

近三百年中国城市的国际知名度 基于大数据的描述与回归

社会
2015·5
CJS
第35卷

陈云松 吴青熹 张 翼

摘 要:本文利用谷歌图书的百万书籍大数据,以中国地级以上(含)城市近300年来英语书籍中出现的词频来展示和分析城市国际知名度的变迁及其特征。研究发现,北京、香港、上海、广州、南京、澳门、天津、台北、重庆和拉萨依次为近300年国际知名度的前十强。在此基础上,本文进一步对基于书籍大数据的国际知名度和媒体提及率进行基于时间序列回归的格兰杰因果检验。研究表明,近代中国大陆城市在国际媒体中的提及率显著影响其国际知名度,而港澳台城市的国际知名度和媒体提及率之间则不存在这种关联。这意味着近代以来大陆城市的国际传播主要通过媒体报道进入西方社会。本文最后总结了城市国际知名度获得过程的模式和特征。

关键词:大数据 国际知名度 城市 文化传播

International Visibility of Chinese Cities in the Last Three Centuries: A Big-Data Based Description and Regression Analysis

CHEN Yunsong WU Qingxi ZHANG Yi

* 作者1:陈云松 南京大学社会学系(Author 1: CHEN Yunsong, Department of Sociology, Nanjing University) E-mail: yunsong.chen@nju.edu.cn; 作者2:吴青熹 江苏省委党校社会学教研部(Author 2: WU Qingxi, Institute of Sociology, Party School of Jiangsu Province); 作者3:张 翼 中国社会科学院社会学研究所(Author 3: ZHANG Yi, Institute of Sociology, Chinese Academy of Social Sciences)

** 本研究得到江苏高校哲学社会科学研究重点项目“中国传统文化的全球知名度:大数据分析”(项目号2015ZDIXM001)的资助。[This paper was supported by The Key Programme of Philosophy and Social Sciences Studies for the Universities in Jiangsu Province“Global Visibility of Traditional Chinese Culture: Big Data Analyses”(2015ZDIXM001).]

感谢《大数据时代》作者、牛津大学维克托·迈尔-舍恩伯格、南京大学周晓虹、张鸿雁、成伯清、闵学勤、王浩斌、吴愈晓,武汉大学罗教讲等教授对本文内容和方法所提出的意见和建议。斯坦福大学严飞、南京大学社会学院研究生黄超、刘伟峰、郭亚静、郝小帅和陈迪波参与了部分数据(地级市)的采集工作。读者如需特定城市国际知名度近300年详细数据和排名结果,可联系作者。

Abstract: This study utilizes Google Books search engine generated big data to examine the development of international visibility of Chinese cities in the last three hundred years. The visibility is measured by the frequency of a city name appeared in millions of English language books. We find that among the 294 cities in our study, Beijing, Hong Kong, Shanghai, Guangzhou, Nanjing, Macau, Tianjin, Taipei, Chongqing, and Lhasa come up as the top ten ones with the highest international visibility. Five factors appear to affect the international profiles of Chinese cities: (1) the city's socioeconomic importance in international relations; (2) the city's association with historical events; (3) the city's visibility has its life cycle; (4) the city's geographical affiliation with other cities; (5) international visibility cannot be definitively related to the city's history, population size and economic elements. We place the data of international visibility and media exposure under the Granger time series regression test and found that media coverage is a significant cause on the international visibility of mainland cities but not Hong Kong and Macau. The finding implies that media coverage is the major entry point for mainland cities to gain international recognition.

Keywords: big data, international fame, city, cultural dissemination

DOI:10.15992/j.cnki.31-1123/c.2015.05.003

一、研究缘起

城市是人类文明在地理空间上的汇聚点,具有文化的贮存、传播、交流、创造和发展等基本功能(Mumford,1961)。在城市的诸多内涵要素中,城市文化是城市的灵魂。作为文化在地理空间上的重要载体,一座城市在全球范围内的知名度是城市综合影响力的重要组成部分。一个国家的城市群体的影响力,是衡量国家的非权力性影响力,也即软实力的重要指标(Nye,1990)。因此,对城市文化影响力尤其是国际知名度的研究,具有重要的经济、社会和政治内涵。不过,囿于数据和测量的局限,学界迄今尚未对这一领域有过系统的分析和探索。例如,即便是对当代城市知名度进行分析,海外抽样问卷的方法也需要较大的成本且面临样本选择的问题,而如果要观察几个世纪以来城市知名度的变迁轨迹,传统的抽样数据和分析方法就更无法实现。

“大数据”尤其是数字化书籍大数据的适时出现,为开展相关的社会科学研究提供了空前的机遇(陈云松等,2015)。目前,谷歌图书

(Google Books)语料库能提供公元 1500 年以来 7 种语言 800 多万种数字化书籍的全文词频检索(Michel, *et al.*, 2011; Lin, *et al.*, 2012)。通过观测和分析关键词在语料库中使用频率的变化,我们可以发现相关的关键词在人类文化发展史中或鲜为人知或饶有趣味的趋势和现象。国际语言学界、历史学界对此已经开始进行跨学科探索(Bentley, *et al.*, 2014; Acerbi, *et al.*, 2013; Twenge, *et al.*, 2012)。国内学者也已利用这一数据,对跨度百年的社会文化现象进行历史轨迹描述和量化分析(陈云松, 2015; Chen and Yan, 2015)。

