Research on Spatio-temporal Behavior Analysis of House Buyers Based on Real Estate Data Mining-A case of a part of cities in Jiangsu Province, China

Master thesis dissertation
To obtain the Graduate Master's degree
at the Faculty of Natural Sciences
the Paris-Lodron-Universität Salzburg

Submitted by
Yu Yang

Supervisor:
Univ.-Dr. Mittlboeck

Subject area:
Interfaculty Department of Geoinformatics – Z_GIS

Salzburg, June, 2018
Declaration

I declare that I have authored this thesis independently, that I have not used any other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

This thesis is a translation of the Chinese language original version as submitted and accepted at Nanjing Normal University under the Master of Science in Geoinformatics dual degree agreement with the University of Salzburg.
Abstract

The ultimate goal of the implementation of the real estate unified registration system is to realize the real estate data information sharing and disclosure. Real estate registration data is a very important part of real estate data, containing important information such as temporal and spatial information. Its unique resource asset attribute has great research and application value, and has important reference value for promoting government and department decision-making, strengthening urban management and maintaining social stability. At present, the research of real estate registration data in China mainly focuses on the integration and construction of real estate registration data, and the research on the application of real estate registration data is still in the initial stage. In the meantime, researches on spatio-temporal behavior of human groups based on industry spatio-temporal data have gradually become a research hotspot in recent years. In view of this phenomenon, this article takes house buyers in the real estate registration data as the research object, and conducts research on the law of spatio-temporal behavior of house buyers, which not only expands the real estate data application field, but also enriches the research content of space-time behavior analysis.

Under the guidance of the theory of spatio-temporal behavior, data mining and spatio-temporal data mining theory, and spatial autocorrelation theory, this paper studies the data processing and mining method of real estate registration, and deeply mines the law of spatio-temporal behavior of house buyers in real estate registration data. The research content of this paper mainly includes the following aspects:

(1) Clarifying the real estate registration data processing method. On the basis of studying the geocoding technology and data cleaning methods, and according to the demand of real estate registration data mining, transforming text containing location information in the real estate registration data into space coordinates; and the real estate registration data is densified and data cleaned against the data quality problems in the real estate registration data. And finally obtain the qualified real estate registration data that meets data mining requirements.

(2) Reaeach on the method of data mining of real estate registration. Combining spatio-temporal behavior analysis method, data mining method, spatiotemporal data
Abstract

mining method, GIS method and visualization method, forming a real estate registration data mining method is formed. Through the data mining method of real estate registration, analyzing the behavior preferences of house buyers, the spatio-temporal distribution of houses, and the spatio-temporal behavior of the house buyers in the field.

(3) Verifying the idea of real estate registration data processing, cleaning and mining by an example. Selecting real estate registration data of Dongtai City in Jiangsu Province, Jintan District of Changzhou City and Wujiang District of Suzhou City from 2009 to 2014 to conduct experiments and finding out the rules of space-time behavior of different types of house buyers.

Key words: Real estate registration, Spatio-temporal behavior analysis, Spatio-temporal data mining, Spatial autocorrelation
Catalog

Abstract .......................................................................................................................... I
Catalog ........................................................................................................................... I
List of Figures .................................................................................................................. 4
List of Tables .................................................................................................................... 6
Chapter 1  Introduction...................................................................................................... 1
  1.1  Background and significance ................................................................................. 1
  1.1.1  Background ........................................................................................................ 1
  1.1.2  Significance ........................................................................................................ 2
  1.2  Research status ....................................................................................................... 3
  1.2.1  Research Status of Real Estate Data ................................................................. 3
  1.2.2  Research Status of Spatio-temporal Behavior Analysis .................................... 5
  1.2.3  Research Status of Data Mining ....................................................................... 6
  1.2.4  Summary of Research Status ......................................................................... 8
  1.3  Research objects and contents ............................................................................. 9
  1.3.1  Research objectives ......................................................................................... 9
  1.3.2  Research contents ......................................................................................... 9
  1.4  Research technical route ..................................................................................... 10
Chapter 2  Real Estate Registration Data Sources and Data Processing Methods ... 12
  2.1  Introduction to the Real Estate Unified Registration System ............................. 12
  2.2  Acquisitions of Real Estate Registration Data .................................................... 13
  2.3  Real Estate Registration Data Storage and Problems ......................................... 14
  2.3.1  Real Estate Registration Data Storage ............................................................ 14
  2.3.2  Problems with existing real estate registration data ..................................... 15
  2.4  Research on Real Estate Registration Data Processing Methods ..................... 16
  2.4.1  Data Geocoding Methods ............................................................................. 16
  2.4.2  Online Geocoding of Real Estate Registration Data ..................................... 17
  2.4.3  Data Cleaning Methods ............................................................................... 19
  2.4.4  Real Estate Registration Data Cleaning ........................................................ 21
  2.5  Classification of real estate registration data ....................................................... 23
  2.6  Summary of this chapter ..................................................................................... 24
Chapter 3  Research on Real Estate Registration Data Mining Methods ................. 25
# Catalog

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Association Rules Mining</td>
<td>25</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Analysis of Association Rules</td>
<td>25</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Apriori Algorithm Overview</td>
<td>29</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Association Rule Mining Table Establishment</td>
<td>32</td>
</tr>
<tr>
<td>3.1.4</td>
<td>Apriori Algorithm Implementation</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>Spatial Hotspot mining</td>
<td>36</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Spatial Clustering Overview</td>
<td>36</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Kernel Density Estimation Method</td>
<td>38</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Spatial hotspot mining of real estate registration data</td>
<td>39</td>
</tr>
<tr>
<td>3.3</td>
<td>Spatial Autocorrelation Analysis</td>
<td>41</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Spatial Autocorrelation Overview</td>
<td>41</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Global Spatial Autocorrelation Algorithm</td>
<td>42</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Realization of spatial autocorrelation analysis of real estate registration data</td>
<td>44</td>
</tr>
<tr>
<td>3.4</td>
<td>Dynamic Visualization of Spatio-temporal Data</td>
<td>45</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Dynamic Visualization of Data</td>
<td>45</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Introduction to Visualization Tools</td>
<td>45</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Dynamic spatio-temporal visualization of real estate registration data</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>Summary of this chapter</td>
<td>48</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>A Case Analysis of Saptio-temporal Behavior of House Buyers</td>
<td>49</td>
</tr>
<tr>
<td>4.1</td>
<td>Overview of Research Area and Data Processing</td>
<td>49</td>
</tr>
<tr>
<td>4.2</td>
<td>Analysis of Basic Features of House Buyers</td>
<td>50</td>
</tr>
<tr>
<td>4.2.1</td>
<td>All house buyers feature analysis</td>
<td>50</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Analysis of Basic Characteristics of Different Types of House Buyers</td>
<td>55</td>
</tr>
<tr>
<td>4.3</td>
<td>Mining Association Rules of Real Estate Registration Data</td>
<td>59</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Association Rules Mining Results</td>
<td>59</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Result analysis</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>Spatial Hotspot Mining and Analysis of Real Estate Registration Data</td>
<td>67</td>
</tr>
<tr>
<td>4.4.1</td>
<td>The results of mining hotspots by all house buyers</td>
<td>67</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Analysis of House Purchase Behavior of Different Types of House Buyers</td>
<td>71</td>
</tr>
<tr>
<td>4.5</td>
<td>Source Analysis of Foreign Buyers</td>
<td>74</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>4.5.1 Time-space dynamic visualization of house buyers' census register location</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>4.5.2 Spatial Autocorrelation Analysis of Foreign Buyers' Census Register Location</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>4.6 Chapter Summary</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Chapter 5 Conclusions and Prospects</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td>84</td>
<td></td>
</tr>
</tbody>
</table>
List of Figures

Figure 1-1 Technical Roadmap ................................................................. 11
Figure 2-1 Real estate registration rights type .............................................. 13
Figure 2-2 Table-to-table association diagram ............................................ 14
Figure 3-1 Lattice structure ........................................................................ 26
Figure 3-2 Theorem 1 .................................................................................. 27
Figure 3-3 The anti-monotone of the measure of support ............................ 27
Figure 3-4 Kernel density maps with different search radius ....................... 40
Figure 3-5 Trend significance classification level map ..................................... 44
Figure 4-1 Dongtai city, Jintan District, Wujiang District House Buyers’ Purchase Age and Gender Distribution .............................................................. 51
Figure 4-2 Dongtai , Jintan and Wujiang House Buyers Purchase Area Distribution . 52
Figure 4-3 Dongtai, Jintan, and Wujiang House Buyers Purchase Area and Purchase Age Distribution ................................................................. 53
Figure 4-4 Proportion of different types of house buyers in different regions ....... 55
Figure 4-5 Different age distribution of the different types of house buyers in Dongtai city .................................................................................. 56
Figure 4-6 Different age distribution of the different types of house buyers in Jintan district ................................................................................. 57
Figure 4-7 Different age distribution of the different types of house buyers in Wujiang district .............................................................................. 58
Figure 4-8 The association rules between house layer and the total house layers in Jintan District ........................................................................ 60
Figure 4-9 The association rules between house layer and the total house layers in Wujiang District ................................................................. 61
Figure 4-10 The Association Rules between Census Register Location and House Area in Dongtai City ........................................................................... 62
Figure 4-11 The Association Rules between Census Register Location and House Area in Wujiang District ............................................................. 62
Figure 4-12 The Association Rules between the Purchase Age of Buyers and the house floor in the Jintan District ....................................................... 63
List of Figures

Figure 4-13 The Association Rules between the Purchase Age of Buyers and the house floor in the Wujiang District ................................................................. 64
Figure 4-14 The Association Rules between the Purchase Age of Buyers and the total floors of houses in the Jintan District ....................................................... 65
Figure 4-15 The Association Rules between the Purchase Age of Buyers and the total floors of houses in the Wujiang District ................................................. 65
Figure 4-16 Dongtai City 2009-2014 house buyers purchase heat distribution map .. 67
Figure 4-17 Jintan District 2009-2014 house buyers purchase heat distribution map 68
Figure 4-18 Wujiang District 2009-2014 house buyers purchase heat distribution map ........................................................................................................ 69
Figure 4-19 Different Types of House Buyers Purchase Heat Distribution Map of Dongtai City ................................................................................................. 71
Figure 4-20 Different Types of House Buyers Purchase Heat Distribution Map of Jintan District ........................................................................................................ 72
Figure 4-21 Different Types of House Buyers Purchase Heat Distribution Map of Wujiang District ................................................................................................. 73
Figure 4-22 Spatio-temporal Dynamic Visualization Map of Foreign House Buyers in Dongtai City ................................................................................................. 75
Figure 4-23 Spatio-temporal Dynamic Visualization Map of Foreign House Buyers in Jintan District ................................................................................................. 76
Figure 4-24 Spatio-temporal Dynamic Visualization Map of Foreign House Buyers in Wujiang District ................................................................................................. 77
Figure 4-25 Moran's I Index Change of Outer province house buyers in each region 80
List of Tables

Table 2-1 Field Information of House Buyers Data Sheet.................................................. 15
Table 2-2 China's mainstream online geocoding service quota ..................................... 18
Table 2-3 Baidu online geocoding core code ................................................................. 18
Table 2-4 Classification of house buyers ........................................................................... 23
Table 3-1 Apriori algorithm generates frequent itemsets ................................................. 29
Table 3-2 Apriori_gen function ......................................................................................... 30
Table 3-3 Rule generation in Apriori algorithm ............................................................... 31
Table 3-4 Ap - genrules function ...................................................................................... 31
Table 3-5 Data set content ................................................................................................. 32
Table 3-6 Association rule mining table between the house floor and the total floors of house .................................................................................................................. 33
Table 3-7 Association rule mining table between the house buyer’s census register location and the house area .................................................................................... 33
Table 3-8 Association rule mining table between the purchase house age and the total floors of house ........................................................................................................... 34
Table 3-9 Association rule mining table between the purchase house age and the house floor ....................................................................................................................... 34
Table 3-10 Association Rules Mining Code ..................................................................... 35
Table 3-11 Real estate registration data dynamic visualization code ............................ 46
Table 4-1 Annual Per Capita GDP of Dongtai City, Jintan District and Wujiang District (Unit: RMB) ................................................................................................................ 49
Table 4-2 Average and median purchase house age by region ....................................... 50
Table 4-3 Average purchase age of different types of house buyers in different regions ............................................................................................................................. 55
Table 4-4 Moran's I index over the years of different types of house buyers in each region .................................................................................................................................. 77
Chapter 1  Introduction

1.1  Background and significance

1.1.1  Background

Immovable Property, in a general sense, refers to property that is immovable according to its natural nature or legal provisions, such as houses and land. Article 186 of the opinions of the supreme people's court of China on the implementation of the general principles of the civil law of the People's Republic of China provides a judicial interpretation of the immovable: land, buildings attached to the land and other fixed ancillary equipment of buildings and buildings as immovable property. In March 2007, China promulgated the Property Rights Law and put forward the unified registration system for real estate. On December 22, 2014, the State Council issued State Council Decree No. 656, “Provisional Regulations on the Registration of Real Estate” (hereinafter referred to as the “Regulations”), which was formally promulgated and will be implemented on March 1, 2015. The official promulgation of the "Regulations" indicates that the top-level design work of the unified registration of real estate in China has basically been completed, and it has entered the formal implementation phase of the unified registration of real estate. On January 1, 2016, the “Detailed Rules for the Implementation of the Provisional Regulations on the Registration of Real Estate” was issued, marking China’s comprehensive regulation of the promotion of the unified registration system for real estate and providing legal protection for the full implementation of the unified registration system for real estate. The implementation of unified registration of real estate can achieve information sharing and disclosure, which will help promote the networking of individual housing information throughout the country, ensure the security of real estate transactions and protect the legitimate rights and interests of the people.

With the implementation of the unified registration of real estate, real estate data has attracted more and more attention. In recent years, the focus of academic circles is the integration of real estate data and the construction of real estate data management platforms. However, with the implementation of the consolidation of depository real estate data and the implementation of the unified registration of real estate, a large
amount of real estate data has been generated. These real estate data contain a large number of information that has great research value. How to use this information becomes another research focus after real estate data integration. Under the trend of humanism, human behavior research gradually pays attention to the practical problems between humans and society. The field of research gradually moves from macroscopic to microscopic perspectives. In research methods, we also began to pay attention to the space-time of human behavior and the systemic nature of human-social interaction. In the study of human behavior, there is an increasing emphasis on the correlation between human behavior and various geographical entities such as land and objects. With the development of economy, people's living standards have gradually increased, and the number of people buying houses has become more and more. The characteristics of house buyers and the behavior of buying house have become a topic which is worthy to study.

In summary, with the implementation of the real estate registration system, real estate registration database construction is gradually complete. The database contains a large number of real estate data which will become a treasure house, so it has a great significance to carry out the mining and application of real estate data. This article defines the research object as house buyers which is a unique group in real estate data, which can not only analyze the law of house buyer's purchase behavior, but also study the movement law of foreign house buyers, and from a new angle to study the law of current population movement.

1.1.2 Significance

(1) Theoretical significance
This article enriches the research content of real estate data and proposes ideas for real estate registration data processing and data mining. With the implementation of the real estate unified registration system, a large amount of real estate data has been generated. Through the research on the real estate registration data structure and the research on data cleaning methods, proposing the idea of real estate registration data processing. Through the study of the characteristics of real estate registration data, proposing the real estate registration data mining method which will provide reference for future real estate data mining and analysis.

This article enriches the research content of spatio-temporal behavior analysis and explores new ideas for house buyers' spatio-temporal behavior. In today's vigorous
development of the national economy, what changes will occur in the behavior of house
buyers, and what new features will emerge. This paper presents the analysis of spatio-
temporal behavior of house buyers for the first time, which enriches the research
content of spatio-temporal behavior analysis.

(2) Practical significance
For the government, it understands the characteristics for different types of house
buyers and provides reference for the government and departments to make relevant
real estate market decisions. For real estate developers, understanding the preferences
of house buyers at different levels and the demand for house buyers can better design
the structure, area, and house type of the built houses, and can better target the different
groups of people for appropriate real estate promotion and attract more house buyers.
For homebuyers themselves, they can know where is the popular areas to buy a house
and future trends of house price, and can better choose the most suitable housing
according to their own needs.

1.2 Research status

1.2.1 Research Status of Real Estate Data
In the 1990s, research on real estate data mainly focused on data collection and
processing. Brown [1] proposed a data acquisition method in a multimedia real estate
database. Knight et al. proposed a Gibbs sampler to fit these missing data for data
missing in real estate data. At the beginning of this century, real estate data research
focused on forecasting real estate value and there are many data mining methods used
in real estate valuation. For example, Dubin et al. [3] proposed a method based on spatial
autoregressive model to predict real estate value and improved the accuracy of
forecasting. Jaen [4] developed a multi-factor real estate price forecast model based on
neural network technology in data mining technology. Yang et al. [5] proposed a
monitoring index system based on the digestion cycle for real estate data using data
mining technology. After 2010, the research direction of real estate data became more
and more diversified, and new technologies were introduced into real estate data
research. For example, Dubé et al. [6] proposed using space-time weight matrix to
evaluate the spatial autocorrelation in real estate data of Quebec City. Fu et al. [7]
proposed a geographical method named Clus Ranking to rank the real estate investment
value from factors such as the location of the real estate and the degree of prosperity of
the business district, so that the buyers can know the value of the real estate. In recent years, research on real estate data has begun to incorporate a large number of emerging technologies, and has expanded the application of real estate data. For example, Forte et al.\cite{8} applied real estate data to immigration studies, which collected housing prices for 112 Italian cities in 2016 and analyzed the correlation between housing price gradients and Italian immigration. Li et al.\cite{9} proposed an interactive visualization analysis system called “HomeSeeker” based on visualization technology. This system collects real estate data from different approaches and visualizes the data which can provide powerful help for house buyers to analyze real estate.

Due to historical reasons, China's land and housing were not registered at the same time and in different department. Before the introduction of the unified registration system for real estate, China's real estate data research mainly focuses on housing part. In the 1980s, the domestic real estate market was just open. Due to the problem of data volume and related policies at the time, research on real estate data was minimal. Until 1998, China completely stopped the housing distribution system and the China’s real estate market began to a new stage of the market. After 2000, the real estate market was vigorously developed and a large amount of real estate data was accumulated. Research on real estate data was mainly focused on forecasting real estate prices and real estate customer relationship management. For example, Wang Jing et al.\cite{10} proposed the real estate price index prediction based on the wavelet neural network; Hu Xiaolong et al.\cite{11} proposed a real estate price forecasting method based on Elman neural network; Yang Hongtao et al.\cite{12} expounded the application of customer relationship management in real estate enterprises; Lü Shengjun\cite{13} used data mining technology to mine association rules in real estate data. And apply it in real estate customer relationship management; Huang Xiaobin et al.\cite{14} also proposed data mining technology as an auxiliary decision tool in the real estate industry. Since the real estate unified registration system was proposed in 2007, the integration of real estate registration data has become a hot research field. For example, Wang Lvhua et al.\cite{15} proposed a real estate data integration method based on cadastral data and unified coding of parcels as an index; Xu Caijiang et al.\cite{16} proposed a technical route for the registration of real estate registration data based on the results of land registration; Zhang Zhengming et al.\cite{17} analyzed the current situation of real estate data management in various places and studied the key technologies for the integration of cadastral and real estate data in real estate registration.

