts high income level which brings more than 1000 TL (A5) with low/mid-income level which brings 151-500 TL in a month (A2, A3) . In component 3, the households, whose income level, income source and education level is unknown (E10, S9, A9), are contrasted with overall data.

The first two principal component scores are calculated with the 'StatTools' software for each parcel and then, these parcels are classified into quintiles showing the overall distribution of variables in study area. These maps can be seen below in Figure 5.5 and Figure 5.6.

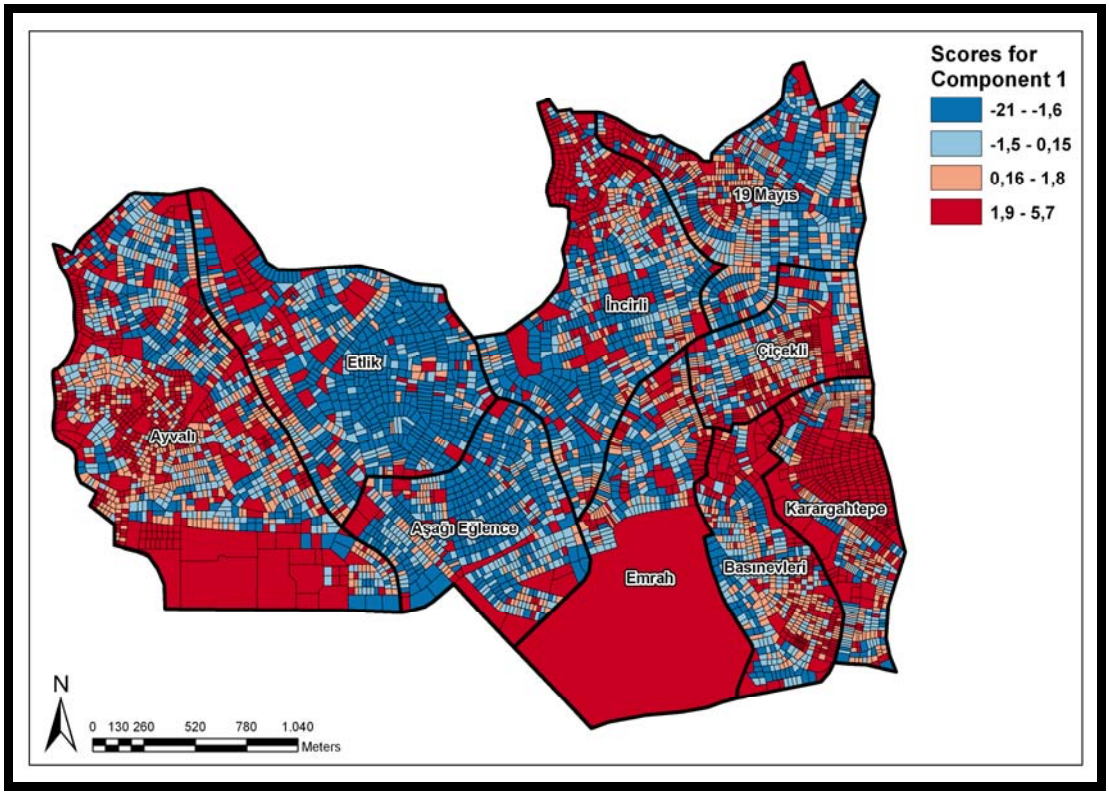


Figure 5.5 First component PCA scores for parcel level data.

It can clearly be seen in Figure 5.5 that in some areas such as south of Etlik, Aşağıeğlence, south of İncirli and east of Ondokuz Mayıs have lower negative scores interpreted as darker areas, suggesting that these zones are relative affluence. In contrast, west part of Ayvalı and İncirli, and some part of Karargahtepe have high positive scores interpreted as light coloured areas. These areas can be characterized as rather deprived areas. But, most areas are rather heterogenous, containing a mix of positive and negative scores which can be labeled as transition zones.

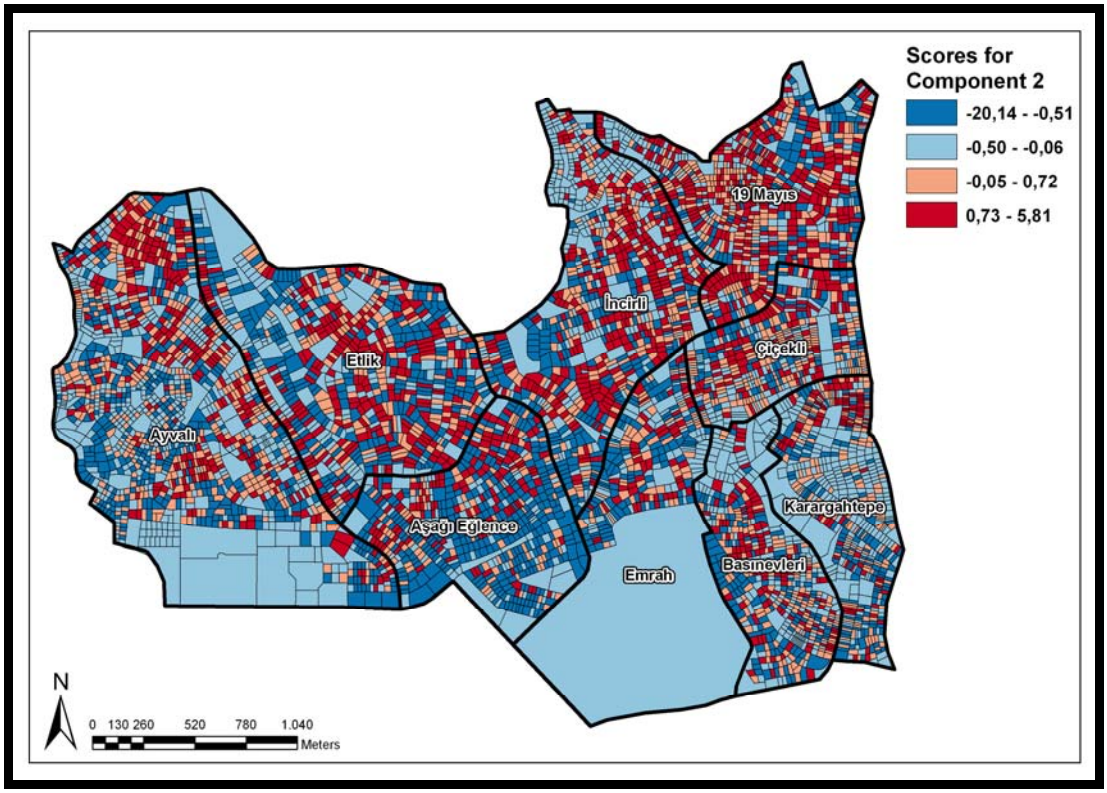


Figure 5.6 Second component PCA scores for parcel level data.

Mapping scores on second component (Figure 5.6) reveals a greater level of spatial fragmentation. Only at south of Aşağıeğlence there is a clumping of negative values.

Considering large number of variables studied (30), for a greater clarity, factor loadings can be plot on correlation circles or axis planes for component 1 versus component 2. For parcel data, we can see that factors load well on the first component on correlation circle (Figure 5.7-a). For example, E3 (primary school), E5 (secondary school), A3 (income level between 351-500 TL) and A4 (income level between 501-1000 TL) contribute to the construction of the first group on the first component. This group consists of households that have middle income and low education level. E6 (high school), S1 (steady income) and S2 (temporal income) makes up the second group on first component. This group consists of households with middle education level that have steady or temporal income source. E7 (university), E8 (master's degree), S12 (both steady and temporal income) and A5 (income more than 1000 TL) represents the third group on first component. This

group is the most prosperous group with high education level and high income from both steady and temporal sources. If we plot each parcels score in nine districts on second component against those on first component, it is possible to see parcels that behave rather differently from others (blue squares on Figure 5.7-b).

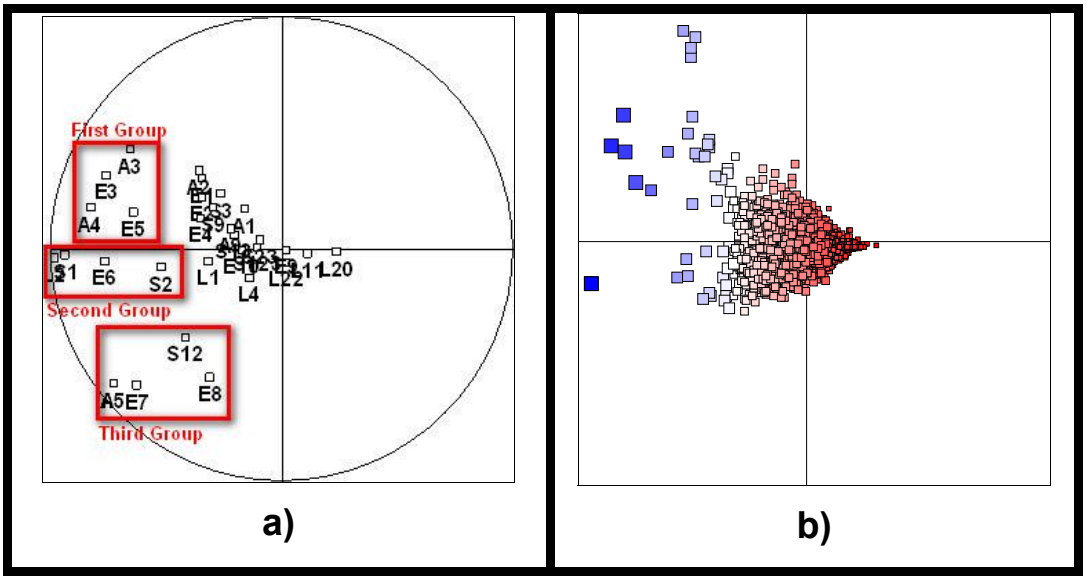


Figure 5.7 **a)** Correlation circle and **b)** second vs. first component PCA scores for parcel level data in nine districts.

For the block level data, 6 components were retained whose eigenvalue is greater than one. These components cumulatively accounted 73.21% of the variance and again as in the parcel level data, component 1 is the most important in the classification, explaining 53.96% of the variance in the block level data (Table 5.7). When we consider the variance derived from parcel data, we can see that the variance derived from block level data is approximately two times more for component 1 and it also continues to increase for subsequent components in block level data (Figure 5.8). As mentioned before, the correlation coefficients generally increase as areas are consolidated into larger spatial units. Likewise, it is found that any given number of components accounts for a higher proportion of the variance at block than at parcel. This fact also results from the notion of ecological fallacy.

Table 5.7 Principle components and histograms of eigenvectors for all districts at block level (Number of blocks=711)

			Principal Components					
			1	2	3	4	5	6
		Eigenvalues:	16,19	1,56	1,14	1,06	1,01	1,01
		% of Variance:	53,96	5,19	3,8	3,54	3,38	3,35
		Cumulative %:	53,96	59,14	62,94	66,48	69,86	73,21
		Eigenvectors:						
		Variable Codes:	1	2	3	4	5	6
Physical Construct	Divisible Unit Types	L1	-0,21	0	-0,07	0,11	0,04	0,06
		L2	-0,25	0,01	-0,01	0,01	-0,03	0
		L4	-0,15	-0,31	-0,09	0,1	0,06	-0,11
		L11	-0,01	-0,33	0,61	0,03	-0,02	0,12
		L20	0,01	0,08	-0,17	0,27	0,22	0,89
		L22	-0,05	-0,33	0,03	-0,43	0,25	0,06
Socio-economic Construct	Education	E1	-0,2	0,16	0,09	0	0	-0,04
		E2	-0,2	0,1	0,04	0	0,05	0
		E3	-0,24	0,15	0,07	-0,01	0,02	0
		E4	-0,2	0,09	0,03	-0,01	0,04	-0,03
		E5	-0,23	0,1	0,04	-0,01	-0,01	-0,01
		E6	-0,24	0,01	-0,01	0,01	-0,01	0
		E7	-0,22	-0,26	-0,17	0,04	-0,09	0,04
		E8	-0,15	-0,44	-0,27	0,04	-0,11	0,04
		E9	-0,01	-0,17	0,52	0,59	0,04	-0,05
		E10	-0,14	-0,04	-0,16	0,12	0,01	-0,13
	Income Source	S1	-0,24	0,01	-0,04	0,02	-0,04	0
		S2	-0,21	-0,01	0,07	-0,02	0,07	-0,02
		S3	-0,18	0,14	0,09	0	0,01	0,1
		S12	-0,2	-0,24	-0,05	-0,07	0,04	0,06
		S123	-0,04	-0,04	0,14	-0,2	-0,86	0,27
		S13	-0,15	-0,01	0,1	-0,24	0,1	0,05
		S23	-0,08	-0,02	0,3	-0,47	0,25	0,19
		S9	-0,19	0,05	0,05	0,09	0,01	-0,04
	Income Level	A1	-0,14	0,25	-0,02	0,07	-0,15	0,03
		A2	-0,2	0,24	0,08	-0,05	0,02	-0,02
		A3	-0,23	0,19	0,06	-0,01	0,03	-0,02
		A4	-0,24	0,1	0,04	0	0	-0,02
		A5	-0,23	-0,23	-0,13	0,03	-0,07	0,05
		A9	-0,17	-0,1	-0,06	0,14	0,05	-0,1
Physical Construct	Divisible Unit Types	L1						
		L2						
		L4						
		L11						
		L20						
		L22						
Socio-economic Construct	Education	E1						
		E2						
		E3						
		E4						
		E5						
		E6						
		E7						
		E8						
		E9						
		E10						
	Income Source	S1						
		S2						
		S3						
		S12						
		S123						
		S13						
		S23						
		S9						
	Income Level	A1						
		A2						
		A3						
		A4						
		A5						
		A9						

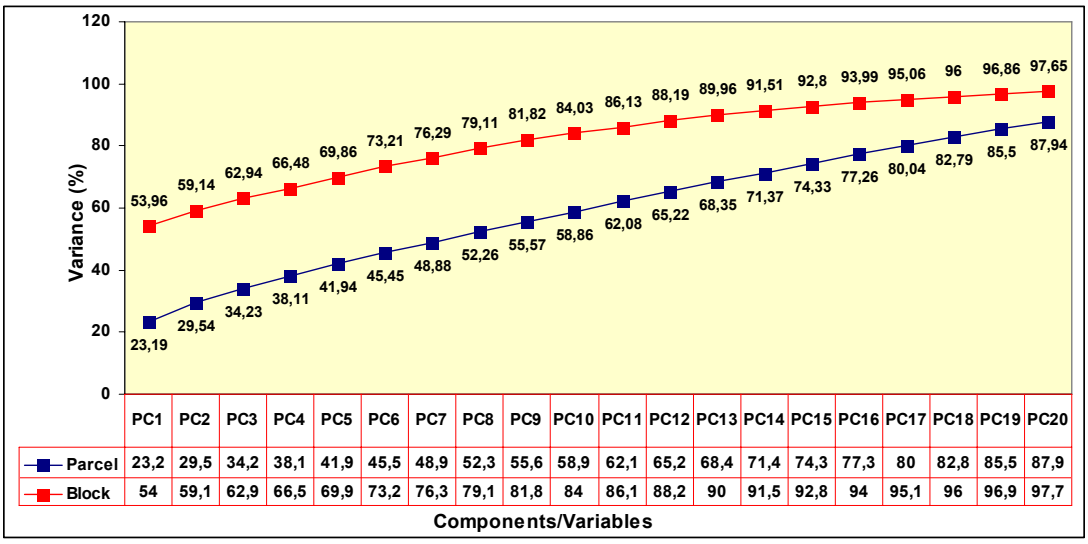


Figure 5.8 Percentage of variance accounted for by number of components/variables

For block level data, component 1 contrasts empty (L1) and occupied residential (L2) parcels with parcels reserved for public service (L11), construction (L20) and other facilities (L22). For education, the conventional education types such as primary (E3), secondary (E5), high school (E6) and university (E7) is contrasted with a higher level education: doctor's degree (E9). For income source, component 1 contrasts steady (S1) and temporal income (S2) sources with income sources that include social welfare besides steady and temporal income sources (S3, S13, S23, S123). For income level, component 1 contrasts middle and high level incomes which bring more than 351 TL (A3, A4 and A5) with overall distribution (Table 5.8).

Table 5.8 Comparison of higher and lower eigenvalues on first three principle components for block level data.

		Lower eigenvalues		Higher eigenvalues	
Component 1	Divisible Unit Types	L1: Empty residences	-0,21	L11: Public services	-0,01
		L2: Occupied residences	-0,25	L20: Construction	0,01
				L22: Other facilities	-0,05
	Education	E3: Primary school	-0,24	E9: Doctor's degree	-0,01
		E5: Secondary School	-0,23		
		E6: High School	-0,24		
		E7: University	-0,22		
	Income Source	S1: Steady income	-0,24	S23: Temporal income+Social welfare	-0,08
		S2: Temporal income	-0,21	S123: Steady income+Temporal Income+Social welfare	-0,04
	Income Level	A3: Income between 351-500 TL	-0,23		
		A4: Income between 501-1000 TL	-0,24		
		A5: Income more than 1000 TL	-0,23		
Component 2	Divisible Unit Types	L4: Commercial	-0,31		
		L11: Public services	-0,33		
		L22: Other facilities	-0,33		
	Education	E7: University	-0,26	E1: Illiterate	0,16
		E8: Master's degree	-0,44	E3: Primary school	0,15
		E9: Doctor's degree	-0,17		
	Income Source	S12: Steady Income+Temporal Income	-0,24	S3: Social welfare	0,14
	Income Level	A5: Income more than 1000TL	-0,23	A1: Income between 0-150 TL	0,25
				A2: Income between 151-350 TL	0,24
				A3: Income between 351-500 TL	0,19
Component 3	Divisible Unit Types	L20: Construction	-0,17	L11: Public Services	0,61
	Education	E7: University	-0,17	E1: Illiterate	0,09
		E8: Master's degree	-0,27	E9: Doctor's degree	0,52
		E10: Unknown education level	-0,16		
	Income Source			S3: Social welfare	0,09
				S23: Temporal income+Social welfare	0,3
				S13: Steady income+Social welfare	0,1
				S123: Steady income+Temporal Income+Social welfare	0,14
	Income Level	A5: Income more than 1000 TL	-0,13	A2: Income between 151-350 TL	0,08

In component 2, commercial (L4), public service (L11) and other facilities (L22) are contrasted with overall data. For education, component 2 contrasts high education levels such as university (E7), master's (E8) and doctor's degree (E9) with low education levels such as illiteracy (E1) and primary school (E3). For income source, households getting both steady and temporal income (S12) are contrasted with households depending only on social welfare (S3). For income level, higher income level bringing more than 1000 TL (A5) is contrasted with lower incomes bringing between 0-150 TL, 351-500 TL and 351-500 TL (A1, A2, A3) (Table 5.8).

In component 3, parcels with construction facilities (L20) are contrasted with parcels reserved for public services (L11). For education, component 3 contrasts university (E7) and master's degree (E8) with two polar; doctor's degree (E9) and illiteracy (E1). For income source, the incomes having social welfare besides steady and temporal incomes (S3, S13, S23 and S123) are contrasted with overall data. For income level, component 3 contrasts low-income households with high-income households gaining between 151-350 TL (A2) and more than 1000 TL (A5) respectively (Table 5.8).

The scores on first two components are obtained and mapped, as shown in Figure 5.9 and Figure 5.10. The shading is based on quintiles (4 intervals) with diverging colors.

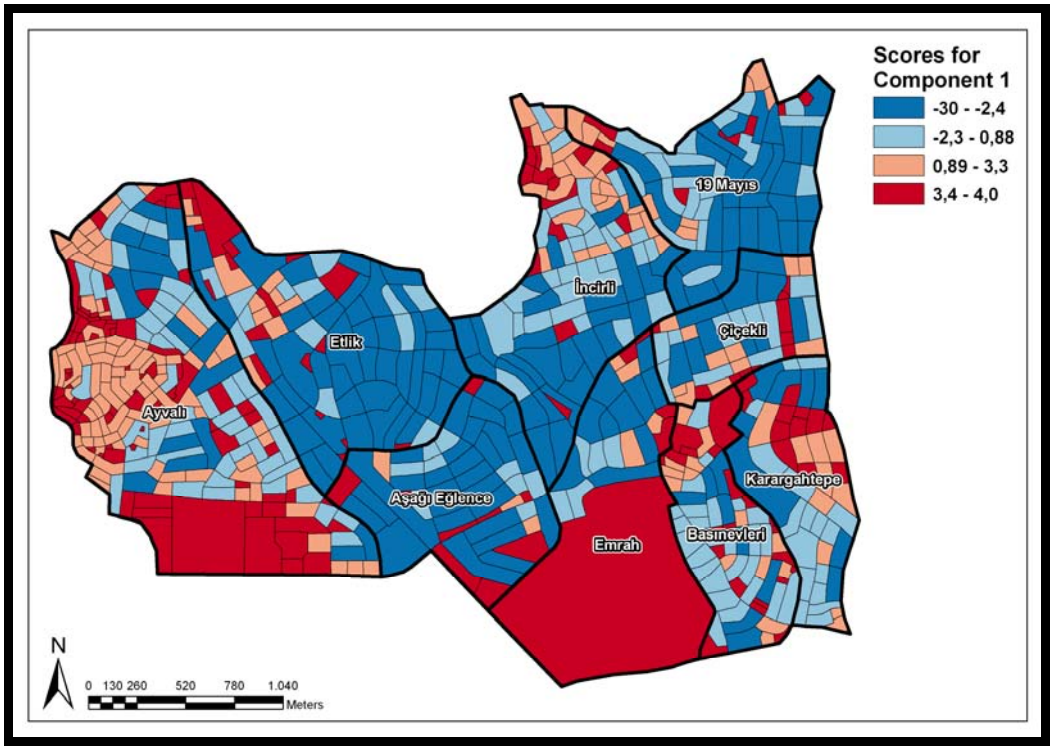


Figure 5.9 First component PCA scores for block level data.

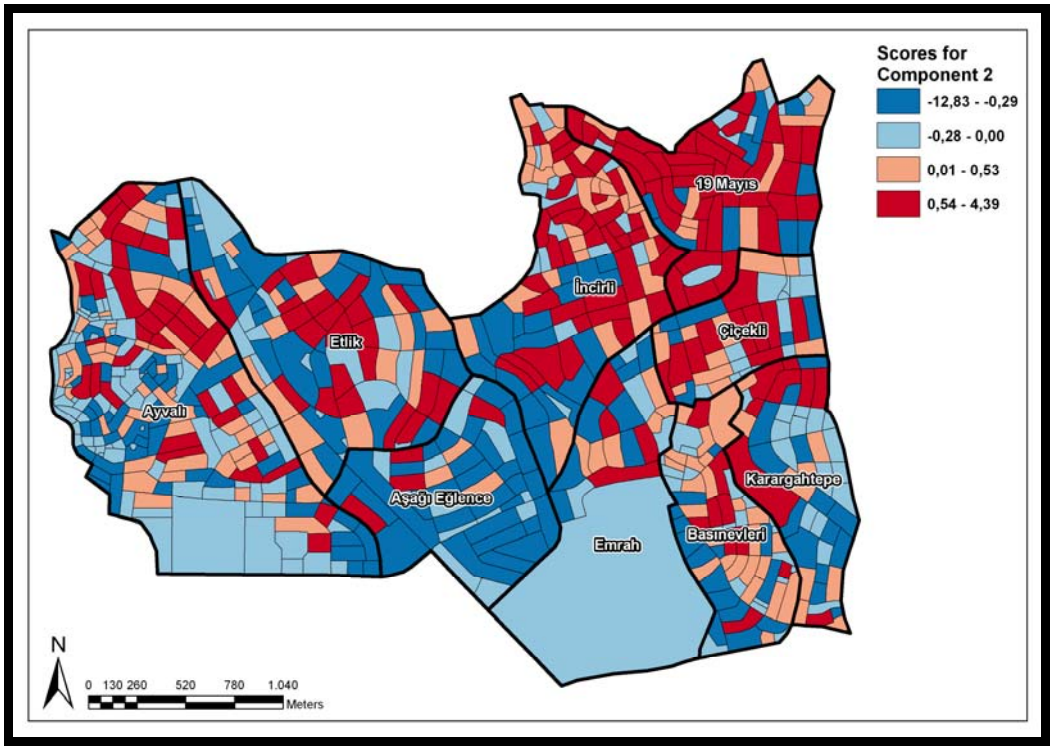


Figure 5.10 Second component PCA scores for block level data.

In Figure 5.9, which represents the first component scores, some areas are virtually entirely characterized by high negative scores. For example, south of Etlik, Aşağıeğlence, South of İncirli, Ondokuz Mayıs and north of Emrah have generally high negative scores. This suggests that these are zones of relative affluence. In contrast, west of Ayvalı, north of İncirli and east parts of Karargahtepe have relatively high positive scores characterized as deprived areas. Some areas are rather heterogeneous, containing a mix of positive and negative scores.

Mapping scores on second component reveals a more spatially fragmented picture than first component (Figure 5.10). But still, Aşağıeğlence and south part of İncirli have high negative scores. On the other hand, this time, Ondokuz Mayıs has high positive scores, Karargahtepe has negative scores and west part of Ayvalı has relatively negative scores. If we consider that component 2 represents a subgroup of component 1, we can say that these districts are far from homogenous units.

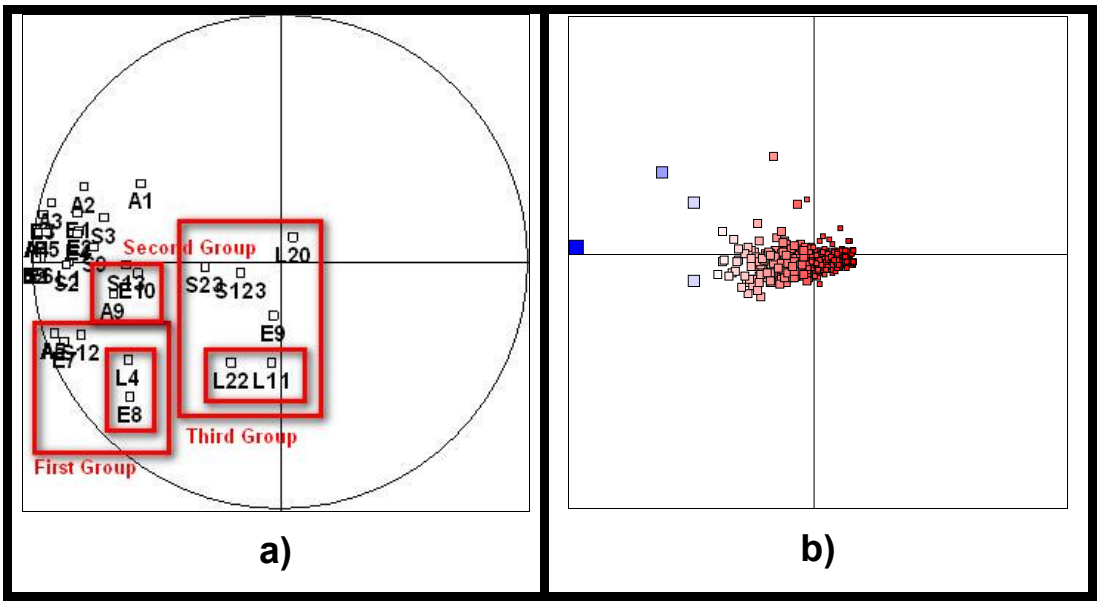


Figure 5.11 **a)** Correlation circle and **b)** second vs. first component PCA scores for block level data in nine districts.

Factor loadings are plot on correlation circles and axis planes for component 1 versus component 2 for block level data. It can be seen that, like parcel level data, factors load well on first component on correlation circle (Figure 5.11-a) and axis

plot (Figure 5.11-b). For example, A5 (income more than 1000 TL), S12 (both steady and temporal income source), E7 (university), E8 (master's degree) and L4 (commercial facilities) contribute to the construction of first group on first component. It can be said that this group consists of households that have high education and income levels and prefer to accommodate near or at central areas. S9 (unknown income source), A9 (unknown income level) and E10 (unknown education level) makes up the second group on first component and represents the households whose education level, income source or income level is unknown. S23 (temporal income and social welfare), S123 (steady income, temporal income and social welfare), E9 (doctor's degree), L11 (public services), L20 (construction facilities) and L22 (other facilities) represents the third group on component 1. This group can be characterized rather heterogeneous and is not easy to interpret. But, we can say that, this group characterizes areas that are reserved for facilities other than residential purposes and households which have income from various resources besides social welfare. E9 (doctor's degree) may be in this group because it is rarely seen in whole nine districts. In addition to correlation circle, if we plot scores of blocks on second component against first component, it is difficult visually to recognize obvious clusters (Figure 5.11-b). However, like in the case of parcel data, it is possible to spot blocks that behave differently from others named as outliers.

5.2.5. Classifying Households into Socio-economic Groups

"Socio-economic status is a multi-faceted concept that is supposed to capture many of the aspects of the relative position and achievements of an individual or a household in the society" (Kolenikov & Angeles, 2004). It is believed to be determined by the resources available to the households as well as the education levels attained by the members of the household and the prestige of their occupation. However, in literature, the researchers had to deal with other proxies for the household wealth or consumption in deriving an index for socio-economic status. Because, numeric measures of welfare such as household income or consumption are not available or reliable. In this thesis, the measures or income source, income level and education level is available based on ABPRS, but there is no variable

which determines occupation of the household head. Thus, income source variables are thought as a proxy for occupation to a degree.

The output from PCA is a table of factor scores or weights for each variable independently. But, in order to attain an aggregate index, the usual way is to construct an index which is simply a linear combination of some observed variables. One of the ways for doing this is to run PCA on data as we did in this thesis and take first component as the indicator of socio-economic status.

One of the earliest papers in population studies for the construction of socio-economic indices that used PCA was Filmer and Pritchett (2001). They used data on household assets (primary importance durable goods such as clocks, bicycle, radio, television, sewing machine, motorcycle, refrigerator, and car), type of access to hygienic facilities (sources of drinking water, types of toilet), number of rooms in dwelling and construction materials used in the dwelling. The methodology was accepted by the World Bank and Demographic Health Surveys (DHS) as the way to assess socio-economic status of a household based on the household assets and facilities. They counted each asset and facility in a house and took this measure as a proxy for SES. This resembles the case in this thesis for which we count the number of person who have currently graduated from different levels of education, earn money from different sources of income and levels of income in a building based on head of household. Then, we aggregated this information to parcel and block level.

The principal components method statistical procedure has been employed by Filmer and Pritchett (2001) basically to determine the weights to be attributed to the variables within an asset index. The procedure “locates and removes the few orthogonal linear combinations of the variables from a large number which best portray the common information”. The first principal component constitutes the linear index of variables with the most information which is common to all the variables. Till this point, the procedure is same with this thesis. But the difference is that, instead of taking the first principal component as it is, they normalized each variable by its mean and standard deviation assuming that household long-run wealth explains the

maximum variance (or covariance) in the asset variables. The mean value of index is zero. In short, the weights stand as a standardized first principle component of the variance matrix of the observed household assets.

To derive a SES index, Filmer and Pritchett's (2001) standardization is adapted using the first principle component derived from PCA. The approach produces a SES index for each household is calculated according to below formula (Equation 5.2);

$$SES_i = f_1\left(\frac{x_{i1} - \overline{x_1}}{s_1}\right) + f_2\left(\frac{x_2 - \overline{x_2}}{s_2}\right) + \cdots + f_n\left(\frac{x_{in} - \overline{x_n}}{s_n}\right) \quad (5.2)$$

Where;

SES_i = SES index for ith household,

f_1 = The scoring factor for the first variable as calculated by PCA,

x_i = The ith household value for first variable,

$\overline{x_1}$ = Mean of the first variable over all households,

s_1 = Standard deviation of the first variable over all households,

n = Total number of variables included in the procedure,

i = 1,...., i households,

n = 1,....,n household variables.

By adapting this method, indices for divisible unit types, education level, income source and income level are derived separately putting their related variables into the Equation 5.2. Then, a composite SES index is calculated by simply summing up these indices for both parcel and block level data.

Studies in literature used cut-off points to differentiate households into broad socioeconomic categories and the approaches are either arbitrarily defined or space/data driven. Commonly used arbitrary cut-off points are classification of the

lowest 40% of households into “poor”, the highest 20% as rich and the rest as the middle group (Filmer et al, 2001), or the division of households into quantiles (Gwatkin et al, 2000) (Vyas et al, 2006). In this thesis, both methods are applied and visualized for composite SES index derived for both parcel and block level data in Figures 5.12 - 5.15 respectively.

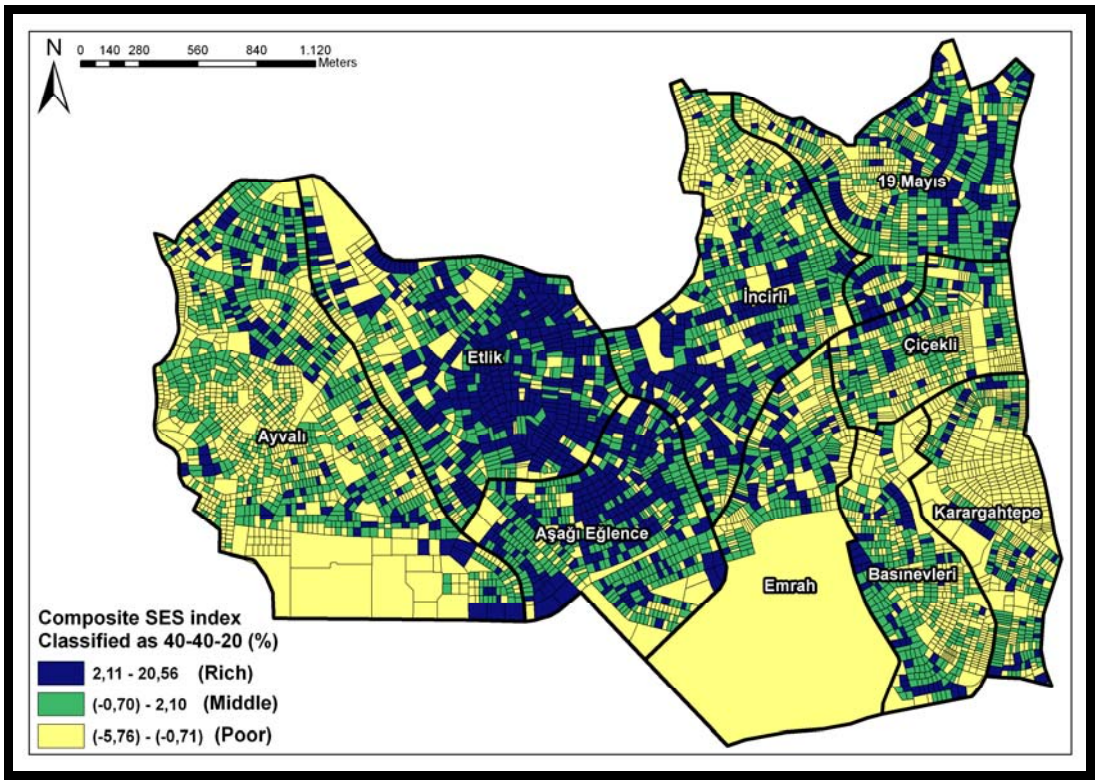


Figure 5.12 Thematic map of composite SES index classified according to 40% (poor), 40% (middle) and 20% (rich) for parcel level data.

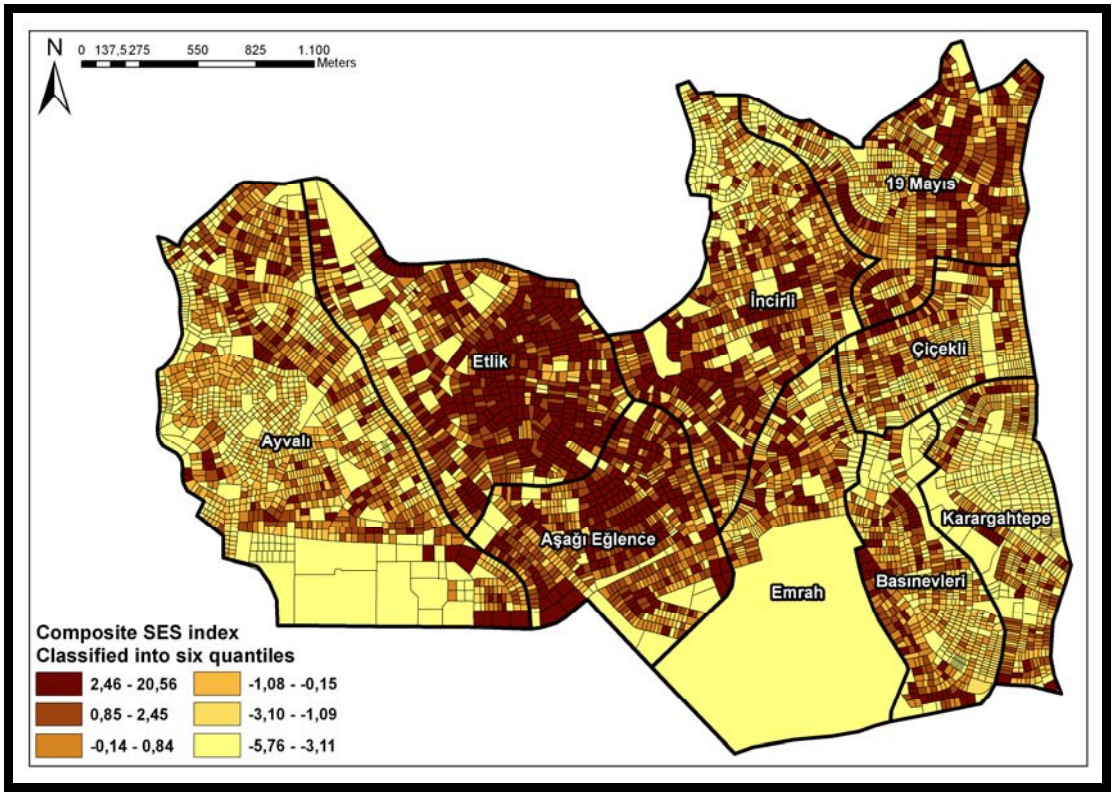


Figure 5.13 Thematic map of composite SES index classified into six quintiles for parcel level data.

