ISTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF INFORMATICS

USING SPATIAL STATISTICS TECHNIQUES TO DETERMINE THE USER PROFILES OF SOCIAL MEDIA

M.Sc. THESIS

İrem ERKUŞ

Department of Informatics Applications

Structure Geographic Information Technologies Programme

AUGUST 2014

ISTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF INFORMATICS

USING SPATIAL STATISTICS TECHNIQUES TO DETERMINE THE USER PROFILES OF SOCIAL MEDIA

M.Sc. THESIS

İrem ERKUŞ
(706121007)

Department of Informatics Applications

Structure Geographic Information Technologies Programme

Thesis Advisor: Assoc. Prof. Dr. Arif Çağdaş AYDINOĞLU

AUGUST 2014

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ BİLİŞİM ENSTİTÜSÜ

SOSYAL MEDYADA KULLANICI PROFİLLERİNİN BELİRLENMESİNDE KONUMSAL İSTATİSTİK TEKNİKLERİN KULLANILMASI

YÜKSEK LİSANS TEZİ

İrem ERKUŞ
(706121007)

Bilişim Uygulamaları Anabilim Dalı

Coğrafi Bilgi Teknolojileri Programı

Tez Danışmanı: Doç. Dr. Arif Çağdaş AYDINOĞLU

AĞUSTOS 2014

İrem ERKUŞ, a M.Sc. student of ITU Institute of Informatics 706121007 successfully defended the thesis entitled "USING SPATIAL STATISTICS TECHNIQUES TO DETERMINE THE USER PROFILES OF SOCIAL MEDIA", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Assoc. Prof. Dr. Arif Cagdas AYDINOGLU** ....……
Gebze Institute of Technology

Jury Members : **Asst. Prof. Dr. Himmet KARAMAN** ....……
Istanbul Technical University

Prof.Dr. Taskın KAVZOGLU ....……
Gebze Institute of Technology

Date of Submission : 05 May 2014
Date of Defense : 28 August 2014

FOREWORD

First of all I would like to express my best thanks and gratitude to my supervisor Assoc. Prof.Dr. Arif Cagdas Aydinoglu for his supervision, numerous advice and inspiration. From the first stages in formulating the aim and core idea of this thesis he helped me in practical realization of my work. His idea was really helpful and his advice were the key factor in my study.

I would like to thank to my teachers in cartography and geoinformation Prof.Dr. Andrew Frank, Amin Abdalla, Paul Weiser, Thomas Fuhrmann. The practical and theoretical knowledge that I get on building a methodology will be definitely helpful in my future career.

August 2014

İrem ERKUŞ
(Mathematician)

TABLE OF CONTENTS

ABBREVIATIONS

AIC	**:** Akaike Information Criteria
GIS	**:** Geographic Information System
SES	**:** Socioeconomic status
ABPRS	**:** Address Based Population Registration System
OLS	**:** Ordinary Least Square
GWR	**:** Geographically Weighted Regression
TurkStat	**:** Turkish Statistical Institute
ESDA	**:** Exploratory Spatial Data Analysis
CDA	**:** Confirmatory data analysis

LIST OF TABLES

Page

LIST OF FIGURES

Page

USING SPATIAL STATISTICS TECHNIQUES TO DETERMINE THE USER PROFILES OF SOCIAL MEDIA

SUMMARY

Social media became hot topic after millennium. Beside communication techniques such e-mail and mobile phone, social media platforms have started to be used in various different social and working discipline, to illustrate this, Facebook for social communication, Twitter for short messages, and Flickr for photo sharing, and so on. Most social media platform has possibility to locate any social media activity spatially. Defining any activity with location or position in social media means that distribution of social media users can be analyzed spatially to determine user profiles.

The purpose of this study is to determine user profiles of social media platforms with the examples of online dating site. The data of users are integrated with in GIS environment using spatial statistics techniques.

In methodology part, spatial statistical methods were used to understand the data relationships. Firstly, the distribution of patterns is analyzed spatially to determine data outliers. Then, user profiles are analyzed by age, gender, income, educational level and occupation to examine socio-demographic relationships. Lastly, target areas are determined by using cluster analysis while the relations of the data is analyzed by using regression techniques.

The applications consist of four parts. As working principles of GIS, related spatial data was captured, queried, analyzed and visualized in ArcGIS program. Data about ten thousands of users who often use this online dating site was retrieved from the MySQL database by using spatial data query. The distribution and outliers of the data were determining by using Exploratory Spatial Data Analysis (ESDA) tools which are histogram, semivariogram, voronoi map, and QQ plots. Furthermore, the population of Istanbul was examined with spatial statistical analysis based on the data of Adress Based Population Registration System and social media site. The results were interpreted and showed in thematic maps, and presented in a diverse range of visualization options such as reports, statistical graphics and tables. The third part examine the hot-spot analysis to find the clustered area of social media users by using mapping cluster toolset. The final part demonstrate the OLS and GWR regression methods to identify regression model by using spatial relationships tools. In this way, this study contributes to research about social media as a new developing discipline. This gives a perspective of user profile in view of education and occupation by using spatial statistical approaches. All in all, this thesis gives an idea about social media and usable statistical techniques.

SOSYAL MEDYADA KULLANICI PROFİLLERİNİN BELİRLENMESİNDE KONUMSAL İSTATİSTİK TEKNİKLERİN KULLANILMASI

ÖZET

Sosyal medya platformları günümüzde en hızlı, en popüler ve en etkili iletişim aracı olmuş ve iletişim dünyasının dinamiklerini farklı bir boyuta taşımıştır. Dünya yüzeyinde neler olduğu hakkında iletişim halinde olduğumuz ve bilgiyi paylaştığımız bu platformda, sosyal medya kullanıcı sayıları ve istekleri giderek artmaktadır. Milyonlarca kullanıcı günlük konularla ilgili paylaşımlarını bu kanallar üzerinde gerçekleştirmektedir. Bu sitelerin her biri farklı amaçlar doğrultusunda oluşturulduğu için, bu platformlar çeşitlere ayrılmaktadır. Örneğin, Facebook insanların başka insanlarla iletişim kurmasını ve bilgi paylaşımını amaçlayan bir sitedir. Twitter ise bir sms karakter sınırı mesajlar için kullanılan mikroblog sitesidir. Bütün sosyal medya kullanıcıları tarafından en çok ihtiyaç duyulan bilgi ise konum bilgisidir. Konum bilgisi sadece sosyal medya kullanıcılarının aktivitelerinin yerini belirtmek için değil, aynı zamanda sosyal medya site yöneticilerinin veya bu konuyla ilgili araştırma yapmak isteyen sosyolog, istatistikçi, haritacı, CBS uzmanları için de bilinmesi gereken en önemli veridir. Günümüzde yazılımların geliştirilmesiyle birlikte, kullanıcılar konum bilgilerini paylaştığında, bu bilgiyi harita üzerindeki dağılımını göstermek veya kullanıcı profilini belirlemek mümkün olmaktadır.

Bu çalışmanın amacı, ciddi ilişki niyetindeki bireyleri internet üzerinde tanıştırma amacıyla kurulan sosyal medya sitesi örneğiyle CBS ortamında konumsal istatistik tekniklerini kullanarak ve siteden temin edilen canlı veriler ile birlikte kullanıcı profillerinin belirlenmesidir. Bu çalışma kapsamında, Türkiye'de popüler olan tanışma sitesindeki sosyal medya kullanıcılarının mekansal analizini sunmak için kullanıcıların sosyo-demografik ve sosyo-ekonomik verileri alındı. Bunlar yaş, cinsiyet, evlilik durumu, iş, eğitim seviyesi, etnik köken, din ve gelir durumu bilgisidir. Bu verilerin her birinde il ve ilçesi bilgisi mevcuttur.

Bu çalışma kapsamında, CBS'nin temel fonksiyonlarından veri toplama, depolama, sorgulama, analiz ve görüntüleme kısımları uygulanmıştır. Sosyal medya kullanıcı verileri özel bir evlilik sitesinden temin edilmiştir. Siteyi sıklıkla ziyaret eden yaklaşık on bin kullanıcının verileri, mekansal sorgulamalar kullanılarak MySQL veritabanından çekilmiştir. Verilerin alındığı yıl ise 2013'tür. Bu veriler ArcGIS'e aktarıldı ve analize hazır hale getirilmek için bir kaç işlemden geçmiştir. Bu işlemlerden bazıları, çalışma alanı İstanbul olduğu için İstanbul ilçe düzeyinde

verilerin düzenlenmesi ve aykırı verilerin çeşitli istatistik yöntemlerle ayıklanmasıyla gerçekleşmiştir. Verilerin toplanıp, ayıklanıp ve bütünleşme aşamaları tamamlandığında veriler analiz için kullanılır hale getirildi. Veriler arasındaki ilişkiyi incelemek için en uygun istatiksel yöntem belirlendi. Metodolojik olarak, kullanıcı profilleri ilçe bazında yaş, cinsiyet, gelir durumu, eğitim düzeylerine bakılarak sosyo-demografik ilişkileri ile irdelendi. Hot-spot ve regresyon analizleri yapıldı ve harita üzerinde sunuldu.

Yoğunluk analizlerini de hesaba kattığımızda çalışma alanımızı İstanbul olarak belirlememiz, hem verilerin düzenlenmesinde, hem de çalışma alanının küçülmesi sebebiyle daha iyi sonuçlar alanacağı öngörüldü. Çünkü, Türkiye genelinde 81 ilimiz, İstanbul genelinde 39 ilçemiz vardır. Ayrıca, İstanbul'daki kullanıcı verilerini daha iyi anlamak için, TÜİK'in Adrese Dayalı Nüfus Kayıt Sistemi veritabanından 2013 yılına ait İstanbul'a ait nüfus bilgileri alındı ve internet üzerindeki sosyal medya kullanıcı verileriyle karşılaştırıldı. Her iki kurumdan alınan verilerden yoğunluk analiz haritaları oluşturmak için ArcGIS yazılım araçları kullanıldı. Sonuç olarak TÜİK'ten alınan verilere gore İstanbul'daki bekar sayısı en yoğun olan ilçeler; Bağcılar, Bahçelievler, Küçükçekmece, Üsküdar, Ümraniye, Kadıköy ve Pendik olarak belirlenmiştir. Sosyal medya kullanıcılarının en yoğun olduğu ilçeler ise; Bahçelievler, Kadıköy, Beyoğlu ve Beşiktaş olarak belirlenmiştir. TÜİK verilerine göre bekar erkeklerin en yoğun olduğu ilçeler ise Bağcılar, Bahçelievler, Güngören, Gaziosmanpaşa, Kağıthane, Beyoğlu ve Fatih'dir. Bekar kadınların en yoğun olduğu ilçeler ise Bahçelievler, Güngören, Gaziosmanpaşa. Ortalama yaş olarak alınan 25-34 yaş arasındaki nüfusun en yoğun olduğu ilçe ise Üsküdar olarak tespit edildi. Sosyal medya kullanıcılarından alınan bilgiye göre ise ortalama yaşın en yoğun olduğu ilçe Bahçelievler olarak belirlenmiştir. Tüm sonuçlar, kullanıcı profilinin analizine yönelik tematik haritalar üretilerek raporlanmıştır.

İstanbul ilçelerinin nüfus bilgileri ve site kullanıcılarının dağılımları dikkate alınarak, site kullanıcılarının nüfus ilişkili yoğunluğu irdelenmiştir. Site kullanıcılarının bekar olduğu dikkate alındığında, İstanbul'daki bekar sayısı en yoğun olan ilçeler belirlenmiş, kadın ve erkek kullanıcı oranlarına göre hedef kitle incelenmiştir. Site kullanıcılarının yoğunluğu orta yaş grubu olarak belirlenerek, nüfus dağılımındaki yaş gruplarına göre öncelikli ilçeler irdelenmiştir. Ayrıca eğitim durumu ve gelir durumuna göre kulanıcı kitlelerinin mevcut durumu belirlenmiştir.

Sosyal medya kullanıcılarının dağılımı ve ilişkilerini test etmek için özelliklerine göre histogramları üretilmiştir. Kullanıcının konumsal dağılımına göre ortalama ve orta konumları belirlenmiştir. Standart sapmasına göre verilerin dağılım doğrultuları belirlenmiştir. Semivariogram, Voronoi haritası ve QQ plot eğrileri ile analiz edilerek, kullanıcıların konumsal dağılımında ve belirli özelliklerine göre konumsal ilişkileri analiz edilmiştir. Sonuçlar tematik haritalarda gösterilerek, gafik ve tablolarla raporlanmıştır.

Kümeleme analizleri kapsamında hot-spot analizi kullanılarak, hedef çalışma alanı belirlenmiştir. Istanbul çalışma alanında yapılan hot-spot analizine göre de ilçe bazlı yüksek yoğunluktaki ve düşük yoğunluktaki yerler belirlenmiştir. Genel anlamda hedef kitlenin nerede kümelendiği belirlenmiştir.

Yapılan mekansal analizler araştırmacıya genel olarak 'Nerede?' sorusuna cevap verir. Bu tez kapsamında cevaplanması gereken soru ise 'Neden?' olacaktır. Bu soruya cevap verecek analiz metodu ise regresyon analizidir. Bu çalışmada regresyon analizinin en çok kullanılan iki metodu tartışıldı. Bunlardan birisi En Küçük Kareler metotu diğeri

Coğrafi Ağırlıklı Regresyon metodudur. Bu iki regresyon yöntemi ArcGIS yazılımının 'Mekansal İlişkilerin Modellenmesi' aracında bulunur. İki metottan ilki olan en küçük kareler yönteminde kontrol etmemiz gereken altı istatistiksel kural vardır. Bu kuralların sırası önemli değildir. Altı kuraldan biri bağımsız değişkenlerinin modeline yardım edip etmediğini bulma yöntemidir. Bu yöntemde verilerin birbiriyle ilişkilerini anlamak için her bağımsız değişkene bir katsayı verilir. Bu katsayıların sıfır ve sıfıra yakın olanları bağımsız değişken olmaktan çıkarılır. Çünkü ilişkiyi belirlemede bir katkıları yoktur ve oluşturulacak modele yardım etmez. Veriler arasındaki ilişki sabit değilse Robust olasılığı dikkate alınır. İkinci metot olan Coğrafi Ağırlıklı Regresyon metotu daha yerel model üretmek için kullanılır ve her özellik için ayrı bir denklem oluşturur. OLS ve GWR arasındaki en büyük farklardan birisi de, OLS regresyon metodunun daha global, GWR regresyon metodunun daha yerel olmasıdır. Bunun dışında OLS bütün özellikler için tek bir denklem oluştururken, GWR her özellik için ayrı ayrı denklem oluşturmaktadır. Bu metotların incelemesinden sonra sosyal medya kullanıcıların İstanbul'daki verilerinin arasındaki ilişkiyi incelemek, araştırmak ve modellemek için iki yöntem de sosyal medya kullanıcı verilerine uygulandı. Sonuç olarak, her iki metot sonucunda sosyal kullanıcı profillerini belirlerken en etkili değişkenlerin gelir ve eğitim olduğu belirlendi. Bu iki yöntem için uygulanan model doğrulukları birbirine çok yakın olduğundan, iki yöntem de kullanıcı profili belirlenmesinde kullanılabilir.

Bu tez kapsamında, CBS'nin konumsal istatiksel tekniklerinin sosyal medya alanı gibi yeni bir disiplinde kullanılabilirliği gösterilmiştir. Bu tez, ileride sosyal medya alanında yapılacak tezlere ışık tutacak ve yeni bir bakış açısı kazandıracaktır. Sonuçta, sosyal medyada konumsal istatistik tekniklerini bütünleştirici bir metodoloji belirlenmiştir.

1 INTRODUCTION

According to UN (2013), “The current world population of 7.2 billion is projected to increase by almost one billion people within the next twelve years, reaching 8.1 billion in 2025 and 9.6 billion in 2050”. The key point is that large amounts of people around the world use a social networking site. Interpreting these results, Gaudin et al. (2013) indicate that “By 2017, when some estimates show the world population reaching 7.44 billion, 31.3% of the world will be using social networks, according to eMarketer” contributing the research by given number of percentage of social network users.

The term of social media is related with online application. It is the type of online networking. The oldest online social media is E-mail. It allowes to users to communicate with each other. Thanks to world wide web, online social networks numbers are increased. Nowadays, social networking sites allow individuals to construct a public profile within the system and view others profiles. These connections may vary from site to site. The most popular social networks such as Facebook through social sharing sites such as YouTube and Flickr. (Boyd and Ellison, 2007).

Quantative revolution was one of the biggest development of modern geography. It led to an increased use of computerized statistical techniques in geographical research. Before the revolution, geographical researches could not explain social, economic, and political situation for a given time and place. Therefore, starting in the 20th century, the concept of social mathematics appeared. The term social mathematics was first used by Rosander (1935). He recognized the relations between social studies and mathematics. Furthermore, Mauch (2005) defined social mathetmatics as “instruction and realistic practice in recognizing, collecting, measuring, judging, and analyzing information presented numerically, recognizing conclusions drawn from data that are faulty, and communicating mathematical and statistical evidence and conclusions to others”.

Geographic Information Systems (GIS) has tools for spatial data management, mapping, cartography, spatial data analysis, and decision-making. Within the past few decades the advent of computer technologies have forced the researchers to understand

and to explain the root causes or processes about spatial patterns of data and their relationships. Spatial data analysis provides the main tools for addressing complex problems about spatial data. Bailey and Gatrell (1995) explain the difference between spatial analysis and spatial data analysis. Spatial analysis includes simple GIS operations such as buffering, querying, etc. Spatial data analysis requires statistical theory and software to analyze data such as regression analysis. Spatial data analysis is the discipline of the spatial statistics called “confirmatory data analysis”. According to John Tukey, There are two types data analysis; exploratory data analysis (EDA) and confirmatory data analysis (CDA) (Tukey, 2002). While EDA is used to look for unpredictable patterns, CDA is used to prove the hypotheses (Brillinger, 2002).

In today’s context from GIS to spatial statistics, GIS were rapidly becoming part of the mass media and social media. Social media are increasingly becoming like GIS, because media are increasingly location-based (Goodchild, 2011). Sharing professional knowledge, social and political thoughts on mobile and web platforms have become another default place in this virtual world. Spatial statistics and data mining platforms are intended to provide decision support mechanisms. Spatial statistics, therefore, can be used to produce a meaningful and valuable information from large-volume data. In addition, social media has become widely popular as a consequence of characterizing the social media users according to socio-economic and socio-demographic features such as age and gender (Bergsma and Durme, 2013).

The new spatial statistical tools are designed to solve location-based problems. Spatial statistics are used for spatial distributions, patterns, process and relationships in a GIS environment. Spatial statistics involves a set of techniques to examine spatial data. These techniques especially use space-area, length, proximity, orientation, or spatial relationships (Scott and Getis 2008).

The spatial statistics provides powerful tools for computing spatial patterns. The spatial pattern analysis tools can help know where something is but regression analysis understanding why happening then try to find solve the problem. Regression is a widely used set of statistical techniques allowing you to examine, model, and explore data relationships (ESRI, 2013). In order to understand the relationships between spatial datasets, it is necessary to produce spatial data in the region of visual information. One such method is to produce hot-spot and density map.