With the advent of the era of big data, research on real estate registration data is also
Chapter 1  Introduction

facing new opportunities. Liu Ye et al. [18] towards the problem of that China have not systematically and comprehensively analyzed the big data of the real estate industry at present, collating the research status of big data in the real estate industry, and proposed future research trends; Zhang Yuwen et al. [19] analyzed China's real estate registration big data, and put forward four major areas of application of real estate registration big data analysis and future directions, indicating that the implementation of real estate registration big data research and practical application work is the vitality of future real estate registration in China.

1.2.2 Research Status of Spatio-temporal Behavior Analysis

The study of spatio-temporal behavior is centered on time geography, highlighting the exploration and understanding of "human" in the study area. It conducts research on human behavior in the context of space-time and conducts in-depth discussions of human behavior in time and space, providing a new perspective for understanding the complex space-time relationship between humans and cities. The analysis of the spatio-temporal behaviors of human beings not only can find the laws of existing events, but also can combine the historical events and existing data to predict the possibility and trend of similar events in the future.

In the late 1960s, the Swedish geographer Haggstrom and his led Lund school proposed the methodology of time geography. The methodology mainly describes the relationship between human behavior and the environment in which it is located in the spatio-temporal process. In the late 1970s, Carlestein et al. [20] published a large number of articles on the promotion and dissemination of time geography. In the mid-1980s, the Japanese geographer Aung Hiroshi translated a book on the main achievements of time geography [21], successfully introduced the concept of time geography to Japan. The study of spatio-temporal behavior also slowly began to sprout from this time.

The focus of research on spatio-temporal behavior has always been the consideration of human inner-behavior activities and the study of the relevance of human life conditions and social status [22],[23]. The analysis of spatio-temporal behavior in the initial stage is based on abstracted and simplified structured data. The data source mainly is the traditional activity log data and activity survey data. This type of data has the drawbacks of incomplete record of time, inaccurate spatial positioning, and high survey cost. The limitations of data sources have caused limitations in the study of early spatio-temporal behaviors, but with the rapid development of geographic information
disciplines, the widespread use of global positioning systems, the rapid development of
the Internet, and the maturity of LAT and LBS technologies, data sources for research
on spatio-temporal behavior and the technical environment has been greatly improved.
Data sources ranged from the traditional activity log data to GPS survey data to mobile
call data and bus IC card data in the era of big data, and spatio-temporal behavior
analysis technology also introduced the measurement of accessibility based on spatio-
temporal prisms \cite{24,25} and spatio-temporal visualization and other technologies.
The study of spatio-temporal behavior has been widely used in transportation
planning \cite{26-28}, research on vulnerable groups \cite{29-32}, and urban spatial structure \cite{33-35} and
many other aspects because of its unique perspective on the spatio-temporal
relationship between human behavior patterns and urban space environment.

1.2.3 Research Status of Data Mining

Data mining refers to a technique for discovering user-interested, unknown information
and knowledge from massive data, and is a process of implementing “data →
information → knowledge” \cite{36}. Data mining originally refers to an important step in
knowledge discovery in database (KDD). However, many fields do not distinguish
between the two meanings and this paper does not distinguish between the two.
"Data Mining" first appeared in the first KDD International Academic Conference held
in Montreal, Canada in 1995. Because this term is very vivid in the description of
mining knowledge from databases, it was quickly accepted by the academic community.
Early data mining is mainly used for mining character-numeric data and information
knowledge in structured data with simple structure; by the late 1990s, the data mining
had some mature algorithms, such as mining, classification, prediction and clustering
analysis.
In 1996, American scholar Fayyad et al. \cite{37} summarized five basic steps of KDD: data
selection (choose the corresponding data) → data preprocessing (cleaning the data,
eliminating errors and missing information in the data). → Data Transformation
(convert the processed data into the format required for data mining) → Data mining
(Using data mining methods to mine data in the data) → Interpretation Evaluation
(excavation results from interpretation and evaluation). The first three steps can be
combined into data preparation, simplifying KDD into three steps: data preparation
phase → data mining phase → interpretation evaluation phase.
The main task of data mining is to discover the patterns in the data; it is mainly divided
into two categories: description mode and predictive mode\textsuperscript{[38]}. The descriptive model is the normative description of the facts that exist objectively in the data, including data summary and induction, association pattern analysis, clustering pattern analysis, classification pattern analysis, and abnormal pattern analysis. Predictive mode is mainly based on the current data, taking the time series as the axis to predict future values or trends.

In terms of theoretical research and related technologies, there are data mining algorithms for the ten classical algorithms: C4.5 algorithm, K-Means algorithm, support vector machine, Apriori algorithm, maximum expectation algorithm, PageRank algorithm, AdaBoost algorithm, KNN algorithm, and naive bayes model, classification and regression tree. The latest directions include: application of regression methods in KDD, improvement of classic algorithms such as Apriori algorithm and Bayesian method. However, China started late in theoretical research and is still in the development stage. The latest directions mainly include: optimization of traditional association rule algorithms, clustering algorithms and classification algorithms, natural language processing, and Chinese text mining.

In applied research, data mining has been widely used in medicine, transportation, urban planning, software engineering, market analysis, banking, telecommunications and other fields. There are also many well-established and stable data mining software, such as Clementine from SPSS, DBMiner from Simon Fraser University in Canada, Intelligent Miner from IBM, and Minset from Stanford University. China’s data mining has also been widely applied to medical, transportation, and other fields. China's data mining software includes Knight developed by Nanjing University, Aminer developed by Fudan University, MS Miner developed by the Chinese Academy of Sciences, Scope Miner and Open Miner developed by Northeastern University and so on.

The task of spatial data mining is to mine useful information and knowledge for users in the spatial database. The types of knowledge that can be mined from spatial data have similarities with traditional data mining but also have its unique features. There are knowledge including geometric knowledge, spatial clustering rules, spatial association rules, spatial distribution rules, spatial sequence rules, spatial evolution rules and object-oriented knowledge.

In applied research, spatial data mining has been applied to many fields such as scientific research, planning, agriculture, forestry, and environment. Yang et al.\textsuperscript{[39]} discovered the distribution structure of river channels based on spatial statistical
techniques in spatial data; Shad et al.\textsuperscript{[40]} used spatial data to study the evaluation methods for predicting pollutant concentrations in GIS environments; Shu et al.\textsuperscript{[41]} used association rules found in geographic data to detect environmental changes; Wanluhe et al.\textsuperscript{[42]} successfully applied spatial data mining techniques to forest fire prevention intelligent decision support systems; Chen Xiaoyu\textsuperscript{[43]} proposed the method of classification in remote sensing image database based on C4.5 algorithm; Zhang Shaoji et al.\textsuperscript{[44]} used the clustering method in spatial data mining to study the urban house price distribution pattern; Huang Hai\textsuperscript{[45]} proposed a method of using fuzzy neural network method to evaluate rural land consolidation potential. In spatial data mining software, there are GeoMiner from Simon Fraser University in Canada, S_{PLUS} interface from ArcView GIS from ESRI, USA, Descartes and Kepler from European Association of Information and Mathematical Research, GISDBMiner and RSImageMiner from Wuhan University, GISMiner developed by China University of Science and Technology.

1.2.4 Summary of Research Status

Comprehensive domestic and international research status, we can find:

(1) In terms of real estate data research, scholars have begun to incorporate a large number of new methods in recent years and have expanded the application level of real estate data; in China, with the full implementation of the unified registration system for real estate and related technologies in the era of big data, research on real estate data is no longer limited to traditional house price forecasting and real estate data integration. The research on real estate data should be combined with multiple technologies, applied in many aspects, and reflect the real value of real estate data.

(2) For the study of human behavior, scholars mainly focus on the combination of individual daily life practice with macroscopic social reproduction, locality and localization, and use new data sources in the era of big data to study human spatio-temporal behavior. In terms of the research object of spatio-temporal behavior, mainly focus on urban commuters, tourists, consumers, but there is no relevant research on house buyers as the research object; in terms of data sources of spatio-temporal behavior research, mainly are the survey log data, mobile phone calls data, taxi data, GPS data, and internet data, but there is no relevant research on real estate data as a data source.

(3) Scholars have done a lot of research on data mining, and related data mining
methods and spatial data mining methods have gradually matured. Recent studies have focused on how to mine spatio-temporal data and related algorithm improvements in the context of big data and the internet, but there are still many deficiencies and there are many mining methods can be improved.

In summary, the mining and application research of real estate data is an inevitable trend, and the study of spatio-temporal behavior has drawn more and more attention. Data mining and spatial data mining technology have also matured, and spatio-temporal data mining has become a research hotspot. The use of real estate registration data as data source to mine the spatio-temporal behavior of house buyers has an important research significance.

1.3 Research objects and contents

1.3.1 Research objectives

With the implementation of the unified registration system for real estate in China, real estate data has received more and more attention from all walks of life. The purpose of this paper is to propose an idea of real estate registration data processing to extract real estate registration data that meets data mining standards, and to study real estate registration data mining methods to explore the wealth of information contained in real estate registration data, and using house buyers as the research object to analyze the spatio-temporal behavior of house buyers.

1.3.2 Research contents

(1) Research on real estate registration data processing methods

Because existing real estate registration data has a series of problems such as single registration information, disorderly registration information, and lack of registration information data confusion, it is difficult to use existing real estate registration data for data analysis and mining. Therefore, based on the existing theory and methods of data cleaning and geocoding, this paper proposes an idea of real estate registration data processing. Through data cleaning, it extracts real estate registration data with analytical value, high quality, and rich content. At the same time, through geocoding translates the address into spatial coordinate information in order to provide a good data foundation for real estate registration data mining and analysis.

(2) Research on Data Mining Methods of Real Estate Registration Data
Chapter 1  Introduction

Real estate registration data as a kind of spatio-temporal data, which contains a lot of valuable space-time information. In this paper, based on the existing data mining and spatial data mining methods and related GIS methods, seek to conform to the real estate registration data mining method, to excavate implicit and meaningful information in the real estate registration data.

(3) Analysis of the Law of Spatio-temporal Behavior of House Buyers

In recent years, the analysis of spatio-temporal behavior has received more and more attention from all walks of life. The data used to analyze spatio-temporal behavior has also become more and more diversified. The real estate registration data contains a large amount of spatio-temporal behavior information of specific populations. This paper analyzes the current status of spatio-temporal behavior research, using real estate registration data to analyze house buyers’ spatio-temporal behavior. At the same time, the research object is divided into all house buyers, the local house buyers, outer province buyers and inner province buyers. It can study different house buyers’ spatio-temporal behavior, analyzing the purchasing characteristics of different house buyers.

1.4 Research technical route

The technical route of this study is mainly composed of four parts: research background and foundation, data processing, research methods, and expected research results. The research background includes the macro background and research accumulation. The research foundation refers to the author’s previous research accumulation, phenomena observation, related literature and theoretical foundation, etc. Under the support of the research background and foundation, the research content of this article is finally determined. The research includes three aspects: real estate registration data processing, real estate registration data mining, and house buyers’ spatio-temporal behavior analysis. The main research methods of this paper include content analysis method, experimental research method and induction summarization method, etc. The expected research results mainly include the following three aspects: First, real estate registration data processing method; second, the data mining method of real estate registration data; third, the results of research on spatio-temporal behavior of house buyers in experimental areas. Figure 1-1 below shows the technical roadmap for this paper.
Chapter 1  Introduction

Figure 1-1 Technical Roadmap

- Research background and foundation
  - Research background
  - Research accumulation
  - Literature review
  - Theoretical basis
  - • Research on real estate data
  - • Data mining/Space/Spatio-temporal data mining
  - • Spatio-temporal behavior analysis
  - • Theory of spatio-temporal behavior analysis
  - • Theory of spatio-temporal data mining

  Research content
  - Real estate registration data processing
  - Real estate registration data mining
  - Analysis of Spatio-temporal behavior of housebuyers

- Data processing
  - Real estate registration data
  - Data acquisition
  - Data cleaning
  - Geocoding

- Research methods
  - Content analysis method
  - Experimental method
  - Inductive method
  - Visualization method

- Expected research results
  - Real estate registration data processing method
  - Real estate registration data spatio-temporal data mining method
  - Analysis results of spatio-temporal behavior of housebuyers
Chapter 2  Real Estate Registration Data Sources and Data Processing Methods

2.1 Introduction to the Real Estate Unified Registration System

The introduction of a unified registration system for real estate has changed the status quo of the deregulation of real estate registration agencies, the inconsistent registration basis, and the inconsistency of registration certificates in China. The unified registration of real estate work will be assigned to the Ministry of Land and Resources (now merged into the Ministry of Natural Resources) and be responsible for the “four unifications” of registration agencies, registration basis, registration book, and information platform. By the end of 2017, China has fully implemented access to the national real estate registration information management platform at all levels and actively promoted the integration and exchange of stock data. There are many types of rights involved in the registration of real estate, including ownership of collective land, ownership of buildings and structures such as houses, ownership of forests and trees, land contractual management rights of cultivated land, forest land, and grassland, right to use construction land, right to use homestead, and other real property rights that are required by law to be registered. Among them, the right type of housing transactions is mainly housing ownership, as shown in Figure 2-1. The types of registration of real estate registration are mainly divided into two categories: acquisition of rights (first-time registration, transfer registration, change registration, correct registration); burden or restriction of rights (deregistration registration, advance registration, opposition registration, seizure registration), among which real estate trading is mainly based on transfer registration. The information generated by the unified registration of real estate is all kept in the unified registration database of real estate.

The real estate unified registration system helps the sharing of real estate data. The real estate registration information at all levels is included in the unified real estate registration information management basic platform to ensure the real-time sharing of registration information at the national, provincial, city and county levels. For government departments, the Ministry of Housing and Urban-Rural Development, the Ministry of Agriculture, the Forestry Bureau, and the Bureau of Maritime Affairs must
implement real-time sharing of approval, transaction, and registration information; for the general public, right holders and interested parties can inquire the real estate registration information according to law.

![Real estate registration rights type](image)

**Figure 2-1 Real estate registration rights type**

### 2.2 Acquisitions of Real Estate Registration Data

The real estate registration data needed in this paper comes from the real estate unified registration database. The real estate unified registration database includes registration information generated by various types of rights registrations. The object of this paper is the house buyers (house right holder). Therefore, the registration information generated mainly from the registration of the house ownership, and the required data are generally stored in the three tables of FW (house), QLR (right holder), and XM (item), each field data in the three tables has a unique identification number, Tables can be related by associated fields. The relationship between tables is shown in Figure 2-2. Therefore, this paper through a series of SQL statements from the real estate unified registration database to obtain the real estate registration data required by this paper and in order to ensure that the privacy of the house right owner is not violated, this paper extracts the research data and simultaneously decrypts the personal data of the house right holder. The steps of SQL processing are the following:

The first step: According to the relationship between the tables, relating three tables of
Chapter 2  Real Estate Registration Data Sources and Data Processing Methods

FW, QLR, and XM through related fields, and filtering out the fields required by this paper in these three tables.

The second step is to use the ID number to extract the location of the right holder's household registration, the year of birth, and the gender number.

The third step is to extract the rights holder's purchase year. The right owner's home purchase year is extracted through the "GXRQ" field.

The fourth step: Obtain the age of right holder when they bought the house by the birth date of the right holder and the time of purchase; the age of right holder when they bought the house = year of purchase - year of birth.

The fifth step: Determine gender by gender coding. The sex code extracted from the ID number is used to determine the gender of the rights holder. Even numbers represent women and odd numbers represent men.

The sixth step: Save the new table completed by the query and save it.

<table>
<thead>
<tr>
<th>FW (House)</th>
<th>QLR (House Holder)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FWID (House ID)</td>
<td>QLRID (House Holder ID)</td>
</tr>
<tr>
<td>JYJG (Transaction Price)</td>
<td>QLR (House Holder Name)</td>
</tr>
<tr>
<td>GHYT (Planning Purposes)</td>
<td>QLRSFZJZL (House Holder ID Type)</td>
</tr>
<tr>
<td>FWXZ (House Property)</td>
<td>QLRZJH (House Holder ID Number)</td>
</tr>
<tr>
<td>FWJG (Building Structure)</td>
<td>QLID (Right ID)</td>
</tr>
<tr>
<td>SZC (House Floor)</td>
<td>PROID (Project ID)</td>
</tr>
<tr>
<td>ZCS (Total House Floor)</td>
<td>GYQK (Mutual Ownership State)</td>
</tr>
<tr>
<td>FWZL (House Address)</td>
<td></td>
</tr>
<tr>
<td>JZMJ (House Area)</td>
<td></td>
</tr>
<tr>
<td>FWLX (Housing Type)</td>
<td></td>
</tr>
<tr>
<td>QLID (Right ID)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2-2 Table-to-table association diagram

2.3  Real Estate Registration Data Storage and Problems

2.3.1  Real Estate Registration Data Storage

This paper uses the Oracle database to store and manage the resulting structured real estate registration data. Oracle Database has become the most popular distributed database in the world with its advantages of strong availability, strong scalability, strong
data security, and strong stability. It can add, delete, modify, and check data in the database.

In the previous section, this paper stored the research data required for this paper in a new table using SQL query code. The structure of the table is shown in Table 2-1:

**Table 2-1 Field Information of House Buyers Data Sheet**

<table>
<thead>
<tr>
<th>FIELD NAME</th>
<th>FIELD TYPE</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>QLID</td>
<td>VARCHAR2(32)</td>
<td>Rights ID</td>
</tr>
<tr>
<td>DJLX</td>
<td>VARCHAR2(100)</td>
<td>Registration type</td>
</tr>
<tr>
<td>GXRQ</td>
<td>DATE</td>
<td>Updated</td>
</tr>
<tr>
<td>FWID</td>
<td>VARCHAR2(32)</td>
<td>House ID</td>
</tr>
<tr>
<td>FWZL</td>
<td>VARCHAR2(500)</td>
<td>House location</td>
</tr>
<tr>
<td>JZMJ</td>
<td>NUMBER(20,4)</td>
<td>Construction area</td>
</tr>
<tr>
<td>FWXZ</td>
<td>VARCHAR2(100)</td>
<td>House nature</td>
</tr>
<tr>
<td>FWJG</td>
<td>VARCHAR2(100)</td>
<td>House structure</td>
</tr>
<tr>
<td>GHYT</td>
<td>VARCHAR2(100)</td>
<td>Planning purposes</td>
</tr>
<tr>
<td>QLRID</td>
<td>VARCHAR2(100)</td>
<td>Right holder ID</td>
</tr>
<tr>
<td>QLR</td>
<td>VARCHAR2(100)</td>
<td>Right holder’s name</td>
</tr>
<tr>
<td>HJXZQ</td>
<td>VARCHAR2(12)</td>
<td>Census register location</td>
</tr>
<tr>
<td>CSNF</td>
<td>VARCHAR2(8)</td>
<td>Right holder's birth year</td>
</tr>
<tr>
<td>DJNF</td>
<td>NUMBER</td>
<td>House registration year</td>
</tr>
<tr>
<td>XB</td>
<td>VARCHAR2(2)</td>
<td>Right holder’s gender</td>
</tr>
<tr>
<td>GFNL</td>
<td>NUMBER</td>
<td>Rights holder's purchase age</td>
</tr>
</tbody>
</table>

2.3.2 Problems with existing real estate registration data

As the government does not have the mandate to replace the existing property certificate of house with real estate certificates, only the real estate registration after the implementation of the real estate unified registration system has obtained the real estate certificates, but some people will take the initiative to change their property certificate of house into real estate certificates. Therefore, in the real estate unified registration database, the previous transition data and the current real estate unified registration data coexist.