本文首先以谷歌图书 1700 年以来的百万英语书籍作为语料库,以中国大陆全部直辖市、副省级以上城市、省会城市、各地级市以及港澳台主要城市英文名称作为关键词,以这些关键词在语料库中每年出现的频率高低为指标,在纵向跨度 300 年、横向跨度达千亿词汇的文化大数据中,精确描绘上述各城市词频位次的变迁及特征。鉴于英语是近百年以来全球使用范围最广和最重要的通用语言之一,而积累数世纪的海量书籍则构成数百年里国际社会知识、观念和体验的最重要载体,因此我们认为,基于谷歌图书百万英语书籍语料库的城市词频,可以用来作为测量城市国际知名度的标尺。

在构建国际知名度测度的基础上,我们进一步对中国城市国际知名度的积累渠道进行分析。近代以降,囿于交通、信息传输技术和成本以及清政府在政治、经济、文化和外交上的闭关锁国政策,绝大多数中国城市与西方社会直接的人流、物流互动非常有限。因此,我们提出中国城市群体国际知名度形成的“差异化”假说。具体而言,近代以来中国大陆城市的国际知名度多受到西方当时主要媒体(报纸)提及率的影响,而对于香港、澳门和台北等曾经有过较长殖民地历史的中国城市而言,该关联并不显著:殖民统治使得这些城市直接成为中西文化对撞的窗口,从而具有与大陆城市(没有殖民地历史或殖民地历史较短)不同的知名度获得途径。为检验这一假说,本研究从《纽约时报》全文数据库中提取中国城市提及率指标,并将其与国际知名度指标进行了基于时间序列回归的格兰杰因果关系分析,以观察两者的统计关联在大陆城市和港澳台三城市之间有无差异。本文既是对城市国际知名度进行大数据测量的首次尝试,也是中国社会科学领域较早利用大数据进行的计量模型回归分析。

二、概念及测量

(一)城市的国际知名度

现有文献中城市的国际知名度仅局限于当代,多被量化为国际展览次数、知名搜索网站可搜索信息条数、利用外资和外资企业数量等指标(如马继刚等,2014;李红波等,2013)。而且,相关研究均局限在某个城市或某个区域,缺乏对全国城市群体的鸟瞰。总体上,现存文献对城市国际知名度的研究既缺乏时间维度上的历史回顾,也没有全国性的系统分析和测量。本文将借助数百年来人类最重要和最广泛使用的书面语言文化载体(书籍)这一大数据,来填补这一研究的薄弱环节。

我们认为,在一个具有足够规模、跨度和代表性的英语书籍语料库中,一座城市在某一时间跨度内出现的次数可以代表该城市在该时段内的国际知名度。该测量方法的严密性和合理性在于:第一,书面语言本身是承载人类观念、意识和价值观的最重要、最全面的载体,而英语是近代以来全球使用范围最广、最重要的国际通用语言。因此,英语书面语言载体是用来研究知名度的最佳数据库;第二,书籍中的语言和词汇既能反映作者/撰稿人的个人观点,又能反映和捕捉社会大众的整体思潮。积累数个世纪的海量英语书籍,实际上容纳和承载了国际社会绝大多数的知识、观念和经验。只要我们使用的书籍语料库具有足够的规模、跨度和代表性,我们就可以认为一个词汇在其中的出现频率能够近似地反映这个词汇本身及其蕴含的社会文化影响力和知名度,甚至折射出某种社会趋势、风尚或思潮(Twenge, *et al.*, 2012)。

不过,考虑到每年书籍中的词汇量不尽相同,因此,我们用词频除以当年的词汇总量,以获得数据的时间可比性。该比例的计算公式为

$$R_i = \frac{C_{it}}{C_t}$$

其中, R_i 为城市名在公元 t 年的词频比例,也即知名度。 C_{it} 表示城市 i 在公元 t 年的出现次数, C_t 为公元 t 年中出版书籍的全部词汇量。

(二)城市的媒体提及率

与基于海量书籍的知名度指标相比,媒体提及率所捕捉的主要是城市突发事件,也即新闻。因此,媒体提及率受新闻规律的影响显然更大,其变化起伏更为剧烈,内容覆盖较之书籍也更为狭窄。不过,当中

国城市和国际社会处于相对隔膜的状态时,国际知名度形成、积累的直接途径往往被堵塞,此时国际媒体的提及与否就会对国际知名度的变化产生非常大的影响,甚至成为主导。这也正是本文关心媒体提及率的缘起:我们试图分析近代以来中国城市的国际媒体提及率与国际知名度之间是否存在紧密联系。

由于互联网等新媒体在 20 世纪 80 年代之后才出现,而本文关心的是近代以来中国城市的国际知名度,因此,本文中的“媒体”是指近代报刊等传统媒体。在一个具有代表性的近代西方报刊全文数据库中,我们测量中国城市的媒体提及率的方法是:在分析时间段内的每一年,我们用该年度提及中国城市的文章数量除以该年度报刊的刊文总量。这个比值,也就是城市的媒体提及率,其计算公式为:

$$Q_{it} = \frac{C'_{it}}{C'_t}$$

其中, Q_{it} 为城市 i 在公元 t 年报刊中的提及率, C'_{it} 表示在公元 t 年的报刊中提及城市 i 的文章数量, C'_t 为整个 t 年中全部文章数。

