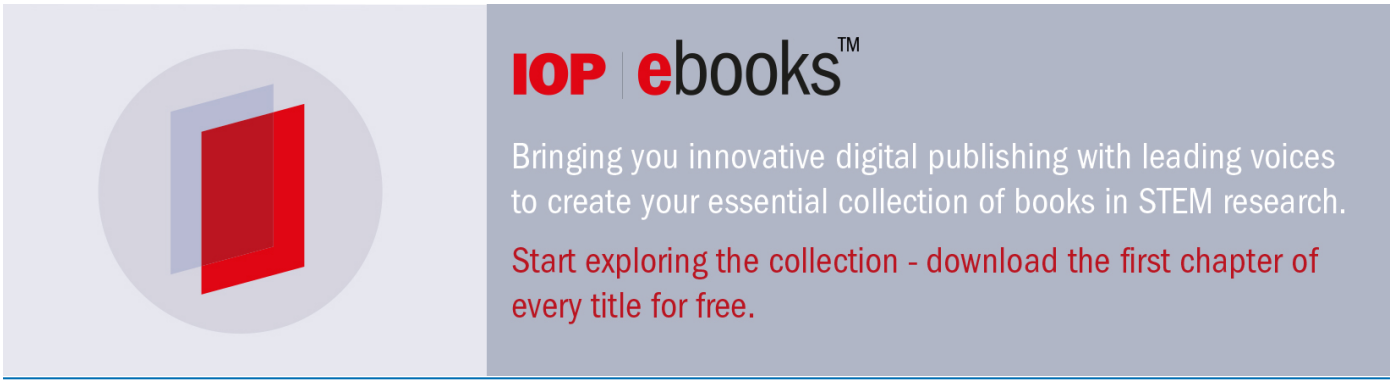
A Study on Unofficial Geographic Location Data Acquisition Technology Path under the Background of Big Data Era



IOP Conf. Series: Materials Science and Engineering **520** (2019) 012018 doi:10.1088/1757-899X/520/1/012018

A Study on Unofficial Geographic Location Data Acquisition Technology Path under the Background of Big Data Era

Ziran Zhang[1, 2], Yujing Tian[3, a], Ying Yu[3] and Shengxi Fan[4]

[1] Tongji University, College of Architecture and Urban Planning, 200092 Shanghai, P.R. China
[2] Shanghai University of Engineering Science, School of Art and Design, 201620 Shanghai, P.R. China
[3] Donghua University, Fashion and Design College, 200051 Shanghai, P.R. China
[4] Tongji University, College of Design and Innovation, 200092 Shanghai, P.R. China

[a] Corresponding author: tinatian5186@163.com

Abstract. The arrival of the era of big data has made people break through the bottleneck of various technologies and resources faced by traditional social sciences, but at the same time, new problems such as data acquisition, data value mining, data storage, circulation and exchange, data privacy and security have also brought new challenges to people. In order to explore a new methodology of geographic location data acquisition, this paper proposed the technical path of unofficial geographic location data acquisition by using web crawler tools and test the feasibility of it.

1. Research background: the arrival of big data era

By 2020, the global data volume is expected to exceed 44 trillion gigabytes [1]. Such a huge amount of data also makes big data gradually become the focus of attention of all walks of life. The arrival of the era of big data has made people break through the bottleneck of various technologies and resources faced by traditional social sciences, but at the same time, new problems such as data acquisition, data value mining, data storage, circulation and exchange, data privacy and security have also brought new challenges to people.

1.1 The increasing concern of big data in China's research field

In 1980, Alvin Toffler, a well-known futurist in the United States, regarded data as the colorful movement of the third wave in his book The Third Wave. Although people did not realize how colorful it was at that time, it was regarded as the first reference to "big data". In 2001, Gartner Group analyst Douglas Laney first defined big data in terms of its characteristics and stressed the 3Vs of big data, namely volume, variety and velocity [2]. Viktor mayer-schonberger, a professor at Oxford University who is known as the father of big data in the industry, published the book Big Data: A Revolution That Transforms How We Work, Live, and Think in 2012 [3], and published a Chinese translation of the book in 2013, which initiated the research wave of big data in China. As can be seen from Figure 1, the development history of big data and the time that people pay attention to it are not long.

According to the statistical results provided by Baidu index, a popular search engine in China, its search volume in China gradually increased from 2013 to the highest around April 2017, and then entered a stable fluctuation state. Table 1 is the result of the author's literature statistics according to

the year after searching the keyword "big data" in China's well-known academic database "CNKI". The results show that after big data was clearly defined in 2001, the concept of big data did not attract extensive attention from the academic circle until 2011. In 2012, the research on big data has gradually become a research hotspot of Chinese scholars around such topics as big data era, big data technology, big data environment, big data analysis, big data industry and data mining. Its level of attention began in 2012, and increased sharply from 2013 to 2014, 4.7 times the number of documents, and 4.4 times in the five years from 2013 to 2018. Obviously, since 2013, big data has become a topic of widespread concern and the research object of scholars in China.