Due to this characteristic of the real estate unified registration database, some registration data records are duplicated. For example, a house registration data may
appears in the transitional data table and the real estate registration data table at the same time. In the process of real estate registration data integration, due to human factors or the data itself, some important fields are also missing. For example, the “FWXZ” (housing properties) field in the real estate registration data of Dongtai City is missing, and “GXRQ” (registration date) field in Wujiang District real estate registration data is missing. Or the structure of certain attributes in the real estate registration data table in different regions is not uniform, such as the structure of “FWJG” (house structure) in Wujiang District real estate registration data and “FWJG” (house structure) in the real estate registration data of Dongtai City is different. The problems in these real estate registration data have a great impact on the subsequent excavation of real estate registration data, so it is very important to clean the real estate registration data.

2.4 Research on Real Estate Registration Data Processing Methods

Although real estate registration data is the data with small fault tolerance rate, there are still some data quality problems. The "address" attribute in the property is not converted into spatial coordinate data, and corresponding data mining work cannot be performed. Therefore, before the data mining, it is necessary to geocode the data and clean the data to provide a data that meets the requirements for real estate registration data mining. This section mainly introduces the methods of geocoding and data cleaning, and studies the processing methods of real estate registration data.

2.4.1 Data Geocoding Methods

Geocoding is a coding method based on spatial location technology. It provides a method of converting geographical information expressed in addresses into geographic coordinates that can be applied to a GIS system[46]. The process of geocoding is mainly divided into four steps: ① input of initial address data; ② address standardization: processing the input initial address into a common, standard format; ③ address matching: the address to be standardized, using the address model and Encoding rules to match the semantic analysis with the address database to establish the spatial coordinate information and geocoding associated process. ④ Output the matching spatial coordinates.

Due to the particularity of Chinese language and Chinese characters, China’s scholars mainly focus on: ① Analytical methods for geographical names and address elements.
For example, Zhang Xueying et al. \cite{47} designed an RBAI Chinese address element parsing algorithm, expounded the mathematical expression method of Chinese addresses, used numbers to represent the characteristic words in the addresses, and established the tree structure of addresses, and finally split the addresses according to the address resolution rules to match space coordinates. Cheng Gang et al. \cite{48} proposed a generic similarity matching algorithm that takes into account the generic name semantics for standard Chinese place names, improving the recall and accuracy of standard place names matching. The annotation of geographic entities and their spatial relationships. For example, Zhang Xueying et al. \cite{49} based on a comprehensive analysis of the differences in the description and expression mechanisms of geographic entities in Chinese texts and GIS, formulated a geographically named entity annotation system and annotation specification for Chinese texts.

With the rapid development of Internet technology and online mapping services, online geocoding services are increasingly attracting the attention of geographers. The principle of online geocoding service is to use the HTTP interface (Geocoding API) to provide conversion services from address to latitude and longitude or from latitude and longitude to address. Users can send HTTP requests and receive JSON or XML return data using various development languages such as C++, C#, and Java. On-line geocoding services have the advantages of faster address matching algorithm optimization, faster reference database update, and so on, and are increasingly attracting the attention of scholars and users.

As of November 15, 2016, 166 companies have obtained the China’s Class A Internet Map Service mapping qualification. At present, the mainstream online geocoding services in China mainly include online geocoding services such as Baidu, Gaode, Tencent, and Sogou. Tian Qin et al. \cite{50} evaluated the quality of these four online geocoding services.

2.4.2 Online Geocoding of Real Estate Registration Data

In the real estate registration data, the field containing the location information is "FWZL" (address) field. The location information in this field is expressed in the form of address text, which results in that the real property registration data cannot be visualized on the map. Therefore, this paper uses the online geocoding method to convert the location information expressed by address into the location information expressed by coordinates, which provides a good data foundation for subsequent data.
In the era of rapid development of Internet technology and Internet map technology, online geocoding has also been increasingly concerned by GIS users. This article selects online geocoding services as a tool for geocoding. At present, China's mainstream online geocoding services include Baidu Maps, Gaode Maps, Tencent Maps, and Sogou Maps’ online geocoding services. The daily quotas for each map geocoding service vary, as detailed in Table 2-2.

Table 2-2 China's mainstream online geocoding service quota

<table>
<thead>
<tr>
<th>Map Name</th>
<th>Personal Developer Quota (times/day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baidu map</td>
<td>300000</td>
</tr>
<tr>
<td>Gaode map</td>
<td>6000</td>
</tr>
<tr>
<td>Tencent map</td>
<td>10000</td>
</tr>
<tr>
<td>Sogou map</td>
<td>No indicated</td>
</tr>
</tbody>
</table>

Tian Qin et al. [50] compared the quality of China's major online geocoding services. The article points out that in the address matching rate, Gaode > Tencent > Baidu > Sogou; in position accuracy, Tencent > Baidu > Sogou > Gaode. Taking into account the quality of online geocoding service and daily quotas, Baidu online geocoding service was chosen to geocode the house's location.

To use Baidu online geocoding service, you first need to apply for a key on the Baidu LBS development platform, and then call the Geocoding API for forward geocoding. This article uses the Baidu Geocoding API interface and the Java development language to program and completes the geocoding of the “FWZL” field in the real estate registration data. The core code is shown in Table 2-3 below:

Table 2-3 Baidu online geocoding core code

```java
public string[] Geocoding(string address, string city)
{
    string ak = "RTrceGaSoHs6vwlxvFCPejCGybXbY6Ny";
    string[] coordinate = new string[2];
    try
    {
        WebClient client = new WebClient();
        client.Encoding = UTF8Encoding.UTF8;
        string url = String.Format("http://api.map.baidu.com/geocoder/v2/?");
```
ak={0}&address={1}&city={2}&output=json", ak, address, city);
 JObject jo = ((JObject)JsonConvert.DeserializeObject
  (client.DownloadString(url)));
 string status = jo["status"].ToString();
 if (status == "0")
  {
   coordinate[0] = jo["result"]["location"]['lat'].ToString();
   coordinate[1] = jo["result"]["location"]['lng'].ToString();
  }
 }
 catch
 {
 }
 return coordinate;
}

2.4.3 Data Cleaning Methods

2.4.3.1 Definition of Data Cleaning

Data cleaning is a way to remove "dirty data" in the original data set by one or more methods, and to improve the data usage rate. But so far, there is no unified definition of data cleaning methods in the academic community. There are three main areas of data cleaning applications: data warehouse (DW), data mining, and data quality management. The following describes the definition of data cleaning in these three areas:

(1) Data cleaning in data warehouse
In the data warehouse, data cleaning is mainly used to detect and clean data that has been duplicated by different data sources after integration and to eliminate inconsistent and erroneous data.

(2) Data cleaning in data mining
Data cleaning is the primary task in data mining, which is the process of preprocessing raw data. In different application fields of data mining, the methods of data cleaning are different.

(3) Data Cleaning in Data Quality Management
The quality of the data determines the efficiency of the next data application, so data quality management is very important in both business and academic terms. In this area, there is no direct definition of the data cleaning process. Data cleaning is the basic method in data quality management. The principle is to use various cleaning algorithms...
to repair data and improve data quality.

In view of the research objectives of this paper, this article only considers the data cleaning problem in data mining, and defines data cleaning as the process of improving data quality through various cleaning methods based on raw data without redundancy, accuracy, and compliance with data mining application standards. The goal of data cleaning is to detect errors, inconsistencies, duplicate data in the original data, and eliminate or correct them.

2.4.3.2 Data cleaning method classification

Data cleaning is mainly divided into the following five data cleaning methods:

(1) Missing data processing method

The reason for the missing data is the empty field that is not stored in the value itself and the data missed for some reasons.

Missing value processing method: Ignoring instances or attributes that contain missing values; mean filling method: using the mean value of the data as a substitute value; missing data filling method based on incomplete data clustering; clustering methods based on evolutionary algorithms;

(2) Similar duplicate object detection

The main reason of duplicate records is the combination of multiple data sources and duplicate records, and similar duplicate records is one of the most critical issues in data cleaning.

Similar duplicate records detection methods include: proximity sorting algorithm; N-gram based duplicate record detection method; segmentation method.

(3) Abnormal data processing

Abnormal data refers to data objects in the database that do not conform to common sense or general laws, and is also called an isolated point \(^{106}\). The abnormal data may be some no-mining noise data, and it may also be data with mining significance.

The methods of handling abnormal data include: data auditing methods to realize automatic detection of abnormal data, namely data quality mining.

(4) Logical error detection

Data logic errors refer to data that is inconsistent with common sense or violates business rules or logic.

Logical error detection methods include the following: Error data cleaning methods based on business rules.
(5) Inconsistent data

Inconsistent data refers to different representations of the same object when multiple data sources are inherited.

The inconsistent data cleaning methods include: ranking method; fusion method; rule-based method; neighboring sorting method.

2.4.4 Real Estate Registration Data Cleaning

The cleaning of real estate registration data mainly includes the following aspects:

(1) Screening of duplicate data

In the real estate registration data, when the unified registration of real estate is implemented, it is not mandatory to replace the original property certificate of house with the real estate certificate, and it will only be changed when the right is changed and re-registered. This results in the addition of data in the new table, and the data in the old table was not deleted in time. Therefore, the phenomenon that the same property information is registered multiple times has occurred. Another reason for duplicate data is that the registration type of the same house is different, and the house address and the right person are the same. For these duplicated registration data, this paper uses the detection of whether the two fields of house address and right holder's name are duplicated to detect the duplicate data of the real estate registration data. If both fields are duplicated, then the record is marked as duplicate and the duplicated record is removed.

(2) Attribute filtering

In the real estate registration data, in addition to the basic information of the right holder and the house information, there are other fields information, some of which are related to the house information or right holder information, and some of which are not related to the research content of this article. This article classifies attribute values and filters out attribute values that are not related to the goals of this article. For this article, the analysis object is house buyers, so by filtering, deleting the value of the rights holder field that is non-personal records such as companies, village committees, factories. Deletion of records for non-domestic purposes such as "garage", "parking space", "underground garage", and "underground parking space" in the GHYT (planning use) field. Deleting the attribute values in the DJLX (register type) field are "property surveying and mapping results", "cancellation registration", "merge processing" and other non-property registration types.
(3) Logical error data and abnormal data processing
Logical error data refers to data that violates common sense or real estate business logic. Abnormal data refers to data that does not conform to general laws. For real estate registration data, each data is unique, and therefore the logical error data and abnormal data cannot replace or infer the correct content of the logical error data or abnormal data through other fields. Therefore, the logical error data or abnormal data in the real estate registration data is only handled by deleting the data. For example, "0000.x" appearing in the QLR (right holder) field is abnormal data and should be removed. "-19" and "-18" appear in the SZC (house floor) field, which obviously does not conform to the general rule. Those values are exception error and they are also necessary to remove it.

(4) Processing of missing data and null values
In the real estate registration data, the missing data comes from the process of integration of real estate registration data. Due to human error, some or all of the fields may be missing. The null value comes from the registration of real estate, and some fields are themselves allow null values. For example, all values of the "FWXZ" (house properties) field in the real estate registration data of Dongtai City are missing, some values of the "GXRQ" field (registration date) in the real estate registration data of Wujiang District are missing, and some of the values in the "SZC" field (house floor) are null values. For the missing data and null values in the real estate registration data, we cannot obtain the correct values through other methods, so we set the missing value and null value to “Null”, but do not delete the record.

(5) Abnormal spatial position data
When the geographic coordinates obtained through the online geocoding service are not very accurate because of the part of the provided address information are not very precise, some fuzzy matching geographic coordinates will appear. Some of these geographic coordinates are correct, and some of them are wrong. For those coordinates whose spatial positions are obviously abnormal, the spatial data records in the study area can be extracted by spatial relation query, overlay analysis, etc., and the spatial coordinates not in the study area can be excluded. For example, when geocoding, for unmatched addresses, the coordinates of "0,0" are displayed, and there are also some fuzzy matching addresses, which are not within the study area. Since the positions of these coordinates are spatially non-contained relationships to the study area, the data in the range coordinates of the non-study area can be eliminated by superposition analysis.
Chapter 2  Real Estate Registration Data Sources and Data Processing Methods

2.5 Classification of real estate registration data

The object of this study is the house buyers, according to the research needs of this article, house buyers are divided into four categories. The first category is “all house buyers”, referring to all house buyers in a certain research area; the second is “local house buyers”, referring to house buyers whose census register are in this study area; the third is “inner province house buyers”, referring to the census register of house buyers and local buyers are in one province, but not in the research area. The fourth type of buyers are "outer province house buyers", referring to house buyers from other provinces in China. Among them, "inner province house buyers" and "outer province house buyers" are collectively referred to as "migrant buyers".

This article will carry out different research analysis for different types of house buyers. For the first type of house buyers, this paper will analyzes their basic characteristics, and also explores the rules of association between house buyers' properties and the spatio-temporal distribution of buyers' houses; for the other three categories of house buyers, this article will analyze and compare spatial distribution law of their house, meanwhile, analyzing the spatial autocorrelation of "inner province house buyers" and "outer province house buyers", as well as the law of census register changes.

<table>
<thead>
<tr>
<th>House buyer’s classification</th>
<th>Classification basis</th>
<th>Research content</th>
</tr>
</thead>
<tbody>
<tr>
<td>All house buyers</td>
<td>All buyers in the study area</td>
<td>The basic characteristics of house buyers; Association rule mining; The spatio-temporal distribution of houses.</td>
</tr>
<tr>
<td>Local house buyers</td>
<td>house buyers whose census register are in this study area</td>
<td>The basic characteristics of house buyers; House spatial distribution.</td>
</tr>
<tr>
<td>Inner province house buyers</td>
<td>In the same province as local house buyers, but not in the research area</td>
<td>The basic characteristics of house buyers; House spatial distribution; Spatio-temporal dynamic visualization; Spatial autocorrelation analysis.</td>
</tr>
<tr>
<td>Outer province house buyers</td>
<td>house buyers from other provinces in China</td>
<td>The basic characteristics of house buyers; House spatial distribution; Spatio-temporal dynamic visualization;</td>
</tr>
</tbody>
</table>
2.6 Summary of this chapter

This chapter first introduced the source and acquisition method of real estate registration data, and explained the storage method of real estate registration data obtained. Next, based on the data quality problems in real estate registration data and the data mining needs in the future, the geocoding and data cleaning methods of real estate registration data were studied. Finally, the real estate registration data was classified to prepare for the subsequent real estate registration data mining.
Chapter 3  Research on Real Estate Registration Data Mining Methods

3.1 Association Rules Mining

Association rules can find the implicit and meaningful relationship effectively between data in the database. This paper uses Apriori algorithm to mine association rules in real estate registration data. It can be used to find the rules of association between house buyers' age, house floor, total floors of house, the location of the census register, and the building area of the house.

This section introduces related concepts of association rules, and introduces the Apriori algorithm in detail, and studies the implementation of Apriori algorithm in real estate registration data.

3.1.1 Analysis of Association Rules

3.1.1.1 Association Analysis Definition and Basic Terms

(1) Item

Items are the most basic elements of correlation analysis. Each record represents a transaction and each field in the record is an item.

(2) Item Set and Support Count

Let \( I = \{i_1, i_2, \ldots, i_s\} \) be a set of all items in the data, and \( T = \{t_1, t_2, \ldots, t_m\} \) be a set of all transactions. Each transaction \( t_i \) contains a set of items that is a subset of \( I \). In association analysis, a collection containing zero or more items is called an item set. If an item set contains \( m \) items, it is called an \( m \)-item set. For example, \( \{item1, item2, item3\} \) is a 3-item set. The meaning of an empty set is a set of items that does not contain any items.

The number of items that appear in a transaction is defined as the width of the transaction. Assuming that item set \( Y \) is a subset of transaction \( t_j \), then transaction \( t_j \) is said to include item set \( Y \). The item set has a very important property: its support count, which is the number of transactions that contain a specific item set. The support count \( \sigma(Y) \) for item set \( Y \) can be expressed by Equation 3.1:

\[
\sigma(Y) = |\{t_i | Y \subseteq t_i, t_i \in T\}|
\]  

(3.1)

(3) Association rule
An association rule is an implication expression of the form $E \rightarrow F$, where $E$ and $F$ are two disjoint item sets. The two parameters, “support” and “confidence”, are used to weigh the strength of the association rules. “Support” is used to determine the probability that $E$ and $F$ item sets appear at the same time, while “confidence” is used to determine the probability that $F$ appears in transactions that contain $E$. The definitions of “support” and “confidence” can be expressed using Equation 3.2 and Equation 3.3:

$$s(E \rightarrow F) = \frac{\sigma(E \cup F)}{M} \quad (3.2)$$

$$c(E \rightarrow F) = \frac{\sigma(E \cup F)}{\sigma(E)} \quad (3.3)$$

(4) Association Rules Discovery

Association rule discovery refers to finding all the rules with the “support degree” greater than or equal to “minsup” and the “confidence degree” greater than or equal to “minconf” in the set $T$ of a given transaction, where “minsup” and “minconf” are the minimum support and minimum confidence that are defined.

(5) Tasks for association rule mining

The task of association rule mining can be divided into the following two aspects:

① Generation of frequent itemsets: The main purpose of this step is to find all itemsets that are larger than the minimum support, and the itemsets found are called frequent itemsets.

The role of the lattice structure is to enumerate all possible item sets. Figure 3-1 shows the item set of $S=\{a, b, c, d\}$. A data set containing $k$ items can produce $2^k-1$ frequent item sets, which do not include empty sets.

![Figure 3-1 Lattice structure](image-url)
Theorem 1 Prior Principle: If a set of items is frequent, then all of its subsets must also be frequent.

As shown in Figure 3-2, any transaction that contains a set of items \{a,b\} must contain its subset \{a\} and \{b\}, then if \{a,b\} is a frequent itemset, then its subset must be frequent.