This thesis consists of five chapters including this introductory chapter. The first chapter includes aim of thesis and methodology. The second chapter is a comprehensive Fundamentals part that examines the spatial data with geographic distributions, histograms and numbers. In addition, this chapter gives an idea about two types of regression analysis which are OLS and GWR. Lastly, Fundamentals chapter includes information about social media researchers related to works. The third chapter focuses on the applications used to analyze for thematic analysis, statistical analysis, hot-spot and regression analysis. This chapter indicates the results of application with graphs and maps. The conclusion part summarizes the results of this master's thesis also explain the data relationships between users. Furthermore, it shows the results of the research, gives an idea about integrating social media sector and GIS.

1.1 Purpose of Thesis

This thesis aims to examine spatial statistics techniques to determine the user profiles of social media users. The purpose of this study guides is to identify the key factors of user profiles and answer these questions:

- Which spatial statistical techniques are used to identify user profiles on social media sites?
- How spatial statistical techniques are used to identify the demographic characteristics of social media users?
- Which regression model is the best to describe the relationships between the data to determine user profile on social media?

The goal is to define the user profile and geographic relationship in databases with spatial statistical analysis and decision support mechanisms will provide information on the approaches.

This thesis addresses spatial relationships issue and includes discussion on identifying key factors of values and method in order to set a model. This thesis aims to prove the relationships between geographic features by utilizing ArcGIS spatial statistics tools. This study examines the relationship between social media users and demographic categories in terms of age and income distribution.

1.2 Methodology

The methodology of this thesis divided into three sections: literature review, data gathering and preparation, analysis and evaluation. Figure 1.1 shows that methodology of this thesis.

The first section comprehensives literature review that examines research related to spatial statistics and usage in GIS. Literature review part informs about social media concepts with area of usage. In addition, it helps to understand current research of social media in GIS and socio-demographics applications using spatial statistical techniques. The best spatial statistics techniques were picked for the analysis.

The second part includes how the data collected and prepared for the analysis. The data gathering and preparation stage mainly concerns the extraction of third-party records from the chosen online social networks site and also district data of Istanbul from TurkStat. These spatial data are stored in ArcGIS geo-database.

The third part guides to analyze and evaluate for understanding relationships between data. This part helps to the interpretation of social media user data by using histograms, thematic maps, spatial statistics, and regression methods.

LITERATURE REVIEW
DATA GATHERING& PREPARATION
ANALYSIS& EVALUATION

Figure 1.1: Methodology sections.

1.3 Related Works

With the popularization of GIS in social context, researchers have began to study about the distribution of social events. Table 1.1 summarized statistics techniques used in selected works below;

Dai, et al. (2013) study illustrates that OLS regression and GWR help to characterize the neighborhoods based on socioeconomic factors such as race, income and education level to find out childhood drowning in Georgia. This particular study determined that "A multi-level model is possible to incorporate both individual factors (e.g., race, age, and gender) and neighborhood factors (i.e., percentage of black population and median income in a census tract). " (Dai et al., 2013).

Park, et al. (2012) focused their study on identifying the factors regarding community residents in rural tourism villages. The study analyzed in three stages which are factor analysis, cluster analysis, and logistic regression analysis. This model shows that the social capital as a dependent variable and socio economic factors were used as independent variables: income, age, gender, education, residence period, the sales of agricultural products etc.

Cheng and Fotheringham (2013) noted that social class and employment rate are two determinants to determine multi-scale effects for comparative analysis. They also used OLS and GWR technique to show differences in social processes.

Karlsdotter, et al. (2012) concentrate on relationship between income inequalities and health. In this study, they used multilevel logistic regression model based on socio-demographic (age, sex, marital status, nationality) and socio-economic (educational level, personal income, unemployment).

Chalkias, et al. (2013) attemps to examine relationship between childhood obesity and socioeconomic in the metropolitan area of Athens, Greece. They used OLS and GWR statistical analysis to test this relation. The results of this analysis indicate that educational level is the most important predictor of childhood obesity and also conclude that local (GWR) method gets more useful information than the global (OLS) method for this research.

Zheng and George (2012) present a study about increasing income inequality on the health in the U.S. This study justifies the two following hypotheses. Zheng and George (2012) suggested that the "Increasing income inequality over time is associated with increasing poor physical functioning and activity limitations" and second, "Increasing income inequality will interact with SES (income, education, and employment status) and with gender, race/ethnicity, and marital status such that income inequality has stronger effects for the disadvantaged than for the advantaged".

Deaton and Lubotsky (2003) attempted to deal with mortality rate and relationships between income inequality in American cities and states. Their study focused on blacks and whites and correlation between their incomes.

Table 1.1: Some online dating websites and their types.

Author(s)	**Regression Analysis**	**OLS**	**GWR**	**Social Behaviour**	**Demographic Analysis**
Dai, et al. (2013)		**X**	**X**		
Park, et al. (2012)	**X**				**X**
Cheng and Fotheringham (2013)		**X**		**X**	
Karlsdotter, et al. (2012)	**X**				**X**
Chalkias ,et al.(2013)		**X**	**X**		**X**
Zheng and George (2012)					**X**
Deaton and Lubotsky (2003)					**X**

2 FUNDAMENTALS

2.1 Spatial Data and Statistic Concept

Information systems store information about people, places within the organization. By information we mean data that have been shaped into a form that is meaningful and useful to human beings. Data, in contrast, are representing events before they have been organized and arranged into a form that people can understand and use.

Information systems can be divided into two main categories: spatial and non-spatial information systems. Non-spatial information systems do not define specific coordinate about place but these systems use information technologies without spatial information. According to Yomralıoglu (2000) non-spatial information systems could be used in the following areas:

- Data Processing
- MIS
- Decision-Support System
- Office Automation
- Artificial Intelligence

Spatial information systems provide spatial and attributes data about object. This system store graphical and non-graphical information in a database. Yomralıoglu (2000) gave examples about spatial information systems used in the following areas:

- Enviromental
- Infrastructure
- Cadastral
- Social-economic

Dale and McLaughlin (1988) describe GIS as type of information management system and it provides visualization for general map information.

Data is the most significant component for GIS. Geographic and attributes data could be gather from essential resources or can be purchased as a ready to use. According to Yomralıoğlu (2000), data gathering process is the most difficult part for GIS, because data sources are plentiful, have different structure. For this reason, data gathering process causes time and cost losses.

GIS data can be separated into two categories: the data represented by vector and raster forms (Figure 2.1). Vector data is split into three types: polygon, line and point data as seen on. Polygon features are used to measure the area. Polygon features are commonly used in mapping symbology, patterns or numeric and color gradation scheme. Line data is used to show linear features. Line features have starting and ending point. Common examples would be road, rivers and street. Point data is used to represent discrete data points. Point features could be city locations or place names.

In its simplest form, a raster data consists of a matrix of cells or pixels organized into rows and columns in simplest form. Where each cell contains a value representing information. Rasters are digital aerial photographs, satellite imagery, digital pictures, or even scanned maps.

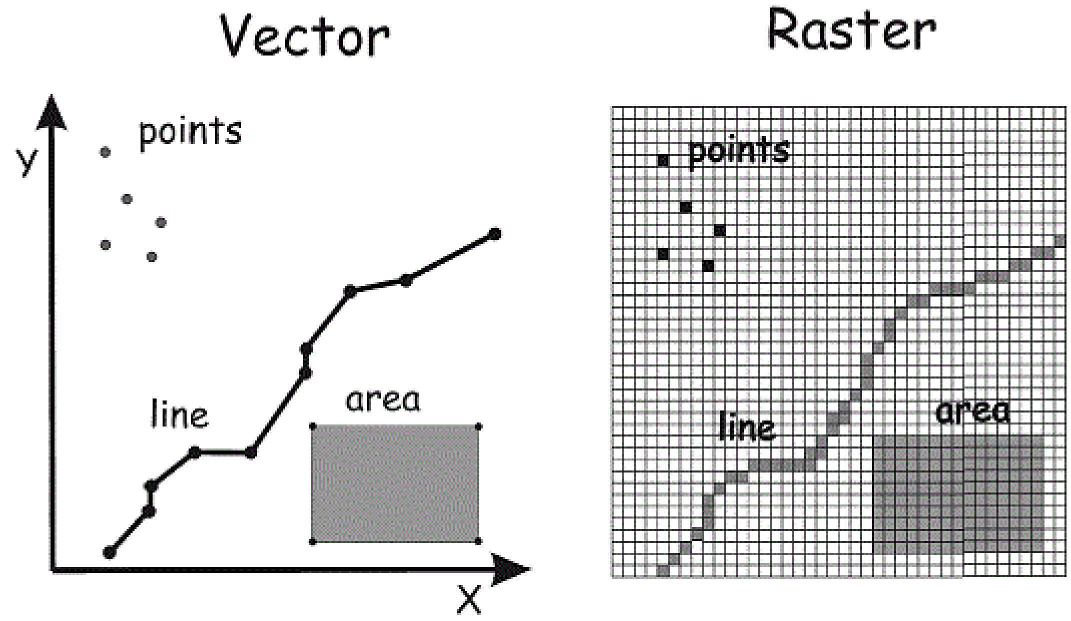


Figure 2.1: Vector and raster data model.

All spatial data must be represented by geographic coordinate systems. According to location of data, there are three types of spatial data: discrete, regional and continuous (Figure 2.2). Modeling of discrete point data is called point pattern analysis, modeling of spatial continuous data is called geostatistics, and modeling of regional data is called lattice analysis.

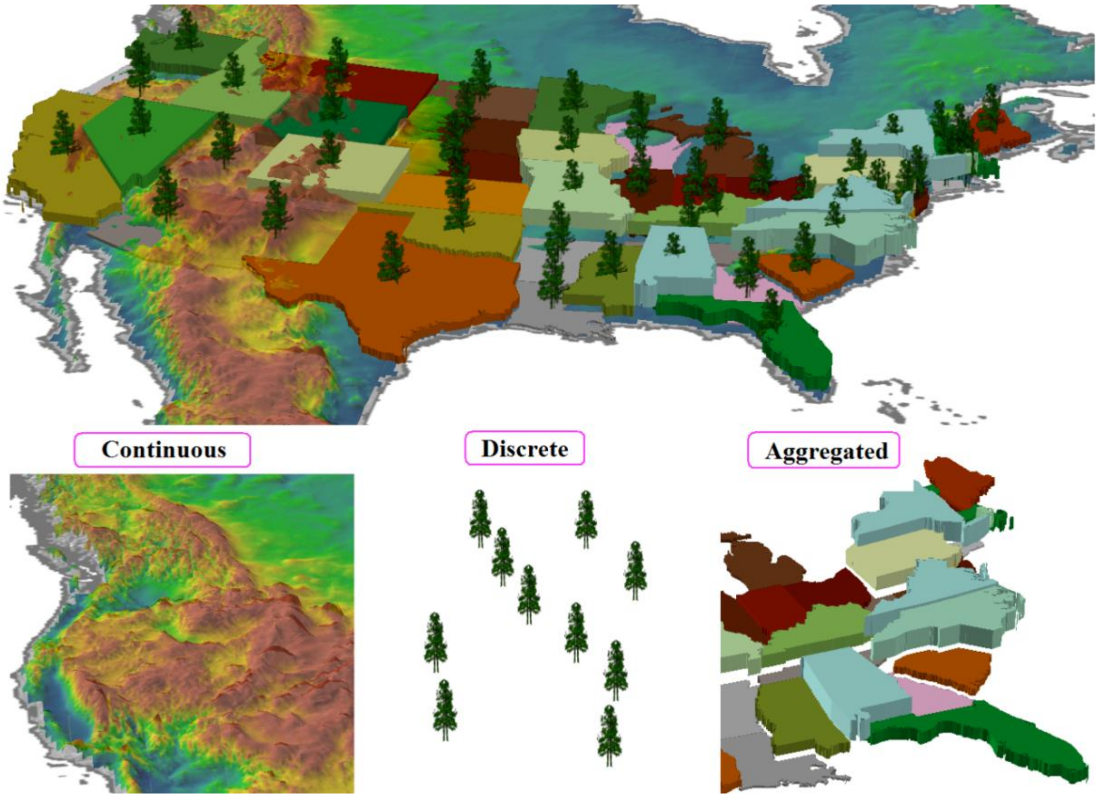


Figure 2.2: Types of spatial data.

Statistics is a numerical summary about a group of observations. GIS used in a wide range of application areas has ability to use statistical data on complex analysis processes. Database can manage massive number of data. The total value, minimum value, maximum value, mean value, and standard deviation can be querable for spatial data. Frequency and normal distribution of the data can be examined. The statistical analysis results in reports and graphs.

2.2 Spatial Statistics

To understand and interpret spatial data, GIS functions are used and this process is called spatial statistics or spatial analysis. According to Goodchild (2011), spatial analysis is a set of methods whose results change when the locations of the objects change.

There is a significance difference between spatial analysis and spatial data analysis. Spatial data analysis uses statistical theory and software to analyze with geographic coordinates.

Spatial statistics includes various techniques using spatial data. It helps to identify the spatial patterns and relationships in data. Various application fields have contributed to improve it. GIS as contributor provides software for revealing complexity of spatial pattern in the analytic toolbox.

The following list briefly summarizes some spatial statistics methods which are used in this thesis study. These statistical methods are explained with mathematically in the next sections. As examples of spatial statistics techniques;

- Hot-Spot
- Descriptive Statistics
- Voronoi
- Normal and QQ Plot
- Koenker Statistics
- Jarque-Bera statistic
- Regression

2.2.1 Exploring Spatial Distribution of The Data

Data exploration is the first step in a data analysis project. Patterns, clusters, and distributions are determined to explore data. Firstly, analyst should ask these questions;

- How features are distributed?
- Where the cluster located?
- What is the pattern?

Measuring techniques can be used to understand the distribution of the data. These techniques are mean center, median center, central feature, standard distance and directional distribution tools (Figure 2.3).

Mean Center determines average x-coordinate and y-coordinate of all features in the study area. Formula 2.1 explains mean X and Y depending on x_i and y_i coordinates for the feature class, n is the total number of features.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2.1}$$

Median Center is the shortest total distance to all features in the study area. It is usually calculated using Euclidean distance.

$$d_i^t = \sqrt{\left(X_i - X^t\right)^2 + \left(Y_i - Y^t\right)^2} \tag{2.2}$$

Formula 2.2 explains at each step (t) in the algorithm, a candidate Median Center is found (Xt, Yt) and refined until it represents the location that minimizes Euclidian Distance d to all features (i) in the dataset.

Central feature is the shortest total distance from all the other features in the study area.

If median center is very different than the mean center, this means data are either dispersed or clustered. If you know the mean and median centers of your data, you should compare these with the directional distribution of your data. Standard distance helps to compare two ore more distributions. It is the measure of degree about features. Standard deviation is defined as the distribution or the variation of values from the mean value.

Directional distribution summarizes the spatial characteristics of features; central tendency, dispersion, and directional trends. Directional distribution measures how concentrated features are around the geographic mean by using standard deviational ellipse. This technique calculates the standard deviation of x-coordinates and y-coordinates from the mean center.

SDE creates a new feature class. This feature class contains an elliptical polygon. This polygon includes two standard distance which are short and long axes. The standard deviational ellipse is formulated as:

$$SDE_x = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}{n}} \qquad SDE_y = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \overline{Y}\right)^2}{n}} \tag{2.3}$$

$\overline{X}$ = Mean Center of X

$\overline{Y}$ = Mean Center of Y

n = Total number of features

The angle of rotation calculated as:

$$\tan\theta = \frac{A+B}{C} \qquad \textbf{(2.4)}$$

$$\text{A} = \left(\sum_{i=1}^{n} \tilde{x}_i^{\,2} - \sum_{i=1}^{n} \tilde{y}_i^{\,2}\right) \qquad \textbf{(2.5)}$$

$$\text{B} = \sqrt{\left(\sum_{i=1}^{n} \tilde{x}_i^{\,2} - \sum_{i=1}^{n} \tilde{y}_i^{\,2}\right) + 4\left(\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i\right)^2} \qquad \textbf{(2.6)}$$

$$\text{C} = 2\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i \qquad \textbf{(2.7)}$$

$\tilde{x}_i$ = Deviation of the x-coordinate from the mean center

$\tilde{y}_i$ = Deviation of the y-coordinate from the mean center

The standard deviations for the x-axis and y-axis are:

$$\sigma_x = \sqrt{2}\sqrt{\frac{\sum_{i=1}^{n} (\tilde{x}_i \cos\theta - \tilde{y}_i \sin\theta)^2}{n}} \qquad \textbf{(2.8)}$$

$$\sigma_y = \sqrt{2}\sqrt{\frac{\sum_{i=1}^{n} (\tilde{x}_i \sin\theta - \tilde{y}_i \cos\theta)^2}{n}} \qquad \textbf{(2.9)}$$

One standard deviation ellipse polygon will cover approximately 68% of features in the cluster. Two standard deviation ellipse polygon will cover approximately 95% of features and three standard deviations will cover approximately 99 percent of the features in the cluster. For example, as seen on Figure 2.3, after analyzing data spatially, mean center, median center, and directional distribution were determined. After mean and median centers of your data are determined, the directional distribution of the data can be calculated. This explains the general orientation of the data based on the rotation of the ellipse. The directional distribution explains the spread of the data based on the standard deviation. When generally outside of ellipse, this data is relatively dispersed.

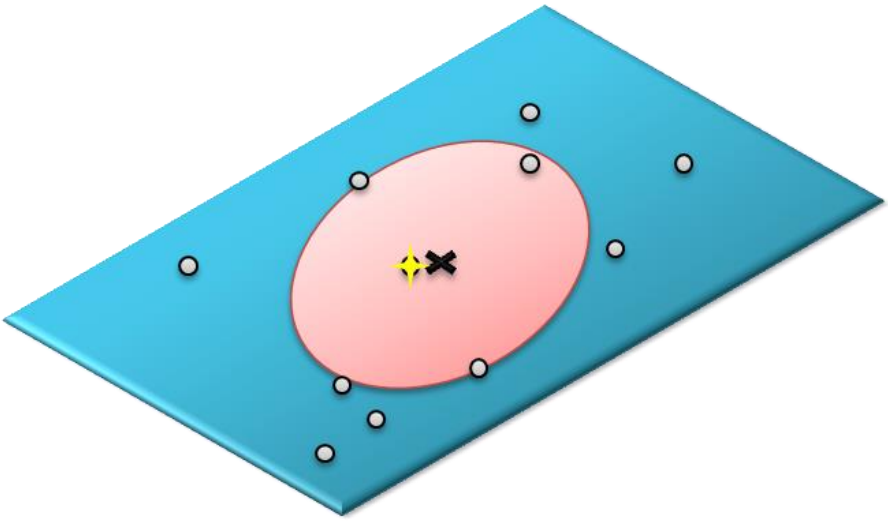


Figure 2.3: Types of measuring techniques (Mean, Median,Directional Distribution).

2.2.2 Identifying Patterns

Quadrat Analysis, Nearest neighbor index, K-function, Join count statistics, Geary's c* Moran's I, and General G-statistic tools are used to identify patterns.

Quadrat analysis provides overlay in areas of equal size and can be used when there are multiple features at a single location. It doesn't consider the distance between the features. Results depend on the size of the quadrats.

Nearest neighbor index is calculating the average distance between features and considers the distance between features. Results may be biased if there are many features near edge of study area. Average nearest neighbor tool is important for comparing various features in study area. It is the ratio of Observed Mean Distance to the Expected Mean Distance. If the nearest neighbor index is less than one, the pattern is clustering; if the index is grater than one, the movement is dispersed.