三、数据和策略

我们使用谷歌图书语料库作为中国城市国际知名度的测量来源数据库。谷歌图书语料库源于谷歌公司自 2004 年底启动的对哈佛、牛津等 40 多所顶级大学图书馆藏书及出版社赠书的数字化工程。2008 年,让-巴蒂斯特·米歇尔(Jean-Baptiste Michel)等人从已被数字化的 1 500 多万种书籍中选择了其中 500 多万种(5 195 769 种)识别质量较高的非期刊书籍,作为其文化定量分析的语料库。这 500 多万种书籍时间跨度从公元 1500 年到 2000 年,含 7 种语言,占古登堡印刷术发明以来人类印刷出版图书总数的 4%,词汇量达 5 380 亿(Michel, *et al.*, 2011)。到 2013 年,超过 3 000 多万种书籍已被扫描和识别,可供分析的最新版语料库书籍高达 800 多万种(8 116 746 种),词汇量更高达 8 613 亿(Lin, *et al.*, 2012)。表 1 展示了谷歌图书语料库 2012 年版的主要构成。考虑到公元 1500 年至 1700 年语料库中的书籍较为稀少,甚至有年度空缺,我们把对谷歌图书语料库的分析历史的上限设定为公元 1700 年。考虑到 2000 年之后的书籍仍然在数字化过程中,为减少样本偏误,我们把分析历史下限设定为公元 2000 年。

表 1:谷歌图书语料库 2012 年版的构成

	811 万种书籍	
	书籍量	词汇量
英语	454 万	4 685 亿
法语	86 万	1 022 亿
西班牙语	79 万	840 亿
德语	66 万	647 亿
汉语(简体)	30 万	269 亿
俄语	59 万	670 亿
希伯来语	7 万	80 亿
意大利语	30 万	400 亿
合计		8 613 亿

我们选取《纽约时报》(*New York Times*)自 1851 年创刊至今 150 多年的数字化语料库作为提取中国城市媒体提及率指标的来源。《纽约时报》是美国社会代表性的主流媒体,每天在纽约出版、全世界发行,在全球范围内具有一定的影响力。2008 年,谷歌和《纽约时报》联手推出了注释语料库,涵盖了 1987—2007 年 180 万篇文章。2012 年,《纽约时报》的研发实验室将语料库进一步扩展到自 1851 年创刊至今的全部文章。

用于检索的“关键词”设定对于本研究也很重要。在关键词的设定过程中,我们注意到,绝大多数城市的英语名称近 300 年来发生了巨大变化:1949 年以前,中国城市在英语世界中多使用韦氏拼音法,而 1949 年之后尤其是改革开放以来多使用汉语拼音——例如北京,英语旧称 Peking,而现在一般是 Beijing。因此,要计算北京的词频,就必须让两者相加;同时,不少城市还不止一个别名,最典型的是厦门,现称为 Xiamen,而国外过去称之为 Amoy。再如大连,既有汉语音译性质的英文名称 Dalian、Dalny,还有日语音译和意译的 Dairen 和 Ryojun,以及在日俄战争中最集中使用的“亚瑟港”(Port Arthur)。因此,对本文所涉全部城市,我们均仔细考证了全部可能的英文名称,然后对检索结果进行加总,以获得最为精确的城市检索数据。¹在数据统计过程中,我们均设置了单词字母大小写的非严格区分,例如,Beijing 和 BEIJING 都可以被统计进来。

1. 需要英文城市名称检索关键词的读者可向作者索取。

四、近三百年中国城市的国际知名度

我们首先对 294 个城市近 300 年、200 年、150 年、100 年、50 年和 20 年的国际知名度指标进行分段分析(均截至 2000 年),计算出平均知名度来进行排名。考虑到文章篇幅,表 2 给出了国际知名度排名前 20 强的城市榜单。²从表中可见,近 300 年、200 年、150 年、100 年、50 年(1949 年之后)和 20 年(改革开放后)等六个不同历史跨度的国际知名度前 20 强均集中在 25 座城市。这表明,城市国际知名度本身是较为稳定和集中的指标,尽管 300 年来经历了清、中华民国和新中国三个政权,但知名城市的构成并无太大变化。

表 2:近代以来各历史阶段中国城市国际知名度前 20 强

	近 300 年	近 200 年	近 150 年	近 100 年	近 50 年	近 20 年
排名	1700—2000 年	1800—2000 年	1850—2000 年	1900—2000 年	1950—2000 年	1980—2000 年
1	北京	北京	北京	北京	北京	香港
2	广州	香港	香港	香港	香港	北京
3	香港	广州	上海	上海	上海	上海
4	上海	上海	广州	广州	广州	广州
5	澳门	南京	南京	南京	南京	台北
6	南京	澳门	澳门	天津	台北	南京
7	天津	天津	天津	澳门	天津	澳门
8	台北	台北	台北	台北	澳门	天津
9	重庆	重庆	重庆	重庆	拉萨	西安
10	拉萨	拉萨	沈阳	沈阳	重庆	拉萨
11	沈阳	沈阳	拉萨	拉萨	西安	重庆
12	厦门	厦门	厦门	大连	武汉	深圳
13	大连	大连	大连	西安	沈阳	武汉
14	宁波	西安	西安	武汉	延安*	沈阳
15	哈尔滨	宁波	宁波	哈尔滨	杭州	杭州
16	西安	武汉	武汉	厦门	厦门	厦门
17	武汉	哈尔滨	哈尔滨	杭州	哈尔滨	成都
18	杭州	杭州	杭州	福州	大连	苏州*
19	福州	福州	福州	青岛	成都	哈尔滨
20	苏州*	苏州*	苏州*	苏州*	福州	昆明