Conversely, if a set of items is not a frequent itemset, its superset must not be a frequent item set. As shown in Figure 3-3, supposing that the set of items \{c,d\} is an infrequent itemset, then all supersets \{a, c, d\}, \{b, c, d\} containing the \{c, d\} itemset, and \{a, b, c, d\} must also not be a frequent itemset. This method of pruning the index search space using support measures is called support-based pruning. Support-based pruning depends on an important property of the measure of support: the support of a set item is less than or equal to the support of its subset. This property is called the anti-monotone of the measure of support.
② Generation of rules: The purpose of this step is to extract rules that are greater than or equal to the minimum confidence from the frequent itemsets generated in the first step. These rules are called strong rules. After all frequent item sets are generated, corresponding rules can be generated based on the confidence threshold. When the rule is generated, we need to remove those rules where the front or back piece is empty ($\emptyset \rightarrow M$ or $M \rightarrow \emptyset$), and then each frequent $k$-item set can generate $2^k-2$ association rules. The association rule can be generated by dividing the $M$ item set into two nonempty subsets $N$ and $M-N$ so that $N \rightarrow M-N$ can satisfy the minimum confidence threshold.

3.1.1.2 Classic Algorithm for Association Rules

The algorithms for association rule mining mainly include the Apriori algorithm proposed by Agrawal et al.\cite{51,52} in 1994, the FP-tree frequent algorithm proposed by Han et al.\cite{53} in 2000, and the DHP algorithm proposed by Park et al.\cite{54} in 1995. And so on, here are some simple introductions to these algorithms:

(1) Apriori algorithm
Apriori algorithm is a width-first algorithm. Its principle is to scan all the data in the database to find all frequent itemsets which is the most classic association rule mining algorithm. The Apriori algorithm is mainly divided into two steps: The first is to find out all the frequent itemsets satisfying the minimum support degree according to the minimum support degree set by the user. The second is to find all the association rules that meet the minimum confidence in the frequent itemsets according to the confidence level set by the user.

(2) FP-tree frequent algorithm
The FP-tree frequent algorithm only needs to scan the database twice during the entire process. After the first scan, a frequent $l$-item set is obtained; after the second scan of the database, infrequent items in the database are filtered out, and a FP-tree is generated, and then a recursive algorithm is performed on the FP-tree to mine all frequent patterns.

(3) DHP (Direct Hashing and Pruning) Algorithm
The DHP (Direct Hashing and Pruning) algorithm uses a hash table to generate a candidate set, and a hash table can generate a smaller candidate set, which is also a key point of the DHP algorithm. After constructing the Hash table for the candidate frequent itemsets, the support degree of each item set can be directly obtained from
Chapter 3  Research on Real Estate Registration Data Mining Methods

the Hash table, and the range of the candidate set and the number of times of scanning the database can be effectively reduced.

3.1.2 Apriori Algorithm Overview

In 1993, Agrawal et al. \[52\] first proposed association rules mining. In the following year, Agrawal et al. \[47\] formally proposed the Apriori association rules mining algorithm. Apriori algorithm is the most classical hierarchical algorithm. Its core concept is widely used by other Boolean association mining algorithms. The Apriori algorithm is generally divided into two steps. The first step is to iteratively find out all the matching frequent item sets in the data set according to the minimum support level set by the user. This step is the core part of the Apriori algorithm and it is also the part that each improvement algorithm wants to improve. The second step is to extract the association rule that is not lower than the minimum confidence level from the frequent item sets according to the minimum confidence level set by the user.

(1) Discovery of frequent itemsets

The Apriori algorithm uses support-based pruning techniques to discover frequent itemsets. The algorithm first enumerates all itemsets, and then uses prior principles and support based pruning techniques to prune infrequent itemsets. Table 3-1 shows the pseudo code generated by frequent itemsets:

<table>
<thead>
<tr>
<th>Algorithm 3.1</th>
<th>Apriori algorithm generates frequent itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>( k = 1 )</td>
</tr>
<tr>
<td>2:</td>
<td>( F_k = { i</td>
</tr>
<tr>
<td>3:</td>
<td>Repeat</td>
</tr>
<tr>
<td>4:</td>
<td>( k = k + 1 )</td>
</tr>
<tr>
<td>5:</td>
<td>( C_k = \text{apriori_gen}(F_k - 1) )       { Generate candidate sets }</td>
</tr>
<tr>
<td>6:</td>
<td>For Each transaction do</td>
</tr>
<tr>
<td>7:</td>
<td>( C_t = \text{subset}(C_k, t) )             { Identify all candidates that belong to ( t ) }</td>
</tr>
<tr>
<td>8:</td>
<td>For Each candidate set ( c \in C_t ) do</td>
</tr>
<tr>
<td>9:</td>
<td>( \sigma(c) = \sigma(c) + 1 )              { Support value count increment }</td>
</tr>
<tr>
<td>10:</td>
<td>End for</td>
</tr>
<tr>
<td>11:</td>
<td>End for</td>
</tr>
</tbody>
</table>
Chapter 3  
Research on Real Estate Registration Data Mining Methods

12: \( F_k = \{ c|c \in C_k \land \sigma(c) \geq N \times \text{minsup} \} \) \hspace{1cm} \{ Extract frequent k-itemsets \}
13: \text{Until } F_k = \emptyset
14: \text{Result} = \bigcup F_k

The algorithm first generates all frequent 1-item sets; at the \( n \)th iteration, use the \( n-1 \) frequent itemsets generated by the \( n-1 \)th iteration to generate a new candidate \( n \)-item set; in order to calculate the support of the candidate item set \( n \)-item set, the algorithm scans the entire data set again (step 6-10); after calculating the support level of the candidate set, frequent \( n \)-item sets are extracted for the next iteration; repeat steps 4-12 until no new frequent itemsets are generated and the algorithm ends.

The main purpose of the \textit{apriori_gen} function in step 5 above is to generate a pruning of candidate item sets and candidate item sets. For the generation of candidate sets, the function \textit{apriori_gen} uses the \( F_{k-1} \times F_{k-1} \) method, that is, merging a pair of frequent item sets \((k-1)\)-itemsets, only if they are the first \( k \)-2 items are the same. That is generate a \( k \)-candidate set that combines \( M = \{m_1, m_2, \ldots, m_{k-1}\} \) and \( N = \{n_1, n_2, \ldots, n_{k-1}\} \) pairs of frequent \((k-1)\)-itemsets, and the pair of frequent item sets satisfy the following conditions: \( m_i = n_i \) \((= 1, 2, \ldots, k-2)\) and \( m_{k-1} \neq n_{k-1} \). For the pruning of the candidate set, the function \textit{apriori_gen} uses a support based pruning technique. Table 3-2 below is the pseudo code for the \textit{apriori_gen} function:

\begin{table}[h]
\centering
\begin{tabular}{|l|}
\hline
\textbf{Algorithm 3.2 apriori_gen function} \\
\hline
1: \textbf{For each itemset } \( f_1 \in F_{k-1} \) \\
2: \textbf{For each itemset } \( f_2 \in F_{k-1} \) \\
3: \hspace{1cm} \textbf{If} \( (f_1[1] = f_2[1]) \land (f_1[2] = f_2[2]) \land \ldots (f_1[k-2] = f_2[k-2]) \land (f_1[k-1] = f_2[k-1]) \) \textbf{then} \\
4: \hspace{1cm} c = f_1 \otimes f_2 \hspace{1cm} \{ \text{Connection step} \} \\
5: \hspace{1cm} \textbf{If} \ has\_infrequent\_subset(c, L_{k-1}) \hspace{1cm} \textbf{then} \\
6: \hspace{1cm} \textbf{delete} \ c \hspace{1cm} \{ \text{Pruning step: Delete candidate elements with a subset of infrequent items} \} \\
7: \hspace{1cm} \textbf{Else add} \ c \ \text{to} \ C_k \\
8: \hspace{1cm} \} \\
9: \textbf{Return} \ C_k \\
\hline
\end{tabular}
\end{table}

(2) Generation of Association Rules
The generation of association rules in the Apriori algorithm is based on a layer-by-layer search method. Each layer corresponds to the number of items in the rule back piece. First of all, extracting rules whose rule back piece only has one item that satisfies the minimum confidence rule. Second, use these rules to derive new candidate rules. Algorithm 3.3 and Algorithm 3.4 are the pseudo-code generated by the apriori algorithm association rules:

### Table 3-3 Rule generation in Apriori algorithm

<table>
<thead>
<tr>
<th>Algorithm 3.3 Rule generation in Apriori algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: For every frequent ( k ) - itemset ( f_k ), ( k \geq 2 ) do</td>
</tr>
<tr>
<td>2: ( H_1 = { i \mid i \in f_k } ) { 1-items after pieces rules }</td>
</tr>
<tr>
<td>3: Call ( ap)-genrules( (f_k, H_1) )</td>
</tr>
<tr>
<td>4: End for</td>
</tr>
</tbody>
</table>

### Table 3-4 Ap - genrules function

<table>
<thead>
<tr>
<th>Algorithm 3.4 Function ( ap)-genrules( (f_k, H_m) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: ( k =</td>
</tr>
<tr>
<td>2: ( m =</td>
</tr>
<tr>
<td>3: If ( k &gt; m+1 ) then</td>
</tr>
<tr>
<td>4: ( H_{m+1} = \text{apriori-gen}(H_m) )</td>
</tr>
<tr>
<td>5: For each ( h_{m+1} \in H_{m+1} ) do</td>
</tr>
<tr>
<td>6: ( \text{conf} = \sigma(f_k) / \sigma(f_k - h_{m+1}) )</td>
</tr>
<tr>
<td>7: If ( \text{conf} \geq \text{minconf} ) then</td>
</tr>
<tr>
<td>8: Output: ( \text{rule}(f_k - h_{m+1}) \rightarrow h_{m+1} )</td>
</tr>
<tr>
<td>9: Else</td>
</tr>
<tr>
<td>10: from ( H_{m+1} ) to delete ( h_{m+1} )</td>
</tr>
<tr>
<td>11: End if</td>
</tr>
<tr>
<td>12: End for</td>
</tr>
<tr>
<td>13: Call ( ap)-genrules( (f_k, H_{m+1}) )</td>
</tr>
<tr>
<td>14: End if</td>
</tr>
</tbody>
</table>

(3) Evaluation of Association Rules

Analyze the following example, the relationship between purchase X and purchase Y in a dataset. There are 10,000 transaction records in the data set, 6000 transactions contain X, 7500 contain Y, and 4000 transactions contain both. The data set is shown
Chapter 3  Research on Real Estate Registration Data Mining Methods

in Table 3-5 below:

<table>
<thead>
<tr>
<th></th>
<th>Buy Y</th>
<th>Do not buy Y</th>
<th>Lines total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buy Y</td>
<td>4000</td>
<td>3500</td>
<td>7500</td>
</tr>
<tr>
<td>Do not buy Y</td>
<td>2000</td>
<td>500</td>
<td>2500</td>
</tr>
<tr>
<td>Columns total</td>
<td>6000</td>
<td>4000</td>
<td>10000</td>
</tr>
</tbody>
</table>

Assuming a minimum support is 30% and a minimum confidence is 60%, the following association rules will be discovered:

\[ X \implies Y \]  
\[ (\text{support}(X,Y)=\frac{4000}{10000}=40\%, \text{confidence}(X\implies Y)=\frac{4000}{6000}=66\% ) \]  \hspace{1cm} (3.4)

Both the degree of support and the confidence of this rule met the requirement that we thought a strong association rule was found. However, this is a misleading rule because the probability of buying Y is \( \frac{7500}{10000}=75\% \), which is higher than 66%. In fact, buying X and Y is negatively related because buying one actually reduces the likelihood of buying another. It is easy to make unwise business decisions based on Rule 3.4 without fully understanding this phenomenon.

The above example shows that the confidence level is deceptive, and it cannot successfully filter out rules that we are not interested in. Therefore, new evaluation criteria are needed to measure the actual strength of association between item sets. Therefore, this paper introduces the concept of lift.

Lift is a simple correlation measure defined as follows. If \( \text{confidence}(A \implies B) = \text{support}(B) \), item set A and item set B are independent of each other, otherwise item set A and item set B are dependent and correlation. The lift between item sets A and B can be obtained by:

\[ \text{lift}(A \implies B)=\frac{\text{confidence}(A\implies B)}{\text{support}(B)} \]  \hspace{1cm} (3.5)

If the value of lift is greater than 1, then item set A and item set B are positively correlated; if the value of lift is less than 1, then A and B are negatively correlated; if the value of lift is equal to 1, then A and B are independent of each other.

The lifting degree "lift" ("X\implies Y") = "0.66" / "0.75" "=0.88" of the above example, the degree of lift is less than 1, so X and Y are negatively correlated.

Therefore, this paper not only based on the degree of support and confidence, but also introduced a degree of promotion to evaluate the association rules.

3.1.3  Association Rule Mining Table Establishment

Before mining association rules of real estate registration data, different association rules mining tables need to be established according to different mining targets. This
article mainly explores the buying tendency of house buyers from the six aspects of house buyers' house purchase age, gender, house area, house floor, total house floor, census register location.

The first is the relationship between the total floors of house and the floor of the house. Analyzing the relationship between the two helps to understand house buyers' choice of housing floors. By analyzing the relationship between the total number of floors of the house and the floor where the house buyer’s house is located, it can be known that house buyers are more inclined to purchase which floor of the house. This article extracts the total number of floors of the house and the floor where the house buyer’s house is located in the real estate registration data as the association rule mining item set. Because the format of the “house floor” field is not uniform, fields such as “attic”, “jump layer”, and “203” will appear. To unify the house's floor, the field will be consolidated and the “attic” and other records of unknown layers are removed, and formats such as “203” are unified. Finally, the association rule mining table between the total number of floors of the house and the floor where the house is located is constructed as follows:

Table 3-6 Association rule mining table between the house floor and the total floors of house

<table>
<thead>
<tr>
<th>Transaction</th>
<th>House floor</th>
<th>Total floor of house</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1st floor</td>
<td>6 floors</td>
</tr>
<tr>
<td>T2</td>
<td>5th floor</td>
<td>9 floors</td>
</tr>
<tr>
<td>T3</td>
<td>3rd floor</td>
<td>6 floors</td>
</tr>
<tr>
<td>……</td>
<td>……</td>
<td>……</td>
</tr>
</tbody>
</table>

Through the analysis of the relationship between the census register location of house buyers and the house area, we can understand the situation of house areas of house buyers from different provinces. Nationwide, this paper selected the provincial administrative district as the item set of the association analysis; in Jiangsu Province, the administrative district of prefecture-level cities was selected as the item set of the association analysis. For the building area, this article divides the three groups of 1~90, 90~140, and 140~ according to the proportion of housing deed tax, as the item set of the association analysis. Table 3-7 shows the association rule mining table structure:

Table 3-7 Association rule mining table between the house buyer’s census register location and the house area

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Census register location</th>
<th>House area</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Zhejiang</td>
<td>1~90</td>
</tr>
<tr>
<td>T2</td>
<td>Anhui</td>
<td>1~90</td>
</tr>
</tbody>
</table>
In addition, this paper also analyzes the relationship between the house buyers' age of purchase and the total floor number of the building, the relationship between the house buyers' age of purchase and the floor of the house, and tries to find out the tendency of house buyers with different characteristics. In this association analysis mining table, this article divides the house buyers' age of purchase into 1 to 24 years old, 25 to 34 years old, 35 to 44 years old, 45 to 55 years old, 55 years old and above total 5 groups; And divide the total floor number of the building and the floor of the house into 8 groups. The association rule mining table structure is shown in Tables 3-8 and 3-9 below:

Table 3-8 Association rule mining table between the purchase house age and the total floors of house

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Purchase house age</th>
<th>Total floors of house</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>25~34</td>
<td>1~3 floors</td>
</tr>
<tr>
<td>T2</td>
<td>35~44</td>
<td>4~6 floors</td>
</tr>
<tr>
<td>T3</td>
<td>45~55</td>
<td>7~9 floors</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 3-9 Association rule mining table between the purchase house age and the house floor

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Purchase house age</th>
<th>House floor</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>25~34</td>
<td>4~6 floor</td>
</tr>
<tr>
<td>T2</td>
<td>35~44</td>
<td>7~9 floor</td>
</tr>
<tr>
<td>T3</td>
<td>45~55</td>
<td>19~21 floor</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

3.1.4 Apriori Algorithm Implementation

This article calls the Apriori function in the Arules package in R language to implement association rule mining. The main call functions are:

1. Apriori (data, parameter = NULL, appearance = NULL, control = NULL)

Wherein, data is a data (transaction) class object or any data structure that can be converted into a transaction; parameter represents an APparameter class object or named list; appearance represents an APappearance class object or named list; control represents an APcontrol class object Or named list, used to control the performance of mining algorithms.
2. The `inspect()` function is used to view the generated association rules. Table 3-10 shows the code for mining association rules in R language (taking Wujiang District as an example):

<table>
<thead>
<tr>
<th>R language code:</th>
</tr>
</thead>
<tbody>
<tr>
<td>library(arules)</td>
</tr>
<tr>
<td>wj&lt;-read.csv(&quot;E:/07 毕业论文/数据/吴江区数据/关联分析.csv&quot;,header=TRUE,sep =&quot;,&quot;)</td>
</tr>
<tr>
<td>str(wj)</td>
</tr>
<tr>
<td>wj$房屋总层数&lt;-as.factor(wj$房屋总层数)</td>
</tr>
<tr>
<td>wj_szczcs&lt;-data.frame(wj$房屋所在层,wj$房屋总层数)</td>
</tr>
<tr>
<td>wj_szczcsrules&lt;-apriori(wj_szczcs,parameter=list(support=0.0001,confidence=0.3,minlen=1))</td>
</tr>
<tr>
<td>wj_szczcsrules&lt;-sort(wj_szczcsrules,by=&quot;lift&quot;)</td>
</tr>
<tr>
<td>inspect(wj_szczcsrules)</td>
</tr>
<tr>
<td>wj_nlzcs&lt;-data.frame(wj$购房年龄分组,wj$总层数分组)</td>
</tr>
<tr>
<td>wj_nlzcsrules&lt;-apriori(wj_nlzcs,parameter=list(support=0.0001,confidence=0.3,minlen=1))</td>
</tr>
<tr>
<td>wj_nlzcsrules&lt;-sort(wj_nlzcsrules,by=&quot;lift&quot;)</td>
</tr>
<tr>
<td>inspect(wj_nlzcsrules)</td>
</tr>
<tr>
<td>wj_nlszc&lt;-data.frame(wj$购房年龄分组,wj$所在层分组)</td>
</tr>
<tr>
<td>wj_nlszc&lt;-subset(wj_nlszc,wj$所在层分组!=&quot;其它&quot;)</td>
</tr>
<tr>
<td>wj_nlszcrules&lt;-apriori(wj_nlszc,parameter=list(support=0.0001,confidence=0.3,minlen=1))</td>
</tr>
<tr>
<td>wj_nlszcrules&lt;-sort(wj_nlszcrules,by=&quot;lift&quot;)</td>
</tr>
<tr>
<td>inspect(wj_nlszcrules)</td>
</tr>
<tr>
<td>wj_sf&lt;-data.frame(wj$全国分布,wj$建筑面积分组 3 组)</td>
</tr>
<tr>
<td>wj_sf&lt;-subset(wj_sf,wj$全国分布!=&quot;江苏&quot;)</td>
</tr>
<tr>
<td>wj_sfrules&lt;-apriori(wj_sf,parameter=list(support=0.0001,confidence=0.3,minlen=1))</td>
</tr>
<tr>
<td>wj_sfrules&lt;-sort(wj_sfrules,by=&quot;lift&quot;)</td>
</tr>
<tr>
<td>inspect(head(sort(wj_sfrules, by = &quot;lift&quot;), n =30))</td>
</tr>
</tbody>
</table>
3.2 Spatial Hotspot mining

The real estate registration data containing location information will have its distribution characteristics in the region, and will show hotspots and thermal distribution. After analyzing different spatial clustering methods, this paper selects the kernel density estimation method to mine the hot spot distribution and heat map of the houses purchased by the house buyers in the real estate registration data according to the characteristics of real estate registration data. Spatial hotspot mining can analyze the selection trend of house location for different types of house buyers.

