THREE ESSAYS ON DATA-DRIVEN PROBLEMS

BY HE ZHANG

A dissertation submitted to the

Graduate School—Newark

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Professor Yao Zhao

and approved by

Newark, New Jersey

May, 2018

ABSTRACT OF THE DISSERTATION

Three Essays on Data-Driven Problems

by He Zhang

Dissertation Director: Professor Yao Zhao

This dissertation comprises three essays on data-driven problems. The first essay addresses the dynamic, evolving academic social network through co-authorship. A descriptive analysis of the research publications appearing in *Management Science* (MS) and *Operations Research* (OR) from 1995 to 2014 is visualized to throw light on the basic collaboration patterns. The second part of the analysis examines the similarities and differences between MS and OR. The social network characteristics are studied and a clique analysis by size, structure, density, and diversity is performed to understand the factors that drive productivity. Finally, new metrics are introduced to assess the journals on their openness to new authors, and on the influence of the status of the editorial board on the chance of publishing a submitted article. The second essay focuses on healthcare analytics. It provides a general macro-level review of the healthcare industry. First, the essay includes an analysis of population health (inpatients) in the New York State, disease mapping associated with the gender, age, and race, and the trend over time is also studied. Second, the relationships are explored between hospital charges or costs and service quality ratings, and between hospital charges or costs and local demographics indicators. Third, hospital profitability and drivers for the profits are investigated. Finally, an optimization model is built to determine the locations and services provided for hospital network expansion. The model is based on the parameters estimated on the New York State inpatient data. The third essay examines the Chinese housing bubble, or, more specifically, on the relationships among housing prices, macro-economic policies, and physical properties. The aim is to determine the validity, if any, of the many popular conjectures circulated in the media. Based on real-time trading data of the second-hand housing market in Beijing and macro-economic indicators, a valuation model is developed to establish the housing market trend in China in order to help investors in choosing the appropriate investment strategy, and the policy makers in regulating the market.

Acknowledgements

I would like to express my special appreciation and thanks to my advisor, Professor Yao Zhao. His guidance helped me throughout my research and writing of this thesis. He has been a tremendous mentor and his advice on both my research and career has been invaluable. At the early stage of my graduate study, he encouraged me to look for an internship to accumulate the industrial experience, attend academic conferences to exchange ideas. He introduced me to other colleagues, and helped me expand my academic network. I also gratefully acknowledge the members of my Ph.D. committee for their time and valuable feedback on the preliminary version of this thesis. I want to express gratitude especially to Professor Cuihong Li, who traveled from Connecticut state to attend my dissertation defense, and Professor Weiwei Chen and Lian Qi for offering me invaluable advice durng the past several years. In addition, I am also grateful to all fellow students, especially to Xin Xu, Hui Dong, Yijun Wang, Zhengwei Wang, Kwon Gi Mun, Ju Myung Song, Irene Akaab, Irene Gerlovin, Phat Luong, Hao Zhong, Jianing Yao, Arim Park, Theekshana Somaratna, Olena Rudna, Maryam Mahdikhani, Qiang Li, and Peng Li, who have helped me throughout the years at Rutgers. Also, I express special appreciation to the Rutgers Business School for offering me the opportunity to carry out my doctoral research and for their financial support. This experience has also given me the ample opportunity to transition from physics into the quantitative field of business management. Finally, I would like to acknowledge the most important persons in my life, Dragon and my parents. They have been a constant source of strength and inspiration, providing me with support under all circumstances.

Table of Contents

Chapter 1

Introduction

Data-driven problem solving is an approach that involves building tools and developing the abilities to explore the information behind the data. During the past decade, advances in information technology have ignited a revolution in decision-making from fields such as business to policy policing. In the past, decisions in such areas had been heavily influenced by qualitative analysis and experiences rather than by quantitative analysis. However, recent improvements in the ability to collect and analyze large amounts of data have allowed decision-makers to reduce the noise in uncovering facts. Meanwhile, the best way to improve government policies is to improve the ability to manage risk and produce results. This could be achieved by a shift towards data-based policy making. The dissertation comprises three essays on data-driven problems. The first essay addresses the dynamic, evolving academic social network through co-authorship. A descriptive analysis of the research publications in *Management Science* (MS) and *Operations Research* (OR) from 1995 to 2014 is visualized to throw light on the basic collaboration patterns. The second part of the analysis examines the similarities and differences between MS and OR. The social network characteristics are studied and a clique analysis by the size, structure, density, and diversity is performed to understand the factors that drive productivity. Finally, new metrics are introduced to assess the journals on their openness to new authors, and on the influence of the editorial board status on the chance of publishing a submitted article. This work aims to offer useful insights to current Ph.D. students or junior faculties about how academic social networks are formed through journal publications. The second essay is on healthcare analytics. It provides a general macro-level review of the healthcare industry. First, the essay includes an analysis of population health (inpatients) in the New York State, disease mapping associated with the gender, age and race, and the trend over time is also studied. Second, the relationships are explored between hospital charges or costs and service quality ratings, and between hospital charges or costs and local demographics indicators. Third, hospital profitability and drivers for the profits are investigated. Finally, an optimization model is built to determine the locations and services provided for hospital network expansion. The model is based on the parameters estimated on the New York State inpatient data. This essay is to help the researchers in the healthcare field understand geographic distribution of the disease, how non-technical service and demographics influent the hospital charge and cost, and take advantage of the profits drivers. The third essay focuses on the Chinese housing bubble, or, more specifically, on the relationships among housing prices, macro-economic policies, and physical properties. The aim is to determine the validity, if any, of the many popular conjectures circulated in the media. Based on real-time trading data of the second-hand housing market in Beijing and macro-economic indicators, a valuation model is developed to establish the housing market trend in China in order to help the investors in choosing the appropriate investing strategy, and the policy makers in regulating the market.

Chapter 2

Co-author Network Analysis of Management Science (MS) and Operations Research (OR)

2.1 Introduction

2.1.1 Definition of Social Network

A social network is a social structure made up of a set of social actors (such as individuals or organizations) and a set of the dyadic ties between these actors in Wasserman and Faust (1994). Social networks support people with common interests, history, and cultures, and they also help strangers connect based on shared interests, political views, or activities.

The history of social networks can be traced back as far as the late 1890s when mile Durkheim and Ferdinand Tnnies foreshadowed the concept of social networks in their research. They argued that potential links exist between individuals who share values and beliefs, and who thus form what they referred to as a community studied by Ferdinand et al. (1887). Simmel and Leggewie (1908) discussed in 1908, Georg Simmel developed a theory on the nature of networks and the effects of network size. Jacob Moreno introduced the sociogram in 1934, a graphical representation of ties between individuals to demonstrate the structure of social groups. Researchers subsequently developed an interest in sociograms and applied them to explore crisscross relationships and their influence over the daily lives of individuals which is studied by Moreno et al. (1934). An increasing number of visual tools applied in quantitative studies are developed to facilitate the analysis of social networks. Since these approaches were mathematically formalized in the 1950s, a growing number of scholars have explored the underlying patterns. Among them was Stanley Milgram, a Harvard sociologist, who developed the renowned six degrees of separation theory in Haggbloom et al. (2002). Freeman (2004) mentioned that in the 1980s, social network analysis (SNA) became pervasive in behavioral sciences , and, currently, it comprises one of the major paradigms of contemporary sociology.

SNA is a strategy applied in exploring social structures through network and graph theories. It is derived from the work of early sociologists such as the above-mentioned Georg Simmel and mile Durkheim, who wrote about the importance of studying patterns of relationships connecting social actors. This research can be found in Freeman (2004).

2.1.2 Co-authorship

One of the essentials in furthering an academic career involves maintaining high levels of research productivity. Following a meticulous research process, scholars are expected to publish their findings and exert a degree of impact on the expert community. An exemplary and often-cited study provides solid evidence of the authors research ability, and it helps build sound reputation, which not only influences receiving future grants and payment but also encourages the scholar to continue conducting their research. Individual knowledge, however, is limited. Therefore, collaboration is preferred in order to share expertise, distribute the workload, and optimize resources, studied by Hauptman (2005).

2.1.3 Literature review

Katz and Martin (1997) co-authorship is a form of collaboration in which people join their efforts and publish their research findings through paper or electronic media. Individual co-authorship networks are threaded together based on co-authors who form a large network, which is referred to as the collective of individual co-authorships, discussed in Ding (2011). de Souza and Barbastefano (2011) found in the network, knowledge can be diffused from one node to the others. Through collaboration, specific expertise or particular skills and hardware resources are shared, studied by Söderbaum (2001) and Bammer (2008) respectively.

That productivity is driven by collaboration is presented by Lee and Bozeman (2005) and it can be rewarding in various scientific disciplines given by Van Rijnsoever and Hessels (2011). Information can be retracted from social networks of co-authorship relationships studied by Said et al. (2008). This network indicates which co-authors directly or indirectly contributed to the field through published papers. Although an interconnected chain of relationships constitutes a social network, the sharing of resources still shared takes place in the forms of knowledge exchange through social interactions. Designing and managing multi-institutional research collaborations poses a potential problem, studied by Corley et al. (2006).

Eaton et al. (1999) performed a structural analysis of co-author relationships and author productivity in selected outlets for consumer behavior research. Abbasi et al. (2011) measured the effects of co-authorship networks on the performance of scholars by conducting correlation and regression analysis. Liao (2011) examined the impacts of collaboration intensity and member diversity in collaboration networks. Han et al. (2009) conducted clique analysis to understand the importance of collaborations.

Social capital associated with impact also influences the publication. Past studies offered by Li et al. (2013) have analyzed co-authorship networks and research impact from a social capital perspective. That the strength of social influence can be quantified and estimated within a model based on real-world large networks was discussed in Tang et al. (2009).

This thesis focuses on research contributions of academic institutions and individual researchers within the field of management science and operation research over a 20-year period (1995-2014). The objective is to discover collaboration patterns, factors influencing individual research productivity, and the diachronic network evolution. Moreover, several studies examined co-author networks in the business-related academic disciplines. In previous work, Young et al. (1996) and Barman et al. (1991) found individual research productivities were evaluated in a set of journals identified as relevant for the field of operations management (OM). Moreover, research productivity was studied in five selected institutions *Management Science* (MS), *International Journal of Production Research* (IJPR), *Decision Sciences* (DS), *Journal of Operations Management*

(JOM), and *Institute of Industrial Engineers Transactions* (IIE) and their influence on the OM discipline in the 1980s and the early 1990s was discussed in Malhotra and Kher (1996). Another study looked at top individual researchers based on their publications in 20 journals from 1959 to 2008 given by Hsieh and Chang (2009). The contributions of academic institutions and individual authors to 11 top journals in the operations field were studied over a period of 26 years(1985-2010) which was presented in Shang et al. (2015). Huber et al. (2014) studied the 50-year history of the *Journal of Marketing Research.* The contribution of PhD graduates has also been examined well in Fry et al. (2013).

The contribution of the present chapter is as follows. First, academic institutions and individual authors in two top journals, *Management Science* and *Operation Research* were analyzed descriptively and the statistical analysis focuses on a period of 20 years (1995-2014). A social network mapping and clique analysis was conducted to explore the factors driving the quantity of publications, as well as clique size and structure as factors improving productivity. New metrics were developed to measure the openness of the analyzed journals to the new authors and to verify whether the editorial board members have personal advantages in publishing their own papers.

2.1.4 Data Collection, Motivation and Contribution

Olson (2005) obtained the mean quality ratings of 33 journals across three academic disciplines. Among them *Management Science* and *Operations Research* were ranked the first and second, respectively, with the highest mean of quality. Taking into consideration their proven solid reputation, for the purposes of this study, the *Management Science* and *Operations Research* publications were collected for the period from 1995 to 2014. The following data limitations should be kept in mind. Despite the best effort to identify the authors, some of the authors names may change (e.g. their names differ before and after marriage, different versions of their names are provided in translation). School names may also change or some authors may have multiple affiliations. Finally, for those authors who passed away, their influence on all publications cannot be delimited.

This work aims to offer useful insights to new-entry authors or current Ph.D. students into how academic social networks are formed through journal publications. The descriptive analysis includes a visualization to understand the basic collaboration patterns, which is followed by an analysis of the similarities and differences between MS and OR. The social network characteristics are then studied and a clique analysis by size, structure, density, and diversity is performed to understand the factors that drive productivity. Finally, new metrics are introduced to assess the journals on their openness to new authors, and on the influence of the status of the editorial board on an authors chance of publishing a submitted article.

2.2 Descriptive Analysis

2.2.1 Descriptive Analysis of *Management Science*

2.2.1.1 Publication Overview

In this part, the co-authorship patterns are studied along with trends in research collaboration. In the period between 1995 and 2014, altogether 4,036 distinct authors contributed to MS. From the BibTex we extract 2,697 effective records and here we only focus on the 2,688 research articles.

Figure 2.1 shows the total number of authors and papers in MS for each year in the indicated period. It is interesting to note, and, perhaps, intuitively obvious, that the total number of papers and the total number of authors follow a similar trend as they increase and decrease together. Although both increased during the analyzed period, the total number of authors saw an even steeper rise. The average number of authors per paper is indicated in Figure 2.2. The same figure reveals the trend of increasing collaboration in the respective time period with the average number of authors per paper rising from around 2.02 to approximately 2.55.

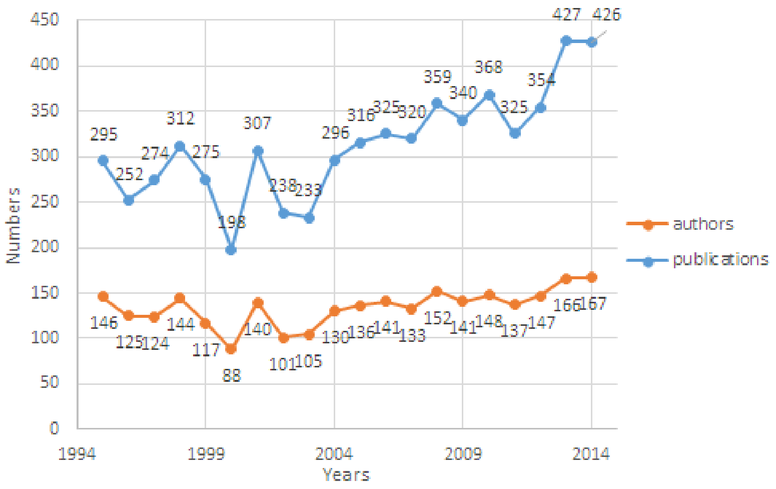


Figure 2.1: Total number of authors and publications per year (1995-2014) (MS).

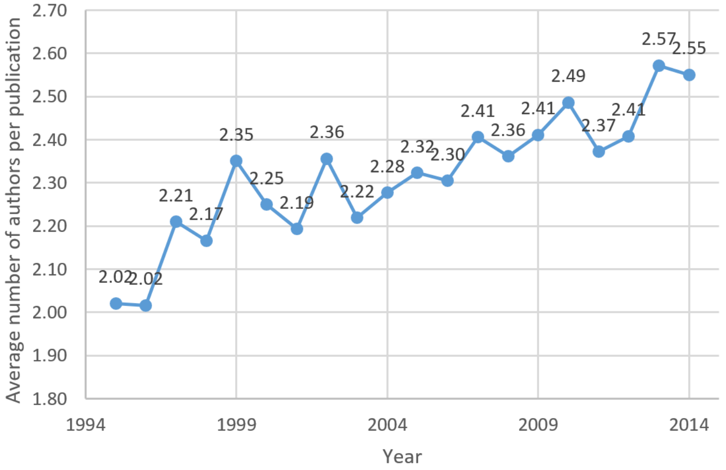


Figure 2.2: Average number of authors per publication (1995-2014) (MS).

In spite of the rising trend, it is still extremely difficult to publish in MS. Figure 2.3 illustrates the total number of publications per author. 74.5% of the authors published only one paper in the 20-year period, and 96.3% of authors published fewer than five papers. Single, two- and three-author publications respectively account for 16.0%, 46.0%, and 29.7% of all papers published. Altogether 98.6% of the papers are authored by fewer than five contributors, as Figure 2.4 shows.

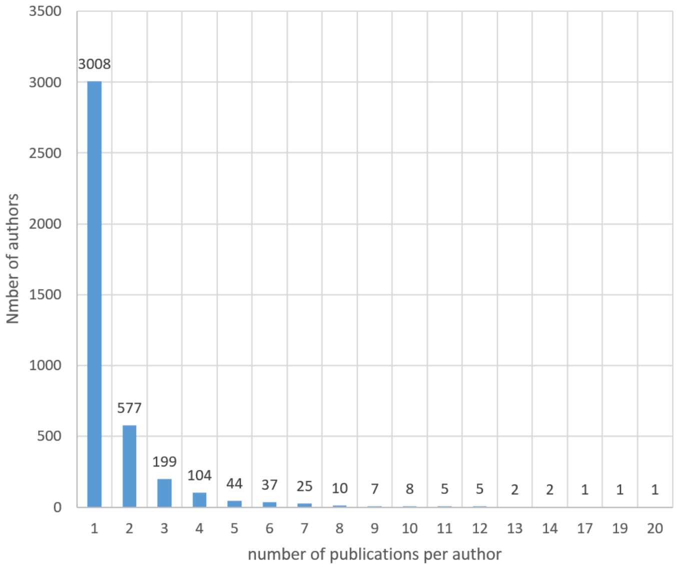


Figure 2.3: Total number of publications per author (MS).

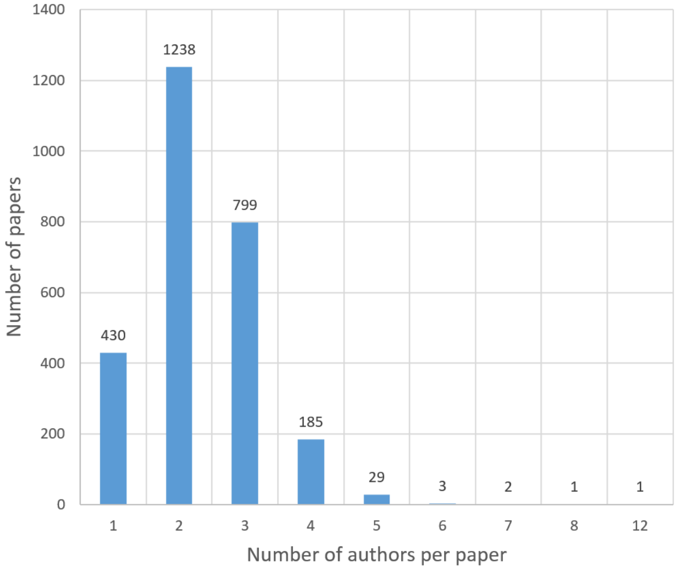


Figure 2.4: Total number of authors per publication (MS).

In what follows, the author ranking, university ranking, and region ranking of MS are analyzed for the period between 1995 and 2014. The aim of this analysis is to identify the institutions and faculty members who significantly contribute to the OM field. These results could thus be useful to Ph.D. students and new faculty members in applying for study programs or in choosing their academic careers.

2.2.1.2 Author Ranking

	1995-2014	
Rank	Author name	Number of papers
1	Teck Hua Ho	20
2	Gerard P. Cachon	19
3	Christian Terwiesch	17
4	Christopher S. Tang	14
4	Serguei Netessine	14
6	Scott Shane	13
6	Ward Whitt	13
8	Elena Katok	12
8	Izak Duenyas	12
8	Lawrence M. Wein	12
8	Noah Gans	12

Table 2.1: Authors with the highest number of publications in MS between 1995 and 2014.

The table 2.1 shows that Teck Hua Ho was the most productive author in MS between 1995 and 2014 with 20 publications. The same author has also been the chief editor of MS as of 2016. The data does not only provide a useful benchmark for individual researchers to compare their research productivity to that of the top researchers in the field, but it also gives the deserved recognition to those researchers who made notable achievements and had the greatest impact on the OM field.

2.2.1.3 University Ranking

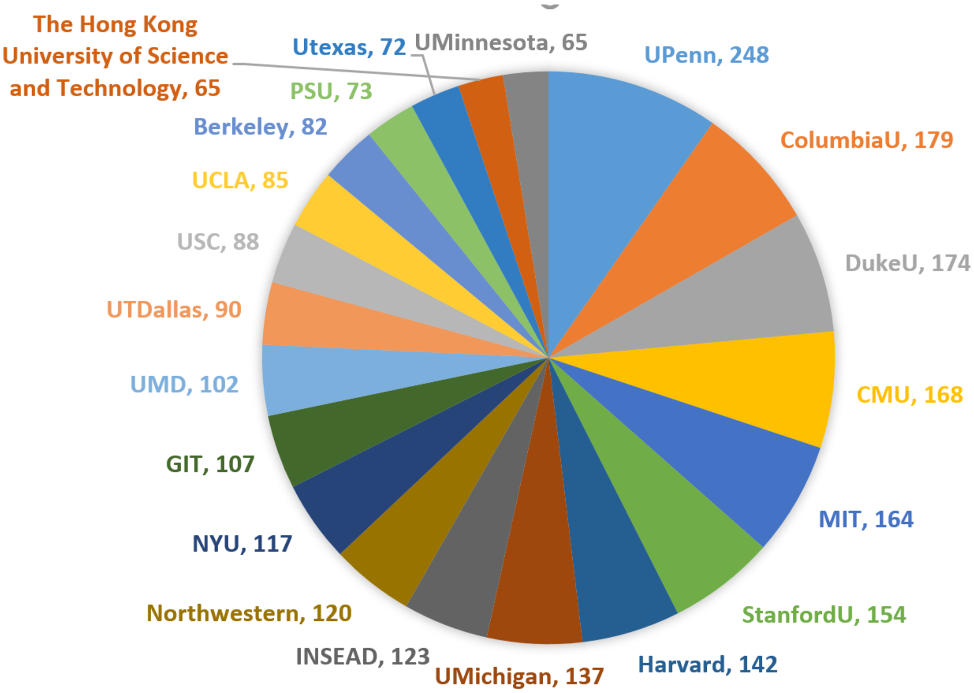


Figure 2.5: Universities with most publications in MS between 1995 and 2014.

Figure 2.5 includes the top 20 universities in terms of the number of MS publications. It is not surprising that these universities have the strongest research reputation. Wharton School at the University of Pennsylvania, for instance, ranks first, and it is followed by Columbia University, Duke Univeristy, Carnegie Mellon, MIT, and Stanford University. The faculty and research groups at these universities have a sustainable research record and a continued momentum of innovation. The rankings of the last 10 and 20 years were compared to each other. The results show that the changes among the top 10 schools are only slight. Therefore, the results imply that these universities have established highly productive long-term research capabilities in the respective field.

2.2.1.4 Region Ranking

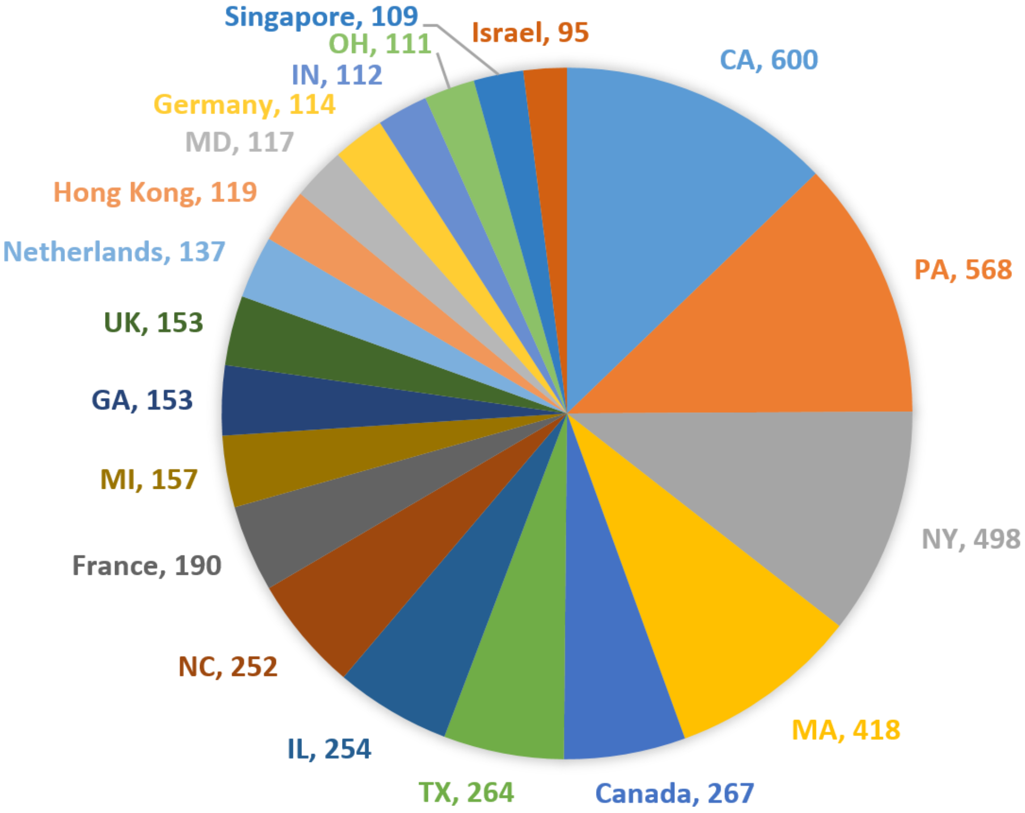


Figure 2.6: Regions with most publications in MS between 1995 and 2014.

All U.S. states and all other countries were categorized as belonging to different regions. Figure 2.6 first shows that between 1995 and 2014, 12 out of 20 top regions are located in the U.S. and eight out 10 top regions are located in the U.S. These results imply that the U.S. has a strong research presence in MS, and it is followed by Canada. Second, among the top five, the most productive regions are located on the West coast of the U.S., such as California. On the East coast, the most productive regions include New York, Pennsylvania, and Massachusetts. Third, California ranks the first in the analyzed period despite the fact that most of the traditional top universities are located in the East coast region. This may be explained by the fact that, in addition to Stanford and the University of California, the U.S. state university system also includes many outstanding scholars.

2.2.2 Descriptive Analysis of *Operations Research*

2.2.2.1 Publication Overview

There are 2,641 distinct authors who contributed to OR from 1995 to 2014. From the BibTex we extract 1,687 effective records and here we only focus on the 1,679 research articles.

Figure 2.7 shows the total number of authors and the number of papers in OR per year between 1995 and 2014. The total number of papers and authors again follow nearly the same trend. Both of them increased in the analyzed period, especially the total number of authors. A small peak is observed in 2010. Figure 2.8 illustrates the average number of authors per paper. The analyzed period is characterized by an increasing trend in collaboration, with the average number of authors per paper increasing from approximately 2.10 to 2.73, which is slightly higher than the numbers for MS.

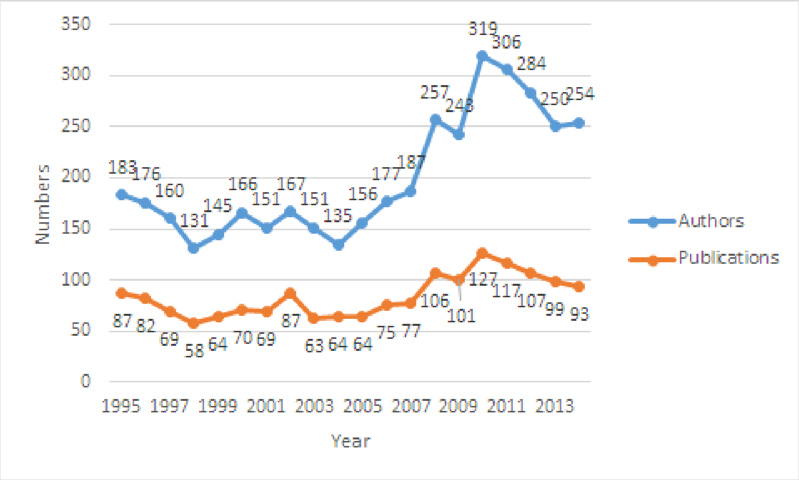


Figure 2.7: Total number of authors and publications per year between 1995 and 2014 (OR).

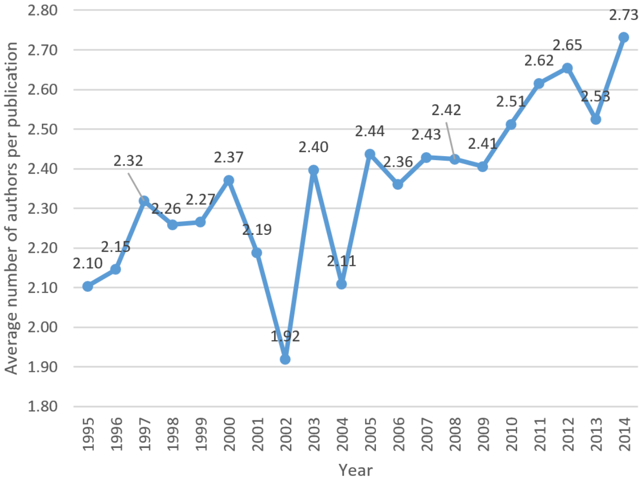


Figure 2.8: Average number of authors per publication between 1995 and 2014 (OR).

It is equally challenging to publish in OR as in the MS. Figure 2.9 shows that 73.3% of the authors only published one paper in the analyzed period, whereas 94.9% of the authors published fewer than five papers. Single, two- and three-author papers respectively account for 15.6%, 44.0%, and 29.7% of all papers published. Altogether 97.0% of the papers are authored by fewer than five contributors, as Figure 2.10 shows.

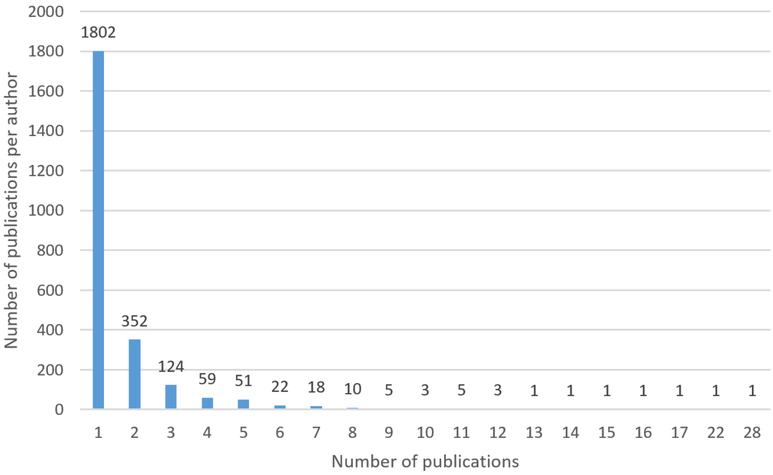


Figure 2.9: Total number of publications per author (OR).

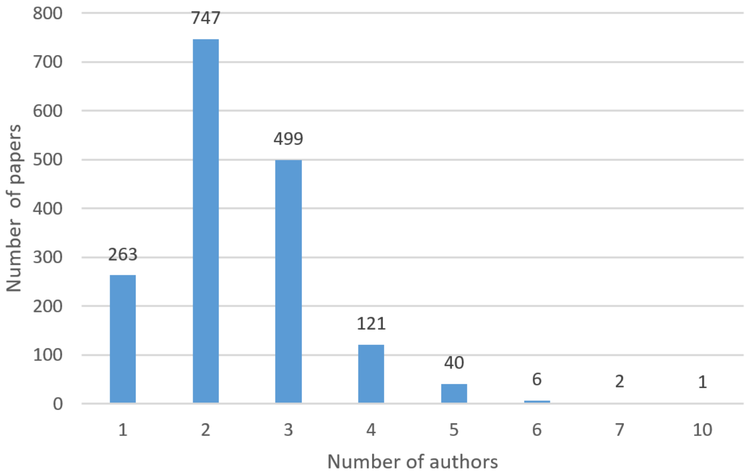


Figure 2.10: Total number of authors per publication (OR).

2.2.2.2 Author Ranking

Table 2.2 shows the 10 top authors in OR. The first-ranked author has 28 publications to his credit. Both of the top ranked authors are affiliated MIT, and their number of publications is significantly higher than that of other authors.

Rank	Author name	Number of papers
1	Dimitris Bertsimas	28
2	David Simchi-Levi	22
3	Awi Federgruen	17
4	Ward Whitt	16
5	Lawrence M. Wein	15
6	Gilbert Laporte	14
7	Jing-Sheng Song	13
8	David D. Yao	12
8	Ashish Arora	12
8	Paul Glasserman	12

Table 2.2: Authors with the highest number of publications in OR between 1995 and 2014.

2.2.2.3 University Ranking

Figure 2.11 indicates that the top three ranked universities are MIT, Columbia University, and Stanford University. All of the top 10 universities are located in the U.S. Universities renowned for engineering and technology, such as Georgia Institute of Technology and UC Berkeley, also appear in the figure. The names of universities listed in Figure 2.11 overlap to a great extent with the most represented ones in MS.

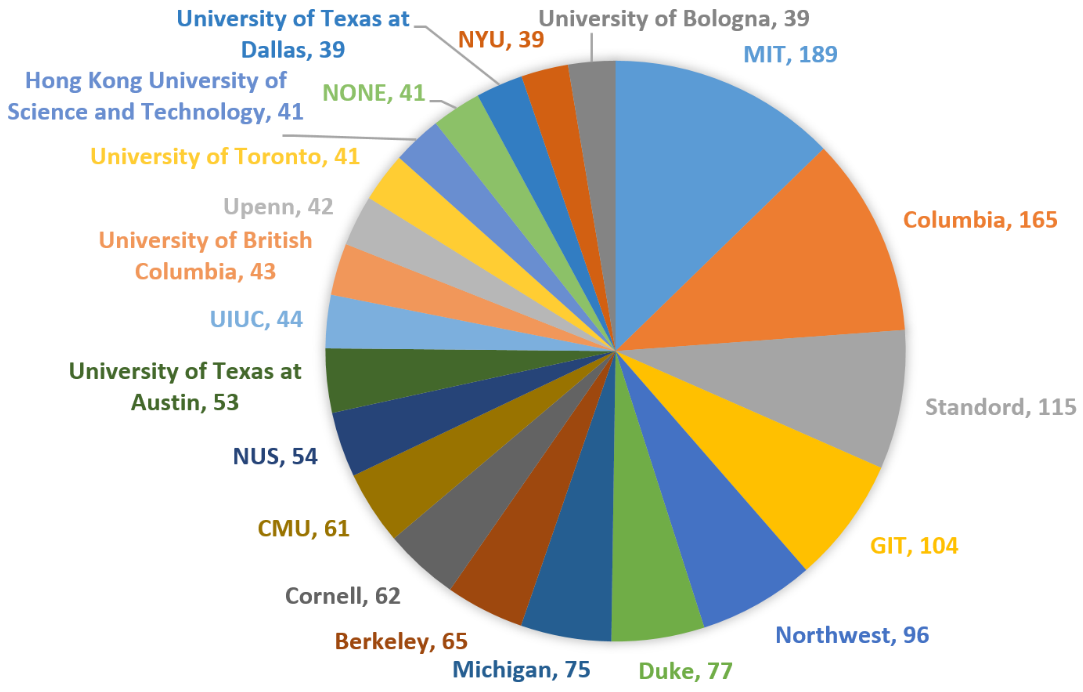


Figure 2.11: Universities with most publications in OR between 1995 and 2014.

2.2.2.4 Region Ranking

Figure 2.12 shows that the top eight out of 10 regions are located in the United States. New York State, California, and Canada are very productive with more than 300 publications.

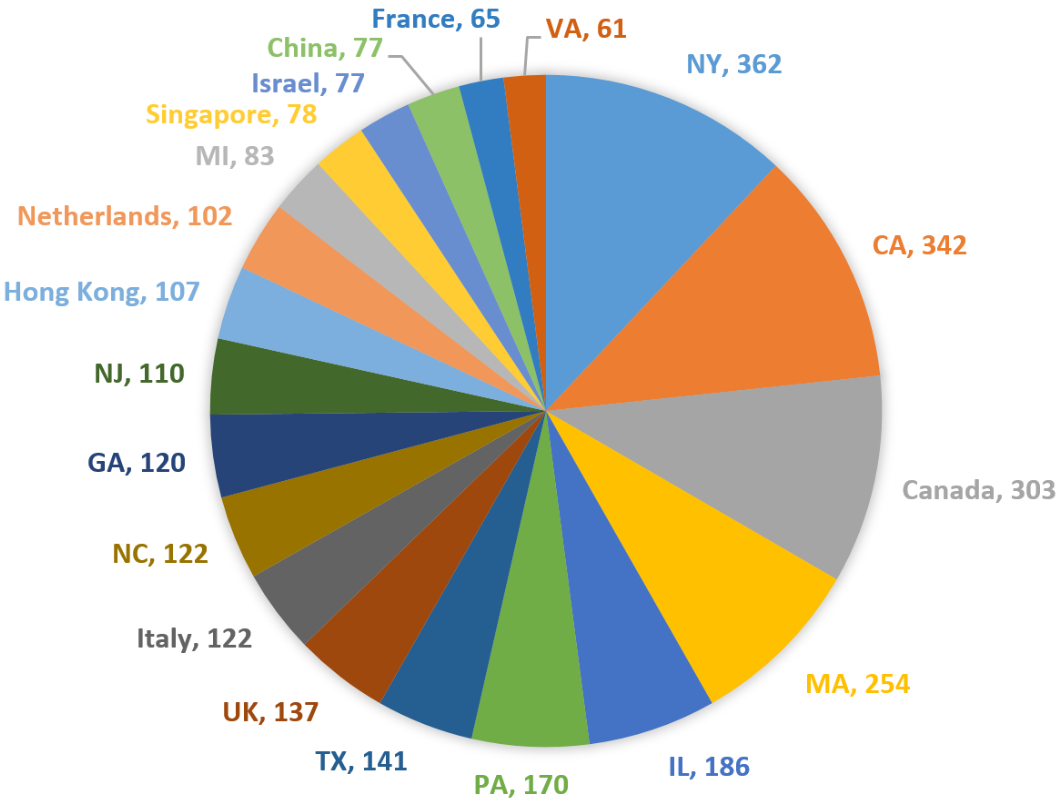


Figure 2.12: Regions with most publications in OR between 1995 and 2014.

2.2.3 Comparison of MS and OR

In comparing MS and OR, many similarities are found regarding the patterns of collaboration. For instance, it is very difficult for a scholar to publish in either of the journals. As indicated above, more than 70% of the authors have published only one paper over a 20-year period. Approximately 95% of the authors have not published more than four papers in the journals. Nearly all of the top 10 authors in terms of their number of publications are affiliated with prestigious universities whose reputation is high. Additionally, the U.S. seems to play a dominant role in the academic fields of management science and operation research.