注:带*者为非省会地级市。

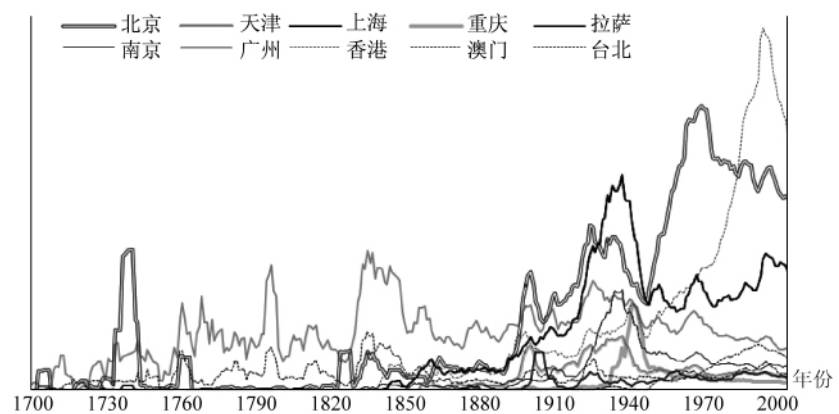
在六个不同历史跨度中全部进入前 20 强的城市有 17 座,³分别是

2. 需要其他 274 座城市国际知名度排名的读者可向作者索取。

3. 排名按照各历史阶段的平均位次而定,也即括弧内数字。

北京(1.17)、香港(2.0)、上海(3.3)、广州(3.5)、南京(5.3)、澳门(6.5)、天津(7.0)、台北(7.2)、重庆(9.5)、拉萨(10.2)、沈阳(11.5)、西安(12.8)、厦门(14)、武汉(14.7)、大连(14.8)、杭州(16.8)和哈尔滨(17.2);五次进入 20 强的为福州(18.5)、苏州(19.6);三次进入 20 强的为宁波(14.7)、深圳(16);两次进入前 20 强的为成都(18);一次进入的为延安(14)、青岛(19)和昆明(19)。除了苏州和延安是仅有的两个地级市,排名也比较靠后外,其他的 23 座城市则由省会、副省级(计划单列)城市、直辖市和香港、澳门、台北构成。

按照不同历史跨度的平均排名,北京、香港、上海、广州、南京、澳门、天津、台北、重庆和拉萨稳居近代以来中国城市国际知名度的前 10 名。考虑到图形识别度和篇幅限制,我们在图 1 中用时间序列曲线来展示这十座中国城市的知名度。⁴图 1 最大的特点就是:城市国际知名度呈现出非常明显的梯次和波动。所谓梯次,是指总体相对高低的层次;所谓波动,是指历史起伏的剧烈幅度。



数据来源:谷歌图书英语语料库,公元 1700—2000 年。

图 1:近 300 年来中国城市国际知名度前 10 强

例如,在 1700—1900 年的两百年间,北京、香港、上海和广州之中仅北京和广州在英语世界的书籍中被规模性地提及,且广州独领风骚,而北京只在 1735—1744 年间短暂地有所超越。直到 1850 年前后,上

4. 考虑到视觉效果,图形绘制中使用了前后各两年数值加以平滑。如 1950 年数据为 1948、1949、1950、1951、1952 共五年的平均值。

海和香港才开始出现在图中。进入 20 世纪以后,除广州长期在 1860 年左右的水平上下徘徊之外,其他三大城市的曲线出现了明显的上升。其中,北京一直强势上升并长期维持在高位运行;上海在 20 世纪 30 年代一度超越北京,但在新中国成立初期进入了低潮期,改革开放后又出现了明显的上升势头;香港的国际知名度自 1948 年起大幅提升,80 年代中后期就赶上并超过了北京,并于 1997 年达到了峰值。此外,在这些梯次和波动中,尤其是峰值和谷底,都富含了大量的历史、政治、经济和社会信息,我们会在下文做进一步诠释。

五、近代以来中国城市的媒体提及率

因《纽约时报》创刊于 1851 年,我们分别整理了近 150 年、近 100 年、近 50 年和近 20 年相关城市在该报的提及率排名,并将其与国际知名度排名一起综合在表 3 内。可以看出:媒体提及率和基于书籍大数据的知名度排名虽略有差异,但总体结构和特征都比较接近。4 个不同历史跨度的媒体提及率排名前 20 强集中在 23 座城市。媒体提及率和国际知名度的前 10 名城市构成几乎没有太大区别(除了澳门在早期的报道较少)。在后 10 名中,除了在近 150 年媒体提及率排名中出现的汕头和烟台,⁵其他城市也都出现在城市知名度排名的后 10 名中,只是在位次上存在差异。⁶表 3 提示我们,总体上中国城市国际知名度的获得和媒体提及率之间呈现出一个比较稳定的相关关系。

为便于和图 1 的曲线比较,我们在图 2 中分别绘制了北京、香港、上海、广州、南京、澳门、天津、台北、重庆和拉萨等知名度前 10 名城市的媒体提及率曲线。不难发现,尽管曲线的变化坡度等与图 1 大相径庭(这显然是由于媒体属性所致),但曲线的总体梯次和时段变化则非常接近。例如,无论是在国际知名度还是在媒体提及率指标中,广州的早期梯次都非常高,甚至超过北京、上海。在清末民初、新中国成立、中美建交等历史关口,北京的提及率和知名度都同样出现了曲线高峰。上海在 20 世纪 30 年代就超过了北京,香港则在 90 年代开始超越北京,如此等等,基本特征和图 1 中知名度的变化曲线非常接近。其他城