The average Nearest Neighbor ratio is given in Formula 2.10,11, and 12:

$$ANN = \frac{\overline{D}_O}{\overline{D}_E} \qquad \textbf{(2.10)}$$

$\overline{D}_O$ = Observed Mean Distance , $\overline{D}_E$ = Expected Mean Distance

$$\overline{D}_O = \frac{\sum_{i=1}^{n} d_i}{n} \qquad \textbf{(2.11)}$$

$$\overline{D}_E = \frac{0.5}{\sqrt{n/A}} \qquad \textbf{(2.12)}$$

n = Total number of features

A = Area value

K-function is counting the number of features within defined distances and uses multiple simulations to create a random distribution envelope. This function calculates the concentration of features at a range of scales or distances simultaneously. Patterns are suspect at larger distances due to edge effects. K-function could be used for point pattern. This function shows how spatial clustering or dispersion of the feature centroids changes when the neighborhood size changes (Formula 2.13).

The K-function given as:

$$\text{L(d)} = \sqrt{\frac{A\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} k_{i,j}}{\pi . n(n-1)}} \qquad \textbf{(2.13)}$$

d = Distance

n = Total number of features

A = Total area of the features

$k_{i,j}$ = Weight

If no edge connection, then weight is equal to one. If distance between i and j is less than d, then weight is equal to zero.

There are two different definition of K-function results. One of them is to define unweigthed K-function result. If observed K-value is larger than the expected K-value, the distribution will be clustered. The second result is weighted K-function result. Weighted K-function results are more clustered than unweight K-function.

Join count statistics is helpful for whether values are clustered or dispersed and straightforward way to identify patterns for areas. It can only applied to categorical (nominal) data.

Geary's c* Moran's I is the similarity of nearby features and provides a single statistics summarizing the pattern. It doesn't indicate if clustering is for high values or low values. Geary's c and Moran's I use the magnitude of feature values to identify and measure the strength of spatial patterns.

The two statistics, Geary's c and Moran's I, identifies similar values and have local versions;

- Geary's* compares the values of neighboring features by calculating the difference between them.
- Moran's I compares each value in the pair to the mean value for all the features in the study area.

General G-statistic is the concentration of high or low values and indicates whether high or low values are clustered. It works the best when either high or low values are clustered (Mitchell, 2005).

The General G statistics (Formula 2.14):

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} x_i x_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j}, \forall_j \neq i \qquad \textbf{(2.14)}$$

x_i = Attribute value for i

x_j =Attribute value for j

$w_{i,j}$ = Spatial weight between features i and j

n = Number of features

Social scientist studying about the integration of different ethnic groups in a region could measure the extent to which the groups are clustered or dispersed, using the percentage of the population of each group in each census tract or block group.

For each feature, the statistics compares neighboring features within a distance that you specify. The statistic indicates the extent to which each feature is surrounded by similarly high or low values.

There are two versions of this statistic; Gi is dispersion of a particular phenomenon from the target feature to the surrounding area over time.

- If you are interested in finding hot spots or cold spots, you would use Gi*. Gi* uses a neighborhood based either on adjacent features or on a set distance.
- Gi is dispersion of a particular phenomenon from the target feature to the

surronding area over time.

- If you are interested in finding hot spots or cold spots, you would use Gi*. Gi* uses a neighborhood based either on adjacent features or on a set distance.

2.2.3 Distribution of Pattern with Histogram, Voronoi Map, and QQ Plot

Spatial autocorrelation is a measure of the degree to which a set of features and their associated data values tend to be clustered together in space (positive spatial autocorrelation) or dispersed (negative spatial autocorrelation). Semivariogram helps to determine the spatial autocorrelation. If the data separate apart spatially, then the difference between the values will increase. The Figure 2.4 shows a semi-variogram. As defined in the Formula 2.15;

$$\text{Semivar(distance h)} = 0.5 \text{ x average } [\ (\text{location i} – \text{location j})^{2}\ ] \qquad \mathbf{(2.15)}$$

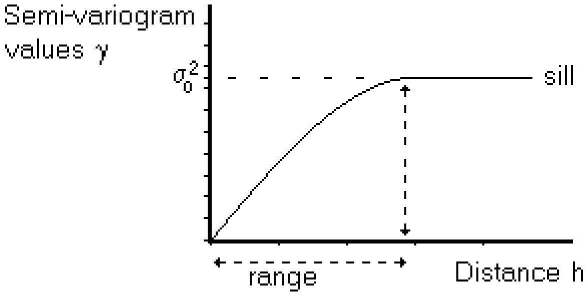


Figure 2.4: A typical semi-variogram.

The frequency is important to understand the distribution of data values. It is easy to explore the frequency distribution of data sets by using histogram. A histogram, using bar and bin, is a graphical representation of the data. Histogram shows the distribution of the data and gives information about data. Each bar represents an interval within the range of the data. The height of the bar represents the number of data values that fall within that interval. It also helps to explore the outliers by looking left or right side of the graph mostly.

As seen on the example of Figure 2.5, the shape of the left histogram matches the bell-shaped (Gaussian) curve of normally distributed data. The right histogram has a visual cue that some data is much lower in value than the rest of the data. This indicates a possible outlier in the data.

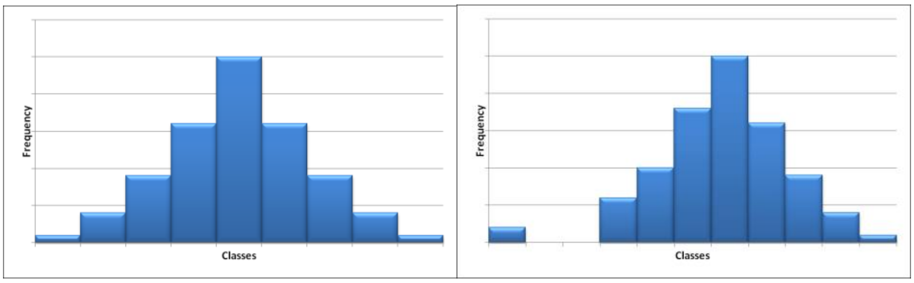


Figure 2.5: Example histograms presenting data distribution.

Voronoi maps are mostly used to find local outliers or check different categories of values. Voronoi maps help to understand and to examine the data based on variation. Voronoi maps demonstrate the polygon which is neighborhood of each point of the data. Almost it is used to find near neighbor for features. These are simple, mean, mode, cluster and entropy methods to calculate polygons. The Figure 2.6 shows that polygon which is neighborhood of each point. Defined on Formula 2.16;

$$\text{Entropy} = - \sum \left(p_i * Logp_i\right) \quad \textbf{(2.16)}$$

p_i = proportion of polygons that are assigned to each class

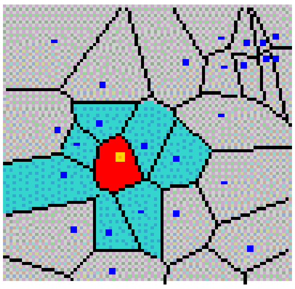


Figure 2.6: Polygons and their neighborhood.

In this thesis, voronoi maps are produced using mean and cluster methods. The cluster method helps to identify the difference classification of data. If the polygons are different from each neighbor, then the polygon color will be gray. So, different types of polygon could be understand easily.

On the example of Figure 2.7, This semivariogram cloud shows several outliers. Cluster Voronoi maps show spatial outliers in your data, simple Voronoi maps can pinpoint data values that are many class breaks removed from surrounding polygons. The gray polygons are potential outliers in this cluster Voronoi map. Both of the potential outliers selected in the semivariogram cloud have gray fill. The polygons that are light and dark green represent little local variation, the polygons that are orange and red represent greater local variation.

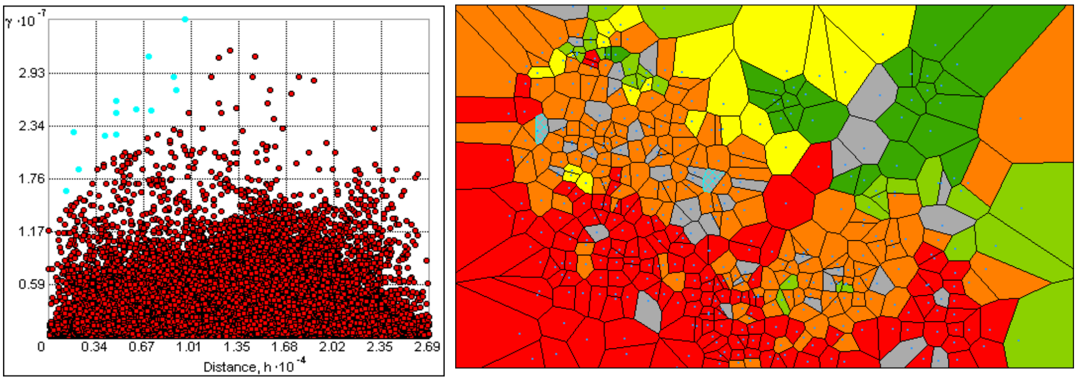


Figure 2.7: A example Semivariogram Cloud and Voronoi Map.

Normal QQ plot and General QQ plot check the relativity of the data by combining distribution graph and data graph. Normal QQ plot use cumulative distribution graph versus normal distribution graph. It has the same steps for general QQ plot but instead of using normal distribution as different data set. Figure 2.8 gives an example of Normal and General QQ plot. General QQ plot is used to understand similarity of distribution of the two different the data set.

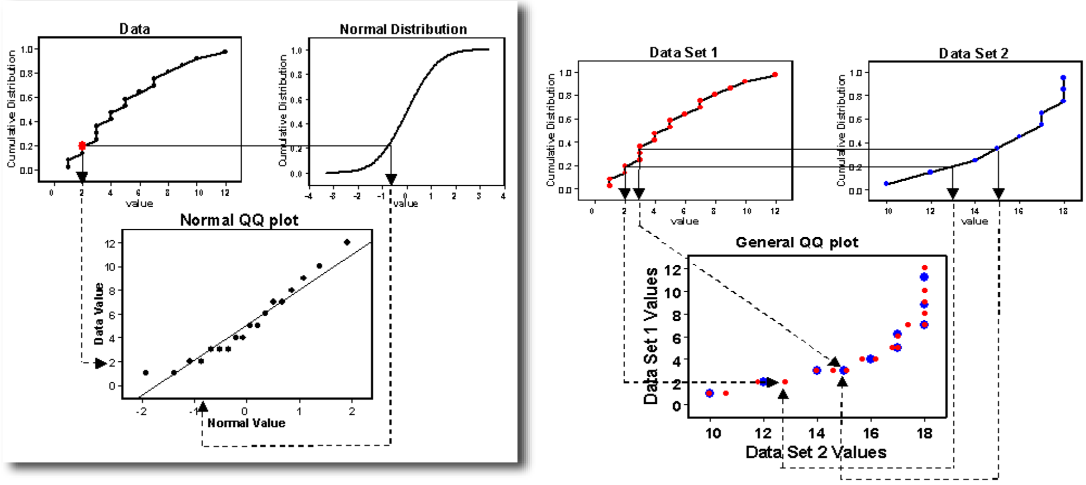


Figure 2.8: An example of Normal QQ plot and General QQ plot.

In the following Formula 2.17, the cumulative distribution values are calculated as;

$$\frac{(i-0.5)}{n} \tag{2.17}$$

for i=value, n= total value. The same equation is used for normal distribution (using $\mu = 0, \sigma = 1$). This two distribution graph have been created Normal QQ-Plot. Furthermore, general QQPlot have been created same way but the second data set could be any set (Mitchell, 2005).

2.2.4 Identifying Cluster

Hot-Spot Analysis identifies statistically significance for high and low values. Getis-Ord Gi* statistics are used. A feature with a high value is interesting, but may not be a statistically significant hot spot. Statistically significant means that the features have a high values and be enclosed by other features with high values as well.

The z-score and p-value explore the values high or low (Figure 2.9). Z score tells you where features with either high or low values cluster spatially. If z-score is positive, then it has hot-spot (high) values. If z-score is negative, then it has cold-spot (low) values. In addition, the p-values are numerical approximations of the area under the curve for a known distribution, limited by the test statistic.

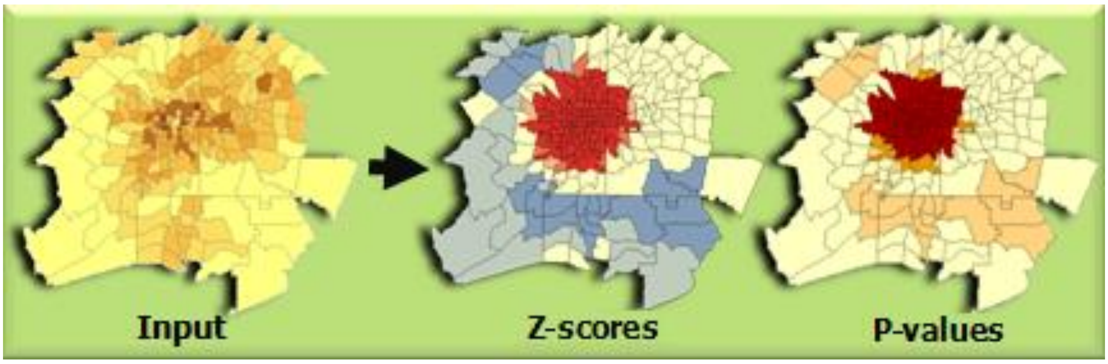


Figure 2.9: An Example for Hot Spot Analysis.

Calculations are defined on Formula 2.18;

The Getis-Ord local statistics is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad \textbf{(2.18)}$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features (2.19 and 2.20).

$$\overline{X} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad \textbf{(2.19)}$$

$$\text{S}=\sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\overline{X}\right)^2} \qquad \textbf{(2.20)}$$

The G_i^* statistic is a z-score so no future calculations are required (Mitchell, 2005). Besides, Nearest Neighbor Hierarchical Clustering, one method for finding clusters of discrete features, is to specify the distance that features can be from each other in order to be part of a cluster, and the minimum number of features that make up a cluster.

2.2.5 Analyzing Geographic Relationships

Pearson's correlation coefficient is the ratio of the joint variation of two variables to the total variation of entire dataset. The value of the correlation coefficient (r) ranges from 1 (indicating a perfect direct relationship) to -1 (a perfect inverse relationship).

Spearman's rank correlation coefficient measures the extent to which two lists of ranked values correspond. The coefficient is based on the difference in rank between each feature for the two variables.

One type of regression analysis, known as linear regression, is a common approach for building simple models to analyze geographic processes. A pair of values for each feature can be plotted as a data point on a chart. Linear regression is to find best fit of a line between the data points on the chart that represent the relationships (Mitchel, 2005).

Regression Equation is defined on Formula 2.21 below:

$$\text{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \qquad \textbf{(2.21)}$$

y = Dependent variable

$\beta =$ Regression coefficients

x = Explanatory variables

$\varepsilon =$ Random Error/ Residuals

In this study regression analysis will be used to figure out characterizing social media users based on demographic issues. Regression analysis allows you to model, examine, and explore spatial relationships, and can help explain the factors behind observed spatial patterns. Regression analysis is also used for prediction. You may want to understand why people are persistently dying young in certain regions, for example, or may want to predict rainfall where there are no rain gauges. There are few studies which attempt to analyze for social media user's relations between spatial data.

Some patterns can help to determine the location of hot spots but some specific statistical techniques find an answer about data by means of regression analysis. During the nineteenth-century in England, Francis Galton originally derived and applied particular statistical technique which is linear regression (Allen, 1997).

According to ESRI (2009), ArcGIS provides a suite of statistical analysis tools called regression that can help determine the causes of these patterns as well as the power to predict future patterns based on current data and trends. Statistical tools are required to analyze geographic relationships.

One common method is called Ordinary Least Square (OLS) which minimizes the squared distance from points to the line and performs global Ordinary Least Squares linear regression to generate predictions or to model a dependent variable in terms of its relationships to a set of explanatory variables (Mitchel, 2005). OLS creates a regression equation to represent the process also compute a coefficient for each variable. There are six checks in finding good model. ESRI (2013) interpreting these six checks respectively.

- First of all, Koenker statistic examining independent variables in the model have a constant state to the dependent variable.
- The secondly, the sign of the coefficient points the positive or negative relations and also shows the problem if there is a unexpected sign of the coefficient.
- Thirdly, multicollinearity occurs when the explanatory variables are highly intercorrelated. Multicollinearity causes redundancy and it prevents to set of reliable model. The best exercise is to remove these variables with Variable Inflation Factor (VIF) over 7.5 and run OLS again.
- Fourth check is that the If the p-value for the Jarque-Bera statistic (test) is statistically significant, your model is biased and the model predictions cannot be fully trusted. Jarque-Bera statistic test defined as ”whether the residuals (the observed/known dependent variable values minus the predicted/estimated

values) are normally distributed with a mean of zero”.

- The other check is to understand the problem which is missing independent variables by running spatial autocorrelation tool. It is the common problem of OLS regression.
- The last check is looking Adjusted R-Squared value how much the dependent variable is being explained by the explanatory variables. Adjusted R-Squared value should be increased. Another important value is to measure model performance, Aikake's Information Criterion (AIC). The lower AIC value provides better model (ESRI, 2013).

One approach for incorporating regional variation in your model is known as Geographically Weighted Regression (GWR). GWR allows model coefficients to vary regionally. Essentially, you run a regression for each location, rather than for the study area as a whole. GWR model creates an equation for each feature in the dataset. GWR constructs these equations by combining the dependent and explanatory variables of features. The shape and size of the bandwidth depend on user input for Distance, Number of neighbors parameters, the Kernel type, and Bandwidth method.

In the examples of Figure 2.10, Regression Analysis explains why something is happening. This analysis models, examines, and explores spatial relationships, and predicts. Left analysis explains coefficients for percent rural and low-weights births. Right analysis gives T-scores where this relationship is significant.

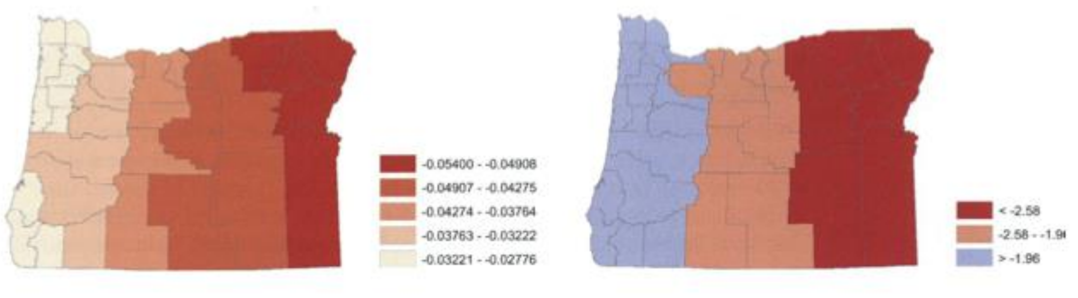


Figure 2.10: An example for Regression Analysis.

2.3 Social Media

Social media has came into our lives and shows very rapid development at the beginning of the 21_{th} century. It brings new dimension to communication between people. New technologies results in revolution in information sharing. Classic media users just listened the news from radio or television but they do not criticize. The most important use of social media is to create and share information and ideas in the electronic networks, with Web 2.0 technologies.