Regarding the pattern of collaboration, the following hypothesis can be made: The number of authors per publication on MS and OR follows the same distribution (Figure 2.13).

H_0:The number of authors per publication on MS and OR follows the same distribution.

H_1: The number of authors per publication in MS and OR does not follow the same distribution.

The non-parametric Kolmogorov-Smirnov test was used to compare sample distributions. The P-value of 1 indicates that there is no evidence to reject the null hypothesis and, therefore, they two journals follow the same distribution.

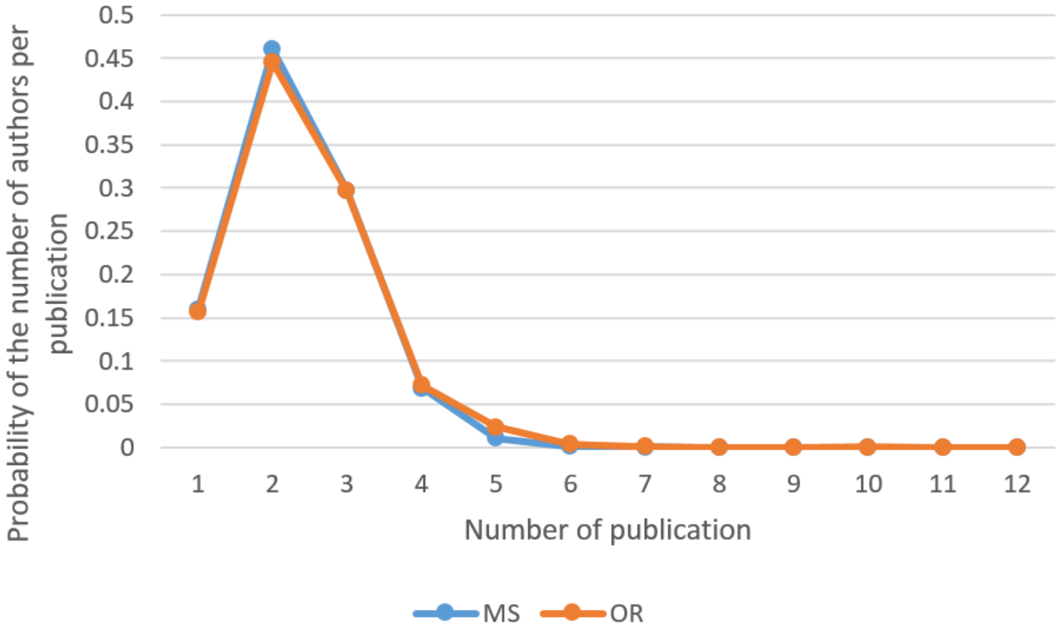


Figure 2.13: The distribution of the number of authors per publication in MS and OR.

2.3 Social Network Mapping and Clique Analysis

A co-author network may have different structures and sizes, so there are different ways of representing networks that encompass all applications. Nevertheless, there are some representations that serve to capture the features of a network. Some standard measures are represented here to analyze the co-author networks of MS and OR. Compared to macro-measurements that focus on profiling the characteristics of a network, micro-measurements used to describe the nodes and links seem more suitable. Isolated sub-networks can be defined as cliques, the basic elements of the network. In what follows, the density, diameter, degree, and betweeness will be discussed; moreover, the factors of cliques will be explored that may drive the productivity.

2.3.1 Density

Density is an indicator of how well the nodes are connected in a real network relative to the possible theoretical number of connections, defined as the number of links of the graph divided by the number of links in a complete graph with the same number of nodes. If every node is directly connected to every other node, a graph is complete. For an undirected graph G with N nodes, density D is defined as:

$$D = \frac{2 * \#L(G)}{N(N-1)}. \tag{2.1}$$

Here $L(G)$ is the link of the graph.

	MS in 20 years
# of nodes	4036
# of edges	4796
Density	0.000589

Table 2.3: Density of MS in 1995-2014

	OR in 20 years
# of nodes	2461
# of edges	3213
Density	0.00106

Table 2.4: Density of OR in 1995-2014

The density is very low, which indicates that the network is very sparse and with few connections. Therefore, the co-authorship in both MS and OR forms a very sparse network.

2.3.2 Diameter

The diameter of a graph is the length of the longest geodesic, which denotes the shortest possible line between two points on a sphere or other curved surface. If the average geodesic distance among actors is small, the information can easily reach everyone, and spread quickly and efficiently. The diameters of MS and OR are 27 and 24 respectively. This result is not surprising as both networks are sparse rather than compact.

2.3.3 Degree

The degree centrality/degree of a node is defined as the number of ties a node has (in graph-theoretical terminology: the number of edges adjacent to this node). It shows how well connected a node is. In mathematical terms, degree centrality $d(i)$ of node i is defined as:

$$d(i) = \sum_{j} m_{ij}. \tag{2.2}$$

Here m_{ij} is 1 if there is a link between node i and j and 0 if no link between each other.

Measuring the degree of each node is a simple way of determining the position of the node. For example, if a node (in a graph of n nodes) has $n - 1$ degree, then it is connected to every other node, and thus is the center of the network. The degree normally does not measure how well located a node is in a network. It is possible for a node to have few links, but have a critical position in the network. Moreover, this measure is very sensitive to a nodes marginal contribution to the network. Table 2.5 shows that Kannan Srinivasan, Ramayya Krishnan, Serguei Netessine, and Scott Shane are listed in the top four for MS.

Number	Name	Degree
1	Kannan Srinivasan	20
2	Serguei Netessine	19
2	Ramayya Krishnan	19
4	Chrysanthos Dellarocas	17
4	Christian.Terwiesch	17
4	Scott Shane	17
7	Han Bleichrodt	16
7	Lawrence M.Wein	16
7	Teck Hua Ho	16
7	Han Bleichrodt	16
7	Z.John Zhang	16

Table 2.5: Top 11 authors with the highest degrees in MS.

In the MS, 198 out of 4,036 authors publish alone, while the majority of authors have two or three co-authors (shown in Figure 2.14).

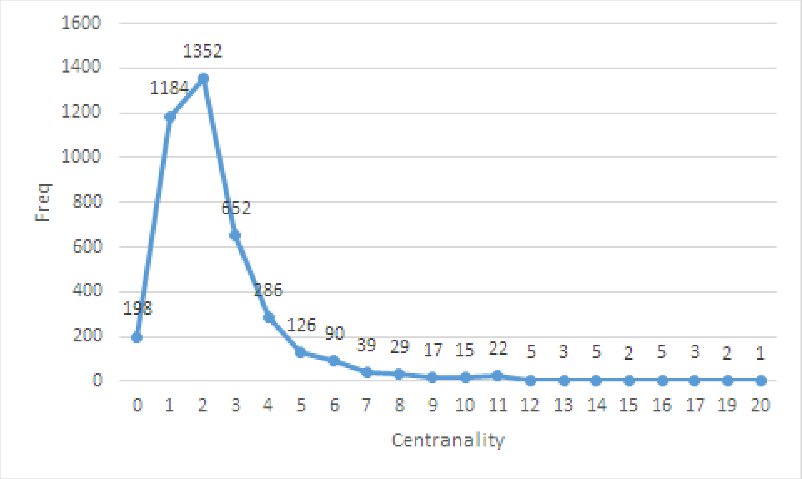


Figure 2.14: The distribution of degree centrality (MS).

Number	Name	Degree
1	Dimitris Bertsimas	37
2	David Simchi-Levi	26
3	Gilbert Laporte	18
4	James B. Orlin	17
4	Melvyn Sim	17
6	Michel Gendreau	16
6	Barry L. Nelson	16
6	Yinyu Ye	16
6	Maurice Queyranne	16
6	Lawrence M. Wein	16
6	Jing-Sheng Song	16
6	Chung-Piaw Teo	16
6	Milind Dawande	16

Table 2.6: Top 11 authors with the highest degrees in OR.

In the OR, 125 out of 2,641 authors publish alone, with the majority of authors working together with two or three co-authors on their OR publications (shown in Figure 2.15).

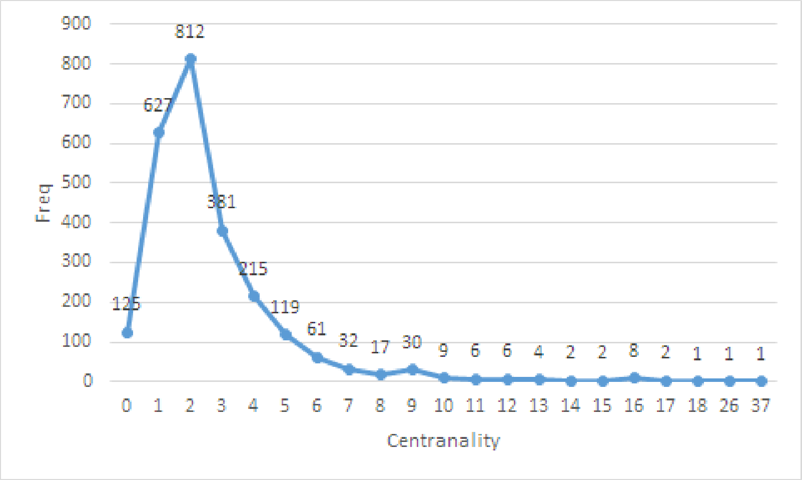


Figure 2.15: The distribution of degree centrality (OR)

The distribution of degree centrality of MS and OR are very similar. The following hypothesis was thus tested.

H_0: The distributions of degree of MS and OR follow the same distribution.

H_1: The distributions of degree of MS and OR follow the different distribution.

Due to the nature of the data, the non-parametric KS-test was used and the p-value of 1 was obtained. Therefore, it failed to reject the null hypothesis, and it was, therefore, concluded that the two journals follow the same distribution.

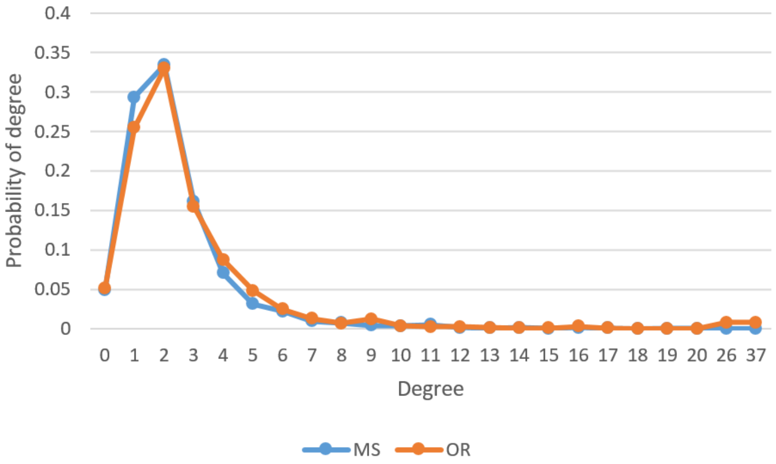


Figure 2.16: Comparing degree centrality between MS and OR.

2.3.3.1 Betweeness

The so-called betweenness centrality is an indicator of a nodes centrality in a network, which is defined as the number of shortest paths from all vertices to all others that pass through that node. A node with high betweenness centrality has a large influence on the transfer of the relationship through the network, under the assumption that the transfer follows the shortest paths studied by Brandes (2008). First, we define i and j to be nodes that subsequently form a pair (i, j). The number of shortest paths for this pair is σ_{ij}. In a mathematical expression, the betweenness of node b is denoted as $C(b)$ and it is obtained as follows:

$$C(b) = \sum_{j \neq i; i,j \neq b} \frac{\sigma_{ij}(b)}{\sigma_{ij}}, \tag{2.3}$$

where $\sigma_{ij}(b)$ is the number of the shortest paths between node i and j through b. For example (cf. Figure 2.17), the number of shortest paths between a and c is 1 ($\sigma_{ac} = 1$) and the number of shortest path between a and c through b is also 1 ($\sigma_{ac}(b) = 1$).

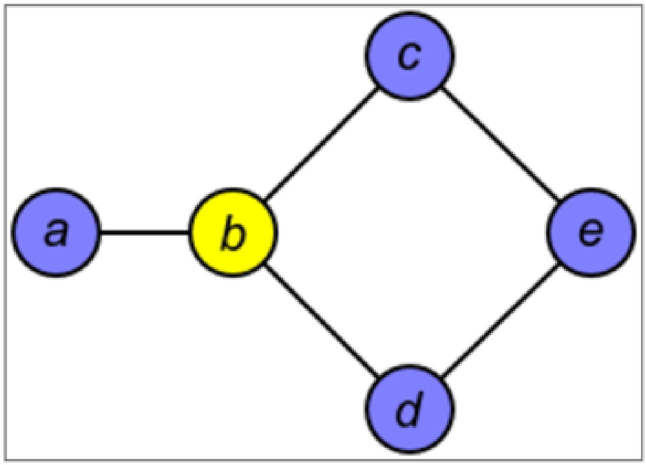


Figure 2.17: An example of Betweeness.

Tables 2.7 and 2.8 show that Teck Hua Ho ranks first for betweenness in MS, and David Simchi-Levi and Dimitris Bertsimas are ranked as top two for betweenness in OR. Authors showing high betweenness are those who may most directly help authors reach out to other authors. In other words, this process represents quantifying the control that one person exerts over the communication among other people within a social network, studied by Freeman (1977).

Order	Authors	Betweenness
1	Christian.Terwiesch	3.41E+05
2	Teck Hua Ho	1.98E+05
3	Peter C.Fishburn	1.72E+05
4	Stefan Thomke	1.67E+05
5	David E.Bell	1.66E+05
6	Eric T.Bradlow	1.24E+05
7	Morris A.Cohen	1.22E+05
8	Serguei.Netessine	1.18E+05
9	YoungHoon Park	1.14E+05
10	Vishal Gaul	8.89E+04

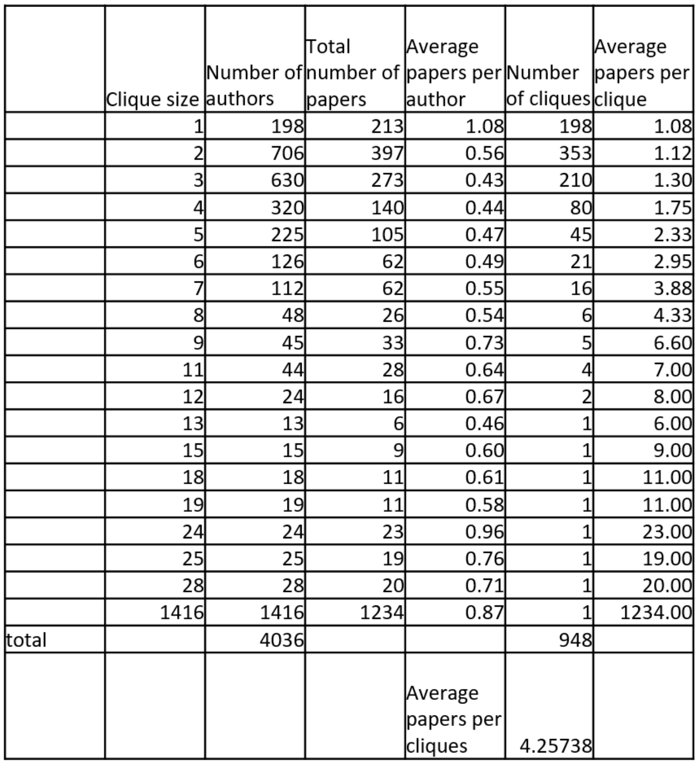
Table 2.7: Betweenness (MS)

Order	Authors	Betweenness
1	David Simchi-Levi	1.88E+05
2	Dimitris Bertsimas	1.69E+05
3	Melvyn Sim	1.42E+05
4	Awi Federgruen	1.08E+05
5	Sebastian Stiller	9.72E+04
6	Paolo Toth	9.14E+04
7	Roberto Roberti	8.21E+04
8	Gad Allon	8.08E+04
9	Chung-Piaw Teo	8.05E+04
10	Enrico Bartolini	7.52E+04

Table 2.8: Betweenness (OR)

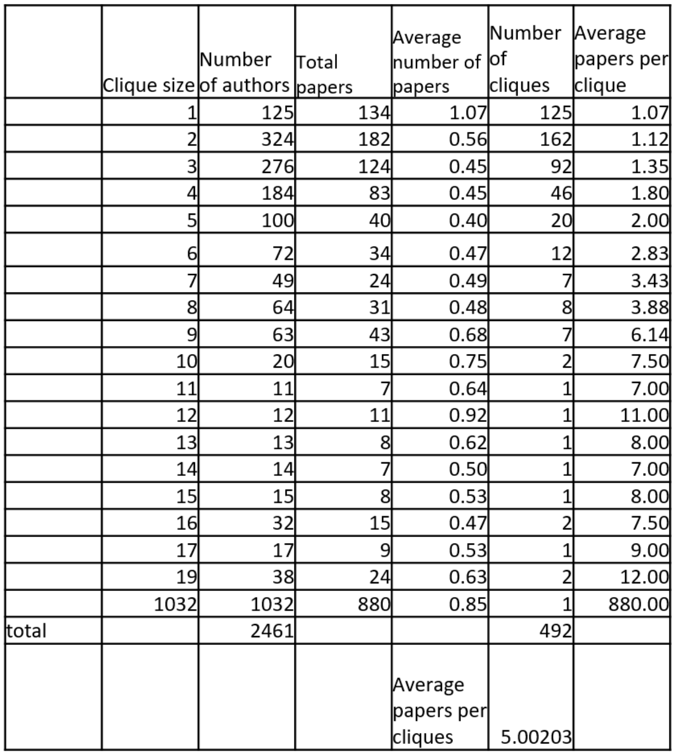
2.3.3.2 Clique Analysis

In graph theory, a clique is defined as a sub-network in which all its nodes are connected, but none of the nodes are connected to the outside network. Clique is one basic concept in graph theory and one of the most interesting topics of structural analysis which was discussed in Alba (1973). It can be important in understanding how the network is likely to behave. Figures 2.18 and 2.19 present its general profile. There is a total of 4,036 authors forming 948 cliques for MS, and 2,461 authors forming 492 cliques for OR. When the clique size equals one, it means that the respective author did not collaborate with any other authors on their MS or OR publications. The average number of papers is defined as the total number of papers over the number of authors. Most of the cliques comprise between one and three members. The largest clique includes more than 1,000 people in both MS and OR, which forms a big community and people in which can be easily introduced to each other and conduct research. The average number of papers per clique is 4.25 for MS and 5.00 for OR.



	Clique size	Number of authors	Total number of papers	Average papers per author	Number of cliques	Average papers per clique
	1	198	213	1.08	198	1.08
	2	706	397	0.56	353	1.12
	3	630	273	0.43	210	1.30
	4	320	140	0.44	80	1.75
	5	225	105	0.47	45	2.33
	6	126	62	0.49	21	2.95
	7	112	62	0.55	16	3.88
	8	48	26	0.54	6	4.33
	9	45	33	0.73	5	6.60
	11	44	28	0.64	4	7.00
	12	24	16	0.67	2	8.00
	13	13	6	0.46	1	6.00
	15	15	9	0.60	1	9.00
	18	18	11	0.61	1	11.00
	19	19	11	0.58	1	11.00
	24	24	23	0.96	1	23.00
	25	25	19	0.76	1	19.00
	28	28	20	0.71	1	20.00
	1416	1416	1234	0.87	1	1234.00
total		4036			948	
				Average papers per cliques	4.25738	

Figure 2.18: MS Clique profile.



	Clique size	Number of authors	Total papers	Average number of papers	Number of cliques	Average papers per clique
	1	125	134	1.07	125	1.07
	2	324	182	0.56	162	1.12
	3	276	124	0.45	92	1.35
	4	184	83	0.45	46	1.80
	5	100	40	0.40	20	2.00
	6	72	34	0.47	12	2.83
	7	49	24	0.49	7	3.43
	8	64	31	0.48	8	3.88
	9	63	43	0.68	7	6.14
	10	20	15	0.75	2	7.50
	11	11	7	0.64	1	7.00
	12	12	11	0.92	1	11.00
	13	13	8	0.62	1	8.00
	14	14	7	0.50	1	7.00
	15	15	8	0.53	1	8.00
	16	32	15	0.47	2	7.50
	17	17	9	0.53	1	9.00
	19	38	24	0.63	2	12.00
	1032	1032	880	0.85	1	880.00
total		2461			492	
				Average papers per cliques	5.00203	

Figure 2.19: OR clique profile.

In what follows, the focus is placed on the number of people in a clique, the total number of publications of a clique, and the average number of publications of a clique. Cliques with different average numbers of publications per author are compared, and the difference in network structures for cliques with different productivity is explored.

Figures 2.20 and 2.21 show the total and average number of papers in each clique in MS between 1995 and 2014 on the Y-axis, and the clique size on the X-axis. For example, the first number in Figure 2.20, 213, indicates that the total number of publications contributed by cliques of size 1 (single-author papers) is 213. The average number of publications per author in the cliques of each size reflects the average productivity of the cliques of a particular size.

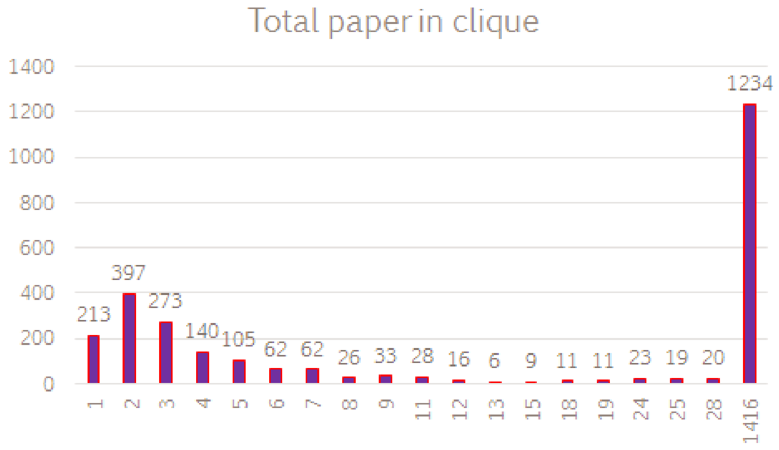


Figure 2.20: Total number of papers for cliques of different sizes in MS between 1995 and 2014.

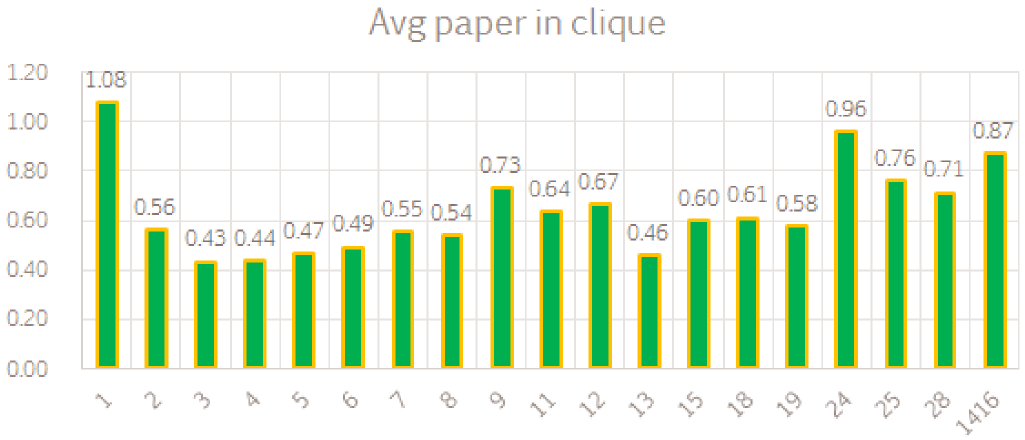


Figure 2.21: Average number of papers for cliques of different sizes in MS between 1995 and 2014.

Figure 2.20 shows that, except for the largest clique, the majority of authors published in MS in groups of 2 or 3 or worked individually. It needs to be noted that some cliques have a much higher average productivity than others; for instance, clique-13 has an average of 0.46 papers, which is very low when compared to 0.96 of clique-24 (cf. Figure 2.21). The network structure of these cliques is explored to showcase their difference in network topology.

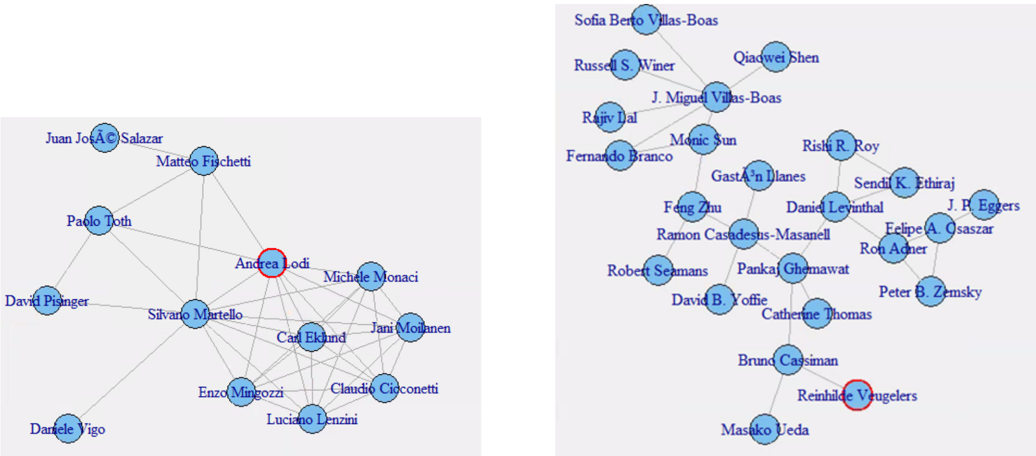


Figure 2.22: Typology comparison between clique-13 and 24.

In Figure 2.22, the figure on the right shows that the most efficient clique is the one including 24 people, whereas nearly all the collaboration took place between two people in all but two cases. The average number of collaborators is 2.04, which supports the previous observation that the most common number of collaborators is either two or three. In the figure on the left, we could see one complete graph with eight members inside which probably hold back the productivity of the clique. A similar study on OR reported a similar pattern, that is, excluding the largest clique, the majority of scholars published in OR in groups of two or three members, or worked individually. Clique-16 has an average number of 0.47 papers in this clique, which is very low when compared to 0.92 of clique-12 (cf. Figure 2.23 and 2.24). The topological structure of a clique in a graph shows that clique-12 has a more star structure or three-person collaborations than the clique-17, which involves several complete sub-networks (cf. Figure 2.25 and 2.26).

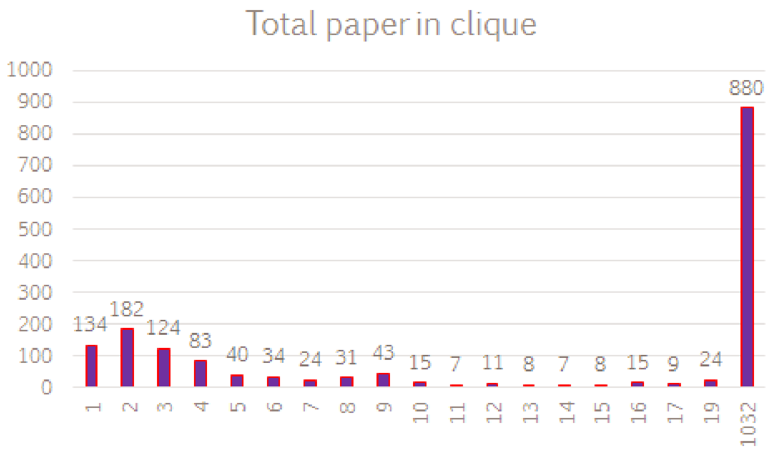


Figure 2.23: Total number of papers in cliques of different sizes in OR between 1995 and 2014.

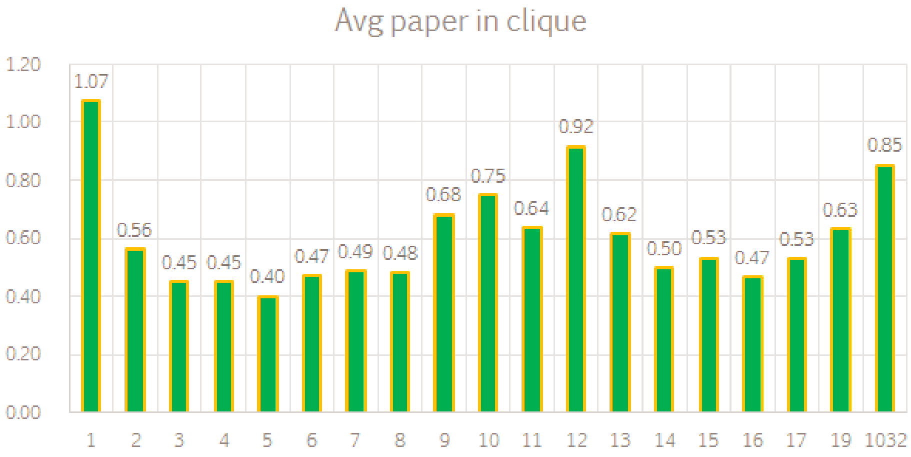


Figure 2.24: Average number of papers in cliques of different sizes in OR between 1995 and 2014.

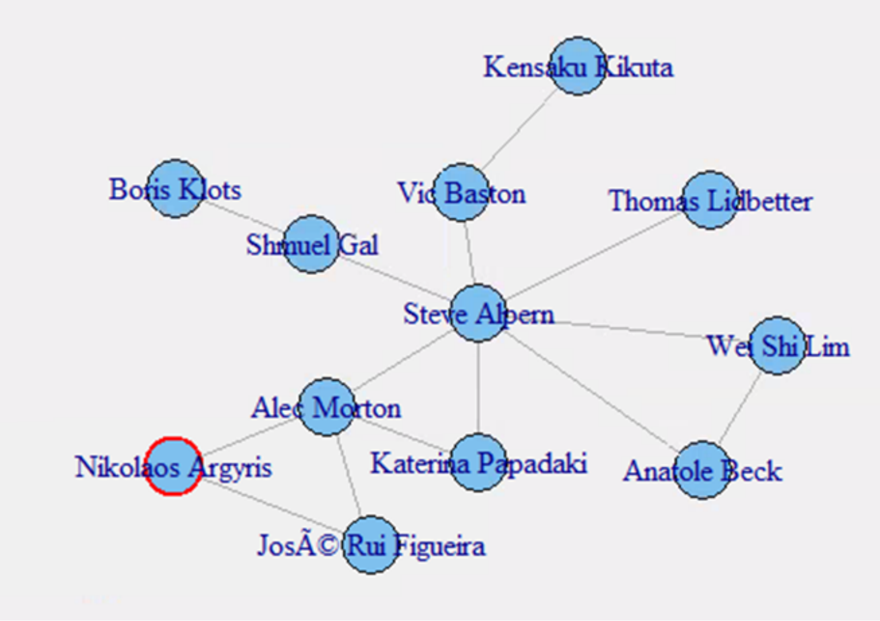


Figure 2.25: Topology of the clique-12 for OR.

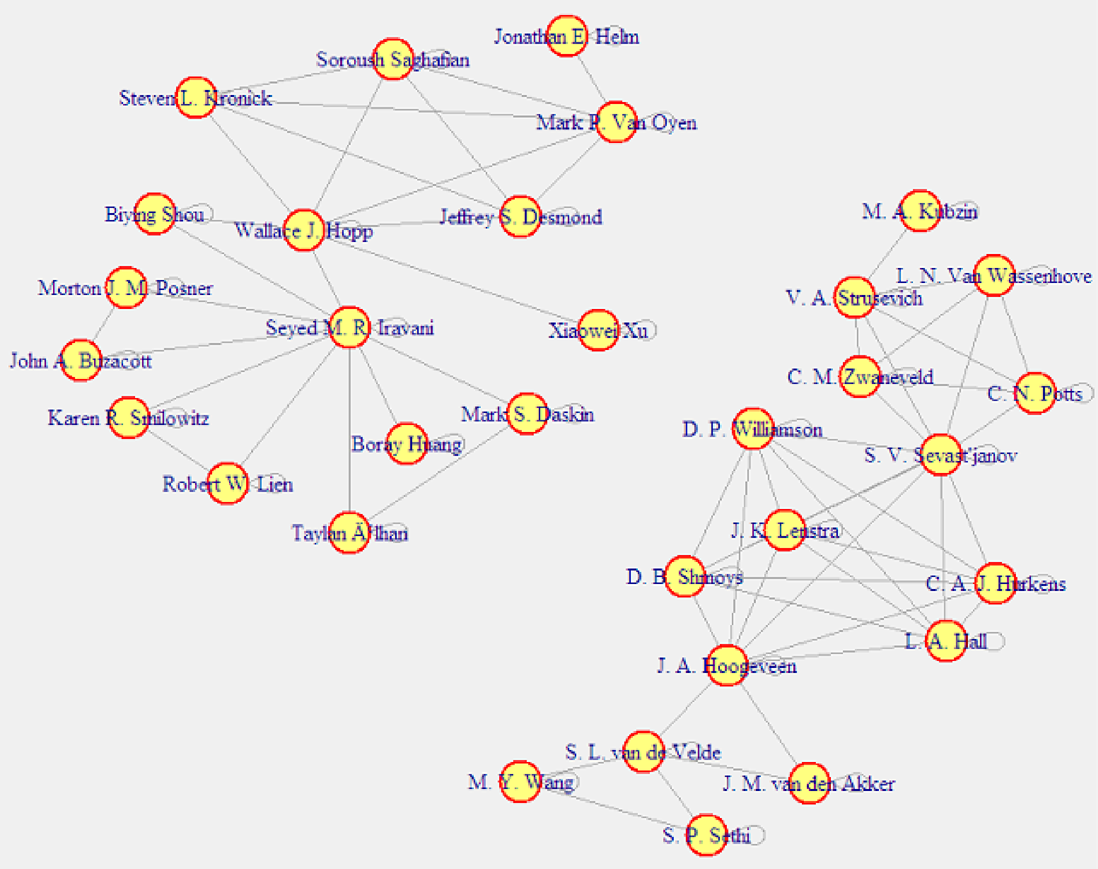
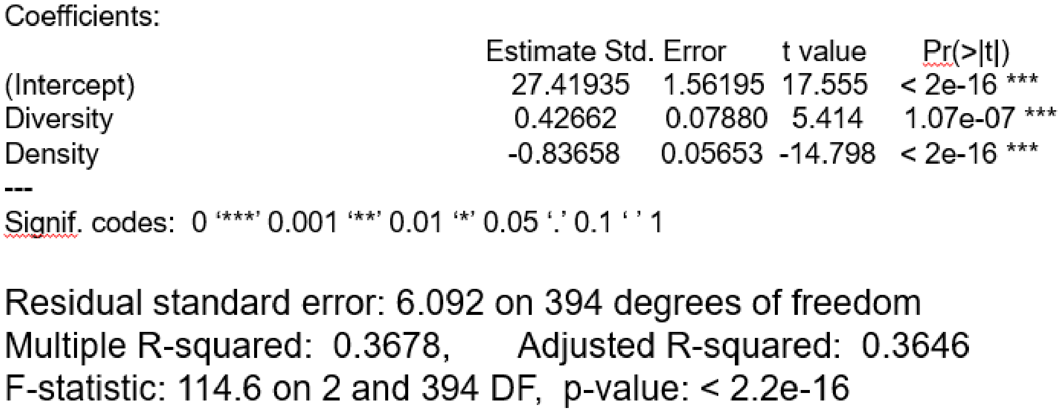


Figure 2.26: Topology of the clique-16 for OR.

Based on the previous analysis, the diversity of a clique may have a positive impact on its productivity, while the density of the clique may have a negative impact on it. We use linear regression here to explore their correlation by defining the dependent variable to be productivity, defined as the ratio of total publication to clique size. The independent variables are diversity, defined as the ratio of the number of regions and schools (of the authors) to clique size and density, defined as the ratio of the actual number of links to the number of all possible links. The regression model is as follows:

$$Productivity = \alpha_0 + \beta_o(diversity) + \beta_1(density). \tag{2.4}$$

After removing cliques with sizes of one or two (because of their constant density), the p-values of the coefficient for both density and diversity are far below the common 0.05 significance level. This thus indicates a strong correlation between the diversity (density) and productivity (cf. Figure 2.27). Furthermore, the coefficient indicates that for each additional diversity, the productivity can be expected to increase by an average of 0.42. For every additional density, you can expect the productivity to decrease by an average of 0.83. To check multi-colinearity, we find that the correlation co-efficent between density and diversity is 0.41.



```
Coefficients:
                              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)                   27.41935   1.56195  17.555   < 2e-16 ***
Diversity                      0.42662   0.07880   5.414    1.07e-07 ***
Density                       -0.83658   0.05653 -14.798   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.092 on 394 degrees of freedom
Multiple R-squared:  0.3678,      Adjusted R-squared:  0.3646
F-statistic: 114.6 on 2 and 394 DF,  p-value: < 2.2e-16
```

Figure 2.27: Linear regression results for MS.

The results of the residual analysis are presented, which indicate that there are certain patterns in the residuals around the critical point of 290 (shown in Figure 2.28).