5. 汕头和烟台也是 19 世纪中期中国的通商口岸,因此获得了较高知名度。

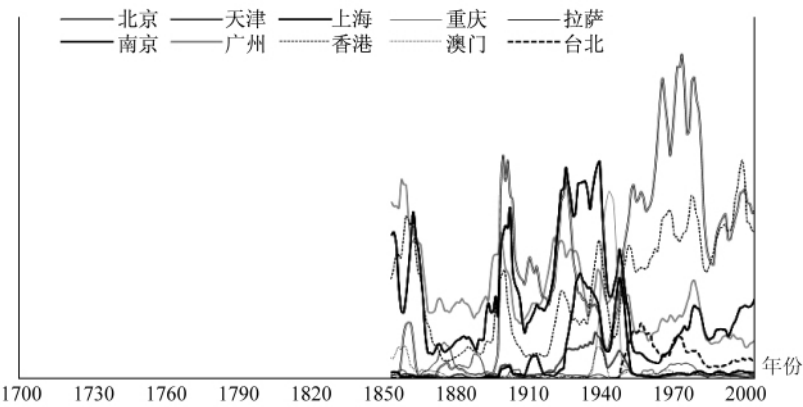
6. 延安和苏州未进入《纽约时报》提及率的前 20 强城市,可能体现了该媒体本身的报道偏向和立场。

市的媒体提及曲线和国际知名度曲线也呈类似的状态。

表 3:近代以来中国城市国际知名度(R)与媒体提及率(Q)排名对比

排名	近 150 年		近 100 年		近 50 年		近 20 年	
	1850—2000 年		1900—2000 年		1950—2000 年		1980—2000 年	
	R	Q	R	Q	R	Q	R	Q
1	北京	北京	北京	北京	北京	北京	香港	北京
2	香港	香港	香港	香港	香港	香港	北京	香港
3	上海	广州	上海	上海	上海	上海	上海	上海
4	广州	上海	广州	广州	广州	广州	广州	广州
5	南京	南京	南京	南京	南京	台北	台北	台北
6	澳门	重庆	天津	重庆	台北	西安	南京	南京
7	天津	台北	澳门	台北	天津	拉萨	澳门	西安
8	台北	天津	台北	西安	澳门	澳门	天津	澳门
9	重庆	拉萨	重庆	天津	拉萨	南京	西安	深圳
10	沈阳	西安	沈阳	沈阳	重庆	天津	拉萨	拉萨
11	拉萨	沈阳	拉萨	哈尔滨	西安	深圳	重庆	重庆
12	厦门	厦门	大连	拉萨	武汉	重庆	深圳	成都
13	大连	澳门	西安	澳门	沈阳	杭州	武汉	天津
14	西安	哈尔滨	武汉	大连	延安*	成都	沈阳	杭州
15	宁波	大连	哈尔滨	厦门	杭州	武汉	杭州	武汉
16	武汉	青岛	厦门	青岛	厦门	沈阳	厦门	哈尔滨
17	哈尔滨	汕头*	杭州	杭州	哈尔滨	厦门	成都	青岛
18	杭州	杭州	福州	长沙	大连	哈尔滨	苏州*	沈阳
19	福州	烟台*	青岛	武汉	成都	大连	哈尔滨	昆明
20	苏州*	长沙	苏州*	成都	福州	昆明	昆明	大连

注:带*者为非省会地级市。



数据来源:《纽约时报》语料库,公元 1851—2000 年。

图 2:近代以来中国城市国际知名度前 10 强的媒体提及率

六、知名度和媒体提及率的关联

利用我们的数据,可以对全部 294 座城市分别进行国际知名度与媒体提及率的时间序列回归(双变量)。考虑到篇幅,我们首先在表 4 中报告国际知名度靠前的 15 座城市(12 座大陆城市和香港、台北、澳门等 3 座有过较长殖民地历史的城市)的格兰杰因果检验结果。⁷考虑到这 15 座城市的知名度和提及率 150 年时间序列的稳定性不一,我们首先分别详细报告各序列的平稳性(即是否通过单位根检验)。如果两个时间序列都是非单位根过程,我们直接拟合 VAR 模型,并进行格兰杰因果测试;如果不是,则对不平稳的时间序列取差分,直至平稳后再拟合 VAR 模型,进行格兰杰因果检验。⁸在单位根检验中,本文使用迪克—富勒检验(DFGLS 检验)和菲利普—帕芬检验(PP 检验)两种方法。VAR 或 VECM 的滞后阶数(lag)均根据 AIC、BIC 和 LR 等信息标准来联合确定。

表 4 清晰地展示了 12 座大陆城市和港澳台 3 城市之间的显著差异。在大陆城市中,无论是开埠较早的广州、上海,还是相对处于内陆的西安、沈阳,它们的国际知名度都受到媒体提及率的影响。准确地说,早前数年的媒体提及率(或其变化)可以解释国际知名度(或其变化),也即媒体提及率是国际知名度的格兰杰原因。同时,部分城市的国际知名度和媒体提及率还呈现双向格兰杰关系。与此形成强烈反差的是,具有较长殖民地历史的香港、澳门和台北,其国际知名度和媒体提及率之间没有显著的统计关联。这意味着,英语世界对港澳台三城市的认知,更多地受到其他因素的影响,而不是通过媒体的中介作用。

那么,前 15 座城市之外的其他大陆城市的回归结果如何?尽管囿于篇幅,我们不再报告其他城市的格兰杰因果检验的结果,但我们发现,其余 279 座大陆城市的知名度和媒体提及率的回归呈现出高度相