Generally, a variable with a positive score is associated with a higher SES and conversely, a variable with a negative factor score is associated with a lower SES or vice versa (Vyas et al 2006). Also, in many studies, the first principal component is taken into account for calculating the SES index (Filmer & Pritchett, 2001; Vyas & Kumaranayake, 2006; McKenzie; 2003).

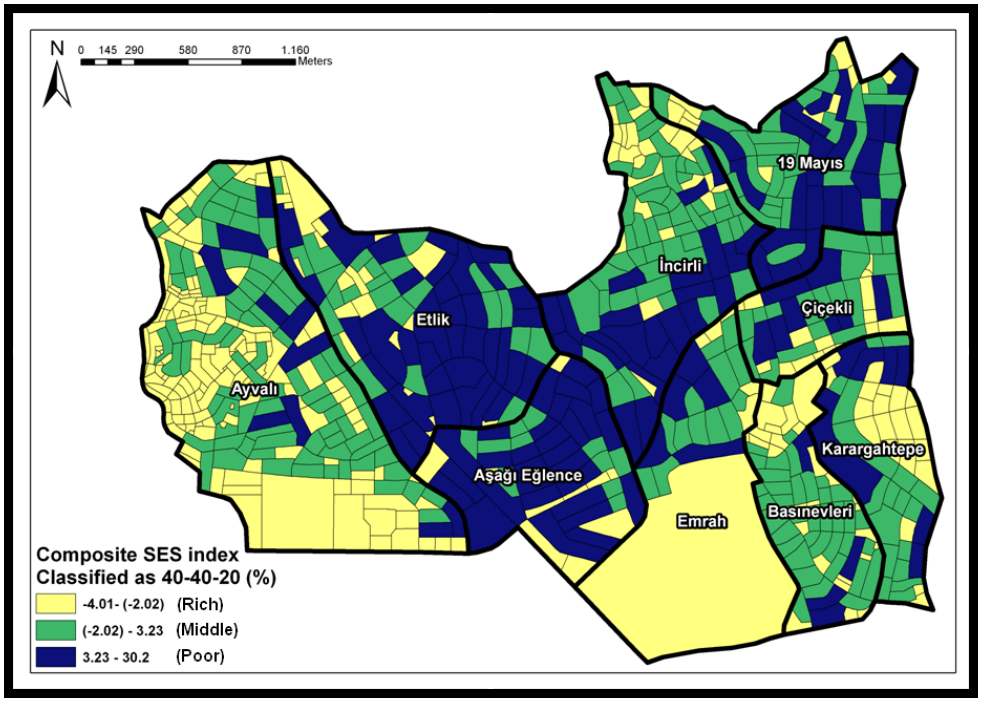


Figure 5.14 Thematic map of composite SES index classified according to 40% (poor), 40% (middle) and 20% (rich) for block level data.

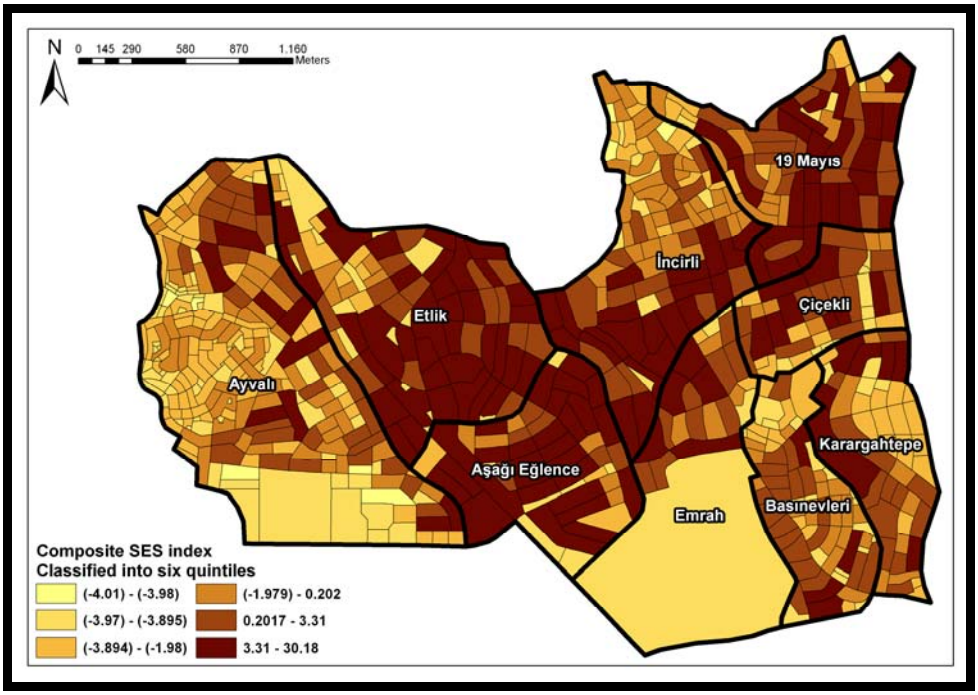


Figure 5.15 Thematic map of composite SES index classified into six quintiles for block level data.

The visualization of composite SES indices on block and parcel data are similar to visualization of first component indices that are found after the PCA study. For both scales it can be said that south of Etlik, Aşağıeğlence, South of İncirli and Ondokuz Mayıs have generally positive SES scores which suggests that these are zones of relative affluence. In contrast, west of Ayvalı, north of İncirli and east parts of Karargahtepe have relatively high negative scores characterized as deprived areas. Some areas are rather heterogeneous, containing a mix of positive and negative scores. These areas can be interpreted as transmission areas.

In the next part of this thesis, coefficient of variation value for parcels and blocks are calculated by using composite SES index.

5.2.6. Analysis of SES index

Standard deviation may be thought of as the average difference of the scores from the mean of distribution, in other words how far they are away from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values. In this case, coefficient of variation, which is a standardized standard deviation value, is calculated in order to find which parcels and blocks differentiate most from the mean of average SES index. Here, average SES index can be thought as an average household status in terms of divisible unit type, income level status, income source level and education level that each parcel and block have.

Coefficient of variation is defined as the ratio of the standard deviation to the mean. Calculation of coefficient of variation is selected instead of standard deviation because the standard deviation of data must always be understood in the context of the mean of the data. However, the coefficient of variation is a dimensionless number. Hence, when comparing between datasets with different units or widely different means such as in the case of parcels and blocks, one should use the coefficient of variation for comparison instead of the standard deviation. The map

outputs of this calculation for parcels and blocks can be seen on Figure 5.16 and 5.17 respectively.

It is observed from Figure 5.16 and 5.17 that in both parcel and block data, variation occurs more or less on same areas. The variation illustrated for block data gives a clearer idea because the variation illustrated for parcels is too fragmented. According to both figures, most parts of Aşağı Eğlence and Etlik have high coefficient of variation measure. This situation most probably arises from the fact that these two neighborhoods are commercial centers for Keçiören. Apart from these areas, south of İncirli, south of Ondokuz Mayıs and some areas of Karargahtepe have high variation. These parcels and blocks exist on transportation corridors where high traffic density exists.

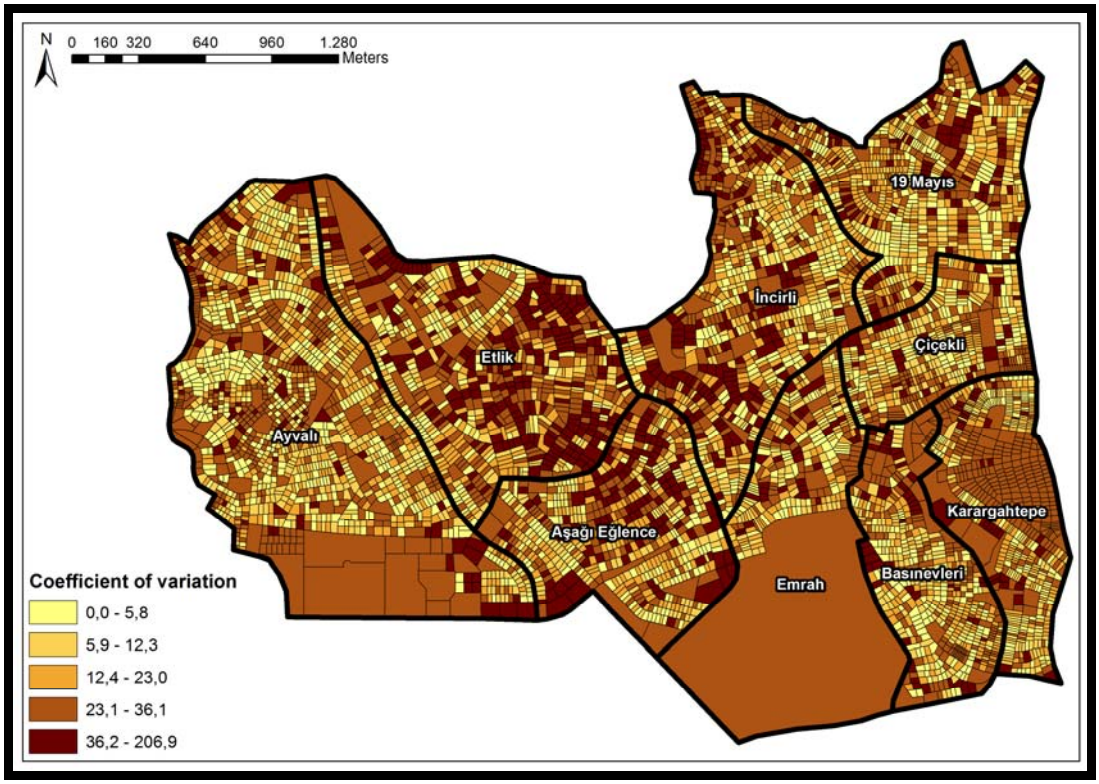


Figure 5.16 Coefficient of variation of SES values for parcel dataset.

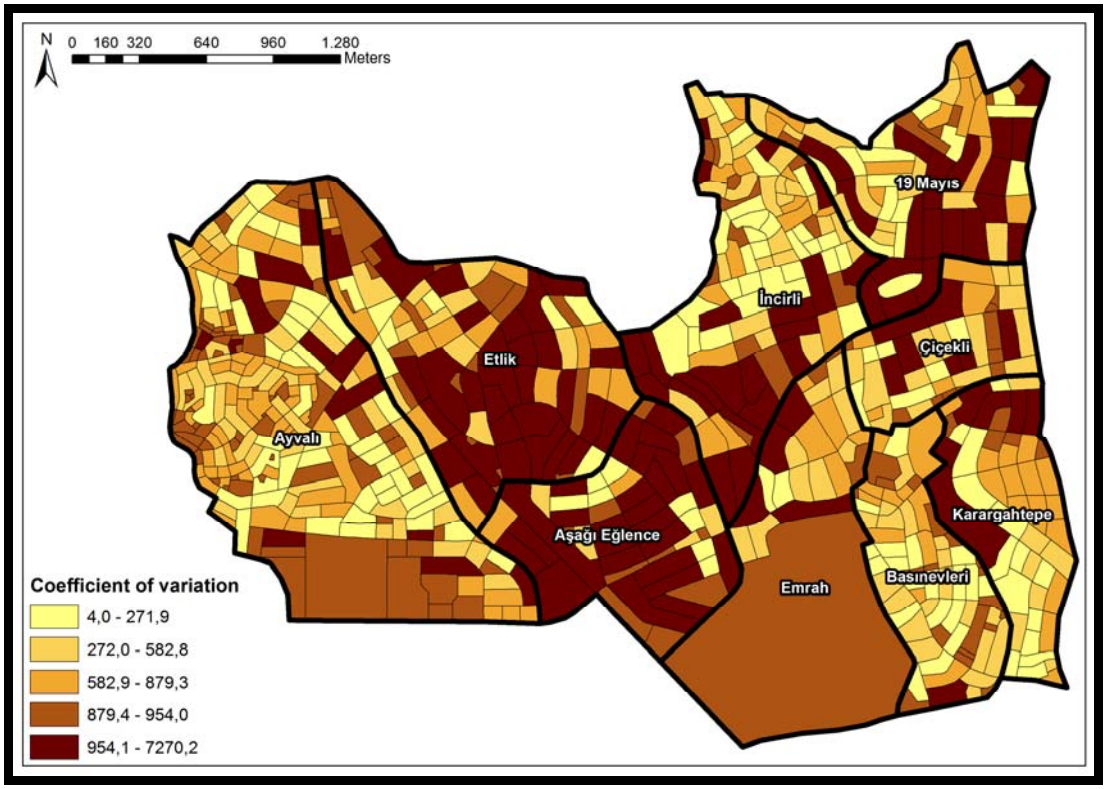


Figure 5.17 Coefficient of variation of SES values for block dataset.

5.3. Establishing SSAs by First Method: Simulated Annealing on Initial K-means clustering outputs

In order to derive SSAs by simulated annealing only raw socio-economic/physical indicators and only unique SES indices are used as separate inputs for both parcel and block dataset. But, for both inputs, population counts are calculated by multiplying the number of occupied dwelling units with the average household number which is accepted as 4. The population counts are necessary in order to create equi-populous small areas (Figure 5.18).

Firstly, the shapefiles of parcels and blocks in each neighborhood are separately prepared in ESRI ArcGIS 9.2 with above inputs in their attribute table (Figure 5.18). For example, Ondokuz Mayıs neighborhood parcel and block shapefiles are created using below attribute table structures either consisting of only raw data or only SES indices (Table 5.9);

Table 5.9 Sample attribute table structures for Ondokuz Mayıs neighborhood.

Ondokuz Mayıs parcel shapefiles attribute table structures		**Ondokuz Mayıs block shapefiles attribute table structures**	
Raw socio-economic and physical data for each parcel	**SES indices for each parcel**	**Raw socio-economic and physical data for each block**	**SES indices for each block**
Physical indicators (Table 5.1)	Composite SES indices	Physical indicators (Table 5.1)	Composite SES indices
Socio-economic indicators (Table 5.1 - Education, Income Source, Income Level)	Population	Socio-economic indicators (Table 5.1 - Education, Income Source, Income Level)	Population
Population		Population	

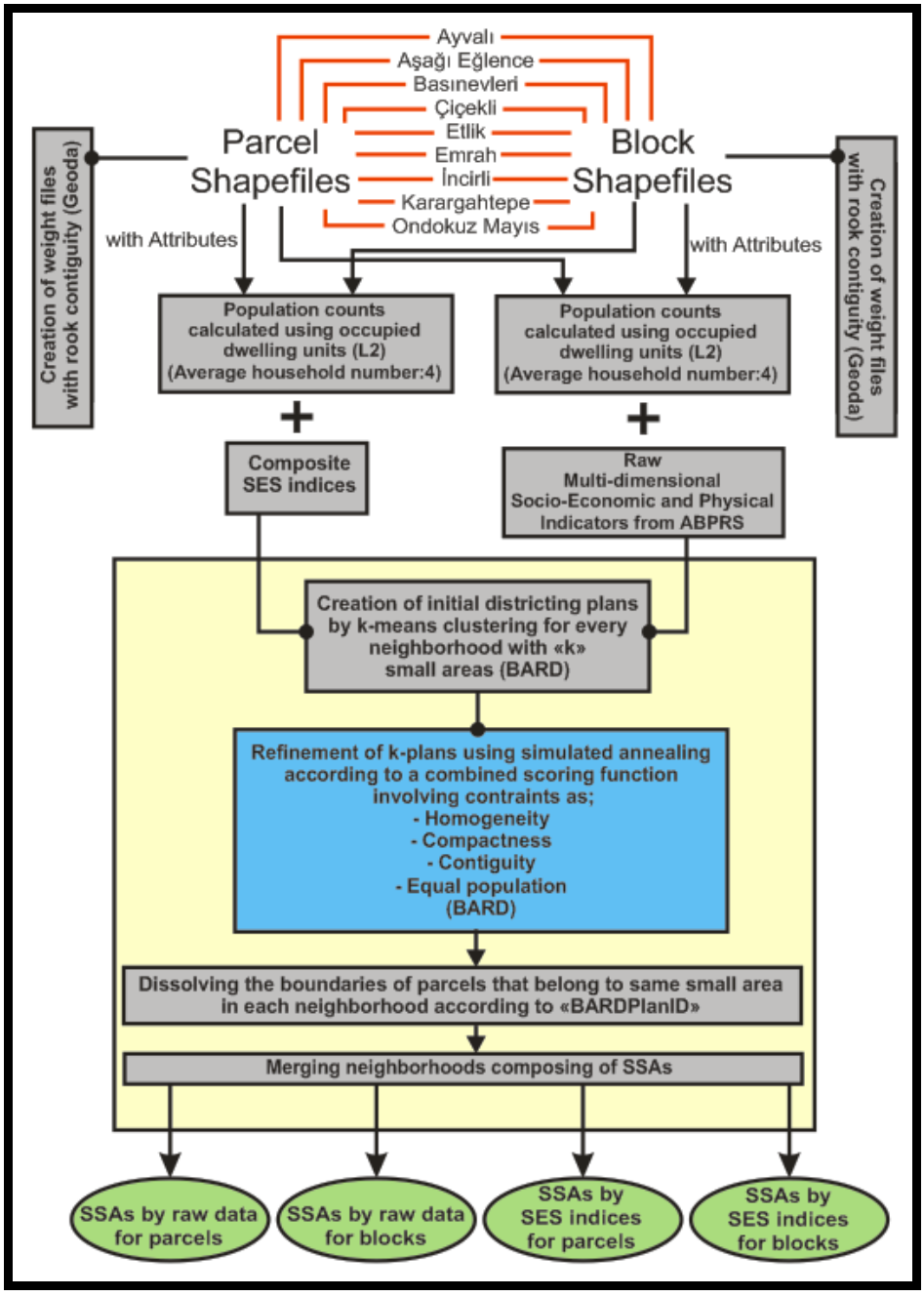


Figure 5.18 Flow-chart for the first method: Simulated Annealing on K-means clustering outputs.

As there were 9 neighborhoods in study area, 36 shapefiles are created for parcel and block datasets either using only raw data or only SES indices (Figure 5.18). Then, by using open source software called “GeoDA” (Web 2), wieight files (“*.gal”) are created by using rook contiguity which indicates a common border as a

contiguity information. Then, these shapefiles and contiguity files are imported into BARD package one by one in R environment by pseudo code written as;

"% Map Alias %"<- importBardShape(file.path
(system.file("shapefiles",package="BARD")," % Name of shapefile % "))

Secondly, numbers of small areas are calculated for each neighborhood by considering the number of household (occupied dwelling unit) that each neighborhood would have. The household numbers of each neighborhood are then divided into "500 households" which was selected as the threshold that each small area would contain (Table 5.10). The small area number, i.e. for Ondokuz Mayıs (OD) neighborhood, "18 cluster" is calculated according to this threshold and entered into BARD as written below;

Numdists_OD<-18

Table 5.10 Calculation of small area numbers for each neighborhood

Neighborhood name	Population of each neighborhood by accepting 4 as the average number of household	Number of household in each neighborhood / 500	Small are counts
İncirli	37568	9392 / 500	18,784=~19
Çiçekli	14176	3544 / 500	7,08=~7
Karargahtepe	15756	3939 / 500	7,8=~8
Etlik	43752	10938 / 500	21,8=~22
Emrah	10336	2584 / 500	5,1=5
Basınevleri	17836	4459 / 500	8,9=~9
Aşağı Eğlence	33968	8492 / 500	16,9=~17
Ayvalı	43372	10843 / 500	21,6=~22
19 Mayıs	35288	8822 / 500	17,6=~18
		Sum=	**127**

Initial plans are created separately using k-means algorithm for later refinement by simulated annealing for those 36 shapefiles by below peudocode;

"%Neighborhood kplan alias%" <- createKmeansPlan("%Map Alias%", numdists)

Then, a combined score function is created for districting with simulated annealing method in BARD. This combined score function involves calculation of scores considering compactness, contiguity, homogeneity and equal population. The pseudo code of this combined score function is written below;

myScore<-function(plan,...){ combineDynamicScores(plan,scorefuns=list(
calcPopScore,
calcLWCompactScore,
calcContiguityScore, function(x,...)
calcHomogeneityScore(x,targrange=c(.01,.99),...))))}

By using this combined score function, simulated annealing refinement is started for the initial k-means plans created for parcels and blocks in each neighborhood by raw data or SES attribute inputs in order to derive SSAs. The annealing function is written below;

"% Neighborhood Anneal Plan Alias %"<- refineAnnealPlan("%Neighborhood kplan
alias%", myScore (combined score function),
displaycount = 30 (shows refined plans in every 30 iteration),
historysize = 10 (holds scores for the last 10 iteration),
dynamicscoring = TRUE (tracking changes between candidate plans),
usecluster = TRUE (uses k-means cluster number information),
tracelevel=3 (gives full information about each iteration while tracing)

BARD's simulated annealing Works similar to AZM tool which was explained in Chapter 3. It swaps parcels and blocks between clusters in order to derive optimum scores for compactness, contiguity, homogeneity and equal population constraints. In BARD, lowest combined score shows the most optimum result. As the iteration number increases, in other words as the temperature of the algorithm decreases, the combined score also decreases.

As a result of annealing iterations for each neighborhood in parcel and block dataset with different inputs, results are exported to ArcGIS 9.2 in shapefile format;

exportBardShape(file.path(tempdir(),"%Exported shapefile name%"),
plan="% Neighborhood Anneal Plan Alias %")

BARD adds a new column to the exported neighborhood parcel and block shapefiles called "BARDPlanID" which shows the distinct cluster number of each parcel and block belong to. However, these small areas are still composed of parcel and block polygons. In order to convert these into merged small area polygons, the parcels or blocks that belong to same small area are dissolved by using "Dissolve" command in ArcToolbox for each neighborhood according to their cluster number written in ""BARDPlanID". However, in order to see all SSA layers colored according to mean SES indices and compare them, SES indices are added by as a column to output block or parcel shapefiles having only raw data in their attribute table and mean SES index is calculated while dissolving those blocks and parcels in a particular small area indicated by "BARDPlanID". In addition, while dissolving parcels and blocks, some statistical calculations are made on the indicator columns in their attribute tables. For shapefiles having raw socio-economic and physical indicators in their attribute tables, sum is calculated for every small area. For example, for a particular small area composing of parcels or blocks, in order to derive total number of the occupied dwelling units or illiterate head of households, the records related to these parcels or blocks of that small area are summed up. For all shapefiles having composite SES indices in their attribute table mean is calculated. And, for all shapefiles sum of population is taken, since population count is used as an indicator for all shapefiles. Then, the parcels and blocks which were dissolved into small areas in each neighborhood are merged by "Merge" tool in ArcToolbox to see the whole 9 neighborhoods together. Finally, thematic maps of these small areas are created according to mean SES index. All these maps are shown in Figures 5.19 - 5.22.

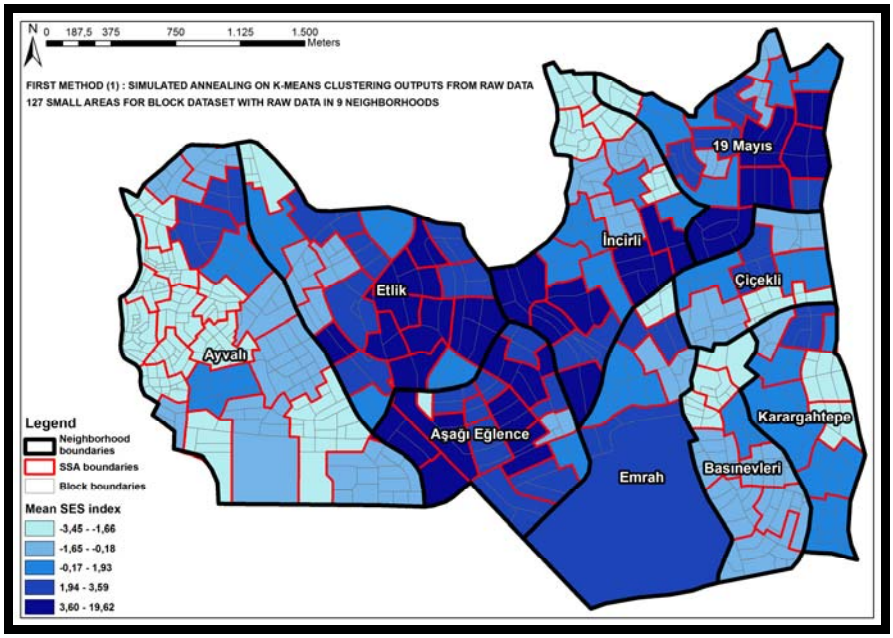


Figure 5.19 Thematic map showing mean SES index in each SSA obtained by simulated annealing on initial k-means clustering for block dataset with raw data.

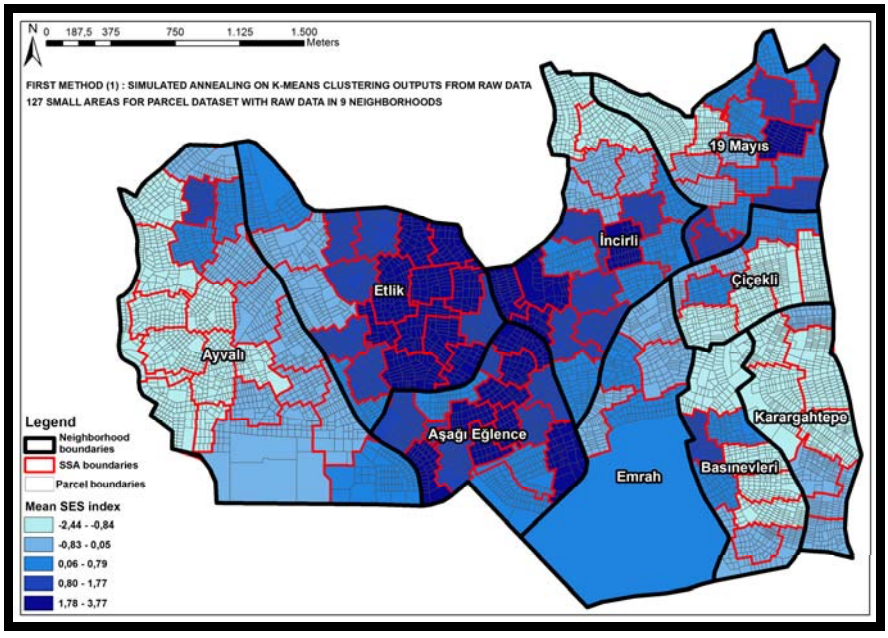


Figure 5.20 Thematic map showing mean SES index in each SSA obtained by simulated annealing on initial k-means clustering for parcel dataset with raw data.

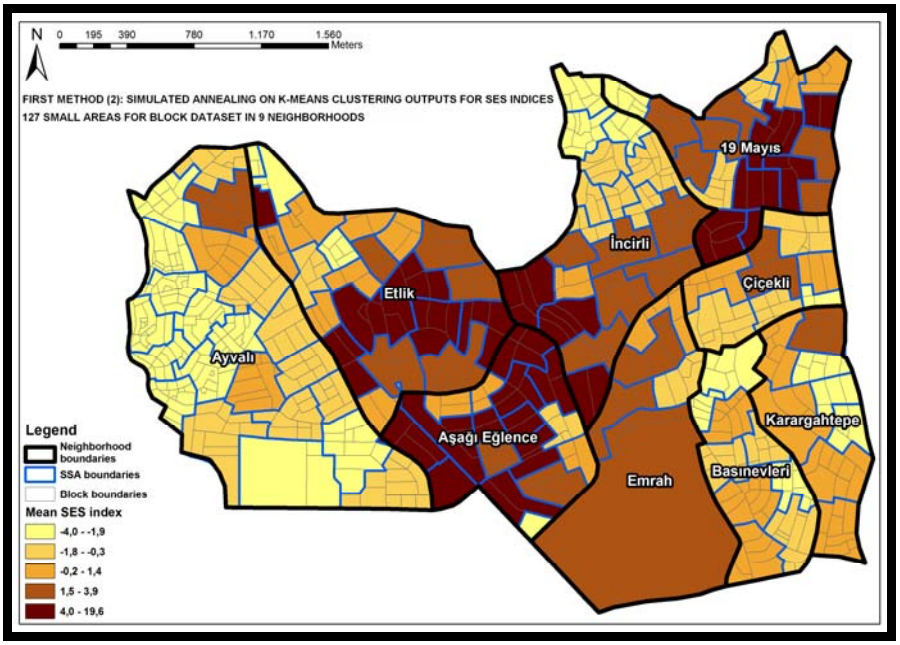


Figure 5.21 Thematic map showing mean SES index of each SSA obtained by simulated annealing on initial k-means clustering for block dataset with SES indices.

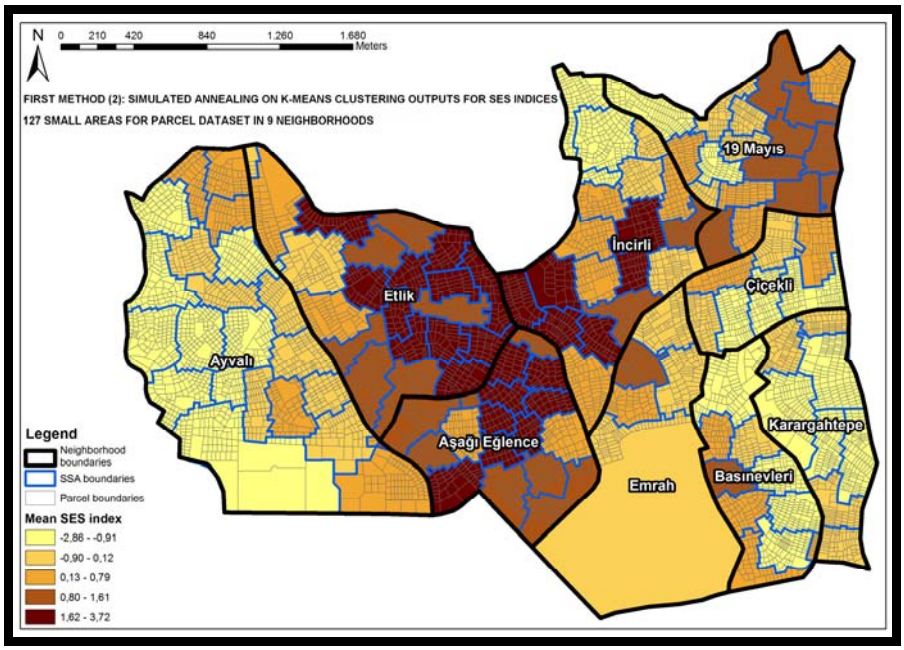


Figure 5.22 Thematic map showing mean SES index of each SSA obtained by simulated annealing on initial k-means clustering for parcel dataset with SES indices

Many other thematic maps can be produced for the small areas retrieved from raw data using different socio-economic or physical indicators. But, SES index gives an overall idea about the socio-economic status of small areas. Thus, all thematic maps are colored according to mean SES index. As can be seen in all Figures 5.19-5.22, the SES index is higher in south of Etlik, south part of İncirli , Aşağıeğlence and some parts in Ondokuz Mayıs indicating wealth. The small area boundaries of parcel data are rough because these small area boundaries follow the boundary lines of parcels, whereas small area boundaries of block data are smoother because these small areas are composed of blocks whose boundaries follow the road centerlines. Also, since parcels are more granular than blocks, it can be observed from the maps that parcel small areas are more compact than block small areas. The small area partitioning is denser in some areas where population is high with the aim to get equally populated areas. But in order to get clearer idea about compactness, contiguity, homogeneity and equal population constraints, quality assessment should be implemented. This application is done at the end of the Chapter for each method.

5.4. Establishing SSAs by Second Method: K-means clustering

Generally, k-means clustering routine groups data into “k” clusters in which “k” is defined by the user. The routine tries to find the best partitioning of the k-centers and then assigns each object to the center that is nearest according to sum of squared error (SSE). K-means assigns objects to one and only one cluster and creates homogenous, compact, contiguous and equi-populous clusters because of the algorithm’s nature which is extensively explained in Chapter 3. For these reasons, as a second method k-means clustering is selected to derive SSAs for the study area.

For k-means clustering, raw data and SES indices are used separately that belong to parcels and blocks in each neighborhood and is implemented by using BARD with the function “createKmeansPlan()” explained in section 5.3 (Figure 5.23). The only difference of second method from the first method is that, in the first method, the k-means plans were accepted initial plans which were later on refined by simulated

annealing. However this time, the k-means clusters are accepted as SSAs. The small area maps derived from k-means clustering of blocks and parcels in each neighborhood are shown in Figures 5.24-5.27 respectively.

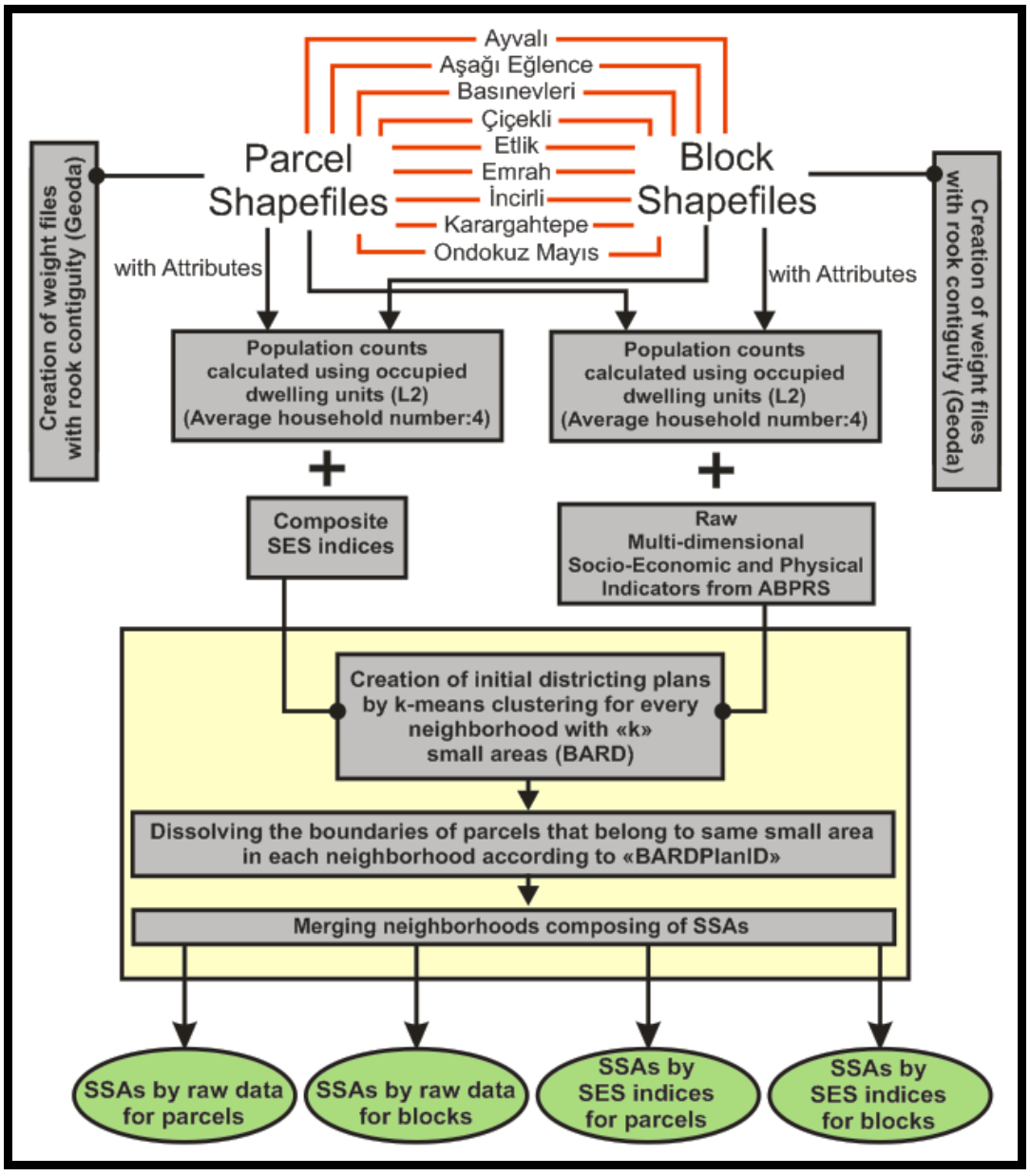


Figure 5.23 Flow-chart for the first method: K-means clustering

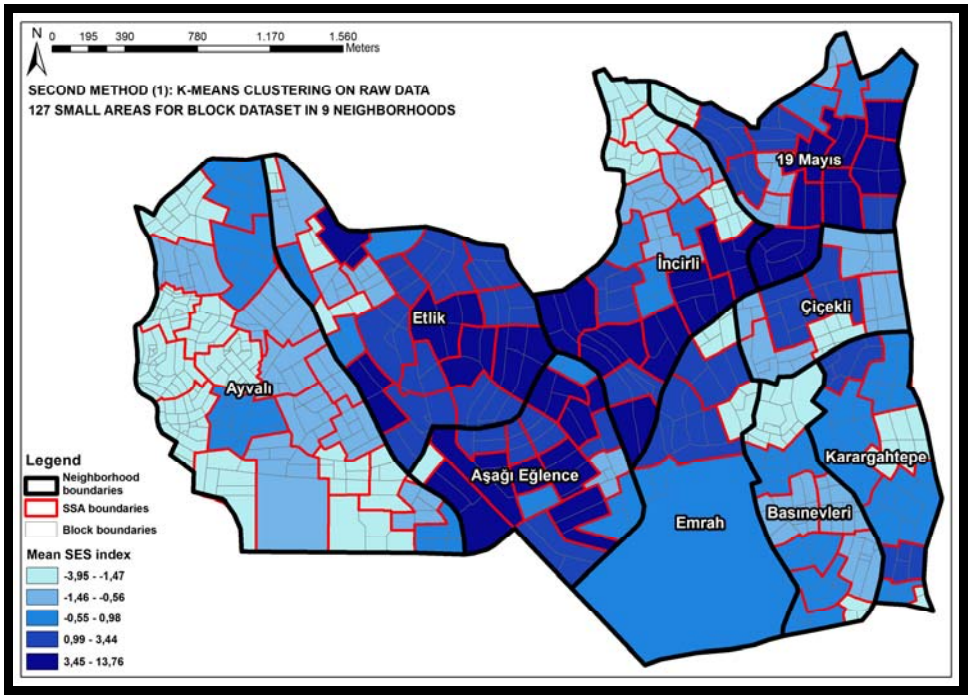


Figure 5.24 Thematic map showing mean SES index in each SSA obtained by k-means clustering on block dataset with raw data.