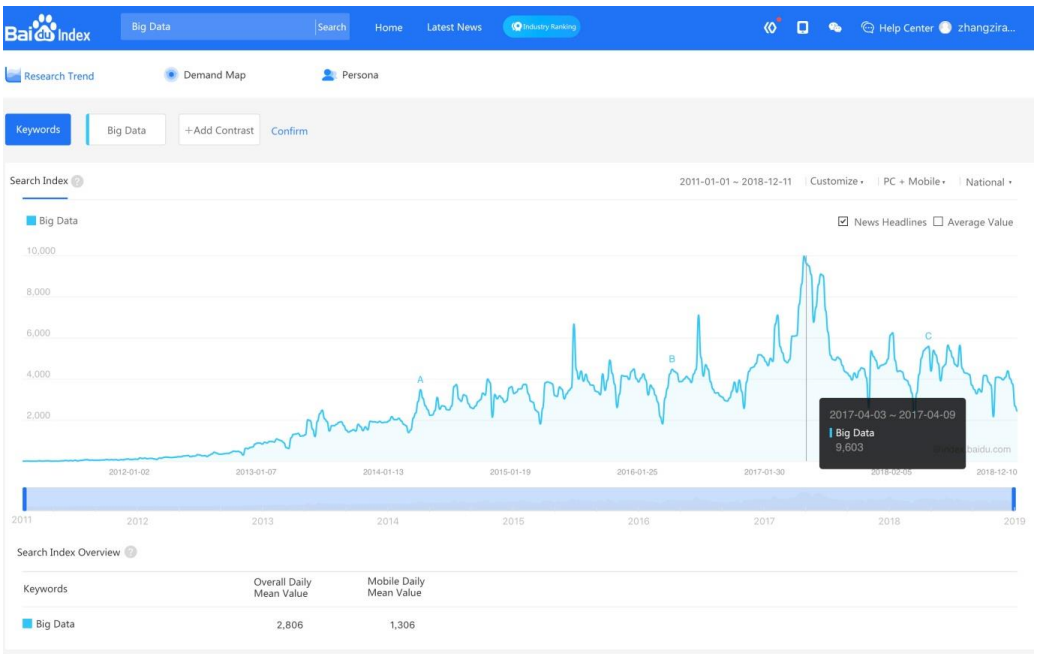


Figure 1. Hot trend of keyword "big data" search (Data source: Baidu index).

Table 1. Statistics on the number of Chinese papers on relevant keywords of "Big Data" on CNKI.

YEAR	BIG DATA	BIG DATA TIME	BIG DATA TECHNOLOGY	BIG DATA EVIRONMENT	BIG DATA ANALYSIS	BIG DATA INDUSTRY
2009	55	0	0	0	0	0
2010	69	1	0	2	0	0
2011	127	8	0	1	8	0
2012	994	214	47	24	55	20
2013	4739	1245	261	173	239	92
2014	10474	2539	600	590	414	312
2015	16708	3415	1020	1224	661	521
2016	22511	4029	1239	1925	876	634
2017	21070	4257	1735	2333	1030	666
2018	25774	4646	1909	2246	1014	468

1.2 Research status and development trend of location big data in China

As the main technology of mobile location data acquisition, LBS has reached a peak between 2012 and 2014 in China with 314 relevant Chinese academic papers in 2014. After 2013, it began to decline gradually, and it was sharply reduced to 45.2% of its peak from 2016 to 2018. This may be due to the gradual shift of people's attention from previous technology research and development to other directions because of the gradual maturity of LBS technology. At the same time, the research on location data has begun. It is not difficult to see from Figure 2 that compared with the two keywords "location data" and "geographical location information", the number of documents containing "big location data" and "mobile signaling data" in the title is relatively small. It started around 2014 and it

is now in a slow climbing stage. However, compared with the three related documents in 2013, "big location data" has produced about 10-fold growth in quantity in 2018, and its development momentum cannot be underestimated.

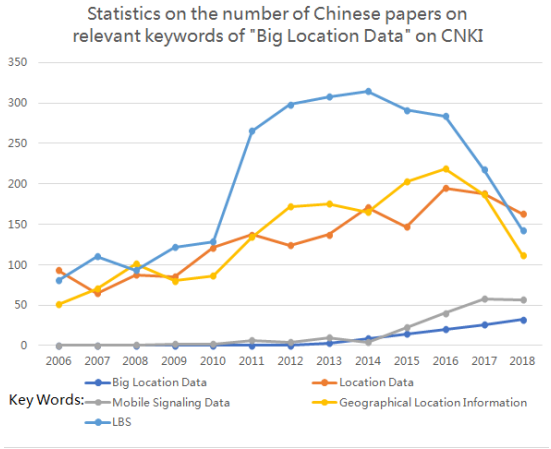


Figure 2. Statistics on the number of Chinese papers on relevant keywords of "Big Data" on CNKI.