At the same time social media has created some questions in mind. What others think about our character? Do they like your ideas or shares? From this point of view, the new study subject was born for researchers; determining the characteristics of users, categorizing based on socio-demographic data for regional spatial analysis, and the implementation of the users on real-time location, etc.

There are six types of social media. These are social networks, bookmarking sites, social news, media sharing, microblogging, and blog comments, and forums.

Social Networks allows the users to interact with other people of similar interests. These users have a profile and ability to different ways to connect with others. The most popular are Facebook and LinkedIn. Online dating sites are also given an example of social networks.

For this thesis research, Table 2.4 gives examples about online dating websites.

Table 2.1: Some online dating websites and their types.

Site Name	Target area
Tastebuds.fm:	Free dating site which matches people based on their tastes in music.
Perfectmatch.com	This dating website is using complex heuristic algorithms for finding matches.
OkCupid	It uses answers from user-generated questions to find matches that conform to a user's stated preference.
Match.com	One of the biggest online singles meeting sites.
Lavalife	One of the oldest dating companies in the world.

BeautifulPeople.com	Dating website where membership is based on a vote, in which existing members rate how attractive they perceive prospective members to be.

Bookmarking sites allow users to organize, keep safe and manage links to online resources. The most popular are Delicious and StumbleUpon.

Social news websites allow user to post news items or stories that are ranked based on popularity. The most popular is Digg.

Media Sharing sites give user a permission to upload and share various media such as video or pictures. The most popular are Facebook and Flickr.

Microblogging sites allow to user short elements of contents such as short sentence, individual images, or links. The most popular is Twitter.

Blog Comments and Forums allow users to discuss about the topic and keep from conservation by messages. There are many popular blogs and forums.

Furthermore, location-based social media can be grouped into three major categories:

- Check-in sites (Foursquare, Gowalla)
- Social Review sites(Yelp, TellmeWhere, Groupon)
- Social scheduling sites(Loopt, Meetup, Eventful)

3 APPLICATIONS

3.1 Selecting Study Area

Social media platforms have data about people distributed all over the world. In Turkey, there are different kinds of social media and communication platforms that are especially global social media sites such as Facebook, Twitter, and Foursquare. Besides, various social media and commercial platforms have been built and on available on different discipline. An online dating site platform was selected as an example of social media site covering Turkey. Detailed analysis is focused on Istanbul metropolitan area (Figure 3.1).

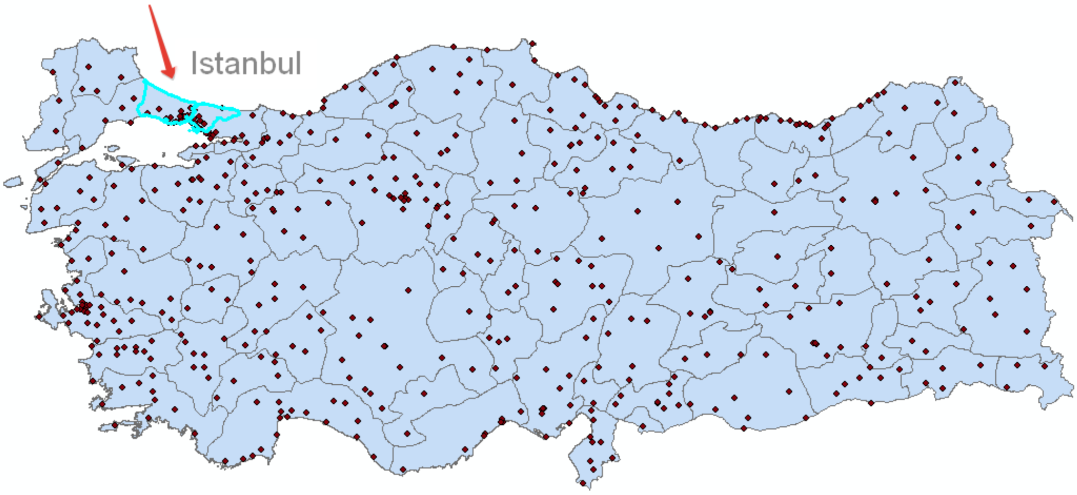


Figure 3.1: Online dating site example platform in Turkey.

According to results of Address Based Population Registration System (2013) “The most populated province was Istanbul with 18.5% (14.160.467 persons) of total population in 2013”. It demonstrates the number of population living in Istanbul increases day by day.

3.2 Data gathering and preparation

In this study, data about social media users was taken from database of social media site. This platform has the data about 2.2 million users. These data was retrieved from the MySQL database including user information as age, gender, income, occupation at county/district levels. To analyze user profiles of this site, the data was eliminated after determining active users that enter the platform twice a week at least.

The data about 9836 users was organized on ArcGIS geodatabase environment with location information. As seen on Figure 3.2, Istanbul metropolitan areas have much more users than other cities.

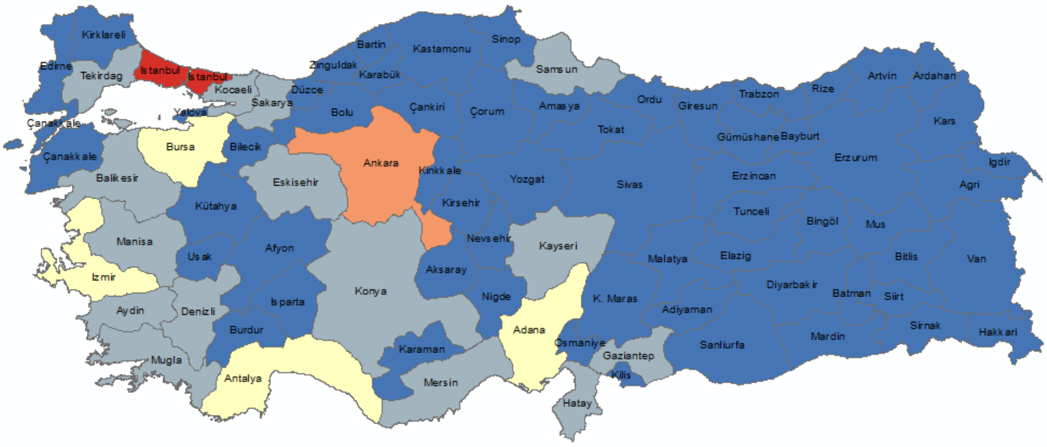


Figure 3.2: Thematic map for social media users in Turkey.

At the beginning of the work, Turkey statistics information will be analyzed at the province level. The data allows to me provide overall perspective in this research. If we just look at the average statistics of the Turkey as a whole, we will not be able to understand that which places have higher incomes, education level, etc. than others. ArcGIS tools will be usable in Istanbul as regional area because it makes analysis easier.

The data was organized and eliminated based on Istanbul districts. 3401 of 9836 users are located in Istanbul Metropolitan area. As seen on Figure 3.3, social media users were situated in the districts with more population. Data collection procedure is not only for social media data but also for statistics data according to Turkish Statistical Institute. This data was taken from an address-based information system and includes demographic information about population, gender, age, and income. County and district data was collected from Istanbul Metropolitan Municipality web service to analyze these data spatially.

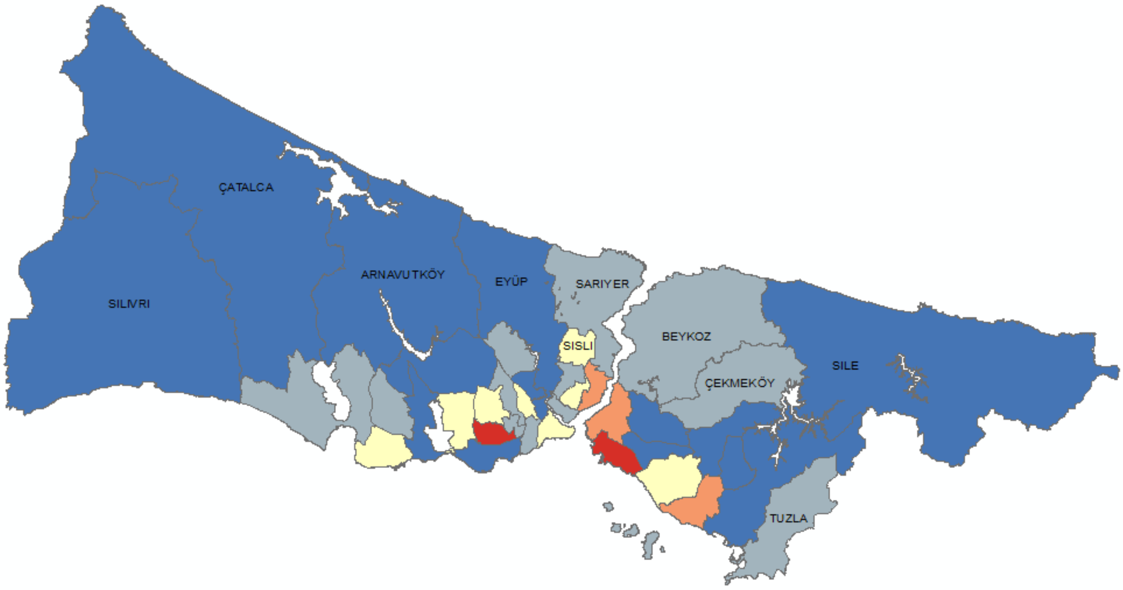


Figure 3.3: Thematic map for social media users in Istanbul.

3.3 Examining Thematic Data

Thematic analysis aims to create meaningful patterns for this study. These patterns represent the main idea of the data more easily rather than reading many transcripts of data. In this study, thematic analysis is used for various reasons. The idea behind that people living in the same region may show similar characteristics. In addition, TurkStat and social media data were analyzed to examine thematic information.

Demographic analysis is a method to build an understanding of the age, sex, and racial property of a population and how it affects through the demographic processes of birth, death, and migration. According to Karlsdotter et al (2012), the individual independent variables correspond to sociodemographic (age, sex, marital status, and nationality) and socioeconomic (educational level, personal income, and unemployment) dimensions. It is important to note that this part informs about sociodemographic and socioeconomic situation in Istanbul.

At the beginning, the number of gender information was defined at district level (Table 3.1). This analysis helps to identify the differences and the distribution of two sets by using thematic map of Istanbul Metropolitan Municipality. As seen on the table, 59% of social media users are male and 41% of which are female while the population percentage of female is 49.8% and the percentage of male is 50.2% in Istanbul.

Thematic mapping helps to classify the data with quantities, color ramps, or charts based on features. It is easy to display the data on the map. Various thematic maps were produced in this study in terms of demographic analysis.

Thematic map on Figure 3.4 shows density of social media users depending on population density. It is understood that people are living in Kadıkoy, Bahcelievler, and Besiktas use this social media site mostly. Urban areas have more user density than rural areas.

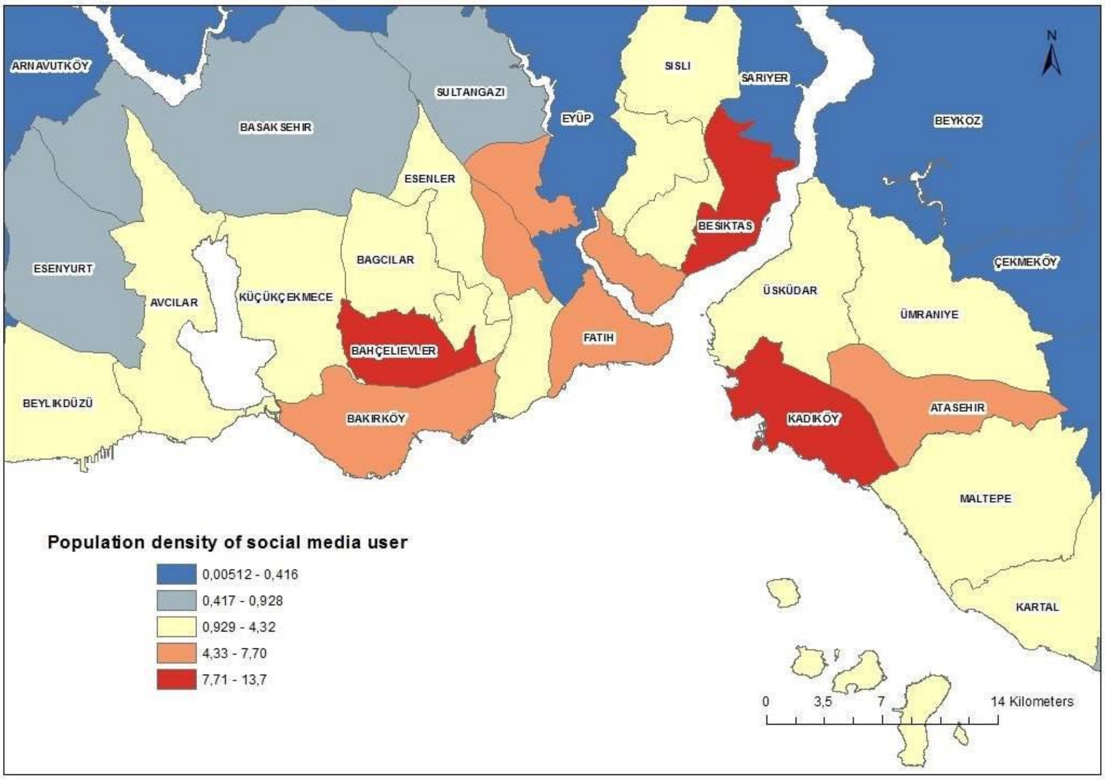


Figure 3.4: Population density of social media user, 2013.

Table 3.1: Population of User (2013).

District	Man	Woman	Total
ÜSKÜDAR	83	54	137
ÜMRANIYE	64	52	116
ÇEKMEKÖY	22	19	41
ZEYTINBURNU	21	12	33
TUZLA	32	14	46
SILIVRI	16	8	24
SULTANGAZI	15	9	24
KÜÇÜKÇEKMECE	70	38	108
GÜNGÖREN	16	15	31
GAZIOSMANPASA	48	35	83
FATIH	61	40	101
EYÜP	44	26	70
ESENYURT	23	17	40
ESENLER	30	26	56
BÜYÜKÇEKMECE	18	24	42
BEYLIKDÜZÜ	74	55	129
BEYKOZ	31	28	59
BASAKSEHIR	30	27	57
BAGCILAR	51	28	79
BAYRAMPASA	35	33	68
BAKIRKÖY	104	93	197
BAHÇELIEVLER	123	73	196
AVCILAR	84	64	148
ARNAVUTKÖY	9	15	24
SILE	4	0	4
ADALAR	21	25	46
ÇATALCA	1	7	8
PENDIK	80	42	122
SULTANBEYLI	7	5	12
KAGITHANE	33	20	53
MALTEPE	80	56	136
KARTAL	81	32	113
SANCAKTEPE	10	8	18
ATASEHIR	99	80	179
KADIKÖY	187	157	344
BESIKTAS	110	74	184
SARIYER	41	17	58
SISLI	94	52	146
BEYOGLU	47	22	69
Total	1999	1402	3401

It is supposed that gender type of this site is all single. The histogram on Figure 3.5 shows single categories. The number of single users is more than divorced users. Single and divorced users especially visit the site. This data is valid for Istanbul users, not for all the rest.

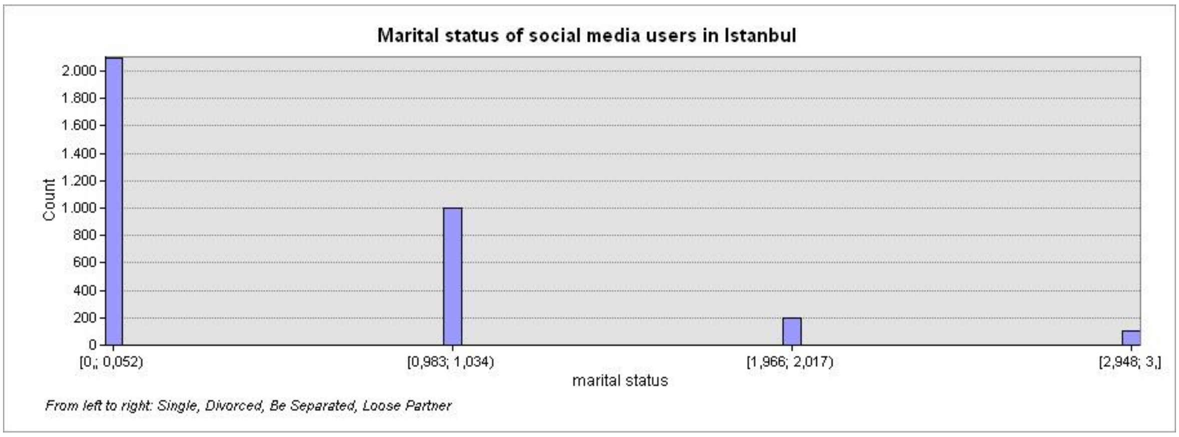


Figure 3.5: Histogram of marital status (social media users)

Therefore, population and the users were compared in view of singles. Figure 3.6 shows single density of population. Some counties like Kadikoy, Besiktas, ve Bakirkoy have the highest user density. Although single population density is the highest in Bagcilar and neighbouring counties, single social population density is low.

The ratio of single users to single population were examined based on gender information (Figure 3.7 and 3.8). According to woman ratio, these districts have more ratio than the rest: Adalar, Kadıkoy, Bakırkoy, Besiktas, and Beylikduzu. Furthermore, these district have less ration than the rest: Esenyurt, Kucukcekmece, Bagcilar, Gungoren, Zeytinburnu, Sultangazi, Sariyer, Sancaktepe, Sultanbeyli, Sile, Kagithane, and Sancaktepe.

According to man ratio, these districts have more ratio than the rest: Adalar, Kadıkoy, Bakırkoy, Besiktas, Beylikduzu. Furthermore, these district have less ration than rest: Esenler, Esenyurt, Arnavutkoy, Catalca, Bagcilar, Gungoren, Zeytinburnu, Sultangazi, Sariyer, Kagithane.

The both gender seems to the same ratio at the same place. If social structure is similar based on location, there would be the same ratio in terms of gender (Figure 3.9).

