主题栏目：中国区域经济开放与发展问题研究 DOI：10.3785/j.issn.1008-942X.2014.07.171

中国城市规模分布实证研究

——基于微观空间数据和城市聚类算法的探索

劳 昕[1,2] 沈体雁[1] 孔赟珑[3]

（1.北京大学 政府管理学院，北京 100871；2.北京大学 科学与工程计算中心，北京 100871；3.中国科学院 遥感与数字地球研究所，北京 100101）

［摘 要］齐普夫定律反映了城市规模与其位序之间简单而准确的关系，也是研究判别城市集聚和城市体系合理性的重要原则。关于齐普夫定律中城市的定义一直颇有争议，由于传统的空间研究尺度过于宏观，不能反映出真实的城市规模，学术界逐渐开始将眼光转向微观空间尺度，突破传统的行政区划界限，研究真正起到城市功能的微观城市组团。引入国外研究用于划分城市界限的新方法——城市聚类算法，对中国微观空间数据进行处理，以得到的功能性城市组团作为研究对象，根据齐普夫定律对中国城市规模分布进行分析，结果表明中国城市规模分布基本上服从齐普夫定律。此外，将基于城市聚类算法的城市规模分布研究结果与中国地级、区县级和乡镇街道级空间层面的研究结果进行比较，证实了城市聚类算法是研究城市规模分布的一种较好的新方法，它成功架设了宏观层面和微观层面研究之间的桥梁。

［关键词］齐普夫定律；中国城市规模分布；城市聚类算法；微观空间尺度

The Empirical Research on China's City Size Distribution: An Exploration Based on the Micro Spatial Data and City Clustering Algorithm

Lao Xin[1,2] Shen Tiyan[1] Kong Yunlong[3]

(1. *School of Government, Peking University, Beijing* 100871, *China*; 2. *Center for Computional Science & Engineering, Peking University, Beijing* 100871, *China*; 3. *Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract: Zipf's Law is an important principle to determine city agglomeration and urban system rationality, which reflects the simple and accurate relationship between city size and its rank. Since the definition of cities in Zipf's Law has roused much controversy due to the too macro spatial scale that cannot exactly reflect the actual city size, scholars have moved on to the functional urban areas (city clusters) at the micro level, which breaks down the traditional

［收稿日期］2014-07-17 ［本刊网址·在线杂志］http://www.journals.zju.edu.cn/soc

［在线优先出版日期］2014-12-23

［基金项目］国家自然科学基金项目(41071076)

［作者简介］1.劳昕，女，北京大学政府管理学院、科学与工程计算中心博士研究生，主要从事区域分析与规划、区域经济学研究；2.沈体雁，男，北京大学政府管理学院教授，博士生导师，主要从事城市与区域规划管理、GIS与空间模拟分析研究；3.孔赟珑，男，中国科学院遥感与数字地球研究所博士研究生，主要从事信号与信息处理研究。

administrative boundaries. To solve this problem, this article introduces a new method of defining city boundaries from abroad—City Clustering Algorithm to analyze China's city size distribution, that is, a "city" is defined as a maximally connected cluster of contiguous populated sites within a prescribed distance l and above a population density cutoff threshold D^*. These established city clusters are used to analyze China's city size distribution, with the sum of population of all populated sites within each city cluster as its population. The main findings of this article are shown as follows: First, China's city size distribution basically obeys Zipf's Law, indicating that the urban system based on employed population has a rank-size distribution, namely, a relatively balanced development of cities with different ranks. Second, by comparing the results of the city size distribution based on City Clustering Algorithm and the results at different scales of prefecture-level cities, counties, townships and streets, it has been proved that City Clustering Algorithm is an effective method to study the city size distribution, which breaks down the traditional administrative boundaries and makes up for deficiencies at both the macro (underestimating the number of small city clusters with a small sample of cities) and the micro (overestimating the number of small city clusters with data errors) level. Third, this method can reflect actual city sizes, making the results more scientific and reasonable; the effectiveness and robustness of this method have been verified by the related analysis of US, Great Britain and China (this article). Last but not least, with regard to China's current new era of the urban and rural dual structure in transition and the abolishment of the boundaries between urban and rural areas, it is of great significance to define the urban functional areas (or city clusters) according to certain rules (just like the combination of the distance threshold and the population density threshold in this paper) and to set up China's cities based on the urban functional areas. However, there are two main deficiencies in this study: One is the lack of data accuracy at the micro level, which is obtained by matching the employment data of the second (2008) economic census data and the spatial map at the level of townships and streets in 2000 after data correction; the other one is lack of population data at the micro level, making it impossible to be compared with the result of employment data at the micro level.

Key words: Zipf's Law; China's city size distribution; City Clustering Algorithm; the micro level

一、引　言

在一个区域或国家，各城市因所处的内外条件不同，会形成不同的功能分工，同时也形成不同的城市规模。1949 年，Zipf [1]提出了齐普夫定律(Zipf's Law)，该定律准确地揭示了城市规模和城市等级之间的数量关系，认为城市规模分布满足以下公式：

$$P(\text{size} > S) = \frac{a}{S^{\zeta}} \tag{1}$$

其中 S 为城市规模，P 为规模大于 S 的城市分布概率，a 为常数，且幂律指数 $\zeta=1$，表明规模为 S_i 的城市的位序 i 与位序大于 i 的城市概率是成比例的。如果 $0<\zeta<1$，表示城市规模分布比齐普夫定律所描述的更为均匀，即位次较低的中小城市比较发达，位次较高的大城市不很突出；如果 $\zeta>1$，表示大城市的规模比齐普夫定律描述的更大，即城市规模的分布更为分散。

自齐普夫定律提出后,国外学者围绕该定律在城市规模分布方面做了大量的理论和实证研究。理论方面主要研究齐普夫定律的理论基础[2-6];实证研究方面关注的是:齐普夫定律是否具有普适性[7-11],齐普夫定律的最优表达形式为帕累托分布还是对数正态分布[12-17],以及齐普夫定律的空间尺度效应——不同城市空间单元的规模分布是否服从齐普夫定律①[12,18-22]。