1.3 Relations between geographical location data and location big data

Location data, as the key data to show people's moving trajectory, can bring great research value and business opportunities for transportation, tourism, population, urban development and other research directions. The relationship between geographic location data and location big data is a deeper rational discussion of urban population flow, people's spatial behavior, spatial structure and preferences and other issues and phenomena. For example, Tencent used Tencent big data and park location data to conduct a series of analyses and research on Shanghai park tourism preference during the National Day holiday in 2018, which provided strong support for young people's vacation tourism preference research. Liang Lin and other scholars established a network model for the characteristics of inter-city population flow in Beijing, Tianjin and Hebei based on Tencent location data [4]; Bai Yan, Zhu Anran and others explored the spatial distribution characteristics of night life in Hefei city based on the location data of Weibo[5]. Thus, it can be seen that selecting appropriate geographic location data according to research needs and combining location big data for relevant analysis and research have certain practical value and research significance for urban spatial structure planning and development.

2. Acquisition methods of geographic location data

If you want to complete a data project, data preparation often takes up 70% or more of the workload.[1] It is also known as "data cleaning" in the industry, *which is one of the 19 research topics on data quality* [6]. Usually, *instance-level data quality problems include similar duplicate records* [7], incomplete records, logical errors, abnormal data, etc. General speaking, *data cleaning is usually an expensive process that requires significant resources, considerable effort, and human interaction* [8]. Therefore, how to obtain relatively "clean" and valuable geographic location data to reduce the cost of cleaning data depends on the choice of data sources and data acquisition technology path.

2.1 Data acquisition methods in traditional social sciences

In the era of underdeveloped information technology, researchers of social sciences usually use computer-aided manual statistics to collect data. Geographical location and its related attribute information are very important research objects in the fields of landscape design, environmental art design and geography, sociology, tourism, traffic information technology, etc. Due to the immaturity of electronic map, remote sensing satellite technology and GIS algorithm, manual statistics or GPS positioning assisted manual statistics combined with scientific sampling method were used to obtain

the specific location of a certain type of building and related attribute information for scientific research.

2.2 Platform paid data purchase

In recent years, due to the rapid development of big data technology in China, many commercial organizations have begun to gradually promote a series of data purchase services. For example, "Tencent Location Big Data" can directly sort out the location data of Tencent users and related attribute data according to the needs of buyers, and then trade or conduct more in-depth business cooperation. If you search Baidu for the keyword "Community Location Data Download", you can find many suppliers of data purchase, including tdata.cn, Shenjian.io, JDCloud, CSDN, etc. Through sample analysis and survey, the average purchase price of community location data in Shanghai is 2,000 Yuan, and the data purchased cannot guarantee its validity, credibility and the last update time, so the data quality is difficult to control.

2.3 Network information data source and crawler technology

The rapid rise of China's Internet economy in the past decade has promoted the maturity and perfection of network information technology in the domestic market of traditional industries. As third-party online platforms such as Dianping, Meituan, Taobao and Jingdong have changed from "being generally accepted by consumers" to "changing consumers' living habits", more and more merchants are forced to choose these third-party online platforms as the channels of customer flow. Therefore, the unshakable commercial status of these platforms is created, and all the information and massive data of consumers and businesses are centralized. Some of these data are published on such platforms' websites to facilitate consumers to make commodity selection decisions as very valuable data resources.

Crawler technology, in simple words, is a technology that can effectively extract web page data according to the logic specified by the program design. With the crawler technology and relatively centralized data resources, China's Internet environment has provided researchers with very favorable research conditions. Nowadays, this way of extracting and analyzing network data by using crawlers has been widely applied in the fields of network consumption preference, hot issues public opinion research and logistics industry.

3. Exploration of unofficial geographical location data acquisition technology path

3.1 Principles for selecting data sources

In order to obtain complete and reliable data, the source of data is particularly important. Usually, according to the research needs, 2-3 websites containing the information needed by the research are selected to extract the crawler data. The website needs to be widely accepted and used, and the amount of data must be nearly complete, or it can be selected and decided according to the market share or the number of users of the website. This is the big data mentioned above. After collecting and cleaning the data in the selected network platform, they can complement and merge with each other, and compare the data results with the data on other websites to test their reliability and validity.