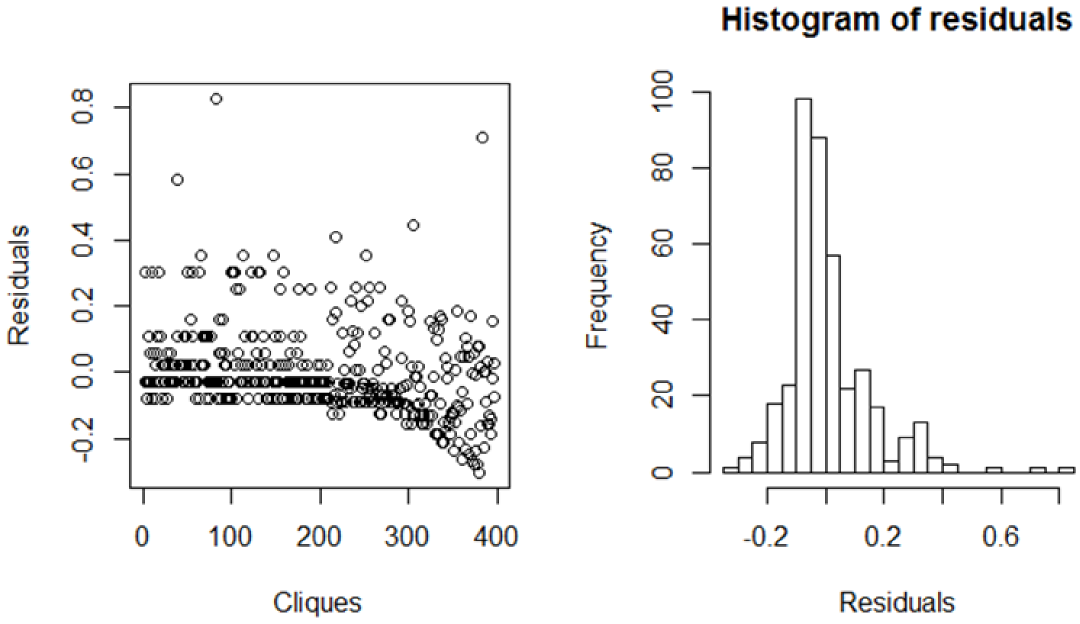
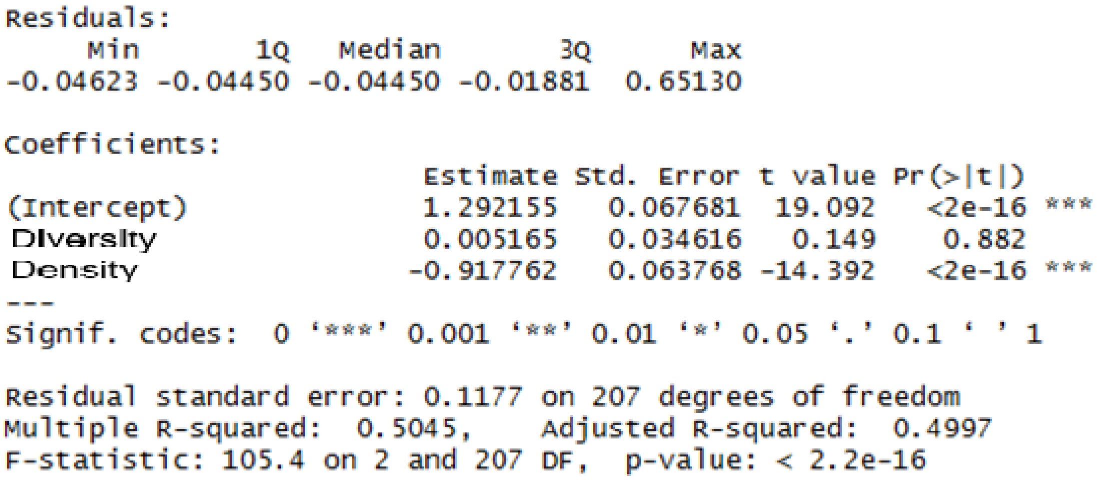


Figure 2.28: Residual analysis of linear regression results for MS.

Therefore, piece-wise linear regression was used upon dividing the dataset into three parts. For the first 210 cliques (their sizes are three), the regression results are shown in Figure 2.29 and 2.30. In this model, the p-value of density is far below the significance level of 0.05, which indicates a statistically significant impact of density on productivity. R-square increases to 0.50. However, the diversity is not significant, with the p-value of 0.88. Every additional density is expected to decrease productivity by an average of 0.91.



```
Residuals:
     Min       1Q   Median       3Q      Max
-0.04623 -0.04450 -0.04450 -0.01881  0.65130

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.292155   0.067681  19.092   <2e-16 ***
 Diversity                0.005165   0.034616   0.149    0.882
 Density                 -0.917762   0.063768 -14.392   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1177 on 207 degrees of freedom
Multiple R-squared:  0.5045,    Adjusted R-squared:  0.4997
F-statistic: 105.4 on 2 and 207 DF,  p-value: < 2.2e-16
```

Figure 2.29: Linear regression on cliques with size equal to three for MS..

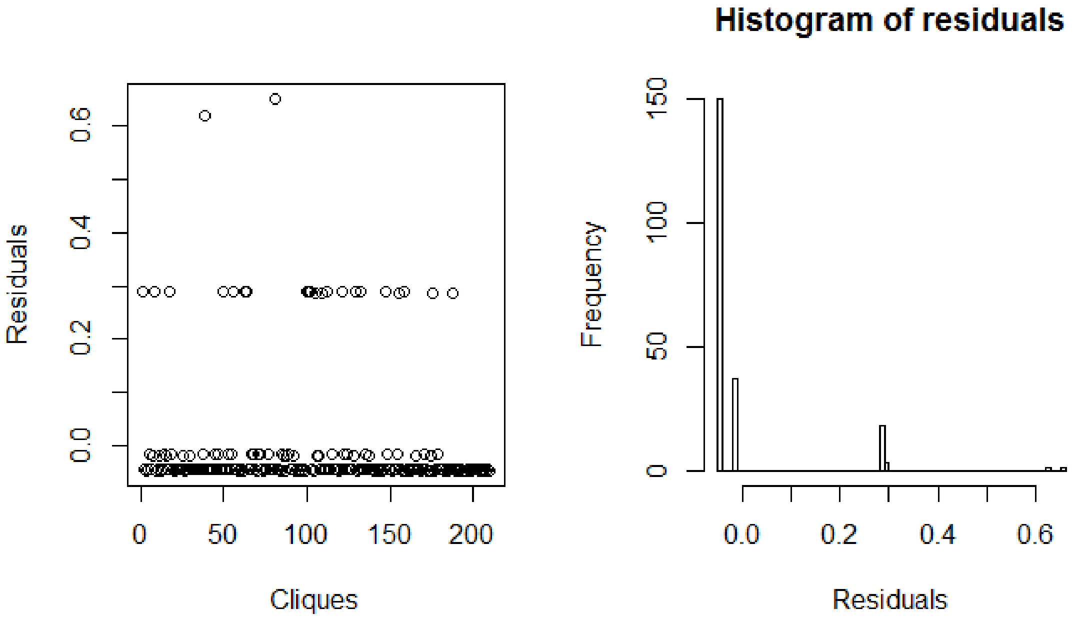
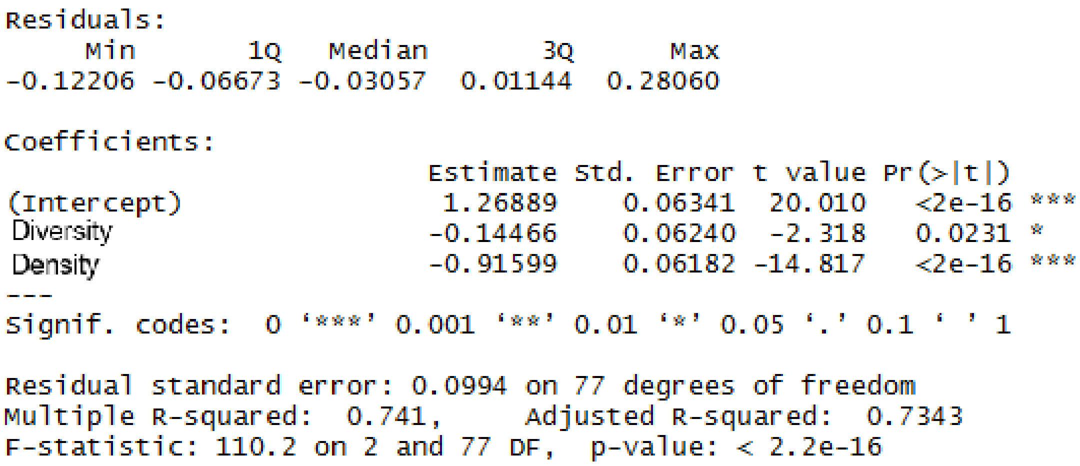


Figure 2.30: Residual analysis of linear regression on cliques with size equal to three for MS.

The second part of data corresponds to cliques sized four, numbered from 211 to 290 in our dataset. The regression results are shown in Figure 2.31 and 2.32. In this model, both the diversity and density are statistically significant at a 0.05 level. R-square increases to 0.74, offering quite a good explanation for the variations in the data.

The model indicates that for each additional diversity (or density), the productivity can be expected to decrease by an average of 0.14 (or 0.91, respectively).



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.12206 -0.06673 -0.03057  0.01144  0.28060 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   1.26889    0.06341  20.010   <2e-16 ***
Diversity                    -0.14466    0.06240  -2.318   0.0231 *  
Density                      -0.91599    0.06182 -14.817   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0994 on 77 degrees of freedom
Multiple R-squared:  0.741,	Adjusted R-squared:  0.7343 
F-statistic: 110.2 on 2 and 77 DF,  p-value: < 2.2e-16
```

Figure 2.31: Linear regression on cliques with size equal to four for MS.

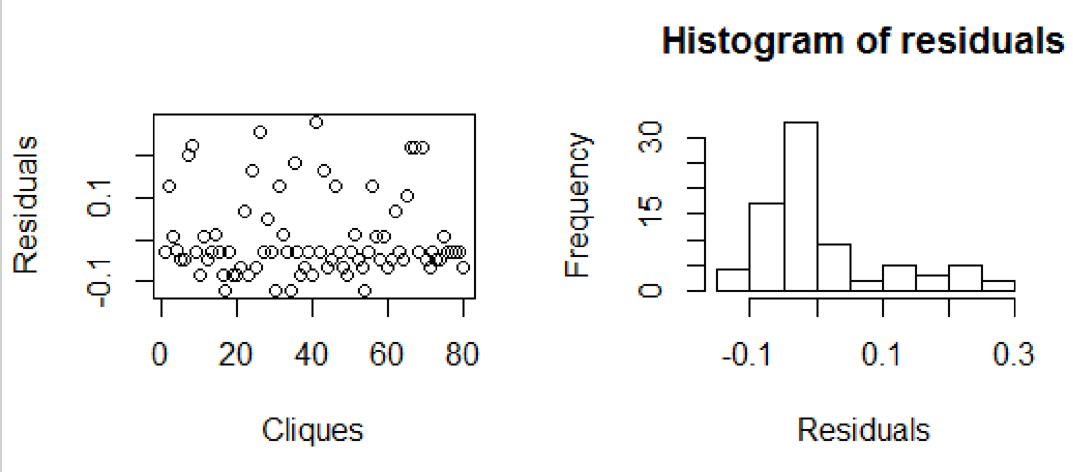
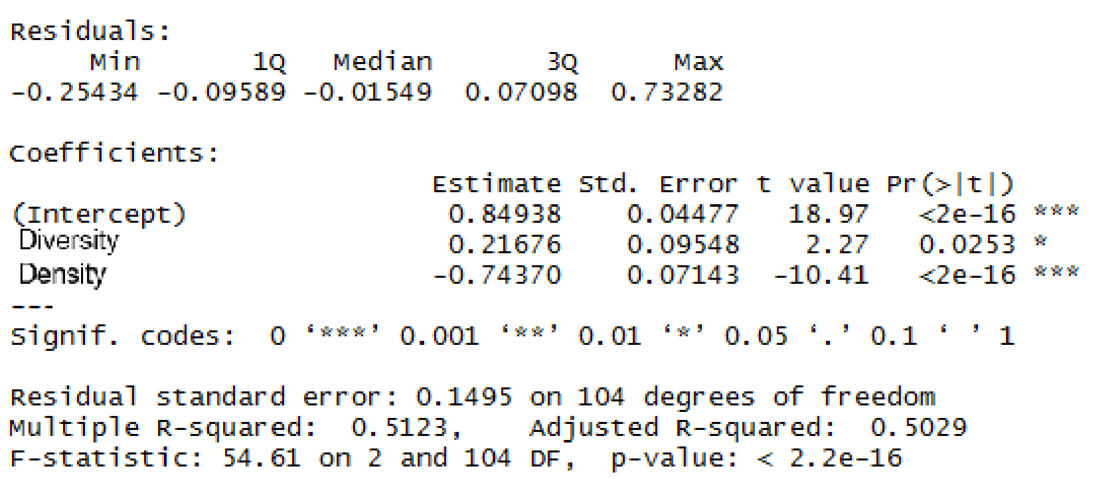


Figure 2.32: Residual analysis of linear regression of cliques with size equal to four for MS.

The last part of the data corresponds to cliques sized five and above, numbered from 291 to 397 in the dataset. The linear regression models are shown in Figure 2.33 and 2.34. In this model, both the diversity and density are statistically significant at the 0.05 level. The R-square is 0.51, offering quite a good explanation for the variations in the data. The coefficients indicate that with every additional diversity, the productivity can be expected to increase by 0.21; for every additional density, however, the productivity decreases by 0.74.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.25434 -0.09589 -0.01549  0.07098  0.73282 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                0.84938    0.04477   18.97   <2e-16 ***
 Diversity                 0.21676    0.09548    2.27   0.0253 *  
 Density                  -0.74370    0.07143  -10.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1495 on 104 degrees of freedom
Multiple R-squared:  0.5123,    Adjusted R-squared:  0.5029 
F-statistic: 54.61 on 2 and 104 DF,  p-value: < 2.2e-16
```

Figure 2.33: Linear regression of cliques with size greater than four for MS.

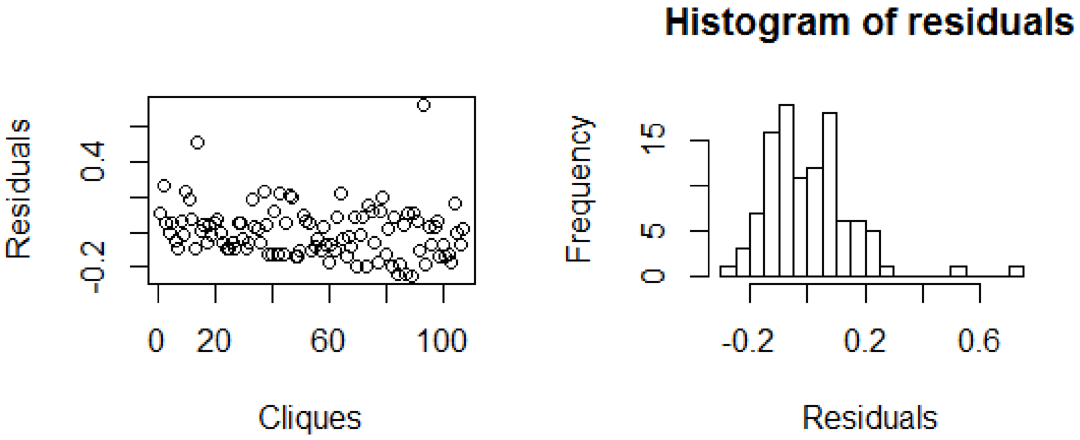


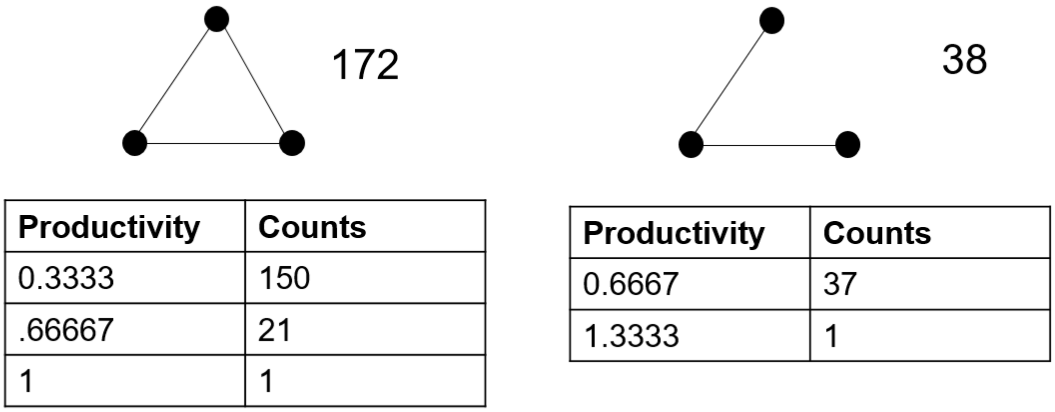
Figure 2.34: Residual analysis of linear regression of cliques with size greater than four for MS.

The impact of clique structure is now explored for cliques sized three and four. Define $S(i, j)$ to be cliques with i nodes and j links. For instance, $S(3, 3)$ refers to cliques with three nodes and three links (a complete graph of three nodes) (cf. Figure 2.35). Figure 2.35 also shows that there are 172 $S(3, 3)$ cliques, and 38 $S(3, 2)$ cliques. In comparing their productivities, it can be hypothesized that the more complete graphs are less productive.

$H_0 : Pr(P_{S(3,3)} \leq p) < Pr(P_{S(3,2)} \leq p)$;

$H_1 : Pr(P_{S(3,3)} \leq p) \geq Pr(P_{S(3,2)} \leq p)$;

where $P_{S(3,3)}$ is the productivity of cliques $S(3,3)$. The p-value of 0 means that the null hypothesis can be rejected, and that thus cliques $S(3,2)$ are stochastically more productive than cliques $S(3,3)$ (cf. Figure 2.36). This is consistent with the previous linear regression results: a higher density correlates with lower productivity.



Productivity	Counts
0.3333	150
.66667	21
1	1

Productivity	Counts
0.6667	37
1.3333	1

Figure 2.35: Topology and productivity of cliques with size equal to three for MS.

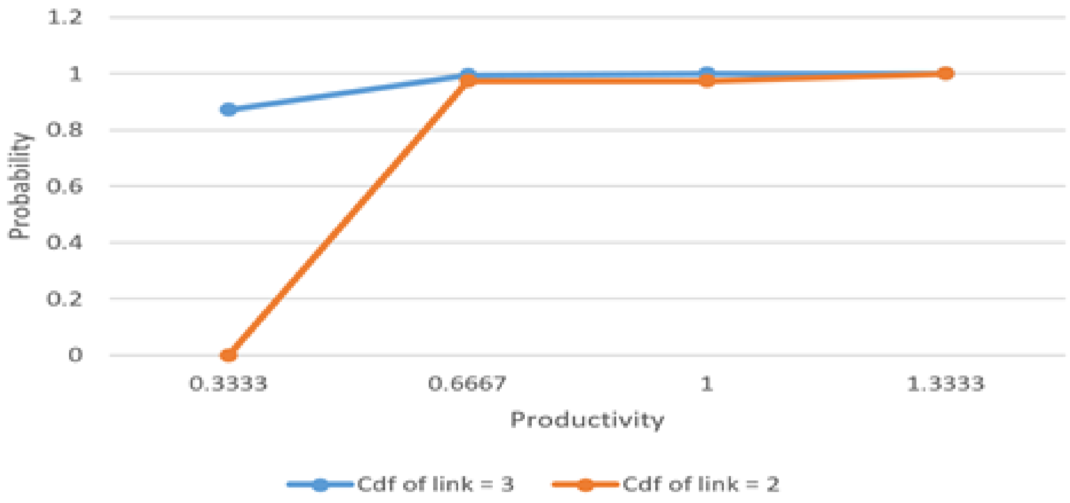


Figure 2.36: The productivity of link = two or three in cliques with size equal to three for MS.

For cliques with size equal to four, there are nine different possible structures as Figure 2.37 shows. In exploring whether more links imply lower productivity, cliques with three and four links can be combined as well as cliques with five and six links.

$H_0 : Pr(P_{S(4,(3,4))} \leq p) < Pr(P_{S(4,(5,6))} \leq p)$;

$H_1 : Pr(P_{S(4,(3,4))} \leq p) \geq Pr(P_{S(4,(5,6))} \leq p)$;

The p-value of 0 means that the null hypothesis can be rejected, and that cliques $S(4,(5,6))$ are thus stochastically more productive than $S(4,(3,4))$ (cf. Figure 2.38). Therefore, more links may not always lead to lower productivity.

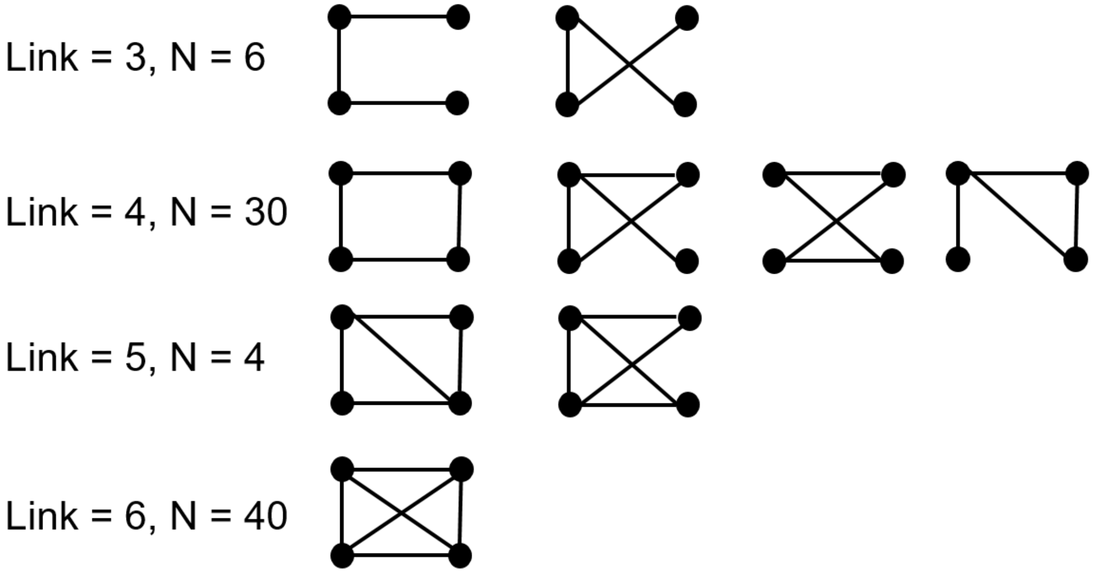


Figure 2.37: Topology and productivity of cliques with size equal to four for MS.

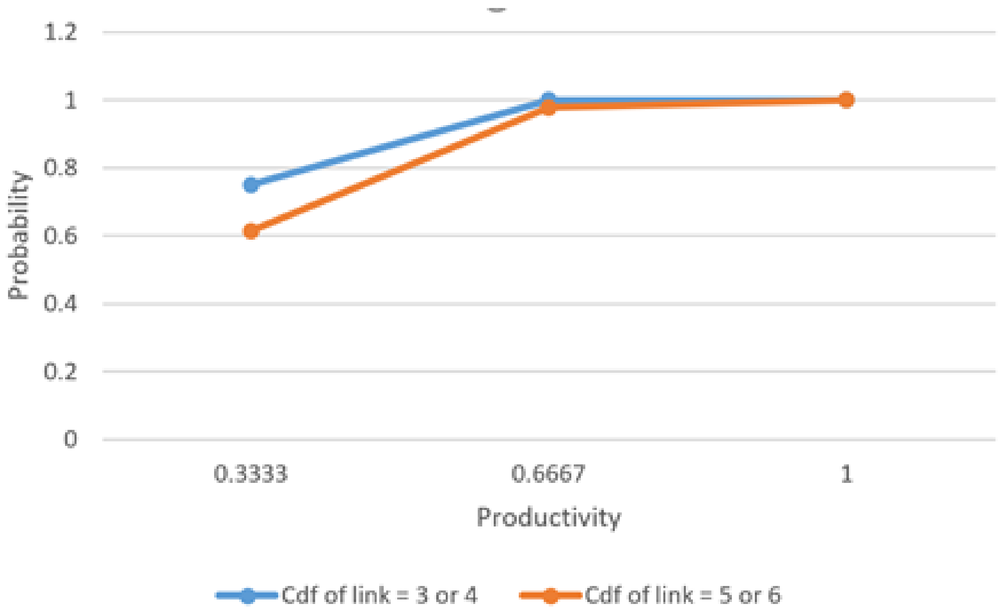
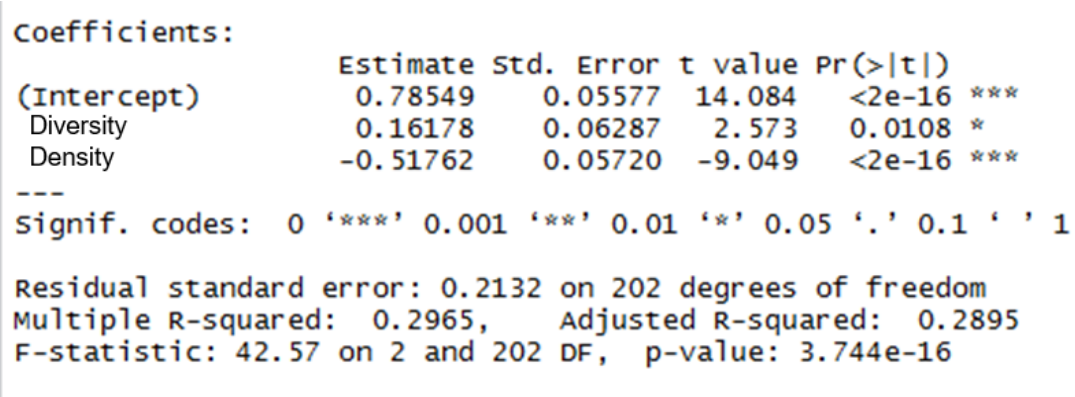


Figure 2.38: The productivity of link = three or four and link = five or six in cliques with size equal to four for MS.

The same analysis is applied to OR. As Figure 2.39 shows, the p-value of density is far below the significance level of 0.05. The null hypothesis that the coefficient is equal to zero can thus be rejected. The coefficient indicates that for every additional density (diversity) you can expect the productivity to increase by an average of 0.51(0.16). The correlation between density and diversity is 0.08.

The correlation between density and diversity is 0.08.



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.78549    0.05577  14.084   <2e-16 ***
 Diversity      0.16178    0.06287   2.573   0.0108 *
 Density       -0.51762    0.05720  -9.049   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2132 on 202 degrees of freedom
Multiple R-squared:  0.2965,    Adjusted R-squared:  0.2895
F-statistic: 42.57 on 2 and 202 DF,  p-value: 3.744e-16
```

Figure 2.39: Linear regression results for OR.

The residual analysis is presented and we found there are certain patterns (cf. Figure 2.40). The cliques dataset is divided according to clique size and a piece-wise linear regression is conducted.

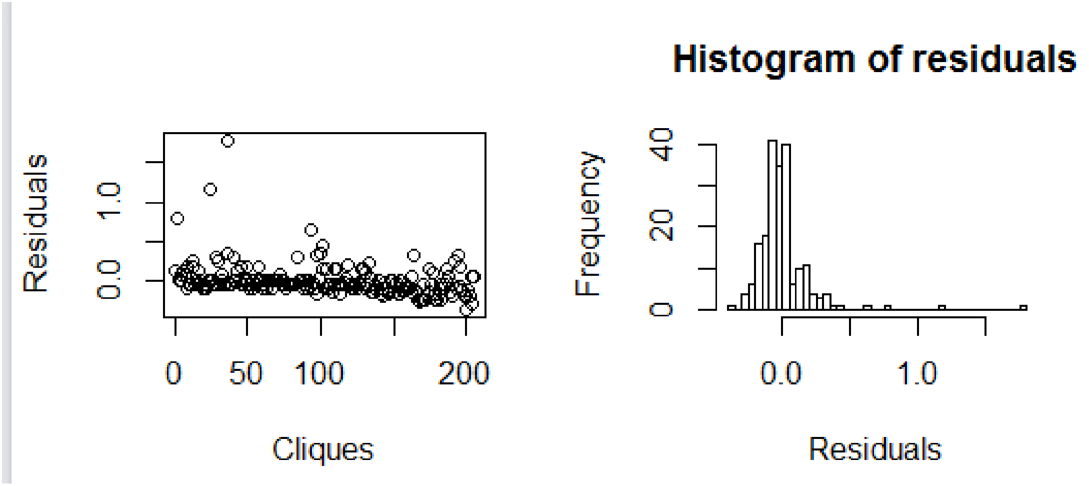
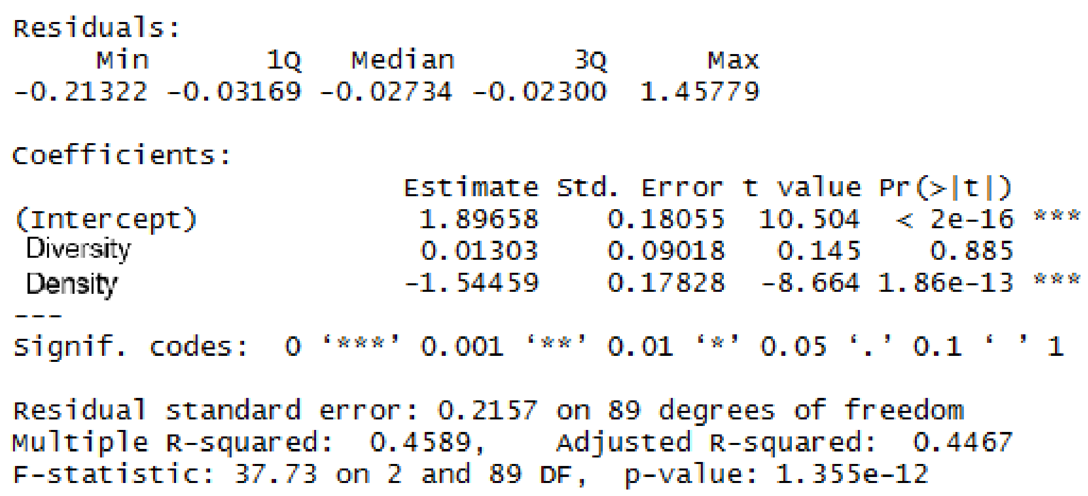


Figure 2.40: Residual analysis of linear regression results for OR

We split the dataset into three parts. For the first 92 cliques, their clique size is three. The regression results are shown in Figures 2.41 and 2.42. In this model, the p-value of density is far below 0.05, which indicates that the coefficient is statistically significantly different from zero. R-square is 0.45.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.21322 -0.03169 -0.02734 -0.02300  1.45779 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               1.89658    0.18055  10.504  < 2e-16 ***
 Diversity                0.01303    0.09018   0.145    0.885    
 Density                 -1.54459    0.17828  -8.664 1.86e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2157 on 89 degrees of freedom
Multiple R-squared:  0.4589,	Adjusted R-squared:  0.4467 
F-statistic: 37.73 on 2 and 89 DF,  p-value: 1.355e-12
```

Figure 2.41: Linear regression of cliques with size equal to three for OR.

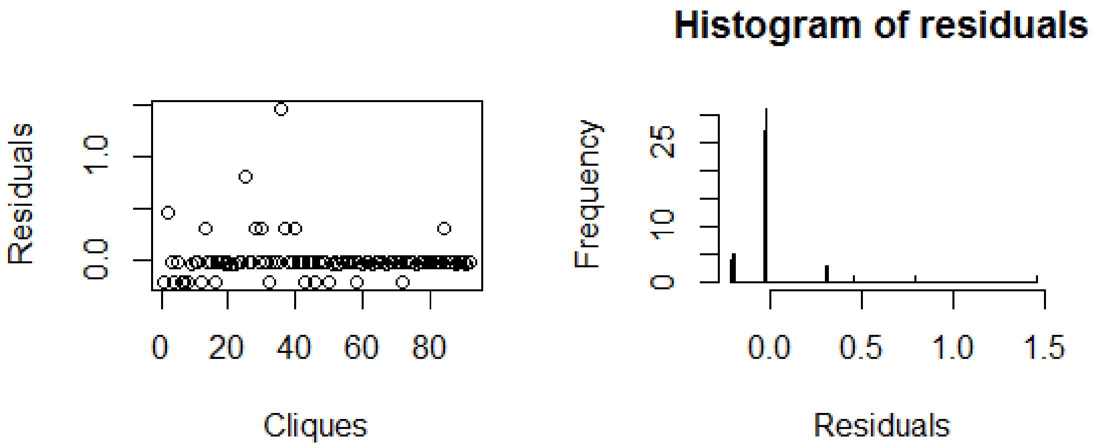
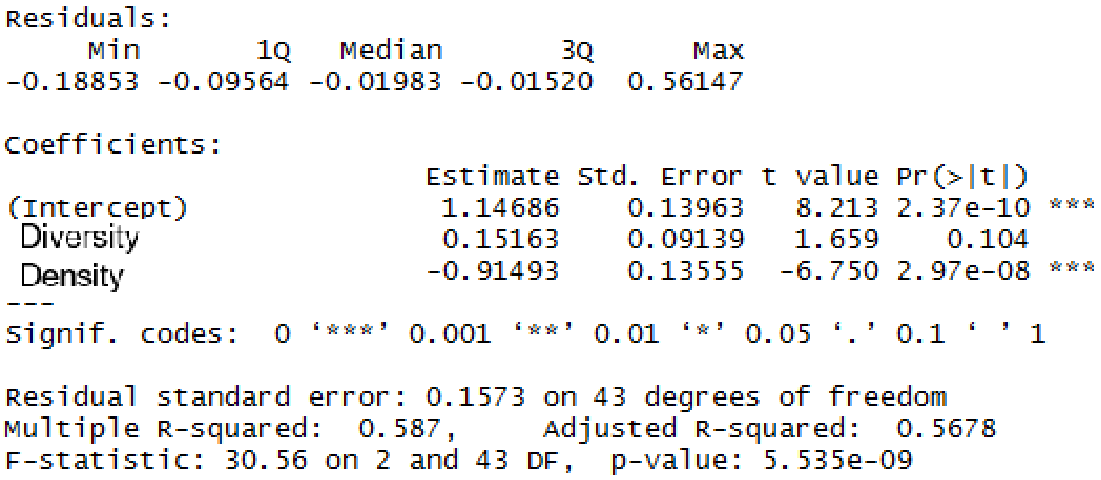


Figure 2.42: Residual analysis of linear regression of cliques with size equal to three for OR.

The second part of data contained cliques numbered from 93 to 138, and their clique size is four. The regression results are shown in Figures 2.43 and 2.44. In this model, the p-value of density is far below 0.05, which indicates that the density is statistically significant. R-square increases to 0.58, offering quite a good explanation for the variation in the data.



```
Residuals:
     Min       1Q   Median       3Q      Max
-0.18853 -0.09564 -0.01983 -0.01520  0.56147

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.14686    0.13963   8.213 2.37e-10 ***
 Diversity                0.15163    0.09139   1.659    0.104
 Density                 -0.91493    0.13555  -6.750 2.97e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1573 on 43 degrees of freedom
Multiple R-squared:  0.587,     Adjusted R-squared:  0.5678
F-statistic: 30.56 on 2 and 43 DF,  p-value: 5.535e-09
```

Figure 2.43: Linear regression of cliques with size equal to 4 for OR.

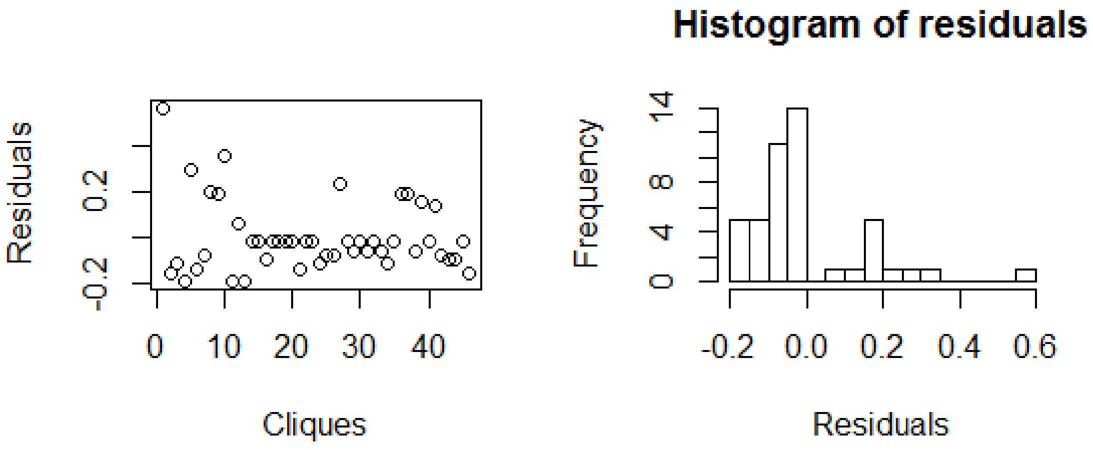
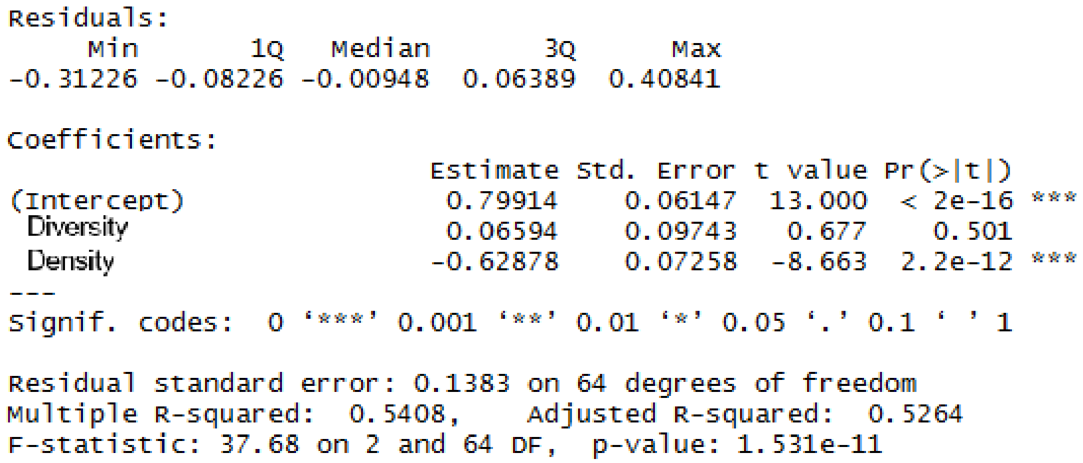


Figure 2.44: Residual analysis of linear regression of cliques with size equal to four for OR.