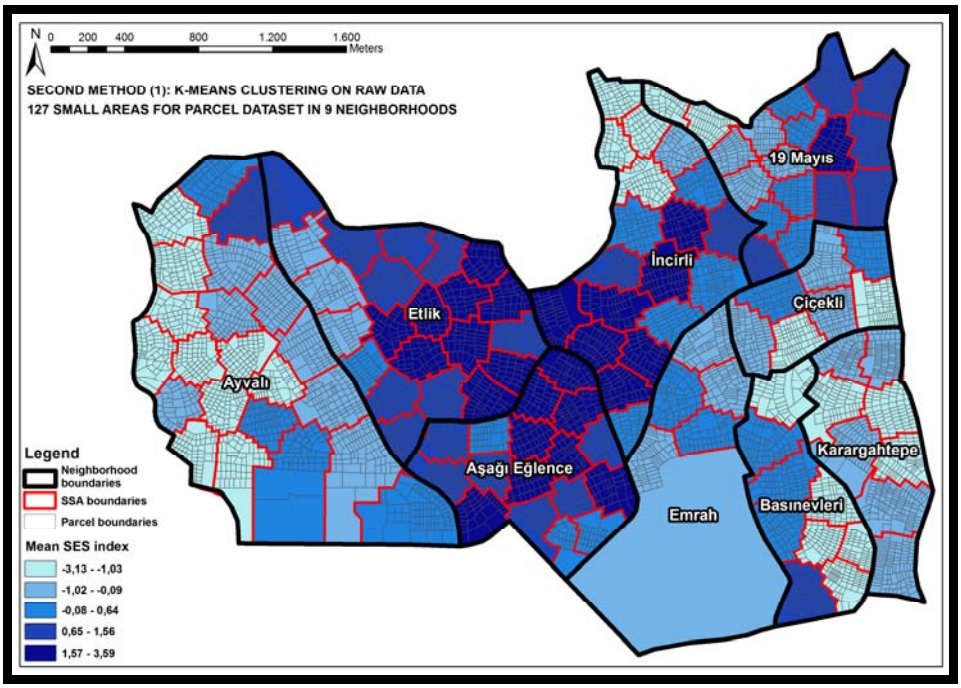


Figure 5.25 Thematic map showing mean SES index in each SSA obtained by k-means clustering on parcel dataset with raw data.

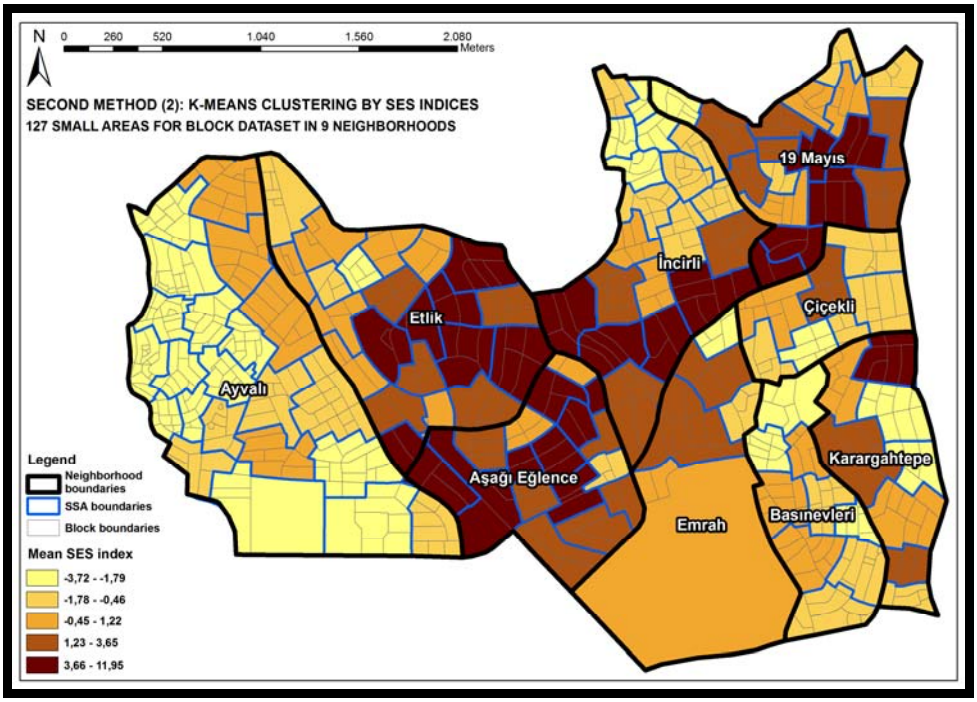


Figure 5.26 Thematic map showing mean SES index of each SSA which are obtained by k-means clustering on block dataset with SES indices.

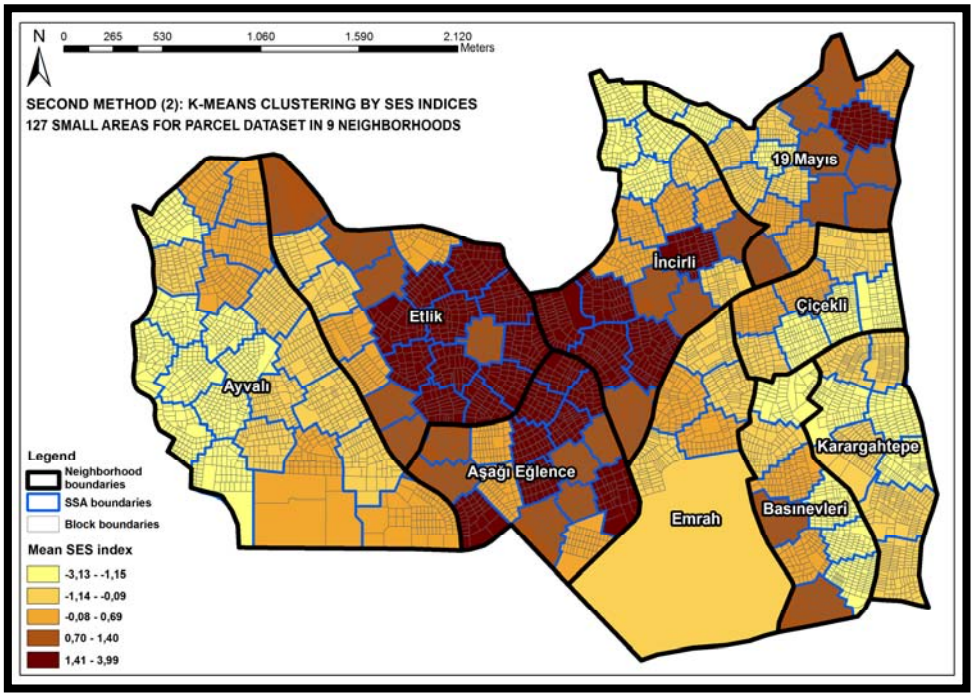


Figure 5.27 Thematic map showing mean SES index of each SSA which are obtained by k-means clustering on parcel dataset with SES indices.

At this point, the comparison between SSA layers derived from first and second method can only be interpreted visually. For example, SSAs from k-means clustering are near to perfect compactness. However, the SSAs from first method are more irregularly shaped. Some of them have elongated form. In addition, the small areas from k-means are more or less the same size. But, it is observed that the SSAs from first method have differentiating sizes which may result from equal population criteria.

5.5. Establishing SSAs by Third Method: Simulated Annealing on K-means clustering of SOM u-distances

In this part of the thesis, a framework is proposed which explores ways to effectively extract homogenous small areas using a data mining technique based on one of the neural network clustering algorithms called SOM and to represent the results using graphical representations for visual exploration. As presented in Figure 5.22, the SOM data mining stages applied here allows us to construct a clustering of the multidimensional input space using the SOM training algorithm tool (SOM Toolbox 2.0) and graphics processing with MATLAB. Also, using the cluster information from SOM, SSAs can be derived using BARD. For all u-matrix (similarity matrix), component planes and cluster products derived from SOM, GIS visualization tools are used to link the SOM grid outputs with the actual geographical data to evaluate the results. From this computational process, global structure and patterns can be represented with graphical representations and maps (geographic view) of similarity results. More exploration can be made on relationships and correlations among the attributes. The framework includes spatial analysis, data mining and knowledge discovery methods, supported by interactive tools that allow users to perform a number of exploratory tasks to understand the structure of the dataset as a whole and also explore detailed information on individual or selected attributes of the dataset.

The kind of data that can be handled by the SOM Tollbox 2.0 is in the format of spreadsheet or table data. Each row of the table is one data unit. The items on the columns are the variables or components of the dataset. The components might be

the properties of an object or set of measurements measured at a specific time Figure 5.28.

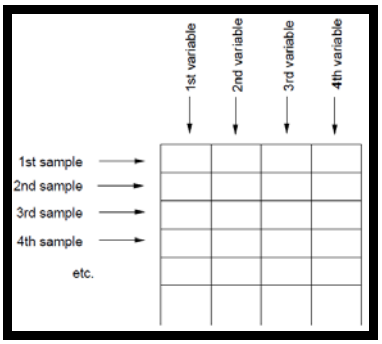


Figure 5.28 Data in table format: There can be any number of data units, but all data units have fixed length and consist of the same variables.

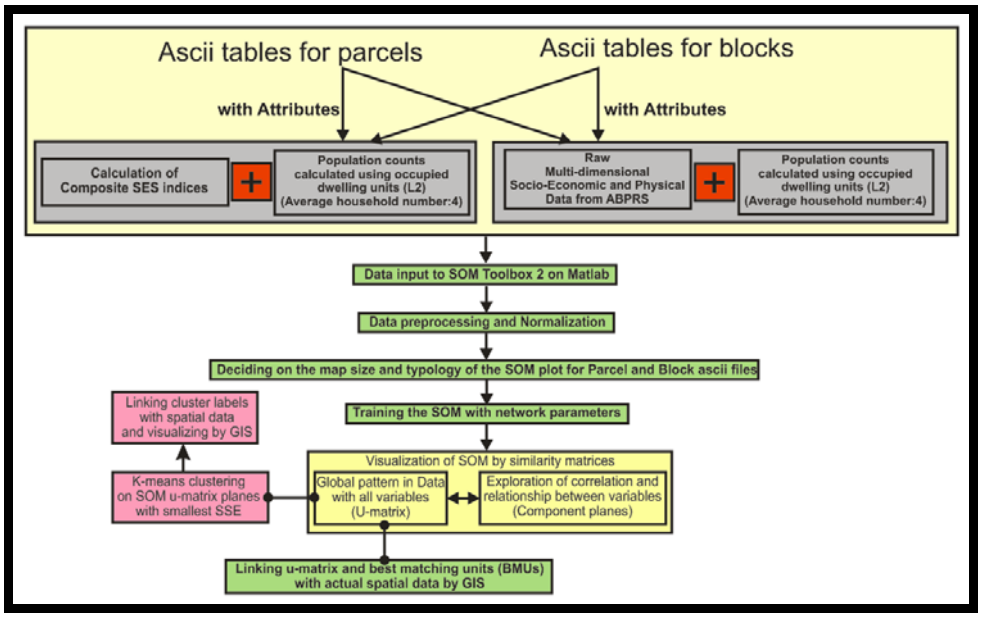


Figure 5.29 Framework proposed for clustering with SOM

As a first step, the data has to be brought into Matlab SOM toolbox workspace. For this process, the toolbox has function *"som_read_data()"* which can be used to read ascii type data files in SOM_PAK format. The attribute data of parcel and block shapefiles has been converted to ascii format including 34 variables for raw physical and socio-economic indicators plus population counts and 2 variables for SES indices plus population counts (Figure 5.29). By this way, 4 ascii files are obtained;

- Parcel ascii file with raw data and population counts for all parcels,
- Parcel ascii file with SES indices and population counts for all parcels,
- Block ascii file with raw data and population counts for all blocks,
- Block ascii file with SES indices and population counts for all blocks.

The ascii files also contain the unique parcel and block identity codes which can later be used to label the SOM plots. These unique codes were given while preparing the data in Chapter 4. The structure of ascii file formats are explained and shown in Figure 5.30.

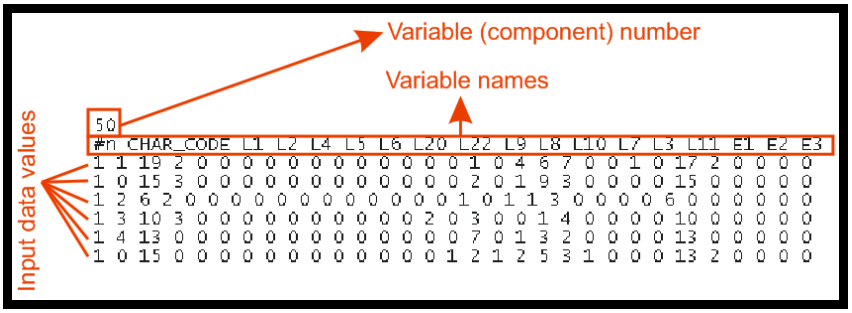


Figure 5.30 Ascii file format used for parcels and blocks as an input for SOM algorithm on Matlab SOM Toolbox 2.0.

The sample source codes for reading block and parcel ascii files in SOM toolbox is written below;

```
B_raw=som_read_data('block_raw.data')
B_ses=som_read_data('block_ses.data');
P_raw=som_read_data('parcel_raw.data')
P_ses=som_read_data('parcel_ses.data')
```

After Matlab successfully read the data for both parcel and block ascii files, the components have been normalized by *"som_normalize()"* function in order to make sure that all components have the same weight and non of them dominates the outcome of SOM training. Also, this normalization function decreases the quantization and topographic errors which are quality indicators for a good SOM. SOM toolbox provides 6 kinds of different normalization methods which are variance, range, logarithmic transformation, softmax normalization, discrete

histogram equalization and continuous histogram equalization. Among these methods, range normalization is used which scales the variable values between [0,1]. It is also called minimum-maximum normalization. This normalization type is selected because, for similar studies in literature listed in Chapter 3, it is seen that range normalization is preferred. A sample source code for parcels with raw data is indicated below;

sP_raw=som_normalize(P_raw,'range'); % for parcels with raw data

When the data for both parcels and blocks are ready, the SOM network typology and parameters has to be determined. Parameterization for each step for SOM has a large effect on the resultant trained map. For all parcel and block ascii files, the suitable SOM parameters were chosen according to rule of thumbs given by literature survey in Chapter 3. These constraints are summarized in Table 5.11.

Table 5.11 Network typology and parameters for SOM algorithm.

Network Typology	Output neurons	For parcels; $k = 5 * \sqrt{6858} = 414$ For blocks; $k = 5 * \sqrt{711} = 134$
	Output SOM plot dimension	2D, 3D
	Distance metric	Euclidean distance
	Neighborhood topology	Hexagonal
Network Parameters	Neighborhood function	Gaussian
	SOM training algorithm type	Batch training
	Weight update function	Linear
	Rough phase learning rate ($\alpha(t)$)	0.5
	Finetune phase learning rate ($\alpha(t)$)	0.05
	Rough phase initial neighborhood radius (r1)	For parcels; 415/4=~104 For blocks; 135/4=~34
	Rough phase final neighborhood radius (r2)	For parcels; 104/4=26 For blocks; 34/4=~9

Table 5.11 (Continued) Network typology and parameters for SOM algorithm.

Network Parameters	Finetune phase initial neighborhood radius (r1)	For parcels; 104/4=26 For blocks; 34/4=~9
	Finetune phase final neighborhood radius (r1)	For parcels; 1 For blocks; 1
	Rough phase training length $(10 * \frac{k}{d})$ (d= dimension of data)	For parcels; $10 * 415/40 \cong 100$ iteration For blocks; $10 * 135/39 \cong 35$ iteration
	Finetune phase training length $(40 * \frac{k}{d})$ (d= dimension of data)	For parcels; $40 * 415/40 \cong 400$ iteration For blocks; $40 * 135/39 \cong 140$ iteration

As mentioned in Table 5.11, batch training is applied for SOM, because batch training has several advantages over sequential training. Firstly, batch training presents the input data at once to the neurons instead of presenting in an ordered fashion and secondly, it is much faster than sequential algorithm. The batch training is triggered by *"som_batchtrain()"* in SOM toolbox. The iterations are started on an Intel 4 CPU 3.40 GHz processor and 2 GB RAM. Iterations took 26 minutes for blocks with raw data, 15 minutes for blocks with SES indices, 1 hour 32 minutes for parcels with raw data and 42 minutes for parcels with SES indices.

Information extraction is usually a bottleneck in neural networks (Ultsch et al., 1990; and Openshaw, 1995), and this is no exception in SOM. There are a number of different ways to summarize SOM outputs. Traditionally, component planes and unified distance matrices (U-matrices) are used.

The U-mat and component planes (where the contribution of each variable is shown) are very useful for visual analysis (Kaski, Nikkila and Kohonen, 2000) mainly when they can be easily related, for example by sharing a common coordinate

system (Kaski and Kohonen, 1998). Figures 5.31 - 5.34 show u-matrix on top left and component planes for either raw socio-economic and physical indicators or SES indices that is used as an input for parcels and blocks in SOM analysis. The advantage of this kind of representation is that all units are linked by position for component planes. So, a neuron or hexagon on a certain position on a random component map corresponds to the same position on another. Another advantage of component planes is that the correlations and relationships between indicators can be seen visually, because similar picture of variation means that two input variables are closely correlated to each other. For example, in both Figure 5.31 and 5.33 variable L20 (Number of divisible units that is on construction process) and L11 (public service areas) are negatively correlated with L1 (empty residents), L2 (occupied residents) and L4 (commercial service areas). Also, in Figure 5.32 and 5.34, C_ind (composite SES indices) is correlated with population, because in blocks and parcels where there is no residential dwelling unit, the population is zero, therefore, C_ind is also zero.

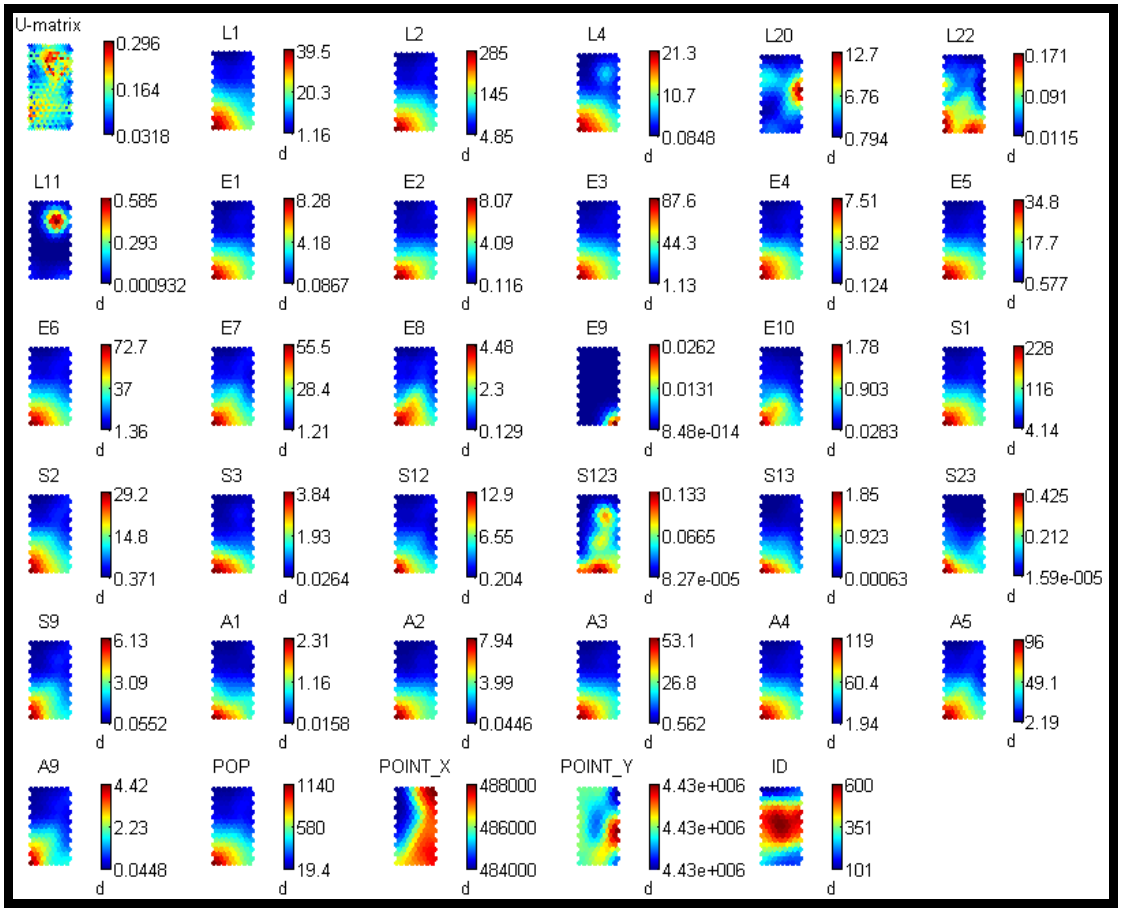


Figure 5.31 Component plane for block ascii file with raw data.

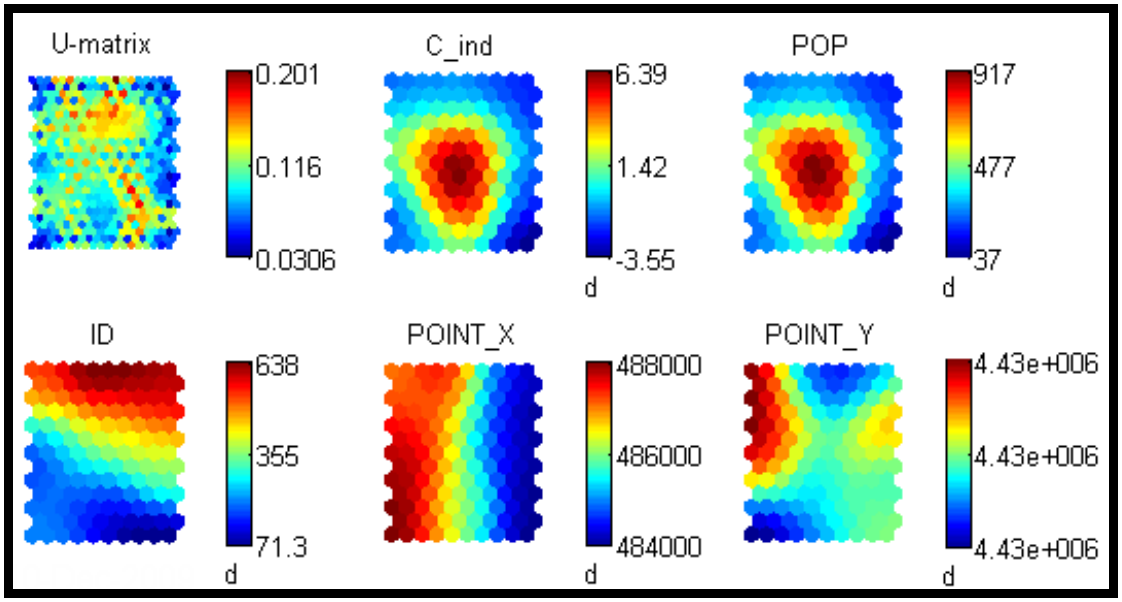


Figure 5.32 Component plane for block ascii file with SES indices.

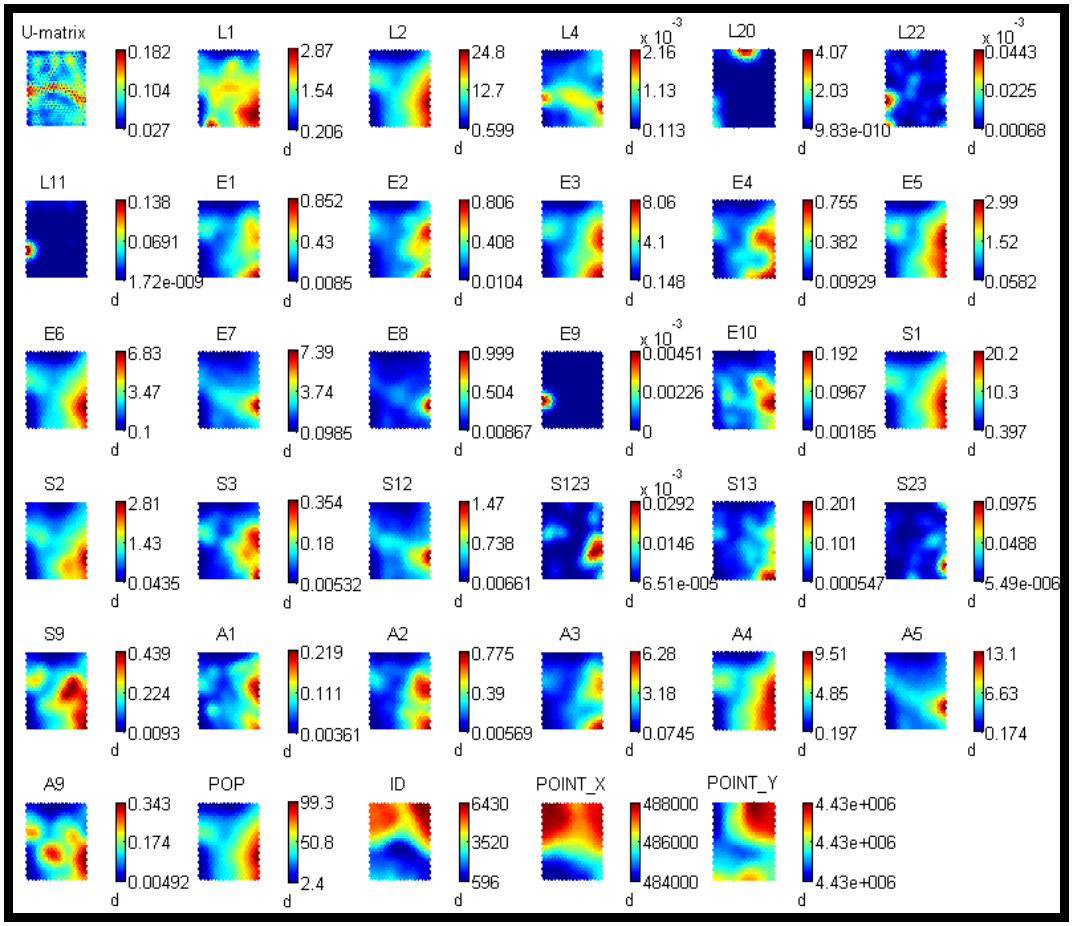


Figure 5.33 Component plane for parcel ascii file with raw data.

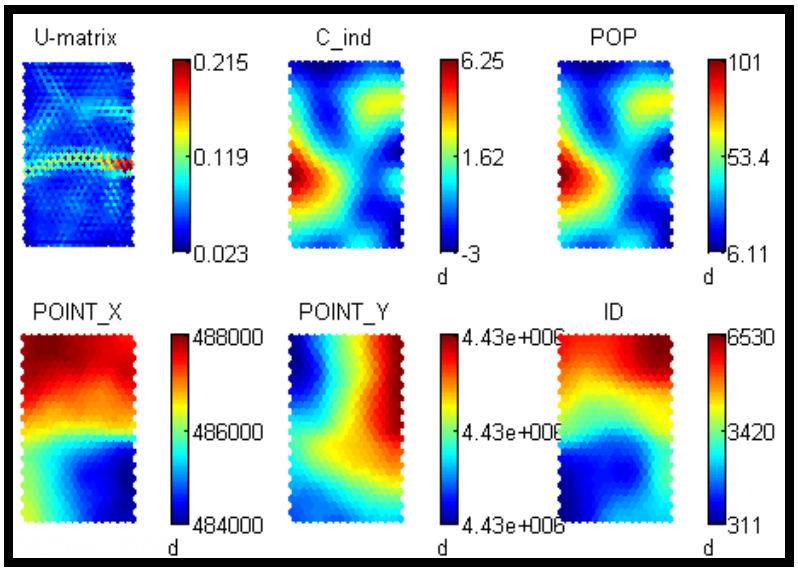


Figure 5.34 Component plane for parcel ascii file with SES indices.

The u-matrix is a visualization of the SOM that illustrates the distance (similarity) between adjacent neurons which carry observations in attribute space. Observations that have similar profiles on input variables are mapped to nearby areas. However, similarity on the synthetic space of SOM is not constant. Some pairs of proximate buckets may hold observations that are more similar than other pairs. U-matrix has hills and troughs which can increase or decrease similarity between proximate buckets. By this way, it shows the cluster structure of data. For example, in Figure 5.35, which shows u-matrices for block ascii files with raw indicators and SES indices, the darker blue areas are most similar since their similarity measure bar on right shows that they have least distance. It is also seen from Figure 5.35 that there are smaller and bigger groups of hexagons with dark blue color in u-matrices indicating similarity. These groups are more or less in the same areas of u-matrices but since different indicators are used for parcel ascii files, the position of these groupings also differentiates in some areas of u-matrices.

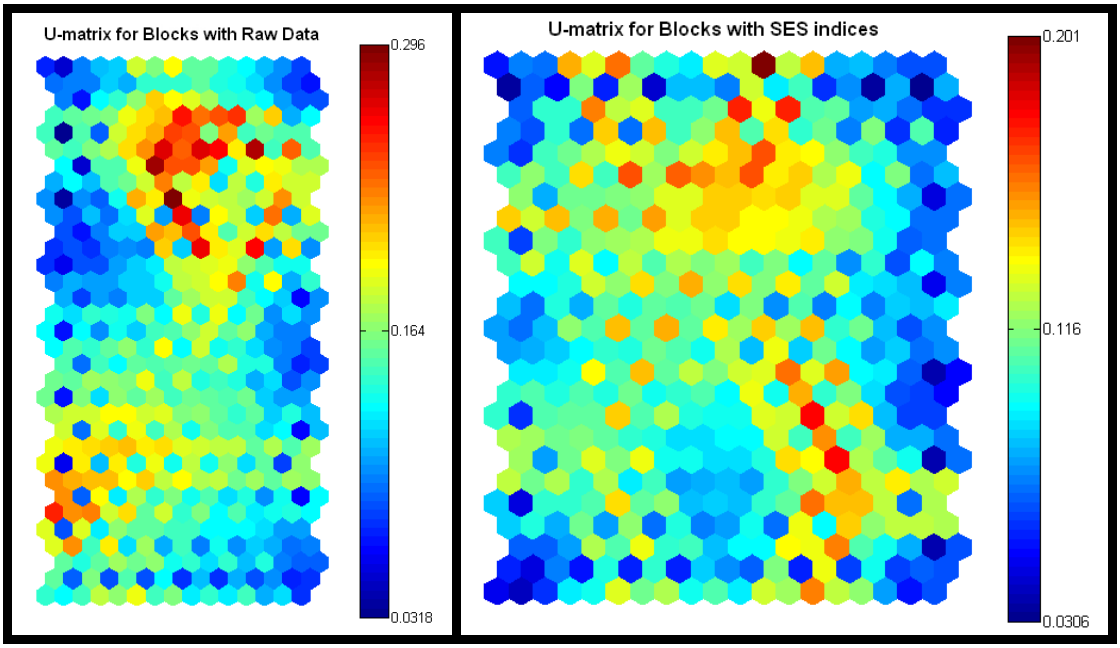


Figure 5.35 U-matrices of block ascii files with raw data and SES indices.

On the other hand, in Figure 5.36, which shows u-matrices for parcel ascii files with raw data and SES indices, the u-matrix with raw data (on the left) has a clear

distinction between south and north as there is a border in the middle. This distinction gets clearer in u-matrix with SES indices (on the right).

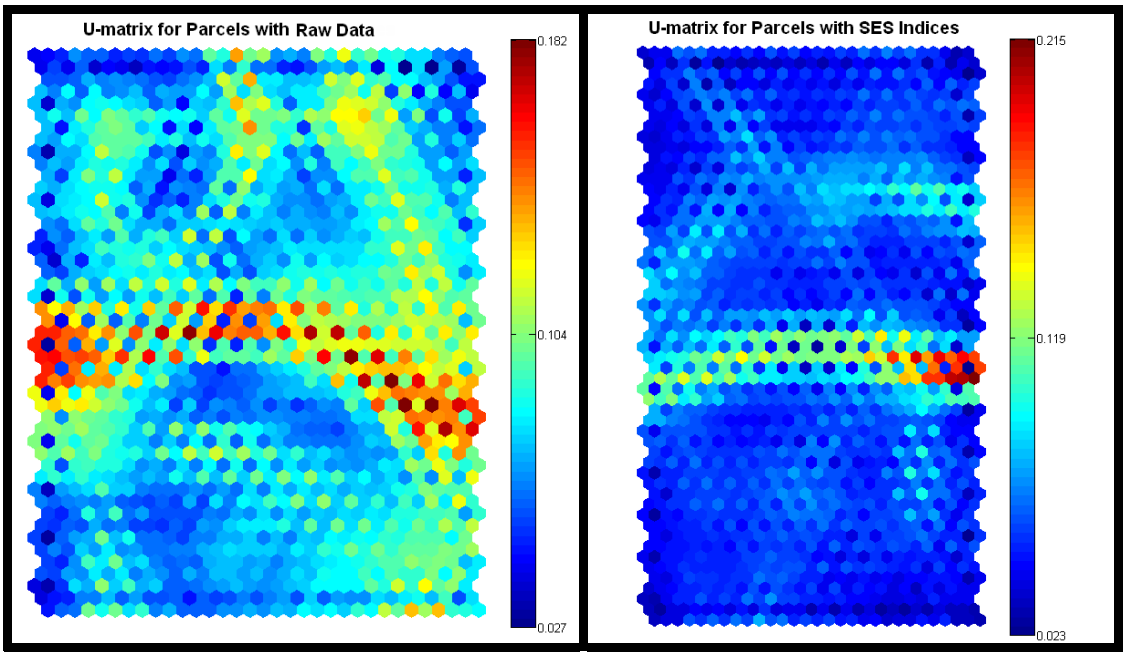


Figure 5.36 U-matrices of parcel ascii files with raw data and SES indices.

Some approaches to automate the interpretation of SOM planes apply additional clustering to the SOM, using classical partitioning methods, for example k-means clustering (Vesanto and Alhoniemi, 2000). Other approaches try to evaluate some common features of several nearby codebook units and color-sizing them. By labeling the number of hits in each SOM unit may also be used where units with low number of hits suggest cluster borders (Kaski and Kohonen, 1990). In this thesis, labeling method, the color-sizing codebook vectors and the k-means partitioning is applied on SOM outputs to interpret results more elaborately.

As a first method, labeling method is implemented. In order to make u-matrices more meaningful and understandable as similarity surfaces, each neuron on u-matrix surfaces can be labeled by the similar observations (number of hits) that a SOM neuron brought together, in this case they are blocks and parcels. In Figure 5.37 and 5.38, examples of colorless surfaces of u-matrices are shown with the unique codes of similar blocks and parcels in each neuron and a particular neuron

zoomed in. We can say that the blocks and parcels in these particular neurons have similar socio-economic and physical indicators or SES indices. The neurons which do not carry any blocks or parcels are the ones that carry low number of hits thus suggesting the cluster borders.

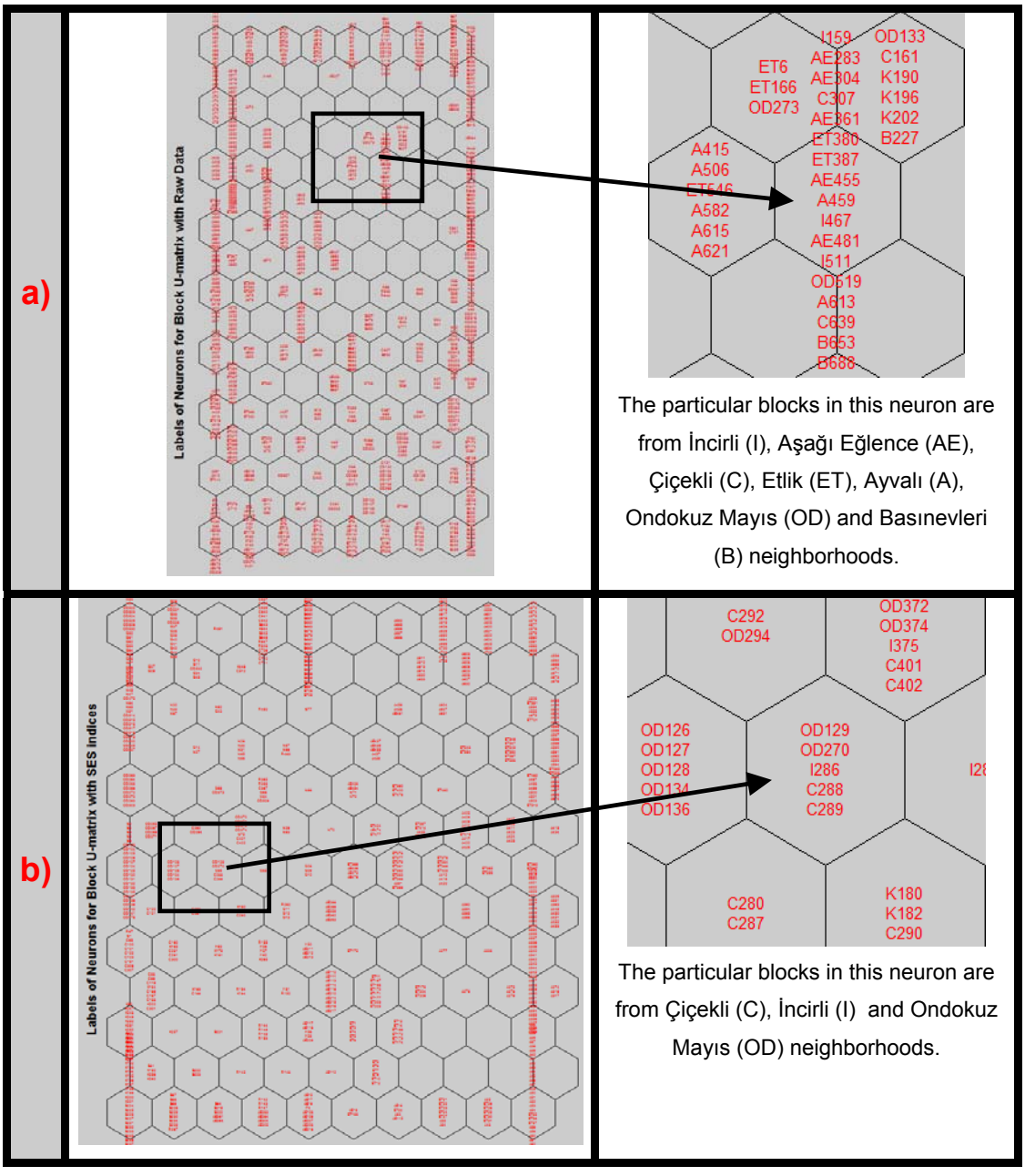


Figure 5.37 Labeling of block u-matrices with a) raw data and b) SES indices by related unique block identity codes.