3.2 Technological feasibility exploration: a case study of related attribute information collection of community in Shanghai

3.2.1 Selection and comparison of geographic location data sources in communities. In China, real estate agencies, as the private enterprises that know the situation of the surrounding communities best, have a large amount of specific information about the communities around the physical store outlets, including the address, construction date, number of residents, housing structure and selling price, etc. Therefore, if the official websites of real estate agencies with high coverage of online stores and large number of users are chosen as data sources to collect location and related information of residential

areas, their data integrity and reliability are better. Major housing intermediary brands have their own free official websites in China, and their information content is updated in real time. Therefore, in this case, keywords "Shanghai Second-hand Housing" were searched on Baidu, and the data of the top 8 real estate agencies' official websites in the first 3 pages of the search results, except the advertising links, were taken as the data source for comparison. The statistical results are shown in Table 2 below. The top three with the largest amount of data were selected as data collection objects and the data information of all the search results in each administrative region under the "community" section was collected. Among them, the website information with the largest amount of data is regarded as the website with the most complete information, and it can be used as the main data source for data acquisition. In the statistical process, missing data information or information without search results was replaced by "0". In addition, Shanghai merged Nanhui District into Pudong New District in 2009. In 2011, "Luwan District" and "Huangpu District" were merged, and in 2018, "Zhabei District" and "Jing'an District" were merged. Therefore, relevant information was also merged in this statistics.

In Table 2, according to the search results, 58.com collected the most information items in Baoshan District, Songjiang District, Hongkou District, Qingpu District and Jinshan District, and they were highlighted in Table 2. Similarly, Fang.com searched for the largest amount of information in six administrative districts, including Pudong New District, Minhang District and Putuo District, and they were also highlighted. However, the highlighted data in the column of Fangdd.com were also the four regions with the largest number of search results compared with the corresponding administrative divisions of other websites. Therefore, this study selected 58.com, Fang.com, Fangdd.com as data sources to collect information on the geographical location of all residential districts in Shanghai, and added 660 district information collected by the "Fengxian District" on Lianjia to the information collected by Fengxian District on 58.com, in order to ensure the integrity of the data as far as possible.

Table 2. Community information statistics of Baidu's top 8 real estate intermediary websites in districts.

District	58.com	Fang.com	Lianjia	Anjuke	qfang.com	5i5j.com	CENTALINE PROPERTY	Fangdd.com
Pudong	2214	3869	3364	3285	1013	3201	2662	3500
Minhang	1655	1739	1596	1480	113	1430	796	1581
Xuhui	1675	1743	1592	1509	6	1455	917	1817
Putuo	1016	1047	985	936	429	960	415	981
Baoshan	1153	1034	1056	1052	415	793	823	1134
Changning	1213	1388	1252	1074	62	1150	745	1236
Yangpu	1290	1205	1342	1174	22	1047	878	1425
Songjiang	1128	898	963	998	0	696	212	995
Hongkou	1315	1150	1185	1076	14	942	849	1315
Jiading	1091	922	1074	999	35	768	550	1141
Huangpu	1397	1227	1321	1143	16	1054	659	1512
Jingan	1777	1883	1653	1547	435	1490	806	1809
Qingpu	806	566	632	752	0	485	91	707
Fengxian	640	477	660	577	0	0	145	446
Jinshan	426	358	330	379	0	0	133	307
Chong Ming	214	222	170	203	0	0	2	145
SUM	19010	19728	19175	18184	2560	15471	10683	20051

3.2.2 The use of web crawler tools and the setting of grabbing technical routes. In this paper, the working principle of crawler software has been introduced in 2.3. This study is to select one of the mature and free crawler software GooSeeker to download website data in batches. The main grab technology roadmap is shown in Figure 3. The main key steps are as follows: 1. Open the designated website; 2. Set cyclic select the designated administrative region; 3. Set cyclic page turning; 4. Grab the attribute information such as "community name", "address" and "building age" in the page (the

specific attribute information can be set according to the research purpose). Among them, 51.com downloaded17781data, Fang.com and Fangdd downloaded 11861and 18502data respectively, and the statistical results are shown in the column of "actual downloads" in Table 3.

