中国姓氏人口的历史计量分析:典型特征、决定因素与作用机制*

陈　强　刘春雨　郝　煜

内容提要:中国姓氏人口中的大姓与小姓差异很大,人口高度集中于大姓。本文首次对中国姓氏人口的决定因素进行深入的历史计量分析,使用2012年中国汉族人口排名前500位的姓氏数据(占汉族总人口约99.8%),在大量描述性分析的基础上,进一步通过回归分析揭示了中国姓氏人口分布的典型特征。首先,中国姓氏人口的分布大致服从齐普夫定律,但也有明显偏差。其次,姓氏历史越久远,姓氏作为国姓时间越长,平均而言姓氏人口也越多。这些实证结果通过了一系列稳健性检验,包括控制姓氏的笔画、声调、是否复姓,使用子样本区分统一与分裂政权的国姓,以及针对国姓组与非国姓组进行倾向得分匹配。进一步,我们发现姓氏采用率(以姓氏起源数目与少数民族姓氏人口数为代理变量)与人口迁移率(以姓氏人口的地理集中度为代理变量)是驱动以上结果的两个作用机制。通过东西方比较发现,中国较高的姓氏集中度,其原因可能在于中国历史上的政治稳定性较高,文化延续性较强,而社会流动的制度性障碍较少。

关键词:姓氏人口　齐普夫定律　地理集中度　姓氏历史久远度　国姓

一、引言

中华姓氏文化源远流长,影响深远。根据《说文解字》,"姓,人所生也。……从女、生。生亦声",[①]这说明"姓"的本义是"生"。人们普遍认为,姓最初是代表有共同血缘、血统、血族关系的种族称号,简称族号。《说文解字》所总结的中国上古八大姓(姬、姜、姚、嬴、姒、妘、妫、姞)均带有女字旁,表明中国的姓可能最早起源于母系社会。据载,"黄帝二十五子,其得姓者十四人"。[②]随着人口繁衍,原有部落可能分为若干子部落,而"氏"就是这些子部落的代号。因此,"氏"可视为"姓"的分支。在秦朝统一六国后,不再区分"姓"与"氏",从而演变为我们现在所熟悉的"姓氏"。

随着几千年的历史发展,中国的姓氏数量与姓氏人口也不断增加。徐铁生编著的《中华姓氏源流大辞典》(中华书局2014年版)所收录的汉姓即高达10523个。但是,中国不同姓氏间的人口分布十分不均匀,大量人口集中于少数的大姓,从而"同姓率"(任意两人姓氏相同的概率)远远高于其他欧美国家。[③]就汉族人口而言,根据本文的数据(详见第二部分),2012年前100名的姓氏人口占汉族

[作者简介]　陈强,山东大学经济学院教授,济南,250100,邮箱:qiang2chen2@126.com。刘春雨,招商信诺人寿保险有限公司,深圳,518040,邮箱:liuchunyu0418@126.com。郝煜(通讯作者),北京大学经济学院副教授,北京,100871,邮箱:maxhao1003@pku.edu.cn。

*　陈强感谢马晓婷在收集数据方面的大力协助。郝煜感谢北京大学经济学院种子基金的协助。

①　许慎撰,徐铉等校:《说文解字》,上海古籍出版社2021年版,第406页。

②　《国语·晋语四》进一步记载,"凡黄帝之子,二十五宗,其得姓者十四人为十二姓。姬、酉、祁、己、滕、箴、任、荀、僖、姞、儇、依是也。"参见陈桐生译注:《国语》,中华书局2013年版,第392页。