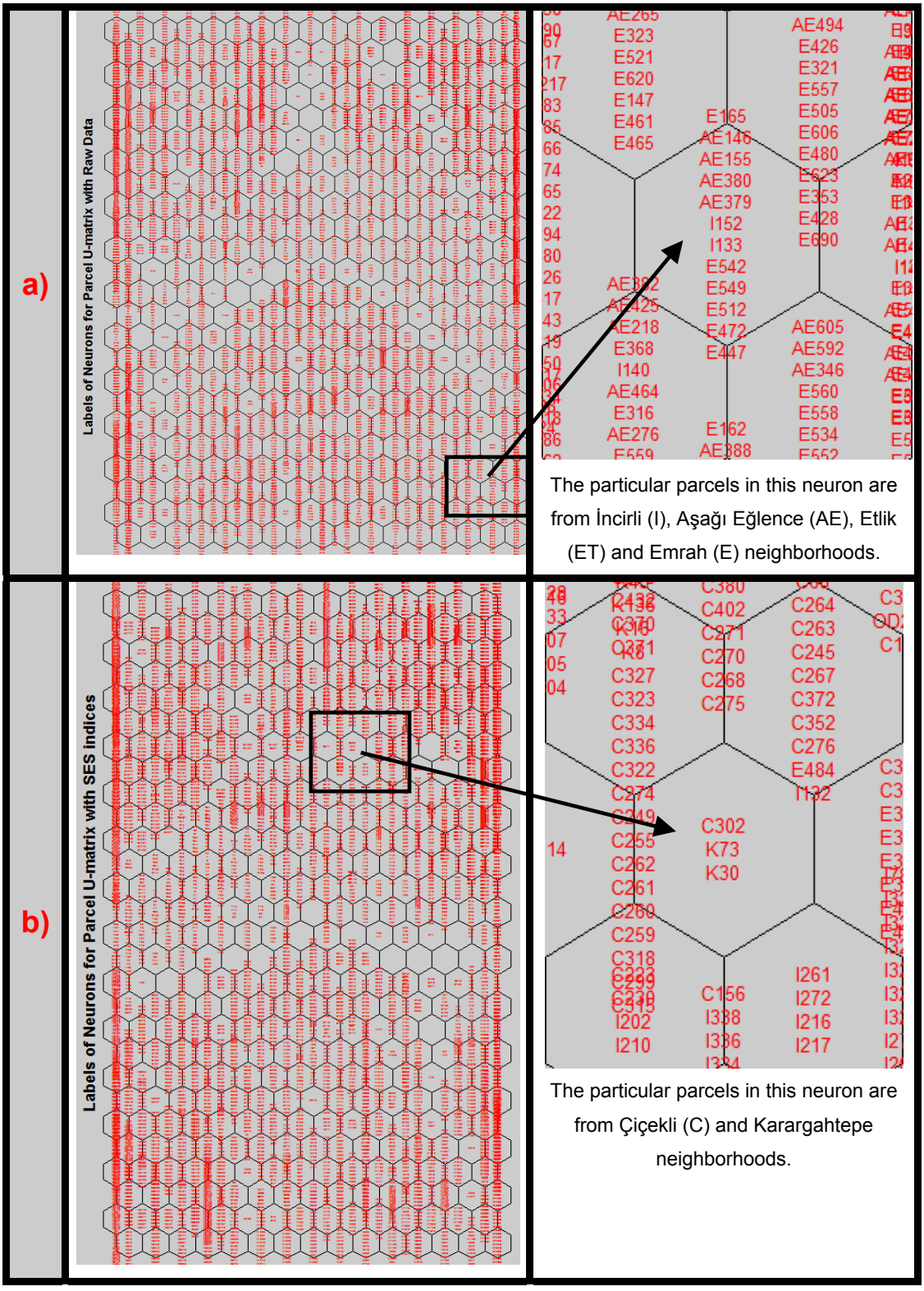


Figure 5.38 Labeling of parcel u-matrices with a) raw data and b) SES indices by related unique identity codes.

As a second method, the projection of colored-coded and resized SOM neurons for block and parcel ascii files offers a clearer view of clustering. Similar data items are grouped together with the same type of color and size. Size, position and color of markers can be used to depict the relationships between data items. For example the small sized neurons, as in the method of labeling, are the ones that carry low amount of blocks and parcels, thus suggesting the cluster borders. This gives an informative picture of the global shape of and overall smoothness of the SOM plane (Figure 5.39 and 5.40 - a and b on the left). On the other hand, dispersion matrices can be used to analyze the neurons which are least and most similar by looking at the position they take on x-y. Also, it is possible to see the outlier neurons on these dispersion matrices (Figure 5.39 and 5.40 - a and b on the right).

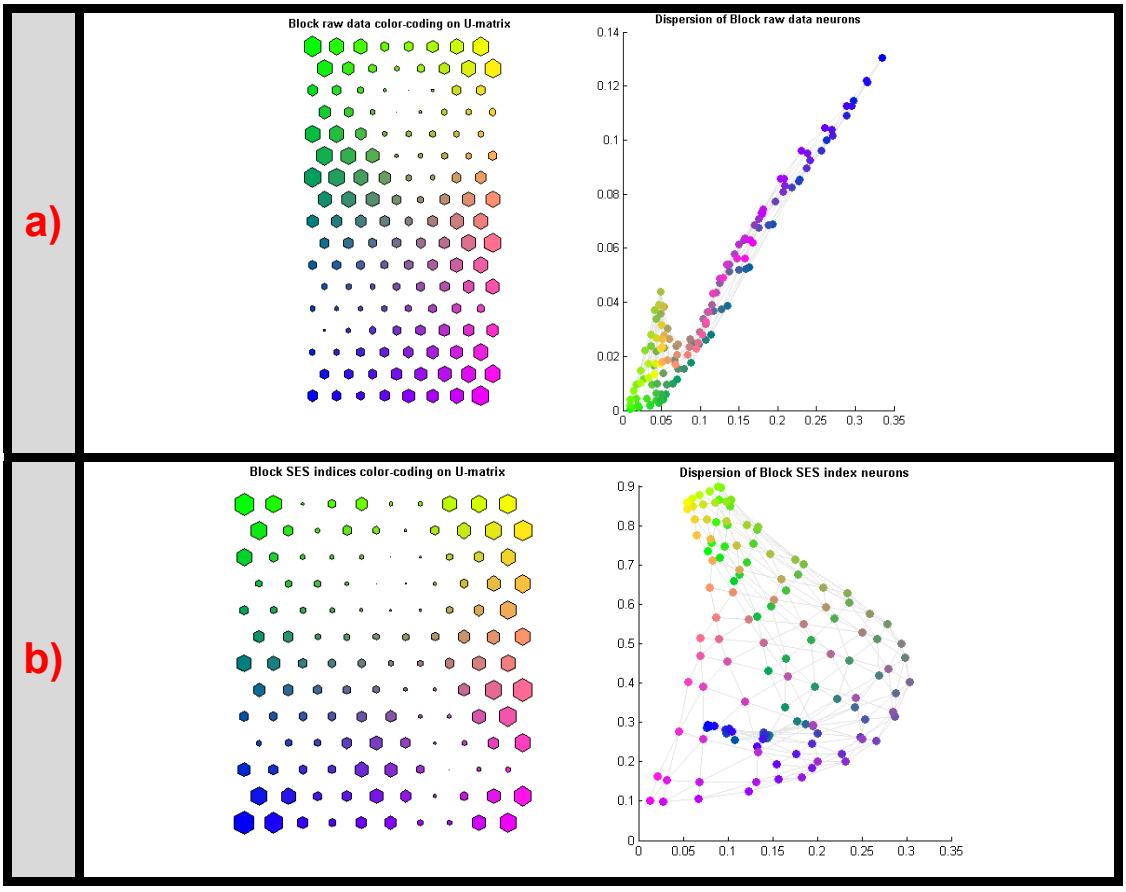


Figure 5.39 a) Block raw data with colorcodes on u-matrix (on the left) and dispersion of neurons on x-y plane (on the right) b) Block SES data with colorcodes on u-matrix (on the left) and dispersion of neurons on x-y plane (on the right).

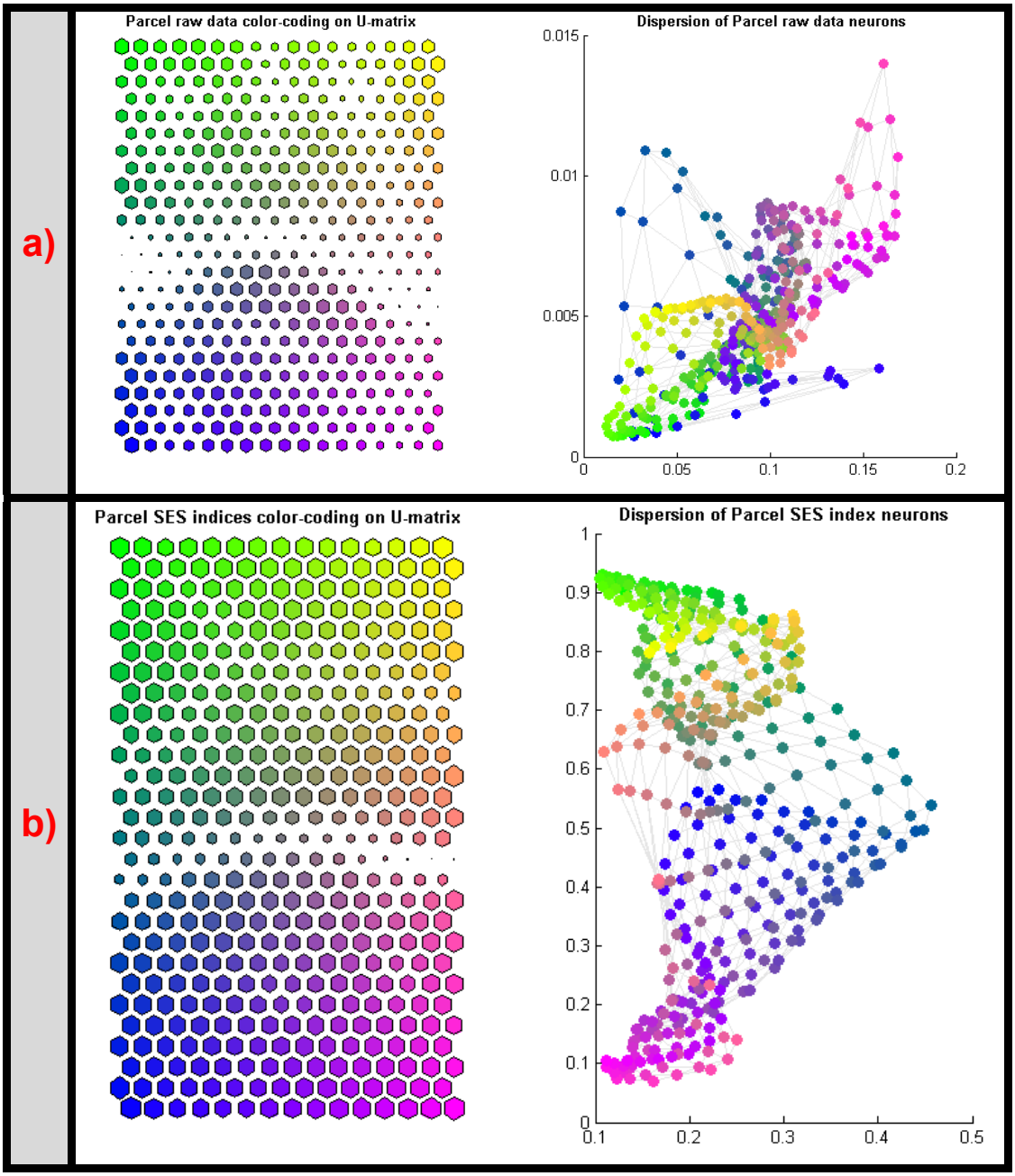


Figure 5.40 a) Parcel raw data with colorcodes on u-matrix (on the left) and dispersion of neurons on x-y plane (on the right) b) Parcel SES data with colorcodes on u-matrix (on the left) and dispersion of neurons on x-y plane (on the right).

As a third method, k-means partitioning is applied. The color-coding can be further enhanced by applying k-means clustering by SOM Toolbox 2.0 on Matlab. K-means clustering automatically found 10 clusters on SOM u-matrices for both blocks with raw data and SES indices, 21 clusters on SOM u-matrices for parcels with raw data data and 14 clusters for parcels with SES indices after selecting the best by considering the quality assessment based on root mean squared error (RMSE) after few trials. The clusters found on u-matrices can be seen in Figures 5.41, 5.43, 5.45 and 5.47 on the right. But these clusters are found on the SOM grids, not on geographical maps. Therefore, the clusters are found based on the locations and similarity of neurons, not considering the similarity and coordinates of actual parcels and blocks. The reflection of clustering results for both parcel and block ascii files including raw data and SES indices on actual maps are retrieved by extra programming on MATLAB for which codes are is given in Appendix C (Figure 5.42, 5.44, 5.46 and 5.48). The results of k-means clustering of u-matrices reflected on actual maps of parcels and blocks also support the idea that this type of clustering is only schematic. Because, the thematic block and parcel maps generated according to cluster numbers show a fragmented structure even if the clustering algorithm that was used upon u-matrices is k-means. But, if we look at the clustering on u-matrices, the clusters are perfectly separated. The thematic maps of parcel and block clusters retrieved from SES indices (Figure 5.44 and 5.48) for both parcels and blocks are far less fragmented than thematic maps of parcel and block clusters retrieved from raw data (Figure 5.42 and 5.46). These results from the fact that SES indices give a composite value related to all raw socio-economic and physical indicators, therefore it is easier to differentiate similar neurons according to a single indicator instead of 31 raw data indicators.

As a result, in this thesis, it is proposed to implement spatial clustering on the similarity measures (u-distances) which are found for each parcel and block as a result of SOM analysis done for raw data and SES indices separately, not on the similarity measures of neurons on u-matrix grids.

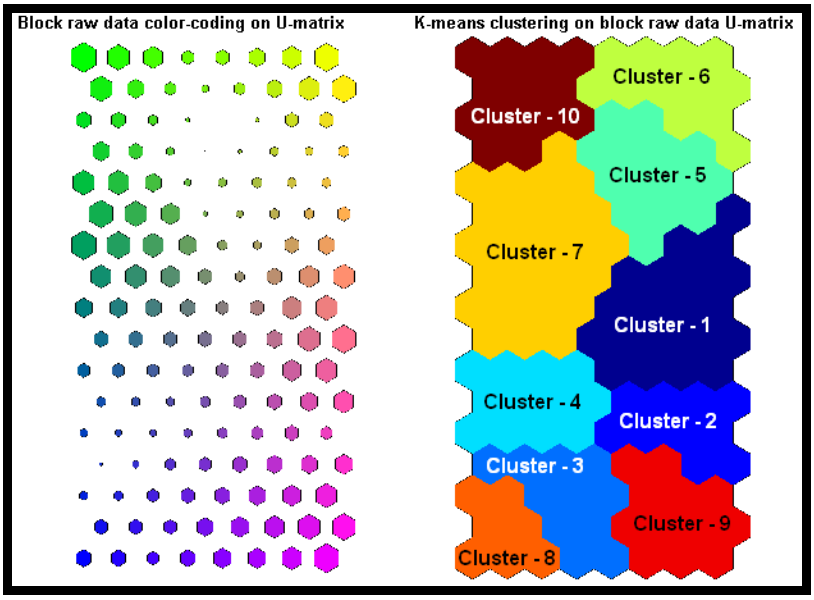


Figure 5.41 Color-coding on block raw data u-matrix (on the left) and k-means clustering on block raw data u-matrix plane (on the right).

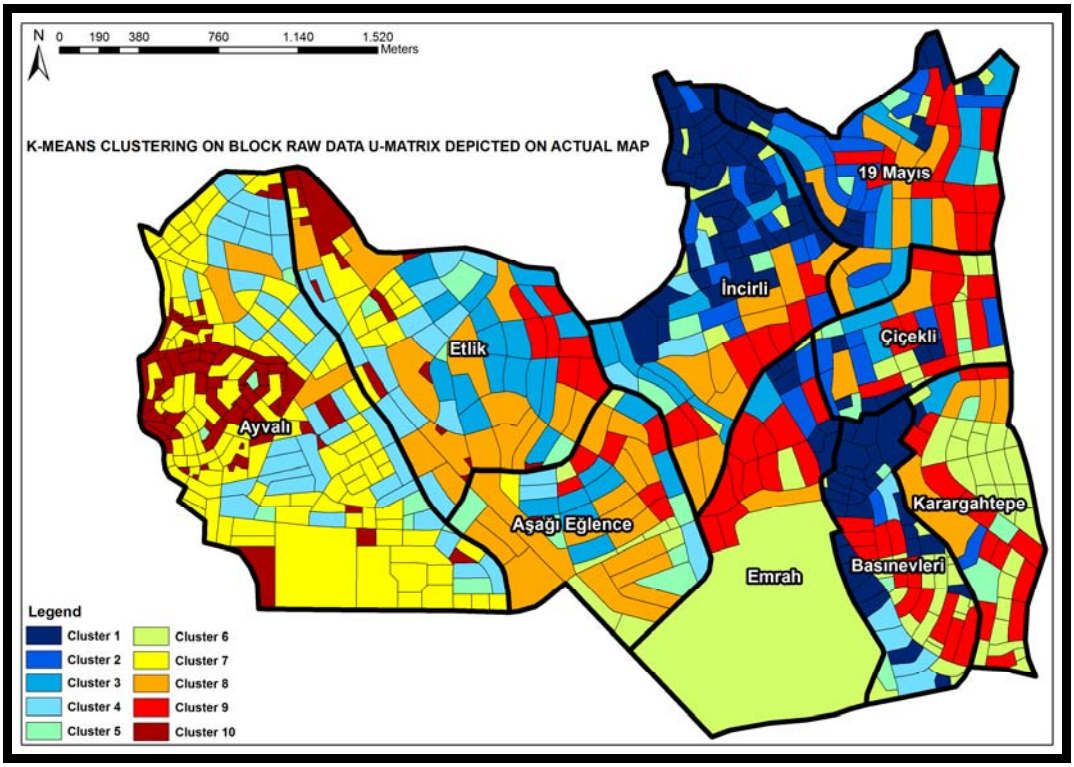


Figure 5.42 Representation of k-means clustering implemented on block raw data u-matrix in actual block map.

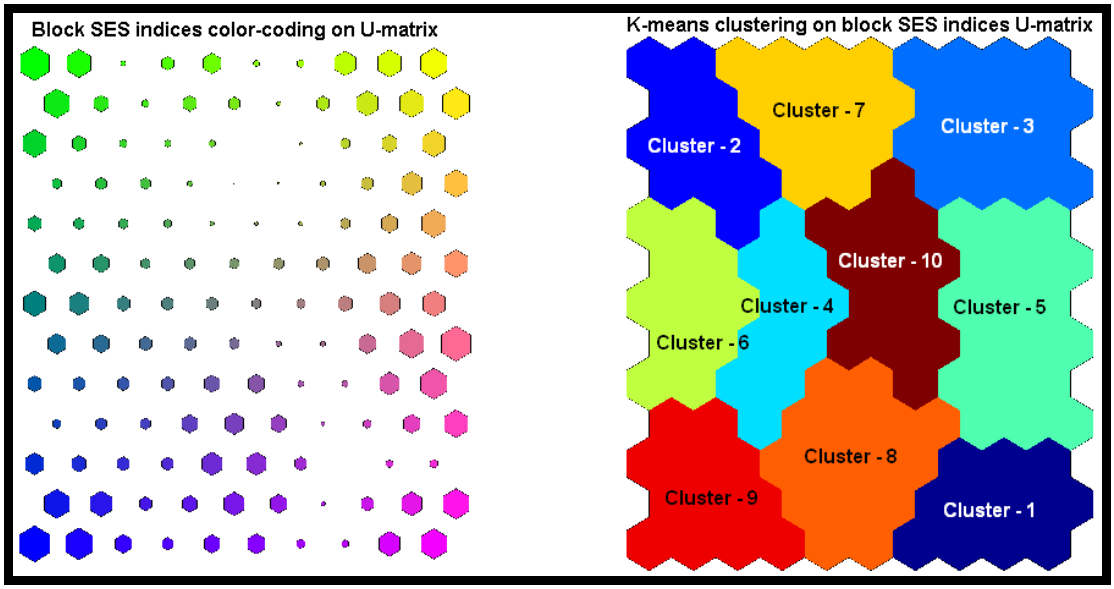


Figure 5.43 Color-coding on block SES indices u-matrix (on the left) and k-means clustering on block SES indices u-matrix (on the right).

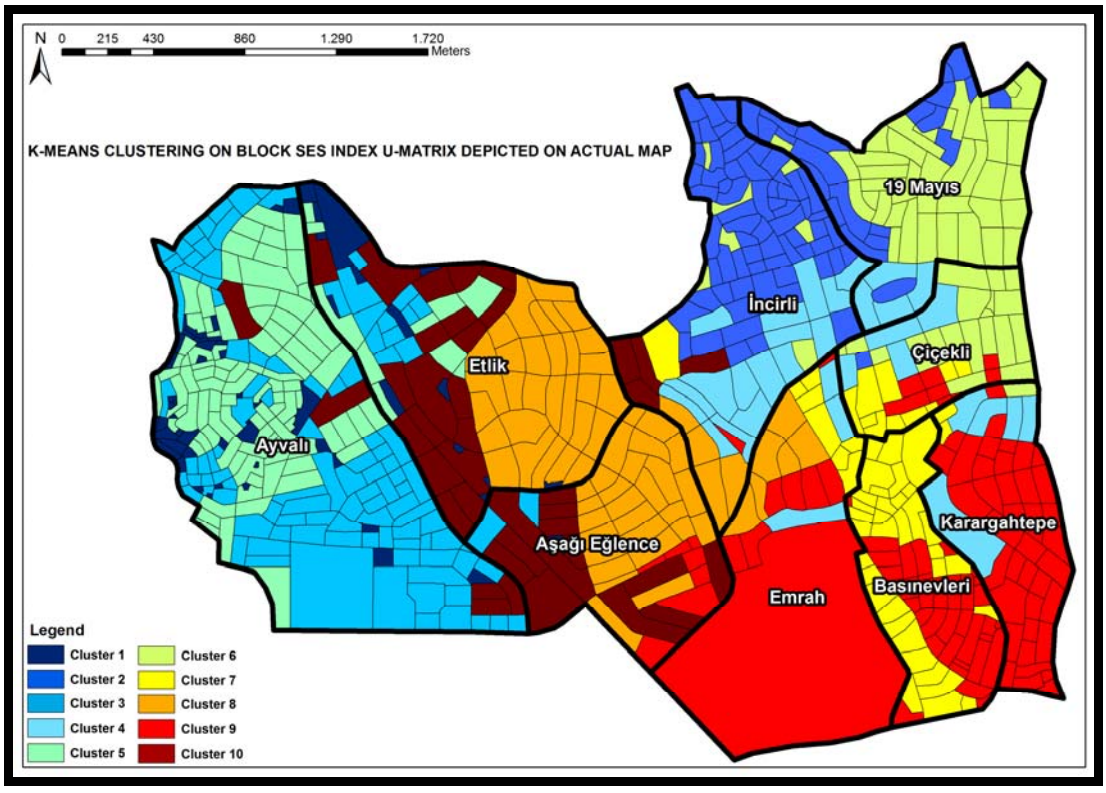


Figure 5.44 Representation of k-means clustering implemented on block SES indices u-matrix in actual parcel map.

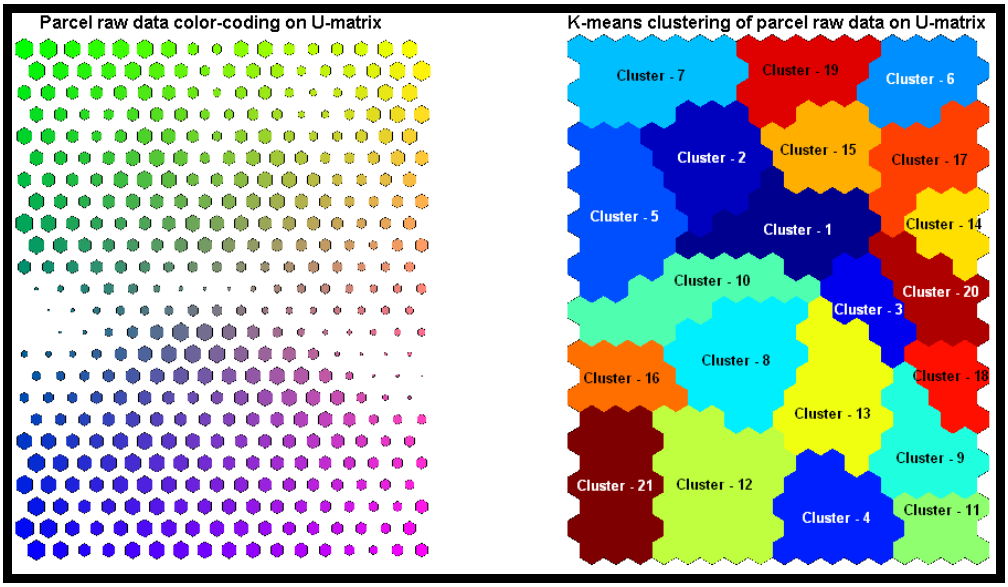


Figure 5.45 Color-coding on parcel raw data u-matrix (on the left) and k-means clustering on parcel raw data u-matrix (on the right).

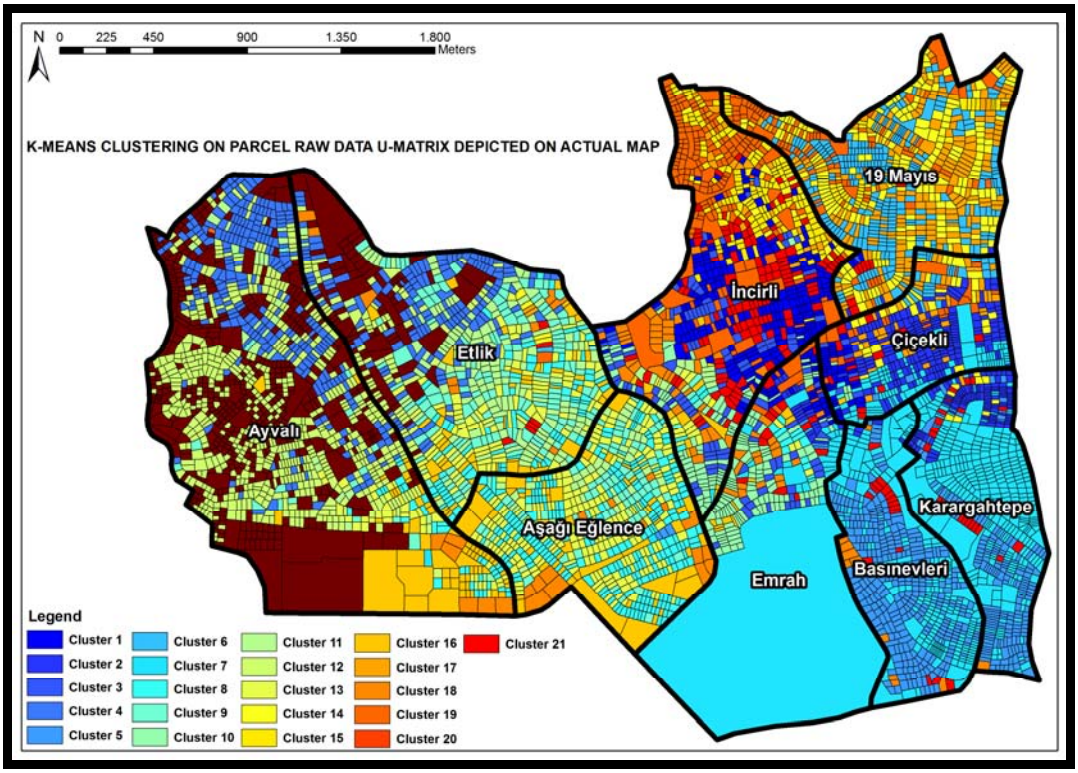


Figure 5.46 Representation of k-means clustering implemented on parcel raw data u-matrix in actual parcel map.

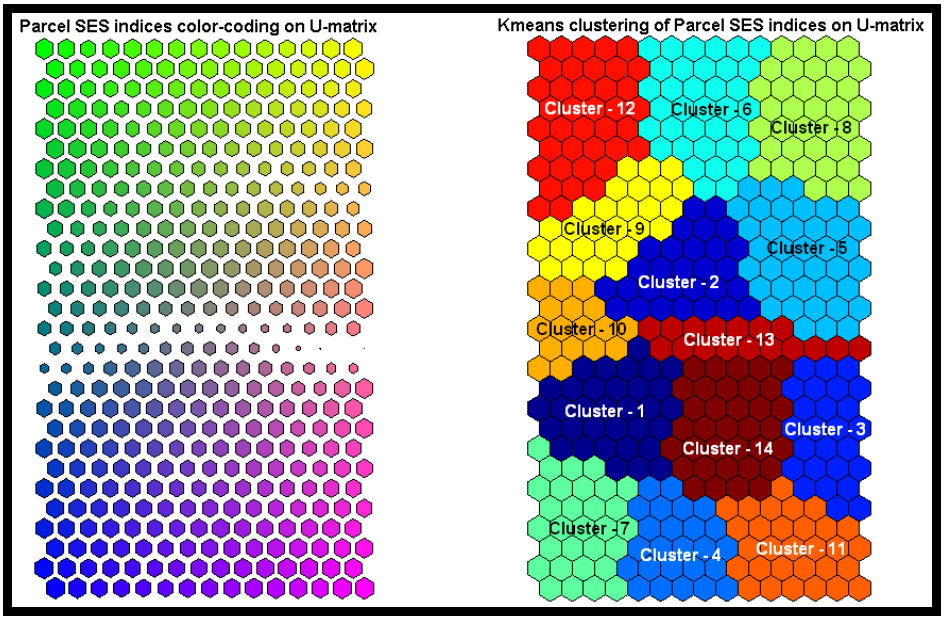


Figure 5.47 Color-coding on parcel SES indices u-matrix (on the left) and k-means clustering on parcel SES indices u-matrix (on the right).

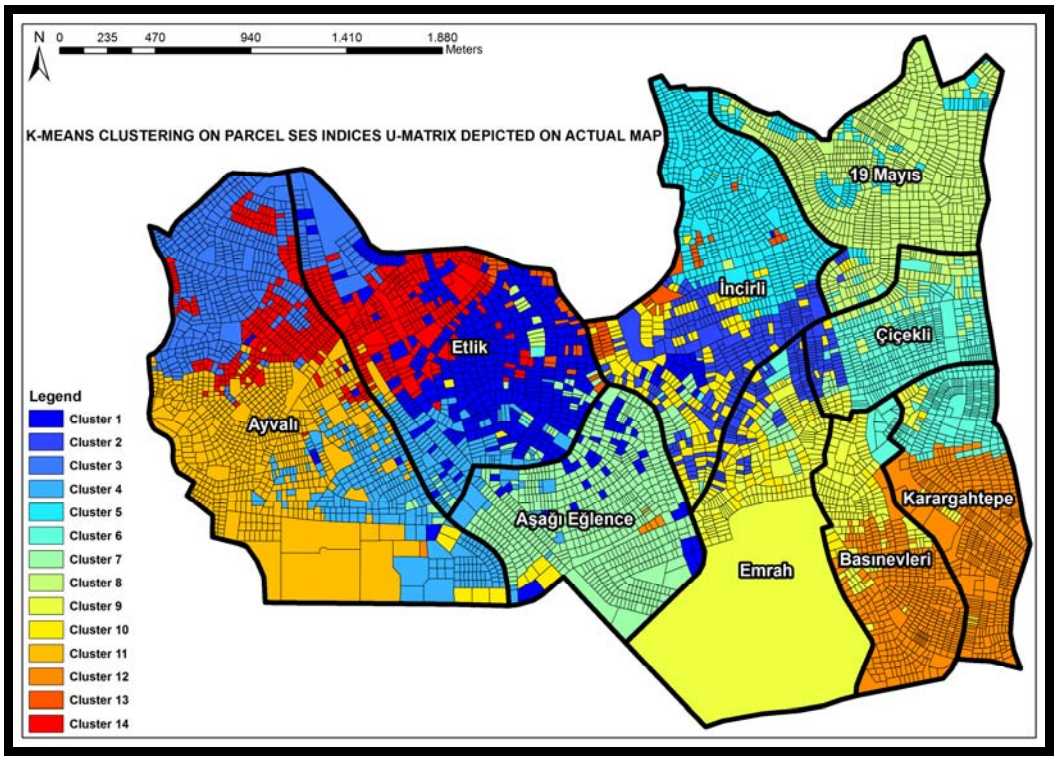


Figure 5.48 Representation of k-means clustering implemented on parcel SES index u-matrix in actual parcel map.

While unified similarity matrix representation is a good method for visualizing clusters, it does not provide a clear picture of the overall shape of the data space because the visualization is tied to the SOM grid. An alternative approach is to work backwards, that is by selecting a group of similarly colored neurons and examining where they fall on the SOM feature map. This process necessitates creating a link between the u-matrices and the actual maps. Through the training process, each parcel or block is supposed to have a best matching unit (BMU) from the set of neurons within the SOM grid. It helps to set up a linkage between SOM neurons and the corresponding parcels and blocks. But finding the BMUs for each parcel and block is not enough, because this thesis requires the similarity values in other words unified distances of each parcel and block in order to use them as an input to obtain SSAs. So, besides available m-files of SOM toolbox 2.0, another m-file has been created by Matlab coding language in order to derive the BMUs and unified distances that each parcel and block value in 4 ascii files supposed to have. These codes can be found in Appendix C. The process can be summarized step-by-step as below;

1) Firstly, the BMUs of all 711 blocks and 6858 parcels with raw data and SES indices separately are derived by a specially written program code as matrices and these BMUs are given an index number (FID) starting from 0. These index numbers follow the same order with parcels' and blocks' ascii data which were used at the beginning of the analysis. These matrices are saved as a text file in the SOM working directory as "bmus_block_raw.dat", "bmus_block_ses.dat", "bmus_parcel_raw.dat" and "bmus_parcel_ses.dat" (Figure 5.49).

BMUs of blocks with raw data

FID	BMUs
0	a
1	a
.	.
.	.
500	d
501	c
.	.
710	c

BMUs of parcels with raw data

FID	BMUs
0	!
1	!
.	.
.	.
500	*
501	*
.	.
6857	#

BMUs of blocks with SES indices

FID	BMUs
0	aa
1	aa
.	.
.	.
500	dd
501	cc
.	.
710	cc

BMUs of parcels with SES indices

FID	BMUs
0	!!
1	!!
.	.
.	.
500	**
501	**
.	.
6857	##

Figure 5.49 Table structure for "bmus_block_raw.dat", "bmus_block_ses.dat", "bmus_parcel_raw.dat" and "bmus_parcel_ses.dat".

2) Then, the parcel and block u-matrix codebook vectors, in other words neurons, are numbered starting from 1 again by a program code. These numbers are also the BMU codes of the neurons on the SOM grid. There were 136 neurons on the block raw data u-matrix, 130 neurons on block SES indices umatrix, 408 neurons on parcel raw data u-matrix and 420 neurons on the parcel SES indices u-matrix.

3) Thirdly, the unified similarity value is derived from each codebook vector/neuron from each u-matrix by a specially written m-file on Matlab. These are the values by which SOM toolbox uses to color the u-matrices and are the key for this thesis.

4) On the fourth step, the BMU codes of neurons and related unified distance values are created as a matrix. Then, this matrix is saved as a text file as "codebook_block_raw.dat","codebook_block_ses.dat","codebook_parcel_raw .dat", and "codebook_parcel_ses.dat". (Figure 5.50).

U-distances of neurons on block raw data u-matrix

BMUs	UDISTANCE
a	0,0694460
b	0,0938190
c	0,1454600
d	0,1275100
e	0,0985030
f	0,0823670
g	0,0762780
h	0,1083400

U-distances of neurons on block SES indices u-matrix

BMUs	UDISTANCE
!	0,0535840
%	0,0665760
*	0,0710150
^	0,1172900
#	0,1052100
&	0,0927270
]	0,0763500
[	0,0901510

U-distances of neurons on parcel raw data u-matrix

BMUs	UDISTANCE
aa	0,0436030
bb	0,0561200
cc	0,0743660
dd	0,0743660
ee	0,0651140
ff	0,0724220
gg	0,0732810
hh	0,0661640

U-distances of neurons on block SES indices u-matrix

BMUs	UDISTANCE
!!	0,0359280
%%	0,0443980
**	0,0424920
^^	0,0425070
##	0,0388890
&&	0,0397360
]]	0,0412290
[[	0,0457600

Figure 5.50 Table structure for "codebook_block_raw.dat", "codebook_block_ses.dat", "codebook_parcel_raw.dat" and "codebook_parcel_ses.dat".

The text files are converted to database files (*.dbf) and imported to ArcGIS 9.2 environment. The codebook tables and bmus tables are one-to-many joined by using the common column named "BMUs". This one-to-many join operation is illustrated for block raw data bmus and codebook tables in Figure 5.51. The output tables retrieved from join operations are saved as "bmu_code_block_raw. dbf", "bmu_code_block_ses.dbf", "bmu_code_parcel_raw.dbf" and "bmu_code_parcel_ses.dbf" respectively.