3.2.1 Spatial Clustering Overview

As the most basic method of data mining, cluster analysis has received attention from scholars of all circles in recent years\[^{[5559]}\]. Clustering analysis is an analysis method that divides data into groups based on the similarity of the attribute characteristics of the data. The greater the similarity between the data within the group, the smaller the similarity between the data between the groups, the better the resulting clustering. Clustering analysis does not need to rely on set a priori knowledge to classify data but on the basis of the characteristics of the data itself. For this reason, cluster analysis is sometimes called unsupervised classification\[^{[36]}\].

Spatial clustering analysis is an extension and extension of traditional clustering analysis. It is an important analysis method in spatial data mining. Spatial clustering identifies clusters or dense areas in a spatial database based on a certain distance metric. Similar to the traditional cluster analysis, the spatial data in the same cluster is as small as possible, and the difference between the spatial data of different clusters is as large as possible\[^{[60]}\]. Spatial clustering algorithms can be divided into five types: Partitioning Clustering, Hierarchical Clustering, Density-based Clustering, and Grid-based Clustering, Model-based Clustering.

1) Partitioning Clustering algorithm

The partitioning clustering algorithm is an algorithm to divide a database with \( n \) objects into \( k \) partitions, each partition represents a cluster, and \( k < n \). Moreover, the \( k \) partitions must satisfy the following conditions: 1) A cluster contains at least one object; 2) An object must belong to only one cluster. From this initial \( k \) division, the grouping is continuously optimized by iterative methods to complete the clustering. Commonly used partitioning clustering algorithms are: PAM algorithm, CLARA algorithm, CLARANS algorithm, K-means algorithm\[^{[61]}\], etc. Among them, PAM algorithm is the
first invented K-center point algorithm, K-means algorithm is the most commonly used and most famous algorithm.

(2) Hierarchical clustering algorithm
Hierarchical clustering algorithm is an algorithm to build clusters according to data hierarchy to form a nested tree with nodes as clusters. There are two types of hierarchical clustering algorithms. One is a top-down splitting algorithm, and the other is a bottom-up aggregation algorithm.

The splitting algorithm first places all objects in a cluster, then splits the cluster into two lower clusters, repeating the above steps until each object is a cluster or reaches a certain threshold. Splitting algorithms are not usually used alone, and are often integrated with other clustering algorithms, such as the BIRCH algorithm.

Convergence algorithm is contrary to the disjunction algorithm. At the beginning, every object is a cluster. Then in the subsequent iteration process, the adjacent clusters are merged until all the data are in a cluster or a certain threshold is reached. The common aggregation algorithms are: single connection algorithm, full connection algorithm, average connection algorithm, CURE algorithm, ROCK algorithm, and CHAMELEON algorithm.

(3) Density-based Clustering algorithm
Most of the algorithms describe the similarity of data by distance, which causes the disadvantages that these algorithms have difficulty finding non-spherical data sets. The density clustering algorithm can make up for this shortcoming. This algorithm starts from the distribution density of data objects and connects adjacent areas whose density exceeds a certain threshold to find any shape of clustering. The strength of the algorithm is that it can find clusters of arbitrary shapes and can effectively remove noise. The disadvantage is that the algorithm has high computational complexity and requires pre-specified areas and thresholds. The common density clustering algorithms are: kernel density clustering algorithm, DBSCAN algorithm, DENCLUE algorithm\textsuperscript{[62]}, OPTICS algorithm, etc. Among them, DBSCAN is the most classical density clustering algorithm.

(4) Grid-based Clustering Algorithm
From the perspective of data space division, the grid clustering algorithm first divides the data space into a grid structure of a number of cells, and then operates on a unit-by-cell basis. The processing speed of this algorithm has nothing to do with the size of the data set, but with the number of units, so one of its advantages is the fast processing
speed, and it can also handle any type of data; its disadvantage is difficult to determine the size of the grid. The common grid clustering algorithms are: STNG algorithm, STNG+ algorithm, CLIQUE algorithm, Wave Cluster algorithm and so on.

(5) Model-based Clustering Algorithm

The model clustering algorithm first sets a model for each cluster, and then finds a data set that can be well matched to this model. This method is based on the assumption that the data conforms to the distribution of potential probability. Model clustering algorithms mainly include statistical methods and neural network methods. Among them, representative methods of statistical methods include AutoClass algorithm, COBWEB algorithm, CLASSIT algorithm, Gaussian mixture model, etc. The neural network algorithm has developed rapidly in recent years and the representative algorithms mainly include SOM algorithm, AVQ algorithm, and CPN algorithm.

3.2.2 Kernel Density Estimation Method

The kernel density estimation method is a nonparametric estimation method proposed by Rosenblatt and Parzen. The kernel density estimation method does not require prior knowledge of data distribution, and does not require any assumption about the data distribution in advance. It is a method for studying the data's own characteristics from the data itself. The kernel density estimation method expresses the point data set in the form of a smooth surface, and strengthens the spatial display, reflecting the difference in spatial position.

The kernel density estimation method generally uses the Rosenblatt-Parzen kernel estimate and can be expressed as:

$$f(s) = \sum_{i=1}^{n} \frac{1}{h^2} k \left( \frac{s-c_i}{h} \right)$$

(3.6)

In Equation 3.6, $f(s)$ represents the calculation function of the nuclear density at the spatial position $s$; $h$ represents the distance attenuation threshold; $n$ represents the number of elements whose distance from the spatial position $s$ is less than or equal to $h$; $k$ is the spatial weight function. The geometrical meaning of this formula is that the density value is the largest at each core element $c_i$, and the density value becomes lower and lower as the distance from $c_i$ is longer, and the density value is reduced to 0 when the distance to $c_i$ reaches the distance threshold $h$. The distance attenuation threshold $h$ is related to the spatial analysis scale and the characteristics of geographical phenomena. A small distance attenuation value $h$ allows the density distribution results to show more high-value and low-value areas, which is more suitable for local features of the display.
density distribution; and if the distance attenuation value $h$ is large, it can make the hotspot area more clear in the global scope. The more clear. The specific steps for nuclear density analysis are as follows:

Step 1: Manually setting the "area" of the search radius and cover it to all spatial points in order;
Step 2: Dividing the area into tiny grids based on the artificially set accuracy requirements for the density output;
Step 3: Using the kernel function to calculate the density contribution value of each grid within each area of the point-to-area grid. The kernel function takes a quadratic decreasing function. The closer the distance to the center of the space is, the larger the density contribution value is, and the contribution of the area edge is zero;
Step 4: Assign a value to the density of each grid. This value is the accumulation of all spatial point contribution values within this search radius.
Step 5: Output the density value of each grid. The nuclear density is the continuous density in the given area and the result is the isoline form. The more densely contoured areas, the denser the distribution of spatial points.

### 3.2.3 Spatial hotspot mining of real estate registration data

This article focuses on spatial hotspots for the house distribution of four types of house buyers.

This paper uses ArcGIS Pro software for kernel density analysis and visualization. The ArcGIS product line provides a comprehensive GIS platform for users. ArcObjects includes a number of programmable components that provide users with comprehensive GIS capabilities.

This article uses the kernel density analysis tool provided in ArcGIS Pro for cluster analysis and visualization; its syntax is: 

$$\text{KernelDensity}(\text{in_features}, \text{population_field}, \{\text{cell_size}\}, \{\text{search_radius}\}, \{\text{area_unit_scale_factor}\})$$

The parameter that affects the spatial feature distribution is search radius.

The analysis of kernel density based on the number of house buyers' homes over the years has reflected the characteristics of the spatial distribution of houses and the characteristics of the distribution of houses over time, which can reflect the spatio-temporal changes of house purchase hotspots in each region. The kernel density analysis of the spatial distribution of different types of house buyers can compare the similarities and differences in the spatial distribution of these house buyers and analyze their spatial distribution characteristics.
3.2.3.1 Determination of Search Radius

In kernel density analysis, the larger the search radius parameter, the smoother the generated density raster and the higher the degree of generalization; the smaller the search radius, the lower the degree of generalization of the generated density raster, and the more detailed the displayed content becomes. The requirement of spatial hotspots in real estate registration data is that the generated density raster generalization degree can clearly reflect the spatial distribution characteristics of houses in the study area. This article uses an iterative method to determine the search radius that best fits the study area. The following example illustrates how to select the most suitable search radius: Take Dongtai city as an example, selecting 0.001~0.008 square map units as the search radius for comparison. The results are shown in Figure 3-4. As can be seen from the figure, when the search radius is too small, the raster generalization degree is not enough to reflect the spatial distribution characteristics; and when the search radius is too large, the raster generalization degree is too high, and it is difficult to reflect the spatial distribution characteristics. Comprehensive comparison of kernel density distribution maps with different search radius, when the search radius is 0.004 square map units, the hot areas for buying houses in Dongtai City can be best reflected, which are mainly in the west of the city and south of the city. And the north of the city also has a certain degree of house-buying hotness. When the nuclear densities are analyzed in different research areas, there will be some differences in the optimal search radius. Therefore, when conducting experiments on different research areas, it is necessary to use the iterative method to select the best search radius.

![Kernel density maps with different search radius](image)

Figure 3-4 Kernel density maps with different search radius
3.3 Spatial Autocorrelation Analysis

Spatial autocorrelation analysis can find the spatial trend of data. This paper based on global spatial autocorrelation analysis to find out the spatial autocorrelation of house buyer's census register location distribution. The parameters used in this paper are the Moran index. The following will introduce the concept of spatial autocorrelation and the specific implementation of Moran’s I specific algorithm in real estate registration data in global spatial autocorrelation analysis.

3.3.1 Spatial Autocorrelation Overview

Spatial data is affected by spatial interactions, and there are mutual correlations between data and are no longer independent of each other. In 1970, Tober proposed the first law of geography: “Everything is related to everything else, but near things are more related to each other” [63]. For traditional quantitative statistical models, either focusing only on correlations between numerical values, such as multivariate correlation analysis and unitary correlation analysis, or focusing only on two-dimensional spatial relationships, such as spatial superposition analysis, etc. The emergence of spatial autocorrelation analysis solves this problem very well. It can well balance both, and can perform numerical correlation analysis based on the analysis of spatial relationships. The definition of spatial autocorrelation is that data containing spatial positions shows a certain degree of spatial correlation due to the mutual influence of spatial positions. And it is a spatial statistical method to study the spatial correlation between certain attributes in neighboring regions. Now it has become an important research method in GIS [64].

Spatial autocorrelation is an important indicator to verify whether the attribute value of a spatial element is related to the attribute value of its neighboring spatial element. According to the nature of spatial autocorrelation, spatial autocorrelation can be divided into positive spatial autocorrelation and negative spatial correlation. (negative spatial autocorrelation). The positive spatial correlation means that the attributes of a certain spatial region and the attributes of its neighboring regions have the same change trend. Negative spatial correlation means that certain spatial region attribute and its neighboring region attribute have the opposite change trend.

In 1995, after Anselin [65] proposed the Local Spatial Autocorrelation Index (LISA) to analyze the spatial correlation patterns of spatial data, spatial autocorrelation can usually be divided into global spatial autocorrelation and local spatial autocorrelation.
Global spatial autocorrelation analyzes whether the specified attribute within the whole scope of the study has spatial autocorrelation characteristics, and the local spatial autocorrelation analyzes whether the specified attribute in a specified spatial range has spatial autocorrelation characteristics.

There are also many indicators used to measure spatial autocorrelation. In 1950, the Moran index [6667] suggested that soon the concept of the Geary coefficient [68] was proposed. The establishment of these two coefficients lays the foundation for the application and development of the spatial autocorrelation index.

### 3.3.2 Global Spatial Autocorrelation Algorithm

The global spatial autocorrelation is used to analyze whether the specified attributes have spatial autocorrelation throughout the scope of the study. This article uses the Moran’s index

For global spatial autocorrelation, Moran’s I definition is:

\[
Moran’s \ I = \frac{\sum_i \sum_{j \neq i} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i \sum_{j \neq i} w_{ij}}
\]  

Where: \( n \) is the number of observations; \( x_i \) is the observation at position \( i \); \( \bar{x} \) is the average of \( x_i, i=1,2,...,n \); \( w_{ij} \) is a symmetrical spatial weight matrix; \( S^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \).

The value of Moran's I is calculated by Equation 3.7. If the value is positive and close to 1, then it indicates that the attribute has a positive spatial correlation, that is, the value of each position is similar in a certain range; if the value is negative and approaches -1, then it shows that the attribute has negative spatial correlation and the data are not similar; if the value approaches 0, then the spatial distribution of the data is randomly distributed and there is no spatial autocorrelation.

1. Spatial weight matrix

One of the preconditions for conducting spatial autocorrelation analysis is to construct a spatial weight matrix. Generally, using a binary symmetric spatial weight matrix \( W_{n \times n} \) to represent the spatial neighbor relationship of \( n \) spatial objects. The weighted space neighborhood metric method usually has adjacency criterion, distance criterion, attribute value \( x_j \), and binary spatial weight matrix. The spatial weight matrix is expressed as:

\[
\begin{bmatrix}
W_{11} & W_{12} & \cdots & W_{1n} \\
W_{21} & W_{22} & \cdots & W_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
W_{n1} & W_{n2} & \cdots & W_{nn}
\end{bmatrix}
\]  

(3.8)
If according to the adjacency criterion, when the spatial object \( i \) is adjacent to the spatial object \( j \), then its spatial weight matrix element \( W_{ij} \) is 1, otherwise it is 0, and the expression is as follows:

\[
W_{ij} = \begin{cases} 
1 & (i \text{ is adjacent to } j) \\
0 & (i=j \text{ or } i \text{ and } j \text{ are not adjacent}) 
\end{cases}
\]  
(3.9)

If according to the distance criterion, when the distance between the spatial object \( i \) and the spatial object \( j \) is less than the specified distance \( d \), then its spatial weight matrix element \( W_{ij} \) is 1 and otherwise 0, and the expression is:

\[
W_{ij} = \begin{cases} 
1 & (i \text{ and } j \text{ is less than the distance } d) \\
0 & \text{(other situation)} 
\end{cases}
\]  
(3.10)

If a weighted space neighbor metric method is defined according to the attribute value \( x_i \) and the binary spatial weight matrix, the corresponding spatial weight matrix is defined as follows:

\[
W_{ij} = \frac{w_{ij}x_j}{\sum_{j=1}^{n}w_{ij}x_j}
\]  
(3.11)

(2) Global spatial autocorrelation significance test

Currently, two-sided tests are generally used for the significance test of Moran’s I. The null hypothesis of the test is \( H_0: \text{Moran’s } I = -\frac{1}{n-1} \), which is the complete spatial randomness. It can be the complete spatial randomness of the observation itself, or the complete spatial randomness of the value associated with the observation. The hypothesis test can be performed with reference to the normal distribution table according to the following standardized statistics:

\[
Z_i = \frac{I-E[I]}{\sqrt{\text{Var}[I]}}
\]  
(3.12)

Among them, \( I \) represents the data obtained from the sample; \( E(I) \) represents the expected value of Moran’s I and its value is “\((-1)/(n-1)\)” (\( n \) is sample size); \( \text{Var}(I) \) represents the variance of Moran’s I and its value is \( E[I^2]-E[I]^2 \). The “Z” score represents the multiple of the standard deviation. Suppose the tool returns a “z” score of +3.9. The result is 3.9 standard deviations. The “P” value represents the probability that the observed spatial pattern was created by a random process. When the value of “P” is small, it means that the observed spatial pattern is less likely to be generated by a random process (a small probability event), and the null hypothesis can be rejected. Z-scores and P-values are associated with the standard normal distribution, as shown in
If you get a small P value and a very high or very low Z score, it means that there is a large probability that the observed spatial pattern is not a random pattern represented by the null hypothesis. But to reject the null hypothesis, we must first choose a confidence level, that is, the probability of correctly determining the spatial pattern. In spatial statistics, the typical confidence settings are 90%, 95%, and 99%. This paper selects 95% as the confidence level.

3.3.3 Realization of spatial autocorrelation analysis of real estate registration data

This paper conducts a spatial autocorrelation analysis of the spatial distribution of census register location of inner province house buyers and outer province house buyers over the years. It not only explores whether there is spatial autocorrelation between these two types of house buyers' census register locations, but also analyzes whether the spatial autocorrelation coefficients of the two types of house buyers' census register locations have relevant rules in terms of time.

This article uses ArcGIS Pro software for spatial autocorrelation analysis, the syntax is:

```
SpatialAutocorrelation_stats(Input_Feature_Class, Input_Field, \{Generate_Report\}, Conceptualization_of_Spatial_Relationships, Distance_Method, Standardization, \{Distance_Band_or_Threshold_Distance \}, \{Weights_Matrix_File\}).
```

Among them, Conceptualization_of_Spatial_Relationships represents the type of the input weight field. This article sets the weight field as the statistical quantity of the foreign buyers' census register location;
Distance Method is used to specify how to calculate the distance between each element. That is, how the spatial weight matrix is set. In this paper, we use the spatial weight matrix according to the adjacency criterion. That is, when the two elements are co-edge, common, or overlap, the space weight matrix is 1.

3.4 Dynamic Visualization of Spatio-temporal Data

The dynamic visualization of spatio-temporal data demonstrates the process of data changes with dynamic effects, allowing users to discover some inherent rules of data in a more intuitive way. This article is based on the ECharts open source visualization library, using map animation method to dynamically visualize the spatial distribution of census register location of inner province house buyers and outer province house buyers over the years.

3.4.1 Dynamic Visualization of Data

Data visualization uses data as a tool and visualization as a means to describe truth and explore the world of data. The meaning of scientific computing visualization is to use the development of computer tools, techniques, and systems to transform abstract data into data that humans can visually perceive as visual or graphical to analyze and explore data. Due to the particularity of spatio-temporal data time information and the dynamics of spatio-temporal changes, it is necessary to visualize spatio-temporal data to reflect its temporal dynamics and user-friendly interaction. Therefore, the technical expression of the spatio-temporal data visualization and its expression effects must satisfy the requirements of good dynamic effects, interactivity, and retrospectiveness.