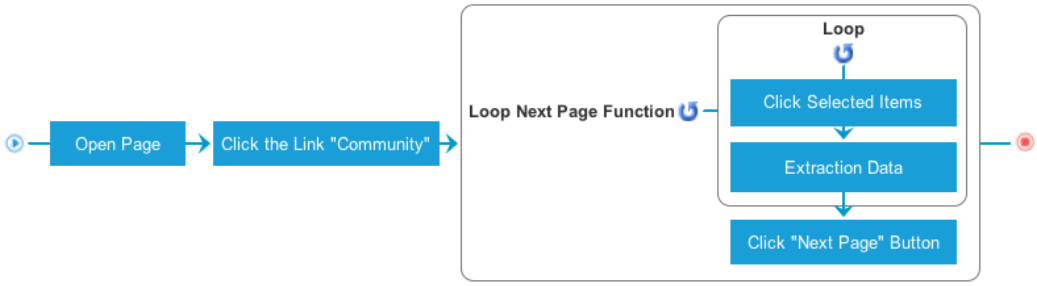


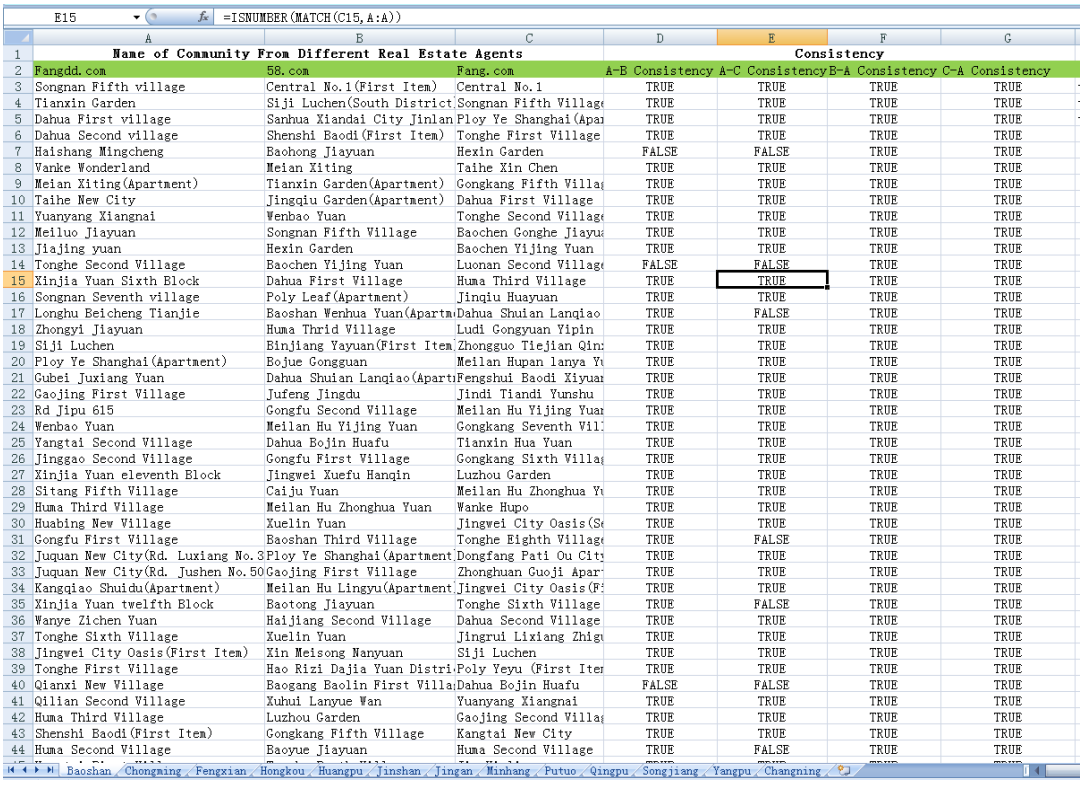
Figure 3. Logical schematic diagram of GooSeeker's main grabbing technology route.

Table 3. Statistics of data downloads from various real estate websites.

District	58.com			Fang.com			Fangdd.com		
	Planned Download	Actual Download	Data Collection Efficiency	Planned Download	Actual Download	Data Collection Efficiency	Planned Download	Actual Download	Data Collection Efficiency
Baoshan	1153	1090	94.54%	1034	707	68.38%	1134	1134	100.00%
Chongming	214	208	97.20%	222	187	84.23%	145	145	100.00%
Fengxian	640	640	100.00%	477	333	69.81%	446	444	99.55%
Hongkou	1315	1315	100.00%	1150	773	67.22%	1315	1309	99.54%
Huangpu	1397	567	40.59%	1227	667	54.36%	1512	1506	99.60%
Jiading	1091	1061	97.25%	922	528	57.27%	1141	1137	99.65%
Jinshan	426	426	100.00%	358	250	69.83%	307	306	99.67%
Jingan	1777	1750	98.48%	1883	1136	60.33%	1809	1799	99.45%
Minhang	1655	1570	94.86%	1739	1064	61.18%	1581	1574	99.56%
Pudong	2214	2104	95.03%	3869	2000	51.69%	3500	1997	57.06%
Putuo	1016	980	96.46%	1047	644	61.51%	981	980	99.90%
Qingpu	806	802	99.50%	566	295	52.12%	707	705	99.72%
Songjiang	2206	1128	51.13%	898	411	45.77%	995	988	99.30%
Xuhui	1675	1674	99.94%	1743	1094	62.77%	1817	1823	100.33%
Yangpu	1290	1253	97.13%	1205	893	74.11%	1425	1420	99.65%
Changning	1213	1213	100.00%	1388	879	63.33%	1236	1235	99.92%
SUM	20088	17781	88.52%	19728	11861	60.12%	20051	18502	92.27%
AVERAGE	1255.5	1111.3125	91.38%	1233	741.3125	62.74%	1253.1875	1156.375	97.06%