For the last part from 291 to 397 cliques, their clique sizes are greater than 4. Our regression results are shown in Figures 2.45 and 2.46. Similarly, the density is statistically significant but the diversity is not.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.31226 -0.08226 -0.00948  0.06389  0.40841 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                0.79914    0.06147  13.000  < 2e-16 ***
 Diversity                 0.06594    0.09743   0.677    0.501    
 Density                  -0.62878    0.07258  -8.663  2.2e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.1383 on 64 degrees of freedom
Multiple R-squared:  0.5408,    Adjusted R-squared:  0.5264 
F-statistic: 37.68 on 2 and 64 DF,  p-value: 1.531e-11
```

Figure 2.45: Linear regression of cliques with size greater than four for OR.

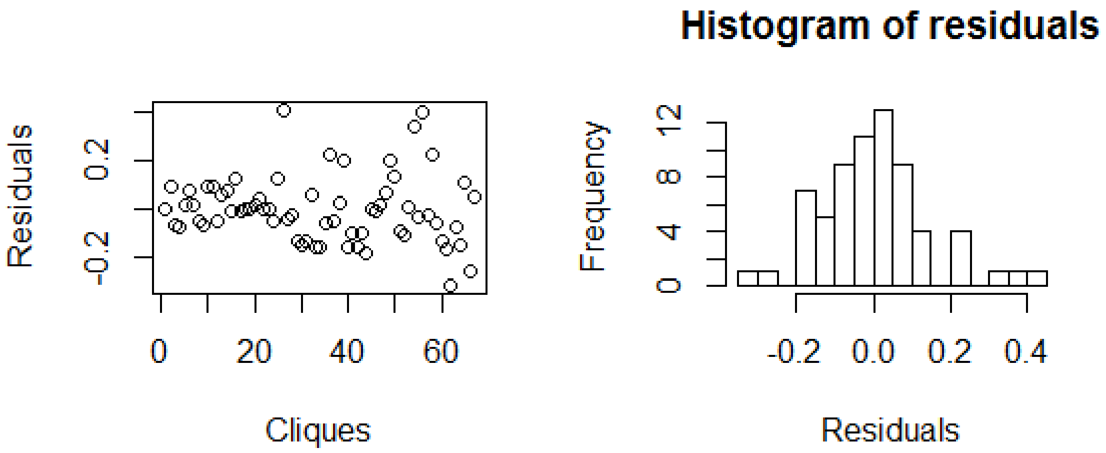


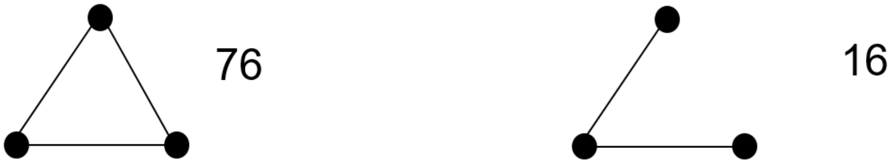
Figure 2.46: Residual analysis of linear regression of cliques with size greater than four for OR.

Similar to MS, the impact of clique structure on their productivity is studied. As Figure 2.47 shows, for cliques with three nodes, the number of $S(3,3)$ cliques is 76, while the number of $S(3,2)$ cliques is 16. Their productivity is then compared (cf. Figure 2.47). The hypothesis is that the more complete graphs are less productive. That is,

$H_0 : Pr(P_{S(3,3)} \leq p) < Pr(P_{S(3,2)} \leq p)$;

$H_1 : Pr(P_{S(3,3)} \leq p) \geq Pr(P_{S(3,2)} \leq p)$;

The p-value of 0 means that the null hypothesis can be rejected, which thus implies that cliques $S(3,2)$ are stochastically more productive than cliques $S(3,3)$.



productivity	Counts
0.3333	70
.66667	6

productivity	Counts
0.6667	13
1.3333	1
1.6667	1
2.3333	1

Figure 2.47: Topology and productivity of cliques with size equal to three for OR.

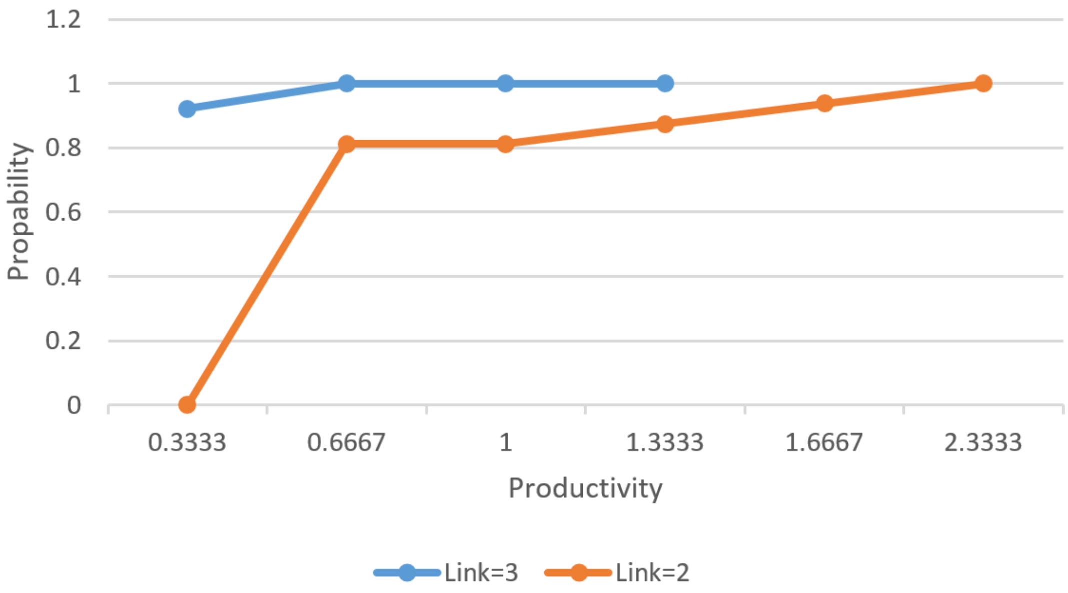


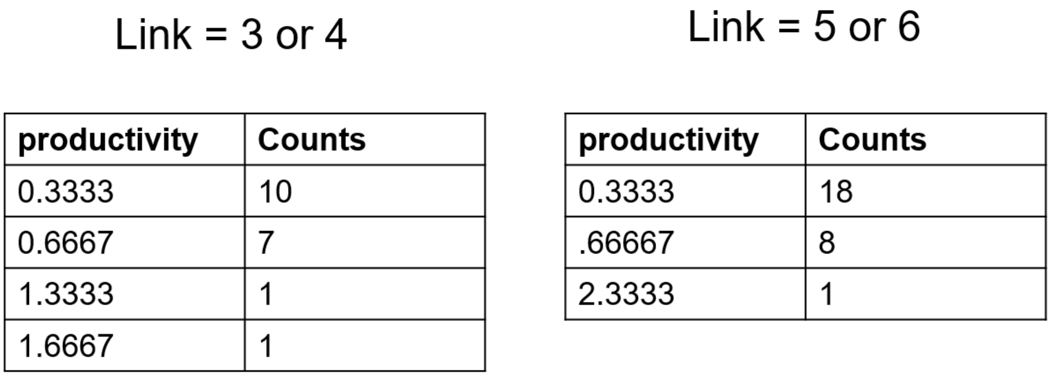
Figure 2.48: The productivity of link = two or three in cliques with size equal to three for OR.

For cliques with size equal to four, there are nine different possible structures shown in Figure 2.37. The following hypothesis are introduced:

$H_0 : Pr(P_{S(4,(5,6))} \leq p) < Pr(P_{S(4,(3,4))} \leq p);$

$H_1 : Pr(P_{S(4,(5,6))} \leq p) \geq Pr(P_{S(4,(3,4))} \leq p);$

The p-value of 0 indicated that the null hypothesis can be rejected, which means that cliques $S(4,(3,4))$ are stochastically more productive than $S(4,(5,6))$.



Link = 3 or 4

productivity	Counts
0.3333	10
0.6667	7
1.3333	1
1.6667	1

Link = 5 or 6

productivity	Counts
0.3333	18
.66667	8
2.3333	1

Figure 2.49: Topology and productivity of cliques with size equal to four for OR.

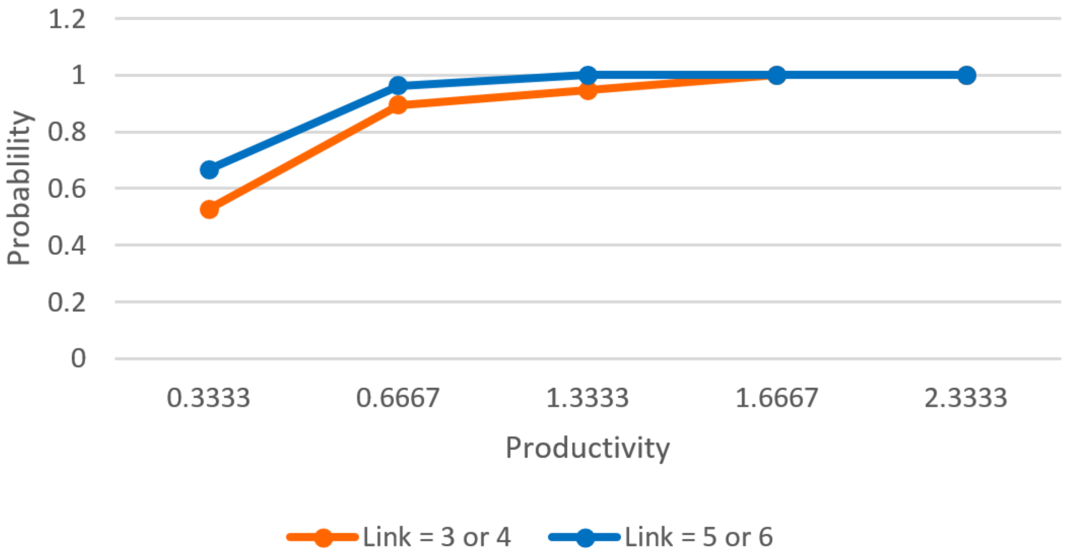


Figure 2.50: The productivity of link = three or four and link = five or six in cliques with size equal to four for OR.

2.4 New Metrics Assessment

Peer review plays an important role in academic journals studied by Armstrong (1997). As gatekeepers, peers or colleagues are requested to evaluate papers and recommend only those that meet the highest scientific standards. There are points of criticism, however, related to the process studied by Bornmann and Daniel (2004). First, reviewers seldom agree with each other concerning their recommendations and thus the reliability of the peer review process can be poor. Second, strong personal bias may result from the reviewers own research and preferences. Ideally, the reviewers recommendations should be based on the papers rather than on the authors who wrote them. Thus, the openness, how welcome the journals to new authors and the influence of the journals editorial board are important indicators of the innovativeness of a journal. The fairness of journals and the advantage of the status of the editorial board may affect the journals ability to achieve the aim of selecting the best studies for publication, and are thus often questioned. In this chapter, two questions are raised, namely, (1) Openness: Are new authors welcomed to publish in the respective journals? (2) Does the editorial board enjoy an advantage in the publication process?

2.4.1 Openness

Transparency and openness are widely recognized as vital features of scientific advances. The lack of openness and transparency means that the publications are constrained to closed cliques where authors outside of the cliques have no chance of publishing their work in the journal, and, thus, the community fails to leverage the entire accumulation of available scientific knowledge and pursue new ideas emerging in the respective fields. This is particularly true for social sciences, or the so-called soft sciences. It can prove to be difficult to judge the work by disciplinary norms and values. Unfortunately, there are no uniform and standardized means of aligning individual and mutual incentives through universal scientific policies and procedures. Editors of journals introduce different standards for researchers. Although standardizing the input may be difficult, the outcome can certainly be measured. In particular, openness and willingness to accept papers from authors who are not a part of an existing network (new authors) can be a strong indicator of the journals innovativeness and unbiased personal preferences. It is strongly linked to the journals willingness to consider new ideas and judge the contributions by the quality of the work rather than by their authors. The fact that an author submitting an article is a new contributor may add to the journals innovation . Consider the complexity, one might expect these to be routine in daily practice. Thus promoting scientific practices toward greater openness requires complementary and coordinated efforts from all stakeholders. In what follows, the development of metrics for measuring the openness of the journals to new authors is described. A journals annual openness index is defined as the percentage of papers contributed by new authors in that year, whereas new authors are defined as those who did not publish a paper in that journal in the past five years. A paper is included in the data only if all of its authors are new.

Figure 2.51 illustrates the annual openness indices for both MS and OR during the past 20 years. The index of MS is relative stable with an average of approximately 47%, while the index of OR fluctuated more than MS with an average of around 45%.

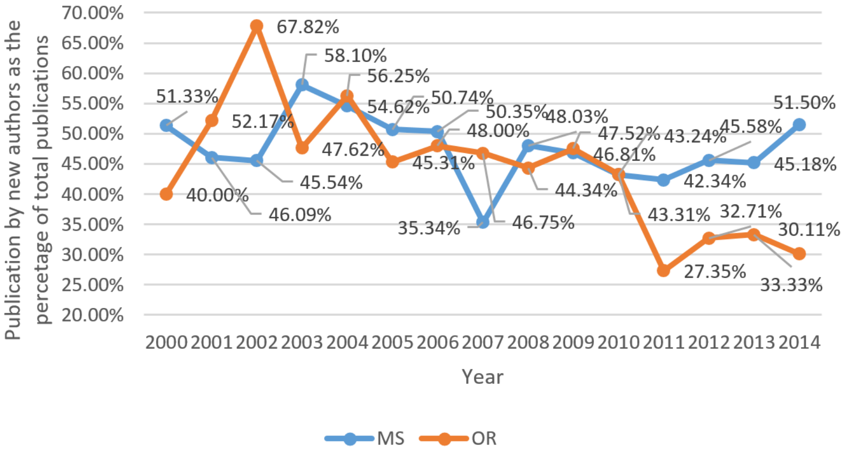


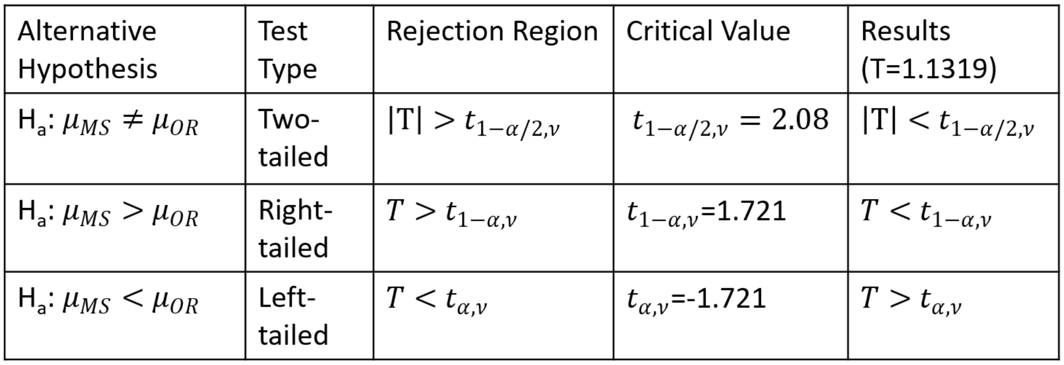
Figure 2.51: Annual openness indices (Annual openness indices (percentage of papers contributed by new authors) in MS/OR between 2000 and 2014.

The t-test is then applied to compare the mean of the openness index between MS and OR to explore whether their openness indices are equal. The mean indices of MS and OR are μ_{MS}=47.65% and μ_{OR}=44.17%. The variances are σ_{MS}=5.47% and σ_{OR}=10.56%. The hypothesis is as follows:

$H_0 : \mu_{MS} = \mu_{OR}$.

$H_a : \mu_{MS} \neq \mu_{OR}$.

First, we find that the openness indices of MS and OR are independent over time. Second, seeing that the datasets are very small (15 samples), normality needs to be assessed. Shapiro-Wilk tests were applied and the p-values were obtained for MS and OR of 0.877 and 0.582 respectively. The null hypothesis thus cannot be rejected, which implies that the data is distributed normally. The T-test is applied to check the null hypothesis H_0. The results of the test are T=1.1319 and p-value=0.2704. Therefore, the null hypothesis cannot be rejected, as the mean of MS and OR are the same. Furthermore, the T-test shows that the single-tailed null hypotheses that the index of MS is greater (smaller) than that of OR cannot be rejected. The alternative hypotheses and test results are shown in Figure 2.52. For all three alternative hypothesis, t-value does not fall into the rejection regions, and the null hypothesis, therefore, cannot be rejected, which implies that for both two-tailed and one-tailed tests, the means of the annual openness indices of MS and OR are the same.



Alternative Hypothesis	Test Type	Rejection Region	Critical Value	Results (T=1.1319)
$\mathrm{H_a}$: $\mu_{MS} \neq \mu_{OR}$	Two-tailed	$\|T\| > t_{1-\alpha/2,v}$	$t_{1-\alpha/2,v} = 2.08$	$\|T\| < t_{1-\alpha/2,v}$
$\mathrm{H_a}$: $\mu_{MS} > \mu_{OR}$	Right-tailed	$T > t_{1-\alpha,v}$	$t_{1-\alpha,v}$=1.721	$T < t_{1-\alpha,v}$
$\mathrm{H_a}$: $\mu_{MS} < \mu_{OR}$	Left-tailed	$T < t_{\alpha,v}$	$t_{\alpha,v}$=-1.721	$T > t_{\alpha,v}$

Figure 2.52: Alternative testing for openness

The analysis shows that a significant percentage of papers published in MS and OR (nearly half) each year were contributed by authors who had not made a single contribution over the previous five years, indicating strong innovativeness and openness of these two journals in the academic field. This situation is in sharp contrast with the extreme case in which none of the papers are contributed by new authors, which implies that if an author has never published in a journal, it is hard for them to get published in the future.

2.4.2 Editorial Board (EB) Influence

Members of editorial boards (EBs) of academic journals usually have high reputations and many successful studies to their credit in their respective fields. They have an obligation to ensure a fair review process and to protect authors from any inappropriate selection strategies. If editors are not held accountable for their actions, increasing incidents of poor peer review and bias are bound to occur. Nevertheless, reviewers may also have personal preferences distorting the results of the process.

In this section, the question is addressed of whether the EB enjoys any advantages in publishing their own papers. Without access to the submission data and unbiased quality measures, the methodology used in this study is based on comparing the ratio between EB members over author population and EB member publications over total publications. The rationale behind such an approach is that if, for example, the EB accounts for 1% of the author population but publishes 50% of the papers, this may indicate a disproportional power of the EB members in publishing their own papers in the respective journal.

Specifically, 20 years worth of data was split into three periods, 1995-2001, 2002-2008, and 2009-2014. Seeing that the editorial leadership changed, we can assume that the members of the EB remained constant in the three periods under the same chief editor. The lists of board members, as well as chief editors, department editors and associate editors were produced for 1998, 2003, and 2009. Then, the number of papers published with at least one board member as a co-author, and the number of board members over the entire author population is identified for the specified years.

Table 2.9 shows the percentage of papers contributed by the EB as a percentage of the total papers of MS for each period between 1995 and 2014. For example, the first figure, 15.4%, means that the number of papers published between 1995-2001 with at least one of the authors who were on the editorial board in 1998 amounted to 15.4% of the total number of publications in the period between 1995 and 2001. The information reported on in diagonal cells of the table, however, paints a different picture. Normally, the professors are the most productive while on the EB or just before joining it. The upper diagonal cells indicate the productivity of the EB before and after assuming board membership. The trend in the table reveals that after they leave the board, their number of publications decreases. The bottom diagonal cells indicate the productivity of the EB before and after joining the board. It is clear that before becoming members of the EB, scholars are very productive. Their trend in authorship is increasing in the period before they become board members.

Rank	1995-2001	2002-2008	2009-2014
Board 1998	15.4%	11.0%	6.5%
Board 2003	11.3%	13.8%	7.5%
Board 2009	6.7%	15.3%	14.8%

Table 2.9: Papers contributed by the EB as a percentage of total number of papers of MS between 1995 and 2014.

Table 2.10 shows the percentage of EB members among authors contributing to MS between 1995 and 2014. For example, 4.4% implies that EB accounts for 4.4% of the total number of authors who published from 1995 to 2001. The percentage of board members among authors in general in 1998, 2003, and 2009 was very consistent. Table 2.11 shows the EB members and their co-authors as a percentage of the total author population.

	1995-2001	2002-2008	2009-2014
Board 1998	4.4%	3.2%	1.6%
Board 2003	3.6%	4.2%	1.9%
Board 2009	1.8%	1.8%	6.4%

Table 2.10: EB as a percentage of the total number of contributing authors for MS in 1995-2014.

#	1995-2001	2002-2008	2009-2014
Board 1998	22.1%	15.5%	8.5%
Board 2003	16.1%	19.6%	10.8%
Board 2009	9.7%	21.1%	25.0%

Table 2.11: EB and their co-authors as a percentage of all contributing authors for MS in 1995-2014.

The analysis was repeated for OR, namely, the lists of board members together with the chief editor, department editors and associate editors were produced for 2002, 2009, and 2015. Twenty years worth of data was divided into three parts, 1995-2001, 2002-2008, 2009-2014 at which the journal had different chief editors. Table 2.12 shows that the percentage of publications with at least one board member in 2002 took was 9.67% and 6.94% in 1995-2001 and 2002-2008 respectively, while the percentage of board members co-authorship in 2009 was 11.43% and 10.08% in 2002-2008 and 2009-2014 respectively. No significant increase in the number of publications was observed in the period after an author became a member of the EB.

Rank	1995-2001	2002-2008	2009-2014
Board 2002	9.67%	6.94%	4.83%
Board 2009	6.41%	11.43%	10.08%
Board 2015	–	9.62%	14.06%

Table 2.12: Percentages of the number of OR papers written by the EB in 1995-2014.

Table 2.13 shows the percentage of EB authors in 1995-2014 and Table 2.14 shows the percentage of EB and their co-authors in 1995-2014. They look quite similar to MS.

board #	1995-2001	2002-2008	2009-2014
Board 2002	4.84%	3.95%	2.54%
Board 2009	3.26%	5.98%	4.40%
Board 2015	–	5.34%	6.86%

Table 2.13: Percentages of the EB authors in OR in in 1995-2014.

board #	1995-2001	2002-2008	2009-2014
Board 2002	24.06%	16.99%	13.12%
Board 2009	14.63%	28.42%	28.37%
Board 2015	–	24.15%	37.93%

Table 2.14: Percentages of EB and their co-authors out of the total number of OR contributing authors in 1995-2014.

To explore the relationship between the percentages of papers contributed by the EB to MS/OR and the percentages of EB authors of MS/OR publications in 1995-2014, the following hypothesis was tested:

H_0: The percentage of papers contributed by the EB to MS/OR is lower than the percentage of EB members to the total number of MS/OR authors in 1995-2014;

H_1: The percentage of papers contributed by the EB to MS/OR is not lower than the percentage of EB members out of the total number of MS/OR authors in 1995-2014.

The p-value for both MS and OR is 0 and the null hypothesis is thus rejected. In other words, the percentage of papers contributed by the EB is always greater than or equal to the percentage of EB.

To explore the relationship between papers contributed by the EB to the total number of MS/OR papers, as well as the percentage of EB members and their co-authors out of the total number of MS/OR authors in 1995-2014, the following hypothesis was tested:

H_0: The percentage of papers contributed by the EB out of the total number of MS/OR papers of MS/OR is greater than the percentage of EB and their co-authors out of the total number of MS/OR authors in 1995-2014.

H_1: The percentage of papers contributed by the EB out of the total number of MS/OR total papers is not greater than the percentage of EB and their co-authors out of the total number of MS/OR authors in 1995-2014.

The p-value for both MS and OR is 0, and the null hypothesis is thus rejected. In other words, the percentage of papers contributed by the EB out of the total number of MS/OR papers is not greater than the percentage of EB and their co-authors out of the total number of MS/OR authors in 1995-2014.

The analysis of the influence of the EB indirectly show that the EBs of MS and OR may not exert a strong influence on publishing their own papers in the respective journals.

2.5 Conclusion

A descriptive analysis and a comparative study were performed on the dataset comprising MS and OR publications. The amount of collaboration kept increasing over the time as did the number of authors per publication; both followed the same distribution in MS and OR. The factors potentially affecting productivity were explored based on social network mapping and cliques analysis. There is no clear evidence showing the clique size matters in the productivity. The degree in MS and OR followed the same distribution, with density as the main and statistically significant factor. In most of the case, density makes negative influence on the productivity while diversity makes positive influence on it. Researchers and scholars from different background should enhance communication and strengthen cooperation. Furthermore, new metrics were developed to assess the openness of the journals to new authors and the influence of editorial board members. Approximately 47% (MS) and 45% (OR) of the papers were contributed by new authors. No evidence was found to show that EB members exert much influence on publishing their own papers.

This work is subject to limitations in that there is no submission data and thus research productivity cannot be measured accurately. The chance of publication could not be measured, and surrogate publication data was used. A challenge remains to separate the impact of individual authors from the social network effect. In future research, the plan is to add social network measurements, such as the diameter, to explore their potential influence over productivity. A further plan is to collect more data from other INFORMS journals to validate and improve the quality of the obtained results.

Chapter 3

Big Data Analytics about Healthcare: A Case Study in New York State

3.1 Introduction

The United States annually spends roughly $3 trillion on the health care industry, which is currently seeing significant changes resulting from the quality and performance improvement initiatives. Improving patient satisfaction has become the overall objective for all health care providers. To achieve a higher degree of hospital performance, increase the demand and rise above the competition, the adoption of a patient-orientated system is necessary. The quality of patient care is usually determined by both medical factors, such as the expertise of the physicians and the quality of medical equipment, and nonmedical factors, such as the infrastructure, quality of services, competence of personnel, and efficiency of operational systems. Therefore, to pursue higher profits and offer higher quality service, a comprehensive system that can improve both factors needs to be designed and implemented. The healthcare industry has historically generated large amounts of data due to the practices related to record keeping, compliance and regulations, and patient care, which was studied by Kudyba (2010). Due to the technological developments, the current trend is developing towards rapid digitization of these large amounts of data. Driven by the high demand for improving the quality of healthcare delivery while reducing the costs, these large amounts of data hold the promise of supporting a wide range of medical and healthcare functions, including, among others, population health management, clinical decisions, and hospital operations decisions for profitability. Achieving these aims may require a concerted effort in the following areas. First, identifying the patient disease demographics and admission type distributions helps to understand the relationship between demographics and a disease trend. Furthermore, the specific healthcare plan can be developed according to a local condition. Second, most hospitals face a great challenge of balancing quality management and cost recovery in patient care. Identifying the factors that influence hospital charges or costs is of great importance. Most clinicians believe that treatment outcomes are regarded as indicators of the hospital reputation and that they are of highest concern among patients. However, whether non-technical indicators such as accessibility of services and cleanness are really of the patients concern remains unclear. They are also relevant in terms of hospital investments and reducing the unnecessary costs and improving the patients experience. Third, the aim is to identify profit drivers for hospitals.

3.1.1 Background and Literature Review

The United States spends more on health than any other country that is economically comparable. Health spending (cf. Figure 3.1) is a measure of the final consumption of health care goods and services (i.e. current health expenditure), including personal health care (curative care, rehabilitative care, long-term care, ancillary services, and medical goods), and collective services (prevention and public health services as well as health administration), but excluding investment spending discussed by OECD (2016). This analysis draws on data from the Organization for Economic Cooperation and Development to compare health care spending across different countries. The U.S. spent nearly $10,000 on health per person in 2016. Switzerland, Luxemburg, and Norway are the next highest spenders, but they all spent at least $2,000 less per person than the U.S. did. The median spending on health care among the other 33 OECD countries was about $3,500 per person, which amounts to only 1/3 of the U.S. spending. In the U.S., the public and private payments are very close in their amounts.

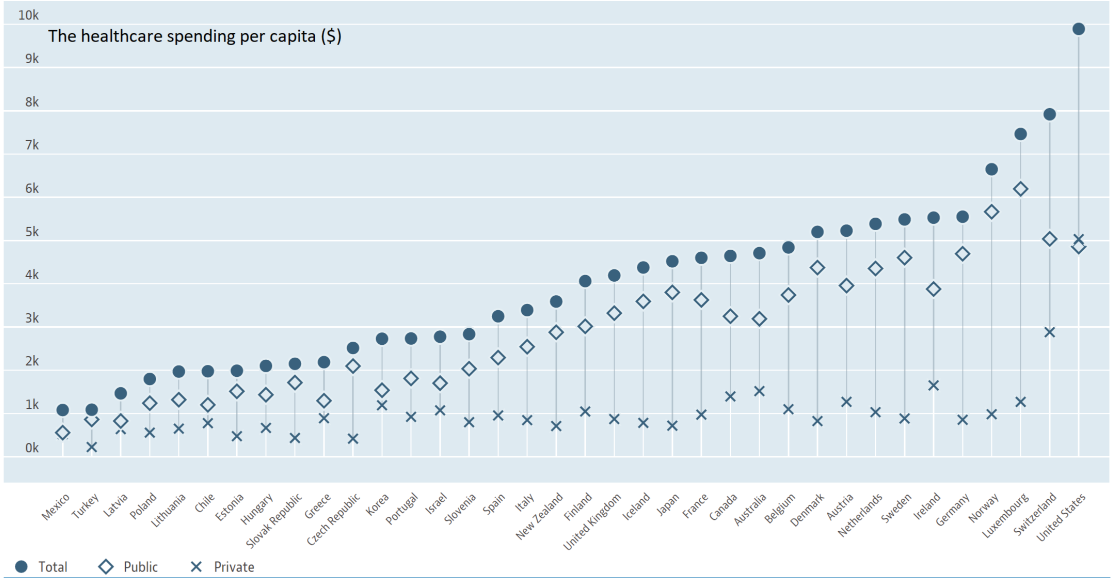


Figure 3.1: The healtcare spending per capita (OECD, 2016)

Moreover, the national health expenditures are projected to grow in the U.S. (as one of the worlds richest countries) at an average annual rate of 5.6 percent from 2016 to 2025. According to this trend, the national health expenditures will amount to 19.9 percent of gross domestic product by 2025, making it the capital-chasing market full of potential. Centers for Medicare & Medicaid Services (CMS) projected that total health care spending for 2016 reached nearly $3.4 trillion, thus showing an increase of 4.8 percent from 2015. According to CMS, the U.S. health care spending is projected to reach nearly $5.5 trillion by 2025. Keehan et al. (2017) showed that the agency attributed the increase largely to the countrys aging population and rising prices for health care services. In addition, even during the financial crisis when the global economy suffered a severe downturn in 2008 and 2009, the blue line in Figure 3.2 indicates that the national health expenditures as a percentage of gross domestic product was still increasing because of the rigid demand from people regardless of the macro-economic impact.

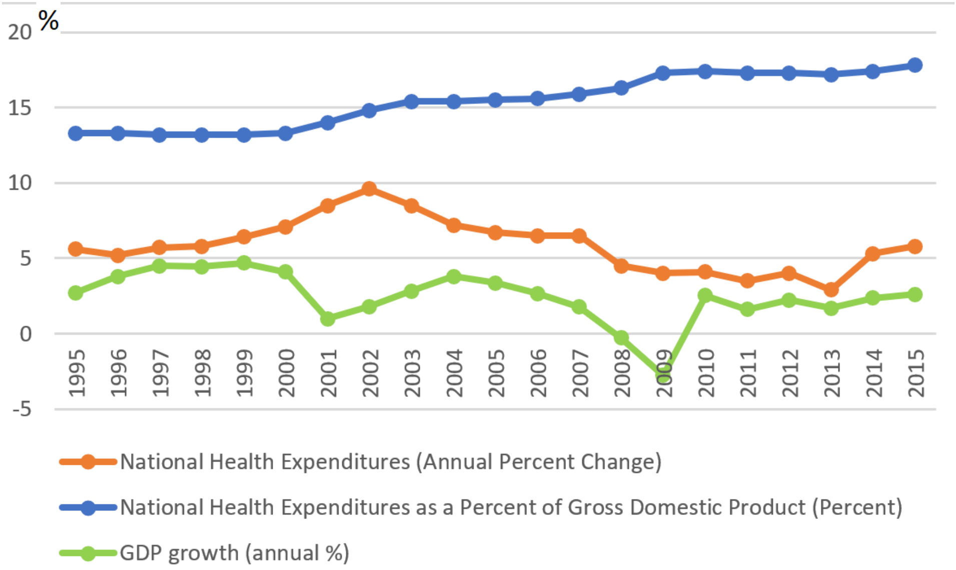


Figure 3.2: Growth in national health expenditures and their share in the GDP, 1995-2015 in Bryan (2016)

Based on the analysis of the dataset of Hospital Inpatient Discharges in New York State, the research areas in the study are the following: 1) studying the disease trend and admission types cluster analysis by gender, race, age group, disease, etc. to understand the geographic patient distribution, 2) exploring the factors influencing the hospital charge or cost, such as non-technique service satisfaction and local demographics, 3) building an optimization model to offer suggestions on location and specialty selection for new medical businesses.

Patient demographics such as race, gender, county, and postal code represent the core information for any medical institution. They facilitate disease identification and patient clustering. A study by Rand and Kuldau (1990) examined the correlation between the epidemiology of obesity and self-defined weight problem and gender, race, age, and social class. Krieger (1990) found racial and gender discrimination to be risk factors for high blood pressure. By location, Wu et al. (2001) studied the incidence rate and stage of disease in colorectal cancer by race, gender, and age group in the United States from 1992 to 1997. Patients can be clustered according to additional elements, such as, admission type and payment methods. Such information can be used to optimize health care utilization, health outcomes, and costs. Stavrakis et al. (2007) predicted the surgical outcomes among in- and outpatients based on admission.