The methods of visualizing space-time data mainly include static map method and dynamic expression method. The static map method mainly uses the map visual variables and graphic charts to convey the changing information. For example, the gradient of a certain symbol in a map or the expression of a moving arrow, or a chronological map. The dynamic expression method is a dynamic visual picture to show the changing process of spatio-temporal data. Dynamic expression can be divided into dynamic map interaction method and map animation representation. The dynamic map interaction method can reflect the development of the research object through the change of the dynamic symbol, and can also control the displayed content and process through the interactive function. The map animation method is a method for viewing a time-sequenced map into an animation through some tools.

3.4.2 Introduction to Visualization Tools

ECharts is an open source visualization library implemented using the JavaScript scripting language that is compatible with most browsers currently on the market. The underlying layer is dependent on ZRender which is a lightweight vector graphics library that provides intuitive, interactive, and highly customizable data visualization diagrams. ECharts provides a variety of charts, including general line charts, scatter charts, bar charts, K-line charts, pie charts, and box charts; heat maps, maps, and line charts for
geospatial data visualization; tree diagrams, sunburst diagrams for relational data visualization; funnel diagrams, dashboards for BI. And Echarts also supports the mix between graphs.

The ECharts displays the produced chart in the form of a webpage, ie HTML hypertext markup language. The steps for creating and displaying a chart using ECharts are:

Step 1: Downloading the ECharts source code package. The source code of ECharts is written in the JavaScript scripting language and stored in JavaScript file with the suffix js. You can obtain the required ECharts package on the ECharts official website or in Github.

Step 2: Introducing the ECharts chart library into the html document by using the script tag in the html as the normal Javascript library.

Step 3: Creating a height and width DOM container for displaying ECharts charts in html;

Step 4: Initializing an echarts instance with the echarts.init method and generate the desired chart via the setOption method.

3.4.3 Dynamic spatio-temporal visualization of real estate registration data

The spatio-temporal dynamic visualization of house buyers inner and outer of the province can more intuitively observe the changes in the number of foreign buyers each year and find out the laws. This article uses the ECharts visualization library, based on the map provided by Baidu as the base map, using the number of house buyers in census register locations as the size of the bubble, and appending the dynamic connection line to display the spatial distribution of census register locations of inner outer province house buyers over the years on the website. The core code is shown below:

```
Table 3-11 Real estate registration data dynamic visualization code

var $debugging = $debugging ? $debugging: {};
$debugging.foreignBuyers = true; // if ($debugging.foreignBuyers) debugger;
var $charts = (!$charts ? $charts: {});
jQuery, echarts].map(function() {
    return (function($charts, $, echarts) {
        var myChart = echarts.init($('#main')[0]);
        console.log(JSON.stringify(createMetaDataOf("盐城市东台市")));
        myChart.setOption(createMapOptionBy(createMetaDataOf("盐城市东台市")));
        if ($debugging.foreignBuyers) debugger;
        $charts.foreignBuyers = myChart;
        return $charts;
        function createMetaDataOf(county) {
            if ($debugging.foreignBuyers) debugger;
```


```javascript
return [{
  destination: "盐城市东台市",
  centerX: 120.314101,
  centerY: 32.853174,
  mapZoom: 6,
  sinceYear: 2009,
  geoCoordMap: mockDongTaiGeoCoordMap,
  seriesData:
  mockDongTaiSeriesData ], as("t").join([county]).as("county").where("t.destination === county").selectOne("t.geoCoordMap = t.geoCoordMap(t.destination, t.centerX, t.centerY), " + "t.seriesData = t.seriesData(), t");

function mockDongTaiData() {
}
function mockDongTaiGeoCoordMap(dest, x, y) {
  if ($debugging.foreignBuyers) debugger;
  var coordMap = mockDongTaiData().reduce(function(prevResult, item) {
    prevResult[item.XZQMC] = [item.coordX, item.coordY];
    return prevResult; }, {});
  coordMap[dest] = [x, y];
  return coordMap; }

function mockDongTaiSeriesData() {
  if ($debugging.foreignBuyers) debugger;
  var years = Array.Range(2009, 2015);
  return mockDongTaiData().as("t").join(years.as("t").select("'buyers' + t[key]")).as("t").groupBy("t.groupKey").select("t.data");

function mockJinTanGeoCoordMap(dest, x, y) {
  if ($debugging.foreignBuyers) debugger;
  var coordMap = mockJinTanData().reduce(function(prevResult, item) {
    prevResult[item.XZQMC] = [item.coordX, item.coordY];
    return prevResult; }, {});
  coordMap[dest] = [x, y];
  return coordMap; }

function mockJinTanSeriesData() {
  if ($debugging.foreignBuyers) debugger;
  var years = Array.Range(2010, 2017);
  return mockJinTanData().as("t").join(years.as("t").select("'buyers' + t[key]")).as("t").groupBy("t.groupKey").select("t.data");
```

3.5 Summary of this chapter

This chapter mainly studies the real estate registration data mining method. The first is the association rules mining method. It introduces the definition of association rules analysis and the classic algorithm. It introduces the Apriori algorithm in detail and expounds the realization of the association rules in real estate registration data. The second is the spatial clustering method. This chapter introduced the spatial clustering method and introduced the kernel density estimation method and the kernel density estimation method in the real estate registration data to realize spatial hotspot mining. The third is the spatial autocorrelation analysis method. It introduces the concept and classification of spatial autocorrelation analysis, and studies the spatial autocorrelation analysis of real estate registration data. Finally, the dynamic visualization method of spatio-temporal data is introduced. The related concepts of data visualization, visualization tools, and spatial-temporal dynamic visualization methods of real estate registration data are introduced.
Chapter 4  A Case Analysis of Saptio-temporal Behavior of House Buyers

4.1 Overview of Research Area and Data Processing

This paper selects Dongtai City, Jintan District of Changzhou City and Wujiang District of Suzhou City in Jiangsu province as the research area of this paper. The three research areas are located in the northern, central and southern parts of Jiangsu Province, and the area GDP is very different. While analyzing the spatio-temporal behavior of different types of house buyers, this article will also compare and analyze whether the regional economic development will affect the spatio-temporal behavior of house buyers.

Dongtai City is located in the middle of the coastal area of Jiangsu Province, the southernmost tip of Yancheng City, with a total area of approximately 3,175.67 square kilometers.

Jintan District of Changzhou is located in the south of Jiangsu Province. It is connected to Wujin District of Changzhou City in the east; Junction of Jurong City in the west; Xiangyang and Yixing City in the south; and the Dantu District of Zhenjiang in the north; The total area is 975.7 square kilometers. In April 2015, with the approval of the State Council, Jintan City at the county level was retired and the Jintan District of Changzhou City was established.

Wujiang District of Suzhou City is located in the southeast of Jiangsu Province. It is connected with Qingpu District in the east and connected with Jiaxing City in Zhejiang Province in the south, Taihu Lake in the west and Wuzhong District in the north. The total area of the district is 1,176.68 square meters. In September 2012, the State Council approved the revocation of Wujiang City at the county level and the establishment of Wujiang District in Suzhou City.

Table 4-1 shows the per capita GDP statistics for the three regions over the years:

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dongtai City</td>
<td>28583</td>
<td>36616</td>
<td>45292</td>
<td>51342</td>
<td>57201</td>
<td>61858</td>
</tr>
</tbody>
</table>
From the table 4-1, we can see that there is a big difference in per capita GDP in the three regions. The economic development of the three regions from low to high is Dongtai, Jintan and Wujiang.

After consultation with the immovable property registration departments in the three regions, the author went to the real estate registration sites in the three regions, under the supervision of the relevant department personnel, using the SQL statement to obtain the required real estate registration data from the real estate registration database on site and it is stored in the Oracle database.

In view of the various data problems in the real estate registration data, this paper using online geocoding and data cleaning method to process the real estate registration data of these three areas, and finally obtains 160000 records of Dongtai City from 2009 to 2014, 330000 records of Jintan district from 2009 to 2014, 340000 records of Wujiang district from 2009 to 2014.

4.2 Analysis of Basic Features of House Buyers

This section mainly analyzes the age characteristics, house purchase area characteristics, and gender characteristics of different types of house buyers, and displays them in a graphical manner to analyzes the laws between them.

4.2.1 All house buyers feature analysis

4.2.1.1 Analysis of Age and Gender Characteristics of House Buyers

An analysis was made of the purchase ages of home buyers in real estate registration data in the three districts of Dongtai City, Jintan District and Wujiang District.

<table>
<thead>
<tr>
<th>Region</th>
<th>Average purchase house age</th>
<th>Median of purchase house age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dongtai city</td>
<td>42.0</td>
<td>41</td>
</tr>
<tr>
<td>Jintan district</td>
<td>41.4</td>
<td>40</td>
</tr>
<tr>
<td>Wujiang district</td>
<td>40.7</td>
<td>39</td>
</tr>
</tbody>
</table>
Table 4-2 shows the average age of purchase houses and median age of purchase houses of house buyers in the three regions. It can be seen that in the three regions, Dongtai has the highest average age of purchase houses for house buyers, Wujiang District has the lowest average age of purchase houses for house buyers, and the median age of purchase houses of house buyers in the three regions is slightly less than the average age, but all are similar, indicating that the distribution of buyers ages of purchase houses in these three regions is basically symmetrical. And the more economically developed areas, the smaller the average age of purchase houses of house buyers.

This paper divides the age of purchase houses of house buyers into 0~24, 25~34, 35~44, 45~54, 55~ five age groups. The specific results are shown in Figure 4-1.

![Figure 4-1 Dongtai city, Jintan District, Wujiang District House Buyers’ Purchase Age and Gender Distribution](image)

Among the house buyers in Dongtai, the 35-44 age group has the most people, followed by the 25-34 age group, again with the 45-55 age group, and the lowest number of buyers is the 0-24 age group. In terms of gender, there are more male buyers than female buyers in all age groups, and the older of the age, the greater the gender gap.

Among the house buyers in Jintan District, the number of house buyers in the 35-44 age group is the highest, followed by buyers in the 45-54 age group, and again in the 25-34 age group and the lowest number of buyers is the 0-24 age group. In terms of gender, there are more male buyers than female buyers in all age groups. Apart from the 1-24 and 55-year age groups, the gap between males and females in other age groups is not large.

Among the house buyers in Wujiang District, the number of home buyers in the 35-44 age group is the highest, followed by buyers in the 45-54 age group, and again in the 25-34 age group and the lowest number of buyers is the 0-24 age group. In terms of gender, there are more male buyers than female buyers in all age groups. Apart from the 1-24 and 55-year age groups, the gap between males and females in other age groups is not large.
age group is the highest, followed by the 25-34 age group, and again the 45-54 age group. The buyers are the youngest among the 1-24 age groups, and with other age groups. The gap is huge. As in the first two regions, there are more male buyers in Wujiang District than female buyers, but the proportion of men and women is almost the same in the 1-24 age group and the 25-34 age group.

The ages of homebuyers in the three regions all exhibited a “pyramid” structure. House buyers in the 35-44 age range are the main buyers, and house buyers in the 24-34 and 45-44 age groups are the sub-main buyers. As for the proportion of male and female house buyers, male of all ages are higher than women. Compared with the degree of economic development, the more economically developed areas, the smaller the difference between the percentage of male and female house buyers.

4.2.1.2 Characteristics Analysis of House Buyers' Purchase Area

According to the proportion of house deed tax payment \cite{69}, this article divides the area of houses into three groups and counts the area of home buyers for the three areas. The statistical results are shown in Figure 4-2 below:

![Figure 4-2 Dongtai, Jintan and Wujiang House Buyers Purchase Area Distribution](image)

As can be seen in Figure 4-2, 90 to 140 square meters of housing are the largest proportion of homes in the three regions, and almost 50% of all homes in the three
regions. The next is 1 ~ 90 square meters of housing, among them, 1 to 90 square meters of housing in Dongtai City accounted for 35.53% of all houses in Dongtai City, accounting for the largest proportion among the three regions. The proportion of houses over 140 square meters is the smallest.

For the distribution of house purchases area in different age groups in the three regions, this article also made a statistical analysis. The statistical results are shown in Figure 4-3 below:

![Figure 4-3 Dongtai, Jintan, and Wujiang House Buyers Purchase Area and Purchase Age Distribution](image)

From Fig. 4-3, it can be seen that among the house buyers of different age groups in Dongtai, the houses purchased by buyers from 1 to 24 years old, the largest proportion of house purchases area is from 90 to 140 square meters, followed by 1 to 90 square meters. House buyers aged 25 to 34 years old have the largest proportion of houses area with 90~140 square meters, followed by 1~90 square meters. The 35 to 44 years old house buyers have the largest proportion of 90-140 square meters, followed by 140 square meters. Among the 45 to 54 years old house buyers, the largest proportion is the
purchase area is 140 square meters or more, followed by 1-90 square meters. Buyers over the age of 55 purchase the largest proportion of houses area is from 1 to 90 square meters, followed by houses area over 140 square meters.

For house buyers of different ages in Jintan District, house buyers in the 1 to 24 age group have the largest proportion of house area with 90 to 140 square meters, and the gap between the other two homes area is small. Among 25 to 34-year-olds house buyers, 90-140 square meters of houses are the largest proportion, followed by 1-89 square meters. The 35 to 44-year-old house buyers have the largest proportion of 90-140 square meters, followed by 140 square meters. The 44 to 54-year-old house buyers have the largest percentage of homes area is 140 square meters or more. Buyers over the age of 55 have the largest proportion of house area is 1 to 90 square meters, followed by 90 to 140 square meters.

Among the buyers in Wujiang District, house buyers from 1 to 24 years old have the largest proportion of houses areas is 140 square meters or more, followed by 1 to 90 square meters. In the 25 to 34-year-old house buyers, the largest proportion of houses areas is 1 to 90 square meters, followed by 90 to 140 square meters. The 35 to 44-year-old house buyers have the largest proportion of house areas is 90 to 140 square meters, followed by 1 to 90 square meters. Among the 45-54-year-old house buyers, the largest proportion of houses areas is 140 square meters or more, followed by 90 to 140 square meters. Buyers over the age of 55 purchase the largest proportion of houses area is the 1 to 90 square meters, followed by houses area over 140 square meters.

The distribution of house purchases area varies from age to age. For the 25~34 age group, the purpose of house purchase is mainly to just need a house, and the buyers do not have sufficient economic strength. Therefore, buyers of this age group mainly purchase houses area is less than 140 square meters. For 35 to 44 age group, there is a certain amount of economic accumulation, so the size of houses purchased by buyers in this age range is scattered, and the proportion of houses in different house areas is not much different. The house buyers of 44 to 54 years old mainly aim to improve their current housing condition, and they also have a certain degree of economic strength. Therefore, house buyers in this age group mainly purchase enjoyable houses of area more than 140 square meters. As for the distribution of housing areas for house buyers aged 1-24 years and over 55 years old, the reasons for this need further verification.
4.2.2 Analysis of Basic Characteristics of Different Types of House Buyers

4.2.2.1 Analysis of Distribution of Different Types of House Buyers

The level of economic development in the three regions is different, so the proportion of different types of house buyers is also very different. The specific statistical results are shown in Figure 4-4:

From Figure 4-4, it can be seen that the vast majority of house buyers in Dongtai are local house buyers which accounting for 94.86%, while the inner province house buyers and outer province house buyers only account for 2.72% and 2.43%. Most of the house buyers in Jintan District are also local house buyers which accounting for 88.06%, while the inner province house buyers and outer province house buyers only account for 5.6% and 6.34%. But the local house buyers in Wujiang District accounted for only about half of the purchases, and the proportion of house buyers inner province and outer province are quite different which accounting for 17.1% and 31.6%. The following conclusions can be drawn: The proportion of foreign house buyers in the region is related to the degree of economic development in the region. The larger the proportion of foreign buyers in the more economically developed regions, the greater the attraction of foreign buyers.

4.2.2.2 Analysis of Age Characteristics of Different Types of House Buyers

This section analyzes the characteristics of the purchase age of different types of house buyers in different regions.

| Table 4-3 Average purchase age of different types of house buyers in different regions |
|-------------------------------------|-------------------------------|-------------------|
| Local house buyers | Inner province house | Outer province |

55
Table 4-3 above shows the average purchase age and the median purchase age for different types of house buyers in different regions. It can be seen that the average purchase age of local house buyers is the highest in all three regions, and the difference between the average purchase age of the local house buyers and the foreign house buyers is 5 to 7 years old. The average purchase age of the inner province house buyers is the smallest, and the average purchase age of the outer province house buyers is not much different from the inner province house buyers. From the above analysis, it can be seen that the average age of different types of house buyers is not significantly related to the economic development of the region.

Similarly, the age of home purchase is divided into five age groups: 0-24, 25-34, 35-44, 45-54, 55-. The following are the specific statistical results:

![Different age distribution of the different types of house buyers in Dongtai city](image)

Figure 4-5 Different age distribution of the different types of house buyers in Dongtai city
Chapter 4  A Case Analysis of Saptio-temporal Behavior of House Buyers

From Figure 4-5, it can be seen that among the local house buyers in Dongtai City, the number of house buyers in the 35-44 age group is the largest, followed by the buyers in the 45-54 age group, so the buyers in the 25-54 age group are the main buyers, and the age of the buyers distribution presents a "pyramid" type. The percentage of 25 to 34 and 35 to 44 age groups in inner province house buyers is not much different and the 25 to 44-year-old house buyers are the main buyers, and the age distribution of buyers is “trapezoidal”. The situation of outer province house buyers is similar to the situation of inner province house buyers. House buyers from 25 to 44 years old are the main buyers.

![Different age distribution of the different types of house buyers in Jintan district](image)

**Figure 4-6** Different age distribution of the different types of house buyers in Jintan district

It can be seen from Figure 4-6 that among the local house buyers in Jintan District, the number of house buyers in the 35-44 age group is the highest, followed by the buyers in the 45-54 age group, the number of buyers in the 25-34 age group and the 45-54 ages have not much difference, so the house buyers in the age range of 25 to 54 years old are the main force in purchasing houses, and the age distribution of purchases presents a “pyramid” type. Among the inner province house buyers, the number of buyers aged 25 to 34 is the highest, followed by buyers in the 35 to 44 age group. These two age groups are quite different from other age groups and are the main force for buying
houses. The proportion of outer province house buyers from the 25 to 34 and 35 to 45 age groups is almost the same, and the number is far higher than that of the other age groups which is the main force for buying homes.

From the previous section, it can be seen that foreign buyers in Wujiang District account for about 50% of all home buyers in Wujiang District. The age distribution of their purchase age is different from Dongtai and Jintan Districts. From Figure 4-7, it can be found that among the local house buyers in Wujiang District, the number of house buyers from 45 to 54 is the largest, followed by those from 35 to 44. Although the 25 to 54-year old house buyers are still the main buyers, but the proportion of buyers aged 25 to 34 is the lowest. Among the inner province house buyers, 25-34-year-old house buyers are the most, followed by 35-44-year-old buyers, the proportion of the two is not much difference and they are the main buyers. Among the outer province house buyers, the proportion of buyers in the 35-44 age group is the largest, followed by the buyers in the 25-34 age group and these two age groups are the main buyers.