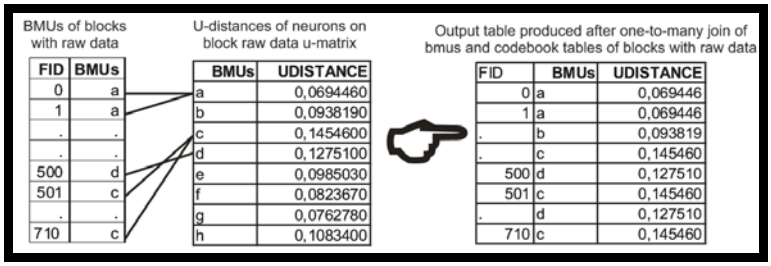


Figure 5.51 One-to-many join operation between codebook and bmus tables related to blocks with raw data.

5) All u-matrix hexagonal surfaces of parcel and block data with raw and SES indices are created by a tool called "Geo-SOM" (Web 3) (Bacao et al, 2005-d)) with same number of neurons and exported as a shapefile to ArcGIS 9.2 environment. The neurons on these planes are labeled by their BMU numbers in their attribute table. The SOM surfaces are joined with the output join tables created in step 4 by using their common column named "BMUs" and the neurons on these SOM surfaces are classified according to unified distance values into 5 classes using natural breaks classification (Table 5.12). In this way, we accomplished to create similar surfaces in ArcGIS 9.2 to the u-matrices for blocks and parcels created by SOM Toolbox 2.0 using raw data and SES indices separately.

Table 5.12 Visualization of SOM a) block raw data; b) block SES index; c) parcel raw data and d) parcel SES index u-matrices by ArcGIS 9.2.

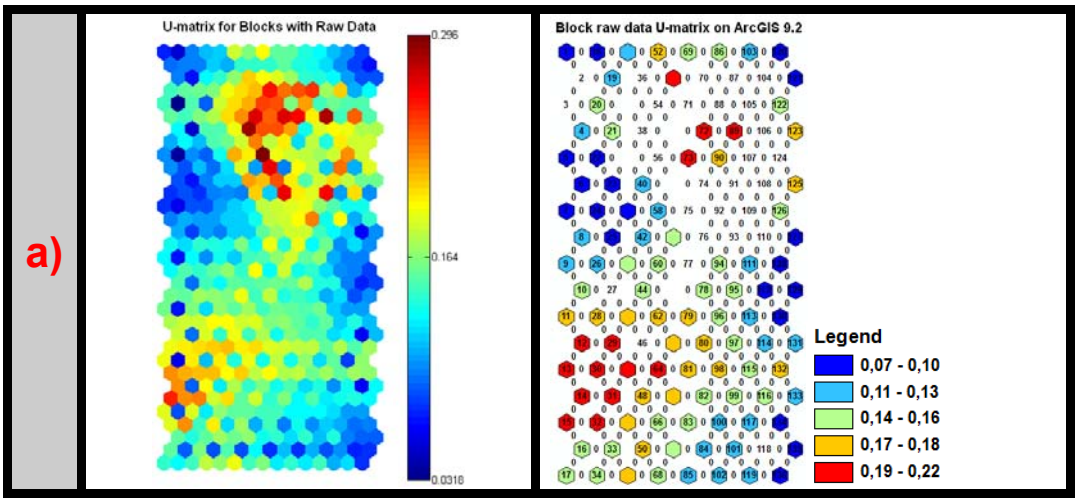
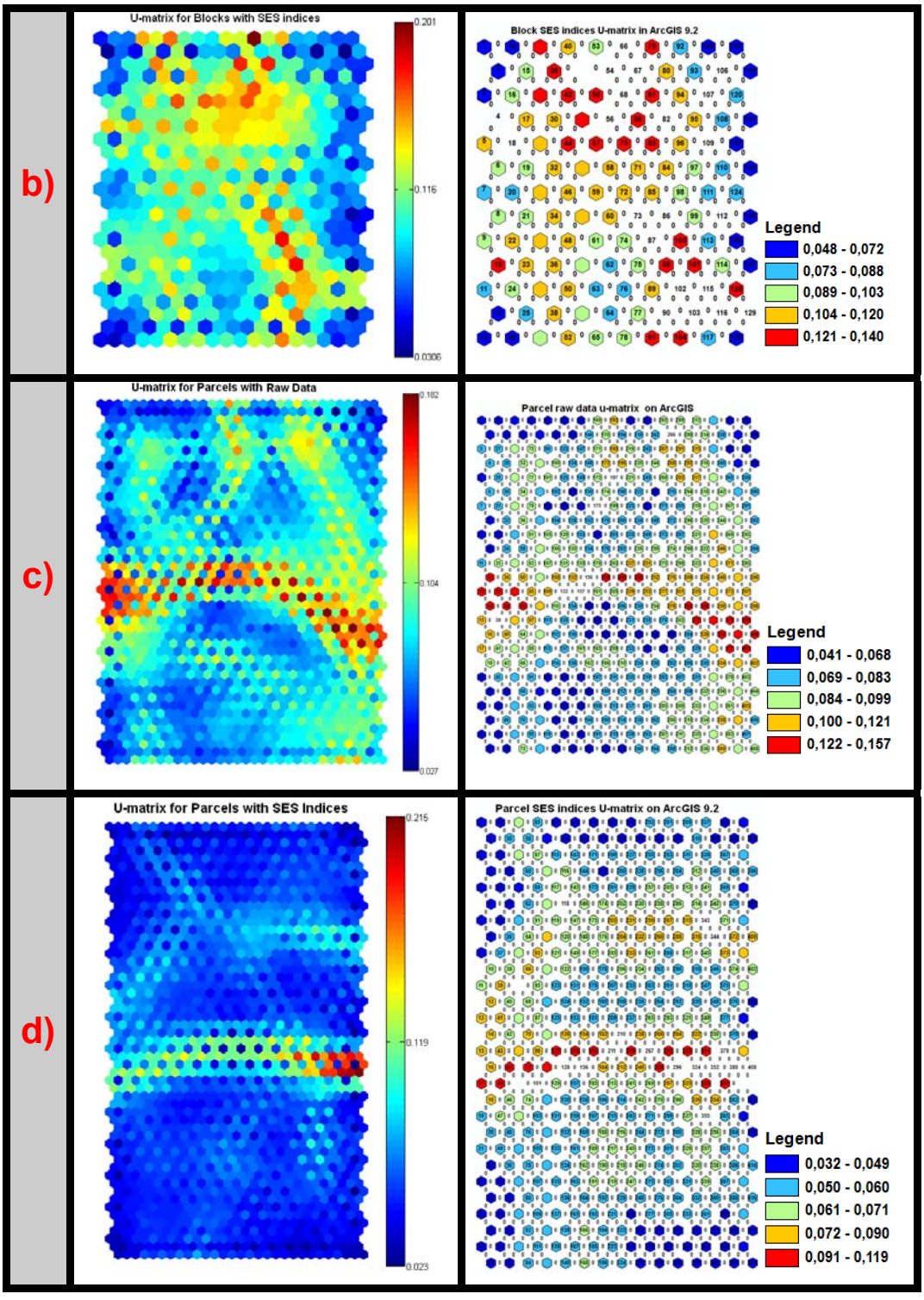


Table 5.12 (Continued) Visualization of SOM a) block raw data; b) block SES index; c) parcel raw data and d) parcel SES index u-matrices by ArcGIS 9.2.



6) The output join tables produced in step 4 are also joined with actual parcel and block shapefiles using the “FID” column commonly. Then, similar to step 5 the parcels and blocks are categorized into 5 classes according to unified distance values using the natural breaks classification (Figure 5.52 – 5.55)

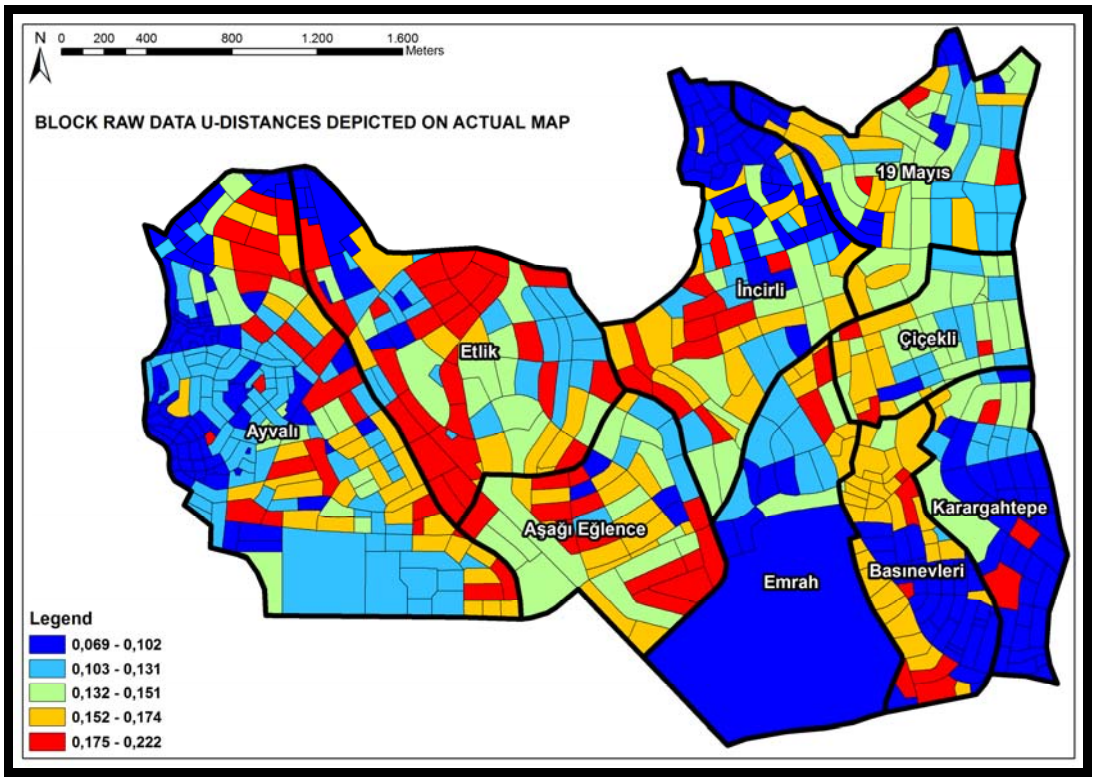


Figure 5.52 Classification of block shapefile according to unified distance values on u-matrix of block raw data.

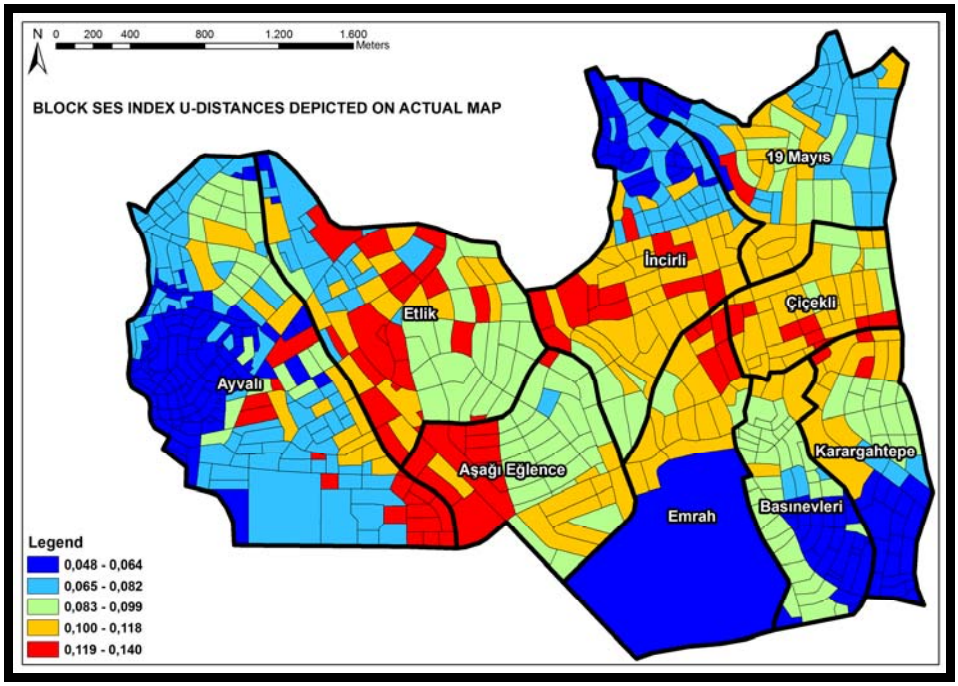


Figure 5.53 Classification of block shapefile according to unified distance values on u-matrix of block SES index.

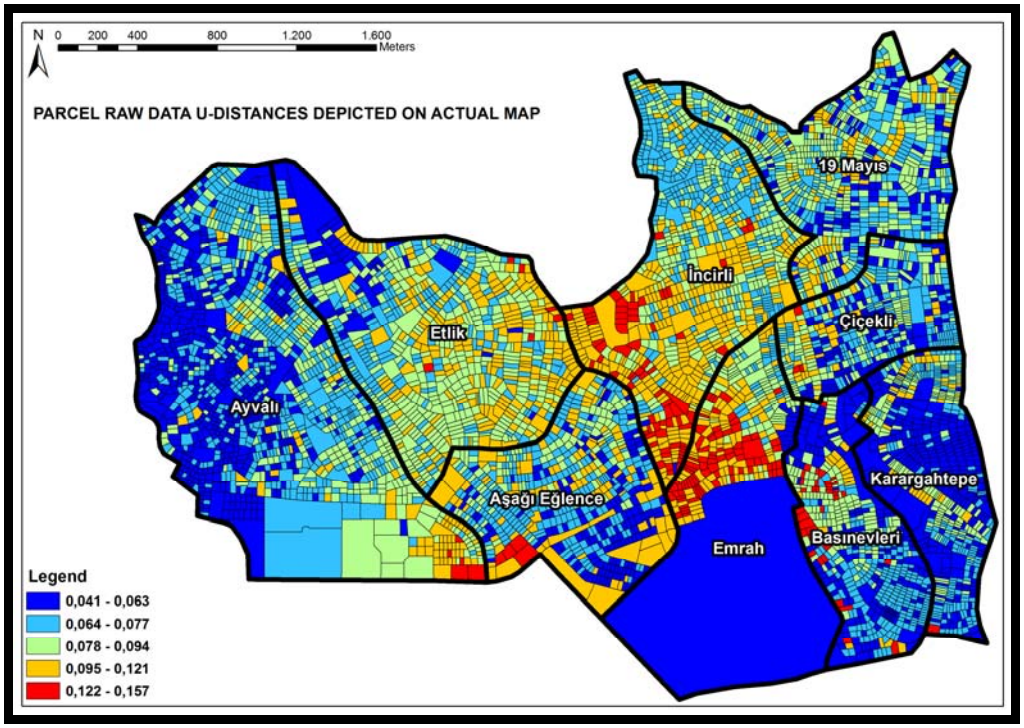


Figure 5.54 Classification of parcel shapefile according to unified distance values on u-matrix of parcel raw data.

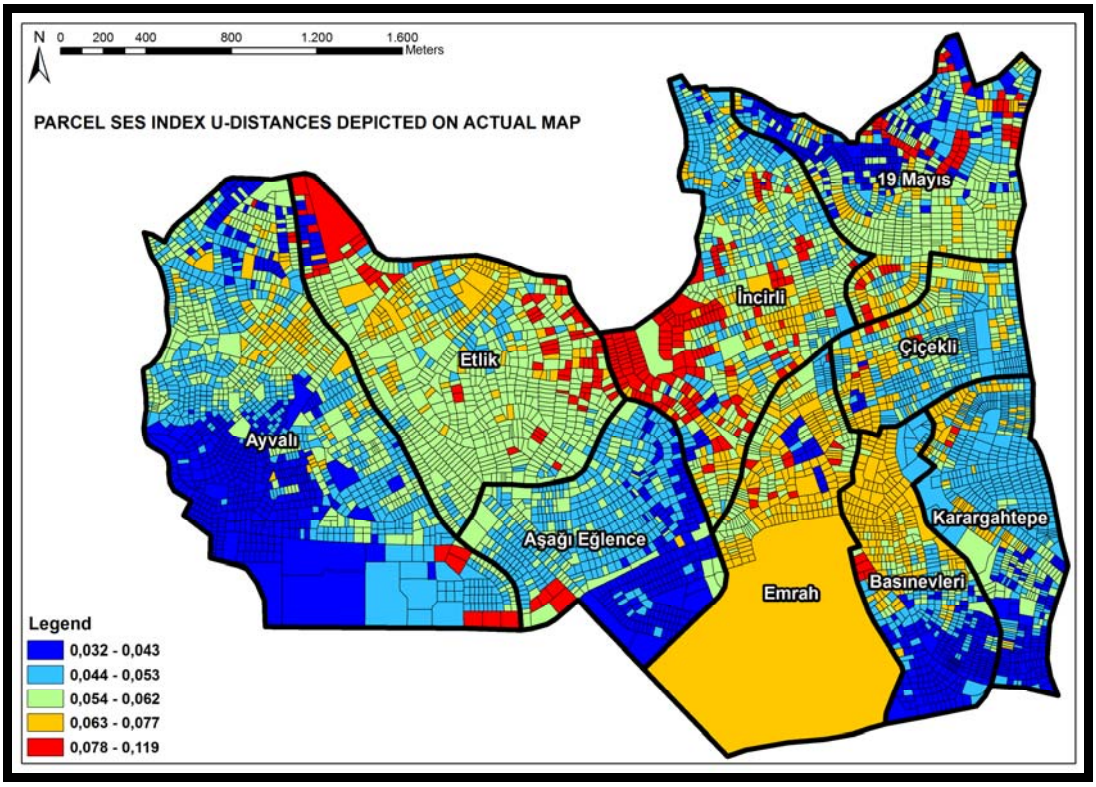


Figure 5.55 Classification of parcel shapefile according to unified distance values on u-matrix of parcel SES indices.

As can be seen from Figures 5.52 - 5.55, contrary to block and parcel shapefiles colored according to u-distances retrieved from block and parcel raw data, the block and parcel shapefiles colored by u-distances from block and parcel SES indices give a more distinctive visualization between classes. Nevertheless, the classes in all figures separate certain areas. For example, in Figure 5.53, most parts of Ayvalı, north of İncirli, east of Ondokuz Mayıs and south parts of Emrah, Basınevleri and Karargahtepe significantly differentiates from the rest. Additionally, in Figure 5.53, the light green areas at Etlik and Aşağı Eğlence form a cluster since these blocks are the areas where there is a high commercial activity. Apart from these, the light green colored blocks in Karargahtepe are the blocks which consist of one or two storey residential buildings. The dark blue colored area in west Ayvalı was formerly a squatter area, but now this location is reconstructed according to construction improvement plans at a great extent. However, the light blues colored blocks in north of Ayvalı, Etlik and İncirli are still in process of transformation from squatter

areas. Besides squatters, there are many ongoing construction sites on these areas. The rest of the map colored with yellow and red in Figure 5.53 is consisting of mainly the transition zones and have a mixed urban pattern.

By selecting an area of interest on the SOM grid created on ArcGIS 9.2, one can examine how it maps on the actual map. Reversing the process, selecting the parcels or blocks that are estimated as similar, lets one quickly visualize where they fall in the SOM grid. For example, in Figure 5.56, the selection which is made on the parcel SES indices u-matrix is shown on actual parcel map.

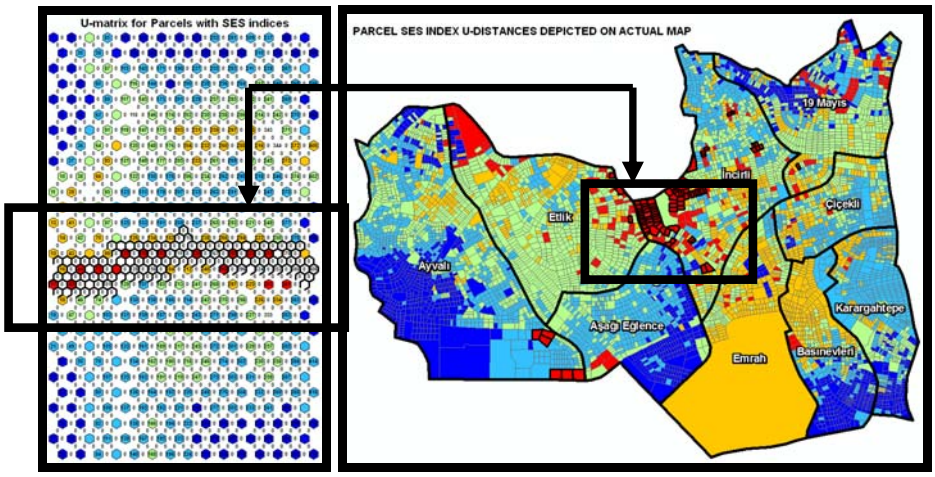


Figure 5.56 Relationship between u-matrix of parcel SES indices and parcel map.

In order to find out the quality of SOM application implemented on parcel and block datasets, quantization and topographic errors are queried by SOM Toolbox. The results are shown on Table 5.13. As it was mentioned in Chapter 3, it is preferred that these quality measures happen between [0, 1]. Therefore, it can be interpreted that all SOM applications are acceptable according to below values. In addition, the quantization errors of SOMs created by using SES indices are relatively smaller than SOMs created by using raw data. However, contrary to this situation, the topographic errors of SOMs with SES indices are larger than the SOMs with raw data. The quantization error is smaller because the SOMs with SES indices use only two indicators lessening the input data vector which are SES indices and population, for this reason the distance (similarity) between a codebook vector (neuron) and an input data vector is smaller. On the contrary, the topographic error is larger because

the topology preservation is lower for the SOMs created with SES indices. This means that for an input data vector the BMUs are not adjacent to each other on some areas of a U-matrix.

Table 5.13 Quality measures for parcel and block SOMs..

	Quantization Error	**Topographic Error**
SOM created for blocks with raw data	0,4330	**0,0155**
SOM created for blocks with SES indices	**0,1561**	0,0225
SOM created for parcels with raw data	0,3495	**0,0190**
SOM created for parcels with SES indices	**0,0995**	0,0236

However, as can be seen from Figures 5.52 - 5.55, the parcels and blocks with u-distances are still not small areas even if they are are clustered I similar classes. In order to obtain SSAs using u-distances as similarity values which were derived from raw data and SES indices as a result of SOM applications on parcels and blocks, simulated annealing method is again implemented by using BARD. Firstly, shapefiles for blocks and parcels in each 9 neighborhood is created including the unified distances derived from SES indices or raw data separately for blocks and parcel ascii files and population counts as attribute information. 36 shapefiles are prepared as a result of this process (Figure 5.57).

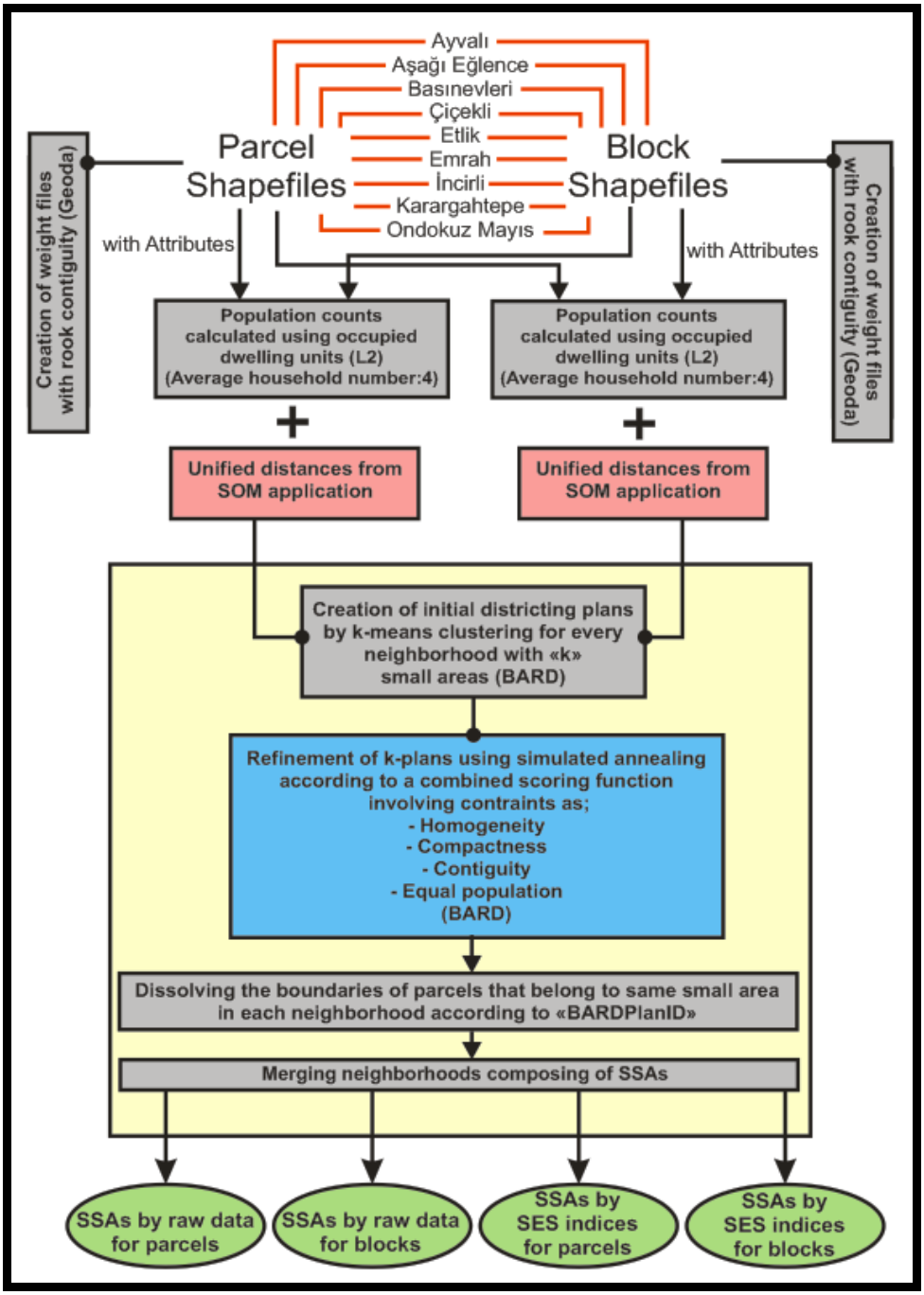


Figure 5.57 Flow-chart for the third method: Simulated annealing on k-means clustering of u-distances from SOM applications.

Secondly, similar to first method, k-means clustering is implemented on these 36 shapefiles using the small area numbers calculated in Table 5.10 for each neighborhood according to the household number that each neighborhood have.

Then, simulated annealing is used to refine the initial k-means plans with the same combined scoring function used for the first method. This combined function is used to calculate the best score that satisfies constraints such as compactness, contiguity, equal population and homogeneity for different small area maps created through annealing iterations. The final small areas are dissolved from parcels and blocks in each neighborhood using the "BARDPlanID". But, before dissolving parcels and blocks, in order to compare the SSAs from third method with the plans from the first and the second method on a common basis, the SES indices are joined with the output parcels and blocks obtained from third method. Eventually, the neighborhoods are merged. The output SSA maps can be seen in Figures 5.58 - 5.61 in which the SSAs are colored by the mean SES indices.

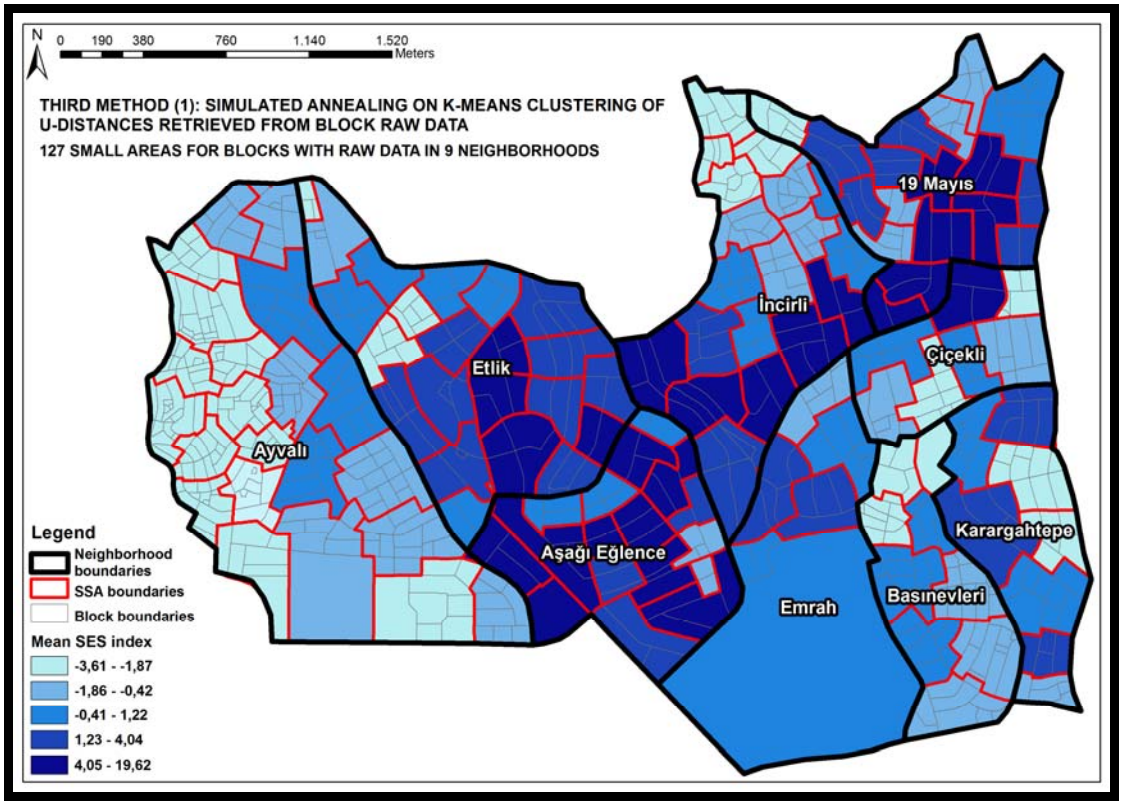


Figure 5.58 Thematic map showing mean SES index in each SSA obtained by simulated annealing on k-means clustering of u-distances retrieved from blocks with raw data.

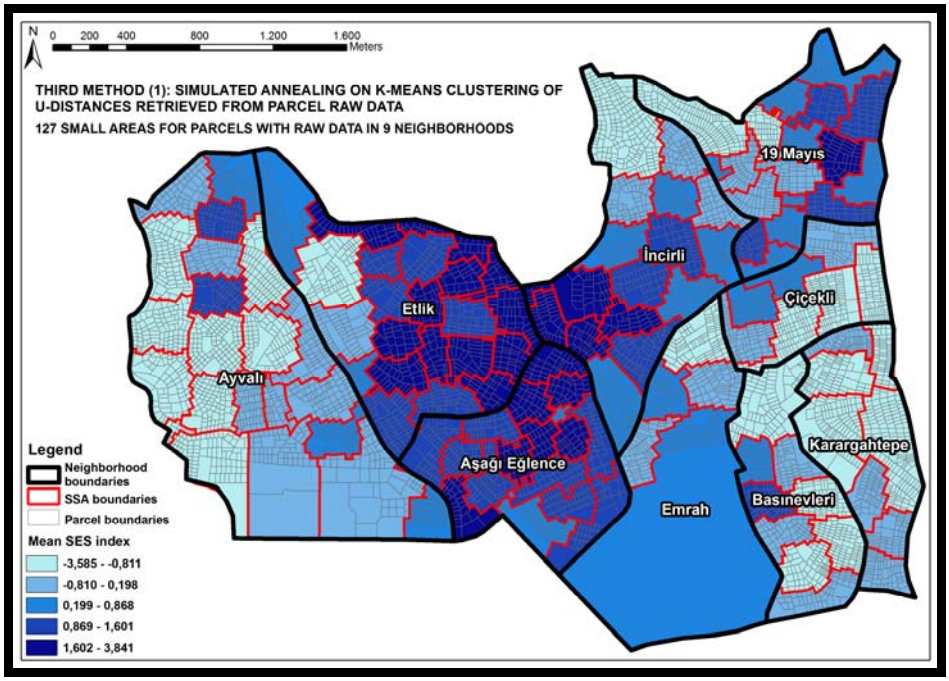


Figure 5.59 Thematic map showing mean SES index in each SSA obtained by simulated annealing on k-means clustering of u-distances retrieved from parcels with raw data.

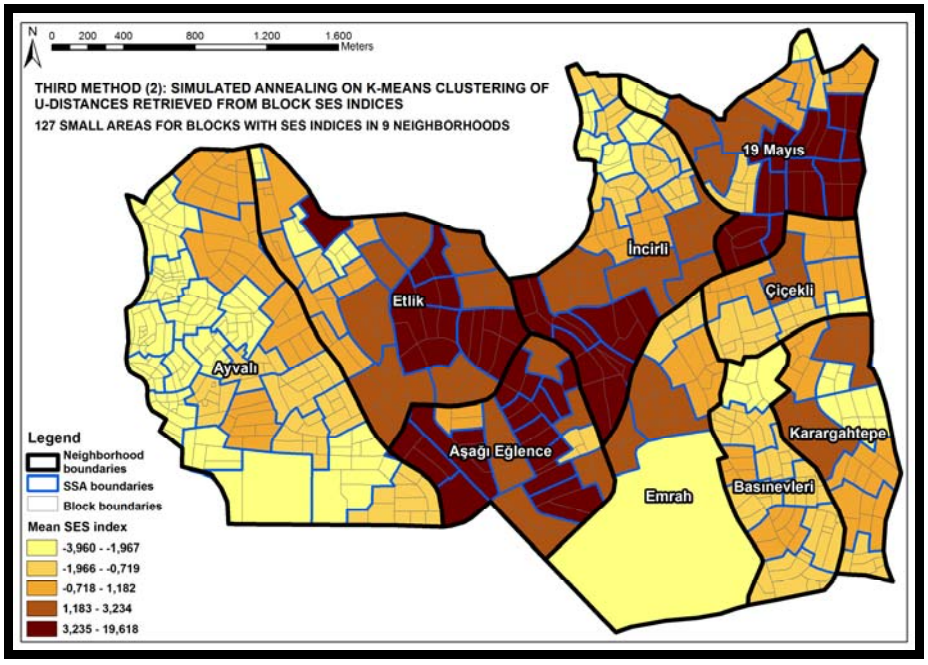


Figure 5.60 Thematic map showing means SES index of each SSA which are obtained simulated annealing on k-means clustering of u-distances retrieved from parcels with SES indices.

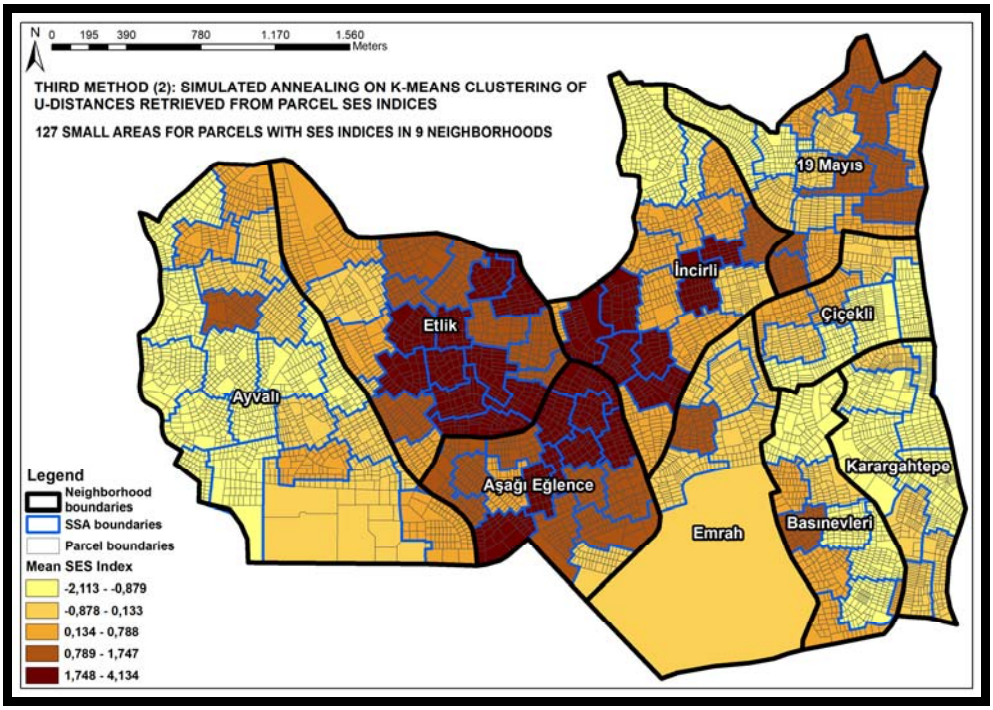


Figure 5.61 Thematic map showing means SES index of each SSA which are obtained simulated annealing on k-means clustering of u-distances retrieved from blocks with SES indices.

The output small areas for parcels and blocks from the third method are similar to those which were found in the first method. For both methods, there are irregular shaped clusters which have elongated forms. Also, for example in south of Ayvalı, the cluster sizes are bigger when compared with the rest which result from the fact that in south of Ayvalı there are huge shopping centers such as “Anteras” and “Metro Market” and in these areas there are no households. So for both methods, simulated annealing tries to find the epui-populous clusters and adds the big blocks of shopping centers to other blocks which have dwelling units nest within. The same situation is also valid for Emrah neighborhood. In south of Emrah, there is a big military health campus (the biggest block), but there are few resident household in this campus. So, some other blocks exist within that particular small area in south of Emrah. Contrary to this situation, the clusters located on west of Ayvalı and north of İncirli, are smaller than rest. The main reason of this situation is that these areas are transformed into regular urban settlements from squatters in near future and high density settlement structure is proposed for these blocks and parcels This means

that the dwelling unit number, in other words household number, proposed for these areas are more than other parts of the study area which causes high population. As a result, these areas have smaller cluster sizes in order to fulfill equal population principle.