7. 格兰杰因果关系为计量经济学术语,并非指反事实框架下的因果关系验证。在两个时间序列 X 和 Y 之间的格兰杰因果关系的最简单定义就是:如果变量 X 有助于解释变量 Y 未来的变化,则认为变量 X 是引致变量 Y 的格兰杰原因。

8. 对不平稳的时间序列直接进行基于 F 和 Wald 检验的标准格兰杰因果分析会产生偏误,因此我们采取取差分的方法。尽管差分的解释意义与原水平变量不同,但仍可体现知名度与媒体提及率之间的关联。此外,本文中的非平稳时间序列均为一阶单整,因此一阶差分稳定。

似的特征:大陆城市中媒体提及率的变化显著影响国际知名度变化的比例分别高达 85%。而且,没有显著的格兰杰因果关系的多为《纽约时报》媒体提及率几乎为零的城市。这进一步验证了我们的假说。此外,尽管数据所限,我们使用的是《纽约时报》的媒体提及率指标而非近代全部西方媒体的提及率,但基于《纽约时报》近代以来在新闻界的重要地位和影响,这一分析结果无疑是富有启发意义的。

表 4:城市国际知名度(R)和媒体提及率(Q)的格兰杰检验

城市	单位根检验		协整	差分单位根检验		格兰杰检验 Q 影响 R	格兰杰检验 R 影响 Q
	R	Q		R	Q	chi 2	chi 2
香港	单位根	平稳	—	平稳	平稳	0.855 lag=5	5.015 lag=5
澳门	平稳	平稳	—	—	—	1.945 lag=3	0.536 lag=3
台北	单位根	平稳	—	平稳	平稳	0.840 lag=2	3.664 lag=2
北京	平稳	平稳	—	—	—	15.338*** lag=3	9.997** lag=3
上海	单位根	平稳	—	平稳	平稳	3.222* lag=1	0.134 lag=1
广州	平稳	平稳	—	—	—	7.222** lag=2	2.005 lag=2
南京	单位根	平稳	—	平稳	平稳	16.019*** lag=2	15.072*** lag=2
天津	单位根	单位根	YES	平稳	平稳	47.615*** lag=3	14.408*** lag=3
重庆	平稳	平稳	—	—	—	194.03*** lag=4	28.616*** lag=4
沈阳	平稳	平稳	—	—	—	29.907*** lag=3	1.231 lag=3
拉萨	平稳	平稳	—	—	—	5.077* lag=2	2.214 lag=2
厦门	单位根	平稳	—	平稳	平稳	13.376* lag=6	12.174 lag=6
大连	平稳	平稳	—	—	—	22.393*** lag=2	0.742 lag=2
宁波	平稳	平稳	—	—	—	59.526*** lag=5	17.755*** lag=5
西安	平稳	平稳	—	—	—	7.522** lag=2	3.791 lag=2

注:1. * $p<0.1$, ** $p<0.05$, *** $p<0.01$;
2. 格兰杰因果检验的零假说为“Q 不是 R 的格兰杰原因”及“R 不是 Q 的格兰杰原因”。

七、中国城市国际知名度的形成模式和特征

利用谷歌图书大数据,本文对 300 年来中国主要城市的国际知名度进行了测量、可视化、排序和时间序列回归分析。由于本文的指标基于百万书籍大数据,因此,它能够较为客观、准确地展示这一时间段内西方公众特别是英语世界对中国城市的总体了解程度。根据前文的时间序列分析,可以发现,大陆城市的国际知名度更多地受到西方媒体提及率的影响,这初步验证了本文提出的“差异化”途径。接下来,本文将进一步完整地提出近代中国城市国际知名度形成的“二元模式”和五大特征。

(一) 国际知名度获得的“二元模式”

国际知名度获得的二元模式,主要是指在文化交流过程中存在的直接与间接的知名度获得过程。直接和间接交流的差异,在中国大陆城市和有较长殖民地历史的中国城市之间表现得非常明显。特别是,由于政治体制、市场结构和文化的差异,近代以来中国大陆城市在政治、经济、文化和人力资源的对外交流方面存在较高壁垒,而港澳台三城市曾长期作为殖民地,与西方社会的经济社会来往、人文政治互动等具有天然的畅通渠道。这种差异通过路径依赖和教育模式的复制而随时间不断强化(特别是在香港、澳门回归之前)。

总体上,这种差别的存在使得中国大陆城市和有较长殖民地历史的港澳台三城市在国际知名度获得方面形成了泾渭分明的两大类型。对于港澳台而言,它们“直接”成为中西文化对撞交流的窗口,较为接受西方文化,而西方社会也把它们视为自己的文化亲族,理所当然地会给予更多关注。对于大陆城市而言,它们在近代西方社会的国际知名度几乎只能通过当时的主要媒体来传递,形成“间接”的知名度获得形式。当然,随着 20 世纪 80 年代以来改革开放和全球化进程的加快以及交通运输、互联网技术的不断提升,这种直接和间接的知名度形成差异会不断弱化。

(二) 国际知名度结构的五大特征

1. 特征之一:总体梯次取决于其在中外经济交流中的地位变迁

广州是典型的案例。在 1700—1900 年这 200 年间,广州的国际知名度在绝大多数时间内甚至超过北京。究其原因,主要是 18 世纪中期开始的“一口通商”政策。乾隆二十二年十一月初十日(1757 年 12 月