3.2.3 Data cleaning and verification. When GooSeeker grabs data, it stores each page data as XML format file separately. Therefore, the number of XML format files in the data storage folder is the number of pages successfully grabbed. Statistical comparison shows that under the premise of accurate field design, the page crawl rate is 91.38%, 62.74% and 97.06%, respectively. As shown in Table 3, Page crawl rate is divided actual crawl data to planned crawl data. According to the commonly used data cleaning methods, the following steps of data cleaning are carried out by software: 1) convert XLM files into CSV data format files; 2) delete duplicate and invalid data, including invalid data items of "shops, office buildings, etc;" 3) convert address information in data into longitude and latitude data through Baidu Map API access; 4) count the amount of data downloaded by various administrative districts and compare it with the planned amount of data downloaded to ensure that there is no missing data.

3.2.4 Verification of data. The verification of the geographical location information data of the community can be transformed into two questions, namely, whether the "community name" in the data collected by each website is uniform and whether the address information of the community is correct. If the consistency is high and the correct rate of address information is high, the data acquisition method can be regarded as effective. Otherwise, it is not.



E15 =ISNUMBER(MATCH(C15,A:A))

	A	B	C	D	E	F	G
1	Name of Community From Different Real Estate Agents			Consistency			
2	Fangdd.com	58.com	Fang.com	A-B Consistency	A-C Consistency	B-A Consistency	C-A Consistency
3	Songnan Fifth village	Central No.1(First Item)	Central No.1	TRUE	TRUE	TRUE	TRUE
4	Tianxin Garden	Siji Luchen(South District	Songnan Fifth Village	TRUE	TRUE	TRUE	TRUE
5	Dahua First village	Sanhua Xiandai City Jinlan	Ploy Ye Shanghai(Apar	TRUE	TRUE	TRUE	TRUE
6	Dahua Second village	Shenshi Baodi(First Item)	Tonghe First Village	TRUE	TRUE	TRUE	TRUE
7	Haishang Mingcheng	Baohong Jiayuan	Hexin Garden	FALSE	FALSE	TRUE	TRUE
8	Vanke Wonderland	Meian Xiting	Taihe Xin Chen	TRUE	TRUE	TRUE	TRUE
9	Meian Xiting(Apartment)	Tianxin Garden(Apartment)	Gongkang Fifth Villa	TRUE	TRUE	TRUE	TRUE
10	Taihe New City	Jingqiu Garden(Apartment)	Dahua First Village	TRUE	TRUE	TRUE	TRUE
11	Yuanyang Xiangnai	Wenbao Yuan	Tonghe Second Village	TRUE	TRUE	TRUE	TRUE
12	Meiluo Jiayuan	Songnan Fifth Village	Baochen Gonghe Jiayua	TRUE	TRUE	TRUE	TRUE
13	Jiajing yuan	Hexin Garden	Baochen Yijing Yuan	TRUE	TRUE	TRUE	TRUE
14	Tonghe Second Village	Baochen Yijing Yuan	Luonan Second Village	FALSE	FALSE	TRUE	TRUE
15	Xinjia Yuan Sixth Block	Dahua First Village	Huma Third Village	TRUE	TRUE	TRUE	TRUE
16	Songnan Seventh village	Poly Leaf(Apartment)	Jinqiu Huayuan	TRUE	TRUE	TRUE	TRUE
17	Longhu Beicheng Tianjie	Baoshan Wenhua Yuan(Apartm	Dahua Shuian Lanqiao	TRUE	FALSE	TRUE	TRUE
18	Zhongyi Jiayuan	Huma Thrid Village	Ludi Gongyuan Yipin	TRUE	TRUE	TRUE	TRUE
19	Siji Luchen	Binjiang Yayuan(First Item	Zhongguo Tiejian Qin	TRUE	TRUE	TRUE	TRUE
20	Ploy Ye Shanghai(Apartment)	Bojue Gongguan	Meilan Hupan lanya Y	TRUE	TRUE	TRUE	TRUE
21	Gubei Juxiang Yuan	Dahua Shuian Lanqiao(Apart	Fengshui Baodi Xiyuar	TRUE	TRUE	TRUE	TRUE
22	Gaojing First Village	Jufeng Jingdu	Jindi Tiandi Yunshu	TRUE	TRUE	TRUE	TRUE
23	Rd Jipu 615	Gongfu Second Village	Meilan Hu Yijing Yuar	TRUE	TRUE	TRUE	TRUE
24	Wenbao Yuan	Meilan Hu Yijing Yuan	Gongkang Seventh Vil	TRUE	TRUE	TRUE	TRUE
25	Yangtai Second Village	Dahua Bojin Huafu	Tianxin Hua Yuan	TRUE	TRUE	TRUE	TRUE
26	Jinggao Second Village	Gongfu First Village	Gongkang Sixth Villa	TRUE	TRUE	TRUE	TRUE
27	Xinjia Yuan eleventh Block	Jingwei Xuefu Hanqin	Luzhou Garden	TRUE	TRUE	TRUE	TRUE
28	Sitang Fifth Village	Caiju Yuan	Meilan Hu Zhonghua Y	TRUE	TRUE	TRUE	TRUE
29	Huma Third Village	Meilan Hu Zhonghua Yuan	Wanke Hupo	TRUE	TRUE	TRUE	TRUE
30	Huabing New Village	Xuelin Yuan	Jingwei City Oasis(S	TRUE	TRUE	TRUE	TRUE
31	Gongfu First Village	Baoshan Third Village	Tonghe Eighth Village	TRUE	FALSE	TRUE	TRUE
32	Juquan New City(Rd. Luxiang No.3	Ploy Ye Shanghai(Apartment	Dongfang Pati Ou City	TRUE	TRUE	TRUE	TRUE
33	Juquan New City(Rd. Jushen No.50	Gaojing First Village	Zhonghuan Guoji Apar	TRUE	TRUE	TRUE	TRUE
34	Kangqiao Shuidu(Apartment)	Meilan Hu Lingyu(Apartment	Jingwei City Oasis(F	TRUE	TRUE	TRUE	TRUE
35	Xinjia Yuan twelfth Block	Baotong Jiayuan	Tonghe Sixth Village	TRUE	FALSE	TRUE	TRUE
36	Wanye Zichen Yuan	Haijiang Second Village	Dahua Second Village	TRUE	TRUE	TRUE	TRUE
37	Tonghe Sixth Village	Xuelin Yuan	Jingrui Lixiang Zhig	TRUE	TRUE	TRUE	TRUE
38	Jingwei City Oasis(First Item)	Xin Meisong Nanyuan	Siji Luchen	TRUE	TRUE	TRUE	TRUE
39	Tonghe First Village	Hao Rizi Dajia Yuan Distri	Poly Yeyu (First Iter	TRUE	TRUE	TRUE	TRUE
40	Qianxi New Village	Baogang Baolin First Villa	Dahua Bojin Huafu	FALSE	FALSE	TRUE	TRUE
41	Qilian Second Village	Xuhui Lanyue Wan	Yuanyang Xiangnai	TRUE	TRUE	TRUE	TRUE
42	Huma Third Village	Luzhou Garden	Gaojing Second Villa	TRUE	TRUE	TRUE	TRUE
43	Shenshi Baodi(First Item)	Gongkang Fifth Village	Kangtai New City	TRUE	TRUE	TRUE	TRUE
44	Huma Second Village	Baoyue Jiayuan	Huma Second Village	TRUE	FALSE	TRUE	TRUE