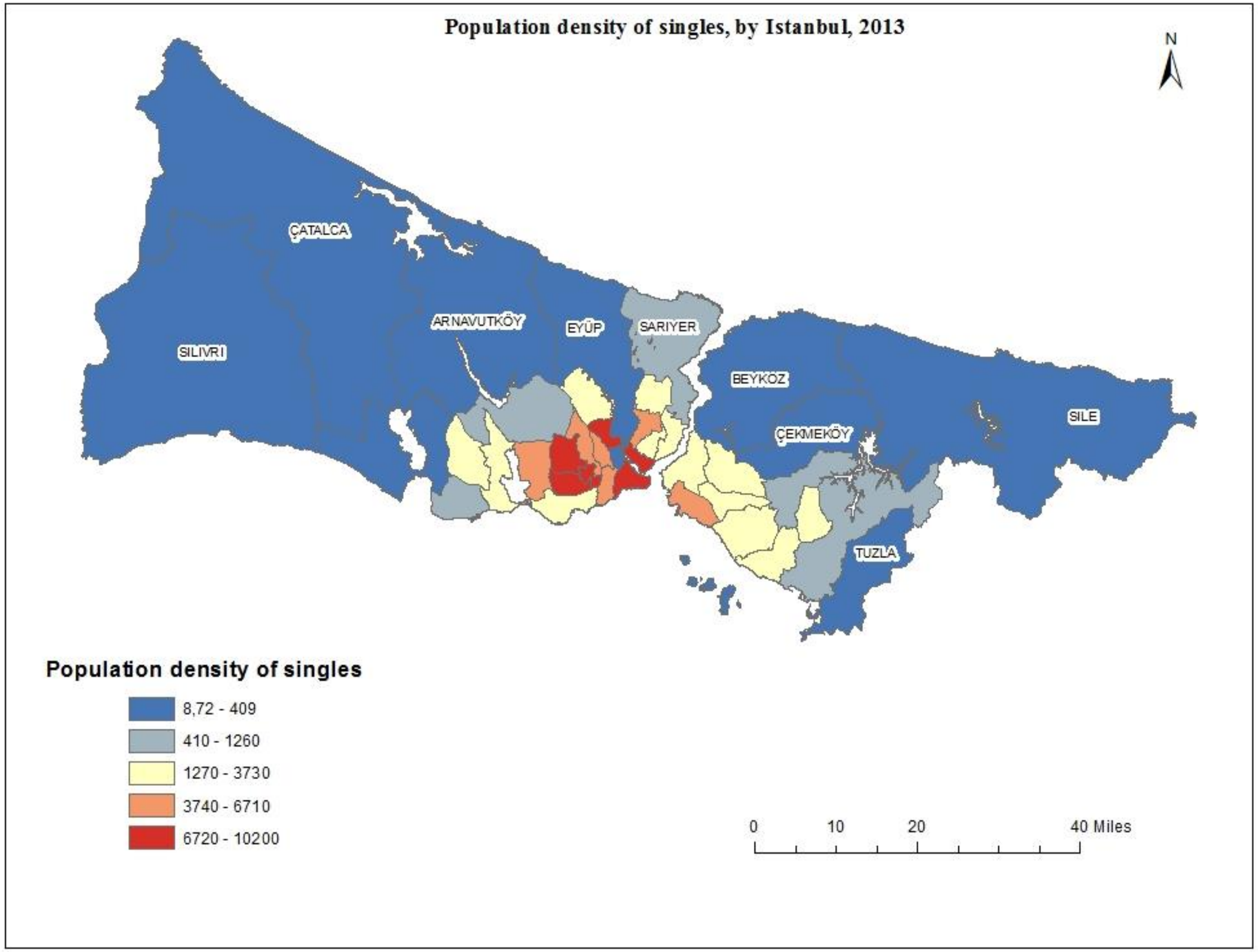


Figure 3.6: Population density of singles, TurkStat 2013.

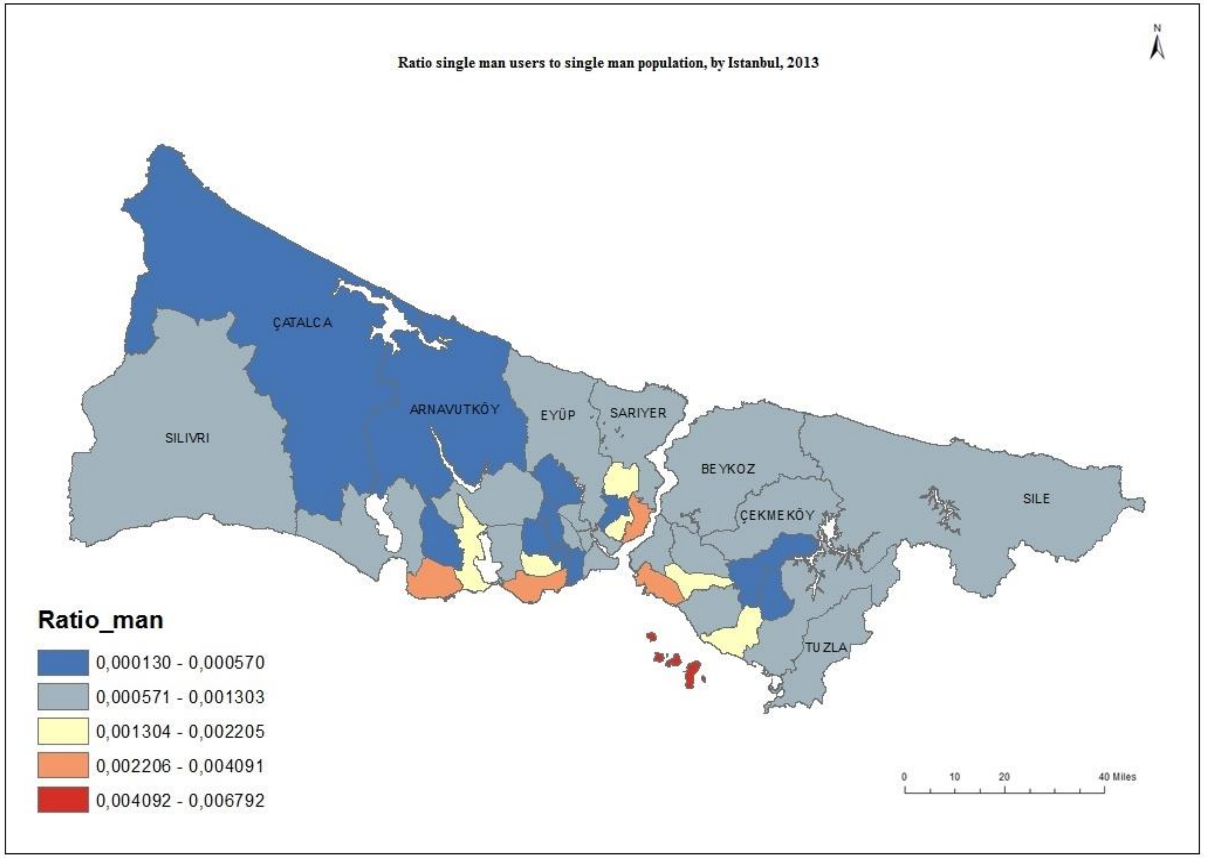


Figure 3.7: Ratio single man users to single man population, 2013.

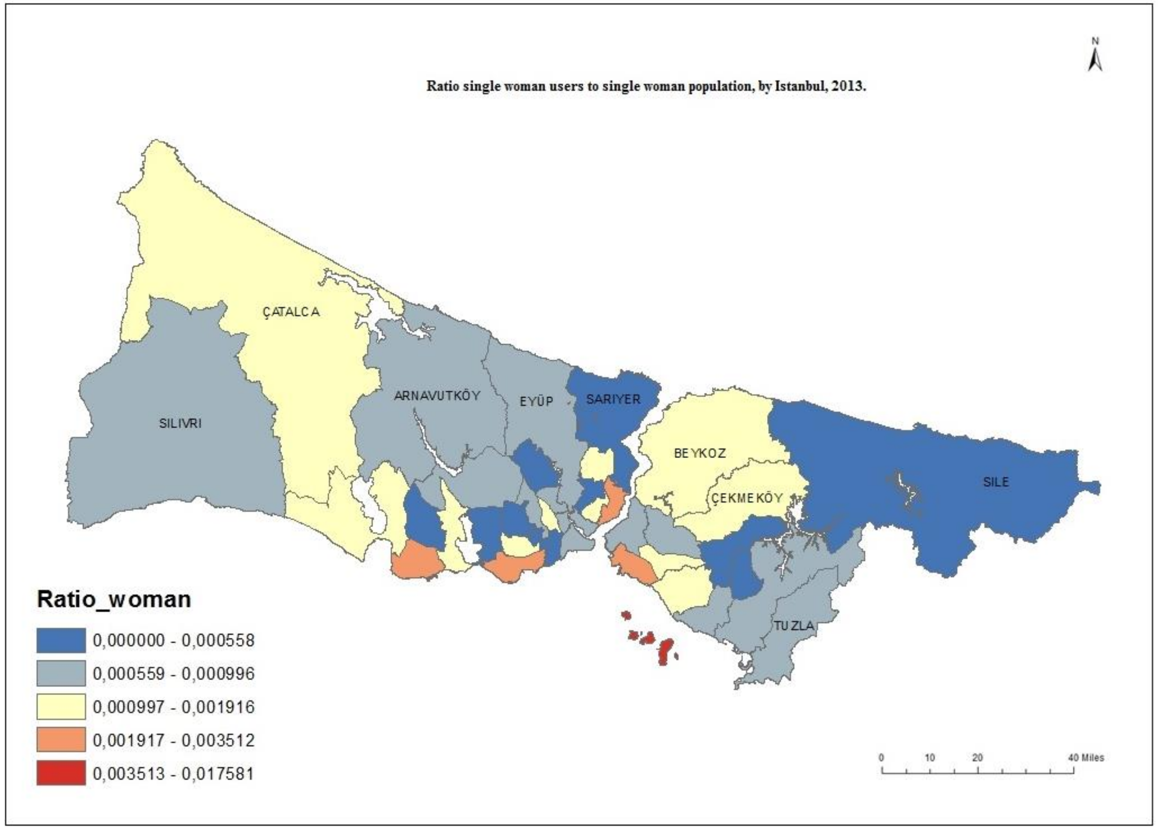


Figure 3.8: Ratio single woman users to single woman population, 2013.

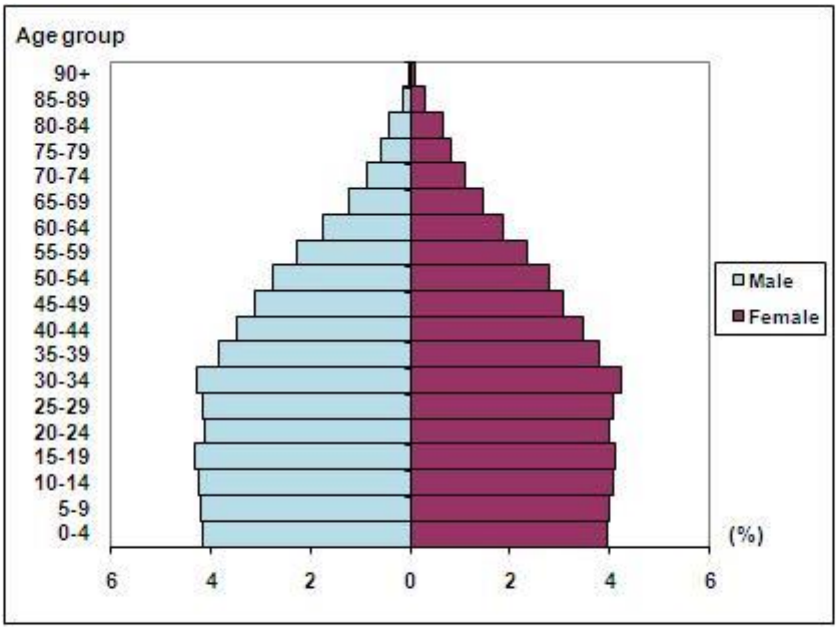
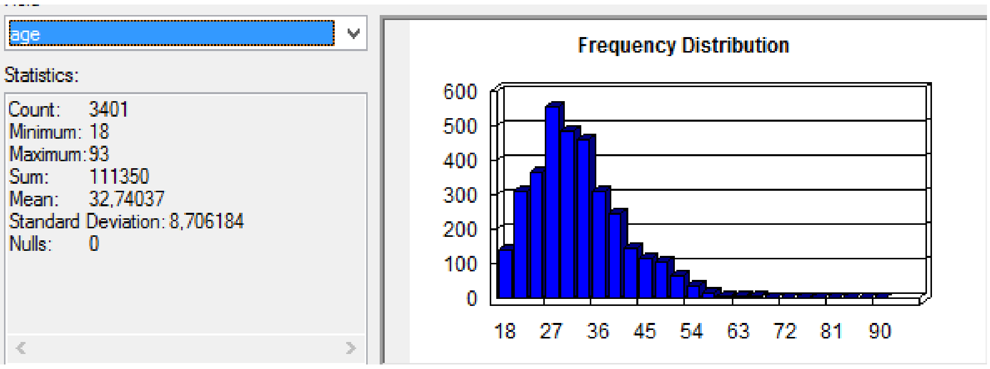


Figure 3.9: Histogram of age (Turkey, 2013).

Population density of median age group is examined in Figure 3.11 following Figure 3.10 that gives an information about minimum and maximum age of the social media users. Adalar has higher median age population density than other counties. If the median age is high in a county, then the number of users is higher than other counties. Although the median age is less in some counties such as Bagcilar and Esenler, the number of the users is still high. The reason for this should be found in the relationships between age and income.



Statistics:

```
Count:    3401
Minimum:  18
Maximum:  93
Sum:      111350
Mean:     32,74037
Standard Deviation: 8,706184
Nulls:    0
```

Figure 3.10: Histogram of user age (female and male)

The ratio of population at working age (15-64) group was increased from 67.6% in 2012 to 67.7% in 2013 (51.926.356 persons) (ABPRS, 2013). The ratio of crude marriage is 6.14% in Tunceli the lowest and 10.12% in Adiyaman the highest. The crude marriage ratio of Istanbul is 7.94%.

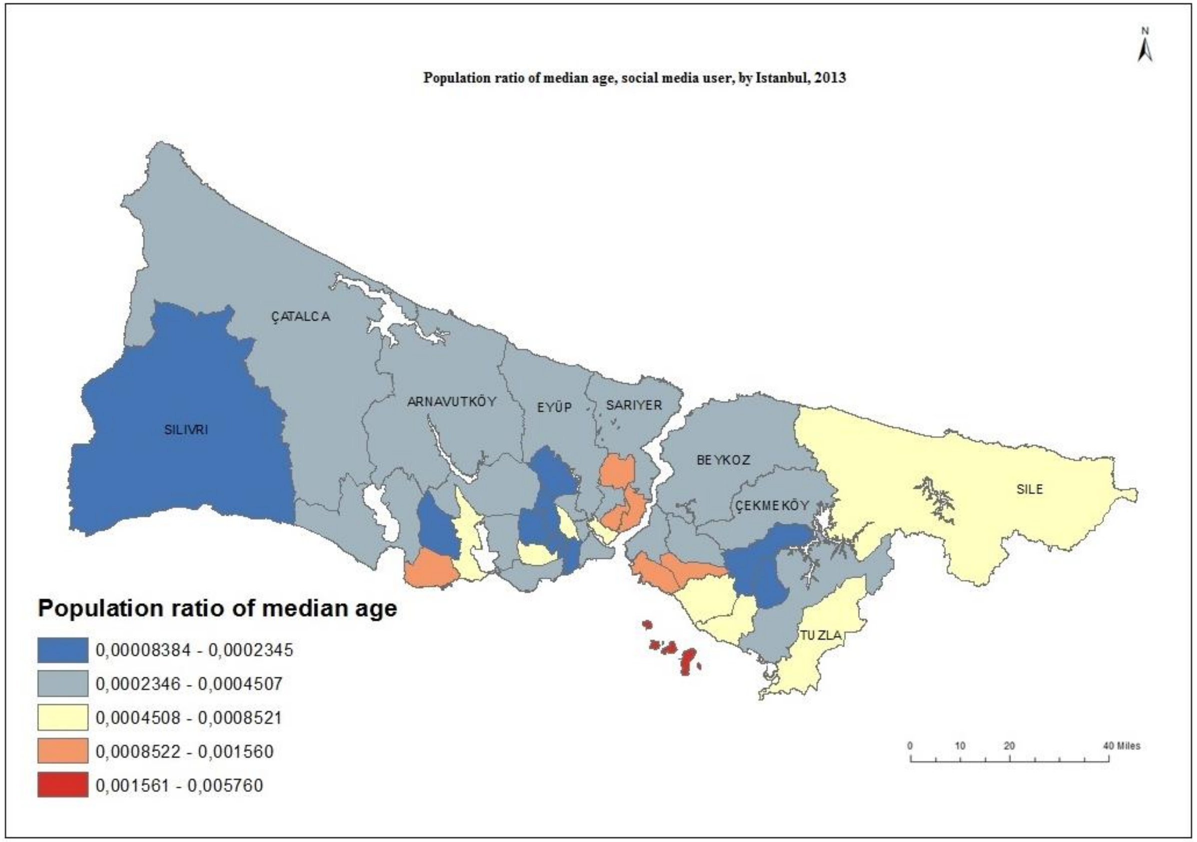


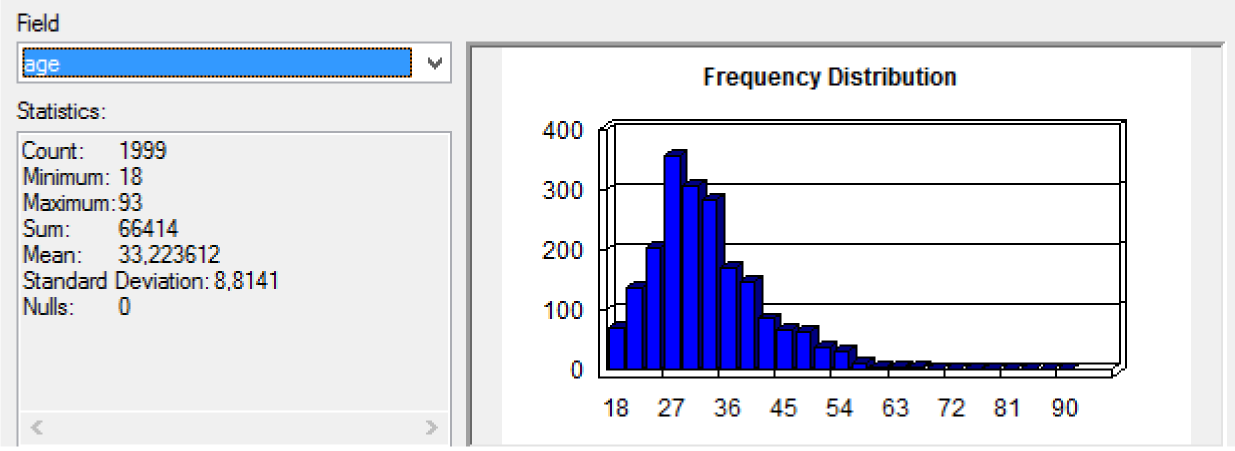
Figure 3.11: Population ratio of median age, social media user, 2013.

As seen on Table 3.2, in terms of gender, average marriage age and average first marriage age of Istanbul were given from 2008 to 2012. According to the table, the average marriage age and the average first marriage age have been raising rapidly.

Table 3.2: Average marriage age and average first marriage age by gender, 2008-2012 (years) in Istanbul.

	Average marriage age		Average first marriage age	
Year	Groom	Bride	Groom	Bride
2008	28,6	24,9	27,0	23,9
2009	29,0	25,3	27,2	24,1
2010	29,1	25,5	27,4	24,3
2011	29,4	25,8	27,5	24,5
2012	29,5	26,0	27,6	24,6

In this study, males have the highest population rate between ages 24 to 36. Also, genders are not distributed equally. The Figure 3.12 shows the age distribution of male users who are using online dating site in Istanbul. The horizontal axis on the graph shows the ages, while the vertical axis shows the frequency.