20日)上谕:“本年乾来船虽已照上年则例办理,而明岁赴浙之船,必当严行禁绝。……此地向非洋船聚集之所,将来只许在广东收泊交易,不得再赴宁波。如或再来,必令原船返棹至广,不准入浙江海口……”(《清高宗实录》乾隆二十二年十一月戊戌)。换句话说,从1757年开始,中国与英国、葡萄牙、荷兰、西班牙等欧洲诸国的贸易往来全部集中到了广东,而不得入江苏、浙江和福建的另外三个海关。⁹这个决定主要出于财政税收的考虑,意图迫使欧洲商人在远离茶、丝产地(如江、浙、闽)的广东进行交易(刘军,2012),这就使得广州成为近两个世纪里中国与欧洲保持较直接接触的几乎唯一的一个大城市。这一局面,直到1842年签订《南京条约》,清政府被迫实行“五口通商”才有改变。此外,上海也是一个重要例证。上海在20世纪初就成为东亚最繁荣的港口和经济、金融中心,近代亚洲唯一的国际化大都市,并在1930—1948年期间成为当时中国国际知名度最高的城市。这主要是因为上海自1843年开埠伊始,就享有完全独立的行政权、司法权和自由贸易权,这为金融业和贸易业提供了难得的发展机遇,促进了其在近代的急速繁荣。实际上,我们注意到:有对外贸易交流史的港口城市、五口通商的开放城市全部名列国际知名度排行榜单的前20位,这些城市同时也是中国东部沿海地区比较发达的城市。据此可以看出,一个城市如果能在中外经济交流中占据枢纽地位,其影响力甚至会超过首都这样的政治身份给城市带来的影响力。

2. 特征之二:局部“波峰”受到重大政治历史事件的影响

城市的国际知名度也与朝代的更迭、政治中心的转换和历史性政治事件的冲击密切相关。仔细观察各城市词频曲线可以发现:曲线的高处和低处总伴有尖锐的起伏和波动,这些波峰与波谷的出现往往与城市历史上的“政治大事件”完美契合。以北京为例,北京是中国近600年来的政治中心,是明、清两代和新中国的首都,正因如此,北京的词频曲线的五个高峰全部和重大政治事件密切相关,包括1840年鸦片战争、1901年《辛丑条约》签订、1927年北伐战争、1937年卢沟桥事变和1949年新中国成立。

再如,南京和重庆分别是中华民国的首都和陪都,二者在20世纪

9. 清代粤海关虽有口岸多处,但总关设于广州。

中期词频曲线的起伏恰好是相反的,这和国民政府的两次迁都有关:只要成为全国政治中心,该城市的国际知名度就开始迅速上升,反之则迅速下降。这清楚地表明,城市国际知名度的变化,尤其是波峰波谷的出现,直接受到国家政治大事件的影响。即便是政治属性看似并不浓厚的香港,政治事件同样影响到国际知名度,比如1997年香港回归就将香港的国际关注度带到了顶峰。而延安由于其政治地位的特殊性,它在50年平均国际知名度排名中名列第14位,这是中国地级市在全体城市中所能达到的最高排名。

3. 特征之三:变迁具有一定的生命周期性

从时间的角度看,城市国际知名度的获得也要经历显著的生命周期,即:蓄势—兴起—发展—峰值四个阶段。从图1的曲线斜率可以看出,快速兴起期一般需历时20—40年。以香港为例,自1842年《南京条约》划香港为英国殖民地始,英国不仅培养和储备了大量熟知西方技术和文化的人才,还创造了大量的就业机会,这是香港发展的蓄势阶段。催化香港兴起的是1948年上海解放,大批内地资本和技术移至香港,给香港注入了巨额资金和大量人才。1949年之后一直到改革开放前,香港作为当时遭到孤立的封闭中国与外部世界联系的唯一窗口,迅速进入了经济繁荣期,很快成为“亚洲四小龙”之一。1997年香港回归,将其国际知名度引向顶峰。

为何一座城市从蓄势到繁荣至少需要20—40年?我们认为,一是因为城市繁荣必须建立在人力、财力、物力以及信息资源高度富裕的基础上,城市发展必须首先提升自己的造血功能,即以自身建设促进经济社会发展,这一过程至少需要20—40年;二是因为城市地位的浮沉往往与政局变动紧密关联,而政治权力的更替、交接或发展周期本身恰恰也在20—40年左右。在周遭条件具备的情况下,城市可以在较短时间内建立起较高的知名度,而在不具备条件的情况下,则需更多时间甚至错过一轮积累知名度的机遇。

4. 特征之四:变化具有地缘联动特征

从地域的视角看,城市国际知名度的波动和变化具有地域联动性。典型的例子发生在北京—天津、上海—南京两对城市之间。北京长期以来是中国的政治中心,仅在20世纪上半叶就经历了清政府、北洋政府、国民政府和中华人民共和国政府四个政权。当20世纪30年代国家政治

中心转移到北京之外时,北京的国际知名度应声而降,而当新中国定都北京后,它的知名度又随即直线回升。在中国,政治中心往往也是信息和资源的交汇中心。作为与政治中心北京地缘最近的大城市,天津在外界看来理所当然地拥有了千载难逢的区位、信息和资金优势,而当政治中心离开北京时,一批具有投机性质的信息和资源也会随之抽离,相应地,对天津的关注度也会随之下降。同样,直至上海在 20 世纪初崛起以及国民党定都南京的双重推动,南京才依托于国家政治中心和上海的金融辐射和产业拉动,开始了国际知名度的快速崛起。而在抗战期间以及 1949 年之后,由于国家政治、经济中心的迁移和政治生态的影响,无论是上海还是南京都承受了大量金融资本和人力资本的外流压力,因此不难发现,南京和上海在 1949 年之后国际知名度的下降曲线是非常接近的。