其中,齐普夫定律在不同空间尺度表现出的特征是存在差异的,各级空间单元是否都能用齐普夫定律来解释,该定律最适用于哪类空间单元,齐普夫定律中城市的内涵及城市边界如何划定(人口阈值超过多少才能定义为城市),都是值得深入探讨的问题[7,23-25]。国外大多数关于齐普夫定律的研究都以人口普查数据来划分城市,一般将大都会区或人口普查区作为城市范围;国内研究则多以城市为基本研究单元,一般采用城市户籍人口、非农业人口或市辖区人口②数据来研究城市规模分布[26-31],存在的问题是:用户籍人口和非农人口作为指标来衡量城市规模都存在一定的偏误(大规模人口流动导致常住人口与户籍人口存在显著差异,而非农人口除城镇人口外还包括了部分居住在农村的居民)[31],用常住人口来反映城市规模相对要准确些,而城市常住人口数据往往较难获取。总的来说,尽管我国目前的空间单元还是用行政建制来划分的,但这些传统的空间研究尺度都因太过宏观,不能反映出真实的城市规模而受到质疑,因此,如何定义真实的城市边界成为需要进一步探索的问题,有学者提出了自然城市(natural cities)这一概念[21],逐渐将对城市规模分布的实证研究转向微观尺度,突破传统的行政区划界限,将实际上起到城市功能的微观城市组团(图斑)作为基本空间单元来研究[18,20],从而使研究的精确度更高,更具有说服力[32]。本文将从微观空间尺度的数据出发,用自下而上的方法来构建新的城市组团,在此基础上拟合中国城市规模分布的齐普夫定律,并与中国传统空间尺度上的城市规模分布进行比较。

二、研究方法与数据说明

国外微观空间尺度的实证研究主要用的是人口普查小区数据或公里格网数据,而我国由于微观空间数据缺乏,目前能获取到的最精细的空间数据仅为乡镇街道级数据。根据城市功能区域的划分及现实数据的可得性,本研究采用的研究数据为2008年第二次全国经济普查乡镇街道级的就业人口数据,以全国43 843个乡镇街道办(不包括我国港澳台地区,下文同)来刻画基本城市单元,用就业人口数据代替传统的人口数据来测度城市规模,更能反映城市作为经济中心的功能特性。

本文的研究方法采用的是Rozenfeld等人的城市聚类算法(City Clustering Algorithm,CCA)[18]。该算法是由Rozenfeld等人[33]在Makse等人[34]工作的基础上改进得来的,是一种基于较精细空间尺度上的人口地理分布来定义城市边界的新算法,它突破了传统的行政区划界限,可有效地弥补基本行政单元低估大城市数量和城市群低估小城市数量的缺陷。该方法假设人口变化是反映城市发展或衰退的一个重要因素,因此,对城市边界的定义可通过对城市人口进行聚类来得到。具体操作步骤是:为了定义一个城市组团(city cluster),首先选定一个人口密集点,将周边邻近的点用递归的方式不断地纳入城市组团中(距中心点的距离小于一个给定范围l,并且人口密度D大于给定阈值D^*);到落在城市组团外的点距离城市组团边界全都大于l,且$D<D^*$的时候,城市组团停止增长。建立好城市组团后,将城市组团内的人口数加总作为整个城市组团的人口总数。

① X. Ye, "Spatializing Zipf's Law in the Dynamic Context: US Cities 1960-2000," UCGIS 2006 Conference, 2006.

② 中共中央政治局2014年7月30日召开会议,审议通过了《关于进一步推进户籍制度改革的意见》,决定建立城乡统一的户口登记制度,取消农业户口与非农业户口的性质区分,统一登记为居民户口。由于是刚刚才颁布的新政策,国内研究根据统计数据的可得性,一般仍采用城市户籍人口、非农业人口或市辖区人口数据来研究城市规模分布,因此本文仍沿用旧称。

故将城市定义为尽可能多微观空间单元连接在一起所产生的城市组团。本文用聚类所得的城市组团进行城市规模分布拟合分析。

如图 1 所示[18]，每个点表示美国的一个人口普查区(FIPS)，使用的人口密度阈值为 $D^*=0$。图 1(A)：选择任意一个人口密集点，围绕该点画一个半径为 l 的圆，浅色实心点为落入此圆范围内的点。图 1(B)、(C)：以此圆中的新点为圆心继续往外画圆，令更多的点落入这个城市组团(各个圆所包含的总区域)中来，将该递归过程不断地进行下去。图 1(D)：直到城市组团外再也没有与城市组团内任一点的距离小于 l 的点时，城市组团停止扩张。

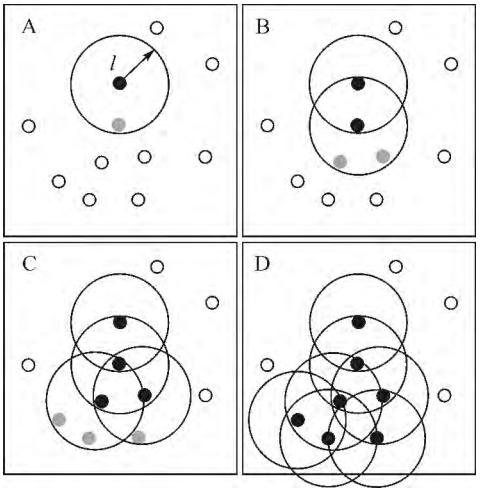

图 1　美国的城市聚类算法示意图

本文使用 CCA 算法对中国 43 843 个乡镇街道办点进行聚类后，参考 Rozenfeld 等对美国数据分析所选取的距离阈值 l(2km、3km 和 4km)[18]，并根据中国实际情况不断进行调试，最终选取距离阈值 l 分别为 6.5km、8km 和 9.5km 进行比较。选取这三个阈值既能有效地避免空间研究单元各成一类($l<6.5$km时)，又能防止出现空间研究单元大部分聚成一类、得到类别较少的情况($l>9.5$km 时)，并且在这三个阈值下的聚类结果特征比较明显，有一定的区分度(间隔少于 1.5km 取点，得到的分析结果之间差异不明显)。

聚类结果表明①，选择的距离阈值越大，得到的城市组团数越少，每个城市组团中包含的单元越多。当 $l=6.5$km 时，各空间单元基本各自成类(单元最多的一类也只有 178 个)，没有形成较大的城市组团。当 $l=8$km 时，空间单元开始出现聚类趋势，其中第一大类(环渤海地区)有 4 680 个单元，第二大类(长三角地区)有 3 956 个单元，第三大类(成渝地区)有 2 366 个单元；此外还形成了以下几个较小的城市组团(包含 200 个单元以上)：西安—咸阳—宝鸡—渭阳，太原—临汾—吕凉，长沙—岳阳—益阳—湘潭—株洲—常德，金华—衢州—上饶，广州—东莞—佛山—中山。当 $l=9.5$km时，第一大类(东部沿海地区，包括环渤海地区、长三角地区及其部分内陆腹地在内，北至北京，南至温州，西至宝鸡)纳入了 13 411 个空间单元，第二大类(成渝地区)包含了 5 170 个单元，第三大类(中三角的湖南省部分)有 2 327 个单元，珠三角的城市(广州、深圳、东莞、佛山、珠海、中山、肇庆、惠州、江门)则聚成第四大类，共有 484 个单元(选取小于 9.5km 的距离阈值时，珠三角地区城市聚类不太明显)。由于所用的微观空间数据存在误差，地图上少数单元没有数据，从而导致得出的城市聚类与现实情况存在一定偏差，但大体规律还是符合中国目前城市群分布情况的，比较重要的几个城市群都被提取出来了。