Field

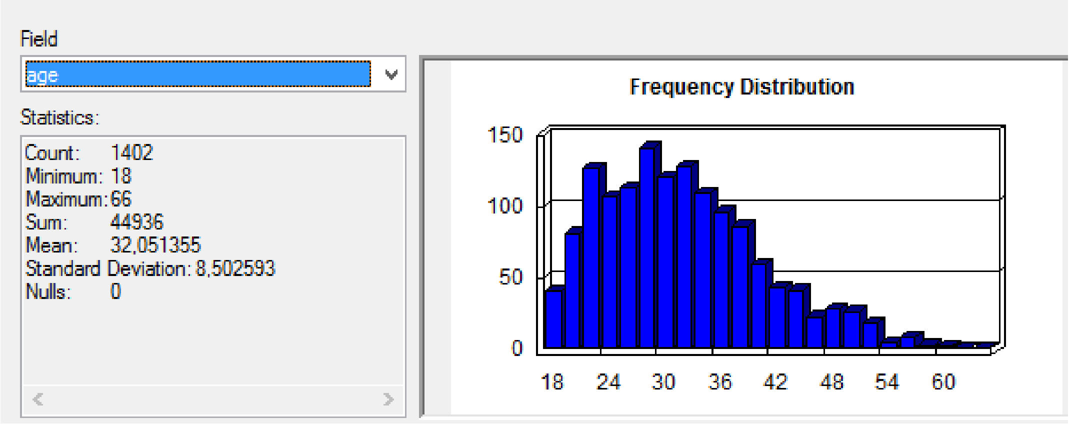
age

Statistics:

Count: 1999
Minimum: 18
Maximum: 93
Sum: 66414
Mean: 33,223612
Standard Deviation: 8,8141
Nulls: 0

Figure 3.12: Histogram of age for male user.

The Figure 3.13 shows the age distribution of female user who are using online dating site in Istanbul. The graph shows that the age distribution of female users were not distributed equally. Income level of the social media users was defined depending on international standards. In this way, users were defined with five levels from poor to high as poor, low, lower middle, upper middle and high.



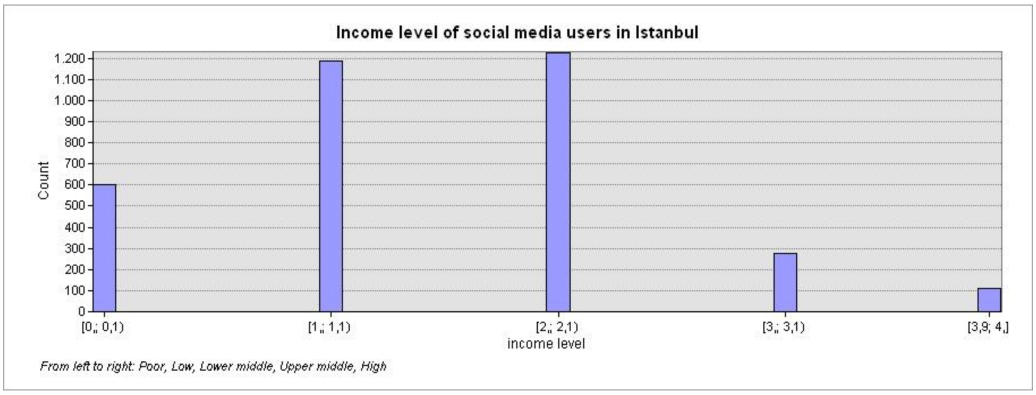
Field
age

Statistics:

Count: 1402
Minimum: 18
Maximum: 66
Sum: 44936
Mean: 32,051355
Standard Deviation: 8,502593
Nulls: 0

Figure 3.13: Histogram of age for female user.

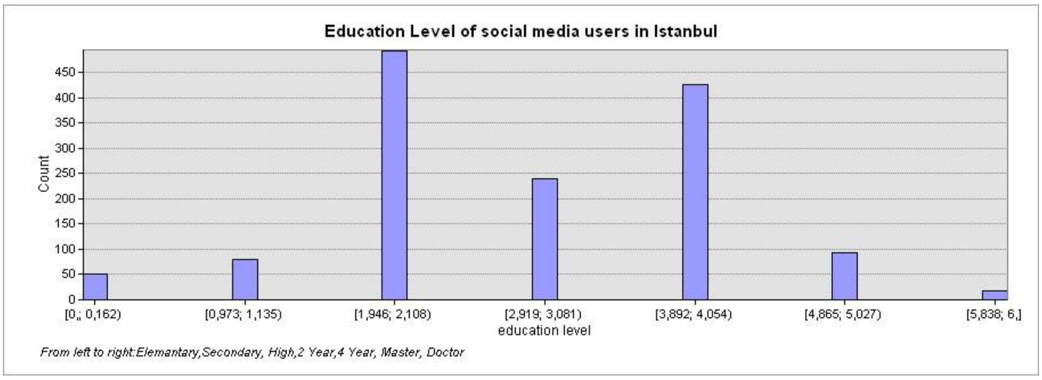
As undertood, high perentage of the social media users was defined at low income level, especially low (1000-2000 TL) and lower middle (2000-5000 TL) (Figure 3.14). The highest user level is at lower middle income level in Istanbul.



From left to right: Poor, Low, Lower middle, Upper middle, High

Figure 3.14: Histogram of income (social media users).

According to education, the users classified into five levels: elementary, secondary, high, 2 year, 4 year, master, and doctor (Figure 3.15).



From left to right:Elemantary,Secondary, High,2 Year,4 Year, Master, Doctor

Figure 3.15: Histogram of education level (social media user).

Thematic map on Figure 3.16 confirms that the users at lower middle income level are situated in the county where the site users live mostly.

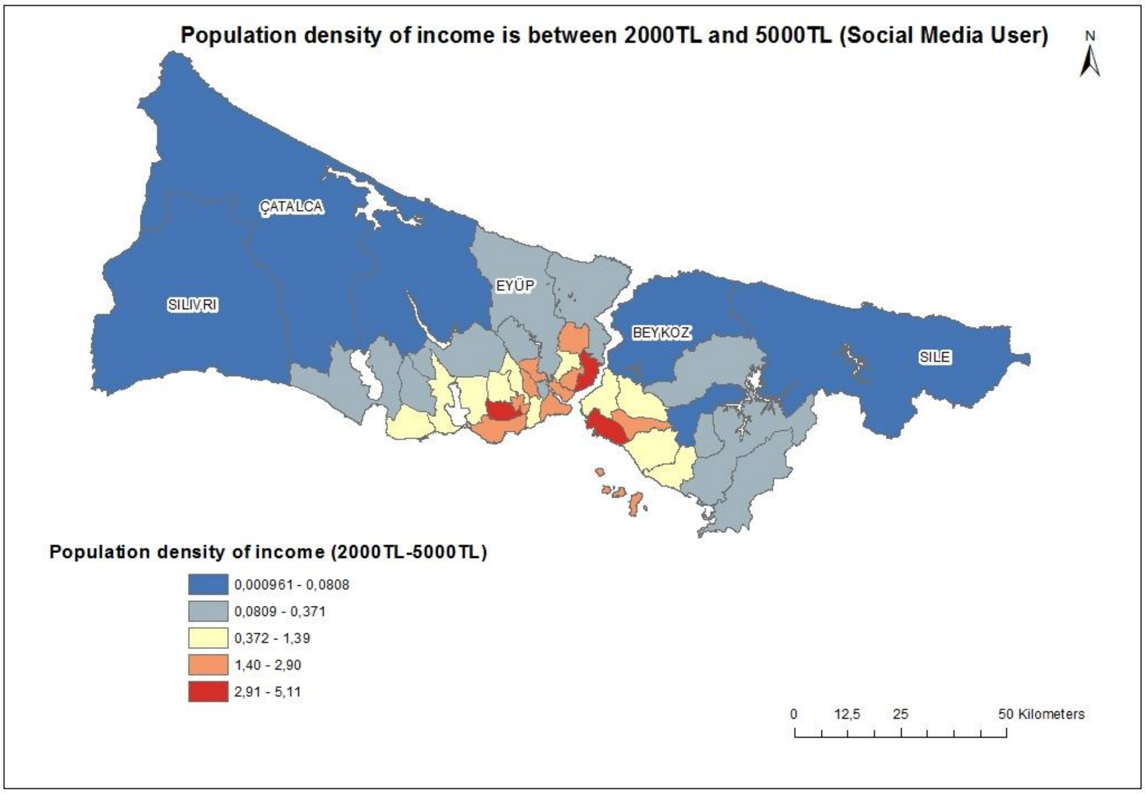

Population density of income is between 2000TL and 5000TL (Social Media User)
N
ÇATALCA
SILIVRI
EYÜP
BEYKOZ
SILE
Population density of income (2000TL-5000TL)
0,000961 - 0,0808
0,0809 - 0,371
0,372 - 1,39
1,40 - 2,90
2,91 - 5,11
0
12,5
25
50 Kilometers

Figure 3.16: Population density lower middle income level, Social Media User, 2013.

It is clear that education level is generally between high school and undergraduate program. Figure 3.17 shows that the high school education level is high in these counties: Adalar, Bakırkoy, Besiktas, Kadikoy, Avcilar, Beylikduzu. Figure 3.18 shows that the high unemployment ratio is high in these counties: Adalar, Bakırkoy, Besiktas. The common counties for education and unemployment ratio are Adalar, Bakirkoy, and Besiktas. So, there is a significant data and a positive relations between education and unemployment ratio. Social profile can be determined by using this information.

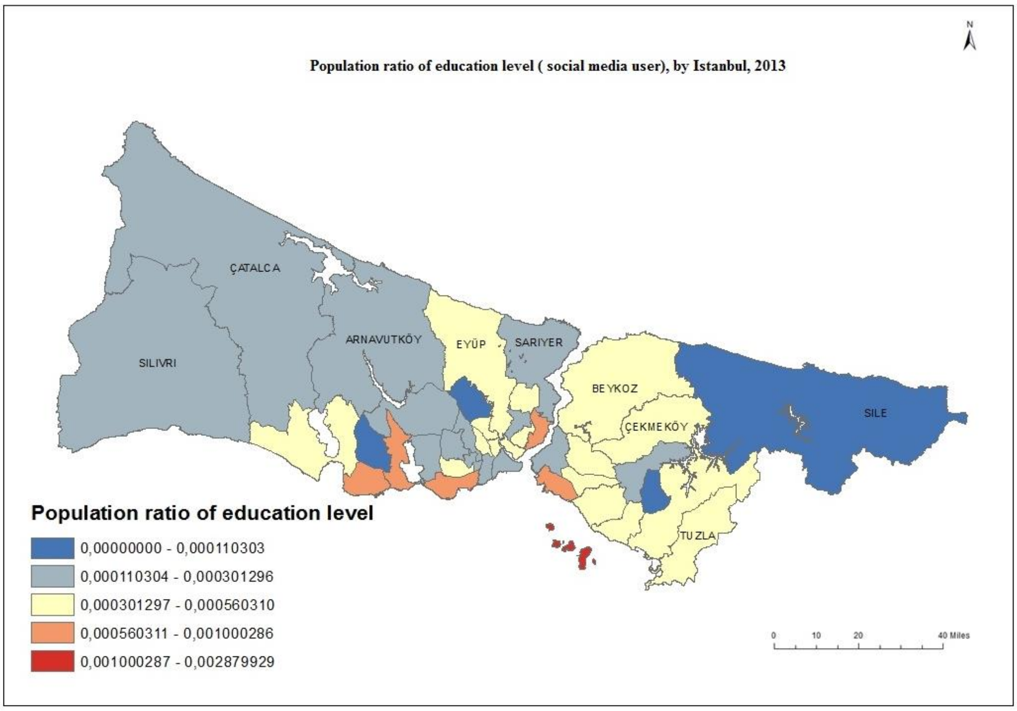


Figure 3.17: Population ratio of education level (social media user), 2013.

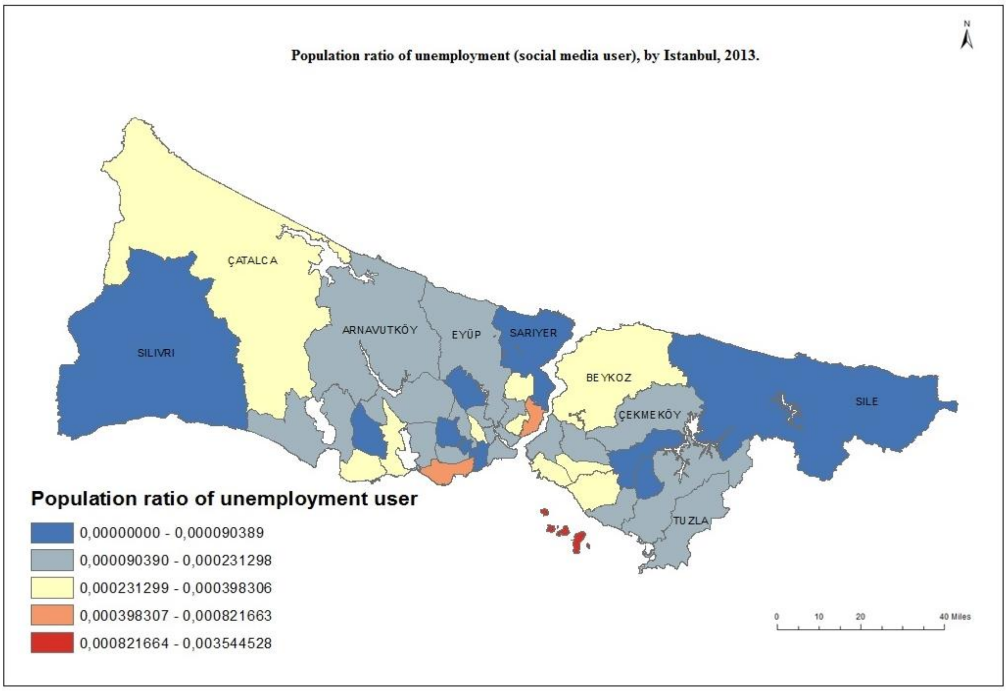


Figure 3.18: Population ratio of unemployment (social media user), 2013.

3.4 Exploring Distribution of Pattern

Firstly, the mean center of the site users was determined in Turkey (Figure 3.19). Eskisehir is the mean center of the locations of social media users. The map on Figure **3.20** shows that Beyoglu is the mean center of the features in Istanbul.

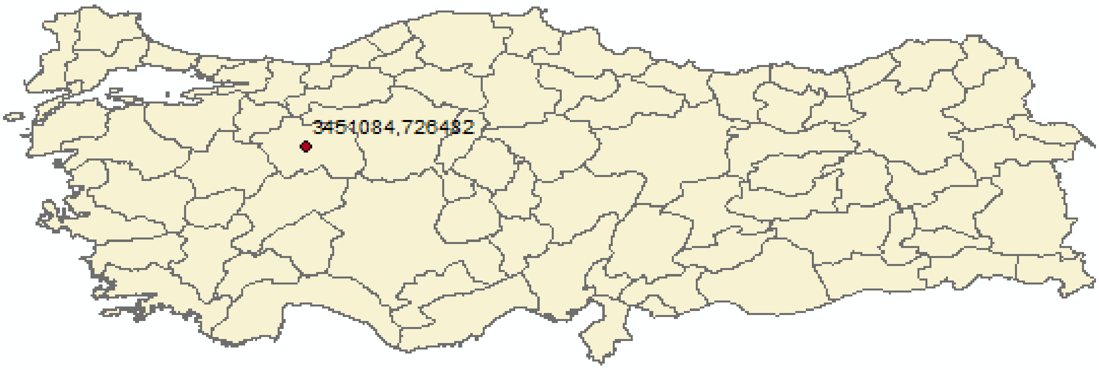


Figure 3.19: Mean Center of Features in Turkey (Social Media Users).

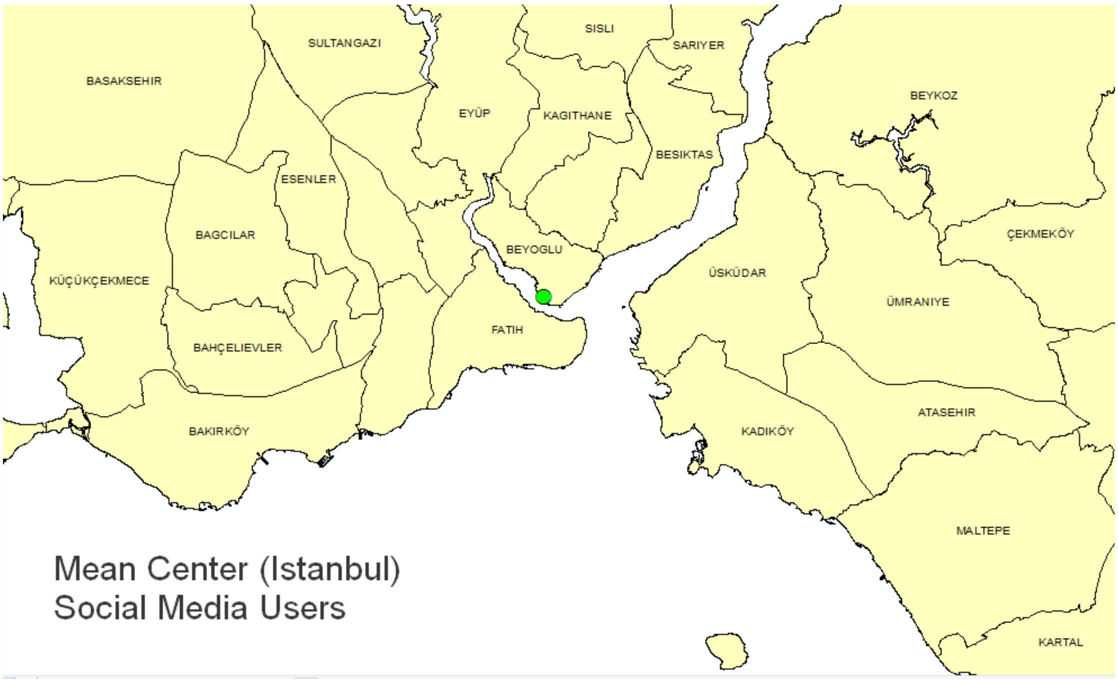


Figure 3.20: Mean Center of Features in Istanbul (Social Media Users)

Secondly, directional distribution was calculated. It shows tendency and dispersion of features (Figure 3.21). North-west and North-east direction was found in social media user region. Thirdly, voronoi map giving an idea about study area was calculated. Voronoi map type were mean and attributes were count. While Figure 3.22 shows Voronoi map in Turkey, Figure 3.23 determines differences between locations in Istanbul. Istanbul has a huge difference between urban and rural area.

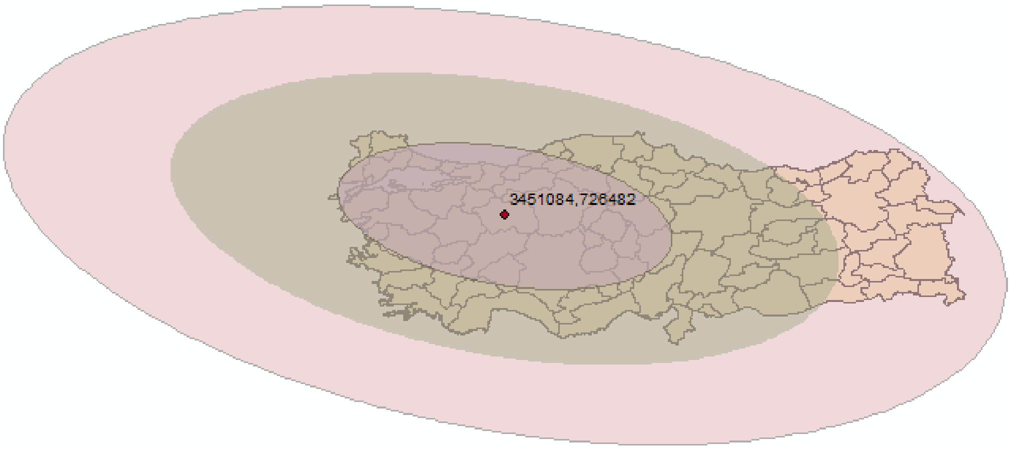


Figure 3.21: Standard Deviational Ellipse.