Baoshan Chongming Fengxian Hongkou Huangpu Jinshan Jingan Minhang Putuo Qingpu Songjiang Yangpu Changning

Figure 4. Comparisons of community data consistency (Baoshan District).

Fig. 4 is an illustration of the sample data used for inspection after cleaning, sorting and merging. Column A is the data column to be tested, and the data source of column A is formed by combining the data with the largest amount of downloaded data in the corresponding administrative region of each website, namely the data marked as high-light color in Table 3 (For example, the community location information data of Baoshan District, Chongming District and other administrative regions downloaded from 58.com; Data of Huangpu District, Jiading District, etc. downloaded from Fangdd). The community name is extracted and merged into the column A data here, and the consistency test is conducted for this column data and other data according to the administrative region. If column A of the data to be tested within the designated administrative region can largely cover the data downloaded from the other two websites (data from the control group), the information integrity and credibility of column A of the data to be tested are relatively high.

Table 4 is the consistency test results of community names in the data of each website. The consistency is divide the total number of communities with the same community name in column B or C as in column A to the total number of communities in column B or C. From the results, it is not difficult to find that the consistency between the tested data columns (column A) and the experimental group data columns (B, C columns) is high. The average consistency is between 95.36% and 96.53%, and the highest consistency is 100% in Baoshan, Chongming and other districts. That is to say, column A contains all the data in column B and column C. In this test, the lowest consistency occurs in Jiading and Jinshan District, and column A contains 84.83% and 85.98% data in column B and column C respectively.

Table 4. Consistency test results of community names in the data of websites.