三、中国城市规模分布实证研究

(一) 结果分析

对中国 2008 年乡镇街道级就业数据进行拟合得到如下概率分布图(图 2)。该图显示了 $l=6.5$km、$l=8$km 及 $l=9.5$ km 的分布概率结果，这三种阈值下的城市组团数分别为 29 130、

① 因篇幅问题，聚类结果的三幅中国地图不在此处显示，若有兴趣，可向作者索取。

19 526和12 862个。我们发现中国的就业人口分布服从以下幂律分布形式：$P(S>S^*)\sim S^{-\zeta-1}$，且幂律指数$\zeta\approx1$，近似符合城市规模分布的齐普夫定律。图2中的三条拟合分布曲线与理论上的齐普夫曲线(幂律指数为1)还是比较接近的，尤其是$l=8$ km时。以$l=8$ km为例，其幂律指数是对符合$S>S^*=10\ 000$人的城市组团(总数为2.4亿人，包含了全国90%的就业人口)进行拟合分析得到的结果，用OLS回归得到在95%置信区间内的幂律指数$\zeta=1.106\pm0.003$。图3显示了在不同距离阈值下拟合所得的幂律指数，可以看出幂律指数波动范围很小，介乎0.95和1.15之间，在$l\in[7\text{km},10\text{km}]$区段，幂律指数围绕标准齐普夫指数(值为1)在5%范围内波动。随着距离阈值的增加，幂律指数是下降的，反映出城市规模是趋于集中分布的，由于聚类数减少，人口逐渐集中在较大的城市组团中，城市体系的人口分布差异较大，集中的力量大于分散的力量。

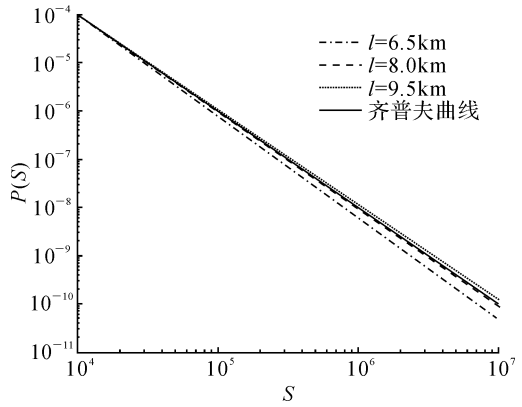

图2 不同距离阈值l下的城市组团人口概率拟合分布①

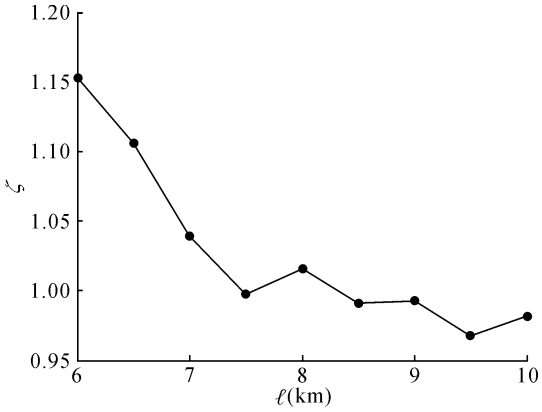

图3 不同距离阈值l下城市组团的幂律指数

为了对齐普夫定律拟合结果进行有效性检查，本文参考了Gabaix[35]、Gabaix和Ibragimov[36]的检验方法，假设实际分布与纯幂律分布之间存在一个很小的二次项偏离，对二次项的检验可用于决定幂律分布(齐普夫定律)能否用于描述城市人口规模分布。方法如下：将城市组团按规模大小排序，对位序i($i=1$为最大的城市)做如下OLS回归：

$$\ln(i-0.5)=\text{constant}-\zeta\ln S_i+q(\ln S_i-\gamma)^2 \tag{2}$$

从该回归中求得ζ和q，其中$\gamma=\frac{\text{cov}((\ln S_j)^2,\ln S_j)}{2\text{var}(\ln S_j)}$。幂律分布假设在渐进极限情况下，二次项系数$q=0$，所以较高的$|q|$意味着实际城市规模分布偏离幂律分布。在幂律分布的原假设下，$\frac{\sqrt{2N}q_N}{\zeta^2}$的值趋于标准正态分布($N$是样本点的数目)，即在99%的概率水平下，标准正态分布值的绝对值小于2.57。则$q_c=\frac{2.57\zeta^2}{\sqrt{2N}}$是二次项$q$在1%置信水平下的临界值，如果$|q|>q_c$，则拒绝原假设；如果$|q|<q_c$，则接受原假设，即服从幂律分布。从表1可看出，在各个距离阈值上，中国的城市规模分布都可以拒绝对齐普夫定律的二次项偏离修正，表明齐普夫定律可以很好地拟合人口大于10 000人的城市组团的城市规模分布，齐普夫定律在中国基于乡镇街道级空间数据的CCA城市组团层面上是成立的。

① 注：图中坐标轴刻度为对数刻度。

表 1　城市组团的幂律分布有效性检验结果

距离阈值(km)	占全国总人数的百分比(%)	$\|q\|$	q_c	检验结果
6.5	82.9	0.014	0.038	服从幂律分布
8.0	90.3	0.016	0.040	服从幂律分布
9.5	94.0	0.036	0.048	服从幂律分布

（二）与美国研究结果相比较：研究方法合理性检验

用城市聚类算法分析所得的研究结果表明，中国与美国的城市规模分布都在一定的阈值范围内基本符合齐普夫定律，城市规模分布属于位序-规模型，这种均衡型的等级规模分布是比较稳定的。中国和美国之间无论是在采用的指标还是人口密度方面都差异较大，且本研究采用的是 2008 年的就业人口数据，美国研究采用的是 2000 年的人口普查数据，对两国的实际城市规模分布情况及就业体制、状况进行比较研究的意义并不大。然而，为了从研究方法的角度上证实本研究所选取距离阈值和人口阈值的合理性，可以参考 Rozenfeld 等人对美国数据的分析过程[18]，因为中美两国的人口规模、国土面积和研究的初始空间单元数都处在同一量级上，较为接近。基于 CCA 的中国与美国城市规模分布研究相关参数比较结果如表 2 所示：

表 2　基于 CCA 的中国与美国城市规模分布研究相关参数比较

中国					美国				
距离阈值(km)	对应城市组团数	初始空间单元数	齐普夫分布的人口阈值	幂律指数为 1±0.05 的距离范围(km)	距离阈值(km)	对应城市组团数	初始空间单元数	齐普夫分布的人口阈值	幂律指数为 1±0.05 的距离范围(km)
6.5	29 130				2	30 201			
8.0	19 526	43 843	10 000	[7，10]	3	23 499	61 224	12 000	[2.5，3.5]
9.5	12 862				4	19 912			