5. 特征之五:与城市历史、人口和经济要素不完全相关

无论是在分析排名前 20 强的其他城市时,还是在后续的时间序列分析中,我们都发现,尽管不少城市的国际知名度与城市规模以及人文要素相关,但这一关联并不十分紧密。在港澳台和大陆城市之间,这一差异非常明显。更重要的是,同样在大陆城市中,有些城市的国际知名度与其自身历史、当代人口及经济规模等我们习惯用来衡量城市的标准并无明显相关。以西安和拉萨为例,虽然西安起源于西周,曾是十朝古都,在中国文化史上的地位极端重要,但它的国际知名度与其悠久的历史并不匹配,直到 20 世纪 70 年代发现秦始皇兵马俑所导致的旅游热点效应之后,西安的排名才开始上升;而拉萨,根据第五次全国人口普查资料,2000 年拉萨的总人口仅有 47.44 万人,人均纯收入仅为 1325 元,但近 300 年来它在城市全球关注度榜单上的排名一直在第 9—11 名之间。我们有理由认为,19 世纪末英法殖民势力渗透入南亚的地缘政治格局以及藏传佛教、旅游圣地等因素,在早期就已经奠定了拉萨的国际知名度。

八、结语

大数据方法的应用在中国社会科学界尚属起步,重要原因之一就是因为大数据难以获得、难以分析。在本文中,我们使用最新版的谷歌图书语料库大数据,对 300 年来的全球数字化英语书籍、报刊进行了关键词检索分析,得到了中国城市国际知名度的量化和排名。更重要的

是,我们进一步利用《纽约时报》数据库获得了城市的国际媒体提及率指标,并将其与基于谷歌书籍的国际知名度指标进行时间序列分析。格兰杰因果测试显示了中国城市近代以来国际知名度形成的差异化模式,也即:具有较长殖民地历史的城市和一般大陆城市的国际知名度形成具有较大差异,后者往往借助于媒体的报道逐渐进入西方社会。

从方法和数据的角度看,本研究表明:第一,从现有的大数据中提取出相关的变量,是当前利用大数据进行社会科学分析的可行途径;第二,词频统计方法是对非结构化的、非为研究定制的文本大数据进行分析的有力武器,其在社会科学领域的用途还将得到进一步的拓展;第三,除了基本描述与可视化,时间序列分析、面板分析将成为社会科学领域大数据分析的重要方法。当然,本文也存在诸多不足:第一,用词频代表城市国际知名度的尝试需要在今后的研究中得到进一步验证,以确保这是合理、综合与科学、可信的测度;第二,对文本大数据进行检索的精度仍然有待进一步提升;第三,限于时间,本文梳理出的一些现象和规律有可能只是挂一漏万,甚至以偏概全。在今后的研究中,我们会加强对上述问题的研究和解决。

参考文献 (References)

- 陈云松. 2015. 大数据中的百年社会学:基于百万书籍的文化影响力研究[J]. 社会学研究(1): 23—48.
- 陈云松、吴青熹、黄超. 2015. 大数据何以重构社会科学[J]. 新疆师范大学学报(3): 54—61.
- 李红波、曾文、周叶青、李悦铮、江海旭. 2013. 中国沿海地区入境旅游经济的时空差异研究[J]. 中国人口资源与环境(S1): 150—153.
- 刘军. 2012. 明清时期“闭关锁国”问题赘述[J]. 财经问题研究(11): 21—30.
- 马继刚、李飞、周彬学、唐忠明. 2014. 旅游集散地:区位合理性与功能提升——以云南昆明为例[J]. 经济地理(2): 174—179.
- 种海峰. 2010. 简论跨文化传播与冲突的四个规律[J]. 深圳大学学报(人文社会科学版)(11): 149—152.
- 庆桂、董浩,等. 1986. 清高宗实录(550 卷 乾隆二十二年十一月戊戌). 北京:中华书局.
- Chen, Yunsong, and Fei Yan. 2015. “Economic Conditions and Public Concerns about Social Class in Twentieth Century Books.” Working Paper, Stanford University.
- Bentley, R. Alexander, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. “Books Average Previous Decade of Economic Misery.” PLoS ONE 9(1): e83147, DOI:10.1371/journal.pone.0083147.
- Acerbi, Alberto, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. “The Expression of Emotions in 20th Century Books.” PLoS ONE 8(3): e59030, DOI:10.

- 1371/journal.pone.0059030.
- Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman-Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. "Syntactic Annotations for the Google Books Ngram Corpus." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (July): 169–174.
- Kloumann, Isable M., Christopher Danforth, Kameron Decker Harris, Catherine A. Bliss, Peter Sheridan Dodds. 2012. *Positivity of the English Language*. PLoS ONE 7(1): e29484, DOI:10.1371/journal.pone.0029484.
- Mumford, Lewis. 1961. *The City in History: Its Origins, Its Transformations and Its Prospects* CCC. New York: Harcourt Brace Jovanovich Company.
- Nye, Joseph, Jr. 1990. "Soft Power." *Foreign Policy* 80:153–171.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): Pages.
- Zhao, Wayne Xin, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. "Comparing Twitter and Traditional Media Using Topic Models." *Advances in Information Retrieval Lecture Notes in Computer Science* 6611: 338–349.
- Twenge, Jean M., W. Keith Campbell, and Brittany Gentile. 2012. "Increases in Individualistic Words and Phrases in American Books, 1960–2008." PLoS ONE 7(7): e40181.

责任编辑:田 青