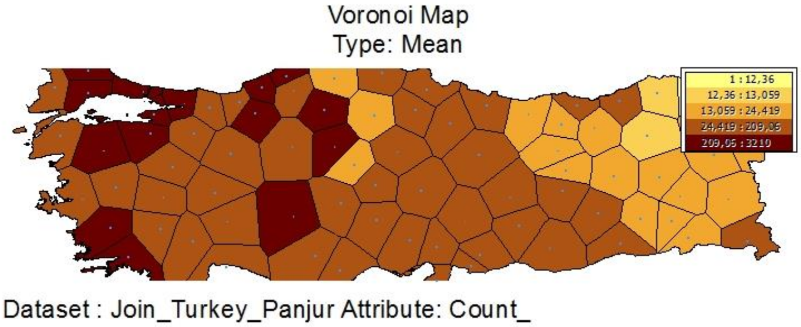


Figure 3.22: Voronoi map in Turkey (Social Media Users).

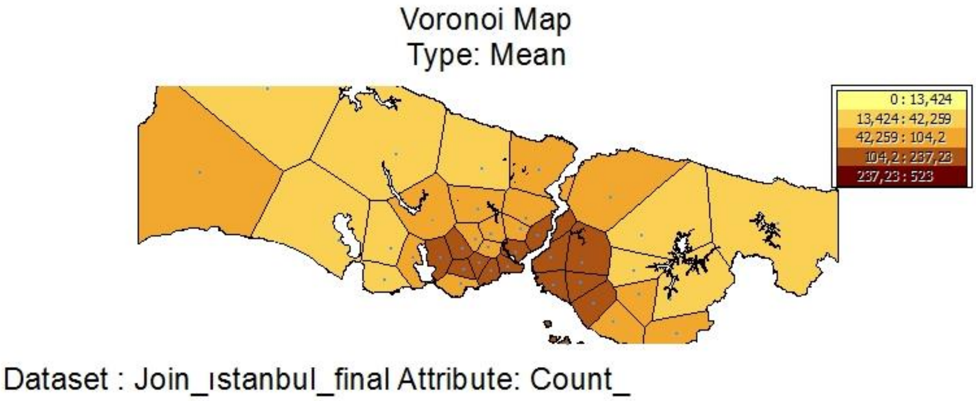


Figure 3.23: Voronoi map in Istanbul (Social Media Users).

In Figure 3.24, it can be easily seen that normal QQ plot helps to determine the huge difference between the number of social media users related with counties. Huge number of users are found in Kadikoy, but Catalca, Sultanbeyli, and Sancaktepe have a less number of users. According to Normal QQ Plot, Bagcilar has the huge number of single man and Kucukcekmece has the huge number of single woman by considering Turkish Statistical Institute data. Figure 3.25 shows that there are significant similarities between income and educational level on counties.

In Figure 3.25, normal and general QQ plots were created by using income and educational level of social media users. Two types of methods were applied to understand the relationships spatially. In conclusion, when used cluster method for Istanbul data, It was determined that education and income levels are not distributed equally on their neighbor. Furthermore, Anatolian side of Istanbul became more equally distribution of income level. The second method of voronoi map is that mean. If we compare the mean value of Istanbul based on educational level, more educated user are living in Bosporus area.

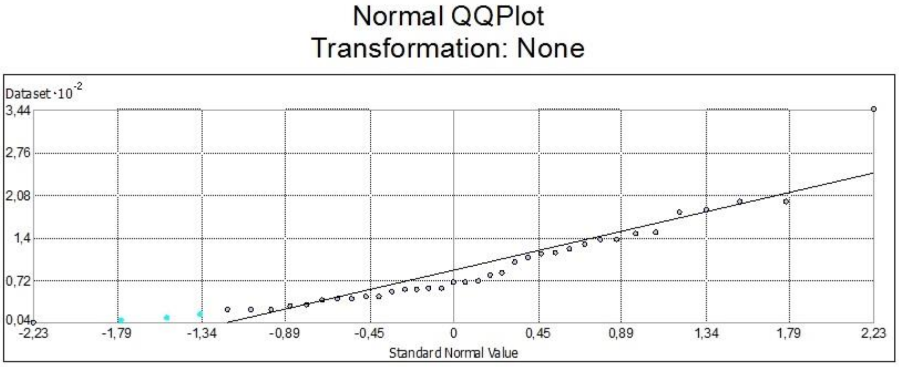


Figure 3.24: Normal QQ Plot (Number of Social Media Users in Istanbul).

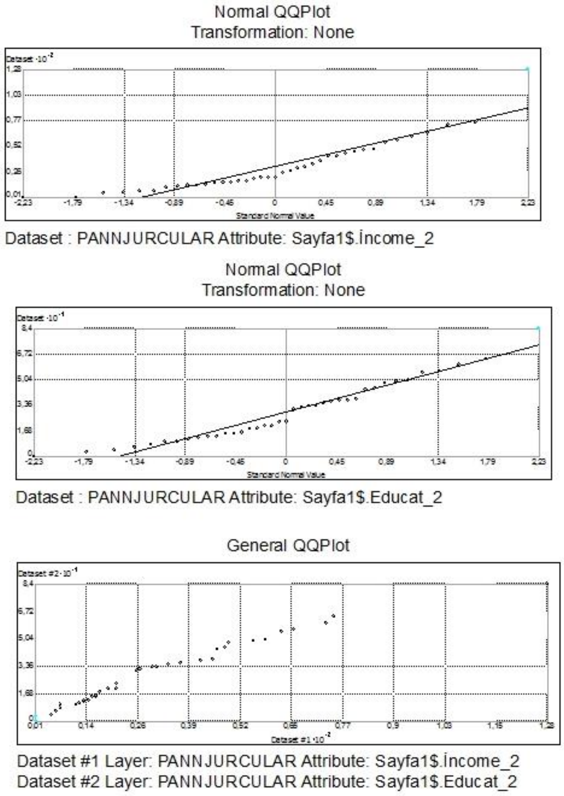


Figure 3.25: Normal and General QQ plot for social media users in Istanbul.

3.5 Hot-Spot Analysis

In this thesis, the distribution of the users is important to know and determine the user profiles. It says that spatial cluster is statistically significant or not. Thanks to Hot-spot analysis, it is the easiest way to identify high and low values as seen on Figure 3.26. It calculates the z-score and p-value. These values help to decide for rejecting null hypothesis. According to results, high values are clustered spatially in Zeytinburnu and Gungoren. Low values are clustered spatially in Catalca and Silivri. In addition, high values are clustered in Europe sides of Istanbul rather than Anatolian sides (Figure 3.27).

After running hot-spot analysis, it gives a z-score. If z-score is positive, there would be a hot-spot map. If the z-score is negative, there would be cold spot map. All these information are related with data neighbors. Furthermore, p-value and z-score are important for understanding the information is statistically significant or not.

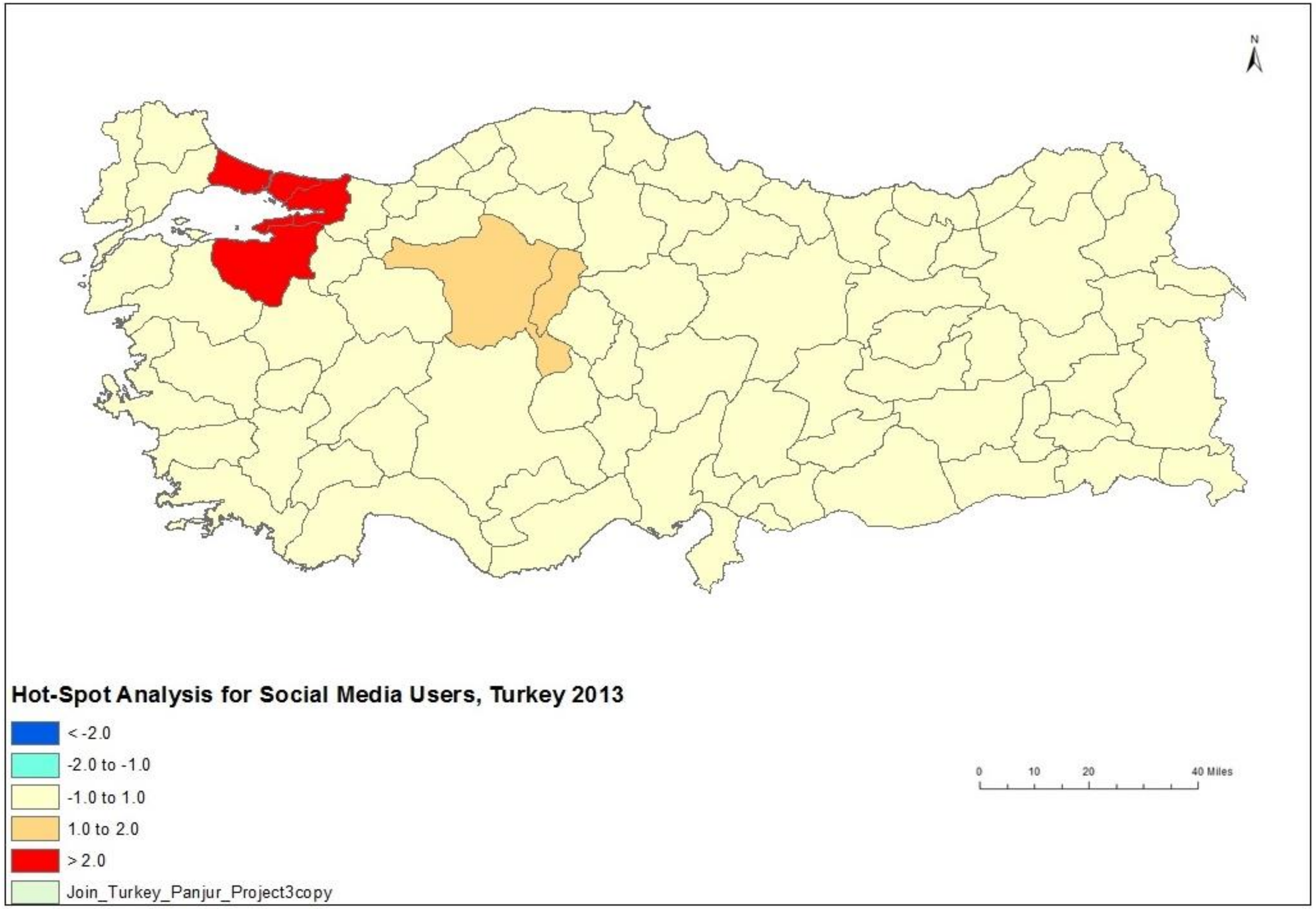

N
Hot-Spot Analysis for Social Media Users, Turkey 2013
< -2.0
-2.0 to -1.0
-1.0 to 1.0
1.0 to 2.0
> 2.0
Join_Turkey_Panjur_Project3copy
0
10
20
40 Miles

Figure 3.26: Hot-Spots Analysis of Social Media Users, Turkey 2013.

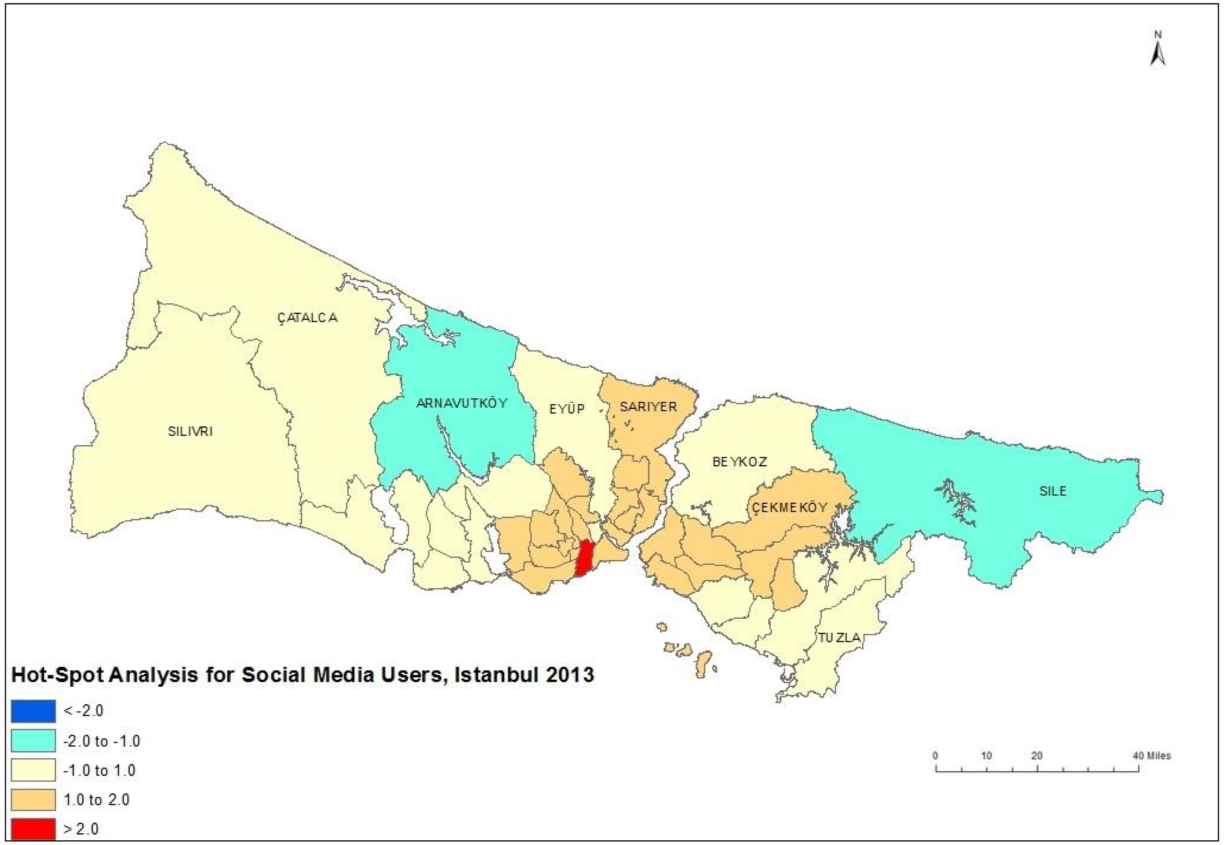


Figure 3.27: Hot-Spots Analysis of Social Media Users, Istanbul 2013.

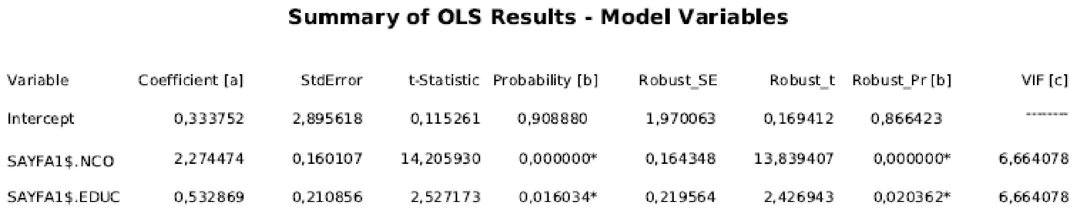
3.6 Regression Analysis

Regression Analysis was used with OLS and GWR method. It was determined that dependent variable is user population, while independent variables are income (2000-5000 TL) and Education (High School).

3.6.1 OLS Method

According to the result of OLS method seen on Table 3.3, coefficients are quite good. Two of them are different from zero. Probability shows that there is a statistically significant because independent variables have asterix bottom of the probability column. There is no large variance inflation factor between independent variables because VIF is less than 7.5. Furthermore, VIF shows that there is no redundancy among explanatory variables. Model performance is good. AIC is 295 and Adjusted R-squared is 0.979620. Koenker (BP) is not statistically significant. So, there is no need to look at Robust Probability. Jarque-Bera statistic is not statistically significant. So, this model is not biased. It means that residuals are normally distributed for this model. The graph seen on Figure 3.28 shows that residuals are normally distributed. This model is not a biased. This scatterplot gives an idea about model. Figure 3.29 shows OLS regression map for the site users in Istanbul.

Table 3.3: OLS Regression Coefficients for social media users, Istanbul 2013.