注：其中距离阈值和人口阈值的选择都是经过反复调试后得到的最服从齐普夫定律的结果。美国的城市规模分布结果来源于 H. D. Rozenfeld，D. Rybski & X. Gabaix et al.，"The Area and Population of Cities：New Insights from a Different Perspective on Cities，" *American Economic Review*，Vol. 101，No. 5(2011)，pp. 2205-2225。

由表 2 可知，首先，中美由于人口总数和人口分布情况不同，在城市规模分布研究中具体所选参数也有所不同。然而，由于中美的人口和土地面积都是在一个量级上的，且分析的初始空间单元数都在 50 000 左右，所选的聚类参数还是有相似点的，如所选城市组团人口阈值都在 10 000 人左右，且最终聚类所得的城市组团数皆介乎初始空间单元数的 1/3—2/3 之间，以保证聚类结果是有效的（不至于使每个点自成一类，或者过多的点聚成一类）。

其次，中国和美国所选的距离阈值都是千米量级的，而中国的距离阈值比美国要大一些，原因有三：一是研究所用的基础数据不同，本研究采用的中国空间单元是乡镇街道办，就业人口从几人到几十万人不等，美国的研究采用的空间单元是人口普查区（FIPS），人口一般为 1 500 至 8 000 人，人口分布比较平均；二是中国的人口空间分布更不均衡，尤其表现在东西差异上，东部乡镇聚落分布极其密集，很容易在聚类过程中连成一大片，西部则面积较大而乡镇聚落较少，在聚类过程中往往各成一类，本研究采用的是就业人口数据，不均衡程度更加严重；三是对中国来说，由于初始空间单元数要少于美国，每个初始空间单元对应的平均面积也比美国要稍大些。

最后,幂律指数较接近1的距离阈值区间,在中国是7—10km,美国为2.5—3.5km,相较而言,中国的城市规模分布幂律指数收敛程度较高。从数据上看,美国的微观空间灵敏度(幂律指数变化/距离阈值变化)比中国要高,即美国齐普夫指数的变动幅度比中国要大,这是由于美国研究的空间数据与本研究相比,尺度更微观,精度更高。

(三)与其他空间尺度的研究结果相比较:研究方法优越性检验

用城市聚类算法处理所得的城市组团与未经处理的原始空间单元,两者的分析结果是否一样呢?根据公式(2)分别对地级、区县级和乡镇街道级的就业数据进行关于齐普夫定律的OLS回归拟合和假设检验,得到的结果如表3所示,其中人口阈值的选择是根据中国各级空间单元的大概人口数,经过不断调试后决定的。

表3 各级空间单元的幂律分布有效性检验结果

空间尺度	人口阈值(S^*)	占全国总人数的百分比(%)	幂律指数(ζ)	$\lvert q \rvert$	q_c	检验结果
地级	100 000	99.6	1.020	0.259	0.106	不服从幂律分布
	200 000	97.2	1.174	0.253	0.152	不服从幂律分布
	500 000	80.7	1.342	0.359	0.271	不服从幂律分布
	800 000	65.3	1.428	0.520	0.412	不服从幂律分布
	1 000 000	60.0	1.597	0.509	0.575	服从幂律分布
区县级	10 000	99.5	0.911	0.151	0.032	不服从幂律分布
	30 000	95.0	1.111	0.119	0.057	不服从幂律分布
	50 000	88.7	1.205	0.106	0.079	不服从幂律分布
	80 000	79.7	1.270	0.114	0.108	不服从幂律分布
	100 000	74.6	1.309	0.119	0.130	服从幂律分布
乡镇街道级	300	79.1	0.709	0.153	0.005	不服从幂律分布
	500	78.4	2.100	0.172	0.046	不服从幂律分布
	1 000	76.6	0.894	0.207	0.010	不服从幂律分布
	5 000	65.0	1.369	0.274	0.035	不服从幂律分布
	10 000	56.1	1.544	0.299	0.055	不服从幂律分布

从表3可以看出,与其他空间尺度的城市人口规模分布情况相比,基于CCA城市组团空间尺度的人口规模分布拟合结果比较服从齐普夫定律,且幂律指数接近于1($\zeta\in[0.95,1.15]$)。而地级、区县级和乡镇街道级的城市人口规模分布在很大程度上不满足幂律分布,即使在服从幂律分布的情况下,幂律指数仍然与齐普夫定律中的标准值1存在一定偏离。该研究结果表明,城市聚类算法在一定程度上弥补了较宏观空间尺度(区县级、地级)低估小城市组团数量和较微观空间尺度(乡镇街道级)低估大城市组团数量的缺陷;经过城市聚类算法处理的微观地理数据,其人口规模分布拟合结果较好地满足了齐普夫定律。此外,通过以上对城市组团、地级和区县级空间单元人口规模分布的研究,可以证明齐普夫定律确实只在满足一定人口阈值之上的空间单元中才成立,即齐普夫定律的成立存在城市规模下限的约束,存在城市规模分布曲线的"上尾奇异性"(a singularity of the upper tail),而这个人口阈值关系着对真实城市的定义(即人口数超过多少才可视为城市)。这

个阈值在不同国家是不一样的，即使在中国，不同级别的空间单元满足齐普夫定律的人口阈值也有数量级上的区别，这是需要反复试验探讨的问题。

尽管在 CCA 中，生成城市组团的距离阈值(l)是可以选择的，但在实际操作过程中对 l 的选择难免存在一定的主观性。为了选出合理的距离阈值，本文将 CCA 城市组团的城市规模分布规律与区县级、地市级的城市规模分布规律进行比较，试图分析三者之间存在的相关关系。

在分析区县级单元与城市组团之间的对应关系时，由于每个区县级单元包含多个 CCA 城市组团，而每个 CCA 城市组团都对应唯一的区县级单元，故将 3 124 个区县级单元与其所包含的城市组团中就业人数最多的城市组团进行匹配分析。在分析区县单元人口规模与对应的城市组团人口规模之间的相关关系时，取其对数，通过 OLS 回归构建两者相关关系为：

$$\ln S_i^{\mathrm{CCA}}(l) = a(l) + b(l)\ln S_i^{\text{county-level}}(l) \tag{3}$$

计算两者之间的 Pearson 相关系数 $\rho(l)$ 及欧氏距离 $d(l)$：

$$d(l) = \sqrt{\sum_{i=1}^{3124}(\ln S_i^{\text{county-level}}(l) - \ln S_i^{\mathrm{CCA}}(l))^2} \tag{4}$$