Cleary and McNeil (1988) used patient satisfaction as an indicator of the quality of care. Shindul-Rothschild et al. (2017) examined hospital characteristics, staffing, and nursing care factors by conducting a secondary analysis of the Hospital Consumer Assessment of Health Care Providers Systems (HCAHPS) survey in California, Massachusetts, and New York hospitals. Ding (2015) discussed the impact of service design and process management on clinical quality. Furthermore, Huerta et al. (2016) measured the relationship between patient satisfaction and hospital cost efficiency.

Finally, Propper et al. (2007) studied the impact of the patients socioeconomic status on the distance travelled for hospital admission in the English National Health Service. Fabbri et al. (2008) studied the geography of hospital admission in the Italian national health service together with patient choice, and found that 35% of the 10 million annual hospital admissions in the country occurs outside of the patients local residence. Akinci et al. (2005) identified key factors on how patients choose hospitals in Turkey.

3.1.2 Essay Organization, Data Collection and Significance of the Study

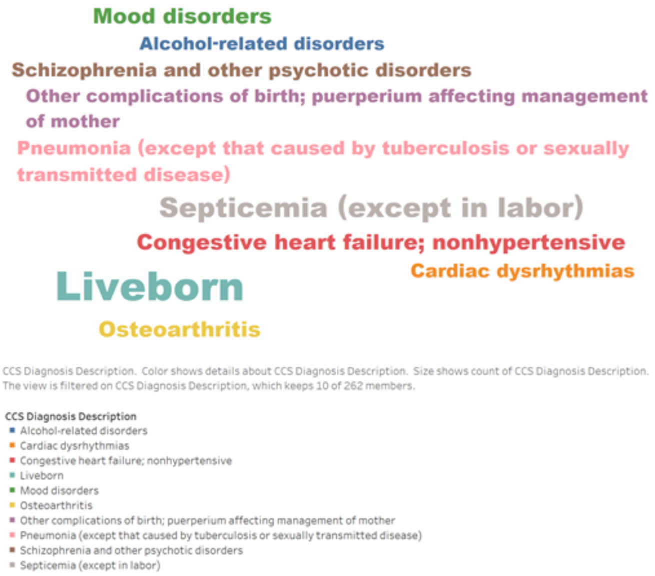
This thesis consists of four parts and the general framework is as follows. First, a general review of the healthcare industry is provided. Second, the population health - disease distribution according to gender, age, and race, and the diachronic trend will be studied. Then, the relationship between hospital charge or cost and non-technical services and demographics indicators is studied. Finally, an optimization model is built for site and specialty selection for new medical businesses. This study is conducted on several datasets. First, the NY state hospital inpatient discharges and the demographic profiles are combined to analyze the disease distribution. In addition, the datasets span over a period of seven years. Trends can be detected by studying the time series data. The hospital inpatient discharges are then combined with the results of the quality measurements survey to learn whether the charge or cost is associated with quality of service and demographic indicators. Finally, based on the dataset, an optimization model for choosing the location and service types is built for hospital expansion. This study has a remarkably wide range of applications, and its potential users come from very diverse backgrounds, including various government agencies, provider associations, individual health care providers, hospitals, health care insurers and individual patients, researchers and policymakers.

3.2 Population Health: Disease Distribution and Trend

In this part, the geographic visualization of disease and admission types are visualized on the map to understand disease distribution and geographic risk. The gender, race and age distribution are studied to understand the disease trend.

3.2.1 The Geographic Visualization of Disease

Heat maps and clustering are frequently used in descriptive analyses for data visualization. Disease maps are visual representations of intricate geographic data that provide a quick overview of said information. They are widely used for explanatory purposes to learn about areas of high risk and help with policy making and resource allocation in such areas. Despite methodological advances and increasing data availability, no systematic evaluation of the available techniques has so far been proposed for the purposes of public health decision-making. Besides disease mapping and descriptive studies, a growing interest has emerged in the evaluation of geographic risk. This essay focuses on the New York State inpatient geographical studies, it explores the incidence and demographic information, and discusses their role in public health decision-making. The following analysis focuses on the 2014 healthcare data. The top 10 diseases (by the number of discharges or incidences) are listed in Figure 3.3. Furthermore, the aim is to identify the areas with the higher number of incidences. The top 20 three-digit zip-code areas were selected and the top 10 diseases were explored (cf. Figure 3.4. The top 10 diseases account for nearly 30% of total disease incidences, whereas Brooklyn, Bronx, and Mid-island have the highest number of incidences.



CCS Diagnosis Description. Color shows details about CCS Diagnosis Description. Size shows count of CCS Diagnosis Description. The view is filtered on CCS Diagnosis Description, which keeps 10 of 262 members.

Figure 3.3: Top 10 CSS description

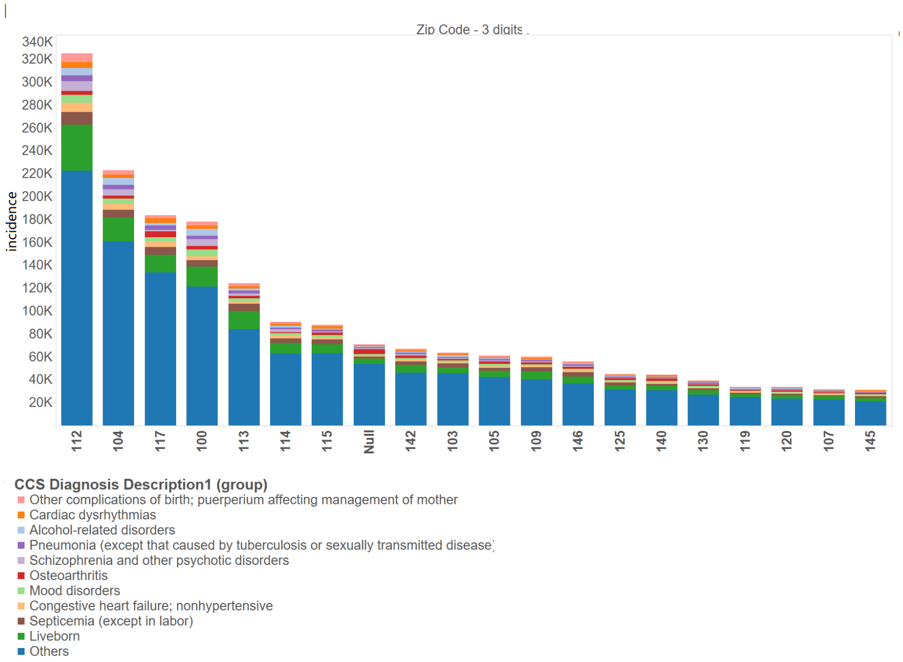


Figure 3.4: Top 10 CSS description in top 20 three-digit areas

Disease mapping focuses on gathering information on high-risk areas. By mapping the incidence to the map, the distribution information can be extracted, helping hospitals target patients and help the patients understand the risk of major diseases associated with the location of their residence. Figures 3.5 and 3.6 indicate the incidence and risk per capita (defined as the number of incidences / local population of a three-digit area) in the New York State. In lower Manhattan and mid Long Island the number of incidences is higher. Taking into consideration the risk per capita, moreover, the overall summary is provided in the Figure 3.7. The three-digit areas 148, 101, 129, 105, and 144 are the five lowest risk regions with the risk of 9.00 . Moreover, 104, 116, 149, 143, and 135 have the highest risk of up to 15.76%.

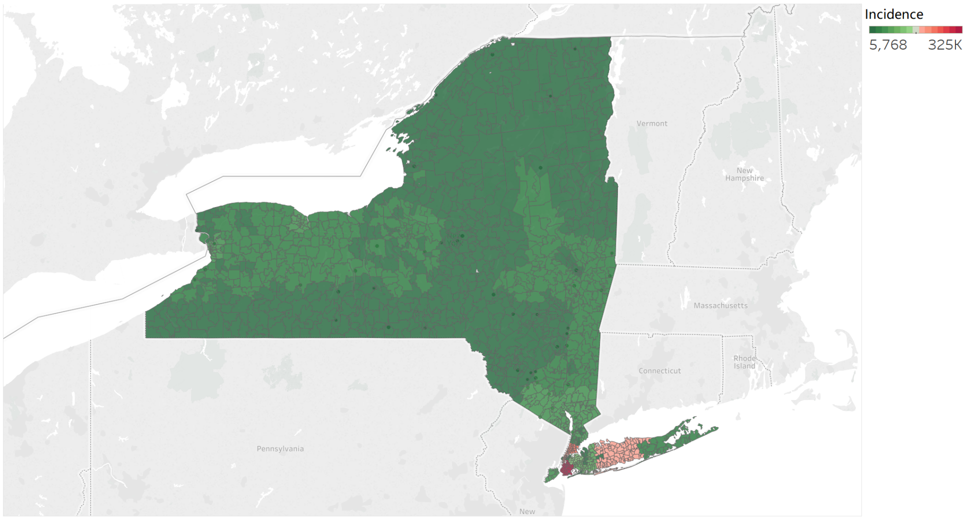


Figure 3.5: Incidence rate

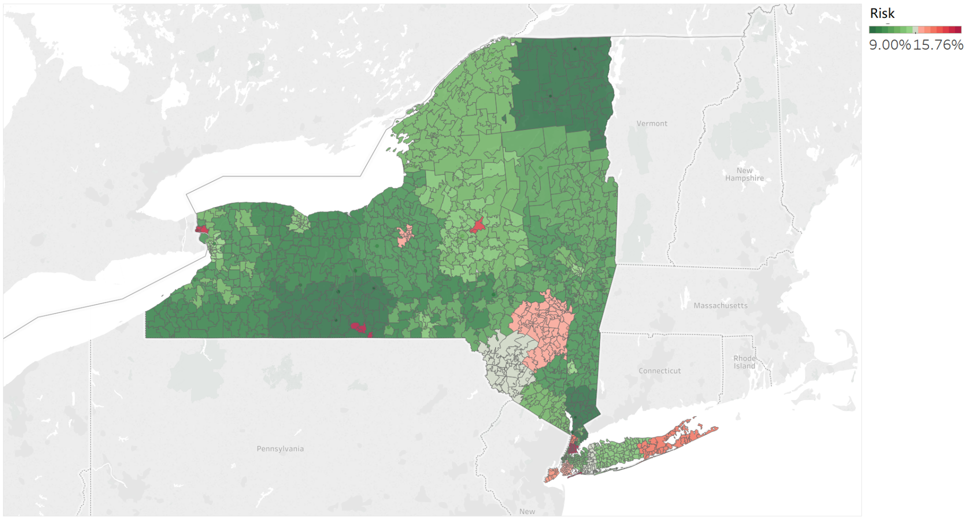
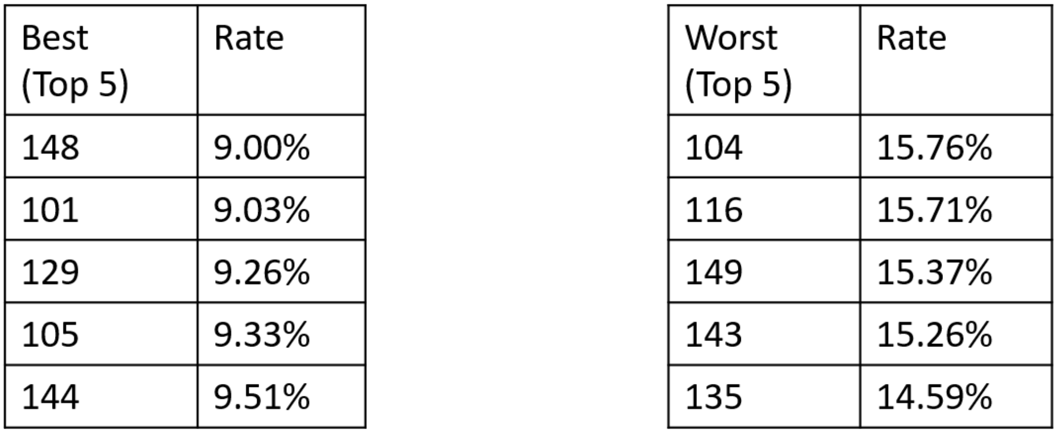


Figure 3.6: Risk



Best (Top 5)	Rate
148	9.00%
101	9.03%
129	9.26%
105	9.33%
144	9.51%

Worst (Top 5)	Rate
104	15.76%
116	15.71%
149	15.37%
143	15.26%
135	14.59%

Figure 3.7: Top five lowest and highest risk three-digit areas

Visualizing instances of cancer can help anticipate future resource needs, evaluate primary prevention strategies, and inform research. In what follows, the regional data visualization for the diagnosed cases of cancer is provided. There are 26 types of cancer out of the 261 diseases appearing in this dataset. Figure 3.8 shows the incidence histogram of the distribution of all types of cancer across top 20 regions. Figure 3.9 and 3.10 illustrate the cancer incidence histogram and risk per capita in New York State. Most counts are identified in lower Manhattan and mid Long Island. The respective figures are summarized in Table 3.12. The three-digit areas 129, 136, 148, 122, and 138 are the five lowest risk regions with the lowest rate of 0.170%. Additionally, 107, 108, 119, 110, and 124 are the top five highest risk regions, with the highest rate of 0.343%.

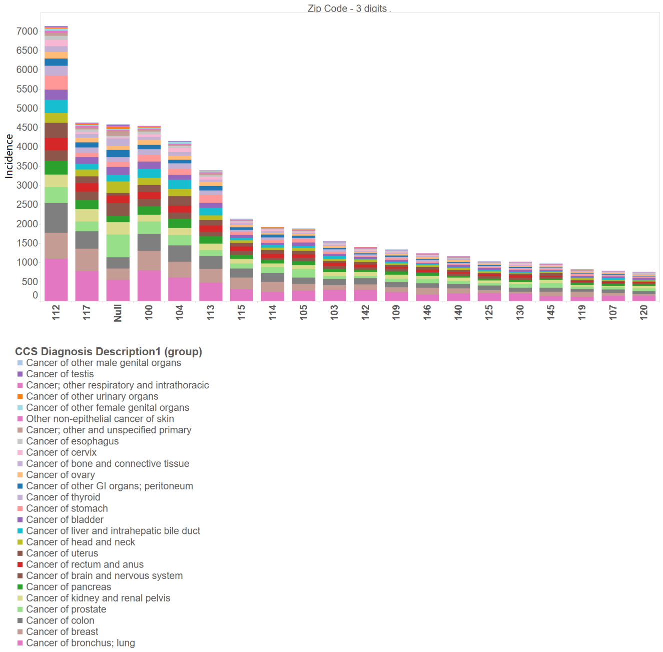


Figure 3.8: Cancer incidence rates in top 20 three-digit areas

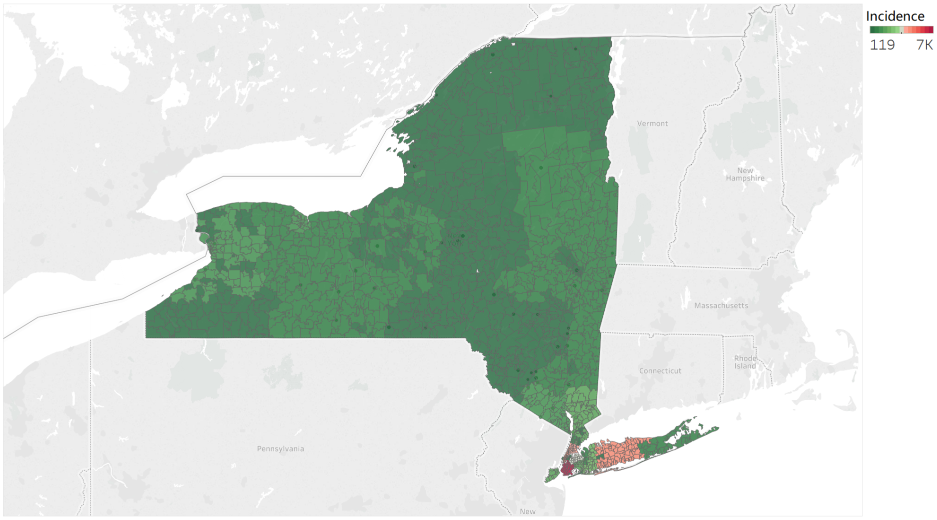


Figure 3.9: Incidence rates controlled by cancer

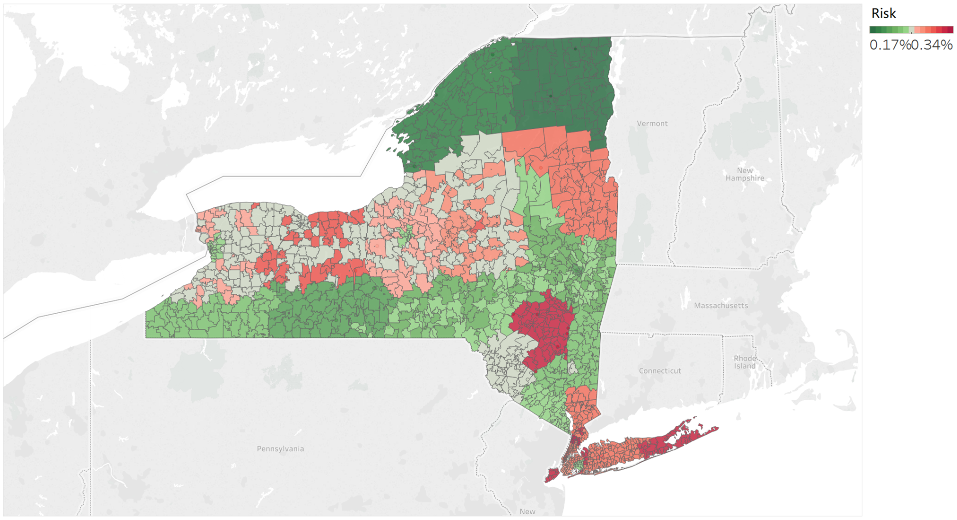
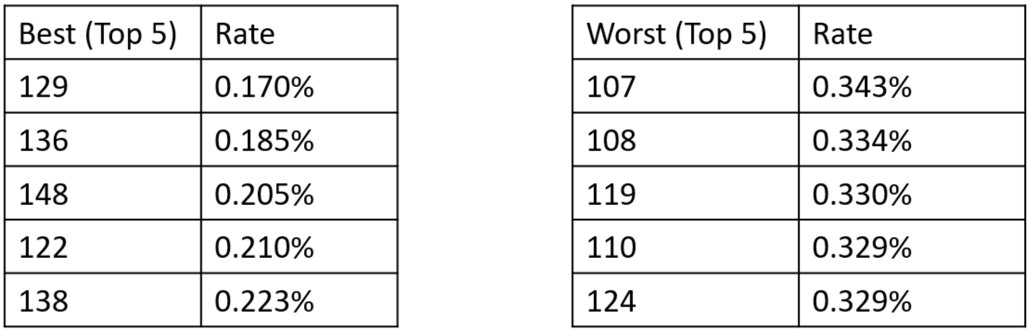


Figure 3.10: Risk controlled by cancer



Best (Top 5)	Rate
129	0.170%
136	0.185%
148	0.205%
122	0.210%
138	0.223%

Worst (Top 5)	Rate
107	0.343%
108	0.334%
119	0.330%
110	0.329%
124	0.329%

Figure 3.11: Risk table controlled by cancer

3.2.2 Geographic Distribution of Admission Types

There are five different admission types in this dataset, namely, emergency, newborn, urgent, trauma, and elective.RecDAC (2017) shows the code indicates the type and priority of inpatient admission associated with the service on an intermediary submitted claim. As shown below, the risk mapping of different admission types differs considerably.

- Emergency - The patient required immediate medical intervention as a result of severe, life threatening, or potentially disabling conditions. Generally, the patient was admitted through the emergency room.

- Urgent - The patient required immediate attention for the care and treatment of a physical or mental disorder. Generally, the patient was admitted to the first available and suitable accommodation.

- Elective - The condition permitted adequate time to schedule the availability of suitable accommodations.

- Newborn - Necessitates the use of a special source of admission codes.

- Trauma - Visits to a trauma center or hospital as licensed or designated by the State or local government authority authorized to do so, or as verified by the American College of Surgeons and involving trauma activation.

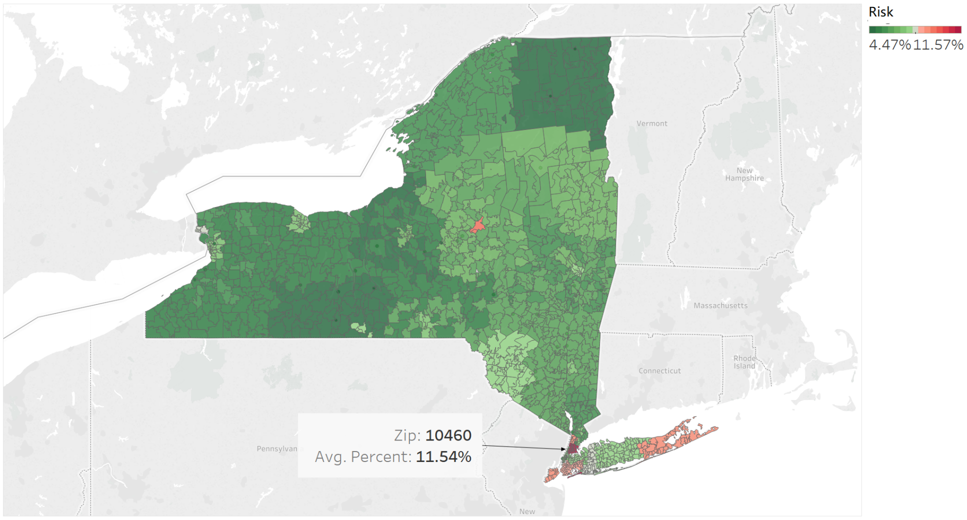


Figure 3.12: Risk mapping for admission type of emergency

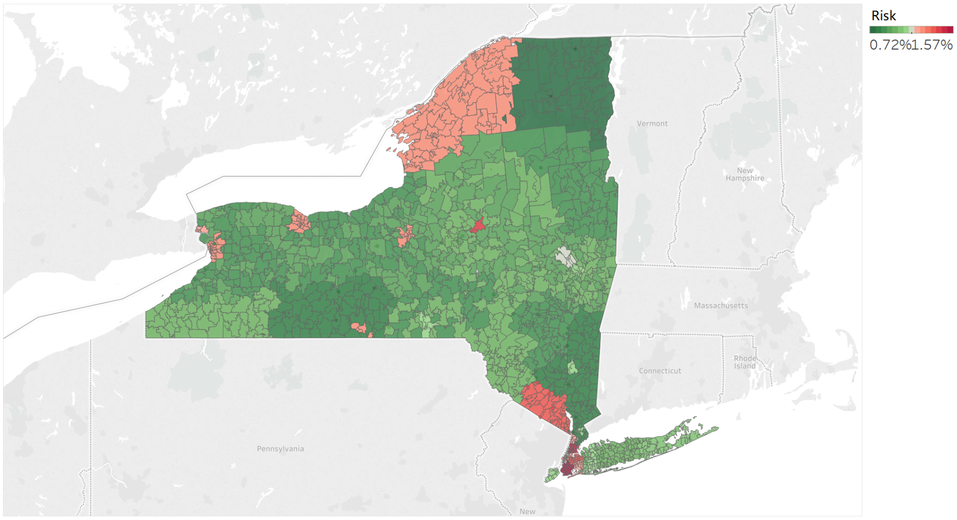


Figure 3.13: Risk mapping for admission type of newbornn

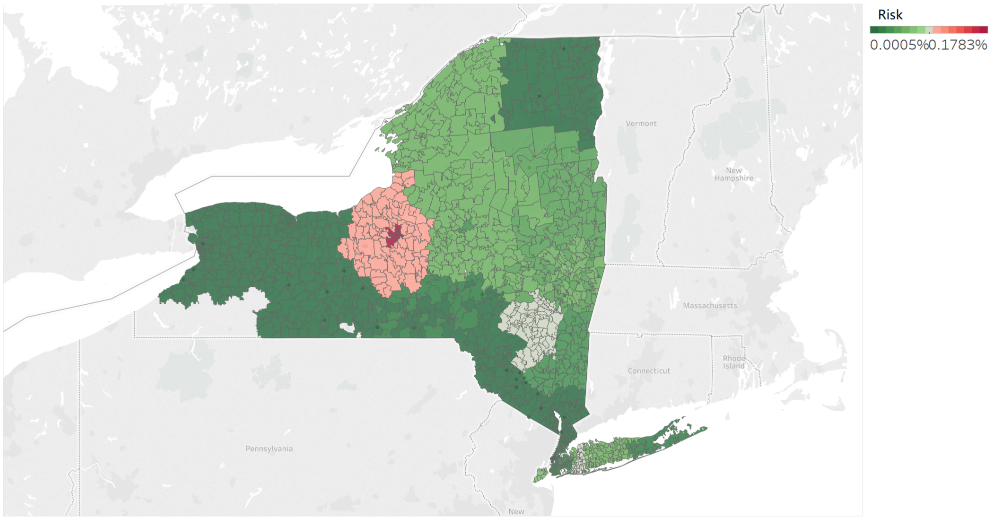


Figure 3.14: Risk mapping for admission type of trauma

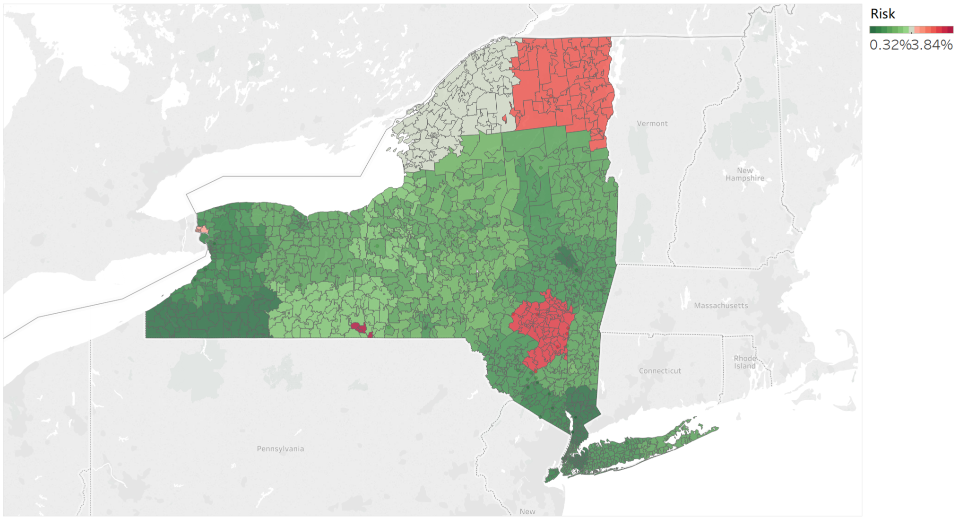


Figure 3.15: Risk mapping for admission type of urgency

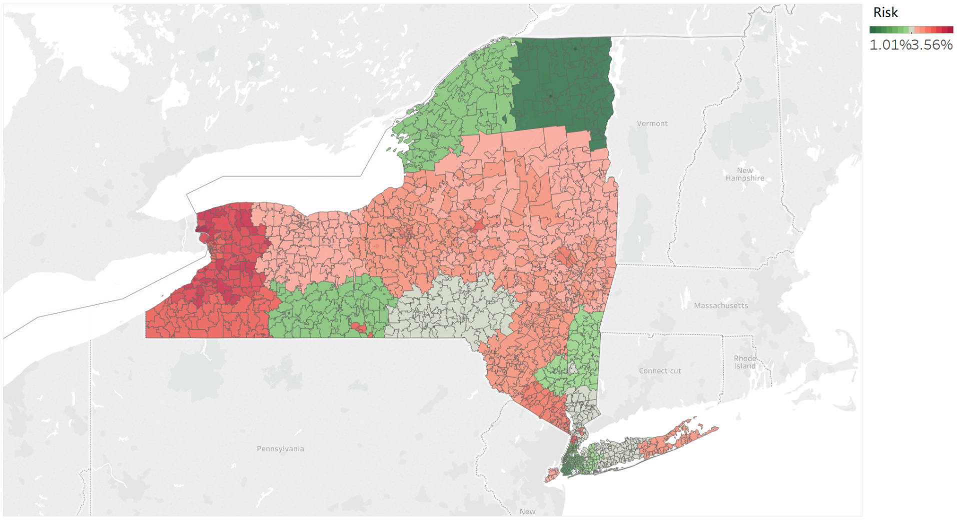
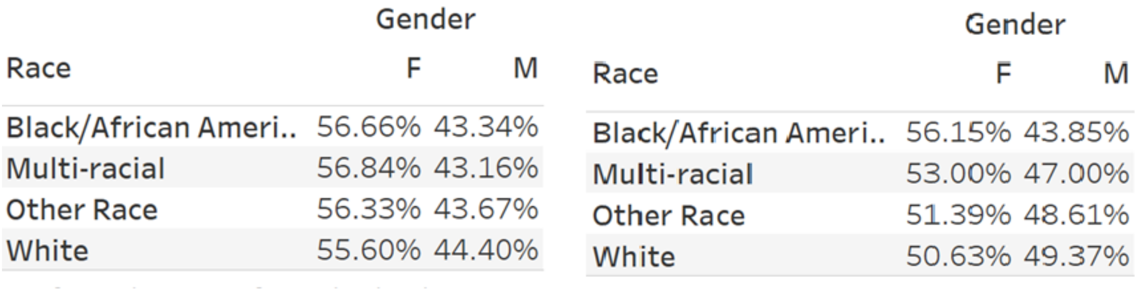


Figure 3.16: Risk mapping for admission type of elective

Five three-digit zip areas are selected to show that there are differences in admission types in different areas.

3.2.3 Gender, Race and Age Distribution

Changes in gender, age, race, and ethnicity affect the necessary health care resources, the cost of care provided, and even the conditions associated with each population group. This section focuses on the analysis of the population-based hospital admissions and incidence rates. In Figure 3.17, the table on the left represents the proportion of females and males of different races. It shows that the percentage of admission rates for females is significantly higher than that of males. Females are thus either more prone to illness, or they are more likely to visit doctors. The pattern is consistent across different races. The table on the right shows the same results controlled by cancer. Interesting here is that the pattern across different races is not the same here. The ratio of females to males is significantly higher among white rather than African American population. It may be the case that white males are at higher risk of getting cancer than other races.



	Gender	
Race	F	M
Black/African Ameri..	56.66%	43.34%
Multi-racial	56.84%	43.16%
Other Race	56.33%	43.67%
White	55.60%	44.40%

	Gender	
Race	F	M
Black/African Ameri..	56.15%	43.85%
Multi-racial	53.00%	47.00%
Other Race	51.39%	48.61%
White	50.63%	49.37%

Figure 3.17: Incidence rate by gender and race in general and controlled by cancer

T-test (Table.3.1) is used to examine the difference in incidence rates between males and females across all three-digit areas. The hypothesis is as follows:

H_0: $p_{Mj} = p_{Fj}$

H_a: $p_{Mj} \neq p_{Fj}$

where p_{Mj} is the mean incidence rate of male in three-digit areas j, p_{Fj} is the mean incidence rate of female in three-digit areas j. P-value for a two-tail test is 0.0023 and thus the null hypothesis is rejected. It can be concluded that the mean incidence rates for males and females across different three-digit areas are not the same. Furthermore, another T-test (cf. Table 3.1) is used to examine the difference in the mean incidence rates between males and females across all different three-digit areas controlled by cancer. The hypothesis is as follows:

H_0: $p'_{Mj} = p'_{Fj}$

H_a: $p'_{Mj} \neq p'_{Fj}$

where p'_{Mj} is the mean incidence rate for males in three-digit areas j and p'_{Fj} is the mean incidence rate for females in three-digit areas j controlled by cancer. P-value for a two-tail test is 0.0045, meaning that the null hypothesis can be rejected and it can be concluded that the incidence rate of males and females in different three-digit areas are not the same when controlled by cancer.

Table 3.1: T-test for incidence by gender

	All disease		Cancer	
	Female in Percentage	Male in Percentage	Female in Percentage	Male in Percentage
Mean	0.5599	0.4400	0.5126	0.4873
Variance	2.050e-4	2.050e-4	1.897e-3	1.897e-3
Observations	50	50	50	50
df	98		98	
t Stat	41.874		2.904	
P(T <= t) one-tail	1.196e-64		2.275e-3	
P(T <= t) two-tail	2.392e-64		4.550e-3	

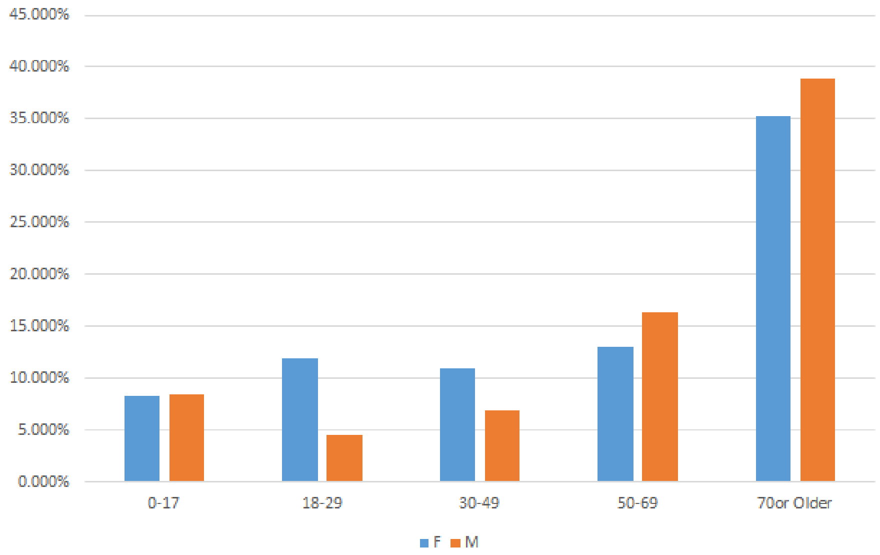


Figure 3.18: Percentage by gender and age groups

Figure 3.18 illustrates the distribution of different gender and age groups for all diseases. The incidence rate increases significantly for people in the ”70 or older” group. In the ”0-17” age group, the incidence rate of males and females prima facie seems equal. In the age group ”18-49”, the incidence rate for females is higher than for males. The reason might be that younger women tend to be more sensitive to bodily changes and more likely to seek medical help. The following hypothesis is thus tested:

H_0: Males and females are distributed equally among the various age groups.

H_a: Males and females are not distributed equally among the various age groups.

The chi-square statistic is 6,1627.9257, and the P-value $< .00001$. The result is significant at $p < .05$, and the null hypothesis is thus rejected. Males and females are not distributed equally across the age groups.

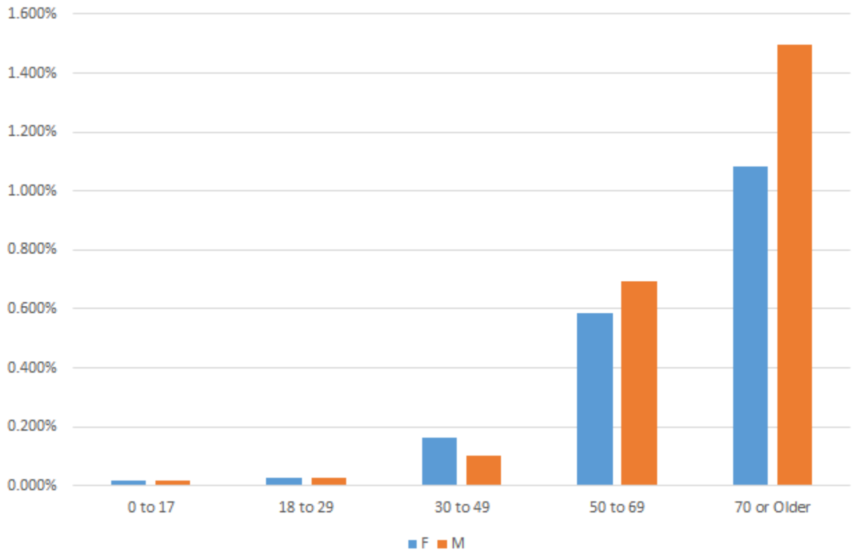


Figure 3.19: Percentages of different gender and age groups controlled by cancer

The distribution of different gender and age groups controlled by cancer is shown in Figure 3.21. The following hypothesis is tested:

H_0: Males and females are distributed equally across the age groups for cancer.

H_a: Males and females are not distributed equally across the age groups for cancer.

The chi-square statistic is 41.8221. P-value is < 0.00001. The result is significant at $p< .05$ and the null hypothesis is thus rejected. Males and females are not distributed equally among the various age groups for cancer.

3.3 Hospital Charge, Cost and Service Quality

In the context of hospital financing, charges, costs, and payments are three separate terms with three distinct definitions. Charges are defined as the initial, individually listed prices that a hospital sets for the items and services it provides, usually including the following: operating room, intensive care unit, nursing, pharmacy, laboratory, radiology, respiratory care, cardiology, supplies, and miscellaneous. Charges are sometimes used as benchmarks or in creating a reference price list to negotiate payment rates with insurers instead of for actual payments.

The dataset in the present study comprises hospital charges from 2009 to 2015 listed in the New York hospital inpatient discharges. As Figure 3.20 shows, the peaks are shifting to the right with the tail elongating and thickening in later years. This illustration indicates that expenses are increasing, especially for the diseases whose treatment is more expensive.

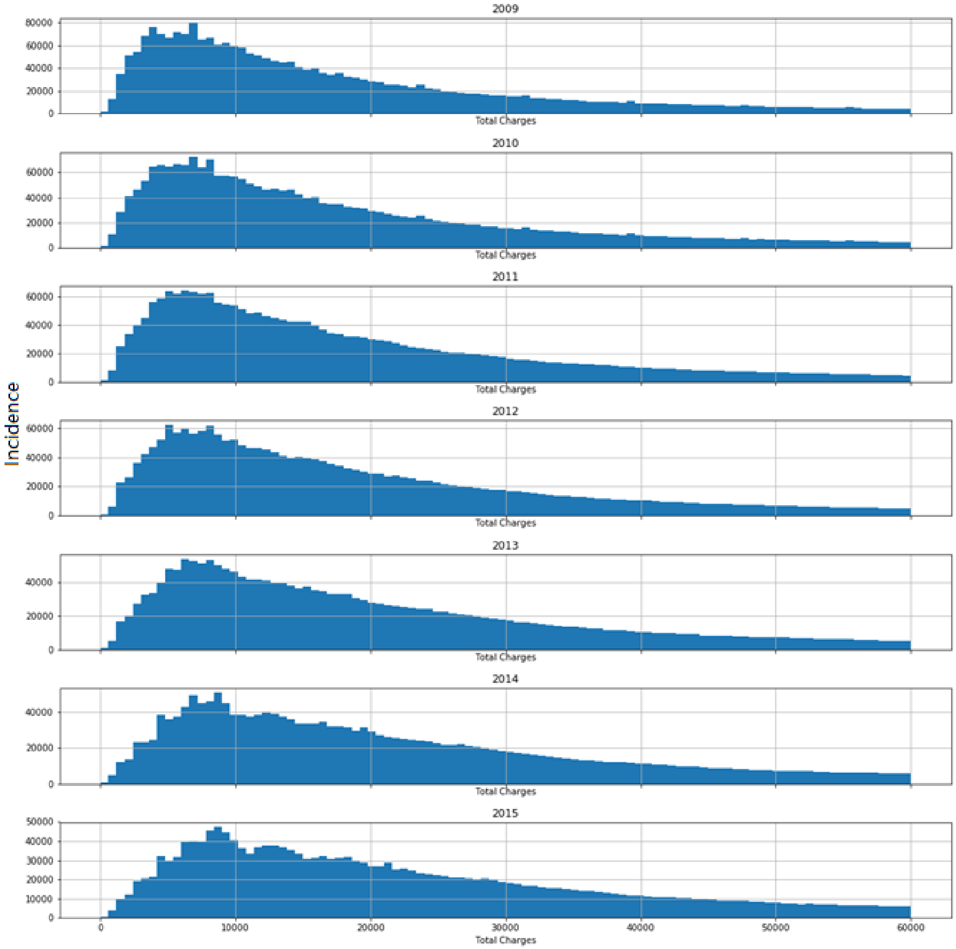


Figure 3.20: Charge by year

One-third of all health care spending in the United States is attributed to inpatient hospital services. Between 1997 and 2011, aggregated inflation-adjusted hospital costs grew by 3.6 percent annually, while average inpatient hospital costs varied substantially depending on the case studied by Audrey J. Weiss and Steiner (2014). Patient satisfaction is an essential part of hospital management and also a necessary part of the work of healthcare providers. Therefore, misunderstanding patient needs may lead to an underutilization of hospital facilities and to hindering the overall development of the healthcare system. In this section, several elements from the patient service quality perception were evaluated, and the relationship between non-technical services on the one hand and hospital charges and costs on the other are studied. The following three hypotheses are proposed.

H_{10}: The charge is positively correlated with cost.

H_{1a}: The charge is not positively correlated with cost.

H_{20}: The cost is positively correlated with service quality (patient satisfaction).

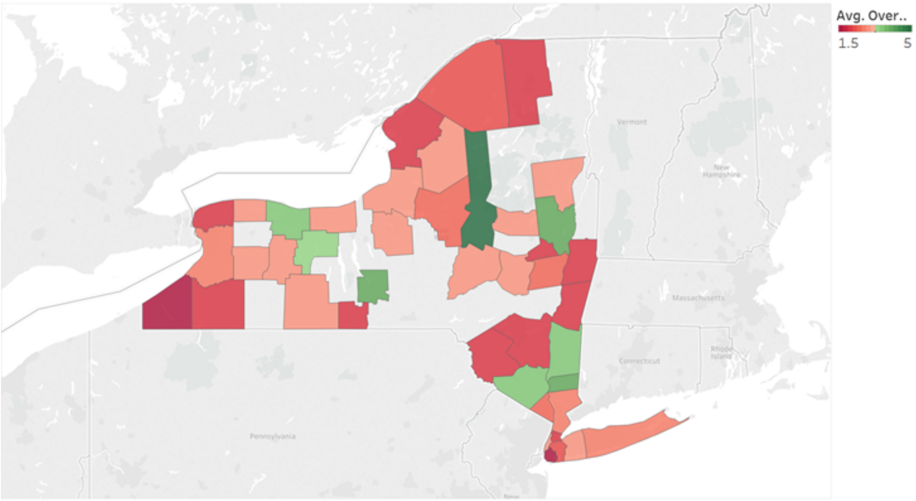
H_{2a}: The cost is not positively correlated with service quality.

H_{30}: The charge is positively correlated with service quality (patient satisfaction).

H_{3a}: The charge is not positively correlated to service quality.

The Center for Medicare & Medicaid Services (CMS) has developed 12 Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) star ratings to enable consumers to compare hospitals against excellence in the quality of healthcare. Figure 3.21 shows that the overall average hospital rating by county. The colors from red to green correspond to ratings from 1.5 to 5. The limitation of the dataset is that not all of the counties have rating values available because hospitals in some counties did not take part in this survey.

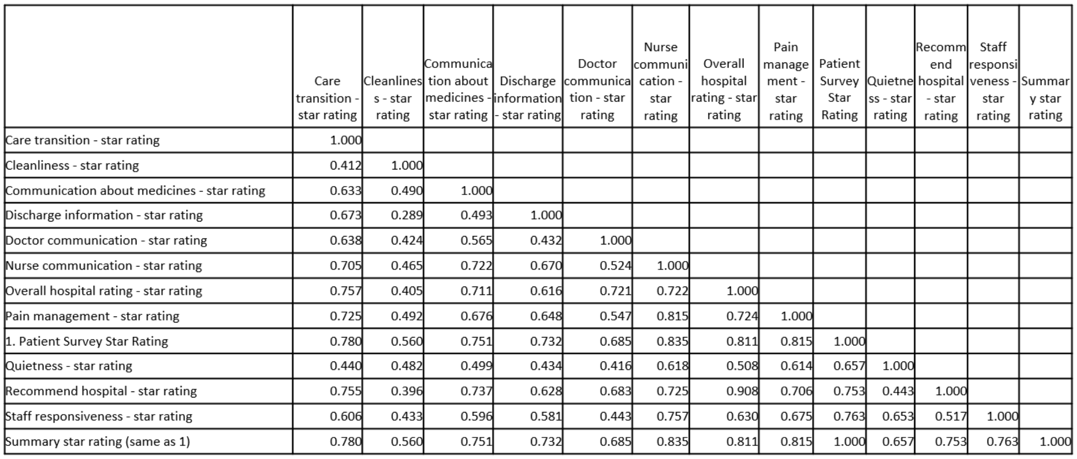
Rating



Map based on Longitude (generated) and Latitude (generated). Color shows average of Overall hospital rating - star rating. Details are shown for County Name.

Figure 3.21: The average overall hospital rating by county in NY State

The correlation among different ratings is analyzed and illustrated in Figure 3.22. Highly correlated variables are defined as those with the absolute correlation coefficient ≥ 0.8. Nurse communication rating highly correlats with pain management and patient survey star rating. Overall patient satisfaction and loyalty positively correlated with patient survey rating and the recommended hospital rating.



	Care transition - star rating	Cleanliness - star rating	Communication about medicines - star rating	Discharge information - star rating	Doctor communication - star rating	Nurse communication - star rating	Overall hospital rating - star rating	Pain management - star rating	Patient Survey Star Rating	Quietness - star rating	Recommend hospital - star rating	Staff responsiveness - star rating	Summary star rating
Care transition - star rating	1.000												
Cleanliness - star rating	0.412	1.000											
Communication about medicines - star rating	0.633	0.490	1.000										
Discharge information - star rating	0.673	0.289	0.493	1.000									
Doctor communication - star rating	0.638	0.424	0.565	0.432	1.000								
Nurse communication - star rating	0.705	0.465	0.722	0.670	0.524	1.000							
Overall hospital rating - star rating	0.757	0.405	0.711	0.616	0.721	0.722	1.000						
Pain management - star rating	0.725	0.492	0.676	0.648	0.547	0.815	0.724	1.000					
1. Patient Survey Star Rating	0.780	0.560	0.751	0.732	0.685	0.835	0.811	0.815	1.000				
Quietness - star rating	0.440	0.482	0.499	0.434	0.416	0.618	0.508	0.614	0.657	1.000			
Recommend hospital - star rating	0.755	0.396	0.737	0.628	0.683	0.725	0.908	0.706	0.753	0.443	1.000		
Staff responsiveness - star rating	0.606	0.433	0.596	0.581	0.443	0.757	0.630	0.675	0.763	0.653	0.517	1.000	
Summary star rating (same as 1)	0.780	0.560	0.751	0.732	0.685	0.835	0.811	0.815	1.000	0.657	0.753	0.763	1.000

Figure 3.22: Correlation among different items

After removing highly correlated variables (absolute value of correlation coefficient ≥ 0.8), regression analysis was performed in which the dependent variable was defined as the average charge per admission (by hospital) and the independent variables as the star rating of care transition, cleanliness, communication about medicines, discharge information, doctor communication, nurse communication, overall hospital rating, quietness, and staff responsiveness. As Table 3.2 shows, according to the regression results, R-squared is 0.333 with the p-value for total cost (5.32E-06) below the common alpha level of 0.05, and overall hospital rating (0.052) below the common alpha level of 0.1. The said results indicate that both of the respective variables are statistically significant. The regression results also show that the coefficient of total cost is 1.28, which indicates that for every additional unit increase in total cost, the total charge is expected to increase on average by \$1.28. The coefficient of overall hospital rating is 6572.86, which indicates that for every additional unit increase in overall hospital rating, the total charge is expected to increase on average by \$6572.86.

Table 3.2: Regression results by hospital

Variable	**Model (dependent variable: Average total charge)**
Doctor communication	-1493.58
	(-0.464)
Cleanness	-1140.22
	(-0.590)
Communication about medicines	-581.29
	(0.163)
Care transition	-4721.91
	(-1.607)
Nurse communication	2908.14
	(0.888)
Overall hospital rating	6572.86*
	(1.962)
Quietness	-5073.28
	(-1.846)
Staff responsiveness	-778.406
	(-0.285)
Average total cost	1.282***
	(4.834)
p-value	1.32e-5
R^2	0.333

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

t statistics in parentheses

3.4 Hospital Charge, Cost vs. Demographics

Hospital charges seem not to follow a logical pattern as it is often difficult to identify the relationship between the charges and the costs of providing medical services. In this section, first, the influence of geographic regions and local demographics on the hospital charge and cost is explored. Figure 3.23 illustrates the geographic variation. The X- and the Y-axis respectively indicate the average total cost and average total charge for different counties. The slope in the red line has the value of 1, indicating the breakeven line for hospitals in each county. Hospitals within the same geographic region likely face similar costs for wages, properties, and resources, despite significant variability in their profit margin. As is shown in Figure 3.24, there are few hospitals (e.g. Coney Island Hospital) below the 45-degree line (red line), which indicates that their charges exceed the costs . For example, the average total charges (per discharge) of Westchester Medical Center is $131,590 but the average total cost (per discharge) amounts to only $28,965.

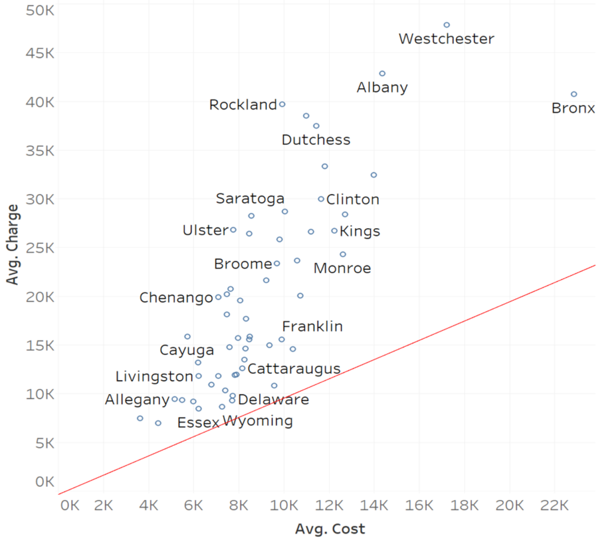


Figure 3.23: Average charge and cost by county

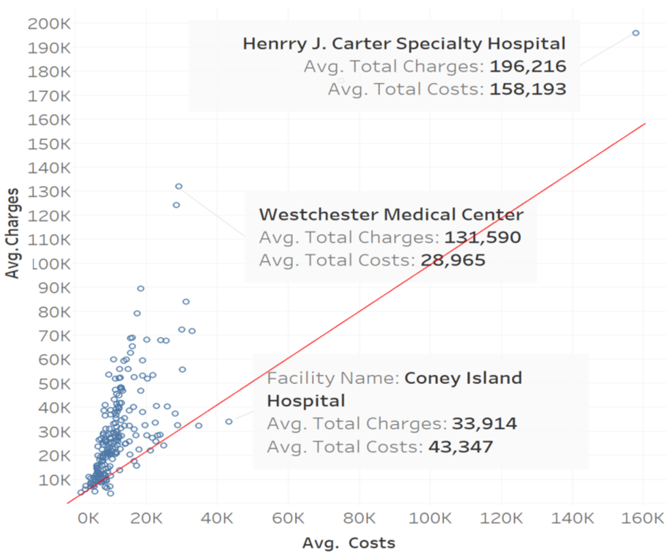


Figure 3.24: Average charge and cost by hospital

The demographic information is collected to study its influence, if any, on the average total charge and cost (per discharge). The color in Figure 3.25 indicates high correlation between income and hospital charges. Higher local income is clearly associated with higher hospital charges in the respective areas.

Figure 3.26 illustrates the relationship between the average median property value and the average hospital charge by county. The average median property value represents the geographic hierarchy distribution. The average property value in Richmond and Queens County are higher, but the lower average charge indicates more affordable medical services. Figure 3.27 shows the relationship between the population and average hospital charge by county. Queens has a higher demand but relatively lower charges, while Albany has a lower demand but relatively higher charges.

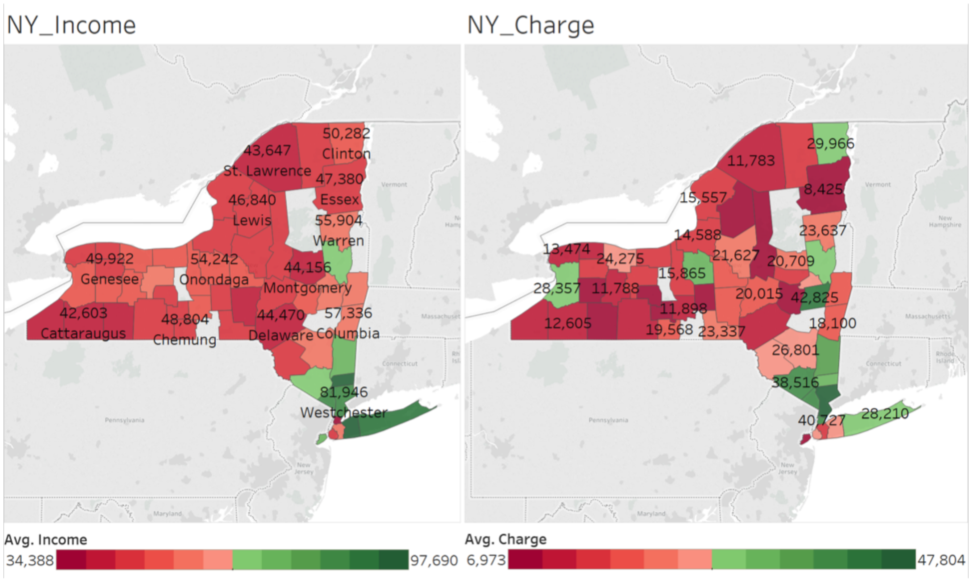


Figure 3.25: Comparison between average income and charge by county

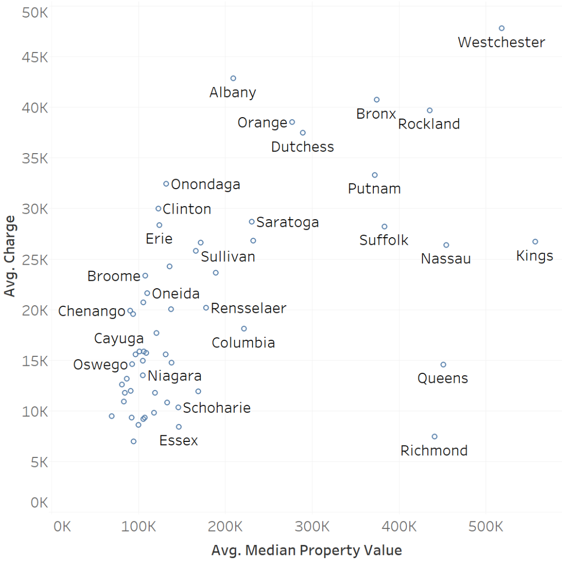


Figure 3.26: The average median property value and average hospital charge

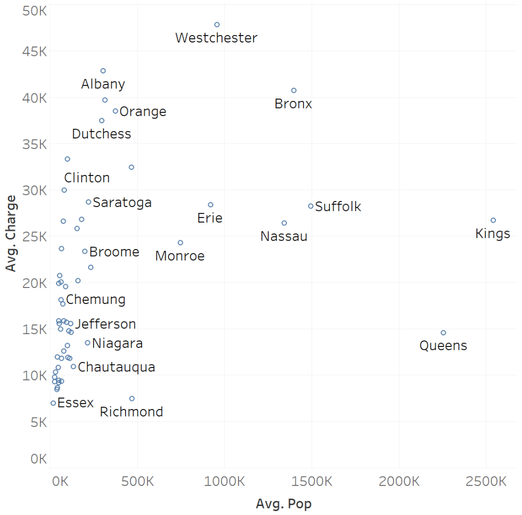
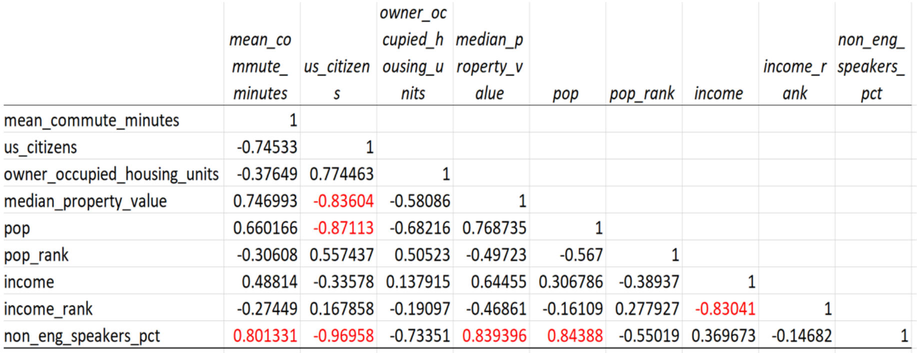


Figure 3.27: The population and average hospital charge by county

The demographic data was collected by counties in New York State. The correlations among demographic features are shown in Figure.3.28. Across different counties, the ratio of U.S. citizens to local population (the variable is named as U.S. citizens) shows a high negative correlation with median property value owner occupied housing unit non-English speaker. And non-English speaker shows high positive correlation with mean commute time. The variables with an (absolute value) correlation coefficient higher than 0.8 are excluded from the analysis. Thus, the independent variables include the number of hospitals per income, mean commute minutes, owner occupation and population. The dependent variable is the average charge (per discharge) by county for model 1 and the average cost (per discharge) by county for model 2. As shown in Table 3.3, according to the regression results, R-squared is 0.59 (model 1) and 0.48 (model 2). The p-value for income and owner occupation is below the common alpha level of 0.05, which indicates that both are statistically significant in both models. The regression results show for both models, the income and owner occupation are statistically significant. Model 1 also shows that the coefficient for income is 0.809, which indicates that for every additional unit of income increase, the average charge can be expected to increase by an average of \$0.809. The coefficient for owner occupied is