From the situation of the purchase age distribution of different types of buyers in three regions, we can draw the following rules: The average purchase age for house buyers is irrelevant to the level of regional economic development, and is related to the census...
register location of house buyers. The average purchase age of local house buyers is greater than the average purchase age of foreign buyers. The main buyers of local house buyers are buyers of 25 to 54 age groups, and the buyers of foreign house buyers are buyers of 24-44 age groups.

4.3 Mining Association Rules of Real Estate Registration Data

Using real-estate registration data association rule mining method, apply to real estate registration data of the three experimental areas in this paper. According to different mining targets, different fields in real estate registration data were selected to study the behavioral association rules among census register location, purchase area, purchase age, floor of house, and total number of floors of houses in three regions of Jiangsu Province.

After the association rule mining table is established, the Apriori function in the Arules package in R language is used to implement the association rules mining, and finally got the related association rules.

4.3.1 Association Rules Mining Results

(1) Association rule between the floor of the house and the total number of floors
This paper studies the association rules between floor of house and total number of floors of house of house buyers in different districts in Jiangsu Province. The real estate registration data of Jintan District and Wujiang District are selected. The support level is 0.0001, and the minimum confidence level is 0.3. The result of the association rule is ranked in descending order of degree of increase. The result of the association rule is
as follows:

<table>
<thead>
<tr>
<th>lhs</th>
<th>rhs</th>
<th>support</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(30, 房屋所在层数=31)</td>
<td>(30, 房屋总层数=34)</td>
<td>0.0012851076</td>
<td>1.0000000</td>
</tr>
<tr>
<td>(30, 房屋所在层数=26)</td>
<td>(30, 房屋总层数=26)</td>
<td>0.0016181363</td>
<td>1.0000000</td>
</tr>
<tr>
<td>(30, 房屋所在层数=30)</td>
<td>(30, 房屋总层数=34)</td>
<td>0.0014513292</td>
<td>0.0000000</td>
</tr>
<tr>
<td>(30, 房屋所在层数=25)</td>
<td>(30, 房屋总层数=26)</td>
<td>0.0022039138</td>
<td>0.6855566</td>
</tr>
<tr>
<td>(30, 房屋所在层数=31)</td>
<td>(30, 房屋总层数=34)</td>
<td>0.0031529238</td>
<td>0.3170727</td>
</tr>
<tr>
<td>(30, 房屋所在层数=29)</td>
<td>(30, 房屋总层数=34)</td>
<td>0.0001765507</td>
<td>0.8121281</td>
</tr>
<tr>
<td>(30, 房屋所在层数=28)</td>
<td>(30, 房屋总层数=34)</td>
<td>0.0022039138</td>
<td>0.6855566</td>
</tr>
<tr>
<td>(30, 房屋所在层数=24)</td>
<td>(30, 房屋总层数=26)</td>
<td>0.0031529238</td>
<td>0.3170727</td>
</tr>
</tbody>
</table>

Figure 4-8 The association rules between house layer and the total house layers in Jintan District

Figure 4-8 shows the association rules between the house layer and the total number of layers of house in Jintan District, summarizing the obtained association rules and explaining the rules. Among buyers who purchase houses in Jintan District:

① If the buyer's house is on the 26th, 27th, 28th, 29th, 30th, 31st, 32nd, and 33th floors, the total number of floors of this house is most likely 34 floors. (The lift is greater than 60)

② If the house of the buyer’s house is on the 28th floor, the total number of floors of this house is most likely 32 floors. (The lift is 109.00.)

③ If the house of the buyer's house is on the 23rd and 24th floors, the total number of floors of this house is most likely 26 floors. (The degree of lift is 44.44, 50.76)

④ If the house of the buyer’s house is on the 18th floor, the total number of floors of this house is most likely 19 floors. (The degree of lift is greater than 9)

⑤ If the floor of the buyer's house is on the 13th, 14th, 15th, 16th, and 17th floors, the total number of floors of the house is likely to be 18 floors. (The degree of lift is greater than 9)

⑥ If the buyer's house is on the 7, 8, 9, and 10 floors, the total number of floors of the house is likely to be 11 floors. (degree of lift greater than 4)

⑦ If the home of the buyer’s house is on the 1st floor, the total number of floors of the house is most likely to be 1 floor. (the lift is 7.70)
⑧ If the home of the buyer’s house is on the 2nd floor, the total number of floors of
the home is most likely to be 2 floors. (the lift is 4.17)
Through these rules, it is not difficult for us to find that house buyers in the Jintan
District, whether they are low-rise or high-rise buildings, house buyers tend to purchase
higher-rise buildings in their buildings. It is not difficult to understand that as urban
land becomes scarce and high-rise residential buildings gradually become mainstream,
high-rise buildings have the advantages of good lighting, broad vision, and low noise,
so house buyers will tend to buy higher-rise buildings.

<table>
<thead>
<tr>
<th>lhs</th>
<th>rhs</th>
<th>support</th>
<th>confidence</th>
<th>lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>{w1, 房屋总层数=3} =&gt; {w2, 房屋所在层=3层}</td>
<td>0.001132183 0.6765281 9.148252</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋总层数=4} =&gt; {w2, 房屋所在层=1层}</td>
<td>0.001137566 0.3414604 6.189947</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=33层} =&gt; {w2, 房屋总层数=34}</td>
<td>0.001517286 0.4307692 5.287608</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=32层} =&gt; {w2, 房屋总层数=34}</td>
<td>0.002752929 0.4221126 5.227214</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=35层} =&gt; {w2, 房屋总层数=34}</td>
<td>0.002049407 0.4093023 5.115322</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=33层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.002004905 0.3652305 5.041319</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=32层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.003251328 0.5581395 4.917597</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=32层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.002794539 0.4705977 4.471760</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=30层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.002964020 0.4952065 4.102693</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=30层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.002592656 0.8093533 4.138559</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=29层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.002655251 0.3034056 3.754677</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=29层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.003711932 0.4241486 3.737044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=28层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.003251328 0.3750000 3.304011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>{w1, 房屋所在层=27层} =&gt; {w2, 房屋总层数=35}</td>
<td>0.003332611 0.3029557 2.669259</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-9 The association rules between house layer and the total house layers in Wujiang District

Figure 4-9 shows the association rule between the house floor and the total number of
floors in Wujiang District, explaining the rules:
① If the total number of floors in the house is 3 floors, it is very likely that house
buyers will buy the house on the 3rd floor. (lift is 9.15)
② If the total number of floors in the house is 4 floors, buyers are likely to purchase
the house on the 1st floor. (lift is 6.19)
③ If the house of the buyers is on the 33rd, 32nd, 31st, 30th, and 29th floor, then the
total number of floors of the house is likely to be 34th floor. (The degree of lift is greater
than 3.7)
④ If the house of the buyers is on the 27th, 28th, 29th, 30th, 31st, 32nd, then the total
number of floors of the house is likely to be 35th floor. (The degree of lift is greater
than 2.6)
Through these association rules, we can see that house buyers in Wujiang District, if
they choose to buy high-rise residential buildings, will tend to choose higher-level
housing; for low-rise residential houses, there are no obvious rules.
(2) Association rules between the census register location and the building area of the
house
This article examines the relationship between the census register location of house buyers and building area of his house and taking Wujiang District and Dongtai City as an example. In the national context, provincial-level administrative areas are taken as the items of association analysis, within Jiangsu Province, prefecture level administrative areas are taken as the items of association analysis, the support degree is set to 0.001, and the confidence degree is set to 0.3. The resulting association rules are ranked according to the degree of lift. The result is shown in Figure 4-10.

**Figure 4-10 The Association Rules between Census Register Location and House Area in Dongtai City**

Summarize the obtained association rules for Dongtai City and explain as follows:

① If the census register location of the house buyer who purchases a house in Dongtai City is Beijing, it is more likely that the house area he purchased is 140 square meters or more.

② If the house buyers’ census register location are located in the Ningxia Hui Autonomous Region and Guizhou Province, the building area he purchases is likely to be 90 to 140 square meters.

③ If the house buyers’ census register location are located in Shanxi Province, Jiangxi Province, Liaoning Province, and Sichuan Province, the building area they purchase is likely to be 1 to 90 square meters.

**Figure 4-11 The Association Rules between Census Register Location and House Area in Wujiang District**

Figure 4-11 shows the association rules between the census register location and the building area of the house of Wujiang District. The rules are explained as follows:

① If the census register location of house buyers who purchase homes in Wujiang
District are located in Wuxi City, Nanjing City, Zhenjiang City, and Changzhou City, it is more likely that the houses they purchase will have a building area of 140 square meters or more.

2. If the census register location of house buyers who purchase homes in Wujiang District are located in Guangdong Province and Zhejiang Province, it is more likely that the house purchased by them will have a building area of 140 square meters or more.

Through the above-mentioned association rules, we can find that the house buyers of different census register location in different regions have relatively large differences in the house area. Among the house buyers in Jintan District, house buyers in Xinjiang and Jilin Province tend to purchase 1~89 square meters of housing. And there is no strong correlations have been found between other provinces and building areas. Among the house buyers in Wujiang District, buyers in Wuxi city, Nanjing city, Zhenjiang city, Changzhou city, Guangdong and Zhejiang Provinces tend to purchase more than 140 square meters of housing. It is not difficult to explain that Wuxi city, Nanjing city, Zhenjiang city, Changzhou city, Guangdong and Zhejiang Provinces are all economically developed provinces and cities. Buyers in these areas also have the ability to purchase more than 140 square meters of housing. Among house buyers in Dongtai district, buyers whose census register location is Beijing tend to purchase more than 140 square meters of housing, which is related to the economic strength of Beijing house buyers. Buyers from Ningxia and Guizhou provinces tend to purchase 90-140 square meters of housing and buyers from Shanxi, Jiangxi, Liaoning, and Sichuan provinces tend to purchase 1 to 90 square meters of housing.

3. Association rules between the age of house buyers and the house floor

This article analyzes the association rules between the ages of house buyers and the house floor in different regions, and attempts to know the choice of house floor for different age groups. In this paper, the data of Jintan District and Wujiang District are selected as experimental data. The support degree is set to 0.001, the confidence level is set to 0.3, and the association rules are arranged in ascending order of descending degrees.

<table>
<thead>
<tr>
<th>LHS</th>
<th>RHS</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>( j.t. \text{所在层分组} = 22 \text{层及以上} )</td>
<td>( j.t. \text{购房年龄分组} = 25-54 )</td>
<td>0.002000908</td>
<td>0.3622527</td>
<td>1.6122261</td>
</tr>
<tr>
<td>( j.t. \text{购房年龄分组} = 55-1 )</td>
<td>( j.t. \text{所在层分组} = 1-3 )</td>
<td>0.113279222</td>
<td>0.6772216</td>
<td>1.3608340</td>
</tr>
<tr>
<td>( j.t. \text{所在层分组} = 19-21 )</td>
<td>( j.t. \text{购房年龄分组} = 26-34 )</td>
<td>0.005109747</td>
<td>0.3034587</td>
<td>1.3553998</td>
</tr>
<tr>
<td>( j.t. \text{购房年龄分组} = 25-34 )</td>
<td>( j.t. \text{所在层分组} = 6-6 )</td>
<td>0.087659886</td>
<td>0.3969377</td>
<td>1.2643228</td>
</tr>
<tr>
<td>( j.t. \text{购房年龄分组} = 1-24 )</td>
<td>( j.t. \text{所在层分组} = 6-6 )</td>
<td>0.023237435</td>
<td>0.3786820</td>
<td>1.2256819</td>
</tr>
</tbody>
</table>

Figure 4-12 The Association Rules between the Purchase Age of Buyers and the house floor in the Jintan District
Figure 4-12 shows the association rule between the age of buyers and the floor where the house is located in Jintan District, and explains the rules:

① If the house where the house is purchased in the Jintan District is located at 22 floors or more, the possibility of the purchaser's age is between 25 and 34. (degree of lift is 1.62)

② If the house buyer who purchases a house in Jintan District is 55 years old or older, then the house of the buyer's is more likely to be located on floors 1-3. (degree of lift is 1.40)

③ If the house where the house is purchased in the Jintan District are located between 19 and 21 floors, the age of the buyers is likely to be between 25 and 34 years old. (The degree of lift is 1.36)

Figure 4-13 The Association Rules between the Purchase Age of Buyers and the house floor in the Wujiang District

Figure 4-13 shows the rules for the association between the age of buyers and the floor where the house is located in Wujiang District, and explains the rules:

① If the house buyer who purchases a house in Wujiang District is 55 years old or older, then the home of the purchaser is more likely to be located on floors 1 to 3. (The degree of lift is 1.54)

② If the house is located in the 19th to 21th floors, it is more likely that the house buyer is between 1 and 24 years of age. (degree of lift is 1.11)

③ If the house is located between 13 and 15 floors in the Jintan District, the age of the house buyer is likely to be between 1 and 24 years old. (degree of lift is 1.10)

Through the above-mentioned association rules, we can find that house buyers in these two districts of Jiangsu Province have a greater correlation between the age of buyers and the floor where the house is located: younger house buyers (under 34 years of age) are more likely to be in high-level housing, which is related to young people like open horizons and quiet environment. The older house buyers (aged 55 and above) who buy a house are more likely to have a low-level house, which is related to factors such as the ease of going out and the sense of security.

(4) Association rules between the age of buyers and the total floors of houses
This paper analyzes the association rules between the age of buyers and the total number of houses in different regions, trying to know the choice of the total floors of houses by different age groups. In this paper, the data of Jintan District and Wujiang District are selected as experimental data. The degree of support is set to 0.001, the confidence is set to 0.3, and the association rules are arranged in ascending order of descending degrees.

Figure 4-14 The Association Rules between the Purchase Age of Buyers and the total floors of houses in the Jintan District

Figure 4-14 above shows the association rules between the age of buyers and the total floors number of house in Jintan District. The rules are explained as follows: Among the home buyers in Jintan District:

① If a house buyer in Jintan District is 55 years old or older, then the house buyer’s house is likely to have a total floor number of 1 to 3 floors. (degree of lift is 2.00)

② If the total number of houses purchased in the Jintan District is between 1 and 3 floors, the likelihood of the house buyers being 55 years old or older. (degree of lift is 2.00)

③ If the total number of houses purchased in the Jintan District is between 13 and 15 floors, the age of the house buyer is likely between 25 and 34 years old. (degree of lift is 1.41)

④ If the total number of houses purchased in the Jintan District is between 1 and 3 floors, the possibility of the age of the house buyer is between 45 and 54 years old. (The degree of lift is 1.32)

Figure 4-15 The Association Rules between the Purchase Age of Buyers and the total floors of houses in the Wujiang District

Figure 4-15 above shows the association rules between the age of buyers and the total floors number of house in Wujiang District. The rules are explained as follows: Among the home buyers in Wujiang District:

① If the total number of houses purchased in Wujiang District is between 1 and 3
Chapter 4  A Case Analysis of Saptio-temporal Behavior of House Buyers

floors, the age of the house buyers is likely to be 35 to 44 years old. (degree of lift is 1.31)

② If the total number of houses purchased in Wujiang District is between 7 and 9 floors, the age of the house buyers is likely to be 35 and 44 years old. (degree of lift is 1.28)

③ If the house where the house is purchased in Wujiang District is between 19 and 21 floors, the age of the house buyer is likely to be between 1 and 24 years old. (Enhancement 1.22)

④ If the house where the house is purchased in Wujiang District is between 16 and 18 floors, the age of the house buyer is likely to be between 1 and 24 years old. (Enhancement 1.22)

Through the above association rules, we can find that in these two regions of Jiangsu Province, younger house buyers (below 34 years of age) have a higher total floor number of houses (13 floors or more), and older house buyers (over 35 years old), the total number of floors in the house is low (less than 9th floor). This is similar to the association rule between the age of house buyers and the floor where the house is located.

4.3.2 Result analysis

Based on Apriori algorithm, this paper explores the association rules of the real estate registration data of Jintan district, Wujiang district and Dongtai city from 2009 to 2014, and obtains some meaningful association rules.

(1) Different age groups also have different choices for the total number of houses and the floor where the house is located: low age house buyers tend to buy top houses in high-rise buildings, and high age house buyers tend to buy top of multi-storey or ground-floor houses. This is because as urban land becomes more and more scarce, high-rise residential buildings have gradually become mainstream. Young house buyers tend to choose high-rise buildings and even higher-rise residential buildings, while young people have the characteristics of strong adaptability and love of new things. They can accept the advantages of high-rise buildings with broad vision, good air, and low noise. They can also accept the disadvantages of lack of convenience and difficulty in escaping when there are unexpected situations.

(2) The purchase behavior of foreign buyers in the three regions is closely related to the economic development level of the census register location of the house buyers: Generally speaking, the more economically developed the census register location of
the buyers, the more bigger of the building area of the houses they purchased.

Through the mining of association rules of real estate registration data, we can understand the preferences of house buyers of different ages. Real estate developers can propose business strategies for buyers of different ages according to these rules.

4.4 Spatial Hotspot Mining and Analysis of Real Estate Registration Data

4.4.1 The results of mining hotspots by all house buyers

Before conducting kernel density analysis in three regions, we should first use the iterative method to determine the best search radius for each region. After experiments, the best search radius in Dongtai City is 0.004 map units, and the best search radius in Jintan District is 0.004 map units, Wujiang District's best search radius is 0.007 map units. This article also compares and analyzes the spatial hotspot findings results of all homebuyers with the urban development directions proposed in the overall urban planning of cities and towns, and checks whether the purchase trend changes according to the development direction proposed in the urban planning.

(1) Dongtai City hot spot mining results

This paper used the real estate registration data of Dongtai City from 2009 to 2014, using kernel density estimation method, using 0.004 map units as the search radius, forming a heat distribution map of house buyers in Dongtai City from 2009 to 2014. The heat map is shown in Figure 4-16:
According to the house distribution changes of house buyers in Dongtai City, it can be seen that from 2009 to 2014, the distribution of houses in the main urban area of Dongtai City was mainly concentrated in the old city, and afterwards it was generally developed to the north and south of the old city. The northern economic and technological development zone is also a hot spot for house purchases, while the “Eastern District” east of the Tongyu River has few hot spots for house buyers.

(2) Results of mining hotspots in Jintan District
This paper used the real estate registration data of Jintan District from 2009 to 2014, based on the kernel density estimation method, using 0.004 map units as the best search radius, forming a heat distribution map of buyers' houses in Jintan District from 2009 to 2014. The figure is shown in Figure 4-17:

According to the house distribution changes of house buyers in Jintan City, it can be seen that the hotspots for purchase houses in Jintan District gradually moved to the east of Jintan District from 2009 to 2014, and the southwestern region also gradually became a hot spot region of housing purchase, but the number of north home buyers is gradually decreasing.

(3) Results of mining hotspots in Wujiang District
This paper used the real estate registration data of in Wujiang District from 2009 to 2014, based on the nuclear density estimation method, using 0.007 map units as the best search radius, forming a heat distribution map of the purchase of houses from 2009 to 2014. The figure is shown in Figure 4-18:
From the thermal map of Wujiang District housing distribution from 2009 to 2014, it can be seen that the purchase of houses in Wujiang District has moved mainly to the south of the urban area in the past few years. In the following years, it mainly moved to the southwest of the city, namely the lakeside area and the south and west of the city, and the general direction is westward and the southwest.