从图 4 和图 5 可以看出，基于 CCA 城市组团的城市人口规模分布与基于区县级空间单元的城市人口规模分布之间存在较强的正相关关系($\rho(l)\in[0.9, 0.95]$)，且随着距离阈值的增加，$\rho(l)$ 增加，$d(l)$ 减少，表明两者的分布情况越来越接近。Rozenfeld 等人认为，$\rho(l)$ 值最大、$d(l)$ 值最小的距离阈值为最优值，对美国来说分别是 $l=3$ km 及 $l=5$ km。而对于中国来说，随着距离阈值的增加，到 $l=9.5$ km 时东部沿海地区的乡镇街道办已经连成一大片。当 $l=10$ km 时，城市组团数变成了 11 181 个，城市组团数低于初始空间单元数的 1/3，聚类的距离阈值过大，东部人口密集区的乡镇街道级空间单元无法区分开来。因此，根据中国实际情况，不能简单地根据 $\rho(l)$ 和 $d(l)$ 的值来选择最优距离阈值(即较大的距离阈值较优)，综合聚类结果和分布曲线拟合结果来看，最优距离阈值应该在 9.5 km 左右；且只有当距离阈值增加到 9.5km 时，中国三大城市群之一的珠三角城市群才在聚类结果中突显出来，此时的聚类结果可基本反映出中国城市群分布现状。

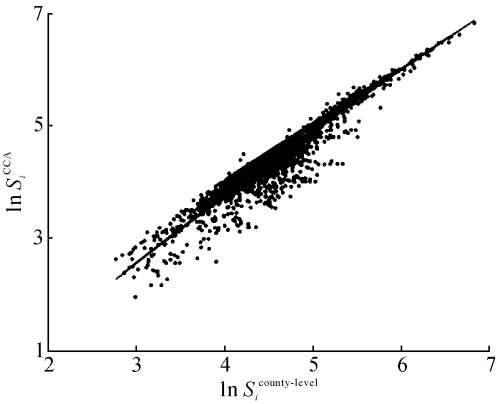

图 4　区县级单元与 CCA 城市组团之间的人口规模对应关系($l=9.5$ km)

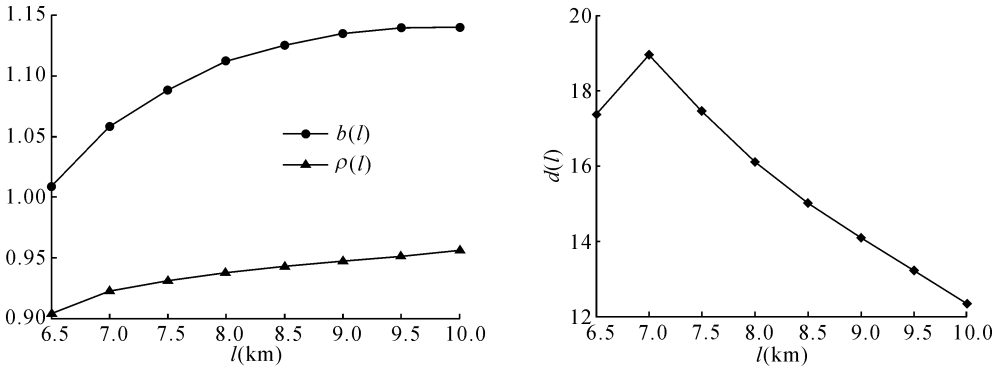

图 5　各距离阈值下城市组团与区县级单元之间的 $b(l)$、$\rho(l)$ 及 $d(l)$ 值

同理,我们分析了地级单元人口规模与对应的城市组团人口规模之间的相关关系,仍然是将333个地级单元与其所包含的CCA城市组团中就业人数最多的城市组团进行匹配分析,结果如图6和图7所示。地级单元与城市组团之间人口规模的相关关系与区县级单元类似。总的来说,用城市聚类算法处理过的微观数据拟合所得的中国城市规模分布规律,与直接用地级市和区县级空间单元刻画的城市规模分布规律大体上一致,且随着所选距离阈值的增加,其空间研究单元变大,人口规模分布自然地与较宏观数据的拟合结果越来越接近。然而,较宏观层面地理数据(地级单元和区县级单元)的空间单元都较大,城市人口规模下限值较高,容易忽视众多较小的城市组团在城市规模分布中的实际地位,从而导致拟合结果不服从幂律分布。此外,由ρ值可以看出,与地级单元相比,区县单元人口规模分布与城市组团人口规模分布情况较为接近,这充分说明,虽然各空间尺度上城市规模的大体分布趋势相同,但所选地理数据越微观,则刻画出的人口规模分布情况越接近于实际情况。