-85969.3, which indicates that for every additional owner occupied average charge can be expected to decrease by an average of $85969.3. Finally, the regression results for model 2 show that the coefficient for income is 0.164 and for owner occupied -28945.8.



	mean_commute_minutes	us_citizens	owner_occupied_housing_units	median_property_value	pop	pop_rank	income	income_rank	non_eng_speakers_pct
mean_commute_minutes	1								
us_citizens	-0.74533	1							
owner_occupied_housing_units	-0.37649	0.774463	1						
median_property_value	0.746993	-0.83604	-0.58086	1					
pop	0.660166	-0.87113	-0.68216	0.768735	1				
pop_rank	-0.30608	0.557437	0.50523	-0.49723	-0.567	1			
income	0.48814	-0.33578	0.137915	0.64455	0.306786	-0.38937	1		
income_rank	-0.27449	0.167858	-0.19097	-0.46861	-0.16109	0.277927	-0.83041	1	
non_eng_speakers_pct	0.801331	-0.96958	-0.73351	0.839396	0.84388	-0.55019	0.369673	-0.14682	1

Figure 3.28: Correlation matrix of demographic indicators

Table 3.3: Regression results by county

Variable	**Model 1 (dependent variable: Average total charge)**	**Model 2 (dependent variable: Average total cost**
Income	0.809***	0.164**
	(4.290)	(2.508)
Mean commutes minutes	-405.06	-25.091
	(-1.341)	(-0.238)
Median property value	-0.017	-0.01
	(-0.581)	(-1.02)
Owner Occupation	-85969.3***	-28945.8***
	(-4.687)	(-4.527)
Population	3.28e-3	-3.4e-5
	(-1.038)	(-0.03)
p-value	9.62e-09	2.22e-06
R^2	0.59	0.48

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

t statistics in parentheses

3.4.1 Average Total Charge, Cost vs Patients' Age, Gender & Ethnicity

Changes in population size, age, race, and ethnicity affect the necessary health care resources, the cost of care provided, and even the conditions associated with each group within the population. Health care organizations will have to adapt quickly in order to meet their patients changing needs while addressing health reform requirements. In this section, the relationship between hospital charges and the patient age, gender, and ethnicity is examined. Figures 3.29 and 3.30 show the average total charges for males and females for all diseases or types of cancer, respectively. Each dot represents one three-digit zip area. The dots are divided into two clusters, with one cluster being relatively more expensive than the other. When observed geographically, it can be seen that the more expensive cluster (in yellow) is located closer to the New York City than the remaining one.

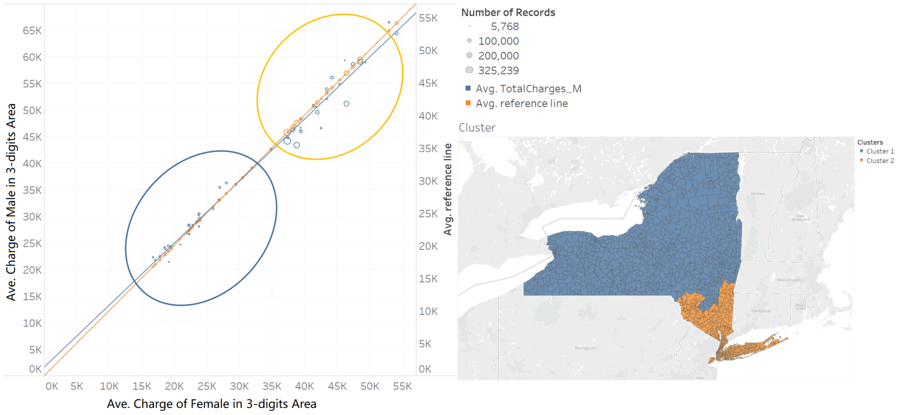


Figure 3.29: The average total charges of male and female across all diseases

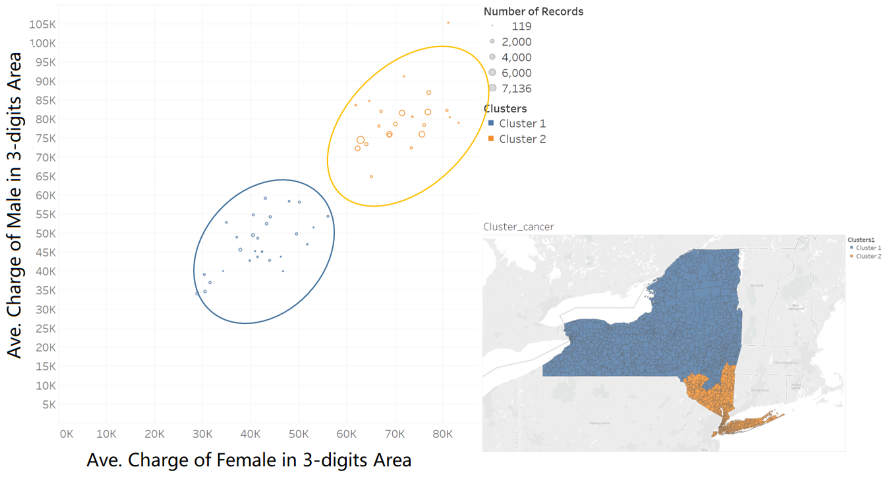
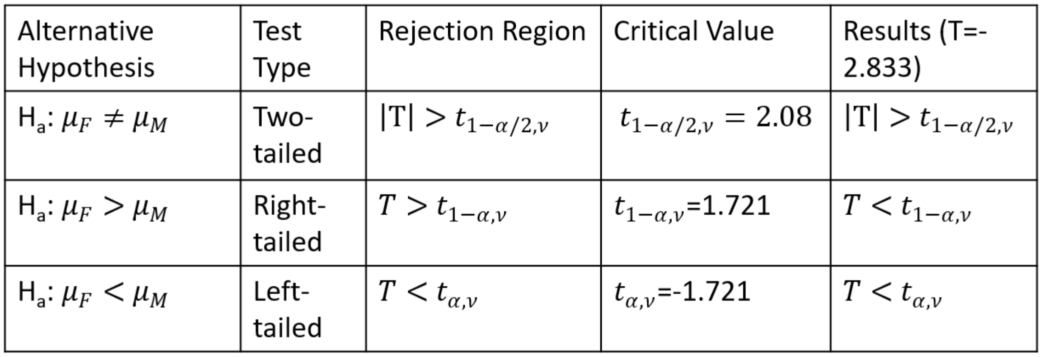


Figure 3.30: The average total charges for males and females across all types of cancer

In addition, the difference between males and females in the average total charges is not great. Two-sample T-test for the mean of the total charges by gender in different zip code areas was carried out based on the following hypotheses:

H_0: The means of total charges for different genders in the three-digit zip codes for all diseases are the same.

H_a: The means of total charges for different genders in the three-digit zip codes for all diseases are not the same.



Alternative Hypothesis	Test Type	Rejection Region	Critical Value	Results (T=-2.833)
H$_a$: $\mu_F \neq \mu_M$	Two-tailed	$\lvert T\rvert > t_{1-\alpha/2,v}$	$t_{1-\alpha/2,v} = 2.08$	$\lvert T\rvert > t_{1-\alpha/2,v}$
H$_a$: $\mu_F > \mu_M$	Right-tailed	$T > t_{1-\alpha,v}$	$t_{1-\alpha,v}$=1.721	$T < t_{1-\alpha,v}$
H$_a$: $\mu_F < \mu_M$	Left-tailed	$T < t_{\alpha,v}$	$t_{\alpha,v}$=-1.721	$T < t_{\alpha,v}$

Figure 3.31: Two-sample T-test for the mean of total charges by gender in different zip code areas

The results of the T-test: T = -2.833 and p-value = 0.005. Figure 3.31 shows that for all three alternative hypotheses, the t-value falls within the rejection regions, and the null hypothesis can thus be rejected. In other words, this implies that the means of the two samples are different. The one-tailed test is left-tailed; therefore, the mean of total charges (per discharge) for males is significantly higher than that of females. In summary, women visit doctors more often than men do, but the average charges per discharge for women are lower than those for men.

After removing the outliers defined as the average total charges over $100,000, comparisons were made and a one-way ANOVA test was conducted for the average total charges across different admission types in different 3-digits area(cf. Figure 3.32):

H_0: The means of total charges across different admission types in different three-digit zip codes for all diseases are the same.

H_a: The means of total charges across different admission types in different three-digit zip codes for all diseases are not the same.

P-value is 3.43E-52, and the null hypothesis is thus rejected. In other words, the means of average total charges across different admission types are not the same.

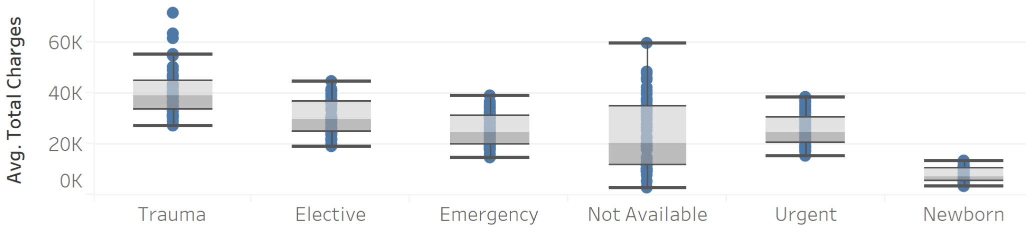


Figure 3.32: Average total charges and costs across different admission types in different 3-digits area

The average total charges across different races in different 3-digits area were compared (cf. Figure 3.33), and a one-way ANOVA test was carried out:

H_0: The means of total charges across different races in different three-digit zip codes for all diseases are the same.

H_a: The means of total charges across different races in different three-digit zip codes for all diseases are not the same.

The P-value is 2.03E-6, and the null hypothesis is thus rejected. In other words, the means of average total charges across different races are not the same.

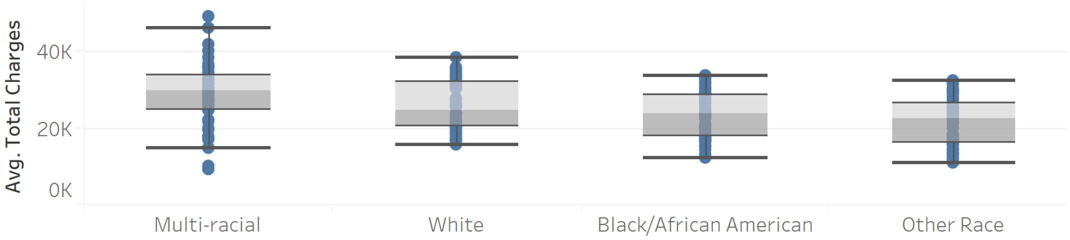


Figure 3.33: One-way ANOVA test for average total charges across different races in different 3-digits area

The average total charges across different age groups in different 3-digits area were compared (cf. Figure 3.34), and a one-way ANOVA test was carried out:

H_0: The means of total charges across different age groups in different three-digit zip codes for all diseases are the same.

H_a:The means of total charges across different age groups in different three-digit zip codes for all diseases are not the same.

The P-value is 8.19E-47, and the null hypothesis is thus rejected. In other words the mean of average total charges across different age groups are not the same.

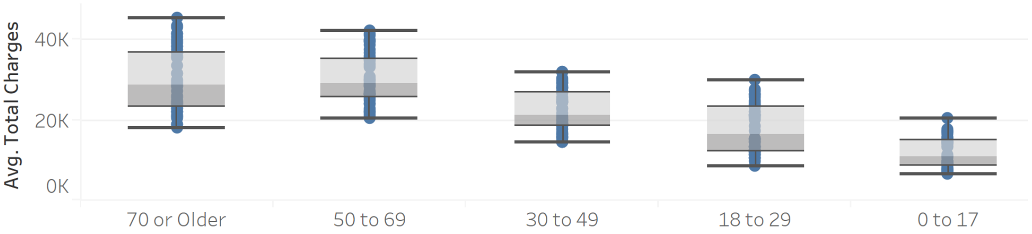
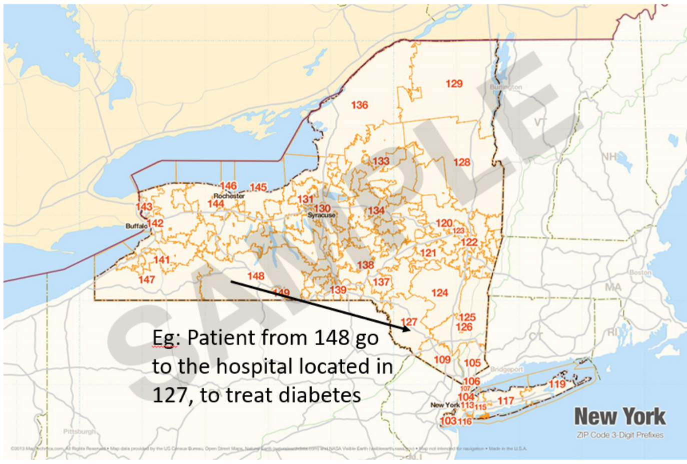


Figure 3.34: One-way ANOVA test for the average total charges across different age groups in different 3-digits area

3.5 Mathematical Model of New Hospital Expansion

3.5.1 Mathematical Model

In considering new clinics or hospital expansion, the questions of where and which diseases are treated to achieve the highest profit as an interesting problem. Figure 3.35 illustrates the general problems: a patient from area 140 visits a doctor in area 127 to receive treatment for an illness. Figure 3.36 and 3.37 shows that the histogram of the average total profits of two examples. The objective is to help the investor find one or multiples sites for building a clinic or hospital for treating a particular illness. Here we assume there is no competition from other hospitals within or outside the same area.



Eg: Patient from 148 go to the hospital located in 127, to treat diabetes

Figure 3.35: Example

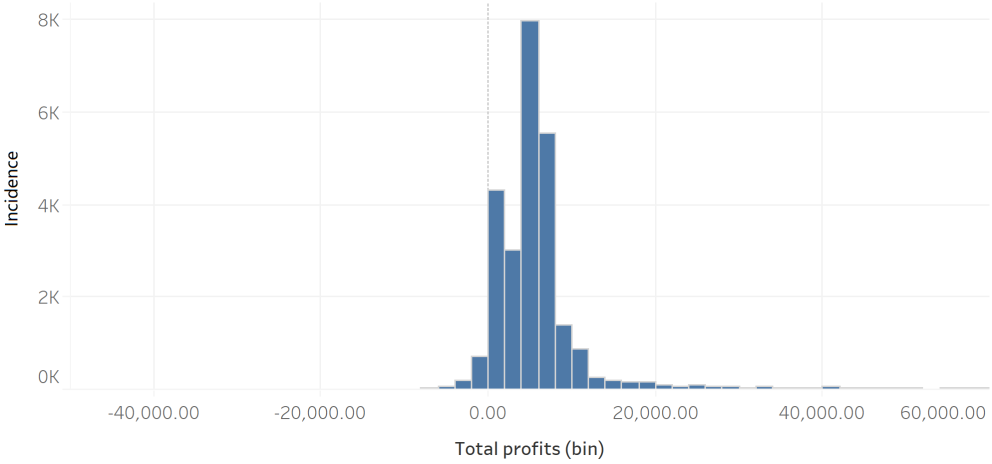


Figure 3.36: The histogram of patients from 112 to hospital located in 112 to treat liveborn infants

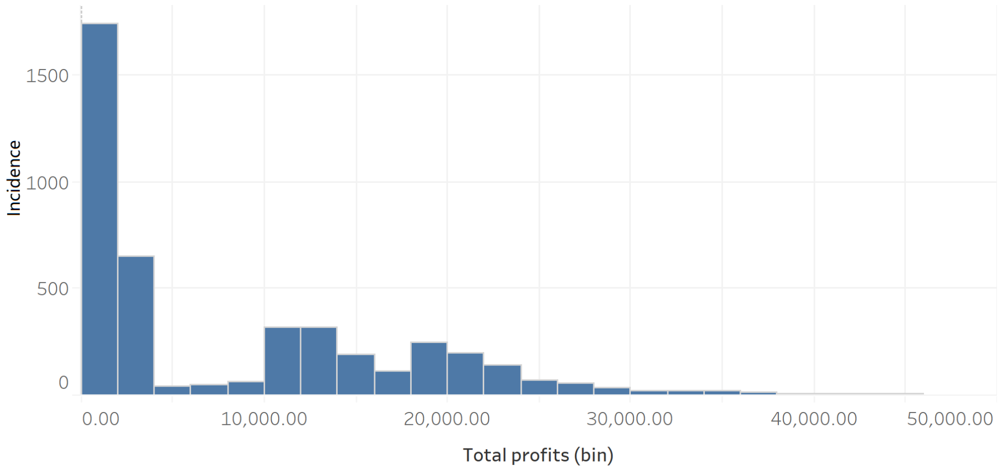


Figure 3.37: The histogram of patients from 140 to hospital located in 142 to treat Osteoarthritis

The notation is defined as follows:

c_{ijk}: the cost of hospital in region j serving a patient from region i on k disease. (Most likely, c_{ijk} is independent of i, so we can use c_{jk} but need to tested).

e_{ijk}: the expected profit earned by hospital in region i by serving patients from region j on k disease.

P_{ijk}: the profit (random variable) for a hospital in j serving a patient from i on k disease;

s_{jk}: setup cost for hospital in region j to treat disease k.

n_{ijk}: the number of patients from region i going to a hospital in region j to treat disease k.

r_0: the requirements for the minimum return S: the limited setup budget O: the limited operating budget

The objective function is to minimize:

$$Min \sum_{j=0}^{J} \sum_{k=0}^{K} x_{jk} \tag{3.1}$$

The decision variable is

$$x_{ij} = \begin{cases} 1 & the\ hospital\ should\ located\ at\ j\ and\ treat\ disease\ k \\ 0 & null \end{cases}$$

Constraint (1): the total setup budget is limited.

$$\sum_{j=1}^{J}\sum_{k=1}^{K} s_{jk}x_{jk} <= S \qquad for\ all\ j,k \tag{3.2}$$

Constraint (2): the total operating budget is limited.

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}x_{jk}c_{ijk} <= O \qquad for\ all\ i,j,k \tag{3.3}$$

Constraint (3): The investment has the minimum return expectation.

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}x_{jk}e_{ijk} >= r_0 \qquad for\ all\ i,j,k \tag{3.4}$$

In addition, an attempt was made to check the value at risk and make sure that the expected return is positive within at least the 95th percentile. The approximation $\sum_{m=1}^{n_{ijk}} P_{ijk}^{m}$ is by normal distribution according to the central limit theorem, and it is assumed that P_{ijk}^{m} is identical for all m values.

$$Profits = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{jk} \sum_{m=1}^{n_{ijk}} P_{ijk}^{m} \qquad for\ all\ i,j,k \tag{3.5}$$

$$VaR(Profits, 0.05) >= 0 \tag{3.6}$$

3.5.2 Numerical Examples

Let us assume the setup budget to be \$500,000, the operating budget \$200,000, and the expected minimum return \$50,000. The respective hospital is located in the 147 area, and the CSS description is Spondylosis; intervertebral disc disorders; other back problems. Upon checking the value at risk, the profits distribution is plotted in Figure 3.38. The five percentile is \$38,714, which means that if an investment is made in a new clinic, there is a 95% probability that the clinic can produce profits equal to or greater than \$38,714.

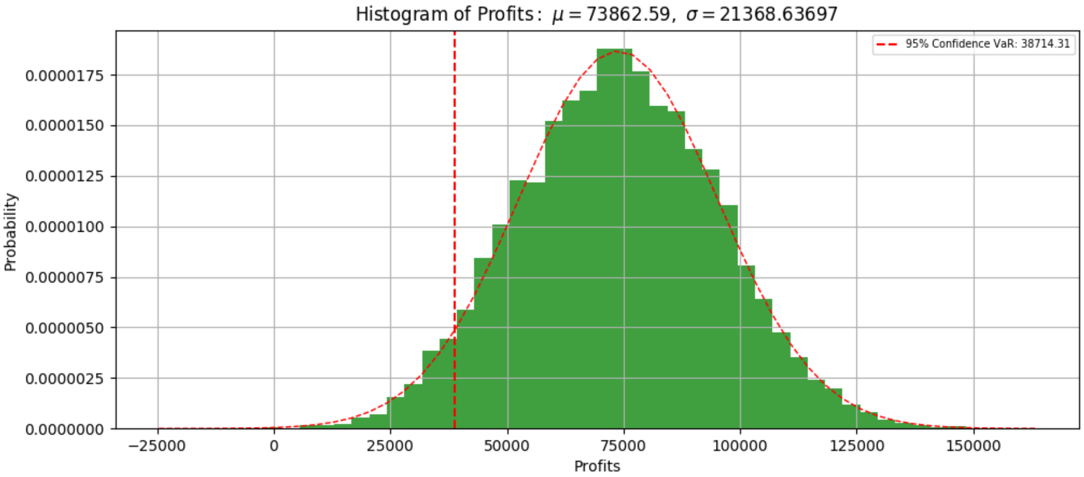
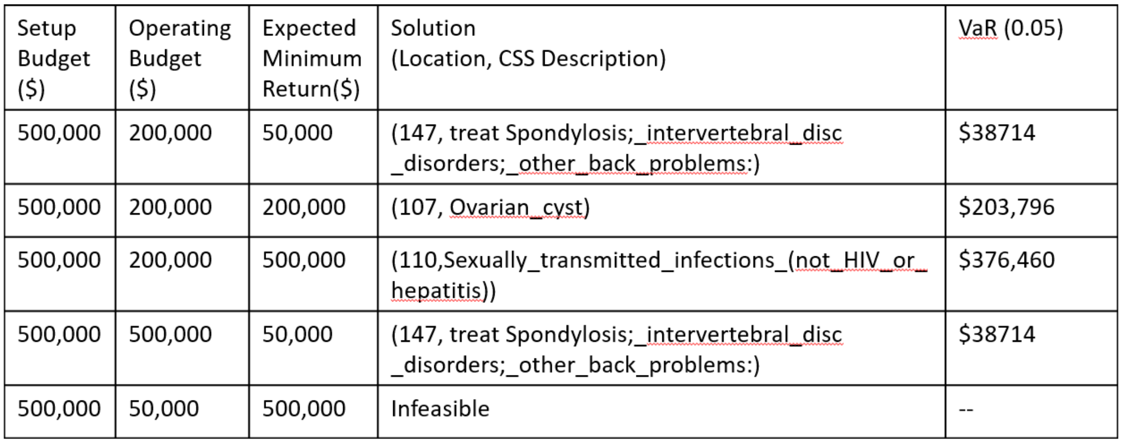


Figure 3.38: Profits distribution

The model parameter sensitivity analysis was also conducted. The unique solutions can always be found in the first four trails and the solution in the last one is not infeasible. The results are summarized in Figure 3.39 including an example on a map (cf. Figure 3.40) and in Figure 3.41 including an example on a map (cf. Figure 3.42).



Setup Budget ($)	Operating Budget ($)	Expected Minimum Return($)	Solution (Location, CSS Description)	VaR (0.05)
500,000	200,000	50,000	(147, treat Spondylosis;_intervertebral_disc _disorders;_other_back_problems:)	$38714
500,000	200,000	200,000	(107, Ovarian_cyst)	$203,796
500,000	200,000	500,000	(110,Sexually_transmitted_infections_(not_HIV_or_ hepatitis))	$376,460
500,000	500,000	50,000	(147, treat Spondylosis;_intervertebral_disc _disorders;_other_back_problems:)	$38714
500,000	50,000	500,000	Infeasible	--

Figure 3.39: Profit distribution for clinic expansion

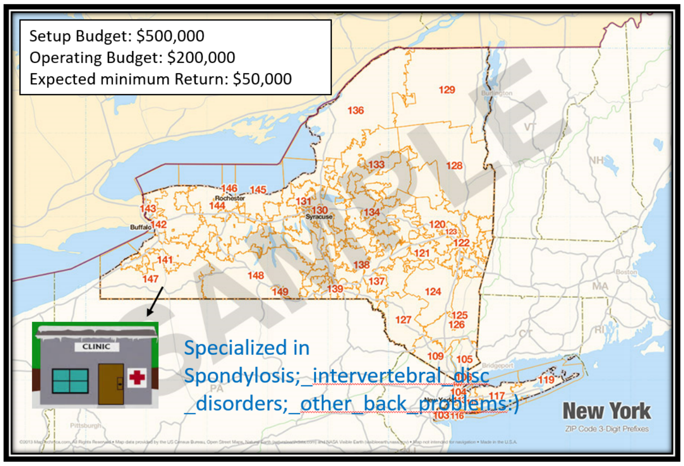
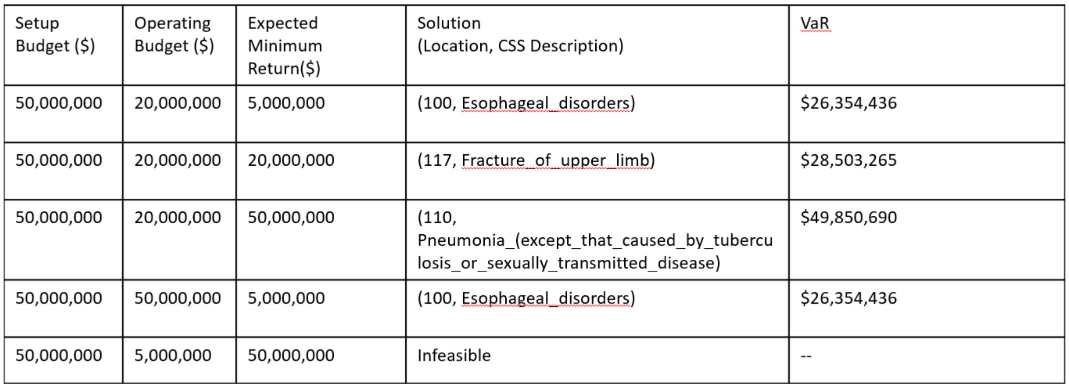


Figure 3.40: An example of clinic expansion



Setup Budget ($)	Operating Budget ($)	Expected Minimum Return($)	Solution (Location, CSS Description)	VaR
50,000,000	20,000,000	5,000,000	(100, Esophageal_disorders)	$26,354,436
50,000,000	20,000,000	20,000,000	(117, Fracture_of_upper_limb)	$28,503,265
50,000,000	20,000,000	50,000,000	(110, Pneumonia_(except_that_caused_by_tuberculosis_or_sexually_transmitted_disease)	$49,850,690
50,000,000	50,000,000	5,000,000	(100, Esophageal_disorders)	$26,354,436
50,000,000	5,000,000	50,000,000	Infeasible	--

Figure 3.41: Profit distribution for hospital expansion

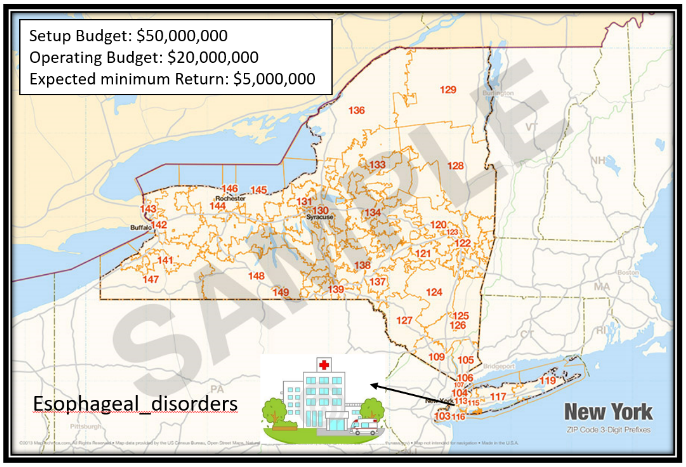


Figure 3.42: An example of hospital expansion

3.5.3 Future Model Extension

The proposed hospital expansion model can be extended in order offer a better description of the investment problems by adding more constraints. Future studies can include, but do not need to be limited to, the following four extensions: 1) Investment fixed number of hospitals ; 2) Treat certain combinations of selected diseases; 3) Add constraints regarding age, gender, and race; 4) Introduce value at risk as a constraint instead of merely checking it. In addition, seeing that the inpatient dataset comprises an additional six years, all of the data can be combined together to include additionally the number of patients as a random variable.

3.6 Conclusion

Despite the hard work of dedicated providers, the health care system arguably remains chaotic, unreliable, inefficient, and crushingly expensive. Seeking quality improvement strategies to provide standardized, safer, and more efficient patient services and medical treatment is the goal pursued by the entire industry. Studying the disease trend and distribution could help improve quality and efficiency, and spur innovation. Meanwhile, exploring key factors that determine the charges could reduce costs. The findings of this paper, based on geographic visualization, indicate the distribution of different diseases and admission types across different areas of the New York State. These findings may provide a useful reference when relocating. The inequities in incidence, prevalence, and complications across gender, race, or ethnicity, and socioeconomic status are also studied. The general gender differences and the gender differences regarding instances of cancer are both statistically significant. In the analysis of differences regarding cancer, interesting is that white male are more probable to get cancer as are people aged over 50. The females are either more prone to illness or they take more care of themselves, as they are tend to visit doctors more often. Nevertheless, the average total charges for males are higher than for females. In addition, during the past seven years, the amounts of payments for medical services have been on the increase. No single service type (such as doctor communication and cleanness) is more statistically significant than reputation and average total cost in the context of the average total hospital charge. The hospital should build the reputation to improve the profits. Furthermore, the average total charges by county depend on the local income, population, and owner occupation. Thus, for patients without health insurance should avoid the hospital located in the rich area. Finally, an optimization model for hospital expansion was built, which offers a good reference for selecting a site and the relevant medical specialties.

Chapter 4

The Empirical Evidence of Second-hand Housing Market : A Case Study of Beijing

4.1 Introduction

The shift from a state-controlled to a market-centered economy in the 1980s has put China on a high growth trajectory. The past three decades have witnessed the rise of the Chinese economy. The International Monetary Fond (IMF) data shows that the real estate industry in China, as one of the pillars of the countrys GDP, contributed with 5% to the GDP in 2000 and its contribution grew to 15% in 2012 studied by Garg (2013). Qualitative analysis has produced many important findings, but the existing studies of the Chinese housing market still lack the support of quantitative research based on the combination of the macro-economic index, monetary policy, and real transaction data.

This study aims to shed light on the comprehensive mechanism and the dynamics on the Chinese housing market. First, the house price fluctuation across neighborhoods within Beijing during the past seven years is examined. An attempt is made to gather macro-economic empirical evidence that offers an insight into the housing market boom. Then, physical attributes such location, square meters, number of bedrooms, and other details are considered within the model. Housing policy and other relative determinants of the house price will also be taken into consideration. The innovation of real estate as a special commodity has both commodity and financial attributes. Its price not only depends on the rigid demand, but it is also influenced by the monetary policy such as M2, government policies such as loans and interest rate adjustments . The present study focuses on the Chinese properties market, which is not a perfectly competitive market, but one of considerable volume and potential. Any purchase restrictions could potentially freeze the market overnight. Such a context offers an excellent opportunity for studying the influence of government policies on the market. The boom in big data analytics in recent years has enabled obtaining real-time transaction data from the real estate brokerage firm through web crawlers. From the marketing perspective, such an approach enables gathering solid evidence.

4.1.1 Background

Chinas residential housing market has seen a rapid growth over the last decade. With the annual growth of 19.4% in trading volume, 2016 has become the fastest growing year since 2010. The property prices per square meter in big cities such as Shanghai, Shenzhen, and Beijing nearly doubled in 2016. Meanwhile, statistics by Peoples Bank of China at the end of 2016 show that more than half (50.03%) of the new loans in the said year, that is, 12.65 trillion CNY, were attributed to mortgages. The objective of this study is to throw light on the housing market trend in China for the purpose of helping the policy makers assess and manage the housing bubble on the one hand, and the investors to design better investment strategies on the other. The aim is to collect data and develop advanced models to identify relationships among housing price, macroeconomic policies (e.g. monetary policy), and physical attributes.

4.1.2 Analysis of Demand and Supply Sides

Based on the elasticity of the housing demand, buyers can be categorized into four classes. (1) First, there are rigid demand buyers purchasing their first house. In China, people traditionally strive to become house owners. Renting an apartment does not provide couples with a sense of security. Second, due to the current mild mortgage interest rates, individuals often choose to buy apartments instead of renting considering the comparative prices of the two. (2) There are also families who want to improve their living conditions. With the rapid increment of income in China, a growing number of families expect to enjoy high life quality and to find larger apartments. (3) The third group includes speculators who only expect to shortly hold the assets. Due to the sharp increase in housing prices, especially compared to the sluggish stock market, speculators could potentially make healthy profits. (4) The last group is composed of investors who are professional in asset management and normally hold the assets in long term to hedge the risk of depreciation of other assets. In the past 20 years, the price of real estate has generally been sharply increasing. Stagnant periods or those experiencing slight drops occurred when new policies were introduced to limit speculation. It is widely regarded that property investment has become an arbitrage investment.

4.1.3 Broker's Special Status in China

Unlike in the U.S., Chinese brokers represent either the buyer or the seller in a real estate transaction. The broker is the medium of the transaction, and does not represent either of the sides but charges the fee once the transaction is complete. Therefore, they may be strongly motivated to promote the transaction, seeing that they profit only from making deals. Against this background, the real estate brokers may sometimes play the role of speculators, and thus stimulate the investors to buy the residential real estate, after heating the housing market.

Rumors are ways of transmitting information among people without verifying their veracity. This is done in a number of ways: overall , electronically, or through media outlets. Rumors can become tools in the hands of brokers, who are speculating the house price. During the past two years , a number of rumors have circulated in social media. The first is related to the slump in transactions. Actually, an occurrence of a slump is time-dependent. For example, monthly transactions normally decrease at the beginning of the year due to the spring festival, and the transaction data lags behind. The February surge is, therefore, inevitable. Prices affect the psychology and emotions of market participants. Brokers exaggerate the market fluctuation, and create panic among the buyers. The second rumor is that an increasing number of people have left large cities to return to their hometowns in order to buy an apartment. A smoke screen is thus created, and support is given to the argument that local apartments have high investment values. Conversely, however, many people have attempted to sell their houses in towns and get down payments for the apartments in large cities. The third rumor is that in large cities, such as Beijing and Shanghai, it is becoming more difficult to buy and sell housing properties. Consequently, creating pressure to close housing deals through for instance low interest rates leads to policy restrictions. The fourth rumor is that the economists predictions should be used. During the past twenty years, the economists and senior executives in the real estate industry made millions of predictions regarding housing prices. The options were divided into call and put. However, it seems that the macro-policy is unpredictable; moreover, it is difficult to compare the Chinese housing bubble with the ones in other countries. The Chinese housing market should have broken out , however, nothing happened. This indicates that there is some kind of a market law in force that is different from other countries and its workings remain unclear. Brokers are proficient at using bullish predictions to make the public believe that housing prices will increase and prompt people to buy real estate.

Brokers create fabricated listings to mislead the customers. Striking is that regardless of whether the said listing is cheaper or more expensive compared to the average price, both potentially result in the profit of the brokers. Some realties are listed at a low price for the purposes of phishing and attracting new customers. Once the customers show interest, other sources are recommended and the details of their demand gradually become clearer. In order to recognize such deceptive attempts, buyers should be cautious after noticing that the price is considerably lower than that of the other listings in the same area. Other factors that should be taken into account are whether the platform is committed to offer false compensation , whether the pictures match the apartment information, or whether any details are included under the listing description. On the other hand, the brokers may post several listings with the same layout in the same district. One of the listings may be fictitious but priced higher than the others. When the owner and the buyer see such a listing, their price expectations increase.

Brokers offer consulting services to help avoid policy restrictions. Interestingly, the customer demands are varied. Whereas some property buyers are seeking divorce others are entering marriage. Couples rushing to divorce tend to buy cheaper properties. Due to the current policies in some cities, if a family wishes to buy a second apartment, the down payment rises from 30% for the first home to 70% for the second. To improve the transaction and make the apartment more affordable, brokers encourage an increasing number of couples to divorce their spouses in order to qualify as first-time buyers. There are some purchase restrictions and apartments can be sold only to local residents. One broker apparently used his local residence got marries and devoice with his clients four time in a single year, only to help them qualify for the restriction . In addition, another directional blasting strategy is to use the legal loopholes. For example, buyers and sellers can sign loan contracts with the property pledged as collateral for the loan during the agreed repayment time. After the contract expires, following the court request for mediation, the housing authority acts as a mediator to the buyer in the transfer, and then signs the new loan contract with the bank pledging the property as collateral for the payment to the seller.