District	B-A Consistency	C-A Consistency	A-B Consistency	A-C Consistency
Baoshan	99.91%	100.00%	94.11%	91.68%
Chongming	100.00%	100.00%	86.45%	97.20%
Fengxian	100.00%	100.00%	89.84%	95.00%
Hongkou	100.00%	100.00%	88.14%	83.73%
Huangpu	88.71%	87.56%	100.00%	89.31%
Jiading	84.83%	85.98%	100.00%	96.14%
Jinshan	84.83%	85.98%	100.00%	96.14%
Jingan	86.82%	86.09%	100.00%	98.45%
Minhang	100.00%	100.00%	96.98%	97.28%
Pudong	100.00%	100.00%	95.53%	85.36%
Putuo	100.00%	100.00%	99.70%	92.42%
Qingpu	100.00%	100.00%	95.29%	95.29%
Songjiang	100.00%	100.00%	98.37%	99.27%
Xuhui	99.64%	99.73%	100.00%	98.41%
Yangpu	91.01%	89.81%	100.00%	95.08%
Changning	89.95%	89.42%	100.00%	93.68%
Average	95.36%	95.29%	96.53%	94.03%
Variance	0.003738991	0.003758881	0.002019555	0.001944446

4. Conclusions

4.1 Technical path of unofficial geographic location data acquisition

To sum up, with regard to the collection of unofficial geographic location data, it is necessary to classify and download the data on the basis of selecting the right data source, and select multiple data download channels at the same time to use crawler technology to download and form data sample 1, data sample 2 and data sample 3…, combine the samples with the largest amount of data in the data sample to form the data to be tested, take the remaining sample data as the data of control group, use the API import function of the map software to convert the address information in the complete data that passes the consistency test into the longitude and latitude information for researchers to use. The flow chart can be summarized as Figure 5.

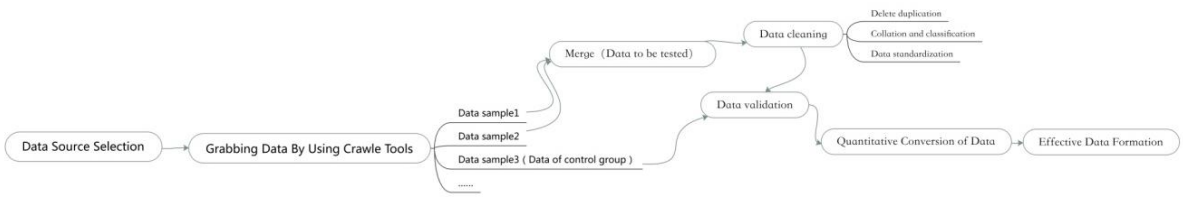


Figure 5. Technical path of unofficial geographic location data acquisition.

4.2 Comparison of several geographical location data acquisition methods

Table 5 shows the differences of economic cost, time cost, technical difficulty and feasibility of several different geographic location data acquisition methods, which are expressed in three grades: high, medium and low. Among them, the traditional manual acquisition method has the highest cost, but because it has no high requirements for technology, it is still used today. Compared with time-saving data purchase service, it also has low technical requirements. However, because data service is a commodity publicly sold by network technology enterprises nowadays, its economic cost

is relatively high, it is impossible to grasp the reliability and integrity of data by oneself. Finally, although the method of collecting and processing unofficial data sources by using crawler technology has certain requirements for technology and requires researchers to learn crawler software by themselves, most crawler software on the market have preset templates for users to choose, so they can quickly learn to operate without programming. Therefore, this method provides a new technical path for researchers who are short of money and time for research.

Table 5. Comparison of several different geographic location data acquisition methods.

	Economic Cost	**Time Cost**	**Technical Difficulty**	**Feasibility**
Data acquisition methods in traditional social sciences	High	High	Low	Medium
Platform paid data purchase	High	Low	Low	Unknow
Network information data source grab with crawler technology	Low	Low	Medium	High

Acknowledgments
Project Research Team of Intelligence Sustainable Package Design Support by Shanghai Summit Discipline in Design; Foundation item: Chenguang Plan (13CG74) sponsor; Shanghai Style Fashion Design &Value Creation Collaborative Innovation Center Support by Shanghai Summit Discipline in Design (DB18212)

References
[1] KPMG China big data team 2018 *From Data to Insights* (Beijing: Tsinghua University Press) p 01
[2] Jing L 2018 The origin and action of big data, *Economics Think Tank* **04** 26-35
[3] Mayer-Schönberger V, Cukier K 2013 *Big Data: A Revolution That Transforms How We Work, Live, and Think* (Boston: Houghton Mifflin Harcourt)
[4] Liang L, Zhao Y B and Liu B 2019 *Northwest Population Journal* **40** 20-28
[5] Bai Y, Zhu A R, Yang Y Y, et al 2018 Spatial Study of Nightlife in Hefei Urban Area Based on the Place Data in Micro-blog *Architecture and Culture* **175** 185-186
[6] Madnick S E, Wang R Y 2009 Overview and Framework for Data and Information Quality Research *ACM Journal of Data and Information Quality* **1** 1-22
[7] Cao J J, Diao X C, Tan M C, et al, 2010 *the 15th International Conference on Inforamtion Quality* (Arkansas USA: UALR)
[8] Cao J J, Diao X C 2017 *Introduction to Data Quality* (Beijing: National Defense Industry Press) p 42