5.6. Visual Interpretation of SSAs Obtained by Methods Employed in Thesis

In this part, the SSAs retrieved from three methods are interpreted visually. For this reason, the properties of clusters are commentated for their population and compactness constraints. In Figure 5.62, the SSAs are shown which are obtained from parcels with raw data. It can be seen that the clusters obtained from the first method and third method are less compact than the clusters obtained from the second method. Especially, cluster 20 in first method and cluster 73 have an elongated form which is not preferred for a small area. On the other hand, clusters 15, 18, 19, 57, 40, 92, 96, 98, 100, 101 and 106 from first method; clusters 1, 13, 14 and 123 from the second method and clusters 15, 19, 18, 48, 57, 62, 83, 102, 105 and110 from third method have relatively big sizes than the other clusters. This clusters are located on the north of İncirli and Ondokuz Mayıs, west of Ayvalı, south of Emrah, and east of Karargahtepe As mentioned before this situation most probably results from the fact that the locations that these clusters cover are the least populous areas. As the algorithms try to find the equal distribution of population, the cluster size gets bigger in order to equate the population size of these clusters to other ones which are smaller but denser at population size. The population counts of clusters from each method and for each dataset (parcel-raw, parcel-SES, block-raw, block-SES) can be examined by maps in Appendix D. In Figures 5-63 and 5-65, it is possible to see more or less the same results for SSAs which is motioned above. However, the better compactness of the second method for block datasets is no longer good as in the case of parcel datasets, because the granularity of blocks is lesser that the granularity of parcel datasets. The population deviation of SSAs retrieved from block datasets are more variable resulting from the same reason of granularity difference between block and parcel datasets.

However, for all three methods proposed for small area establishment in this thesis, the visual interpretation is not enough. For comparing these methods effectively in

means of compactness, contiguity, equal population and homogeneity, there must be some quality measures. In the following section, these measures are discussed.

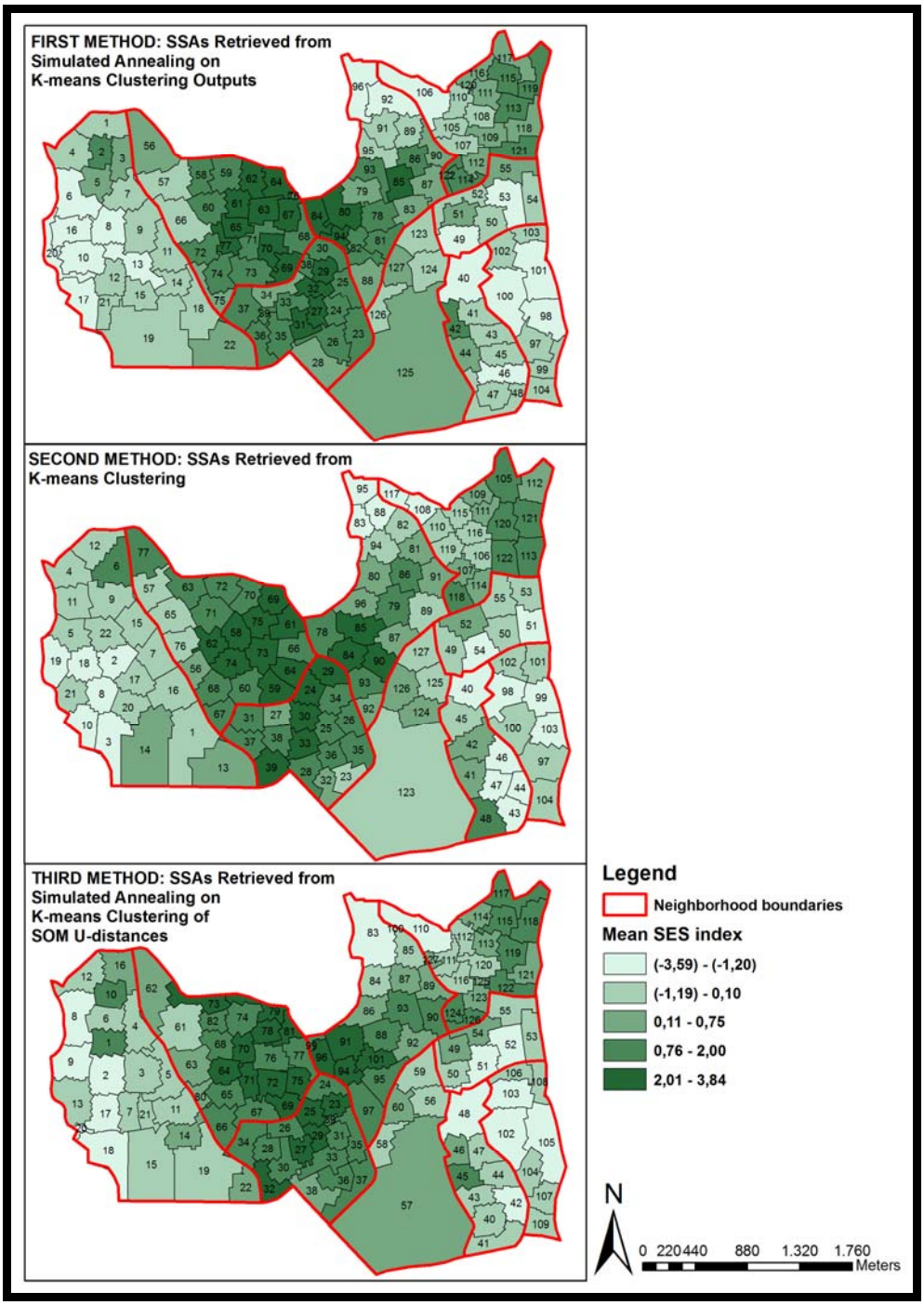


Figure 5.62 SSAs for parcels with raw data retrieved by methods employed in thesis.

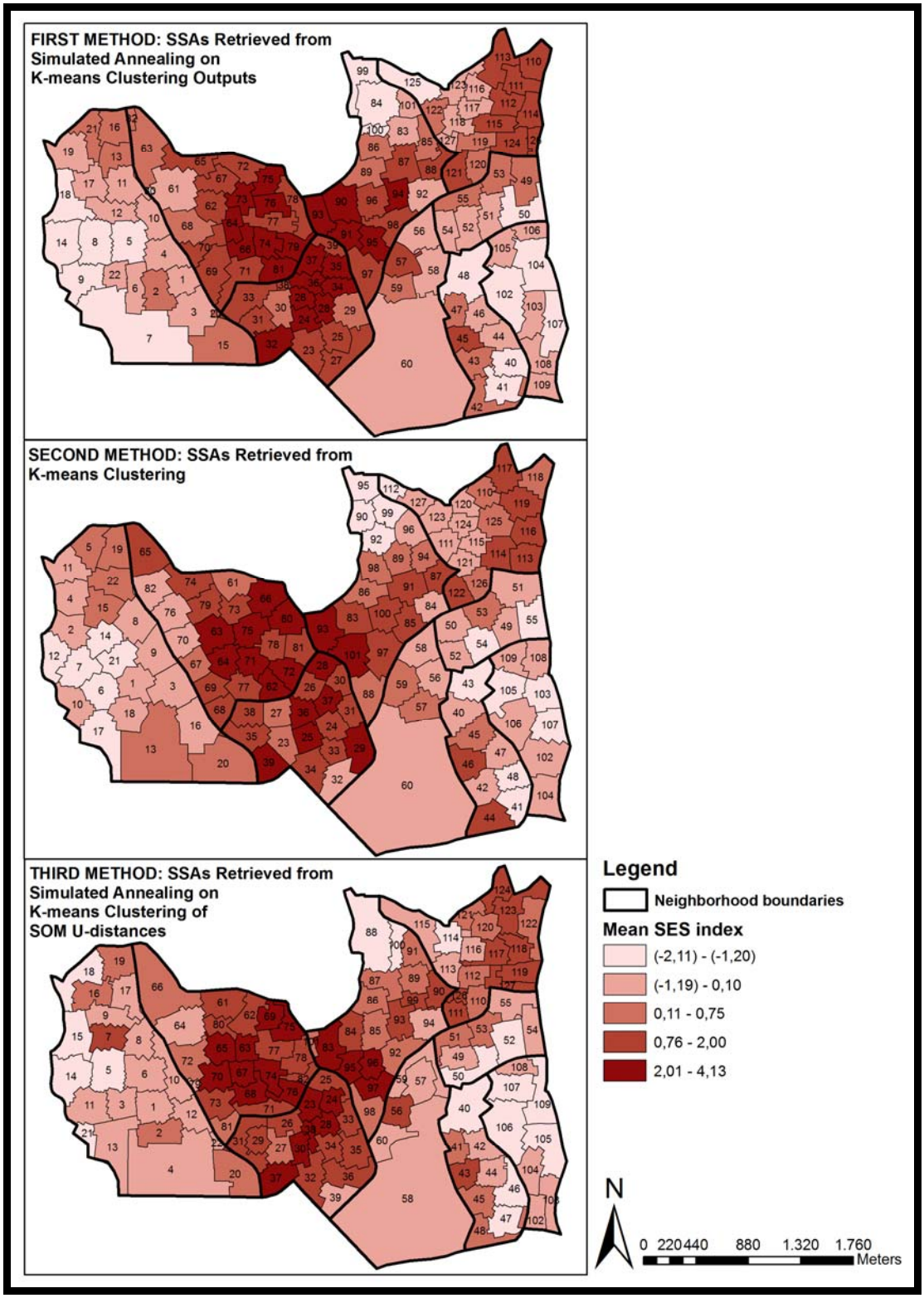


Figure 5.63 SSAs for parcels with SES indices retrieved by methods employed in thesis.

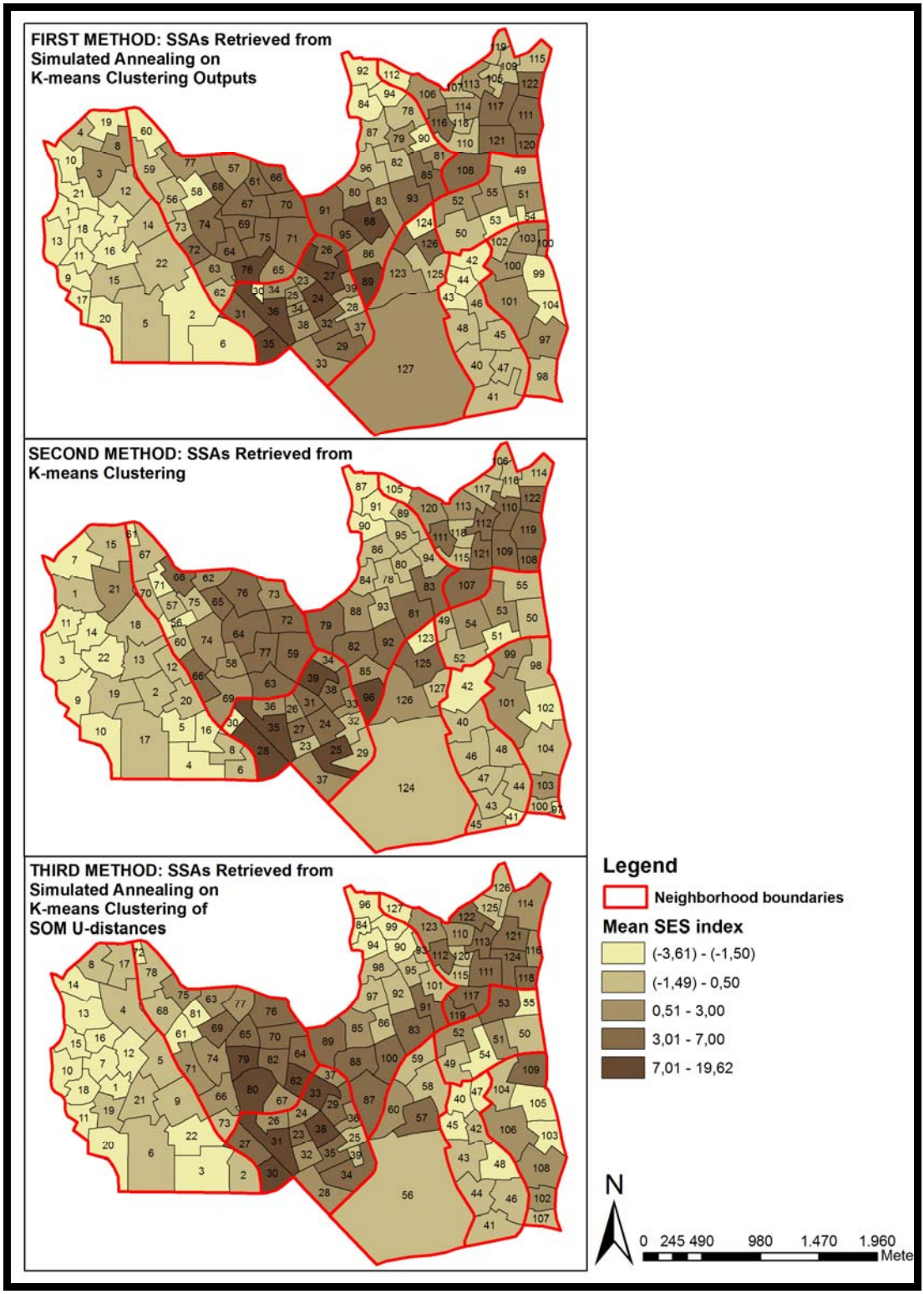


Figure 5.64 SSAs for blocks with raw data retrieved by methods employed in thesis.

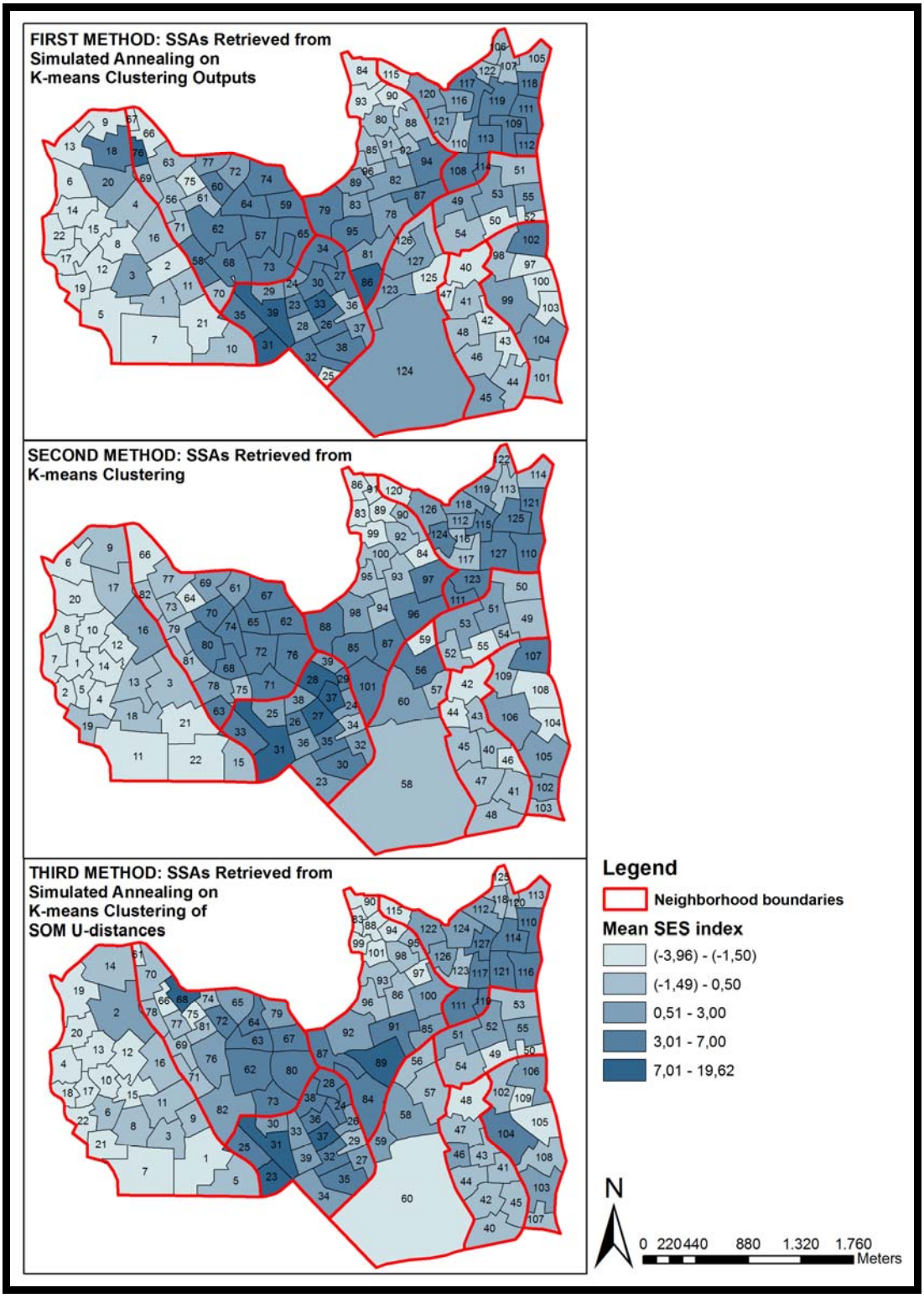


Figure 5.65 SSAs for blocks with raw data retrieved by methods employed in thesis.

5.7. Quality Assessments for Methods Employed in Thesis

As mentioned many times in this thesis, a good SSA layer should contain small areas that fulfill conditions of being compact by shape, contiguous to each other, socially homogenous and having equal population. This is a multi-criteria optimization problem that no best solution exists. But, finding an optimum, in other words, a near-best solution is possible (Azimi & Delavar, 2007).

In this thesis, three different clustering and districting algorithms are applied for raw data and SES indices separately evaluated for parcels and blocks. As there exist a number of methods for clustering, a comparative study to select the best one according to validity for districting problem must be assessed. For this reason, quality assessment is done in terms of the small area identification constraints. Basically, four quality measures and an overall quality measure are calculated for the small areas composing of parcels and blocks from three different methods by help of BARD.

Firstly, compactness score is calculated by BARD based on the ratio of the sides of the bounding rectangle for each particular small area obtained from parcels and blocks. This is accomplished by a function called “calcLWCompactScore()”. Then mean of these scores are calculated for an SSA layer. This score is accepted as the overall compactness of that layer. This calculation is done for every SSA layer derived from three different methods for parcels and blocks separately.

Secondly, population score is calculated by BARD using “calcPopScore()” function which returns the score of a small area based on the deviation from population equality of the districts.

Thirdly, contiguity score is calculated by BARD using “calcContiguityScore()” function which returns a score based on the number of separate contiguous regions in the district. The ideal district comprises a single contiguous region.

BARD also calculates homogeneity score based on the standard deviation of all variables from their mean belonging to a particular small area composing of parcels and blocks. Then, these standard deviations are used to calculate the homogeneity score for a small area.

For calculating an overall quality measure, a combined scoring function is used which was also explained in section 5.3. The results of these calculations are shown in Table 5.14 and Table 5.15 by marking the first optimum result with darker shade and the second by a lighter shade.

Table 5.14 Quality assessment results for SSA layers created by parcels.

	First Method: Simulated Annealing on K-means clustering	**Second Method:** K-means Clustering	**Third Method:** Simulated Annealing on K-means clustering for SOM u-distances
	SSA PLANS FROM PARCELS WITH RAW DATA		
Compactness	0,2256	**0,1432**	**0,2159**
Contiguity	0,924	0,924	**0,912**
Equal Population	**0,0207**	0,0412	**0,0239**
Homogeneity	**2,257**	2,262	**2,253**
Overall Score	**275,62**	278,75	**273,32**
	SSA PLANS FROM PARCELS WITH SES INDICES		
Compactness	**0,2242**	**0,1258**	0,2321
Contiguity	**0,926**	0,927	**0,925**
Equal Population	**0,0223**	0,0392	**0,0223**
Homogeneity	**2,26**	2,28	**2,25**
Overall Score	**274,76**	275,76	**272,77**

Table 5.15 Quality assessment results for SSA layers created by blocks.

	First Method: Simulated Annealing on K-means clustering	**Second Method:** K-means Clustering	**Third Method:** Simulated Annealing on K-means clustering for SOM u-distances
	SSA PLANS FROM BLOCKS WITH RAW DATA		
Compactness	**0,1607**	**0,17**	0,1747
Contiguity	**0,599**	0,609	**0,591**
Equal Population	**0,0492**	0,0527	**0,0487**
Homogeneity	**2,96**	**2,96**	**2,91**
Overall Score	**229,74**	232,62	**227,52**
	SSA PLANS FROM BLOCKS WITH SES INDICES		
Compactness	**0,1601**	0,1721	**0,1685**
Contiguity	0,607	**0,594**	**0,593**
Equal Population	**0,0532**	0,0534	**0,0528**
Homogeneity	**2,84**	3,02	**2,84**
Overall Score	231,24	233,95	**230,47**

In Table 5.14, if we consider overall quality for parcels with raw data, third method gives the most optimum result. In addition, third method also, gives the best score for homogeneity and contiguity constraints for this particular dataset. On the other hand, as expected, k-means clustering provides most compact partitioning of small areas. However, for equal population constraint, first method gives the optimum result. For parcel dataset with SES indices, third method again gives the most optimum result when the scores are evaluated for homogeneity, equal population, contiguity and overall quality. However, for compactness constraint, second method gives the best result as expected, because second method uses k-means clustering. For equal population constraint, first method gives the same result as the third. As a result, it can be said that either for both raw data or SES index attribute inputs used in parcel dataset, simulated annealing on k-means clustering of SOM u-distances gives optimum results.

In Table 5.15, similar to the situation with parcel datasets, third method gives the optimum result for blocks with raw data if the overall quality is taken into account. However, it is very difficult to make a compromise between the first method and the third method if the other constraints are considered. In instance, for blocks with raw data, the third method has a considerable advantage over the rest when the value of homogeneity constraint is evaluated. On the other hand, the optimum value for compactness constraint is obtained by the first method. For block dataset with SES indices, third method again gives the most optimum result. However, for compactness constraint first method gives the best value. As a conclusion, similar to the results obtained for parcel dataset (Table 5.14), simulated annealing on k-means clustering of SOM u-distances gives the best results.

At this point, it should be mentioned that he quality assessment results are obtained from one-time iterations for all three methods in all different datasets. However, in literature, it is preferred to make iterations not less than 100 times (Martin, 1997). This process can be approximated to Monte Carlo technique. The optimum selection is then made after implementing a quality assessment for all the results retrieved from these iterations. For this reason, in this thesis, it is not guaranteed to make a clear assumption about the best method when the first and the third method are considered in which both uses simulated annealing. However, for the particular case study area, it is proved that simulated annealing on k-means clustering of SOM unified distances is an applicable method for establishing SSAs when the results are taken into account. In addition, it can be indicated that SOM u-distance (similarity values) information is a valuable input for a districting algorithm based on simulated annealing.

CHAPTER 6

CONSCLUSIONS AND RECOMMENDATIONS

In this thesis, it is aimed to establish a small statistical area (SSA) layer with a case study in 9 neighborhoods of Keçiören, Ankara by using different spatial clustering and districting methodologies. This study is undertaken fulfilled with the understanding that in SSA construction policies and practices there is a huge gap in Turkey. Such a study can be helpful to understand how to utilize different spatial clustering methodologies and small area identification rules when constructing small areas which happens to be the core components that make up a complete census geography from bottom-to-top. Additionally, such an analysis may be helpful for the policy-makers when it is decided to establish a census geography for Turkey. This study analyzes different solutions that would contribute in distinct aspects of spatial clustering in order to deal with the small area identification problem and therefore, the conclusion part deals with each of these aspects respectively. In this part, these aspects are evaluated according to the problems, solutions, difficulties and recommendations which are observed throughout the thesis in order to enlighten further studies in this area.

In Chapter 2, a comprehensive literature survey is given about the concept of census geography and small area identification rules which is further researched with national applications from UK, US, South Africa and with an international application called Tandem utilized by European countries. Some research is also done about the current status of Turkey in terms of relationship between geography and statistics. From this research it can be inferred that there is also an already continuing process in Turkey to upgrade the current regional statistical system. Some concrete steps have also been taken such as creating the regional scale NUTS 1-2-3 statistical units and preparing important reports and draft laws necessary for further actions through the adaptation process to the European Union.

However, NUTS classification is composing of large statistical areas and the lowest level in this classification is corresponding to cities, therefore these studies are not enough to build up a basis for census geography creation in Turkey. When it is searched further, only Kırlangıçoğlu's (2005) study is found about creation of a census geography and establishment of small statistical areas in Turkey. However, his study lacks resolution because of the unavailability of household information from TURKSTAT, because TURKSTAT disseminates census data at neighborhood level at most. Nonetheless, his study has a considerable place among the similar others because it takes the first concrete step for defined further actions of Turkey about its statistical system in terms of census geography and small area identification. Despite this pessimistic picture, this thesis increases the resolution by using the current census information from ABPRS which has a spatial dimension contrary to former censuses applied in Turkey, since new census system includes address information related to a households which allows one to geocode this information for dwelling units. This thesis brings a new approach to census geography concept in Turkey and emphasizes the importance of local/small area knowledge by utilizing this new spatial dimension of current census methodology. By small areas, it is foreseen that, it will be possible to produce and disseminate high resolution information that will meet the demands from society.

In addition, by using small area information, policies may be monitored and assessed more accurately; because local and regional policies, and also the budgets, of all data users need to be allocated accurately. Therefore it is very important to target the resources effectively and efficiently to the related areas. In addition to the policy makers, local inhabitants also want to know more about the area where they live, and how it compares with other areas. Current dissemination policy of TURKSTAT about censuses and survey statistics do not contain detailed information about local areas and do not give information about all the descriptive indicators.

Moreover, the kinds of current statistical geographic units are insufficient for serving different people from different organizations and professions, especially working in local scale; because the new geographic units will include place-specific information.

This will make good contributions to many projects dealing with the local. A new hierarchical structure of census geography built upon small areas by varying sizes of census geography levels would make it flexible and more easily usable by many people.

As a result of the information based on small areas, the pinpointing of events will be easier and interactions among factors like unemployment, health, safety and education will be better understood at the local area level. For example, police forces will serve the community better and increase safety level using better geographic information to use in their geographic crime analyses. In addition, allocation of many public services may be targeted more precisely and their operational benefits will be increased, too. For instance; locations of clinics, schools, police stations, fire stations etc. will be determined after examining the local needs.

A census geography built upon large scale small areas may also be used for better integration and use of existing nation-wide information held by private and public organizations. If all of separate data from different sources are combined on the same geographical levels and the necessary data are collected according to the new standards, nation-wide overlapping information will be eliminated. By a new common database, all the information may be brought together in a database management system, and for instance, it may be possible to relate different kinds of data from different sources. For example, a small statistical area may be identified in the new census geography as having high population density, overcrowded dwelling units also having children with low educational levels (data from Turkey Ministry of Education), poor health conditions (data from The Ministry of Health), and high unemployment ratio (data from State Institute of Statistics).

The proposed census geography also encourages interventions at local areas and based on small statistical areas. Individually these interventions may be looked small, but when they come together, the overall impact will be very large and benefits will be higher than costs. The availability of the new census geography will increase the efficiency of important private and public investments either by having

the same cost and increasing the benefits or having the same benefits and decreasing the cost.

The small area identification rules can be accessed from many sources and these rules introduce an optimization problem where more than one constraint should be fulfilled. Main constraints are compactness, equal population, homogeneity and contiguity. These constraints necessitate to group similar households on a spatial basis. For this reason, in Chapter 3, spatial clustering algorithms are examined in order to utilize some of them for optimizing the small area identification problem. However, it is seen that, despite the small area identification rules are very clear, tools and methods to construct them are very rare. Therefore, the selection of suitable clustering algorithms is done according to their abilities to fulfill the constraints at the most possible extent for the particular data in hand. Therefore, Simulated annealing on k-means clustering (first method); only k-means (second method) and simulated annealing on k-means clustering of SOM unified distances (third method) are selected as suitable by considering their successful implementation in constructing small statistical areas in literature (Martin, 1997), the nature of the algorithm that work behind, automation capabilities and handling spatial proximities. However, there are also other clustering or districting methods such as genetic algorithms, tabu search, multi-criteria decision making (MCDM) methods, or Analytic Hierarchy Processes (AHP) (Kırlangıçoğlu, 2005) which can be used to construct small areas. This thesis utilizes a small portion of these methodologies. Nevertheless, it might be useful to apply the same procedure with all of the suitable clustering or districting methods to see which one gives the best results. While implementing all these clustering algorithms SOM toolbox and BARD software have a considerable contribution, otherwise advanced computer programming is needed to create software which automatically creates small areas according to small area identification rules and also takes geography into account.

Especially, BARD had a major contribution by its readily prepared functions to create small areas. However, the major deficiency of the package is that there is no function to fulfill the appropriate boundaries principle to construct small areas. SSA boundaries generally follow permanent, visible features, such as streets, roads,

highways, rivers or statutory boundaries. BARD package does not allow one to import a layer that includes such a boundary. Therefore, for each neighborhood, parcels and blocks are exported as a separate shapefile and iterations are made upon these separate neighborhood shapefiles composing of parcels and blocks. Even if the neighborhood boundaries are respected by this process, the streets or roads could not be evaluated as natural boundaries that would surround a small area created by parcels. However, parcels are very important pieces of information which are commonly used for constructing the smallest level in a census geography composing of generally 125 households. On the other hand, the blocks in the study area composes of 300 households on the average and are created considering the streets and roads as boundaries. For these reasons, the small areas are created by approximately 500 households. Therefore, the small areas created in this thesis are not the smallest level for a census geography. Nevertheless, the methodologies work efficiently for blocks if one considers them as a starting point.

For future studies, it is aimed to use a more professional tool which is in the development phase by Prof. Dr. David Martin called AZtool formerly known as AZM (Martin, 1997). AZM tool is a freeware tool and is used by National statistical office of UK for establishing small statistical areas. However, as mentioned in Chapter 2, this tool works with the old ArcINFO command prompts to create the necessary coverage files including topology and especially boundary information whose file structure is changed by new versions of ArcINFO. This tool also uses simulated annealing for creating small areas according to small area identification rules including appropriate boundary principle. Since, it is seen that simulated annealing is a good method for fulfilling the optimization constraints for establishing small areas; AZtool would be a great contribution for the future studies of small area creation in which the small areas are created by taking all the necessary rules into account.

In Chapter 4, data preparation stages are explained, because data is the most important part for establishing small areas. Esoecially statistical data related to granular parts parts of a society forms the bottleneck of a study related to construction of small areas. Because, one of the constraints for realizing this aim is

to define small areas as homogeneous as possible areas according to demographic and socio-economic characteristics of the inhabitants, and then to draw the boundaries of the new census geography units, each of which have specific population thresholds, from this point of view. However, in this thesis, only 3 statistical indicators were available from ABPRS study; which are income level, income source and education level. However, it is necessary to increase the number of these indicators to have more accurate results. The additional indicators of 'quality of life' issues should be considered, including housing conditions, history, culture, ethnicity, geography, health, physical environment, crime, noise pollution, geology, air quality, aesthetics, urban pattern etc.

For instance, the major point of the methodology applied in the thesis is firstly to Here, homogeneous areas have been defined based on 10 indicators; which are Literacy Ratio, University and Primary School Graduate Ratios, Proximity to Police Stations-Hospitals-Fire Stations, Average Household Size, Population Density, Unemployment Ratio, and Young Age Employment Ratio. These indicators and data about them have been relatively sufficient to realize the main aims of the thesis, and to create a new census geography.

In addition, this study takes an urbanized area as a case study into account. However, rural areas of Turkey have a number of distinct characteristics that make it different from other countries. Turkey has a highly dispersed and differentiated rural settlement pattern with different characteristics and very low population numbers, which may not be adequate to satisfy the determined minimum population thresholds. Therefore, new rules and methodologies have to be created and tested for these areas. Not only the rural areas but also the seasonal settlements like secondary housing settlements, tourist areas etc. have to be evaluated again in another research project because of their changeable population characteristics.

As seen in Chapter 4, all the necessary cadastral basemaps are created manually in this thesis as parcels and blocks by relating them with necessary raw socio-economic and physical data from ABPRS. However, this was a challenging task even if the study area was including only 9 neighborhoods. Therefore these up-to

date basemaps are very important for creating small area creation. These digital maps should be produced by the responsible governmental offices.

Another important subject is the construction of small areas in squatter settlements. They are defined as temporary statistical units in this study and temporary boundaries are drawn for them by using 1/1000 development plans obtained from municipality. Therefore it is proposed to redraw their boundaries in the direction of their development, and finally, converting some of them, which are completely rebuilt, to permanent statistical units. Nevertheless, the practical application may not be easy. The boundaries of squatter settlements may not be always drawn with discrete lines, and some areas may contain one within the other and mixed urban patterns. Therefore, while determining the temporary areas and redrawing the boundaries, it is very important to precisely separate the squatter settlements from the other urban areas.

Turkey suffers frequent administrative boundary changes. This situation is alos valid for UK which is mentioned in Chapter 2. These changes, of course, will also cause changes on the drawn boundaries of small areas through years and seem to be opposite to the stability principle of statistical units. The number of these changes must be hold at minimum to protect the reliability of time-series data. Therefore, the areas which have important potential to increase or decrease their population numbers, must be previously defined as temporary statistical units; and they must be converted to the permanent ones when they are mostly complete their urbanization process. Another solution to these problems is proposed by UK Office for National Statistics which is areal weighing.

In conclusion, the methodology for the creation of SSAs proposed in this study is only a small (but important) contributory step for the establishment of a complete census geography in Turkey, because Turkey needs it in order to reach to a highly developed nations' level. If the rules and methodologies are carefully examined and deficiencies are completed through utilizing and comparing more complex algorithms, Turkey may have a much better spatial census and statistical infrastructure than today for the future censuses.

REFERENCES

Aronoff, S., (1991), Geographic Information Systems: A Management Perspective, WDL Publications, Ottawa, Canada.

Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P., (1998), Automatic Subspace Clustering of High-dimensional Data for Data Mining Applications, Presented at the International Conference of Management of Data (SIGMOD'98), pp.94-105, Seattle, WA. Retrieved from: http://www.cs.cornell.edu/johannes/papers/1998/sigmod1998-clique.pdf (Last accessed on 21.06.2009).

Aldenderfer, M.S. & Blashfield, R.K., (1989), Cluster Analysis, Quantitative Applications in the Social Sciences, Sage University Publications, USA.

Altman, M., McDonald, K., McDonald, M. P., (2005), From Crayons to Computers: The Evolution of Computer Use in Redistricting, Social Science Computer Review, Vol.23, pp. 334- 346.

Altman, M., McDonald, M. P., (2009), BARD: Better Automated ReDistricting, Journal of Statistical Software (Forthcoming). Retrieved from: http://www.stats.bris.ac.uk/R/bin/windows/contrib/r-release/BARD_1.05.zip (Last accessed on 09.11.2009).

Ankerst, M., Breunig, M., Kriegel H.P., & Sander, J., (1999), OPTICS: Ordering Points to Identify the Clustering Structure, Presented at the International Conference of Management of Data, Philadelphia, PA, pp. 49–60. Retrieved from: http://www.dbs.ifi.lmu.de/Publikationen/Papers/OPTICS.pdf (Last accessed on 22.06.2009).

Anselin, L., (1990), What is special about spatial data? Alternative perspectives on spatial data analysis, in Spatial Statistics, Past, Present and Future, D.A. Griffith Eds., Institute of Mathematical Geography: Ann Arbor, p. 63-77. Retrieved from: http://www.odum.unc.edu/odum/content/pdf/anselin%201989.pdf (Last accessed on 30.06.2009).

Anselin, L., (2000), GIS, Spatial Econometrics and Social Science Research: The Link between GIS and Spatial Analysis, Journal of Geographical Systems, Vol.2, pp.11-15.

Anselin, N., (1992), Spatial Data Analysis with GIS: An Introduction to Application in the Social Sciences, NCGIA Technical Report Series, Retrieved from: http://www.ncgia.ucsb.edu/Publications/Tech_Reports/92/92-10.pdf (Last accessed on 03.07.2009).

Arslan, O., (2008), Su Kalitesi Verilerinin CBS ile Çok Değişkenli İstatistik Analizi (Porsuk Çayı Örneği), HKM Jeodezi, Jeoinformasyon ve Arazi Yönetimi Dergisi, Vol.99. Retrieved from: http://www.hkmo.org.tr/resimler/ekler/b89a2e980724cb8_ek.pdf (Last accessed on 19.06.2009).

Aydınoglu, A.C., Yomralıoglu, T., (2009), Developing Geospatial Data Specification Following INSPIRE with Turkey Case, Presented at the Joint International Workshop of ISPRS WG IV/1, WG VIII/1 and WG IV/3 - Geospatial Data Cyber Infrastructure and Real-time Services with special emphasis on Disaster Management, Nov 25-27, Hyderabad, India. Retrieved from: http://www2.itu.edu.tr/~tahsin/yayinlar.htm (Last accessed on 02.12.2009).

Azimi, A. & Delavar, M.R., (2007), Quality Assessment in Spatial Clustering of Data Mining, Presented at the 5th International Symposium of Spatial Data Quality, Netherlands. Retrieved From: http://www.itc.nl/issdq2007/proceedings/Session%202%20Spatial%20Statistics/paper%20DELAVAR.pdf (Last accessed on 20.05.2009).

Bacao, F., Lobo, V., & Painho M., (2005-a), On the Particular Characteristics of Spatial Data and Its Similarities to Secondary Data Used in Data. Retrieved from: http://www.isegi.unl.pt/ensino/docentes/fbacao/gisplanet2005.pdf (Last accessed on 29.06.2009).

Bacao, F., Lobo, V. & Painho, M., (2005-b), Self-Organizing Maps as Substitutes for K-means Clustering, Lecture Notes in Computer Science, V. 3516, pp.476-483, Springer-Berlin, Heidelberg.

Bacao, F, Painho, M., Pedrycz, W. & Vasilakos, A., (2005-c), Exploring Spatial Data Through Computational Intelligence: A Joint Perspective, Soft Computing - A Fusion of Foundations, Methodologies and Applications, Vol. 9, pp. 326-331, Springer-Verlag, Berlin, 2005.

Bacao, F., Lobo, V. & Painho, M., (2005-d), The Self-Organizing Map, The Geo-SOM, and Relevant Variants to Geosciences, Computers and Geosciences, Vol.31, pp: 155-163.

Bacao, F., Lobo, V. & Painho, M., (2008), Applications of Different Self-Organizing Map Variants to Geographical Information Science Problems, Self-Organizing Maps: Applications in Geographic Information Science, Eds. Agrawal, P. & Skupin, A., John Wiley & Sons, England.