Brokers also offer the capital to promote the transactions. Brokers may offer bridge loans. A bridge loan is essentially a short-term loan received by a borrower to finance purchasing a new property. A speculator invests in the property and sells it in a short amount of time to make a profit. When a speculator cannot afford the down payment, additional external capital is necessary in form of, for instance, a bridge loan. Such a loan can be considered as a down payment loan. Many critics view bridge loans as highly risky. The borrower essentially takes on a new loan with a higher interest rate, and there is no guarantee that the property will sell at a competitive price within the allotted time, unless the market prices continue to increase. A similar pattern can be observed when the speculator receives a down payment from a crowd fund. For all that, brokers use their professional knowledge to hoard and profiteer.

Buy low, sell high is a well-known investing adage, which refers to taking advantage of the markets propensity to overshoot on the downside and upside. Although it is very simple, it is difficult to execute. For brokers, however, its execution is easy, primarily due to their monopoly on resources. All listings have to go through the realty system, which means that brokers have first-hand information on properties and can prevent a target house from entering the market. Second, depending on the brokers professional experience, they can easily evaluate whether the price determined by the owner is overestimated or underestimated. If underestimated, the broker could hoard it, and wait for the opportunity to make another sale.

4.1.4 Literature Review

The housing market has received a lot of attention from researchers. Macroeconomic studies have focused on the link between the increase in housing wealth, financial wealth, and consumer spending studied by Case et al. (2005). Goodhart and Hofmann (2008) carried out panel vector auto-regression to explore the links among money, credit, house prices, and economic activity in industrialized countries over the period of three decades. Yue and Hongyu (2004) investigated the city-level interactions of housing prices and economic fundamentals in 14 cities in China between 1995 and 2002.

The hedonic price method known as hedonic regression is used in consumer and market research, studied by Holbrook and Hirschman (1982) and in the calculation of consumer price indices studied by Moulton (1996). Lancaster (1966) established microeconomic foundations for analyzing characteristics and applied it to the housing market. Law (2017) estimated the geographically weighted effect on housing prices in London. Swoboda et al. (2015) conducted hedonic analysis over time and space. Diewert et al. (2011) provided a review of the capitalization of school quality into house values. Diewert et al. (2011) contributed to the decomposition of a house price index into land and structures components through a hedonic regression approach. In recent years, machine-learning algorithms are widely used in housing price prediction. Limsombunchai (2004) compared the hedonic price model and the artificial neural network. Park and Bae (2015) applied C4.5, RIPPER, Nave Bayesian, and AdaBoost to compare their classification of accuracy performance.

In addition, the impact of tax policy and demography on housing prices and environmental variables were studied by Poterba et al. (1991) and Boyle and Kiel (2001). Lynch and Rasmussen (2001) measured the impact of crime on housing prices. Several studies have found that certain segments of the housing market display more volatility than others, even at very specific geographic levels. For example, Guerrieri et al. (2013) examined zip code level data and found the areas on the boundaries of affluent neighborhoods to be the most price sensitive (i.e. rising the highest during booms and falling the lowest during busts), while Landvoigt et al. (2013) found that the individual homes in poorer San Diego neighborhoods appreciated fastest during the run-up to the housing bubble. Case and Shiller (1990) used quarterly indices of existing single-family home prices estimated with micro-data on properties that sold more than once in order to estimate excess returns to investment in owner-occupied housing. Finally, Jianglin (2010) studied the bubble in the urban Chinese housing market.

4.1.5 Data collection and motivation

The physical features data and real-time trading data on second-hand apartments in Beijing was collected from the largest real estate brokerage firm, Lianjia in China, which included millions of records. Additionally, the macro-level data was collected from the official government websites and other financial institutes. The motivation behind the present analysis is in its potential of detecting the housing market trend in China and helping the investors choose the appropriate investing strategy and policy makers to regulate the market. The present research question is, but is not limited to the following: How to build models to estimate the future price on a macro and a micro level?

Real estate is a special type of commodity including both commodity and financial attributes. Its price might depend not only on the supply and demand, but also on monetary policies such as M2, and government policies such as loans and interest rate adjustments. This study aims to make significant theoretical and practical contributions. First, it focuses on the Chinese real estate market, which is not fully market-oriented, but mainly influenced by factors related to government policies and speculative demand. Second, following the recent developments in big-data analytics and machine learning in recent years, real trading data can be obtained, which offers solid evidence that the physical attributes of a house play a significant role in pricing.

4.2 Macro-economics Analysis

The importance of the interactive nexus between the housing markets and macroeconomy has recently received growing recognition. The aim of this section is to explore the interdependence between the second-hand housing prices based on macro-economic determinants.

4.2.1 Economics Index

The gross domestic product (GDP) is one of the primary indicators used to gauge the health of a countrys economy. It represents the total dollar value of all goods and services produced over a specific time period. It is often conceived as an indicator of the size of the economy. While Western economies have floundered in the aftermath of the global financial crisis in 2008 4.1, Asian economies, and the Chinese economy are continuing to flourish. The growth of Chinese economy has always been sustained at no lower than 6% as Figure 4.2 shows. A study by Chow et al. (2008) has confirmed the positive relationship between the housing market, the GDP, and disposable income. Madsen (2012) agrees that a strong short-term relationship exists between housing market and GDP; in the long term, however, this nexus weakens. Adams and Füss (2010) found that the GDP growth has an increasing impact on the housing market. Peng et al. (2008) suggested a two-way linkage between GDP and the growth in housing prices.

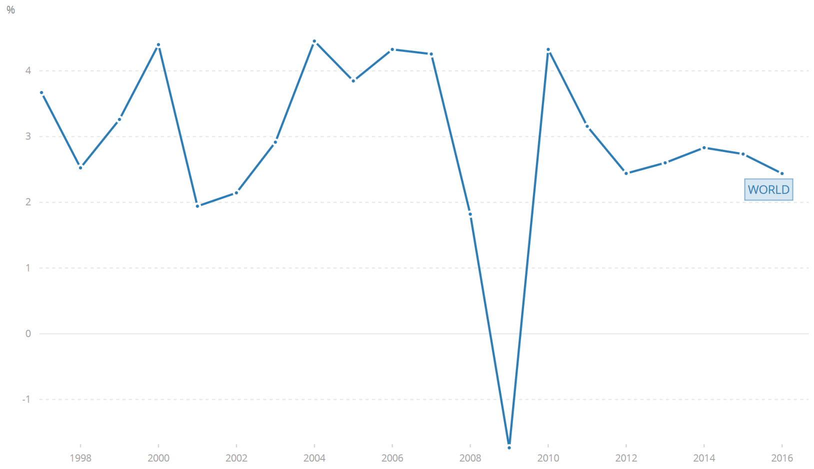


Figure 4.1: GDP worldwide average studied by WorldBank (2017)

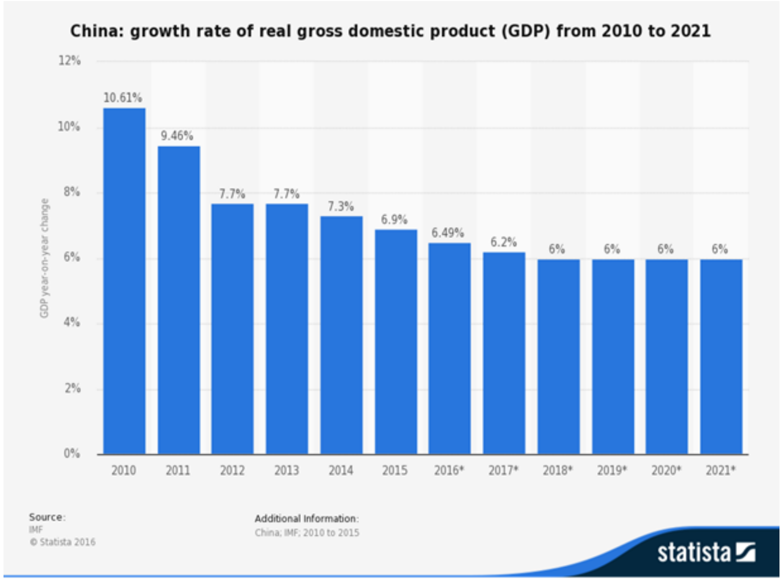


Figure 4.2: China: Growth rate of real GDP from 2010 to 2021 studied by IMF (2017)

Meanwhile, the living standards have caught up with the increase in GDP, and the disposable income has increased sharply. As shown in Figure 4.3, the disposable income of Beijing residents per capita nearly doubled when the data from 2011 and 2016 were compared. The rate increased annually from 8.4% to 13.2%.

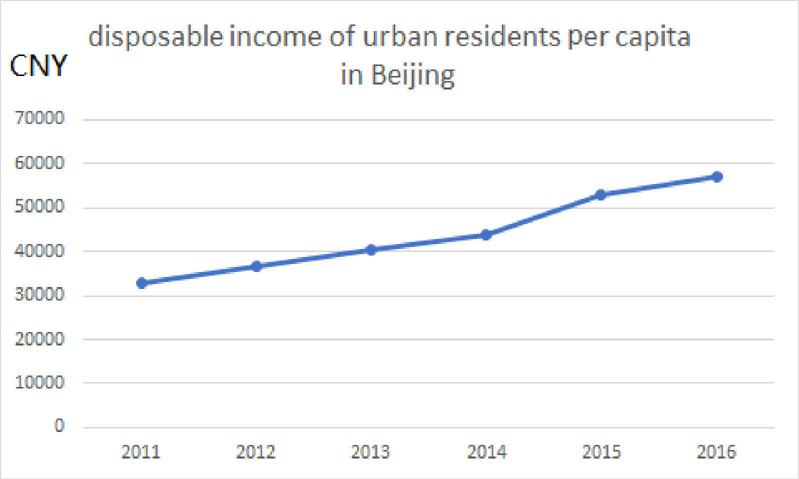


Figure 4.3: Disposable income of urban residents per capita in Beijing from NBS (2017)

4.2.2 Monetary Policy

In general, the monetary policy has impacted the real estate market through two main channels: money supply and interest rates. Figure 4.4 shows the M2 increment annually sustained above 10% from January 2011 to January 2017. Normally, with the growth of money supply, the relative cost of housing drops and the demand for housing rises. Zhang (2014) showed that the rise in the money supply is strongly supported by the great increase in the property development and individual housing consumption. Additionally, the money supply supports the credit line for the housing industry by contributing to the fast growth of individual housing mortgage loans. According to an opposing view studied by Peng et al. (2008), however, bank credit expansion arguably did not accelerate the property price inflation.

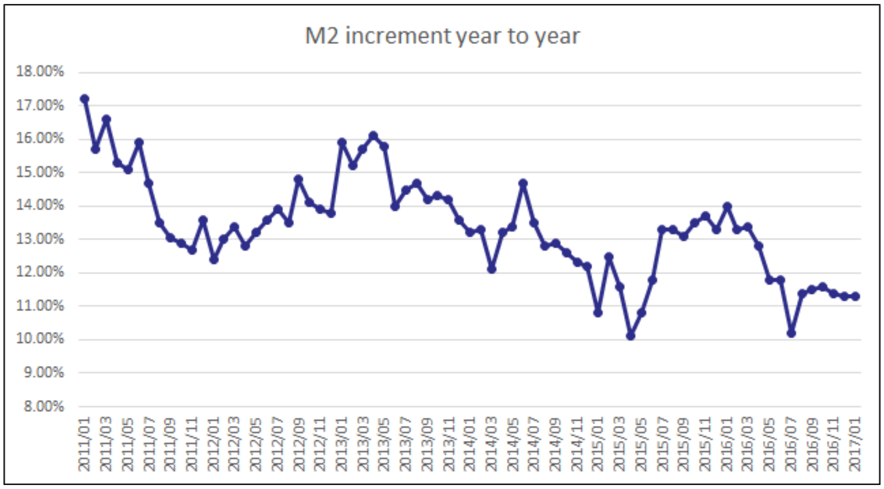


Figure 4.4: Change in M2 money supply in China from Eastmoney (2017)

Besides money supply, interest rates also play an important role in the context of the housing prices. According to Tsatsaronis and Zhu (2004), housing prices depend on inflation and credit, which are strongly linked to short-term interest rates. Andrews (2010) argued that the correlation between housing prices and the loan interest rate is negative. Figure 4.5 charts Chinas lending rates of 5 years and above (that is, commercial mortgage) between 1996 and 2015. We can observe a decreasing trend for the overall interest rates. Since 1999, there have been two rate hiking cycles and two rate cut cycles. Nowadays first-time home-buyers in Beijing have to pay a minimum benchmark mortgage rate of around 4.9 percent.

From the point of view of consumer psychology, home-buyers acquainted with historically high interest rates may seize the unprecedented opportunities with much lower financial costs than before. Moreover, the housing market boom is normally not driven by economic prosperity, but by recession. There are few investment opportunities in real economy, so the capital influx into housing market . Ordinary people prefer to choose relatively safe assets, such as real estate, for investments. The leaders such as banks are also seeking cushions against risk. Mortgage is a safer asset when compared to a commercial loan, which provides lenders with a higher incentive to loosen the standards and offer loans.

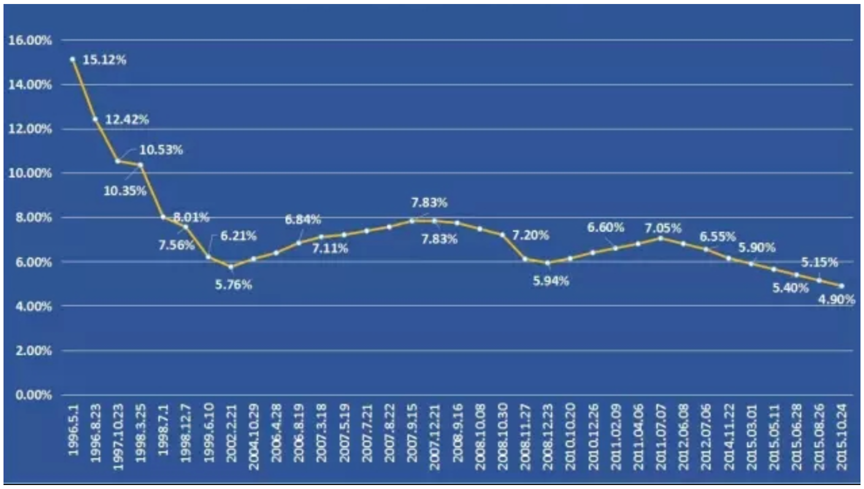


Figure 4.5: Interest rates of commercial loans between 1996 and 2015

4.2.3 Taxation

China introduced an annual residential property tax on some homes in Shanghai and Chongqing in 2011, but it has not yet seen an increase. Until now, China has had no plans to implement a nationwide property tax, despite the high expectations that such measurements would restrain surging property prices. It is beyond doubt that the properties taxation could deter speculation in real estate. The lack of any annual taxes on payments means that it is more sensible for investors to leave homes empty rather than negatively affect resale values by renting out the homes in the period in which the prices are rising. However, little progress has been made due to the resistance from stakeholders, such as local governments, who heavily rely on land sales for revenue. Only in 2015, the revenue of land transfer reached 193 billion CNY (approximately 29 billion USD).

4.2.4 Local Restrictions on Property Purchases in Beijing

Between January 2011 and January 2017, local governments in Beijing announced three times new restrictions on property purchases, which were designed to dampen speculative buying and curb soaring prices.

In February 2011, a restriction called The Beijings 15 Guidelines was introduced to cool the residential property market by banning the purchases of second and third homes. According to the restriction, a family of Beijing residents who already owned an apartment unit or who held a valid temporary residence permit in Beijing without housing and paid social insurance or personal income tax of non-Beijing residents for more than five years (including) could purchase an apartment (including new commodity housing and second-hand housing). The transactions of families of Beijing residents who already owned two apartment units in Beijing, or of the ones who did not have a valid temporary Beijing residence permit and paid social insurance or personal income tax in Beijing for more than five consecutive years (inclusive) were suspended studied by Beijing-Daily (2011). In March 2013, the new restriction called national five guidelines was announced to curb trading by improving the purchasing requirements and the transaction costs. A special restriction was enforced for Beijing introducing a ban as of March 31 for single persons to buy their second suite. The government thus collects a 20% transfer duty, levied on the value of the property sold by the seller. In addition, the banks also increased the interest rate of mortgage loans on the purchases of the second apartment studied by Baidu (2013). The Beijing government announced the 930 restriction on September 30, 2016. It aimed to increase the land supply and differentiate housing credit policies to control prices. In purchasing the first ordinary owner-occupier apartment, the down payment ratio should thus not be lower than 35%, and in purchasing the first ordinary non-owner-occupier apartment, the down payment ratio should be not be lower than 40%. For those households that already own an apartment, but wish to improve their living conditions by applying for commercial individual housing loans to buy ordinary self-housing, with or without loan records, the proportion of down payment should not be lower than 50%; in purchasing non-ordinary self-housing, the down payment should not be lower than 70%.

4.2.5 Other Relative Determinants of House Price

Case and Shiller (2003) highlighted the strength of the housing sector since the stock market crash of 2000 and the recession of 2001. The same trend was detected in China. Since 2011, the Shanghai (Securities) composite index kept decreasing and it remained low for a long time; the capital relocated from the stock market to real estate. The situation remained the same until the turbulences occurred in the Chinese stock market between October 2014 and December 2015, which revealed the snowballing effect of the actions of disparate players. The plunge of 2015 is not just the result of an ordinary correction after a year of great gains, but of a high leverage speculation. With the improvement in life quality and an income increase, the growing middle class is eager to protect their capital, and there is a great demand for ideal investment targets. However, the said group tends to view the market as speculative and risky, and the investment environment as poor. A growing number of people turned to the housing market, which poses fewer risks and ensures stable returns.

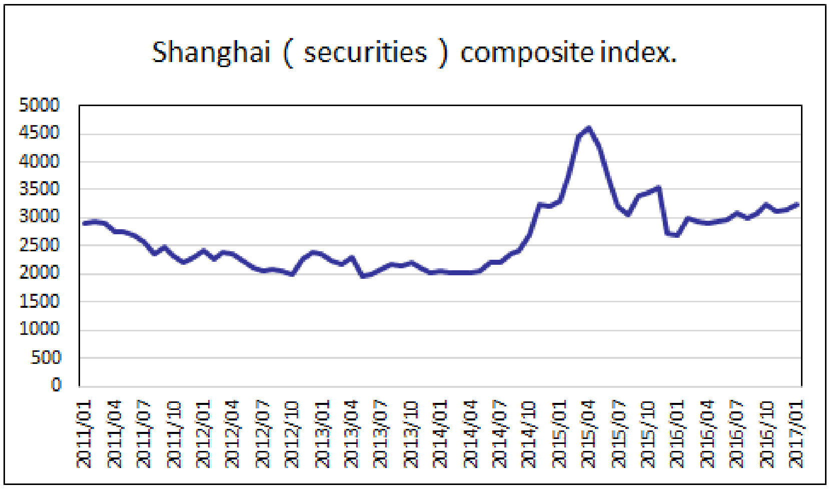


Figure 4.6: Shanghai securities composite index

In addition, Miller et al. (1988) examined the effect of exchange rates (JPY/USD) and Japanese buyers on selected residential market prices and turnover. Benson et al. (1999) studied the constant-quality of the Bellingham housing price index as the dependent variable in a reduced-form model of market price to estimate the impact of the exchange rate. Consequently, the present study also focuses on the influence of the exchange rate from USD to CNY in the housing market.

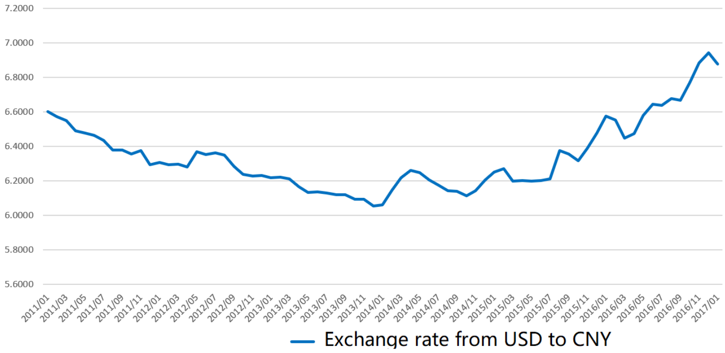


Figure 4.7: Exchange rate from USD to CNY

4.2.6 Regression Analysis

The linear regression model is as follows. The dependent variable is the unit of the housing price and the independent variables include increasing rate of long-term residential loan, M2 increasing rate, domestic and foreign currency deposits increasing rate, Shanghai index increasing rate, and the exchange rate from USD to CNY. The regression analysis was applied in two ways. First, all the macroeconomics indicators are taken for the same month as the unit housing price. Second, the unit-housing price was shifted three months after the macroeconomics indicators to examine the macroeconomic impacts. As is shown in Table 4.1, in model 1, the dependent variable is the increasing rate of average unit price m^2; in model 2, the dependent variable is the increasing rate of average unit price per m^2 but shift three months. Only the exchange rate from USD to CNY is statistically significant in the shift three-month model, which indicates that for every additional exchange rate from USD to CNY, the unit housing price was expected to increase by an average of CNY 0.08.

$$\begin{aligned} housing\ price = \alpha_0 &+ \beta_o(increasing\ rate\ of\ long\text{-}term\ residentialloan) \\ &+ \beta_1(M2\ increasing\ rate) \\ &+ \beta_2(domestic\ and\ foreign\ currency\ deposits\ increasing\ rate) \\ &+ \beta_3(Shanghai\ index\ increasing\ rate) \\ &+ \beta_4(exchange\ rate\ from\ US\ dollar\ to\ CNY) \end{aligned} \tag{4.1}$$

Table 4.1: Regression results by macroeconomics indicators

Variable	**Model 1**	**Model 2**
Increasing rate of long-term residential loan	3.145e-3 (0.332)	5.77e-4 (0.068)
M2 increasing rate	2.078e-1 (0.495)	0.483 (1.317)
Domestic and foreign currency deposits increasing rate	-9e-5 (-0.135)	1.04e-3 (-0.597)
Shanghai index increasing rate	4.44e-2 (0.526)	9.259e-2 (1.245)
Exchange rate from US dollar to CNY	4.414e-2 (1.433)	8.605e-2*** (2.849)
p-value	0.783	0.120
R^2	0.035	0.12

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

t statistics in parentheses

4.3 Housing Price and Investment Analysis

In the previous section it was indicated that, between January 2011 and January 2017, the average annual increasing rate of M2 was 13.38%, which was equal to the annual return rate of 19.25%. Moreover, the annual increasing rate of disposable income of urban Beijing residents was 12.65%. A hedge strategy is needed to maximize the profits through protection against inflation caused by the constant high money supply. The target assets include, but are not limited to, real estate, gold, oil, stocks, and inflation-indexed bonds. Let us examine the annual return rate of some typical assets during the same period. First, as mentioned above, the annual return rate of Shanghai (securities) composite index is only 1.9%, which implies that probably even a good trader would not be able to identify the alpha return. The respective trend can be characterized as unpredictable . Retail investors were engaging in high-risk activities with the speculators and the government. On another note, on the American stock market, the annual return rate of Amazons stock price is 63.36%. In the emerging market, Bitcoin represents one of the currently attractive investment assets with the annual return rate of 30573.36%. As is shown in Figure 4.9, the price oscillations for investors are highly stressful. The crypto-currency may have created a dangerous speculative bubble . Most investors, however, are pursuing long term, stable returns. From this perspective, real estate investments represent a better option. The average price per square meter and the median price per square meter in Beijing are shown in Figures 4.10 and 4.11 respectively. With the annual return of 30.92%, if the average unit price per square meter is considered, this type of investment is competitive against other similar investments but it includes lower risks and good liquidity. Actually the housing prices in Beijing in the past 30 years have never plummeted, although they did shortly decrease immediately after the government launched some purchase restrictions. There is no doubt that the residential real estate generally represents a good investment target. However, purchasing an apartment is usually one of the most important decisions a person makes, since it is often among the largest and most expensive purchases. Offering the homeowners a way to evaluate the assets is of great value. In the next section, the factors influencing the housing prices are discussed.

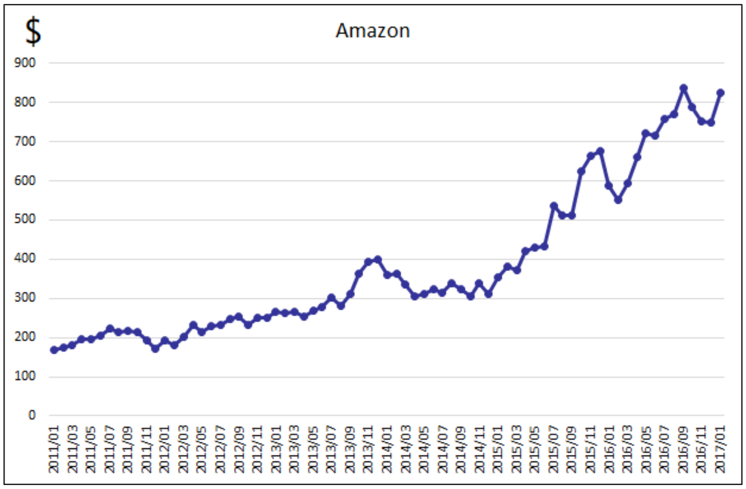


Figure 4.8: Amazon stock price changes

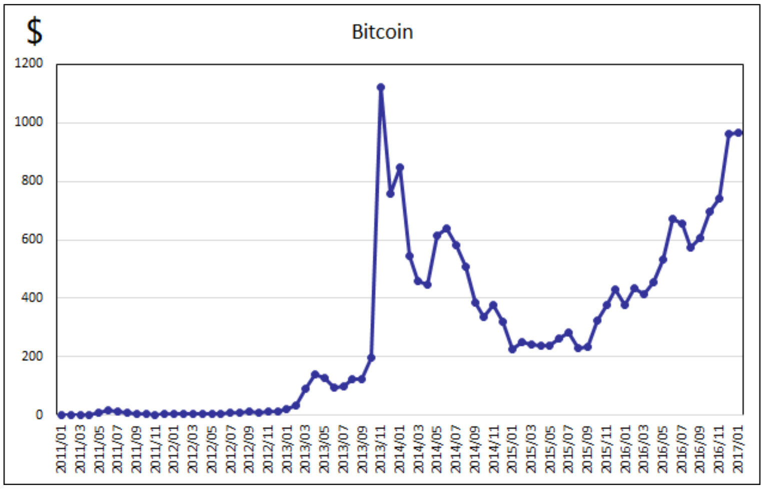


Figure 4.9: Bitcoin stock price changes

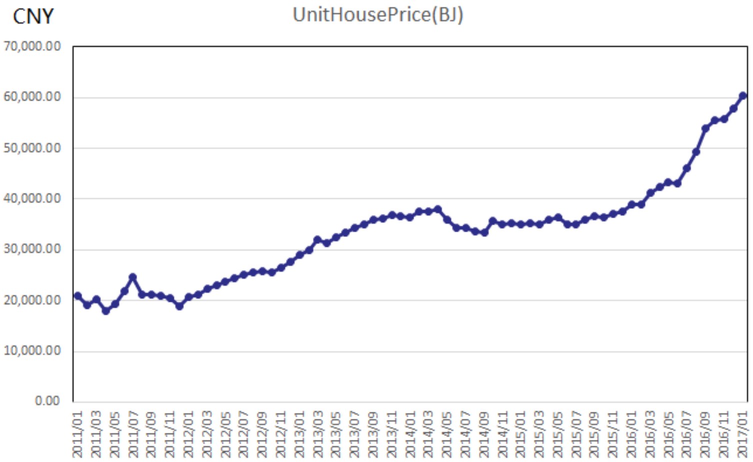


Figure 4.10: The average price per square meter in Beijing from 2011 to 2016

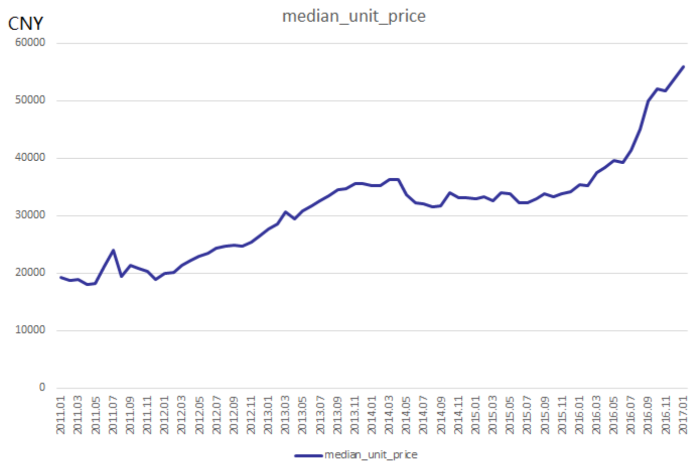


Figure 4.11: The median price per square meter in Beijing from 2011 to 2016

4.3.1 Housing Types in Beijing Second-hand Market

On the Beijing second-hand housing market, apartments can be divided into non-residential and residential. The non-residential apartments, according to their residential use, include, for example, office buildings, shops, garages, and debris. Residential apartments can also be divided into high-end residential and ordinary residential apartments depending on their location. Depending on the source, housing can be divided into commercial housing, housing reform housing (Purchased public housing and Purchased central government owned housing), affordable housing, right-to-use public housing, and private property.

- Commercial housing: refers to the type of housing owned by the real estate development company, which has invested in the construction for the purpose of profit, and according to the laws of the housing business. From the sales perspective, commercial housing can be divided into existing housing and housing under construction. From the sales object perspective, the housing is divided into domestic and export housing. Based on its use, it can be divided into ordinary houses, apartments, and villas.

- Purchased public housing: the type of housing affected by the housing reform brought about by the national housing policy. Employees buy the house according to the cost or standard price. The original property belonged to the companies or to the housing projects responsible for the construction.

- Purchased central government owned housing: the type of housing affected by the housing reform introduced through the national housing policy; the employees buy the house according to the cost or standard price. The original property belonged to the central Beijing government, state-owned large and medium-sized enterprises, the public housing State Council departments, housing projects, or fund-raising constructions of the central government.

- Affordable housing: this type of housing is intended for low-income families. According to the national housing construction standards, ordinary residential are guaranteed. Land supply is in principle distributed by the government. Housing design reflects the standard economic and aesthetic principles, while satisfying the basic needs of residents. The price of affordable housing is determined based on the construction costs and according to the administrative department of affordable housing.

- The right to use public housing: refers to housing investments made by the state or state-owned enterprises or institutions. The owner has only partial property rights. The government rents public housing to the residents. Commonly known as small property rights, the owners have only partial housing property rights. They can only live in the property, but cannot transfer, rent, or award it to a third party. The most significant feature of this type of housing is that it includes rent.

- Private property Cottage: also known as a private residential, is a house bought or built by a private person or a family.

The present study focuses on the second-hand housing trading. It includes all of the housing types mentioned above the cost of which differs significantly depending on whether they have had previous owners.

4.3.2 Descriptive Analysis

In what follows, prediction models are introduced for particular homes based on the information relating to their physical properties. The analysis aims to assist the first-time buyers, or someone expecting to improve their living conditions, or investing in an apartment. The descriptive statistical analysis is based on 580,000 transaction records. Then, the regressions models are built to access the factors.