③　袁义达、张诚、杨焕明:《中国人姓氏群体遗传Ⅱ:姓氏传递的稳定性与地域人群的亲缘关系》,《遗传学报》2000年第7期。

总人口的 87.0%(参见图 1)。而法国前 100 名常见姓氏仅占总人口的 8.1%,美国前 100 名常见姓氏仅占总人口的 16.0%。①

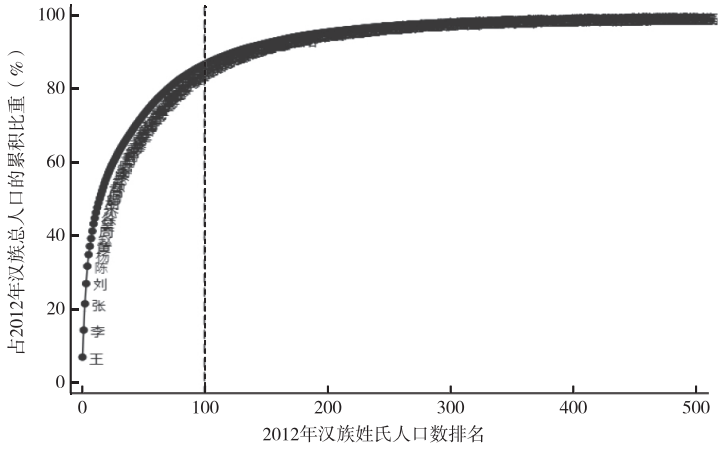

图 1　2012 年前 500 名汉族姓氏人口的累积分布

资料来源:据"全国公民身份证号码查询服务中心(NCIIC)"相关人口数据整理。

究竟哪些因素驱动了中国姓氏间人口的巨大差异?为何中国姓氏人口的集中度(同姓率)远高于欧美国家?对于这些问题的解答无疑可增进对于中国姓氏文化乃至中国文化与历史的理解,一定程度上也可揭示东西方差异的来源。为此,本文使用历史计量方法,首次深入地定量分析中国姓氏人口的典型特征、决定因素与相关作用机制。

在理论上,影响姓氏人口的因素可分为两大类,即生育率与采用率。生育率指某姓氏人口本身的增长率,采用率指原来无姓的人采用某姓或原有姓的人改用某姓。具体而言,从图 1 可见,2012 年中国最大的姓氏依次为王、李、张、刘、陈等,观察这几个姓氏何以成为超级大姓,不难发现以下几个特点:首先,它们的起源时间都很早,其中张姓与刘姓起源于三皇五帝时期,王姓与陈姓起源于商朝,而李姓则起源于周朝。显然,姓氏起源越早,则累积生育率越高,人口数量也会越多。其次,它们中有些建立过中国历史上的统一政权(刘汉、李唐),有些则建立过分裂(非统一)政权(王、张、陈)。一个自然的假设是,作为国姓的姓氏,占有更多的经济和政治资源,其生育率高于总体人口平均水平。② 最后,起源较早的姓氏和作为国姓的姓氏,更有可能被其他姓氏或无姓氏的民众所采用。总之,姓氏起源较早和曾作为国姓都可能对该姓氏的人口有显著的正向影响。当然,姓氏人口的数量可能也受姓氏本身固有特征的影响,例如姓氏的复杂程度(是否复姓、笔画)与声音特性(声调)。

本文使用 2012 年中国汉族人口排名前 500 位的姓氏数据(占汉族总人口约 99.8%),在大量描述性分析的基础上,通过深入的回归分析揭示了中国姓氏人口的以下典型特征:首先,中国姓氏人口的分布大致服从齐普夫定律,但也有明显偏差,人口集中于大姓,但是集中度高于该定律的预测。其次,姓氏诞生朝代越久远,姓氏作为国姓时间越长,则平均而言姓氏人口越多。这些实证结果通过了一系列稳健性检验,包括控制姓氏的笔画、声调、是否复姓,使用子样本区分统一与分裂政权的国姓,以及针对国姓组与非国姓组进行倾向得分匹配。最后,我们发现姓氏采用率(以姓氏起源数目与少

① Yan Liu, et al., "A Study of Surnames in China through Isonymy," *American Journal of Physical Anthropology*, Vol. 148, No. 3, 2012, pp. 341 – 350.