Backer, L.H.., Tammilehto-Luode, M. & Guiblin, P., (2002), A (feasibility) study towards a common geographical base for statistics across the European Union, Tandem_GIS- I. Retrieved from: http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-AN-02-001/EN/KS-AN-02-001-EN.PDF (Last accessed on 06.06.2009)

Bailey, T.C. & Gatrell, A.C., (1995), Interactive Spatial Data Analysis, Longman, Harlow, UK.

Basara, H.G. & Yuan, M., (2008), Community Health Assessment Using Self-Organizing Maps and Geographic Information Systems, International Journal of Health Geographics, Vol.7, pp.1-8.

Birimcombe, A.J., (2003), A Variable Resolution Approach to Cluster Discovery in Spatial Data Mining, Computational Science and Its Applications (ICCSA'03), Lecture Notes in Computer Science, Springer-Berlin, Heidelberg.

Bock, H.H., (2007), Clustering Methods: A History of K-means Algorithms, Selected Contributions in Data Analysis and Classification, pp.161-172, Springer-Berlin, Heidelberg.

Burrough, P.A., (1986), Principles of Geographic Information Systems for Land Resource Assessment, Clarendon Press, Oxford, GB.

Buttenfield, B., Gahegan., M, Miller, H. & Yuan, M., (2000), Geospatial Data Mining and Knowledge Discovery, An UCGIS White Paper on Emergent Research Themes. Rterieved from: http://www.ucgis.org/priorities/research/research_white/2000%20Papers/emerging gkd.pdf (Last accessed on 21.06.2009).

Chawla, S., Shekhar S., Wu W. & Ozesmi U., (2001), Modeling Spatial Dependencies for Mining Geospatial Data: An Introduction, Geographic Data Mining and Knowledge Discovery, pp.139-171, Taylor and Francis, London.

Cheung, Y.M., (2003), K*-means: A New Generalized K-means Clustering Algorithm, Pattern Recognition Letters, Vol.24, pp.2883-2893.

Chrisman, N., (1996), Exploring Geographic Information Systems, John Wiley & Sons, Inc., Canada.

Coleman, A.M., (2008), An Adaptive Landscape Classification Procedure Using Geoinformatics and Artificial Neural Networks, Dissertation submitted to Faculty of Earth and Life Sciences, Vrije University, Amsterdam, The Netherlands. Retrieved from: http://www.unigis.nl/downloads/msc/Andre%20Coleman.pdf
(Last accessed on 10.09.2009).

Conley, J., Gahegan, M. & Macgill, J., (2005), A Genetic Approach to Detecting Clusters in Point Data Sets, Geographical Analysis, Vol.37, pp.286-314.

Coombes, M., (1995), Census User's Handbook, edited by Stan Openshaw, Bell and Brain, Glasgow.

Dai, D. & Oyana, T.J., (2009), Automatic Cluster Identification for Environmental Applications Using the Self-Organizing Maps and a New Genetic Algorithm, Geocarto International, pp: 1-17.

Demir, Ö. and Toprak, A.Ö., (2004), Turkish Statistical System: Current CurrentSituation and New Challenges, Presented at the United Nations Economic and Social Commission for Asia and Pasific Subcommittee on Statistics, Bangkok, Thailand, 18-20 February 2004. Retrieved from:
http://www.unescap.org/stat/sos1/sos1_turkey.pdf (Last accessed on 25.11.2009).

Demirci, M., & Taştı, E., (2009), Changing the System From Traditional Census To Register Base Census In Turkey, United Nations, Census Knowledge Base. Retrieved From:
http://unstats.un.org/unsd/censuskb/attachments/2009TUR_ISI_GUIDc1d61c956ec44d05b91a69aae6057ee9.pdf (Last accessed on 25.11.2009).

Ding, Y. & Fortheringham, A.Z., (1991), The Integration of Spatial Analysis and GIS: The Development of the STATCATS Module for ARC/INFO, National Center of Geographical Information and Analysis. Retrieved From:
http://www.ncgia.ucsb.edu/Publications/Tech_Reports/91/91-5.pdf
(Last accessed on 12.06.2009).

Edmonston, B., (1999), Challenges For The U.S. Census In The Information Age. Retrieved from:
http://usinfo.state.gov/journals/itsv/0699/ijse/census.htm
(Last accessed on 16.09.2009).

Essa J. A. & Nieuwoudt, (2003), Socio-Economic Dimensions of Small-scale Agriculture: A Principal Component Analysis, Development Southern Africa, Vol.20, pp: 67-73.

Ester, M., Kriegel H., Sander, J. & Xu, X., (1997), Clustering and Knowledge Discovery in Spatial Databases, Vistas in Astronomy, Vol. 41, pp.397-403, Great Britain.

Ester, M., Kriegel, H., Sanderi J. & Xu, X., (1996), A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Presented at the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), pp.226-231, Portland, Oregon.

Estivill-Castro, V. & Houle M.E., (2000), Robust Distance-Based Clustering with Applications to Spatial Data Mining, Revised Submission to Special Issue of Algorithmica, Algorithms for Geographical Information.

Fayyad U., Piatetsky-Shapiro G. & Smyth P., (1996), From Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol.17, pp. 37 – 54.
Retrieved from: http://infovis.uni-konstanz.de/papers/1998/kdd98.pdf
(Last accessed on 22.06.2009).

Filmer, D. & Pritchett, L.H., (2001), Estimating Wealth Effects Without Expenditure Data - or Tears: An Application to Educational Enrollments in States or India, World Bank Publications, Demography, Vol.38, pp:115-132.

Fincke, T., Lobo, V. & Bacao, F., (2008), Visualizing Self-Organizing Maps with GIS, Presented at the International Conference of Geoinformatika 2008, Munich, Germany. Retrieved From: http://www.gitage.de/archive/2008/downloads/acceptedPapers/Papers/Fincke,Lobo,Bacao.pdf (Last accessed on 20.06.2009).

Franzini, L., Bolchi, P. & Diappi, L., (2001), Self-Organizing Maps: A Clustering Neural Method for Urban Analysis, Presented at the New Areas in Theoretical and Quantitative Geography Symposium (TheoQuant'01). Retrieved from: http://thema.univ-fcomte.fr/theoq/pdf/2001/franzini.pdf
(Last accessed on 25.09.2009).

Frosztega, M. & Estibals, A., (2004), Neighbourhood Statistics – Developing Better Information for Small Areas in England and Wales, Presented at the One Day Meeting about Small Area Statistics – Potential and Challenges, Belfield. Retrieved from: http://www.tcd.ie/Statistics/smallareastatistics/ONS.ppt
(Last accessed on 15.09.2009).

Gehlke C. & Biehl, E., (1934), Certain Effects of Grouping Upon the Size of Correlation Coefficients in Census Tract Material, Journal of American Statistical Association, Suppl.29, pp.169-170.

Grobbelaar, N., (2005), The Development of Small Area Spatial Layer to Serve as the Most Detailed Geographic Entity for the Dissemination of Census 2001 Data, Presented at the 7th Africa GIS Conference, Tshwane (Pretoria), South Africa. Retrieved from: http://mapserver2.statssa.gov.za/geographywebsite/Docs/AfricaGIS/685_Development%20of%20the%20Small%20Are%20Layer.pdf (Last accessed on 06.10.2009).

Guiblin, P., Tammilehto-Luode, M., & Backer, L. H., (2001-a), In Search of a Common Geographical Base to Compare Statistics Across the EU: The TANDEM GIS Project, Presented at the Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of Statistics and Geography, Estonia. Retrieved from: http://www.unece.org/stats/documents/2001.09.gis.htm (Last accessed on 03.09.2009).

Guiblin, P., Tammilehto-Luode, M., & Backer, L. H., (2001-b), In search of a system of small statistical areas, Presented at the Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of Statistics and Geography, Estonia. Retrieved from: www.unece.org/stats/documents/2001/09/gis/crp.1.e.pdf (Last accessed on 03.09.2009).

Guo, D., (2002), Spatial Cluster Ordering and Encoding for High-Dimensional Geographic Knowledge Discovery, GeoVISTA Center and Department of Geography, Pensylvania State University, USA, 2002. Retrieved from: http://www.cobblestoneconcepts.com/ucgis2summer200 2/guo/guo.html (Last accessed on 18.06.2009).

Guo, D., Gahegan, M., MacEachren, A.M. & Zhou, B., (2005), Multivariate Analysis and Geovisualization with and Integrated Geographic Knowledge Discovery Approach, Cartography and Geographic Information Science, Vol. 32, No.2, 2005, pp. 113-132.

Gwatkin, D.R., Rustein, S. and Johnson, K., (2000), Socio-economic differences in Brazil, World Bank Publications, HNP (Health, Nutrition and Population)/Poverty Thematic Group of World Bank, Washington, DC.

Han, E.H., Karypis, G. and Kumar V., (1999), CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling, Computer, Vol.32, pp.68-75.

Han, J. & Kamber, M., (2000), Data Mining: Concepts and Techniques, Chapter 8: Cluster Analysis Course Slides, Morgan Kauffman Publishers. Retrieved from: http://www.cs.sfu.ca/~han/dmbook (Last accessed on 01.07.2009).

Han, J. & Miller J.H., (2009-a), Geographic Data Mining and Knowledge Discovery: An Overview, Geographic Data Mining and Knowledge Discovery, Chapter 1, pp.1-27, CRC Press, Taylor & Francis Group, Boca Raton, London, NY.

Han, J. & Ng, R.T., (1994), Efficient and Effective Clustering Methods for Spatial Data Mining, Presented in Very Large Databases (VLDB) Conference, pp. 144-155, Santiago, Chile.

Han, J., Lee, J. & Kamber , M., (2009-b), An Overview of Clustering Methods in Geographic Data Analysis, Geographic Data Mining and Knowledge Discovery, Chapter 7, pp.149-189, CRC Press, Taylor and Francis Group, Boca Raton, London, NY.

Hatzichristos, T., (2004), Delineation of Demographic Regions with GIS and Computational Intelligence, Environment and Planning B: Planning and Design, Vol. 34, p.39-49.

Henriques, R.A., (2005), CARTO-SOM: Cartogram Creation Using Self-Organizing Maps, Dissertation submitted to Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa. Retrieved from: http://www.isegi.unl.pt/servicos/documentos/TSIG010.pdf (Last accessed on 04.10.2009).

Hinneburg, A. & Keim, D.A., (1998), An Efficient Approach to Clustering in Large Multimedia Databases with Noise, Presented at the International Conference of Knowledge Discovery and Data Mining (KDD'98), pp. 58-65, NY.

Holt, D., Steel, D. and Tranmer, M., (1996), Areal Homogeneity and The Modifiable Areal Unit Problem, Geographical Systems, Vol.3, pp.181-200.

Izenman, A.J., (2008), Modern Multivariate Statistical Techniques, Chapter12: Cluster Analysis, Springer Science+Business Media, LLC, 2008.

Jacquez, G.M., (2008), Spatial Cluster Analysis, Chapter 22 in "The Handbook of Geographic Information Science", S. Fortheringham and J. Wilson (Eds), Blackwell publishing, pp.395-416.

Jain, A.K. & Mao, J., (1996), Artificial Neural Networks: A tutorial, IEEE Computer, V. 29, pp. 31-44.

Jain, A.K., (2008), Data Clustering: 50 Years Beyond K-means, Machine Learning and Knowledge Discovery in Databases, Lecture Notes from Computer Sciences, Vol: 5211, Springer-Berlin, Heidelberg.

Jain, A.K., Murty, M.N. & Flynn, P.J., (1999), Data Clustering: A Review, ACM Computing Surveys, Vol.31, No.3.

Jiang, B., (2004), Extraction of Spatial Objects from Laser Scanning Data Using a Clustering Technique, Presented at the International Society of Photogrammetry and Remote Sensing (ISPRS'04), İstanbul. Retrieved from: http://www.isprs.org/congresses/istanbul2004/comm3/papers/270.pdf (Last accessed on 10.09.2009).

Jiang, B., (2004), Spatial Clustering for Mining Knowledge in Support of Generalization Processes in GIS, Presented at the Workshop on Generalization and Multiple Representation, Leicester.

Kalaycıoğlu, M, (2006), Poverty Mapping with Geographic Information Systems: A Case Study in Keçiören District, a Master of Science thesis submitted to the department of Geodetic and Geographic Information Technologies in graduate school of Natural and Applied Sciences of the Middle East Technical University, December 2006.

Kaski, S. & Kohonen, T., (1996), Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World, In Refenes, A.N., Abu Mostafa, Y. Moody, J. & Weigend A. (Eds.), Neural Networks in Financial Engineering, pp.498-507.

Kauko, T., (2005), Using the Self-Organizing Map to Identify Regularities Across Country Specific Housing Market Contexts, Environment and Planning B: Planning and Design 2005, Vol. 32, pp: 89-110.

Kauko, T., (2009), Classification of Residential Areas in the Three Largest Dutch Cities Using Multidimensional Data, Urban Studies, Vol.46, pp: 1639-1663.

Khumalo, B., (2009), The Dwelling Frame Project as a Tool of Achieving Socially-Friendly Enumeration Areas' Boundaries for Census 2011, South Africa, Presented at the 57th Session of International Statistical Institute. Retrieved from: http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/1238.pdf (Last accessed on: 11.09.2009).

Kırlangıçoğlu, C., (2005), A New Census Geography for Turkey Using Geographic Information Systems: A Case Study on Çankaya District-Ankara, ,a Master of Science thesis submitted to the department of Geodetic and Geographic Information Technologies in graduate school of Natural and Applied Sciences of the Middle East Technical University.

Kohonen, T., (1996), Self-Organizing Maps, 2nd edition, Springer, Berlin-Heidelberg, pp: 362.

Kohonen, T., (2001), Self-Organizing Maps, 3rd Edition, Springer-Verlag, Berlin.

Kolenikov, S. & Angeles, G., (2004), The Use of Discrete Data in PCA: "Theory, Simulations and Applications to Socio-economic Indices, Technical Report, MEASURE/ Evaluation Project, Carolina Population Center, University of North Carolina, Chapel Hill.

Koperski, K., Adhikary, J. & Han, J., (1996), Spatial Data Mining: Progress and Challenges Survey Paper, Presented at the SIGMOD Workshop on Research Issues on data Mining and Knowledge Discovery (DMKD). Retrieved from: https://eprints.kfupm.edu.sa/66097/1/66097.pdf (Last accessed on 15.06.2009).

Koua, E.L. & Kraak, M.J., (2004), Geovisualization to support the exploration of large health and demographic survey data, International Journal of Health Demographics, Vol.3, pp: 1-13.

Kumar, V., (2000), An Extensive Survey of Clustering Methods for Data Mining. Retrieved from: http://www-users.cs.umn.edu/~han/dmclass/
(Last accessed on 25.06.2009).

Kumar, V., Tan, P.N. & Steinbach, M., (2006), Introduction to Data Mining, Chapter 8 - Cluster Analysis: Basic Concepts and Algorithms. Retrieved from: http://www-users.cs.umn.edu/~kumar/dmbook/index.php
(Last accessed on 15.06.2009).

Lacayo, M. & Skupin, A., (2007), A GIS Based Visualization Module for Self-Organizing Maps, Presented at the 23rd International Cartographic Conference, Moscow, Russia. Retrieved From: http://cartography.tuwien.ac.at/ica/documents/ICC_proceedings/ICC2007/html/Proceedings.htm (Last accessed on 22.08.2009).

Laldaparsad, S., (2007), Census mapping and the use of geo-spatial technologies: A case of South Africa, Presented at the United Nations Expert Group Meeting on Contemporary Practices in Census Mapping and Use of Geographical Information Systems, New York. Retrieved from: http://unstats.un.org/unsd/Demographic/meetings/egm/CensusMappingEGM07/docs/ESA-STAT-AC.115-2.pdf (Last accessed on 11.09.2009).

Lehohla, P., (2005), Statistics Needs Geography; Geography Needs Statistics, Presented at the 7th Africa GIS Conference, Tshwane (Pretoria), South Africa. Retrieved from:
http://mapserver2.statssa.gov.za/geographywebsite/africaGIS.html
(Last accessed on 06.10.2009).

Levine, N., (2009-a), CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v:3.2), Chapter 6: 'Hot Spot' Analysis 1, Ned Levine and Associates, Houston, TX, and National Institute of Justice, Washington, DC. Retrieved from:
http://www.icpsr.umich.edu/CRIMESTAT/files/CrimeStatChapter.6.pdf
(Last accessed on 30.06.2009).

Levine, N., (2009-b), CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v:3.2), Chapter 7: 'Hot Spot' Analysis 2, Ned Levine and Associates, Houston, TX, and National Institute of Justice, Washington, DC. Retrieved from:
http://www.icpsr.umich.edu/CRIMESTAT/files/CrimeStatChapter.7.pdf
(Last accessed on 30.06.2009).

Li, D. & Wang, S., (2005), Concepts, Principles and Applications of Spatial Data Mining and Knowledge Discovery, Presented at the International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion, Beijing, China.

Li, Y. & Shanmuganathan, S., (2007), Social Area Analysis Using SOM and GIS: a Preliminary Research, Ritsumeikan Center for Asia Pacific Studies (RCAPS) Working Paper. Retrieved From:
http://www.apu.ac.jp/rcaps/modules/webpublication/content/RCAPS.WP07-3.pdf
(Last accessed on 06.09.2007).

Lombaard, M., (2005), Geographical Support for Social Statistics, Statistics South Africa. Retrieved from: http://mapserver2.statssa.gov.za/geographywebsite/Docs/AfricaGIS/678_Geography%20support%20to%20Social%20Stats.pdf (Last accessed on 06.10.2009)

Martin, D., (1997), From Enumeration Districts to Output Areas: Experiments in the Automated Creation of a Census Output Geography, Population Trends, Vol. 88, pp.36-42.

Martin, D., (2000), Towards an Integrated National Socioeconomic GIS: The Geography of the 2001 Census in England and Wales, Presented at the 3rd AGILE Conference on Geographic Information Science, Helsinki. Retrieved from: http://plone.itc.nl/agile_old/Conference/2000-helsinki/66.pdf
(Last accessed on 10.09.2009).

Martin, D., (2002), Geography for the 2001 Census in England and Wales, Population Trends, Vol. 108. Retrieved from: http://www.statistics.gov.uk/geography/downloads/georoadshowpaper.pdf
(Last accessed on 16.09.2009).

Martin, D., (2003), Extending the Automated Zoning Procedure to Reconcile Incompatible Zoning Systems, Journal of Geographical Information Science, Vol.17, pp.181-196.

Mataracı, O., Yomralıoğlu, T. & Çete, M., (2009), AB'de Kadastro Parselinin INSPIRE Direktifleri Kapsamında Değerlendirilmesi ve Türkiye'nin Yeri, TMMOB Harita ve Kadastro Mühendisleri Odası 12. Türkiye Harita Bilimsel ve Teknik Kurultayı.

Matejicek, L., (2006), Modeling of Water pollution in Urban Areas with GIS and Multivariate Statistical Methods. Retrieved from: http://www.iemss.org/iemss2006/papers/s2/25_Matejicek_3.pdf
(Last accessed on 22.06.2009).

McAdams, M. & Demirci A., (2006), The Use of Principle Component Analysis in Data Reduction for GIS Analysis of Water Quality Data, Presented at the 4th GIS Days in Turkey. Retrieved From:
http://dis.fatih.edu.tr/store/docs/mcadams_cbssukalanvlLbIaZg.pdf
(Last accessed on 19.06.2009).

McKenzie, D.J., (2003), Measure Inequality with Asset Indicators, Bureau for Research and Economic Analysis for Development, Center of International Development, Harvard University.

Office for National Statistics, (2005), Neighbourhood Statistics Geography Policy. Retrieved from:
http://www.neighbourhood.statistics.gov.uk/HTMLDocs/images/GeographyPolicy_tcm97-51009.pdf (Last accessed on 10.09.2009).

Office for National Statistics, (2009-a), Beginner's Guide to UK Geography. Retrieved From: http://www.statistics.gov.uk/geography/census_geog.asp
(Last accessed on 10.09.2009).

Office for National Statistics, (2009-b), Geography. Retrieved from:
http://www.statistics.gov.uk/geography/default.asp (Last accessed on 10.09.2009).

Openshaw, S. & Rao, L., (1995), Algorithms for Reengineering 1991 Census Geography, Environment and Planning – A, Vol.27, pp.425-466.

Openshaw, S., & Alvanides, S., (1999), Applying geocomputation to the analysis of spatial distributions, Longley, P.A.., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (Eds), (1999). Geographical information systems, Wiley, New York, pp. 267–282.

Openshaw, S., (1991), Developing appropriate spatial analysis methods for GIS, in Maguire, D., Goodchild, M. F., Rhind, D. (eds) GIS Principles and Applications Volume 1, Longman, London.

Openshaw, S., (1996), Census Users Guide, GeoInformation International, Cambridge, pp 239- 268.

Openshaw, S., (1997), Artificial Intelligence in Geography, John Wiley & Sons, Chichester, Sussex.

Openshaw, S., (2001), Geocomputation, Geocomputation, Edited by Openshaw S. and Abrahart R.J. (2000), Taylor & Francis, London.

Openshaw, S., Blake, M. & Wymer, C., (1995), Using Neurocomputing Methods to Classify Britain's Residential Areas. Retrieved from: http://www.geog.leeds.ac.uk/papers/95-1/ (Last accessed on 25.07.2009).

Openshaw, S., Charlton, M., Wymer, C. & Craft, A., (1987), Geographical Analysis Machine for the Automated Analysis of Point Data Sets, International Journal of Geographical Information Systems, Vol.1, pp.335-358.

Parinet, B., Lhote, A. & Legube, B., (2003), Principal Component Analysis: An Appropriate Tool for Water Quality Evaluation and Management – Application to a Tropical Lake System, Ecological Modeling, Vol. 178, pp.295-311.

Parinet, B., Lhote, A., Legube B., (2004), Principal Component Analysis: An Appropriate Tool for Water Quality Evaluation and Management – Application to a Tropical Lake System, Ecological Modeling, Vol. 178, pp. 193-213.

Pena, J.M., Lozano, J.A. and Larranaga, P., (1999), An Empirical Comparison of Four Initialization Methods for the K-means Algorithm, Pattern Recognition Letters, Vol.20, pp.1027-1040.

Pettorelli, N., Dray, S. and Maillard, D., (2005), Coupling Principal Component Analysis and GIS to Map Deer Habitats, Wildlife Biology, Vol.11, pp. 363-370.

Qi, F. & Zhu, A.X., (2003), Knowledge discovery from soil maps using inductive learning, International Journal of Geographical Information Science, Vol. 17, pp. 771-795, Taylor and Francis.

Ratcliffe, J.H. & McCullagh, M. J., (1999), Hotbeds of crime and the search for spatial accuracy, Geographical Systems, Vol.1, pp.385-398.

SEU, (2000), National Strategy for Neighborhood Renewal, Report of Policy Action Team 18: Better Information, Social Exclusion Unit, London. Retrieved from: http://www.cabinetoffice.gov.uk/media/cabinetoffice/social_exclusion_task_force/assets/publications_1997_to_2006/pat_report_18.pdf (Last accessed on 16.09.2009).

Sheikholeslami, G., Chatterjee, S. & Zhang, A., (1998), WaveCluster: A Multiresolution Clustering Approach for Very Large Databases, Presented at the International Conference of Very Large Databases (VLDB'98), pp.428-439, NY. Retrieved from:
http://www.cs.sfu.ca/CC/459/han/papers/sheikholeslami98.pdf
(Last accessed on 22.06.2009).

Silva, M. A. S., Monteiro, A. M. V. & Medeiros, J.S., (2004), Visualization of Geospatial Data by Component Planes and U-Matrix, Presented at the 6th Brazilian Symosium on GeoInformatics (GEOINFO 2004). Retrieved From:
http://www.geoinfo.info/geoinfo2004/papers/6419.pdf
(Last accessed on 20.08.2009).

Skupin, A. & Hagelman, R., (2005), Visualizing Demographic Trajectories with Self-Organizing Maps, Geoinformatica, Vol. 9, pp: 159-179.

Spielman, S.E. & Thill, J.C., (2008), Social Area Analysis, Data Mining and GIS, Computers, Environment and Urban Systems, Vol.32, pp.110-122.

Statistics South Africa, (2007), Statistics Using the 2001 Census: Approaches to Analyzing Data, Pretoria: Statistics South Africa. Retrieved From: www.agirn.org/documents/1-South_Africa_census_2001_handbook.pdf (Last accessed on 06.10.2009).

Statsoft Electronic Textbook, (2008), Cluster Analysis. Retrieved from: http://www.statsoft.com/textbook/stcluan.html (Last accessed on, 29.06.2009)

Şenyapılı, T. (2004), Baraka'dan Gecekonduya, Ankara'da Kentsel Mekanın Dönüsümü 1923-1960, İletisim Yayınları, İstanbul.

Şenyapılı, T. (2005), Ankara Kenti İkili Yapısında Dönüsümler, in the book "Cumhuriyet'in Ankara'sı" edited by Şenyapılı, T., ODTÜ Gelistirme Vakfı Yayınları, Ankara.,

Tammilehto-Luode, M., (2003), Towards a Common Geographical Base for Statistics across Europe, Presented at the workshop of standing committee of the International Association of Official Statistics (IAOS). Retrieved from: www.busmgt.ulst.ac.uk/scorus/potsdam/078.pdf (Last accessed on 03.09.2009).

Tobler, W.R., (1979), Cellular Geography, Philosophy in Geography, Gale and Olsson Eds., Dordrecht, Reidel.

Tuia, D., Christian, K., De Cunha, A., Kanevski, M., (2009), Detection of Socio-economic Patterns Using Clustering Techniques, Studies of Computational Intelligence, Vol.176, pp.19-36 Springer-Verlag Berlin, Heidelberg.

TURKSTAT, (2002), Country Paper: Republic Of Turkey, Presented at the Economic and Social Commission for Asia and the Pacific (ESCAP) Committee on Statistics - Thirteenth Session, Bangkok, Thailand, November 2002. Retrieved from: http://www.unescap.org/stat/cos13/cos13_turkey.pdf (Last accessed on 25.11.2009).

TURKSTAT, (2008), New Method for 2010 Population and Housing Census of Turkey - Considerations About Data quality and Coverage, Presented at the United Nations Joint Economic Commission/Eurostat Meeting on Population and Housing Censuses, Geneva. Retrieved From: http://unstats.un.org/unsd/censuskb/attachments/2008TUR_ECE-GUID0d090cc0e5c24cb6bb4e6f0a119e4eaf.pdf (Last accessed on 25.11.2009).

Ultsch, A., & Siemon, H. P., (1990), Kohonen's Self-Organizing Feature Maps for Exploratory Data Analysis, Presented at the International Neural Network Conference (INNC'90), pp. 305-308, Paris.

Uncorrelated principal components as perpendicular vectors. Retrieved from: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=eurekah&part=A3780&rendertype=figure&id=A3821 (Last accessed on 23.06.2009).

United Nations – Department of Economic and Social Affairs / Statistics Division, (2004), Integration of GPS, Digital Imagery and GIS with Census Mapping, New York. Retrieved From: http://unstats.un.org/unsd/Demographic/meetings/egm/CensusEGM04/docs/AC98_14.pdf (Last accessed on 02.08.2009)

United Nations – Department of Economic and Social Affairs / Statistics Division, (2008), Principles and Recommendations for Population and Housing Censuses, Revision 2, New York.

United Nations – Department of Economic and Social Affairs / Statistics Division, (2009), Handbook on Geospatial Infrastructure in Support of Census Activities, New York.

United States Census Bureau, (2005-a), Geographical Areas Reference Manual, Chapter 1 – Census Bureau Geography. Retrieved From: http://www.census.gov/geo/www/garm.html (Last accessed on 02.09.2009).

United States Census Bureau, (2005-b), Geographical Areas Reference Manual, Chapter 2 – Geographic Overview. Retrieved From: http://www.census.gov/geo/www/garm.html (Last accessed on 02.09.2009).

United States Census Bureau, (2005-c), Geographical Areas Reference Manual, Chapter 3 – Small Area Geography. Retrieved From: http://www.census.gov/geo/www/garm.html (Last accessed on 02.09.2009).

United States Census Bureau, (2005-d), Geographical Areas Reference Manual, Chapter 10 – Census Tracts and Block Numbering Areas. Retrieved From: http://www.census.gov/geo/www/garm.html (Last accessed on 02.09.2009).

United States Census Bureau, (2005-e), Geographical Areas Reference Manual, Chapter 11 – Census Blocks and Block Groups. Retrieved From: http://www.census.gov/geo/www/garm.html (Last accessed on 02.09.2009).

United States Census Bureau, (2008), Redistricting Data Prototype, Appendix A – Geographic Terms and Concepts. Retrieved From: http://www.census.gov/geo/www/geoareas/GTC_08.pdf (Last accessed on 02.09.2009).

United States Census Bureau, (2009), Census Bureau Legal and Statistical Geographic Entities. Retrieved from: http://www.census.gov/geo/www/geodiagram.pdf (Last accessed on 02.09.2009).

Unwin, D.J., (1996), GIS, Spatial Analysis and Spatial Statistics, Progress in Human Geography, Vol. 20, pp.440-441.

Vesanto, J. & Alhoniemi, E., (2000), Clustering of the Self-Organizing Map, Transactions on Neural Networks, Vol.11, No.3, pp: 586-600.

Vesanto, J., Himberg J., Alhoniemi, E. & Parhankangas, J., (2000), Technical Report on SOM Toolbox 2.0. Retrieved from: http://www.cis.hut.fi/projects/somtoolbox/package/papers/techrep.pdf (Last accessed on 02.06.2009).

Voas, D. & Williamson, P., (2001), The Diversity of Diversity: A Critique of Geodemographic Classification, Royal Geographical Society, Area, Vol.33, pp.63-76.

Vyas, S. & Kumaranayake, L., (2006), Constructing Socio-Economic Indices: How to use Principal Components Analysis, Health Policy and Planning, Vol.21, pp. 459-468, London, UK.

Web 1: Automated Zone Matching (AZM) Software, (2009). Retrieved from: http://www.public.geog.soton.ac.uk/users/martindj/davehome/software.htm (Last accessed on 22.06.2009).

Web 2: Geoda Software, (2009). Retrieved from: http://geodacenter.asu.edu/ (Last accessed on 05.06.2009).

Web 3: Geo-SOM, (2009). Retrieved from: http://www.isegi.unl.pt/labnt/geosom/ (Last accessed on 12.07.2009).

Web 4: Keçiören Municipality Photo Gallery (2008). Retrieved from: http://www.kecioren.bel.tr/index.php?option=com_content&view=article&id=94 (Last accessed on 29.11.2009).

Web 5: R for statistical Computing, (2009). Retrieved from: http://www.stats.bris.ac.uk/R/index.html (Last accessed on 18.04.2009).

Web 6. Retrieved from: http://neural.cs.nthu.edu.tw/jang/books/dcpr/example/dataClustering/output/dendrogram01.png (Last accessed on 30.06.2009).

Web 7.Retrieved from: http://people.revoledu.com/kardi/tutorial/kMean/WhatIs.htm (Last accessed on 21.07.2009).

Web 8. Retrieved from: http://www.cis.hut.fi/somtoolbox/documentation/somalg.shtml (Last accessed on 01.09.2009).

Web 9: SOM Toolbox v2.0. Retrieved from: http://www.cis.hut.fi/projects/somtoolbox (Last accessed on 02.06.2009).

Web 10: The scale and aggregation effect of MAUP. Retrieved From: http://www.geog.ubc.ca/courses/geob479/notes/Why_geography_is_important.pdf (Last accessed on 09.07.2009).

Web 11.Retrieved from: http://tuikapp.tuik.gov.tr/adnksdagitapp/adnks.zul

Web 12: Basic Principles of NUTS, (2009). Retrieved from: http://ec.europa.eu/eurostat/ramon/nuts/basicnuts_regions_en.html (Last accessed on 03.09.2009).

Wikipedia, (2009-a), Artificial Neural Network in en.wikipedia.org. Retrieved from: http://en.wikipedia.org/wiki/Artificial_neural_network (Last accessed on 02.07.2009).

Wikipedia, (2009-b), k-means clustering in en.wikipedia.org. Retrieved from: http://en.wikipedia.org/wiki/K-means_clustering (Last accessed on 21.07.2009).

Wu, S. & Chow, T.W.S., (2003), Self Organizing Map Based Clustering Using a Local Clustering Validity Index, Neural Processing Letters, Kluwer Academic Publishers, Netherlands.

Xiao, N., (2008), A Unified Conceptual Framework for Geographical Optimization Using Evolutionary Algorithms, Annals of the Association of American Geographers, Vol.98, pp.795 – 817.

Yan, J. & Thill, J.C., (2009), Visual Data Mining in Spatial Interaction Analysis with Self-Organizing Maps, Environment and Planning B: Planning and Design, Vol.36, pp.466-486.

Yetik, A., (2003), The Use of Geographic Information Systems in Decision Support System & An Application Concerning Small Area Districting for Statistical Purposes in Zafer Sub-District of Istanbul-Bahcelievler, Turkish Statistical Institute – Expertness Thesis, Ankara.

Yost, K., Perkins, C., Cohen, R., Morris, C. and Wright, W., (2001), Socio-economic Status and Breast Cancer Incidence in California for Different Race/Ethnic Groups, Cancer Causes and Control, Vol. 12, pp. 703-711, Kluwer Academic Publishers, Netherlands.

Zhang T., Ramakrishnan R. & Livny M., (1996), BIRCH: An Efficient Data Clustering Method for Very Large Databases, Presented at the 1996 ACM-SIGMOD International Conference of Management of Data (SIGMOD'96), pp.103-114, Montreal, Canada.

APPENDIX A

BOUNDARIES OF NUTS REGIONS IN TURKEY

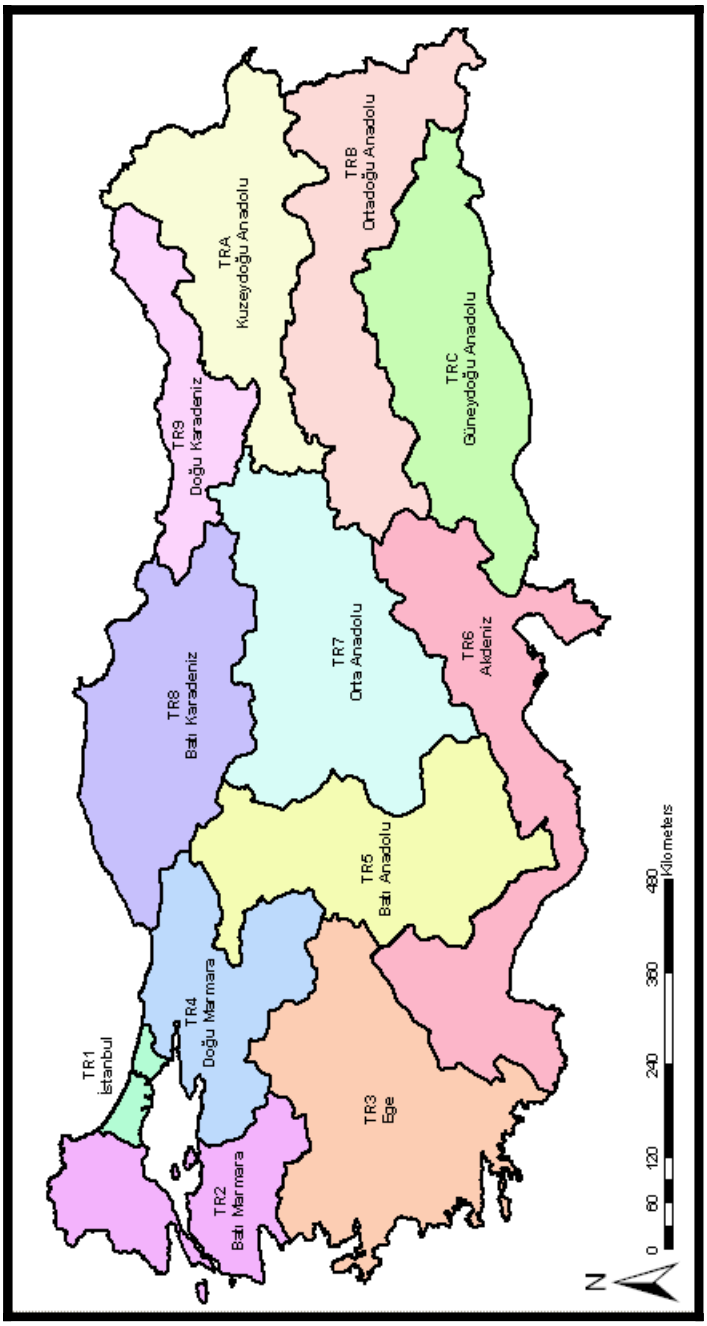


Figure A.1 Boundaries of NUTS regions in Turkey - NUTS 1 (12 Units)

Figure A.2 Boundaries of NUTS regions in Turkey - NUTS 2 (26 Units)

Figure A.3 Boundaries of NUTS regions in Turkey - NUTS 3 (81 Provinces)

APPENDIX B

B. FORMS USED IN ABPRS STUDY:

B.1. ADDRESS FORM USED WHERE MUNICIPALITY EXISTS

T.C.
BAŞBAKANLIK
TÜRKİYE İSTATİSTİK KURUMU
BAŞKANLIĞI

2006
ADRES FORMU
(Belediye Teşkilatı Bulunan Yerlere Ait)

Sayfa No :

İl adı : .....
İlçe adı : .....
Bucak adı : .....
Belediye adı : .....
Köy adı : .....
Mahalle adı : .....
MAHALLE SABİT TANITIM NO:.....
Mevki adı : .....

BİLGİSİ DOLDURULAN YER MEYDAN, BULVAR, CADDE, SOKAK VEYA KÜME EVLERDEN HANGİSİ İSE AŞAĞIDA ADINI YAZINIZ, BELİRTİLEN TÜRÜ SEÇENEKLERİNDEN BİRİNİ VE GELİŞMİŞLİK DURUMUNU İŞARETLEYİNİZ. (CADDESİ, CD., SOKAĞI, SK., KÜME EVLER VB. UZANTILARINI YAZMAYINIZ)

.....