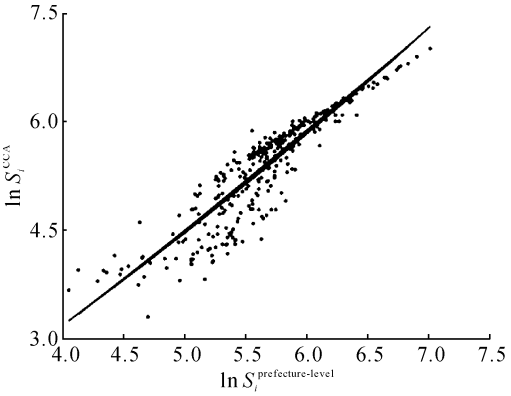

图6 $l=9.5$ km时地级单元与CCA城市组团之间的人口规模对应关系

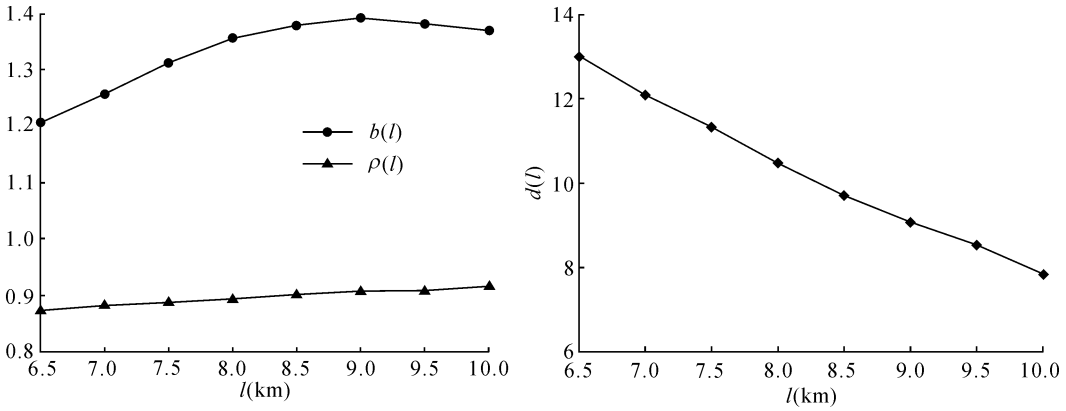

图7 各距离阈值下城市组团与地级单元之间的$b(l)$、$\rho(l)$及$d(l)$值

(四)研究方法稳健性检验

下面来验证城市聚类算法的稳健性,即验证本文的基本研究结果并不是由城市聚类算法(CCA)促成的,或者换句话说,不是由城市聚类算法产生的人工伪造结果。关于城市聚类算法的稳健性,Rozenfeld等人已经做出了验证[18],具体做法如下:以美国为例,将美国61 224个人口普查区(FIPS)空间点的实际位置打乱后,令其随机分布在一个与美国国土面积相等的矩形内,即将点的空间属性打乱,而非空间属性(人口数)不变,然后用城市聚类算法处理新生成的空间点。所得的城市规模分布拟合结果表明:打乱后的数据拟合结果并不符合齐普夫定律,与真实数据的拟合结果存在一定程度的偏离;打乱后的数据中最大城市组团的人口数只有196 112,即随机处理过程阻止了较大城市组团的形成。该验证结果表明,原始数据之所以呈现出符合齐普夫定律的城市规模分布规律,是由原始数据本身的内在特性决定的,而非城市聚类算法决定的,城市聚类算法只起到了对数据进行优化加工的作用,充分说明该算法只是一种研究手段,而非决定因素。Rozenfeld等人

用英国公里格网微观数据来验证 CCA 的稳健性，结果也一样[18]。因此，用由城市聚类算法产生的城市组团来分析城市规模分布规律是科学可行的。

四、总结与讨论

本文的研究结果表明：首先，目前可用齐普夫定律来较好地拟合中国城市规模分布，表明中国现阶段基于就业人口的城市体系符合位序-规模分布，各位次城市的发展较为均衡，城市规模分布趋于集中和趋于分散的力量基本平衡。其次，城市聚类算法是一种研究城市规模分布的较好的研究方法，其突破传统的行政区划界限，有效地弥补了宏观层面研究（低估小城市组团数量，样本较小）和微观层面研究（高估小城市组团数量，数据存在误差）的不足，架设了两者之间的桥梁。用该方法研究城市规模分布，更能反映真实的城市规模，所得结果更具科学性和合理性。该方法的有效性和稳健性已经得到了美国、英国[18]和中国数据（本文）的验证：美国超过 12 000 人的城市组团、英国超过 5 000 人的城市组团，以及中国超过 10 000 人的城市组团，其城市规模分布规律均能用齐普夫定律来较好地刻画。再次，中国和美国的城市规模分布基本上服从齐普夫分布，两国的拟合结果相似是由土地与人口量级上的一致性所促成的，不同之处则是由人口分布均衡度不同与微观数据精确性差异所造成的。最后，在城市规模分布的齐普夫定律实证研究中，对实际城市的定义仍然是一个需要深入探讨的问题。本研究发现，选取不同的人口阈值所得到的分布规律是不一样的，很难说这个城市规模的下限值是否存在普适性规律，需要用不同国家的不同空间尺度数据来进行反复的试验和探索。

然而，本研究与国外研究相比，还存在以下不足：一是微观数据不够精确，受数据可得性限制，研究数据是用 2008 年的经济普查就业数据与 2000 年的乡镇街道办空间底图相匹配得来的，虽然经过数据校正，但仍存在一定误差；二是缺少微观人口数据，无法与微观就业数据的拟合结果进行比较。

基于本文的研究发现与不足之处，笔者提出进一步的研究方向，具体如下：

首先，从数据角度来看，一是将基于人口数据的行政性城市实证研究与基于就业人口数据的功能性城市实证研究相比较，进一步验证齐普夫定律。因为从理论上来说，功能性城市概念更符合城市的本义，且 Rosen 和 Resnick 的研究表明：按照行政性城市测算的齐普夫指数的变动幅度，要比按照功能性城市测算的齐普夫指数的变动幅度大得多；而且，按照功能性城市拟合的结果比按照行政性城市拟合的结果更接近齐普夫定律，体现为齐普夫指数更接近于 1[7]。二是收集时间序列的数据，研究中国城市规模分布的动态演化规律，分析幂律指数的影响因素，从而为未来中国城市体系规划提出有针对性的政策建议。

其次，从方法来看，目前学术界定义城市边界的较新方法除了 CCA 算法外，还有 Jiang 和 Jia 对美国的街道节点（包括交叉点及末端）进行聚类形成自然城市的方法[21]。与他们的算法相比，CCA 算法继承了传统的空间单元定义法——大都会统计区划分方法（Metropolitan Statistical Areas，从一个人口密集的中心区开始，将周边与其存在密切社会经济联系的县域联结起来）的优点，并弥补了 MSA 存在一定主观性的缺陷（由人工逐个构建），基于人口在微观地理单元的空间分布自动化且系统性地构建城市组团，可用于研究不同空间尺度上的人口增长和集聚过程。与 CCA 算法及 Jiang 和 Jia 的算法在原理上有类似之处的还有空间聚类算法（Spatial Clustering Algorithm，主要分为划分法、层次法、基于密度的方法和基于网格的方法）。作为空间数据挖掘的主要方法之一，它对处理海量空间数据、提取大型空间数据库中有用的信息和知识具有十分重要的现实意义[37]，但目前尚未有研究应用空间聚类算法来划分城市界限，这可以作为未来的一个研究方向。此外，由于用城市聚类算法提取出的城市聚类结果与中国实际城市群分布存在一定偏差，该算法在中国的应用仍存在一定的探索空间，未来可尝试用相对空间距离（如交通通达度）来代替绝

对空间距离来进行研究,并对城市聚类算法中的各个点根据实际情况赋以不同的权重,或对各个区域设置不同的距离阈值来进行反复试验,使聚类结果更符合真正的功能性城市组团分布情况。

最后,从政策意义来看,按照一定的标准(本文仅以距离阈值和人口阈值相结合作为标准来试验)来划定城市功能性组团,然后按照城市功能性组团来设置中国的城市,对于中国城乡二元格局正在转变、政府正逐步取消城市和农村界限的现阶段来说,具有重要的现实意义。目前中国的城市还是按行政建制来设立的,然而,近些年在城市研究领域中兴起了一个城市划分的新概念——城市功能区(functional urban area, FUA),可以大致定义为城市地区(或核心市区)和邻近的通勤区(边缘市区)[38]。城市功能区最重要的特性是超越了行政边界。出于统计上的原因,数据分析通常是基于行政单元的,但目前人们不断尝试从功能导向出发将较小的行政单元结合起来形成城市功能区,这样可以更有效地刻画城市经济活动的实际范围,从而使区域战略规划更加连贯和理性。目前城市间合作逐渐加强,FUA逐渐成为区域、社区决策以及地方乃至国家规划制定的基本单位之一,因此FUA除了在统计分析层面外,在政府层面也变得日益重要。对中国来说,早在1995年,由于中国的行政地域远大于城市功能地域,为解决传统的城市概念(行政地域)不能准确刻画真正起到城市功能的地域这一问题,周一星和史育龙提出了"城市功能地域"的概念:一般是以一日为周期的城市工作、居住、教育、商业、娱乐、医疗等功能所涉及的范围,它以建成区为核心,还包括与城市建成区存在密切社会经济联系并有一体化倾向的城市外围地域,以县为基本组成单元[39]。

[参 考 文 献]

[1] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Cambridge: Addison-Wesley Press, 1949.