Summary of OLS Results - Model Variables

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	0,333752	2,895618	0,115261	0,908880	1,970063	0,169412	0,866423	--------
SAYFA1$.NCO	2,274474	0,160107	14,205930	0,000000*	0,164348	13,839407	0,000000*	6,664078
SAYFA1$.EDUC	0,532869	0,210856	2,527173	0,016034*	0,219564	2,426943	0,020362*	6,664078

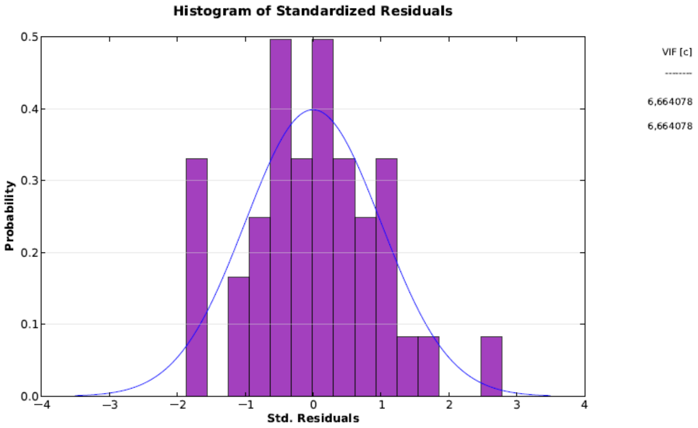


Figure 3.28: Histogram of Standardized Residuals of Social Media Users, 2013

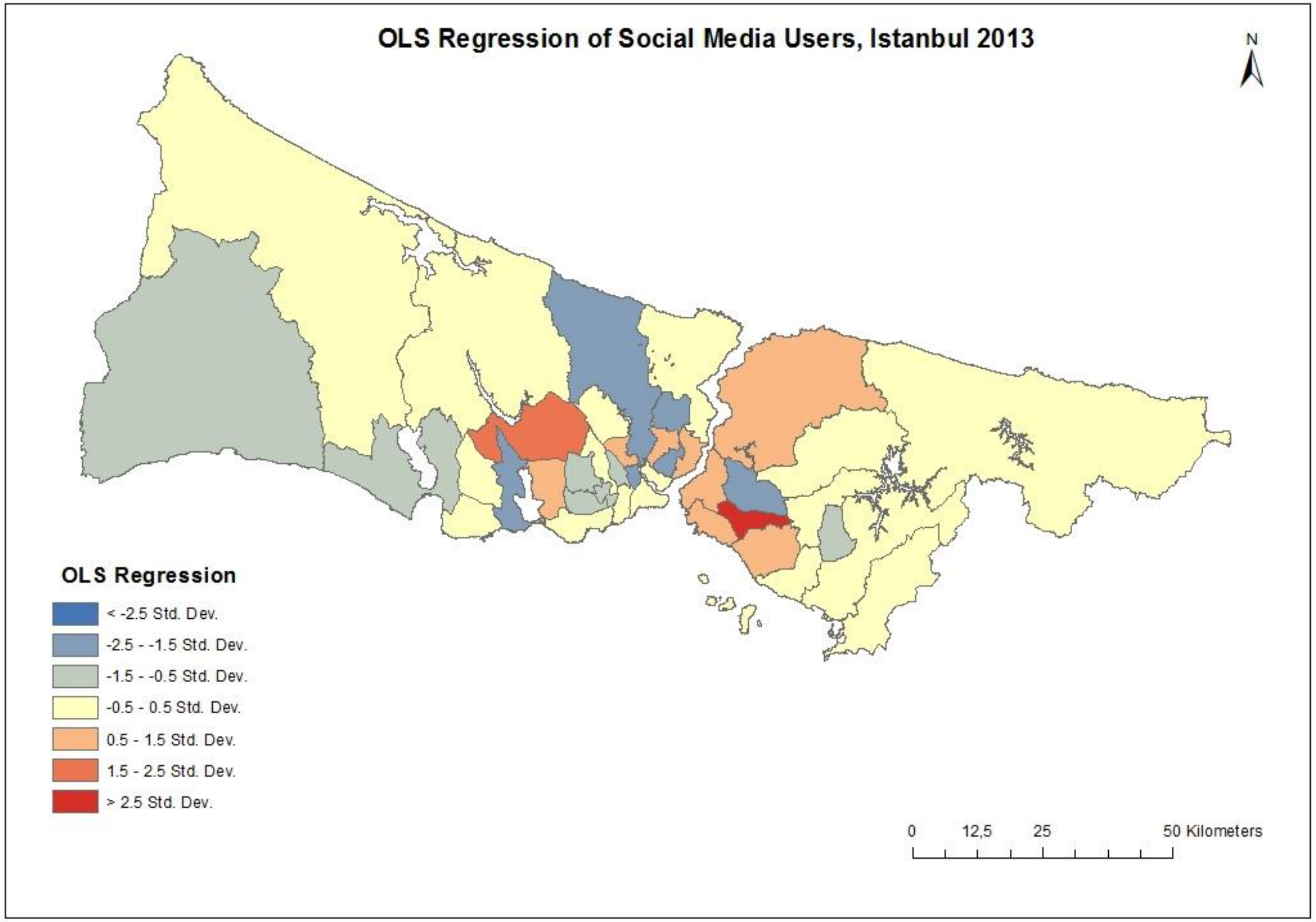


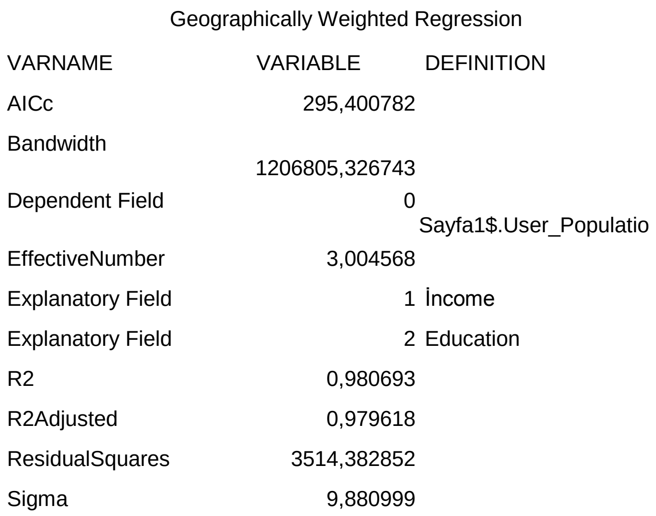
Figure 3.29: OLS Regression of Social Media Users, Istanbul 2013.

3.6.2 GWR Method

According to the result of GWR method, AIC is 295,400782 and Adjusted R2 is 0.979618 as seen on Table 3.4. In addition, all explanatory variables have different coefficients. Figure 3.30 shows GWR regression map for the site users in Istanbul.

According to these results, OLS and GWR methods are reliable. There is no huge gap between these results. Both techniques should be used to determine social media profile.

Table 3.4: GWR of social media users, Istanbul 2013.



Geographically Weighted Regression

VARNAME	VARIABLE	DEFINITION
AICc	295,400782	
Bandwidth	1206805,326743	
Dependent Field	0	Sayfa1$.User_Populatio
EffectiveNumber	3,004568	
Explanatory Field	1	İncome
Explanatory Field	2	Education
R2	0,980693	
R2Adjusted	0,979618	
ResidualSquares	3514,382852	
Sigma	9,880999	

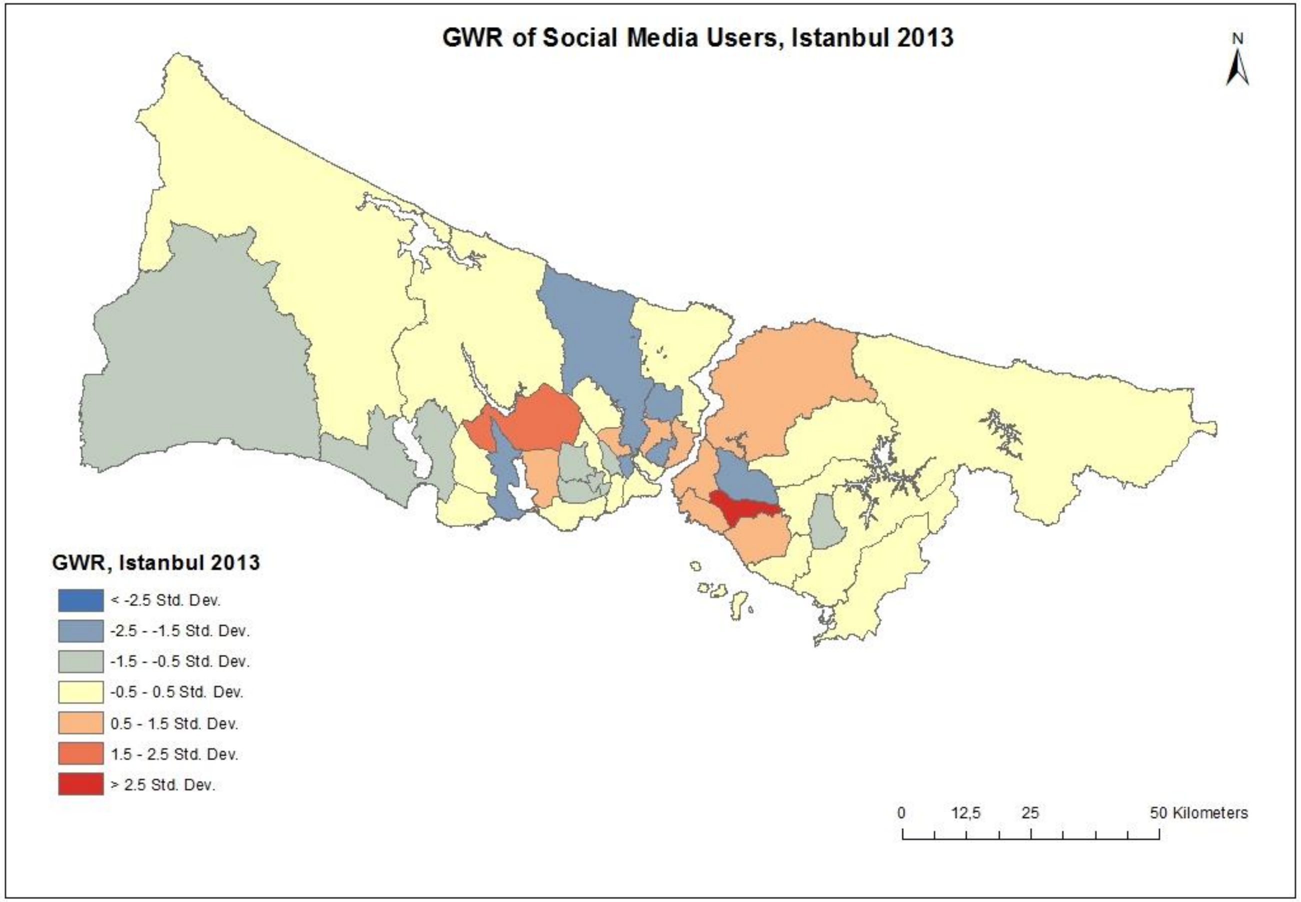


Figure 3.30: GWR of social media users, Istanbul 2013

4 CONCLUSION

In this study, a variety of spatial statistical techniques are used to identify social media users profile. These spatial statistical tools are used not for only find a relations between demographic data, but also determine the study area. The application consist of six parts. These parts are selecting study area, data gathering and preparation, examining thematic data, exploring distribution of pattern, hot-spot analysis, and regression analysis.

Many ways can be used to select study area, but in this study the number of people is considered to identify user profile. If the number of people is high in study area, then these would be usable to find meaningful information. Turkstat data were used for selecting study area. The data about population living in Turkey was found in Address Based Population Registration System.

Social media users are living in Istanbul mostly like the distribution of Turkey's population. According to social media cite example, 59% of male internet users use social media and 41% of female internet users use social media while the population percentage of female is 49.8% and percentage of male is 50.2% in Istanbul. It shows that male users prefer to use social media rather than female users.

Online social dating platform users analyzed in this thesis have 2.2 million users. These data sets are stored in MySQL database. Firstly, the data was retrieved for whole users in Turkey but then decided to change study area for focusing on Istanbul. This was selected to improve study quality because Istanbul is the metropolitan area. Thematic maps were used to understand the distribution and the density of social media users. Thematic maps show that Istanbul is the most crowded city according to Turkstat and also given social media site example.

The overall pattern of the results shows strong relations between user profiles. Many statistical analysis techniques were used for this thesis. Each analysis is not appropriate for each data. First of all, it is important to find the best analysis method for data. For this aim, ESDA tools helps to identify the explanatory variables for regression analysis. These explanatory variables are educational level and income level in this study.

Thematic maps help to compare Turskish Statistical Institute data and social media users data. Thematic maps show that most of social media users are living in Kadıkoy, Bahcelievler, and Besiktas. Also, Turkstat statistics find the same result. There is a positive relation between Turkstat and social media data. Futhermore, social media users are high in urban areas. They do not prefer to live in rural areas.

One of the highest population county in Istanbul is Bagcilar but single social media users are low. Furthermore, economic level is too low in Bagcilar. So, many young people would prefer to marriage in their earlier age.

This ratio was calculated by dividing number of social media users to general number of people in a county. Adalar, Kadıkoy, Bakırkoy, Besiktas, Beylikduzu have more woman ratio rather than Esenyurt, Kucukcekmece, Bagcilar, Gungoren, Zeytinburnu, Sultangazi, Sariyer, Sancaktepe, Sultanbeyli, Sile, Kagithane, and Sancaktepe.

Adalar, Kadıkoy, Bakırkoy, Besiktas, and Beylikduzu have more man ratio than Esenler, Esenyurt, Arnavutkoy, Catalca, Bagcilar, Gungoren, Zeytinburnu, Sultangazi, Sariyer, and Kagithane. Furthermore, the woman ratio is less and man ratio is high in Sile and Sancaktepe. The man ratio is less and woman ratio is high in Arnavutkoy and Catalca. So, anatolian side of Istanbul has more woman ratio and Europe side of Istanbul has more man ratio.

Besides, man and woman ratio are high in same counties. It is said that living in the same place could be give same gender ratio. In the next step of this thesis try to find a relation between people are living same region may show similar characteristics or not. So, other demographic features are examined such as age, education and income level. The age differences between gender were examined by histogram. According to age histogram, males have the highest population rate between ages 24 to 36. Also, it is distributed equally. Females have the highest population rate between ages 18 to 66. Also, it is not distributed equally. It is understood that the number of female users are less than male users and their age is not distributed equally. In addition, age histogram pointed out that the mean age of male user is 33 and female user is 32. There is no huge gap between ages. In some counties such as Bagcilar and Esenler, median age is low. The population of these counties are high but economical structure is low.

The education level is close between high school and undergraduate program for social media users. This study also looked into unemployment rate and education level relations. In Adalar, Bakırkoy, and Besiktas, there is a high unemployment rate. In Adalar, Bakırkoy, Besiktas, Kadikoy, Avcilar, and Beylikduzu, social media users have high school educational level. As a result, social media users are living in the counties which have high unemployment rate and high school education level.

Most of social media users was defined at low income level, especially low (1000-2000 TL) and lower middle (2000-5000 TL).

After examining thematic maps, spatial pattern was explored with ESDA tools. Firstly, mean center of Turkey was determined as Eskisehir. Then, mean center of Istanbul was determined as Beyoglu. Directional distribution of features were found as North-Weast and South-East direction. It gives an in idea about social media users where distributed in Turkey. The manager of social media site can decide where the users should be encouraged. The second ESDA tool is Voronoi map used to identify differences between regions. Thiessen polygons are used to create Voronoi map. These polygons considered their neighbour relations. Voronoi map of Turkey shows that Istanbul is highly related with its neighbour based on spatial location. Also, Voronoi map for Istanbul shows that there is a more strong relations in seaside of region. As a result, Istanbul has a huge difference between urban and rural area. Voronoi map helps to find a general idea about it but Normal QQ plot shows the which counties are not related spatially. There is a huge differences between some counties such as Kadikoy and Catalca. Kadikoy has the maximum number of social media users. Catalca has the minimum number of social media users. There is also available comparison of two different data features. It is done by General QQ Plot. In this study, income and education level is examined and it is found that there is a strong relations between them.

The distribution of patterns is examined by using ESDA tools. First of all, hot-spot techniques are used to determine study area. It helps to select study area as Istanbul city. The hot-spot tools are very useful because it helps to cluster the users spatially. So, where the users are clustered is easily found. High and low values are found easily by using hot-spot techniques. It calculates z-score and p-value. These determine the null hypothesis statistically significant or not. Z-scores shows the standart deviations. If the z-score is negative than it called cold spot. If the z-score is positive, than it called hot-spot. Also, p-value is probability. If p value is very low, then null hypothesis can be reject. When hot-spot analysis is applied on Turkey map, it is shown that Istanbul and neighbor counties are clustered significantly. Hot-spot analysis is also used for Istanbul city. According to the results, high values are clustered spatially in Zeytinburnu and Gungoren. Low values are clustered spatially in Catalca and Silivri. In addition, high values are clustered in European sides of Istanbul rather than Anatolian sides. So, hot-spot analysis helps to answer how many and where questions.

The next step is to understand the relations by using Regression Analyst technique. First of all, OLS technique gives Adjusted R-squared value to measure model performance. In this study, model performance is good because Adjusted R-squared is 0.979620. OLS regression analysis gives VIF number measuring redundancy among explanatory variables. In this study model, there is no large variance inflation factor between independent variables because VIF is less than 7.5. Spatial Autocorrelation tool is used to ensure residual distributed randomly. In this study, OLS results shows that residuals are distributed randomly. The second method is GWR for regression analysis. The most important differences between them, GWR builds a local regression equation for each feature in the dataset.

OLS and GWR methods defined the same results. Both models can be used to analyze user profiles. A strong correlation between income and educational level is determined.

For future researches, there is no need to store all data for this kind of analysis. People can predict which data is required to determine the user profiles of any site before the analysis. Similar to this thesis methodology, required data should be collected for research questions and the data and their statistics analyst should be managed in geographic database.

REFERENCES

Bailey, T. C., & Gatrell, A. C. (1995). Interactive spatial data analysis (Vol. 413). Essex: Longman Scientific & Technical.

Brillinger, D. R. (2002). John W. Tukey: his life and professional contributions. Annals of Statistics, 1535-1575.

Mauch, J.W. (2005). Social mathematics in the curriculum of American civics: An analysis of selected national and state standards and of Magruder's American government. Unpublished doctoral dissertation, Pennsylvania State University.

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 210-230.

Rosander, A. C. (1935). A quantitative study of social attitudes. *The School Review*, 614-620.

Scott, L., & Getis, A. (2008). Spatial statistics. In K. Kemp (Ed.), Encyclopedia of geographic information science. (pp. 437-441). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412953962.n199

ESRI, 2013. Regression Analysis Using ArcGIS, ESRI Virtual Campus Course, Redlands, USA

Bergsma, S., & Van Durme, B. (2013). Using Conceptual Class Attributes to Characterize Social Media Users. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (pp. 710-720).

Zheng, H. (2012). Do people die from income inequality of a decade ago?. Social science & medicine, 75(1), 36-45.

Dai, D., Zhang, Y., Lynch, C. A., Miller, T., & Shakir, M. (2013). Childhood drowning in Georgia: A geographic information system analysis. Applied Geography, 37, 11-22.

Park, D. B., Lee, K. W., Choi, H. S., & Yoon, Y. (2012). Factors influencing social capital in rural tourism communities in South Korea. Tourism Management, 33(6), 1511-1520.

Cheng, J., & Fotheringham, A. S. (2013). Multi-scale issues in cross-border comparative analysis. Geoforum, 46, 138-148.

Karlsdotter, K., Martín Martín, J. J., & López del Amo González, M. (2012). Multilevel analysis of income, income inequalities and health in Spain. Social Science & Medicine, 74(7), 1099-1106.

Chalkias, C., Papadopoulos, A. G., Kalogeropoulos, K., Tambalis, K., Psarra, G., & Sidossis, L. (2013). Geographical heterogeneity of the relationship between childhood obesity and socio-environmental status: Empirical evidence from Athens, Greece. Applied Geography, 37, 34-43.

Zheng, H., & George, L. K. (2012). Rising US income inequality and the changing gradient of socioeconomic status on physical functioning and activity limitations, 1984–2007. Social Science & Medicine, 75(12), 2170-2182.

Deaton, A., & Lubotsky, D. (2003). Mortality, inequality and race in American cities and states. Social science & medicine, 56(6), 1139-1153.

Dale, P. F., & McLaughlin, J. D. (1988). Land information management: an introduction with special reference to cadastral problems in Third World countries. Oxford: Clarendon Press.

Yomralıoğlu, T. (2000). *Coğrafi bilgi sistemleri: Temel kavramlar ve uygulamalar*. Karadeniz Teknik Üniversitesi.

UN (2013). UN, World Population Prospects: The 2012 Revision, Press Release (13 June 2013). Retrieved 21.04.2014 from http://esa.un.org/unpd/wpp/Documentation/pdf/WPP2012_Press_Release.pdf.

Allen, M. P. (1997). The origins and uses of regression analysis. Understanding Regression Analysis, 1-5.

Mitchell, A. (2005). The ESRI Guide to GIS analysis, Volume 2: Spartial measurements and statistics. ESRI Guide to GIS analysis.

Turkish Statistical Institute (2013). The Results of Address Based Population Registration System, 2013. Retrieved 28.04.2014 from http://www.turkstat.gov.tr/PreHaberBultenleri.do?id=15974 .

Dramowicz, K. (2010, July 13). *Using spatial statistics and geostatistical analyst as educational tools*. Paper presented at the ESRI User Conference, San Diego, California.

Goodchild,M.&Sui,D. 2011. The convergence of GIS and social media:challenges for GIScience

CURRICULUM VITAE

Name Surname	: İrem Erkuş
Date of Birth and Place	: June 25,1989 and Istanbul
E-mail	: erkusi@itu.edu.tr

Education

B.Sc degree : 2011, Isik University, Faculty of Science and Letters, Mathematics

M.Sc degree : …. , ITU, Institute of Informatics, Geographic Information Technology