TÜRÜ: MEYDAN ☐ BULVAR ☐ CADDE ☐ SOKAK ☐ KÜME EVLER ☐

GELİŞMİŞLİK DURUMU GELİŞMİŞ ☐ ORTA GELİŞMİŞ ☐ GELİŞMEMİŞ ☐

MEYDAN/BULVAR/CADDE/SOKAK/KÜME EVLER SABİT TANITIM NO...

	İlk Numarası	Son Numarası
Meydan ise		
Bulvar, cadde veya sokağın tek numaralı tarafı		
Bulvar, cadde veya sokağın çift numaralı tarafı		
Küme evler ise		

SIRA NO	ADRESİN				NUMARALI YERİN NİTELİĞİ	NİTELİK KODU	BİNANIN BAŞKA MEYDAN, BULVAR, CADDE veya SOKAĞA AÇILAN KAPISI VAR İSE			ADA/PAFTA/PARSEL			BELEDİYE NOTLARI		
	DIŞ KAPI NO (ARSA İSE TAHSİS NO)	İÇ KAPI NO (DIŞ KAPIYA BAĞLI)	SİTE ADI	BLOK ADI	Kullanım amacını açıkça yazınız. Örneğin; konut, bakkal, kasap, berber,okul, resmi daire,cami, hastane, samanlık, ahır, depo, garaj, arsa, tahsis, boş yerler için, boş konut veya boş işyeri yazınız.	1. Konut 2. Özel işyeri 3. Kamu işyeri 4. İnşaat 5. Arsa 6.Tahsis 7. Yazlık/ Mevsimlik 8. İmara açılan arsalar 9.Diğer	Meydan, Bulvar, Cadde veya Sokağın adını yazınız.	Sabit tanıtım No	Kapı No	Ada no	Pafta no	Parsel no	Not1	Not2	Not3
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															

Figure B.1 Address form used in places where municipality exists for ABPRS study.

B.2. ABPRS HOUSEHOLD INFORMATION FORM

FORM DOLDURULURKEN DİKKAT EDİLMESİ GEREKEN HUSUSLAR

1. Formu doldururken nüfus cüzdanı bilgilerini esas alınız.
2. Yazıları tükenmez kalem kullanarak büyük harfle yazınız.
3. Yazınızın okunaklı olmasına dikkat ediniz.
4. Şu anda bulunan adres, hanehalkının daimi ikametgah adresi ise daha önce bu hane için form doldurulmuş olsa bile formu mutlaka yeniden doldurunuz.
5. Şu anda bulunulan adres, hanehalkının ikinci konutuysa (yazlık, kışlık, yayla evi vb.) ve bu hane için daha önce daimi ikametgahında form doldurulmuşsa form doldurmayınız.
6. Konut dışında işyeri, depo vb. yerlerde ikamet edenler için de form doldurunuz.
7. Hanedeki kişi sayısı 10'dan fazla ise bu hanehalkı için birden fazla bilgi formu doldurulması gerekmektedir. Birden fazla bilgi formu doldurduldu ise ilk formun "Form No"sunu, sonraki formların ön sayfasının sağ alt köşesinde yer alan "İlk Form No" bölümüne yazınız ve formları iç içe koyarak teslim ediniz.
8. Formu anketöre herhangi bir şekilde teslim edemediğiniz taktirde, ön sayfada verilen büro veya muhtarlık adresine imza karşılığı teslim ediniz.

SORULARA İLİŞKİN AÇIKLAMALAR

01. Bu adres hanehalkının daimi olarak ikamet ettiği (yıl içerisinde en uzun süre kaldığı) adres midir? Hanehalkının sürekli veya yılın çoğunluğunda yaşadığı adres "daimi ikamet adresi"dir. Bilgi formunun "Adres Bilgileri" kısmında yazılı olan adres, bu hanehalkı için daimi ikamet edilen adres ise "Evet" kutusuna, bu adres hanehalkının daimi ikametgah adresi değil ise (yazlık, kışlık, yayla evi vb.) "Hayır" kutusuna "X" işareti konulacaktır.

Daimi ikametgah: *Bir kişinin daimi ikametgahı sürekli veya yılın çoğunluğunda yaşadığı adrestir. Eğer bir kişi yıl içerisinde birden fazla adreste ikamet ediyorsa, en uzun süre kaldığı adres daimi ikamet adresi olarak alınacaktır. Ayrıca, bir kişinin en az 6 ay süreyle kalmak niyetiyle bulunduğu yer de daimi ikametgah olarak kabul edilecektir.*

İkinci konut: *Hanehalkının daimi ikamet ettiği yer dışında, yılın belirli zamanlarında yaşadığı yazlık, kışlık, yayla evi vb. yerlerdir.*

02. Bu hanehalkı kaç kişiden oluşmaktadır? Bu hanehalkının oluşturduğu kişi sayısı açıkça yazılacaktır.

Hanehalkı: *Aralarında akrabalık bağı bulunsun veya bulunmasın aynı konutta ikamet eden bir veya birkaç kişinin oluşturduğu topluluktur.*

03. Hanenizde gelir elde etmek amacıyla kendi hesabına veya işveren olarak tarımsal (bitkisel-hayvancılık) faaliyette bulunan kimse var mı? Hanede esas işinde veya ikinci işinde, kendi hesabına veya işveren olarak tarım arazisi işleyen ve/veya hayvancılık (sadece balıkçılık ve kümes hayvancılığı yapanlar hariç) yapan en az bir kişi var ise 1. seçenek, yok ise 2. seçenek işaretlenecektir.

04. Bu hanenin aylık olarak ortalama net geliri ne kadardır? Tüm hanehalkı üyelerinin aylık ortalama net gelirlerinin toplamı dikkate alınarak ilgili kutu işaretlenecektir.

05. Hanenizin geçim kaynağı nedir? Bu soruda, hanenin geçim kaynağı; son bir yıl içinde ücret, maaş gibi düzenli gelirlerden oluşuyor ise 1. seçenek, ticari, hizmet, sanayi, tarımsal, menkul veya gayrimenkul gelirlerinden oluşuyorsa 2. seçenek, kişi veya kurumlardan sağlanan sosyal yardımlardan (2022 sayılı kanun kapsamında yaşlılık-özürlü maaşı dahil) oluşuyor ise 3. seçenek işaretlenecektir. Birden fazla seçenek işaretlenebilir.

06. Ad ve soyadınız? Bu hanede sürekli ikamet edenler ile yılın büyük bir kısmını bu hanede geçirenler, hanehalkı üyesi olarak kaydedilecektir. Hanehalkı sorumlusu 1. sıraya yazıldıktan sonra bu hanede yaşayan (askerde olanlar dahil) diğer fertlerin ad ve soyadları yaş sırasına göre yazılacaktır. Kişilerin ad ve soyadları nüfus cüzdanı bilgisi ile aynı olmalıdır.

Hanehalkı sorumlusu: *Hanehalkının sosyo-ekonomik durumu ve hanede yaşayan tüm fertlerin kişisel özellikleri hakkında en doğru bilgiye sahip, hanenin yönetim ve geçiminden sorumlu olan yetişkin hanehalkı üyesidir.*

07. Uyruğunuz? Türkiye Cumhuriyeti vatandaşları için "Türkiye Cumhuriyeti", yabancı ülke uyrukluları için "Diğer" kutusu işaretlenecektir. Yabancı uyruklu kişilerin vatandaşı olduğu ülkenin adı ve pasaport numarası pasaporttaki bilgilerden yazılacaktır.

Türkiye'de ikamet eden yabancı uyruklu kişiler: *En az 6 ay süre ile Türkiye'de ikamet eden veya en az 6 ay süre ile Türkiye'ye kalmak niyetiyle gelen yabancılardır.*

08. T.C. kimlik numaranız? Türk vatandaşlarına verilen 11 haneli bir numaradır. T.C. kimlik numarası olmayanlar veya bilmeyenler, ilçe nüfus müdürlüklerinden numaralarını öğrenerek forma yazacaklardır.

09 - 13. Sorular: Bu sorulara ilişkin bilgiler, Türkiye Cumhuriyeti vatandaşları için nüfus cüzdanı, yabancı uyruklular için pasaport bilgileri esas alınarak doldurulacaktır.

14. Hanehalkı sorumlusuna yakınlık dereceniz? Hanede yaşayanların hanehalkı sorumlusuna yakınlık derecesine ilişkin seçeneğin kodu ilgili kutuya yazılacaktır.

15. En son tamamladığınız eğitim düzeyiniz? 6 ve daha yukarı yaştaki her fert için en son bitirdiği (ferdin halen devam ettiği değil mezun olduğu okul) eğitim düzeyine ait kod ilgili kutuya yazılacaktır. Yabancı uyrukluların eğitim düzeyi, bitirilen okulun eğitim süresi dikkate alınarak doldurulacaktır.

T.C.
BAŞBAKANLIK
TÜRKİYE İSTATİSTİK KURUMU
BAŞKANLIĞI

T.C.
İÇİŞLERİ BAKANLIĞI
NÜFUS VE VATANDAŞLIK İŞLERİ
GENEL MÜDÜRLÜĞÜ

ADRESE DAYALI NÜFUS KAYIT SİSTEMİ
HANEHALKI BİLGİ FORMU

BU ÇALIŞMA, 5490 SAYILI NÜFUS HİZMETLERİ KANUNU GEREĞİNCE ÜLKE SINIRLARI İÇİNDE YAŞAYAN TÜRKİYE CUMHURİYETİ VATANDAŞLARI İLE YABANCI ÜLKE VATANDAŞLARININ İKAMET ADRESLERİNİN BELİRLENMESİ AMACIYLA YAPILMAKTADIR.

HANEHALKININ DİKKATİNE !

1. *Bilgi formu, hanede ikamet eden kişiler hakkında yeterli bilgiye sahip yetişkin hanehalkı üyelerinden biri tarafından, formun arkasındaki açıklamalar dikkate alınarak doldurulacak ve imzalanacaktır.*
2. *Form, kişilerin Türkiye Cumhuriyeti kimlik numarası ve nüfus cüzdanı bilgileri esas alınarak doldurulacaktır.*
3. *Nüfus Cüzdanı olmayan kişiler, ilçe nüfus müdürlüklerine müracaat ederek nüfus cüzdanlarını hiçbir cezai işlem uygulanmadan alacaklardır.*
4. *5490 sayılı "Nüfus Hizmetleri Kanunu" gereğince, Türkiye'de ikamet eden herkes için bu formun doldurulması zorunludur. Aynı kanunun 68. maddesi gereği formu doldurmayanlara 250 YTL, gerçeğe aykırı beyanda bulunanlara 500 YTL idari para cezası verilir.*

ADRES BİLGİLERİ

Etiket üzerindeki adreste kesinlikle düzeltme yapmayınız.
Hata olduğunu düşünüyorsanız bu hatayı anketöre bildiriniz.

Ev telefon no: ☐☐☐ - ☐☐☐☐☐☐☐

Bu bölüm anketör tarafından doldurulacaktır.

Form doldurulmadı ise nedeni;

- Boş konut (Bu konutta kimse ikamet etmiyor) ☐ 1
- Bu adres konut değil (ikamet eden kimsenin olmadığı işyeri, depo, arsa, yıkılı vb.) ☐ 2
- Bu konutta bilgi formu doldurma süresince kimse bulunamadı ☐ 3
- Daha önce ikamet adresinde bilgi formu doldurulmuş ve bu adres ikinci konut (yazlık, kışlık, yayla evi vb.) olarak kullanılıyor ☐ 4
- Diğer (açıklayınız) ☐ 5
  .....

01. Bu adres hanehalkının daimi olarak ikamet ettiği (yıl içerisinde en uzun süre kaldığı) adres midir?
- Evet ☐ 1
- Hayır (yazlık, kışlık, yayla evi vb.) ☐ 2

02. Bu hanehalkı kaç kişiden oluşmaktadır?
☐☐

03. Hanenizde gelir elde etmek amacıyla kendi hesabına veya işveren olarak tarımsal (bitkisel-hayvancılık) faaliyette bulunan kimse var mı?
- Evet ☐ 1
- Hayır ☐ 2

04. Bu hanenin aylık olarak ortalama net geliri ne kadardır?
- 0 - 150 YTL ☐ 1
- 151 - 350 YTL ☐ 2
- 351 - 500 YTL ☐ 3
- 501 - 1000 YTL ☐ 4
- 1001 YTL ve üzeri ☐ 5

05. Hanenizin geçim kaynağı nedir?
(Birden fazla seçenek işaretlenebilir)
- Maaş, ücret gibi düzenli gelir ☐ 1
- Ticari, hizmet, sanayi, tarımsal, menkul veya gayrimenkul gelirl ☐ 2
- Kişi veya kurumlardan sağlanan sosyal yardımlar (2022 kapsamında yaşlılık-özürlü maaşı dahil) ☐ 3

Bu formdaki bilgiler tarafımdan verilmiş olup, doğruluğunu taahhüt ederim.

Beyan eden kişinin;

Adı ve Soyadı : .....

İmzası : .....

Tarih : ...../...../ 200.....

Form haneye bırakılacak ise bu bölüm anketör tarafından doldurulduktan sonra haneye teslim edilecektir.

Büronun;
Adresi : .....
Tel : ☐☐☐ - ☐☐☐☐☐☐☐

Muhtarlığın;
Adresi : .....
Tel : ☐☐☐ - ☐☐☐☐☐☐☐

Anketörün:
Adı ve Soyadı:.....
İmzası :.....
Form No

Kontrolörün:
Adı ve Soyadı: .....
İmzası : .....
İlk Form No ☐☐☐☐☐☐☐

Figure B.2 Household information form used for ABPRS study.

FERT SIRA NO	AD VE SOYADINIZ? A ve B SEÇENEKLERİNİ DİKKATLİCE OKUYARAK HANEDE İKAMET EDENLERİN ADI VE SOYADINI YAZINIZ. A. Hanehalkı sorumlusunu 1. sıraya yazdıktan sonra bu hanede yaşayan diğer fertleri (askerde olanlar dahil) yaş sırasına göre yazınız. B. Hususi, yetiştirme yurdu, cezaevi ve öğrenci yurdunda yaşayanları ve on az 6 ay süre ile eğitim, iş, vb. nedenlerle başka bir hanede veya yurtdışında yaşayanları bu hanede yazmayınız.	UYRUĞUNUZ?	T.C. KİMLİK NUMARANIZ? (11 haneli T.C. kimlik numarasını her kutuya bir rakam gelecek şekilde okunaklı olarak yazınız)
	(06)	(07)	(08)
01	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
02	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
03	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
04	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
05	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
06	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
07	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
08	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
09	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	
10	Adı : ... Soyadı: ...	Türkiye Cumhuriyeti ☐1 Diğer ☐2→ Ülke adı : ... Pasaport no: ... (Yabancı uyruklular için soru 9'a geçiniz)	

NÜFUS CÜZDANI BİLGİLERİNİ KULLANARAK DOLDURUNUZ					HANEHALKI SORUMLUSUNA YAKINLIK DERECENİZ? (Hanehalkı sorumlusu; Hanenin yönetim ve geçiminden sorumlu yetişkin hanehalkı üyesidir) 1. Hanehalkı sorumlusu 2. Eşi 3. Oğlu/kızı 4. Babası/annesi 5. Kardeşi 6. Kayınpederi/ kayınvalidesi 7. Gelini/damadı 8. Torunu 9. Diğer (Uygun seçeneğin kodunu aşağıdaki kutuya yazınız)	(6 ve daha yukarı yaştaki kişiler için cevaplayınız) EN SON TAMAMLADIĞINIZ EĞİTİM DÜZEYİNİZ? 1. Okur yazar değil 2. Okur yazar fakat bir okul bitirmedi 3. İlkokul mezunu 4. İlköğretim mezunu 5. Ortaokul veya dengi mezunu 6. Lise veya dengi mezunu 7. Yüksekokul mezunu 8. Fakülte mezunu 9. Yüksek lisans ve üstü (Uygun seçeneğin kodunu aşağıdaki kutuya yazınız)
			YABANCI UYRUKLULAR İÇİN BOŞ BIRAKINIZ			
CİNSİYETİNİZ?	DOĞUM TARİHİNİZ?	BABANIZIN ADI?	ANNENİZİN ADI?	NÜFUSA KAYITLI OLDUĞUNUZ İL VE İLÇE ADI?		
(09)	(10)	(11)	(12)	(13)	(14)	(15)
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐
Erkek ☐1 Kadın ☐2	Gün: ☐☐ Ay : ☐☐ Yıl : ☐☐☐☐			İl adı : ... İlçe adı: ...	☐	☐

Figure B.2 (Continued) Household information form used for ABPRS study.

APPENDIX C

CODES DEVELOPED IN THESIS

function U2 = umatrix2colormap(sM, varargin)

% Function umat2colormap

% sM is a map structure

%%%%%%% DETAILED DESCRIPTION

% umatrix2colormap

% PURPOSE: Uses U-matrix and colormap to produce a coloring for nodes and connections of SOM grid; default colormap is used; som_grid function is called to procuce a figure of the lattice on geographic space x and y coordinates are presumed at 1.st and 2.nd positions, e.g. sMap.codebook(:,[1 2])

% SYNTAX: U2 = umatrix2colormap(sM,LineWidth,MarkerSize)

% DESCRIPTION:

% REQUIRED INPUT ARGUMENTS

% sM (struct): map structure

% 'LineWidth' (scalar): gridlines width; default=1

% 'MarkerSize' (scalar): codebook nodes size; default=1

% 'mask' (vector): mask to be used in calculating the interunit distances, size [dim 1]. Default is the one in sM (field sM.mask) or a vector of ones if only a codebook matrix was given.

% OUTPUT ARGUMENTS: U2 (matrix) the unified distance cross matrix of the SOM

% EXAMPLES:

U2 = umat2colormap(sM,'LineWidth', 15,'MarkerSize', 50, 'mask', [0 0 1 1 1]);

error(nargchk(1, Inf, nargin)); % check no. of input args

%default values

mask=sM.mask;

LineWidth = 1;

MarkerSize = 1;

```
% varargin
i=1;
while i<=length(varargin),
argok = 1;
if ischar(varargin{i}),
switch varargin{i},
% argument IDs
case 'LineWidth', i=i+1; LineWidth = varargin{i};
case 'MarkerSize', i=i+1; MarkerSize = varargin{i};
case 'mask', i=i+1; mask = varargin{i};
   otherwise, argok=0;
end
else
argok = 0;
end
if ~argok,
disp(['(som_umat) Ignoring invalid argument #' num2str(i+1)]);
end
i = i+1;
end; % while
U=som_umat(sM, 'mask', mask);
U2=som_umat2dist(U);
cmax=max(max(U2));
cmin=0;
cm_length=length(colormap);
cm=colormap;
colormap_index = fix((U2-cmin)/(cmax-cmin)*(cm_length-1))+1;
msize=prod((sM.topol.msize));
for t=1:3, for i=1:msize,for j=1:msize,
Uc(i,j,t)=cm(colormap_index(i,j),t);end;end;end
% denormalize sM
sMd=som_denormalize(sM);
% plot the lattice using geographic coordinates
```

```
% som_grid(sMd,'coord', sMd.codebook(:,[1,
2]),'LineColor',Uc,'marker','.','LineWidth',  LineWidth*U2,'MarkerSize',
MarkerSize*diag(U2));
% introduce labels
msize;
a=1:msize;
label = num2str(a');
som_grid(sMd,'coord',  sMd.codebook(:,[1  2]),'LineColor',Uc,'marker','.','LineWidth',
LineWidth*U2,'MarkerSize',  MarkerSize*diag(U2),  'Label',  label,  'LabelSize',  8,
'LabelColor','k');
h=colorbar;
set(h,'YTickLabel','');
```

function umatrix_codebook_tables(sM, sD, varargin)

```
% Function umatrix_codebook_tables
% sM is a map structure
%%%%%%%%% DETAILED DESCRIPTION
% umatrix_codebook_tables
% PURPOSE: Uses U-mat and colormap to produce a coloring for nodes and
connections of SOM grid; default colormap is used; som_grid function is called to
procuce a figure of the lattice on geographic space x and y coordinates are
presumed at 1.st and 2.nd positions, e.g. sMap.codebook(:,[1 2])
% Two figures are produced showing Umat with and without x y components
% Two tabes are output
% SYNTAX: umatrix_codebook_tables(sM, sD, varargin)
% DESCRIPTION
% REQUIRED INPUT ARGUMENTS
% sM (struct): map structure
% sD (struct):  data structure
% 'LineWidth' (scalar): gridlines width; default=1
% 'MarkerSize' (scalar): codebook nodes size; default=1
```

```
% 'mask' (vector): mask to be used in calculating the interunit distances, size [dim
1]. Default is the one in sM (field sM.mask) or a vector of ones if only a codebook
matrix was given.
% OPTIONAL INPUT ARGUMENTS
% OUTPUT ARGUMENTS
% (none)
% OUTPUTS:
% Files:
% 'codebook.dat' - codebook with fields UNIT U_MATall U_MATwoXY X Y
% 'csvbmus.dat' - table of data samples BMU with fields FID BMU
% EXAMPLES
% umatrix_codebook_tables(sM, sD,'LineWidth', 15,'MarkerSize', 50, 'mask', [0 0 1
1 1]);
% check no. of input args
%default values
mask = sM.mask;
LineWidth = 5;
MarkerSize = 5;
% varargin
i=1;
while i<=length(varargin), argok = 1;
if ischar(varargin{i}),
switch varargin{i},
% argument IDs
case 'LineWidth', i=i+1; LineWidth = varargin{i};
case 'MarkerSize', i=i+1; MarkerSize = varargin{i};
case 'mask', i=i+1; mask = varargin{i};
   otherwise, argok=0;
end
else
argok = 0;
end
if ~argok,
```

```
disp(['(som_umat) Ignoring invalid argument #' num2str(i+1)]);
end
i = i+1;
end; % while
figure(1)
Uall = umat2colormap(sM,'LineWidth',LineWidth,'MarkerSize',MarkerSize); % all
comp.
[r c]=size(sM.codebook);
m=ones(1,c);
m(1:2)=0;
figure(2)
UwoXY =
umat2colormap(sM,'LineWidth',LineWidth,'MarkerSize',MarkerSize,'mask',m); %
without x y
umatbus_all=diag(Uall);
umatbus_woXY=diag(UwoXY);
sMd=som_denormalize(sM);
ind=(1:prod(sM.topol.msize))'; % index of codebook vector
% table with U-mat_all_comp, U-mat_without x y, x and y of codebook vectors
codebook=[ind umatbus_all umatbus_woXY sMd.codebook(:,1) sMd.codebook(:,2)];
% UNIT U_MATall U_MATwoXY X Y
csvwrite('codebook.dat',codebook)
%find BMU of input data
bmus=som_bmus(sM,sD);
ind=(0:prod(size(bmus))-1)'; % index of data samples
m=[ind bmus];
% DATA/FID BMU
csvwrite('csvbmus.dat',m)
% join in ArcGIS these 2 tabelas through fields (data)FID->(Bmus)BMU-
>(Codebook)UNIT
```

APPENDIX D

POPULATION COUNTS FOR SSAs
RETRIEVED BY METHODS EMPLOYED IN THESIS

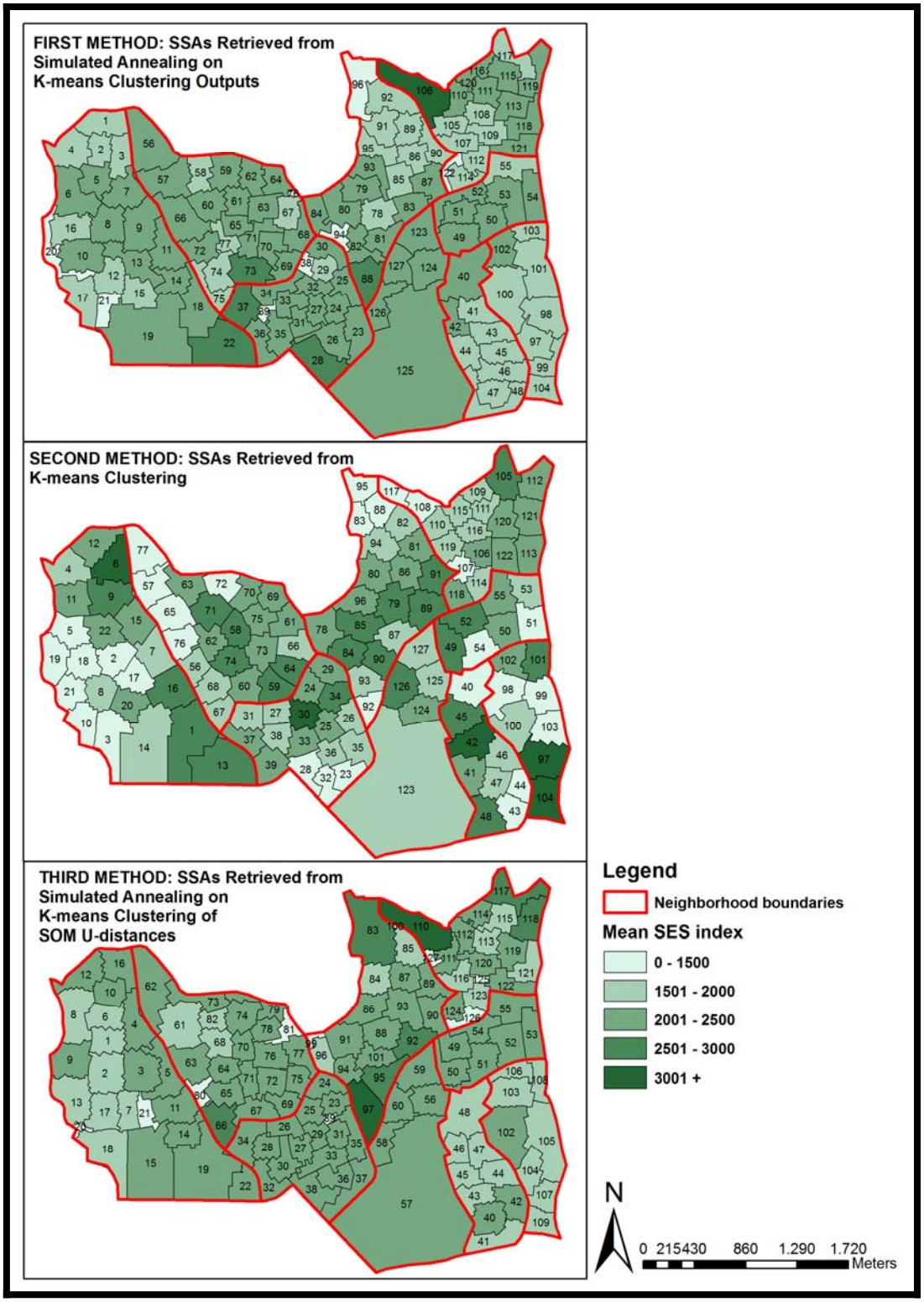


Figure D.1 Population counts for SSAs obtained by using parcels with raw data and by methods employed in thesis.

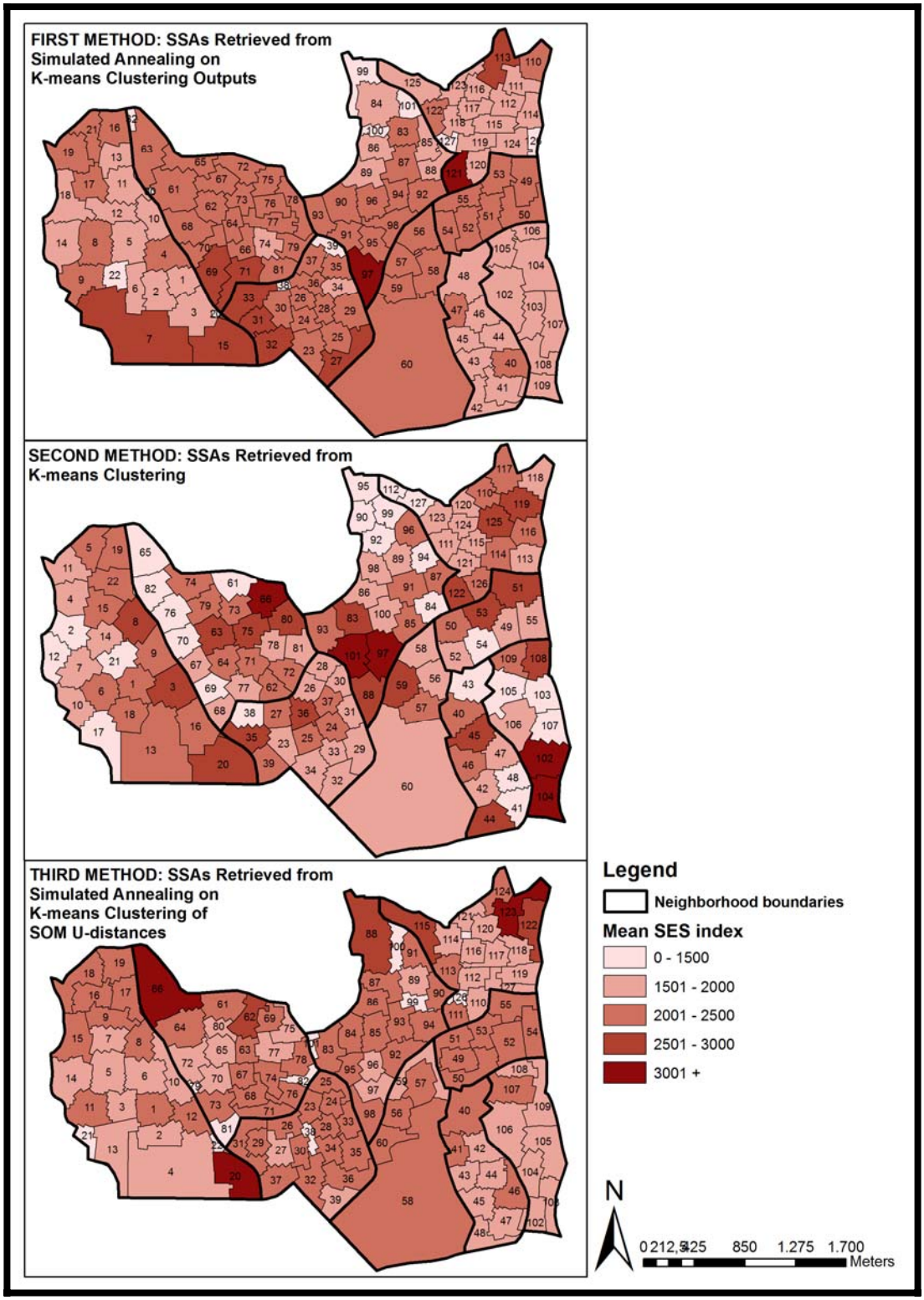


Figure D.2 Population counts for SSAs obtained by using parcels with SES indices and by methods employed in thesis.

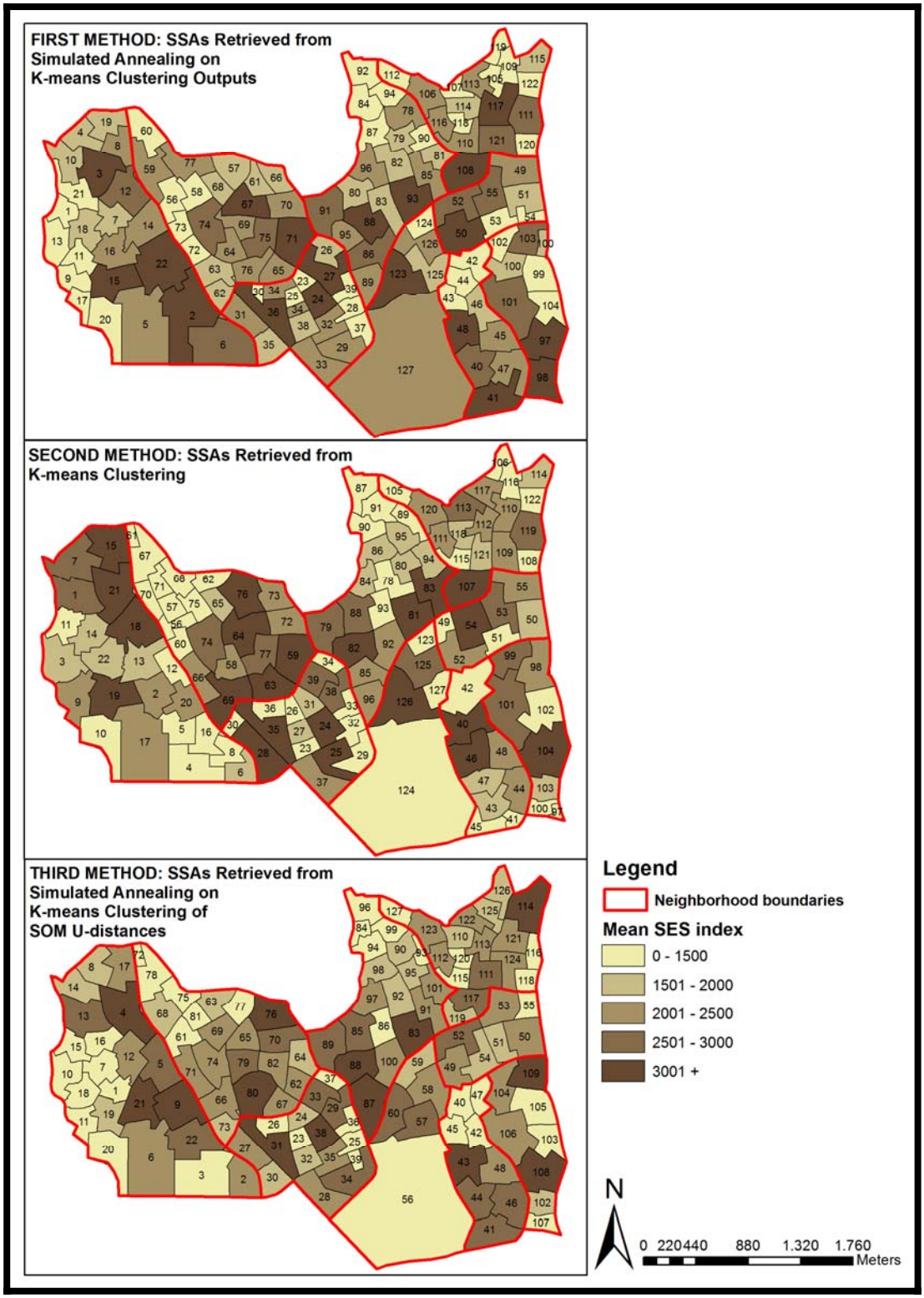


Figure D.3 Population counts for SSAs obtained by using blocks with raw data and by methods employed in thesis.

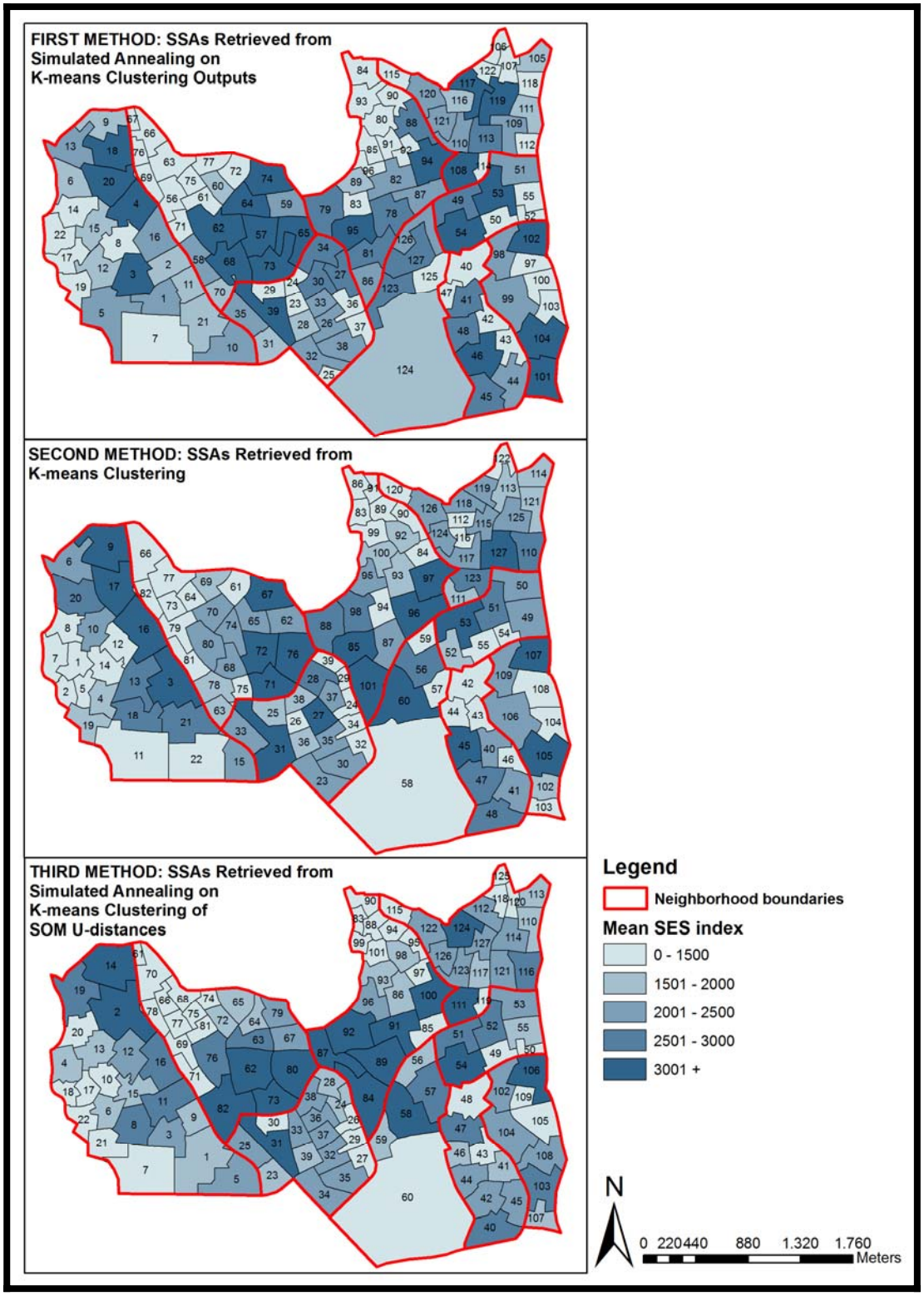


Figure D.4 Population counts for SSAs obtained by using blocks with SES indices and by methods employed in thesis.