② 李中清、王丰:《人类的四分之一:马尔萨斯的神话与中国的现实(1700—2000)》,陈卫、姚远译,史建云校,生活·读书·新知三联书店 2000 年版。

数民族姓氏人口为代理变量)与人口迁移率(以姓氏人口的地理集中度为代理变量)是驱动以上结果的两大作用机制。作为对比,欧洲的大部分姓氏历史只能追溯到中世纪,"国姓"也没有被大量人口采用,所以姓氏人口的集中度远远低于中国(其原因将在结论部分讨论)。

中国的姓氏研究历史悠久。东汉思想家王符的专著《潜夫论》卷9《志氏姓》一文,或许是研究收纳中国姓氏的开篇之作,但所收姓氏不足500个。形成于北宋初年的《百家姓》,原收集姓氏411个,后增补到504个(其中单姓444个,复姓60个),成为家喻户晓的中国传统蒙学经典。① 后世关于中国姓氏研究的专著日益增多,所收姓氏也不断增长,尤以当代的袁义达、邱家儒编著的《中国姓氏大辞典》(江西人民出版社2010年版)与徐铁生编著的《中华姓氏源流大辞典》为集大成者。其中,《中国姓氏大辞典》收录汉姓10523条,译姓(少数民族姓氏)21050条,译姓演变为汉姓111条,合计31684条,是迄今收录姓氏条目最多的姓氏辞典。目前,传统的姓氏研究着重于每个姓氏的起源、演变、人口迁移、郡望、堂号、历史名人等,多为描述性内容,②关于中国姓氏的系统性定量研究还较少。具有代表性的定量研究有:袁义达、张诚从群体遗传与人口分布的角度对中国姓氏的起源数量、地域集中度、血型分布等指标进行了描述性研究;刘岩等对我国全国、省级、地级市与县级的"同姓率"进行了描述性研究;克拉克(Gregory Clark)等用精英与平民人口的姓氏分布差异随时间变化的速度研究历史上的社会流动性;郝煜采用上述姓氏方法,研究了1645—2010年期间中国的长期社会流动性;陈永伟等计算了中国省份之间的"姓氏距离"作为文化距离的代理变量,讨论了其对于跨省贸易的影响;张慧云计算了县级的姓氏多样性作为宗族多样性的代理变量,讨论了其对于县级经济绩效的影响。③ 以上这些研究都是利用姓氏人口在时空分布上的差异性来度量其他社会经济变量,进而研究其影响和后果。然而,关于中国姓氏人口本身的决定因素,目前还几乎没有定量研究,而此正是本文的研究重点。

本文其余部分安排如下:第二部分为数据介绍与描述性分析,第三部分验证中国姓氏人口的齐普夫定律,第四部分研究姓氏诞生朝代对姓氏人口的影响,第五部分研究国姓对于姓氏人口的影响,第六部分针对国姓组与非国姓组的姓氏进行倾向得分匹配,第七部分探讨国姓与姓氏诞生朝代的作用机制,第八部分为结论。

二、数据说明与典型特征

由于中国姓氏起源和发展的主要推动力来自汉族,而少数民族有很多音译姓氏,且不具有历史追溯性,所以本文仅研究汉族的姓氏人口。如无特别说明,下文的姓氏人口均为汉族姓氏人口。本部分除了介绍数据来源,还进行了较多的描述性分析。由于这是首次针对中国姓氏人口的深入定量分析,故描述性统计本身也较有价值,特别是通过可视化揭示姓氏人口的典型特征。描述性统计的另一好处在于,它并不取决于模型,而回归分析则依赖于模型设定。

---

① 参见《三字经·百家姓·千字文》,上海古籍出版社2017年版。

② 参见陈意浓编著:《中国人姓氏源流分析和归类》,上海三联书店2014年版;钱文忠:《钱文忠解读〈百家姓〉》,凤凰出版传媒股份有限公司、江苏凤凰文艺出版社2014年版;袁义达、邱家儒编著:《中国姓氏大辞典》。其中,《中国人姓氏源流分析和归类》曾以李唐、刘汉为例,说明国姓对于姓氏人口的促进作用,详见下文。