4.4.1.1 Results Analysis

This article compares hotspot mining results with the overall urban planning of each region. The analysis results are as follows:

The urban master plan of Dongtai City from 2008 to 2020 points out that the core urban area of Dongtai City develops mainly eastward, extends northwards, southwards, and westwards, and forms the “Xicheng Dongqu” pattern with the Tongyu River as the boundary. “Xicheng” is mainly based on the old town and the economic development zone in the north, and it has developed into a “one-core, three-cluster” urban area with
integrated functions and complete facilities. The “Dongqu” is mainly based on the secondary industry and expects to form “Northern Engineering South Habitat” space pattern. According to the map of house distribution changes of house buyers in Dongtai City from 2009 to 2014, it can be seen that the trend of buying houses in Dongtai City is generally developed in the north and south of the old town. The economic and technological development area in the north is also a hot spot for buying houses, while the “Dongqu” has few hot spots for buying houses. The distribution of “Dongqu” is mainly due to the fact that the “Dongqu” is dominated by the secondary industry and has not vigorously developed the real estate industry.

In April 2015, with the approval of the State Council, Jintan City was retired at the county level and Jintan District of Changzhou City was established. The urban master plan of Jintan City from 2013 to 2030 points out that the core urban area mainly refers to the area surrounding the Xindan Jinyi River, Jinyi Highway, No. 340 Provincial Highway, Quantang River, Shuibe Road, Jintang Road, and Changhe Expressway. The total area is approximately 118.4 square kilometers. The Old town refers to the area enclosed by Ximen Street, Xihuan First Ring Road, Hengjie Street, Dongmen Street, and Donghuan Ring Road. The total area is about 5.77 kilometers. The central urban area as a whole constitutes the spatial structure of the “Two Cities and Two Areas” where the production and the city are integrated. “Two Cities” are the old town and Lakeside New City. The “Two Areas” is Chengdong Hi-tech Zone and Chengbei Development Zone. The development direction of Jintan District is: the main direction is oriented towards the development of the South, positive development towards the east, moderate development towards the north, and control of the westward development. According to the map of the hot spots distribution of houses purchased in Jintan District from 2009 to 2014, the trend of home purchase in Jintan District moved to the east and southwest, which is generally consistent with the general urban planning of Jintan District.

In September 2012, the State Council agreed to repeal Wujiang City at the county level and establish Wujiang District in Suzhou City. The overall urban planning of Wujiang City from 2006 to 2020 points out that the spatial development direction of Wujiang District is mainly westward and southward, and the southern new city area is planned to be a comprehensive area with functions of modern cultural life, commercial finance, education and scientific research, and living. The lakeside area will be planned as a high quality residential community with supporting facilities. According to the distribution
Chapter 4  A Case Analysis of Saptio-temporal Behavior of House Buyers

map of the purchase of houses in Wujiang District from 2009 to 2014, it can be found that the buying trend in Wujiang District is generally moving in the south and north, which is roughly in line with the overall urban planning. And with the construction of Suzhou Metro Line 4, it is also obvious that the trend of buying houses is concentrated along the subway.

In this section, through the kernel density analysis method, we have obtained maps of the spatial distribution of house purchases in various regions. The purchase trend is basically in line with the overall urban planning direction of each region.

4.4.2 Analysis of House Purchase Behavior of Different Types of House Buyers

The kernel density analysis method is used to find out whether there are certain differences between different types of house buyers in the places of purchase, and use ArcGIS Pro software visualization and image rendering functions to display the results. The analysis of the kernel density of three different types of house buyers in three regions was conducted and visualized. The results are shown below.

As can be seen from the spatial heat distribution maps of the three types of house buyers in Dongtai City, the local house buyers’ hotspots are mainly concentrated in the west and south part of the old town, and the economic and technological development zones in the north have little purchase density. The “T”-shaped strips were formed in the hotspots for inner province house buyers and the highest density of buyers was in the southern part of the city, and the strips were distributed in the west and north part of the city. The purchase hotspots of the outer province house buyers are mainly the economic and technological development zones in the north and south of the city, and the purchase density of the houses in the west of the city is not high. From the above analysis, we can see that there are significant differences in the hotspots for buying houses of different types of house buyers in Dongtai City: the hot spots of local house buyers are
mainly concentrated in the traditional “old town”. The purchase hotspots of inner province house buyers are gradually transferred to the economic and technological development zones in the north. The outer province house buyers focus on the northern economic and technological development zones in the north and the southern part of the old town.

Figure 4-20 above shows the house purchase spatial distribution of different types of house buyers in Jintan District. It can be seen from the figure that the local house buyers in Jintan District mainly focus on the old town in the west and the economic development zones in the east. The hotspots for inner province house buyers are mainly in the old town in the west and the economic development zones in the east and the density in the west is higher than that in the east. The hotspots for inner province house buyers are mainly concentrated in the economic development zone in the east of the city and the density of the old town in the west of the city is low. From the above analysis, we can see that the hotspots for local house buyers in Jintan District are mainly concentrated in the central area of the city, and the density of houses purchased in the outlying areas of the city is relatively low. The purchase hotspots of house buyers in the inner and outer province are mainly scattered in the outlying areas of urban areas and economic and technological development zones.
According to Figure 4-21, the spatial heat distribution maps of different types of home buyers in Wujiang District can be seen that the hot spots of local house buyers in Wujiang District are mainly distributed in the integrated urban centers of the old town commercial center (top left corner hot spot), administrative service center (lower right corner hot spot), and east Taihu Lake CBD (lower left corner hot spot), in which the administrative center has the highest purchase density. The hotspots for inner province house buyers in Wujiang District are mainly located in the economic and technological development zones on the outskirts of the urban comprehensive center and in the north near Wuzhong District. The distribution of hotspots for outer province house buyers is roughly the same as the distribution of inner province house buyers, and is mainly distributed outside the urban integrated center. From the above analysis, we can see that the hot spots of local house buyers in Wujiang District are mainly concentrated in the urban comprehensive center. The hotspots for foreign house buyers are mainly located in the economic and technological development zones on the outskirts of the urban comprehensive center and in the northern part of the city near Wuzhong District.

By analyzing the different types of house buyers in the three regions, we can find the following rules: The spatial distribution of homes for house buyers is not significantly related to the level of economic development in the region, but is related to the census register location of the house buyers. Local house buyers tend to buy houses in the old town or in the main city, while foreign house buyers tend to buy houses in the development area or new town.
4.5 Source Analysis of Foreign Buyers

4.5.1 Time-space dynamic visualization of house buyers' census register location

This paper selects the real estate registration data of inner and outer province house buyers from 2009 to 2014 in three regions, and uses the ECharts visualization library to dynamically visualize the spatial data of the three regions. The spatio-temporal dynamic visualization results are displayed in the form of web pages.

The experiment selected the distribution data of the annual census register location of inner and outer province house buyers from 2009 to 2014 in three regions, and dynamically visualized them. The size of the bubble represents the number of home buyers in that census register location who purchased the house in the study area, and lines of convergence is where the purchase houses are. The article only shows screenshots of the dynamic migration maps of inner and outer province house buyers.

(a) Spatio-temporal Dynamic Visualization Map of Inner Province House Buyers in Dongtai City
From Figure 4-22, it can be seen that the outer province house buyers in Dongtai City from 2009 to 2014 are mainly from the neighboring provinces of Jiangsu Province and Sichuan Province. There are fewer house buyers in the northeast, northwest, south and west regions provinces which are farther away from Jiangsu Province and the annual distribution is roughly the same. From 2009 to 2014, the inner province house buyers in Dongtai City were mainly from Yangzhou City, Nantong City and Taizhou City. There were fewer house buyers from cities far from Dongtai City such as Changzhou city, Wuxi city, Suzhou city, Xuzhou city, Suqian city and Lianyungang city.
Figure 4-23 shows the spatio-temporal dynamic visualization of foreign house buyers in Jintan District, and interprets the visualization results. The outer province house buyers in Jintan District from 2009 to 2014 are mainly from Anhui, Hubei, Sichuan, Zhejiang, and Shanghai. Provinces and cities, and other provinces, have fewer buyers; inner province house buyers are mainly from three cities: Zhenjiang, Wuxi, and Nanjing; Huai’an, Yancheng, Yangzhou, and Nantong also have some buyers.
4.5.2 Spatial Autocorrelation Analysis of Foreign Buyers' Census Register Location

This paper uses the spatial autocorrelation (Global Moran's I) tool in ArcGIS Pro software to calculate the spatial autocorrelation index (Moran's I) of foreign house buyers of Dongtai, Jintan and Wujiang from 2009 to 2014 and performed a Z-value test (P-value ≤ 0.05, Z-value ≥ 1.96) on the results, and analyzed whether the census register location distribution of foreign house buyers has spatial aggregation. The calculation results are shown in the following table 4-4:

Table 4-4 Moran's I index over the years of different types of house buyers in each region

<table>
<thead>
<tr>
<th>Year</th>
<th>Inner province house buyers</th>
<th>Outer province house buyers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Moran’s I</td>
<td>z-score</td>
</tr>
<tr>
<td>Dongtai</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>0.159599</td>
<td>1.429399</td>
</tr>
</tbody>
</table>
## Chapter 4  A Case Analysis of Saptio-temporal Behavior of House Buyers

<table>
<thead>
<tr>
<th>Year</th>
<th>Moran’s I</th>
<th>z-score</th>
<th>p-value</th>
<th>Autocorrelation</th>
<th>Moran’s I</th>
<th>z-score</th>
<th>p-value</th>
<th>Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inner province house buyers</td>
<td></td>
<td></td>
<td></td>
<td>Outer province house buyers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>0.177778</td>
<td>1.527249</td>
<td>0.126699</td>
<td>0.300613</td>
<td>4.090484</td>
<td>0.000043</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2011</td>
<td>0.127021</td>
<td>1.240957</td>
<td>0.214622</td>
<td>0.264003</td>
<td>3.526244</td>
<td>0.000421</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>0.282773</td>
<td>2.160663</td>
<td>0.030721</td>
<td>0.20549</td>
<td>2.905504</td>
<td>0.003667</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.148608</td>
<td>1.421035</td>
<td>0.155306</td>
<td>0.303021</td>
<td>3.792803</td>
<td>0.000149</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>0.175273</td>
<td>1.599810</td>
<td>0.109641</td>
<td>0.29016</td>
<td>3.437933</td>
<td>0.000586</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>0.284073</td>
<td>2.102974</td>
<td>0.035468</td>
<td>0.231312</td>
<td>2.768949</td>
<td>0.005624</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>0.223331</td>
<td>1.736715</td>
<td>0.082438</td>
<td>0.236105</td>
<td>2.987785</td>
<td>0.00281</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2011</td>
<td>0.18089</td>
<td>1.504714</td>
<td>0.132398</td>
<td>0.236082</td>
<td>2.654989</td>
<td>0.007931</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td>0.300865</td>
<td>2.042943</td>
<td>0.041058</td>
<td>0.345958</td>
<td>3.957744</td>
<td>0.00076</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>0.241642</td>
<td>1.854458</td>
<td>0.063674</td>
<td>0.272252</td>
<td>3.194964</td>
<td>0.001398</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>0.130216</td>
<td>1.234784</td>
<td>0.216911</td>
<td>0.246893</td>
<td>2.87015</td>
<td>0.004103</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>Wujiang</td>
<td>-0.1474</td>
<td>-0.36382</td>
<td>0.715992</td>
<td>0.327717</td>
<td>3.858256</td>
<td>0.000114</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>-0.135802</td>
<td>-0.30324</td>
<td>0.761708</td>
<td>0.361985</td>
<td>4.147797</td>
<td>0.000034</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
<tr>
<td></td>
<td>2011</td>
<td>-0.11708</td>
<td>-0.18883</td>
<td>0.850223</td>
<td>0.303482</td>
<td>3.604793</td>
<td>0.000312</td>
<td>Strong spatial positive autocorrelation</td>
</tr>
</tbody>
</table>
From Table 4-4, we can see that the Global Moran's I index of inner province house buyers in Dongtai City from 2009 to 2014 is positive, but only the index in 2013 data have weak spatial autocorrelation, and there is no spatial autocorrelation in other years. The Global Moran's I index of inner province house buyers in Jintan District from 2009 to 2014 is positive but only the index in 2009 and 2012 have weak spatial autocorrelation and there is no spatial autocorrelation for the remaining years. The Global Moran's I index of inner province house buyers in Wujiang District is positive, but all years have no spatial correlation. From Table 4-4, we can see that for the census register location distribution of inner province house buyers, there is no obvious spatial autocorrelation.

From Table 4-4, it can be seen that the census register location distribution of outer province house buyers in all regions has strong spatial autocorrelation. To visualize the data, we can see from Figure 4-25 that the Global Moran's I index of Dongtai City dropped first and then rose from 2009 to 2014, and the Global Moran's I index was the lowest in 2012. From 2009 to 2014, the Global Moran's I index of Jintan District rose first and then declined and the Global Moran's I Index was the highest in 2012. From 2009 to 2014, the Global Moran's I index in Wujiang District rose first, then fell and then rose, but the fluctuation of the value of the Global Moran's I Index was not very large. In the three regions, the Moran’s I transformation has no obvious pattern over the years. From the above analysis, it can be seen that there is a clear spatial positive correlation among the outer province house buyers’ census register location, that is, the provinces with a large number of purchases are adjacent to the provinces with a large number of purchases, and the provinces with a small number of purchases are adjacent...
to the provinces with a small number of purchases which presents spatial aggregation characteristics. From the changes in the Global Moran's I Index over the years, it can be known that the indices of Dongtai City and Jintan District fluctuate greatly, while the Wujiang District has the smallest index fluctuation and the largest average value.

![Figure 4-25 Moran's I Index Change of Outer province house buyers in each region](image)

4.6 Chapter Summary

This chapter takes Dongtai City, Jintan District and Wujiang District of Jiangsu Province as experimental regions, and verifies the method of real estate registration data processing proposed in this paper and the data mining method of real estate registration.

First of all, this paper use the geo-encoding and data cleaning methods to process the real estate registration data in each of the three regions, and get the data that meet the excavation standards required for this experiment. Second, using the statistical analysis method to analyze the basic characteristics of house buyers, found the basic characteristics of house buyers. Thirdly, using association rule mining methods, we find that different types of house buyers’ preferences for buying houses provide references for real estate developers to implement business strategies. In addition, using kernel density analysis to analysis the spatio-temporal characteristics of the purchase of hot spots for all house buyers and different types of house buyers, and to verify the spatial distribution characteristics of all buyers and the overall urban planning. Finally, spatio-
Chapter 4  A Case Analysis of Saptio-temporal Behavior of House Buyers

temporal dynamic visualization method and spatial autocorrelation analysis are used to analyze the spatio-temporal characteristics of census register location of foreign house buyers, and analyzing China's population migration patterns from a new perspective. In short, this chapter uses the methods of real estate registration data processing and real estate registration data mining methods proposed in this paper, analyzes the rules of spatio-temporal behavior of house buyers, and verifies the validity of the ideas proposed in this paper.
Chapter 5  Conclusions and Prospects

This paper is in the context of the full implementation of the unified registration system for real estate in China, based on the analysis of spatio-temporal behavior analysis theory and data mining theory, analyzes the existing problems of real estate registration data and researches the processing method of real estate registration data and based on the characteristics of real estate registration data, studies the real estate registration data mining method. Through the study of this article, we can deeply understand the characteristics of real estate registration data, and through real estate registration data mining methods, study the spatio-temporal behavior of house buyers, provide reference for government and real estate developers to make decisions, and expand the application of real estate registration data. The main findings and results of this article are as follows:

(1) Processing method of real estate registration data. Through in-depth analysis of the quality problems of real estate registration data, based on data cleaning methods and online geo-encoding methods, form the real estate registration data processing methods. For address space information in real estate registration data, online geocoding technology is used to convert it into spatial coordinates. For problems existing in the real estate registration data, the data is checked and cleaned using spatial location characteristics and attribute value ranges, and excluding the invalid, missing, and incorrect real estate registration data and unrelated to the subject research data, finally get the real estate registration data that meet the data mining standards. At the same time, the real estate registration data of Dongtai City, Jintan District and Wujiang District are processed. The results show that the real-estate registration data mining standard data can be obtained by using the real estate registration data processing method proposed in this paper.

(2) Real-estate registration data mining method. According to the characteristics of real estate registration data, under the guidance of the theory of spatio-temporal behavior and data mining theory, the data mining method of real estate registration is studied. Through Apriori algorithm, kernel density analysis method, spatial autocorrelation algorithm and dynamic visualization method, a comprehensive analysis of the spatio-temporal behavior of house buyers is conducted. Using the real estate registration data
of Dongtai City, Jintan District and Wujiang District for method verification, and the results show that the real estate registration data mining method proposed in this paper can excavate the spatio-temporal behavior of different types of house buyers.

(3) Verification of real estate registration data processing and mining methods. Using real estate registration data processing and mining methods proposed in this paper, excavating the real estate registration data of Dongtai City, Jintan District and Wujiang District of Jiangsu Province from 2009 to 2014. The results show that the real estate data processing method can obtain real estate registration data that meet the mining standards. Through the Apriori algorithm, it is found that the purchase preferences of house buyers at different ages are different. Through the kernel density analysis method, it is found that the purchase tendency of house buyers is roughly the same as the development direction proposed in the urban master plan. Through visual analysis and spatial autocorrelation analysis, it is found that the census register location of the outer province house buyers exhibit strong spatial autocorrelation characteristics. The discovery of the laws of spatio-temporal behavior of house buyers also verifies the feasibility of the real estate registration data processing and mining method proposed in this paper.

The deficiencies and improvements in this paper mainly include the following aspects:

(1) Single data source. The research data in this paper is a single real estate registration data and does not combine real estate registration data with other data. In the future, we should try to combine real estate registration data with internet data to enrich the data sources.

(2) For the factors affecting the spatio-temporal behavior of house buyers, only the level of regional economic development has been considered and other factors have not been considered.

(3) Insufficient analysis of real estate registration data attributes. Although this article uses a variety of methods to conduct mining and analysis of real estate registration data, it mainly focuses on the attributes of the house buyer and the properties of the houses, but there are some interesting attributes such as the planning use and the properties of the building that are not analyzed. In the future, other attribute information should be fully considered to enrich the content of spatio-temporal behavior of house buyers.
References


References


References


References


[47] Zhang Xueying, Lü Guonian, Li Boqiu, et al. Rule-based Chinese Address...


[57] Shi Yun, Chen Zhong, Sun Bing. Network intrusion detection based on mean clustering analysis and multi-level core set aggregation algorithm[J]. Journal of
References


[69] Zhang Ying, Deng Liping. The Ministry of Finance, State Administration of Taxation, Ministry of Housing and Urban-Rural Development jointly issued the “Circular on Adjusting the Preferential Policies on Deed Tax Business Tax on Real