[2] X. Gabaix, "Zipf's Law for Cities: An Explanation," *Quarterly Journal of Economics*, Vol. 114, No. 3 (1999), pp. 739-767.

[3] G. Duranton, "Some Foundations for Zipf's Law: Product Proliferation and Local Spillovers," *Regional Science and Urban Economics*, Vol. 36, No. 4(2006), pp. 542-563.

[4] G. Duranton, "Urban Evolutions: The Fast, the Slow, and the Still," *American Economic Review*, Vol. 97, No. 1(2007), pp. 197-221.

[5] E. R. Hansberg & M. L. J. Wright, "Urban Structure and Growth," *Review of Economic Studies*, Vol. 74, No. 2(2007), pp. 597-624.

[6] J. C. Córdoba, "On the Distribution of City Sizes," *Journal of Urban Economics*, Vol. 63, No. 1(2008), pp. 177-197.

[7] K. T. Rosen & M. Resnick, "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy," *Journal of Urban Economics*, Vol. 8, No. 2(1980), pp. 165-186.

[8] S. Brakman, H. Garretsen & C. V. Marrewijk, *An Introduction to Geographical Economics: Trade, Location and Growth*, Cambridge: Cambridge University Press, 2001.

[9] K. T. Soo, "Zipf's Law for Cities: A Cross-country Investigation," *Regional Science and Urban Economics*, Vol. 35, No. 3(2005), pp. 239-263.

[10] S. Terra, "Zipf's Law for Cities: On a New Testing Procedure," http://www.cerdi.org/uploads/ed/2009/2009.20.pdf, 2014-11-20.

[11] K. Giesen & J. Südekum, "Zipf's Law for Cities in the Regions and the Country," *Journal of Economic Geography*, Vol. 11, No. 4(2010), pp. 667-686.

[12] M. E. J. Newman, "Power Laws, Pareto Distributions and Zipf's Law," *Contemporary Physics*, Vol. 46, No. 5(2005), pp. 323-351.

[13] A. S. Garmestani, C. R. Allen & C. M. Gallagher, "Power Laws, Discontinuities and Regional City Size Distributions," *Journal of Economic Behavior & Organization*, Vol. 68, No. 1(2008), pp. 209-216.

[14] K. Giesen, A. Zimmermann & J. Suedekum, "The Size Distribution across All Cities—Double Pareto Lognormal Strikes," *Journal of Urban Economics*, Vol. 68, No. 2(2010), pp. 129-137.

[15] M. Bee, M. Riccaboni & S. Schiavo, "Pareto versus Lognormal: A Maximum Entropy Test," *Physical Review E*, Vol. 84, No. 026104(2011), pp. 1-11.

[16] Y. Malevergne, V. Pisarenko & D. Sornette, "Gibrat's Law for Cities: Uniformly Most Powerful Unbiased Test of the Pareto against the Lognormal," *Swiss Finance Institute Research Paper*, No. 40(2009), pp. 9-40.

[17] Y. Malevergne, V. Pisarenko & D. Sornette, "Testing the Pareto against the Lognormal Distributions with the Uniformly Most Powerful Unbiased Test Applied to the Distribution of Cities," *Physical Review E*, Vol. 83, No. 036111(2011), pp. 1-11.

[18] H. D. Rozenfeld, D. Rybski & X. Gabaix et al., "The Area and Population of Cities: New Insights from a Different Perspective on Cities," *American Economic Review*, Vol. 101, No. 5(2011), pp. 2205-2225.

[19] J. Eeckhout, "Gibrat's Law for (All) Cities," *American Economic Review*, Vol. 94, No. 5(2004), pp. 1429-1451.

[20] T. J. Holmes & S. Lee, "Cities as Six-by-Six-Mile Squares: Zipf's Law?" in E. L. Glæser (ed.), *Agglomeration Economics*, Chicago: University of Chicago Press, 2010, pp. 105-131.

[21] B. Jiang & T. Jia, "Zipf's Law for All the Natural Cities in the United States: A Geospatial Perspective," *International Journal of Geographical Information Science*, Vol. 25, No. 8(2011), pp. 1269-1281.

[22] J. Roca & B. Arellano, "Does the Size Matter?Zipf's Law for Cities Revisited," http://www-sre.wu.ac.at/ersa/ersaconfs/ersa11/e110830aFinal00374.pdf, 2014-11-20.

[23] B. J. L. Berry, F. E. Horton & J. O. Abiodun, *Geographic Perspectives on Urban Systems: With Integrated Readings*, Englewood Cliffs: Prentice Hall, 1970.

[24] L. H. Dobkins & Y. M. Ioannides, "Dynamic Evolution of the U. S. City Size Distribution," in J. M. Huriot & J. F. Thisse(eds.), *The Economics of Cities*, Cambridge: Cambridge University Press, 2000, pp. 217-260.

[25] D. Black & V. Henderson, "Urban Evolution in the USA," *Journal of Economic Geography*, Vol. 3, No. 4 (2003), pp. 343-372.

[26] S. Song & K. H. Zhang, "Urbanisation and City Size Distribution in China," *Urban Studies*, Vol. 39, No. 12 (2002), pp. 2317-2327.

[27] G. Anderson & Y. Ge, "The Size Distribution of Chinese Cities," *Regional Science and Urban Economics*, Vol. 35, No. 6(2005), pp. 756-776.

[28] Z. Xu & N. Zhu, "City Size Distribution in China: Are Large Cities Dominant?" *Urban Studies*, Vol. 46, No. 10(2009), pp. 2159-2185.

[29] K. T. Soo, "Zipf, Gibrat and Geography: Evidence from China, India and Brazil," *Papers in Regional Science*, Vol. 93, No. 1(2014), pp. 159-181.

[30] 高鸿鹰、武康平：《我国城市规模分布 Pareto 指数测算及影响因素分析》，《数量经济技术经济研究》2007 年第 4 期，第 43-52 页。[Gao Hongying & Wu Kangping, "The Estimates and Influential Factors of the Pareto Exponent of City Size Distributions in China," *The Journal of Quantitative & Technical Economics*, No. 4 (2007), pp. 43-52.]

[31] 梁琦、陈强远、王如玉：《户籍改革、劳动力流动与城市层级体系优化》，《中国社会科学》2013 年第 12 期，第 36-59 页。[Liang Qi, Chen Qiangyuan & Wang Ruyu, "Household Registration Reform, Labor Mobility and Urban Hierarchical System Optimization," *Social Sciences in China*, No. 12(2013), pp. 36-59.]