4.3.2.1 Trading Volume via Month

Figures 4.12 and fig.4.13 show the actual trading by month and year. We can observe is that an annual increase in the trading volume of Lianjia. Trading in February is approximately at the same low level due to the spring festival. The market, however, recovers immediately in March. The high demand starts and continues steadily until fall. Seeing that National Day is on the first of October, the seven-day holiday period provides potential buyers with sufficient time to visit the housing properties. At the end of the year, there is another trading peak, which is due to the companies distributing bonuses at the end of year. More people can afford the down payment after receiving the bonus. The trading volume in October 2016 decreased, which may be due to the restrictions on purchases introduced at the time.

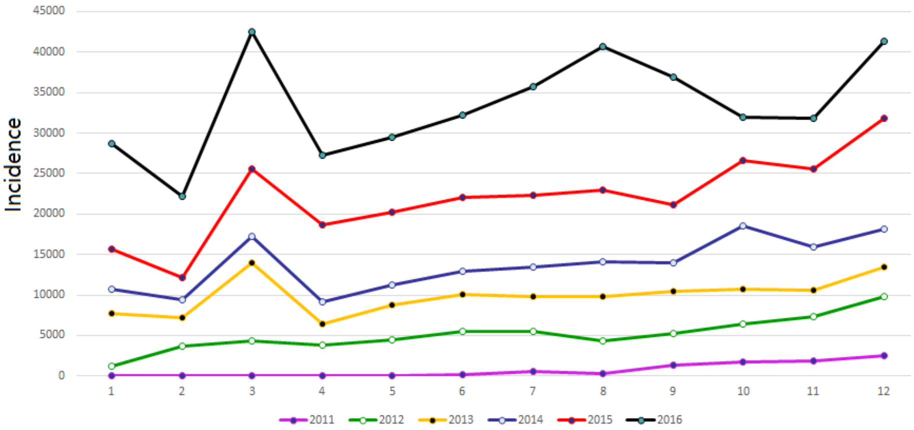


Figure 4.12: The monthly transaction in Beijing from 2011 to 2016

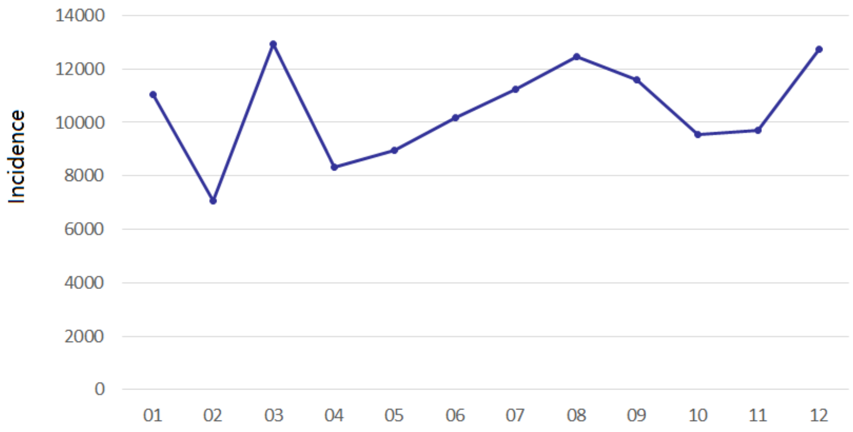


Figure 4.13: The average monthly transaction in Beijing from 2011 to 2016

4.3.2.2 Factors Influencing Average Trading Price

In this section, the factors are examined that may influence the unit price. Figure 4.14 shows that the highest average price per square meter is concentrated in the city center, in the districts of Xicheng, Dongcheng, Haidian, in which many schools with solid reputations and high enrollment rates are located. Fengtai is also located in the center of Beijing, but there are few good schools in the area, and the prices are not high when compared to the three above-mentioned areas. Figure 4.15 indicates the nearest subway line to the apartment. Three peaks can be observed. The average price is relatively higher if the apartments are near lines 2, 7, or the airport line. Line 2, which was opened in 1984, is a rapid transit rail line in central Beijing that runs in a rectangular loop around the city center. It is the only one leading to the Beijing railway station. This line surrounds the Ming dynasty inner city walls, which were demolished and paved over by the 2nd Ring Road and Qianmen Avenue. Line 7 is the rapid transit rail crossing the west to east of central Beijing. The airport line offers more convenience to the downtown lines and it is therefore expensive. In addition, Figure 4.16 indicates the price changes based on the walking distance to subway. When the distance exceeds 400 meters, the price lowers together with the growing distance.

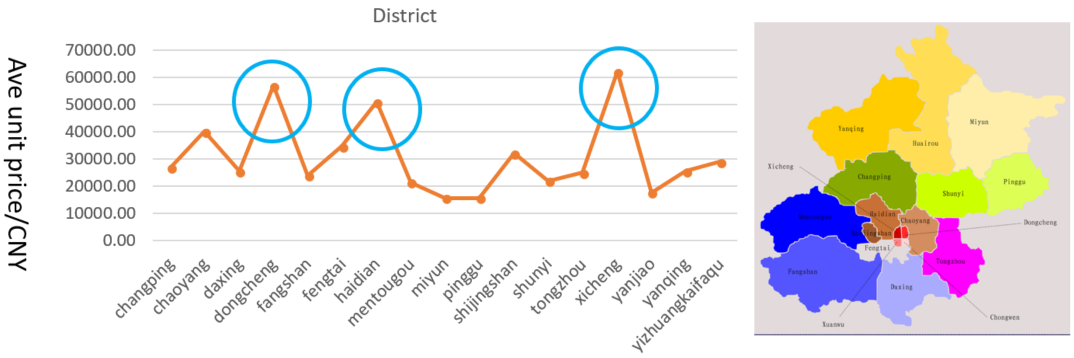


Figure 4.14: Average trading price per square by district

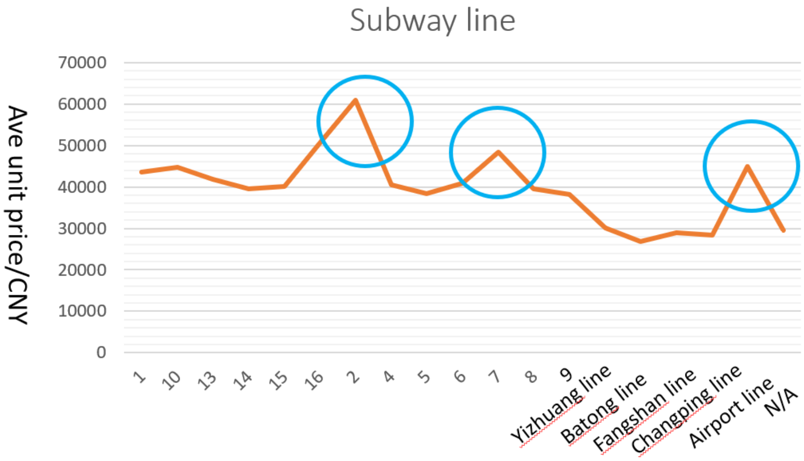


Figure 4.15: Average trading price per square by subway line

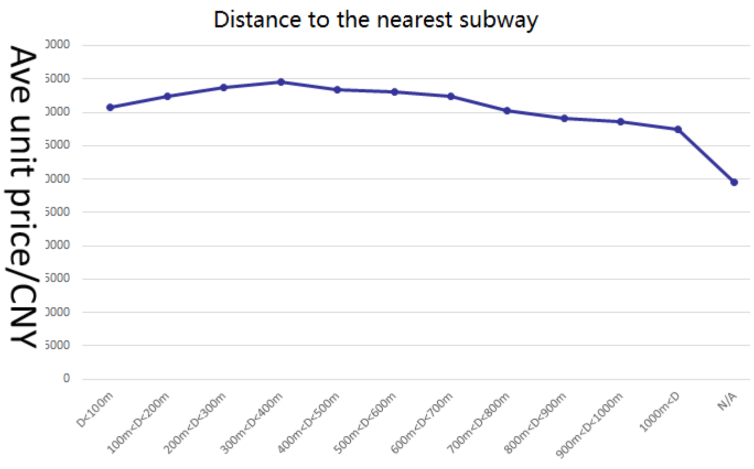


Figure 4.16: Average trading price per square by distance to subway

Figure 4.17 shows that the buildings with elevators are more expensive than the ones without. As is shown in Figure 4.18, the relative height of the building also influences the price. Apartments on the lower floors are more expensive than those on higher floors. The average price of the bedrooms facing south is slightly higher than others (cf. Figure 4.19).

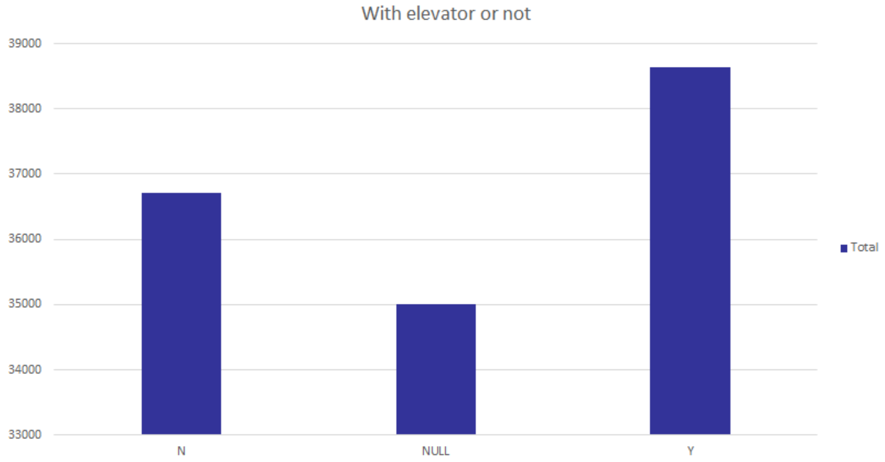


Figure 4.17: Average trading price per square by elevator

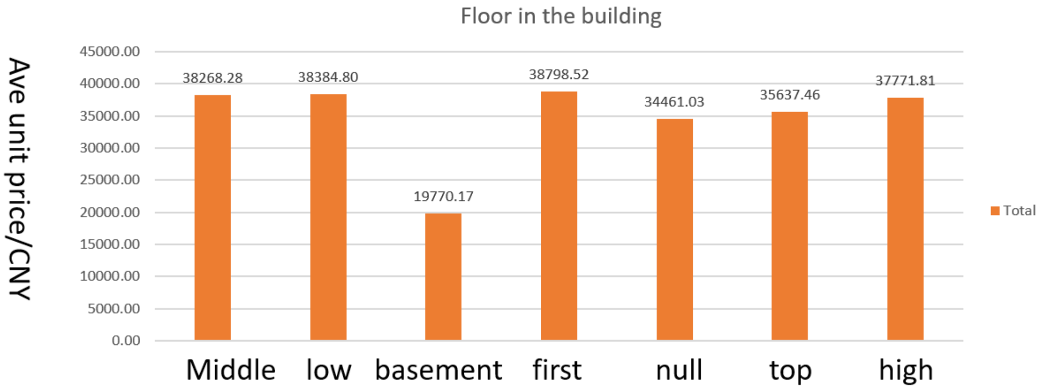


Figure 4.18: Average trading price per square by floor

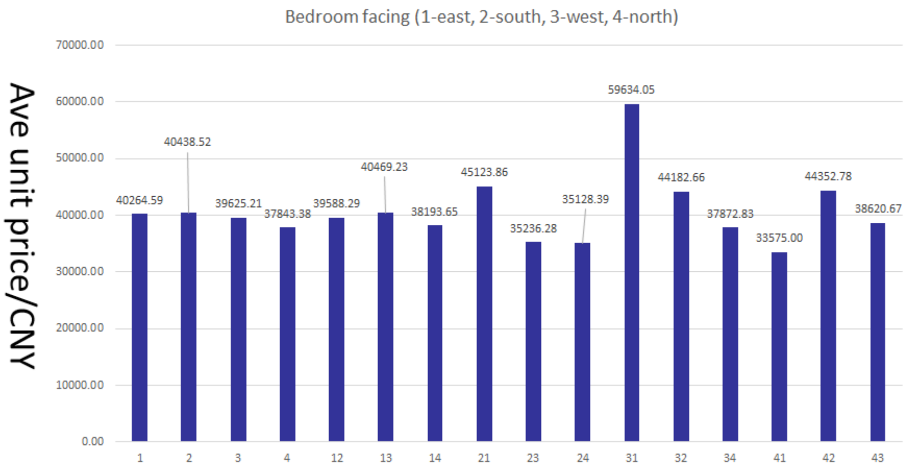


Figure 4.19: Average trading price per square by direction

4.3.3 General Linear Regression Model

The forecast is combined with the data on the individual property characteristics. This section focuses on the aggregate forecasts with property characteristics to construct the forecast for a particular property. The dependent variable is average trading price per square meter. The list of independent variables is as follows. Based on the results of the general linear regression, the R-squared is 71.15%, accounting well for the variations in the data. The p-value of all independent variables is below 0.001, which means that all of them are statistically significant in this model (cf. Table 4.2). The area (of the apartment) and distance to subway are the only continuous variables. Their coefficients are -2.849 and -82.48 respectively, which means that with each square meter the price increases by CNY 2.84 and with each meter further away from the subway the price decreases by CNY 82.48. In addition, house type, direction, decoration, floor location, policy age, the nearest subway line, location and built year are also influencing the price.

- Unit_price: price per square meters
- DistanceToSubway: distance to the nearest subway in meters
- Area: Total area in square meters

- House_type: bedroom and living room number (eg: 2_1, 2 bedroom and living room)
- Direction: bedroom direction (1: east, 2:south, 3: west, 4: north)
- Decorate: building decoration (0: no decoration, 1: simple decoration; 2: refined decoration
- Elevator: the building has an elevator or not (Y: yes; N: no)
- Deal year month: the transaction time (e.g. 2015.02)
- Floor: apartment location in the building (top, high, middle, low, ground floor, basement)
- Policy_age: last transaction took place before special policy year (2, 5, NULL [NULL is not known])
- Subway_num: the nearest subway line number
- District: building location
- Build_era_new: the decade in which the property was built (1950, 1960, 1970, 1980, 1990, 2000, 2010)

Table 4.2: Regression results by apartment details

Term	Coef	T-Value
Constant	45816	30.72***
DistanceToSubway	-2.849	-26.35***
Area	-82.48	-38.18***
house_type		
0_1	-21837	-5.97***
0_2	-4554	-0.63
1_0	-10658	-14.5***
1_1	-5320	-7.4***
1_2	-2343	-3.01***
2_0	-7587	-9.71***
2_1	-4954	-6.99***
2_2	-1077	-1.52
2_3	-2441	-0.98
3_0	-5558	-6.8***
3_1	-3195	-4.52***
3_2	-98	-0.14
3_3	-217	-0.13
3_4	7617	1.49
4_0	-5470	-2.67***
4_1	-1175	-1.31
4_2	3918	4.97***
4_3	3640	2.09**
4_4	14234	2.78***
5_0	325	0.05
5_1	-3767	-1.62
5_2	2461	1.94*
5_3	23083	9.38***
5_4	28120	3.89***
direction		
1	-1473	-2.21**
2	1963	2.99***
3	-1979	-2.96***
4	-1334	-1.98**

Term	Coef	T-Value
12	-110	-0.17
13	-1028	-1.54
14	-2334	-3.48***
21	3206	0.85
23	-86	-0.13
24	2478	3.78***
31	5485	1.77*
32	-597	-0.9
34	-2226	-3.31***
41	-7181	-0.97
42	2792	1.99**
43	-6557	-1.24
112	-2602	-0.99
114	-6253	-1.66*
121	-6185	-1.31
122	7354	2.96***
123	355	0.44
124	1654	2.34**
132	3139	1.89*
134	-1901	-2.44**
142	6076	1.89*
143	684	0.14
212	1435	0.44
214	633	0.41
232	4583	1.97**
234	685	0.95
241	3907	2.94***
243	2295	1.69*
312	-1107	-0.44
314	-1602	-0.59
321	-211	-0.04
322	-7287	-0.69
323	5834	1.81*
324	1482	1.77*

Term	Coef	T-Value
332	2258	0.3
334	-778	-0.18
341	-14242	-1.92*
342	433	0.19
344	4741	0.45
412	-1645	-1.27
414	-15090	-2.86***
421	1041	0.1
432	-2209	-1.83*
434	-5984	-0.98
1122	1550	0.36
1123	2569	0.35
1124	7156	0.96
1212	11477	1.89*
1214	-2304	-0.92
1221	-10730	-2.03**
1224	5572	1.39
1232	2272	1.11
1234	2224	2.34**
1243	-4868	-0.46
1314	-5234	-0.5
1324	-5591	-1.06
1334	-2091	-0.2
1414	-2076	-0.2
1432	-2471	-0.24
1434	906	0.26
2323	3089	0.51
2324	-9265	-1.52
2334	-3723	-0.35
2412	7662	1.45
2414	972	0.09
2432	2254	0.21
2434	2772	0.26
3212	882	0.23

Term	Coef	T-Value
3231	13550	1.29
3234	-771	-0.36
3322	2078	0.39
3324	-3371	-0.45
3342	-12205	-1.16
3344	12852	2.44**
3412	405	0.05
3414	-4095	-1.15
3432	-4256	-0.57
4112	-6036	-1.14
4332	-7894	-1.06
11214	-8966	-0.85
12412	13212	1.26
21232	-2238	-0.21
23434	8366	0.8
32121	4192	0.4
121434	-14407	-1.37
123234	31548	3.01***
123414	247	0.04
decorate		
0	-1338	-9.8***
1	-770	-10.12***
lift		
N	-782.9	-17.36***
deal_year_month		
2011.01	-2308	-0.22
2011.05	432	0.04
2011.06	-1366	-0.13
2011.07	-7990	-0.77
2011.10	-14912	-2.49**
2014.04	-6313	-9.07***
2014.05	-8087	-12.42***
2014.06	-9510	-15.09***
2014.07	-9850	-16.1***

Term	Coef	T-Value
2014.08	-9223	-15.33***
2014.09	-9900	-16.11***
2014.10	-9023	-15.55***
2014.11	-8553	-14.46***
2014.12	-8525	-14.28***
2015.01	-8293	-13.95***
2015.02	-8473	-13.57***
2015.03	-7689	-13.27***
2015.04	-7217	-12.52***
2015.05	-6372	-11.03***
2015.06	-5828	-10.05***
2015.07	-5308	-9.14***
2015.08	-4790	-8.26***
2015.09	-4472	-7.64***
2015.10	-3927	-6.75***
2015.11	-3798	-6.57***
2015.12	-2585	-4.53***
2016.01	-1231	-2.15**
2016.02	323	0.56
2016.03	3478	6.11***
2016.04	5034	8.65***
2016.05	5861	10.11***
2016.06	6598	11.43***
2016.07	8340	14.6***
2016.08	11358	20.02***
2016.09	17027	29.88***
2016.10	19471	32.56***
2016.11	19992	33.9***
2016.12	22850	39.48***
2017.01	24704	42.11***
Floor		
basement	-19209	-48.86***
ground_floor	5136	41.69***
high	3781.6	37.9***

Term	Coef	T-Value
low	4100	40.47***
middle	4458.4	47.76***
policy_age		
-1	-230.6	-3.46***
2	-273.3	-3.84***
Subway_Num		
1	2499	18.18***
2	3772	27.46***
4	4733	36.42***
5	-433	-4.11***
6	-4407	-41.48***
7	-7333	-59.35***
8	-1485	-12.7***
9	1535	8.88***
10	2004.9	22.63***
13	2173	16.31***
14	-2515	-21.21***
15	2213	14.38***
district		
changping	-10820	-15.61***
chaoyang	4713	6.88***
daxing	-17165	-24.35***
dongcheng	21634	31.13***
fangshan	-11886	-1.75*
fengtai	-3782	-5.5***
haidian	16479	24.01***
shijingshan	-8595	-12.15***
shunyi	-17451	-23.27***
tongzhou	-4927	-6.7***
build_era_new		
NULL	652	0.14
1950	-4275	-4.67***
1960	1441	1.76*
1970	394	0.56

Term	Coef	T-Value
1980	2012	2.98***
1990	352	0.52
2000	1103	1.63

R square 71.15%

***p<0.01, **p<0.05, *p<0.1

In the residual analysis, the normal probability plot is to check whether outliers exist in the data. There are still some outliers. The histogram of residuals could determine the residuals are normally distributed. The residuals versus fits plot is to verify the assumption that the residuals have a constant variance and in the plot, the errors have non-constant variance, with most of the residuals scattered randomly around zero. In the future work, The method of weighted least squares is proposed to explore the other possibility. We also use the residuals versus order plot to verify the assumption that the residuals are uncorrelated with each other. In the future study, we may try to remove the outliers and also may try to explore other machine learning algorithms to analyze the data.

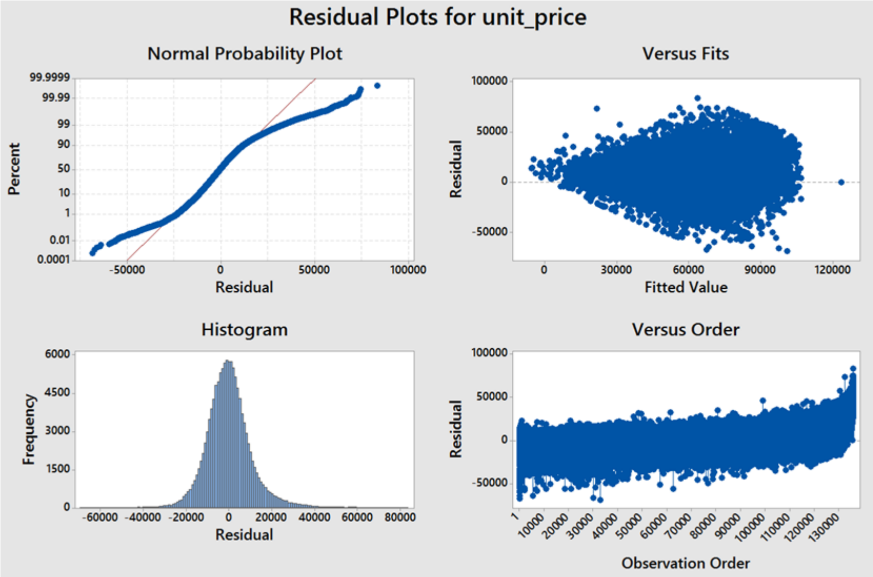


Figure 4.20: Residual analysis of GLM model for prediction of price per square

4.4 Conclusion

In this essay, the relationships among house prices, macroeconomics policy, and physical properties were analyzed. Based on the real time trading data in the second-hand housing market in Beijing and the macro-economic indicators, a valuation model was developed to detect the housing market trend in China, to assist the investors in choosing the appropriate investment strategy, and policy maker to regulate the market. One of the interesting findings is that, compared to other macro-economic indicators, only the exchange rate from USD to CNY is statistically significant in the shift thee-month model. This means that for every additional exchange rate from USD to CNY, a unit-housing price is expected to increase by an average of 0.086. An attempt was made to visualize the physical properties and most of them were found to be statistically significant in the general linear regression model, with the R-squared reaching the high value of 71.15%. Square (the area of the apartment) and distance to subway are the continuous variables. With an increase by every one square meter and with one meter further away from the subway, the price decreases by CNY 2.84 and CNY 82.48 respectively.

Vita

The author of the thesis

2018	Ph.D. in Management, Rutgers Business School, Rutgers University
2016	M.A. in Physics, SUNY-Stony Brook University
2012	M.S. in Nuclear Technology and Application, Peking University
2009	L.L.B. in Law, Beijing University of Aeronautics and Astronautics
2008	B.S. in Electrical Engineering, Beijing University of Aeronautics and Astronautics

2013-2017	Part-time Lecturer/Teaching assistant, Rutgers Business School, Rutgers University
2017-2017	Adjunct Associate, Columbia University
2012-2013	Teaching assistant, SUNY, Stony Brook University
2009-2012	Teaching/Research assistant, Peking University

Bibliography

Abbasi, A., Altmann, J., and Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4):594–607.

Adams, Z. and Füss, R. (2010). Macroeconomic determinants of international housing markets. *Journal of Housing Economics*, 19(1):38–50.

Akinci, F., Esatoglu, A. E., Tengilimoglu, D., and Parsons, A. (2005). Hospital choice factors: a case study in turkey. *Health Marketing Quarterly*, 22(1):3–19.

Alba, R. D. (1973). A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1):113–126.

Andrews, D. (2010). Real house prices in oecd countries: the role of demand shocks and structural and policy factors. *OECD Economic Department Working Papers*, (831):0_1.

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and engineering ethics*, 3(1):63–84.

Audrey J. Weiss, M. L. B. and Steiner, C. A. (2014). Trends and projections in inpatient hospital costs and utilization, 2003-2013. *STATISTICAL BRIEF 175*.

Baidu (2013). National 5 guidelines. *http://baike.baidu.com/item/, Access: May, 09, 2017*.

Bammer, G. (2008). Enhancing research collaborations: Three key management challenges. *Research Policy*, 37(5):875–887.

Barman, S., Tersine, R. J., and Buckley, M. R. (1991). An empirical assessment of the perceived relevance and quality of pom-related journals by academicians. *Journal of Operations Management*, 10(2):194–212.

Beijing-Daily (2011). Beijing's 15 guidelines. *http://zhengwu.beijing.gov.cn/bmfu /bmts/t1155210.htm*.

Benson, E., Hansen, J., Schwartz, A., and Smersh, G. (1999). Canadian/us exchange rates and nonresident investors: Their influence on residential property values. *Journal of Real Estate Research*, 18(3):433–461.

Bornmann, L. and Daniel, H.-D. (2004). Reliability, fairness and predictive validity of committee peer review. *BIF Futura*, 19:7–19.

Boyle, M. and Kiel, K. (2001). A survey of house price hedonic studies of the impact of environmental externalities. *Journal of real estate literature*, 9(2):117–144.

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145.

Bryan, B. (2016). Gdphealthcare. *http://www.businessinsider.com/healthcare-spending-as-percent-of-gdp-recession-2016-12*.

Case, K. E., Quigley, J. M., and Shiller, R. J. (2005). Comparing wealth effects: the stock market versus the housing market. *Advances in macroeconomics*, 5(1).

Case, K. E. and Shiller, R. J. (1990). Forecasting prices and excess returns in the housing market. *Real Estate Economics*, 18(3):253–273.

Case, K. E. and Shiller, R. J. (2003). Is there a bubble in the housing market? *Brookings papers on economic activity*, 2003(2):299–342.

Chow, K. K., Yiu, M. S., Leung, C. K., and Tam, D. (2008). Does the dipasquale-wheaton model explain the house price dynamics in china cities?

Cleary, P. D. and McNeil, B. J. (1988). Patient satisfaction as an indicator of quality care. *Inquiry*, pages 25–36.

Corley, E. A., Boardman, P. C., and Bozeman, B. (2006). Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research policy*, 35(7):975–993.

de Souza, C. G. and Barbastefano, R. G. (2011). Knowledge diffusion and collaboration networks on life cycle assessment. *The International Journal of Life Cycle Assessment*, 16(6):561–568.

Diewert, W. E., HENDRIKS, R., et al. (2011). The decomposition of a house price index into land and structures components: A hedonic regression approach. *The Valuation Journal*, 6(1):58–105.

Ding, X. D. (2015). The impact of service design and process management on clinical quality: An exploration of synergetic effects. *Journal of Operations Management*, 36:103–114.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5(1):187–203.

Eastmoney (2017). Disposable income per capita in beijing.

Eaton, J. P., Ward, J. C., Kumar, A., and Reingen, P. H. (1999). Structural analysis of co-author relationships and author productivity in selected outlets for consumer behavior research. *Journal of Consumer Psychology*, 8(1):39–59.

Fabbri, D., Robone, S., et al. (2008). The geography of hospital admission in a national health service with patient choice: Evidence from italy. *Health Econometrics and Data Group*.

Ferdinand, T. et al. (1887). *Gemeinschaft und Gesellschaft*. .

Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Fry, T. D., Donohue, J. M., Saladin, B. A., and Shang, G. (2013). The origins of research and patterns of authorship in the international journal of production research. *International Journal of Production Research*, 51(23-24):7470–7500.

Garg, V. (2013). Basics of china's real estate industry. *http://www.investopedia.com/articles/investing/091814/basics-chinas-real-estate-industry.asp, Access: Jul, 09, 2017.*

Goodhart, C. and Hofmann, B. (2008). House prices, money, credit, and the macroeconomy. *Oxford Review of Economic Policy*, 24(1):180–205.

Haggbloom, S. J., Warnick, R., Warnick, J. E., Jones, V. K., Yarbrough, G. L., Russell, T. M., Borecky, C. M., McGahhey, R., Powell III, J. L., Beavers, J., et al. (2002). The 100 most eminent psychologists of the 20th century. *Review of General Psychology*, 6(2):139.

Han, Y., Zhou, B., Pei, J., and Jia, Y. (2009). Understanding importance of collaborations in co-authorship networks: A supportiveness analysis approach. In *SDM*, pages 1112–1123. SIAM.

Hauptman, R. (2005). How to be a successful scholar: Publish efficiently. *Journal of Scholarly Publishing*, 36(2):115–119.

Holbrook, M. B. and Hirschman, E. C. (1982). The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of consumer research*, 9(2):132–140.

Hsieh, P.-N. and Chang, P.-L. (2009). An assessment of world-wide research productivity in production and operations management. *International Journal of Production Economics*, 120(2):540–551.

Huber, J., Kamakura, W., and Mela, C. F. (2014). A topical history of jmr. *Journal of Marketing Research*, 51(1):84–91.

Huerta, T. R., Harle, C. A., Ford, E. W., Diana, M. L., and Menachemi, N. (2016). Measuring patient satisfactions relationship to hospital cost efficiency: Can administrators make a difference? *Health care management review*, 41(1):56–63.

IMF (2017). Gdp in china. *http://www.imf.org/en/Data.*

Jianglin, L. (2010). The measurement of the bubble of urban housing market in china [j]. *Economic Research Journal*, 6:28–41.

Katz, J. S. and Martin, B. R. (1997). What is research collaboration? *Research policy*, 26(1):1–18.

Keehan, S. P., Stone, D. A., Poisal, J. A., Cuckler, G. A., Sisko, A. M., Smith, S. D., Madison, A. J., Wolfe, C. J., and Lizonitz, J. M. (2017). National health expenditure projections, 2016–25: price increases, aging push sector to 20 percent of economy. *Health Affairs*, 36(3):553–563.

Krieger, N. (1990). Racial and gender discrimination: risk factors for high blood pressure? *Social science & medicine*, 30(12):1273–1281.

Kudyba, S. P. (2010). *Healthcare informatics: improving efficiency and productivity.* CRC Press.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of political economy*, 74(2):132–157.

Landvoigt, T., Piazzesi, M., and Schneider, M. (2013). Housing assignment with restrictions: Theory and and evidence from the stanford campus. Technical report, Working Paper, Stanford University.

Law, S. (2017). Defining street-based local area and measuring its effect on house price using a hedonic price approach: The case study of metropolitan london. *Cities*, 60:166–179.

Lee, S. and Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social studies of science*, 35(5):673–702.

Li, E. Y., Liao, C. H., and Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9):1515–1530.

Liao, C. H. (2011). How to improve research quality? examining the impacts of collaboration intensity and member diversity in collaboration networks. *Scientometrics*, 86(3):747–761.

Limsombunchai, V. (2004). House price prediction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference*, pages 25–26.

Lynch, A. K. and Rasmussen, D. W. (2001). Measuring the impact of crime on house prices. *Applied Economics*, 33(15):1981–1989.

Madsen, J. B. (2012). A behavioral model of house prices. *Journal of Economic Behavior & Organization*, 82(1):21–38.

Malhotra, M. K. and Kher, H. V. (1996). Institutional research productivity in production and operations management. *Journal of Operations Management*, 14(1):55–77.

Miller, N., Sklarz, M., and Real, N. (1988). Japanese purchases, exchange rates and speculation in residential real estate markets. *Journal of Real Estate Research*, 3(3):39–49.

Moreno, J. L., Jennings, H. H., et al. (1934). Who shall survive?

Moulton, B. R. (1996). Bias in the consumer price index: what is the evidence? *Journal of Economic perspectives*, 10(4):159–177.

NBS (2017). Disposable income per capita in beijing. *https://data.worldbank.org/indicator/.*

OECD (2016). Healthcare spending. *doi: 10.1787/8643de7e-en.*

Olson, J. E. (2005). Top-25-business-school professors rate journals in operations management and related fields. *Interfaces*, 35(4):323–338.

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934.

Peng, W., Tam, D. C., and Yiu, M. S. (2008). Property market and the macroeconomy of mainland china: a cross region study. *Pacific Economic Review*, 13(2):240–258.

Poterba, J. M., Weil, D. N., and Shiller, R. (1991). House price dynamics: the role of tax policy and demography. *Brookings Papers on Economic Activity*, 1991(2):143–203.

Propper, C., Damiani, M., Leckie, G., and Dixon, J. (2007). Impact of patients' socioeconomic status on the distance travelled for hospital admission in the english national health service. *Journal of Health Services Research & Policy*, 12(3):153–159.

Rand, C. S. and Kuldau, J. M. (1990). The epidemiology of obesity and self-defined weight problem in the general population: Gender, race, age, and social class. *International Journal of Eating Disorders*, 9(3):329–343.

RecDAC (2017). Claim inpatient admission type code. *RecDAC*.

Said, Y. H., Wegman, E. J., Sharabati, W. K., and Rigsby, J. T. (2008). Retracted: Social networks of author–coauthor relationships. *Computational Statistics & Data Analysis*, 52(4):2177–2184.

Shang, G., Saladin, B., Fry, T., and Donohue, J. (2015). Twenty-six years of operations management research (1985–2010): authorship patterns and research constituents in eleven top rated journals. *International Journal of Production Research*, (ahead-of-print):1–37.

Shindul-Rothschild, J., Flanagan, J., Stamp, K. D., and Read, C. Y. (2017). Beyond the pain scale: Provider communication and staffing predictive of patients satisfaction with pain control. *Pain Management Nursing*.

Simmel, G. and Leggewie, C. (1908). *Exkurs über den Fremden*.

Söderbaum, F. (2001). Networking and capacity building: the role of regional research networks in africa. *The European Journal of Development Research*, 13(2):144–163.

Stavrakis, A. I., Ituarte, P. H., Ko, C. Y., and Yeh, M. W. (2007). Surgeon volume as a predictor of outcomes in inpatient and outpatient endocrine surgery. *Surgery*, 142(6):887–899.

Swoboda, A., Nega, T., and Timm, M. (2015). Hedonic analysis over time and space: the case of house prices and traffic noise. *Journal of Regional Science*, 55(4):644–670.

Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM.

Tsatsaronis, K. and Zhu, H. (2004). What drives housing price dynamics: cross-country evidence.

Van Rijnsoever, F. J. and Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy*, 40(3):463–472.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

WorldBank (2017). Worldside gdp. *https://data.worldbank.org/indicator/*.

Wu, X. C., Chen, V. W., Steele, B., Ruiz, B., Fulton, J., Liu, L., Carozza, S. E., and Greenlee, R. (2001). Subsite-specific incidence rate and stage of disease in colorectal cancer by race, gender, and age group in the united states, 1992–1997. *Cancer*, 92(10):2547–2554.

Young, S. T., Baird, B. C., and Pullman, M. E. (1996). Pom research productivity in us business schools. *Journal of Operations Management*, 14(1):41–53.

Yue, S. and Hongyu, L. (2004). Housing prices and economic fundamentals: A cross city analysis of china for 1995-2002. *Economic Research Journal*, 6:78–86.

Zhang, F. (2014). *Modelling the Housing Market and Housing Satisfaction in Urban China*. PhD thesis, University of Bath.