[32] 沈体雁、劳昕：《国外城市规模分布研究进展及理论前瞻——基于齐普夫定律的分析》，《世界经济文汇》2012 年第 5 期，第 95-111 页。[Shen Tiyan & Lao Xin, "The Progress and Theoretical Perspective of the Foreign Research on the City Size Distribution: An Analysis Based on Zipf's Law," *World Economic Papers*, No. 5 (2012), pp. 95-111.]

[33] D. H. Rozenfeld, D. Rybski & J. S. Andrade et al., "Laws of Population Growth," *Proceedings of the National Academy of Sciences*, Vol. 105, No. 48(2008), pp. 18702-18707.

[34] H. A. Makse, S. Havlin & H. E. Stanley, "Modelling Urban Growth Patterns," *Nature*, Vol. 377(1995), pp. 608-612.

[35] X. Gabaix, "Power Laws in Economics and Finance," *Annual Review of Economics*, Vol. 1, No. 1 (2009), pp. 255-294.

[36] X. Gabaix & R. Ibragimov, "Rank-1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents," *Journal of Business & Economic Statistics*, Vol. 29, No. 1(2011), pp. 24-39.

[37] J. Xi,"Spatial Clustering Algorithms and Quality Assessment," in *Proceedings of the* 2009 *International Joint Conference on Artificial Intelligence*, Los Alamitos: IEEE Computer Society, 2009, pp. 105-108.

[38] J. Antikainen, "The Concept of Functional Urban Area," *Findings of the ESPON Project*, Vol. 1, No. 1 (2005), pp. 447-452.

[39] 周一星、史育龙:《建立中国城市的实体地域概念》,《地理学报》1995年第4期,第289-301页。[Zhou Yixing & Shi Yulong, "Towards Establishing the Concept of Physical Urban Area in China," *Acta Geographica Sinica*, No. 4(1995), pp. 289-301.]

"浙江区域史研究的回顾与展望"学术研讨会综述

2014年12月12日,由浙江大学历史学系、浙江大学中国古代史研究所、浙江省人民政府地方志办公室共同主办的"浙江区域史研究的回顾与展望"学术研讨会在浙江大学召开。来自浙江大学、宁波大学、厦门大学、浙江省社科院等省内外高校和科研院所及浙江省内方志办单位等相关学者六十余人参加会议。研讨会主题发言从宏观到微观,全面地了审视浙江区域史研究的进展,分组讨论则围绕四个主题进行了广泛而深入的探讨。《传承与创新:浙江地方历史与文化学术研讨会论文集》发行仪式亦在期间举行。现将会议主要内容综述如下:

(1)主题发言:浙江区域史研究的思考。刘进宝就区域史研究者知识背景、区域划分、史料时空分布的不平衡、区域史与整体史的关系等问题发表看法。陈国灿对区域文化与地方文化进行了区分,回顾了浙江区域文化研究的历程,并就政府参与下的研究容易导致某些问题提出了思考。钱茂伟认为,文化上的浙东具有全国性意义,并从多个方面回顾与总结了2005年以来的浙东学术研究,指出重提"浙学"的可能性。李学功对《补农书》进行介绍,提出应当重视其对于明清江南经济史研究以及认识小农经济的重要意义。宫云维指出"浙商"研究存在着厚今薄古的倾向,需从史料入手加强对传统"浙商"的研究。何勇强提出"越文化"在地域、时间、民族等不同层面意义的不确定性,同时指出区域文化研究中的同质化问题。

(2)浙江地方与人物。方新德以上虞为例探讨了如何处理地方自我美化与史实之间的矛盾,并指出境外材料对地方史研究的重要性。龚剑锋、金晓刚对百年来全祖望研究进行了梳理和总结。尤育号考察了温州旅沪同乡会与其家乡社会间互动的渠道和方式,同时提出区域间的联系亦应作为区域史研究的关注内容。尤东进介绍了浙东学派以及章学诚的学术思想。杜正贞借助浙西南宗族材料考察"禁立异姓为嗣"观念的产生与地方习俗形成的过程,并以之回应习惯法的产生问题。柴伟梁阐释了徐志摩的爱国情怀和爱乡情怀。朱珏关注了城市化进程中浙江历史文化名村的生存、保护和发展。

(3)文献整理与方志编纂。陈志坚梳理了六朝隋唐时期的"江南"等地理概念,指出上溯式研究思路存在的问题。鲍永军关注近年来浙江地方文献的收藏、整理和研究情况,并指出应重视当代文献的收藏整理及数字化。秦桦林介绍了黑水城出土元代杭州官刻本《元一统志》残页的考证过程,分析其方志学研究价值及历史文化意义。莫艳梅提出新时期亦应专家修志与众手修志相结合。周田田介绍了瑞安孙氏四代潜心方志收集、整理和研究的情况及其家族文化传统。方先勇就如何编好志稿以及史志关系进行探讨。徐逸龙以温州为例阐述了区域学术风气对方志编修的意义。刘峰以海宁为例讨论了地域文化传承的时代价值。

(4)浙江近现代变迁。梁敬明以义乌胡宅村为例探讨了工商化村庄的转型问题及土地、人口、区位等因素所扮演的角色。徐立望关注近代浙江教育转型,钩稽史料力图呈现求是书院之前一段被遗忘的历史。夏卫东讨论了1928年浙江人口调查及户口调查制度的变迁。熊彤考察了抗战时期浙江省政府的税收改革问题。颜志通过描述辛亥革命前后杭州三大经济事件中普通商人的行为展现秩序与利益的博弈。徐亮辨析浙江行政督察专员制度在抗战期间的变化及影响。叶君剑以政府与广播的关系为中心讨论民国时期浙江的广播事业。李凡考察了浙江大学史地系的发展。

(5)学术思潮的审思。陶磊认为江南早期开发研究过于强调农业的作用,而整体性研究相对不足。周运中讨论了海洋文明史研究的相关概念、现状及发展等,并就浙江开展此项研究提出设想。杨雨蕾探讨了将对外关系史研究与江南区域史研究有机结合的可能性和必要性。陈健梅从环境史角度梳理了针对"江南"人地关系问题的讨论。孙杰、孙竞昊强调江南社会经济史研究中的问题意识及比较研究的意义,并就相关研究进行检讨。屈啸宇、彭连生通过两个具体研究案例,提出了市场、知识与权力是宗教文化史研究中重要的思考方向。陆敏珍借助"思想市场"理论反思了区域史研究及看待各种相关理论的态度。张凯从路径与志趣视角探讨近代浙江文化史研究。

近年来浙江区域史研究逐渐受到重视,本次研讨会为专家学者提供了交流与合作的平台,共同推进了浙江区域史研究。

(浙江大学中国古代史研究所张权、赵卓供稿)