③ 袁义达、张诚:《中国姓氏:群体遗传和人口分布》,华东师范大学出版社2002年版;Yan Liu, et al., "A Study of Surnames in China through Isonymy," *American Journal of Physical Anthropology*, Vol. 148, No. 3, 2012, pp. 341 – 350; Gregory Clark, et al., *The Son Also Rises: 1,000 Years of Social Mobility*, Princeton: Princeton University Press, 2014;郝煜:《中国的姓氏、籍贯和长期代际流动性(1645—2012)》,《经济学(季刊)》2021年第3期;Yongwei Chen, et al., "Cultural Differences and Interprovincial Trades in China: Effect of Surname Distance and its Mechanisms," *Pacific Economic Review*, Vol. 23, No. 4, 2018, pp. 609 – 631; Huiyun Zhang, "The Influence of Clan Surname Diversity on County Economic Development Performance in China: An Empirical Study Based on Chinese Genealogy Data," *Modern Economy*, Vol. 10, No. 4, 2019, pp. 1073 – 1089。

(一)姓氏人口

我们从全国公民身份证号码查询服务中心(NCIIC)①获得了 2012 年汉族人口排名前 500 位的姓氏人口数量,记为变量 *pop*;②并记相应的姓氏人口数排名为变量 *rank*。例如,李姓人口数在 2012 年排名第 2,则其 *rank* 为 2。本文的被解释变量为姓氏人口数的对数,记为 *lnpop*。2012 年排名前 500 名的汉族姓氏人口总数达 12.2359 亿,而根据国家统计局编的《中国统计年鉴·2012》(中国统计出版社 2013 年版),该年汉族总人口约为 12.2593 亿(占全国人口总数的 91.5%)。这意味着排名前 500 位的汉族姓氏人口约占汉族总人口的 99.8%。因此,本文所选取的前 500 个汉族姓氏已具有足够的代表性。

图 2 提供了姓氏人口数(pop)与姓氏人口数排名(*rank*)的散点图。从图 2 可见,少数几个姓氏(比如王、李、张、刘、陈)的人口特别多,故姓氏人口的集中度很高。若仅以汉族姓氏人口数前 500 名计算赫芬达尔指数(HHI),可得

$$HHI_{China} = \sum_{i=1}^{500} (pop\_share_i)^2 = 0.0273 \quad (1)$$

其中,pop_share_i 为第 i 个姓氏人口占前 500 个姓氏总人口的比重。这意味着,在这前 500 个姓氏构成的总人口中,任意选取两人,则他(她)们同姓的概率高达 2.73%。如果这 500 个姓氏的人口均匀分布,则任意两人同姓的概率仅为 1/500,即 0.20%。

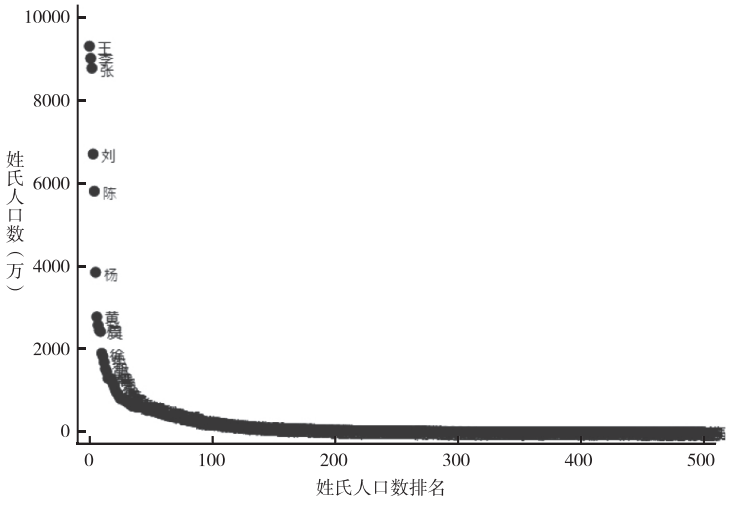

图 2　姓氏人口数与姓氏人口数排名的散点图

资料来源:据"全国公民身份证号码查询服务中心(NCIIC)"相关人口数据整理。

(二)姓氏历史久远度

一般地,姓氏历史越久远(姓氏诞生的朝代越久远),则经过更多年的繁衍生息,且有更多机会被无姓民众或他姓民众改姓时所采用,故姓氏人口通常更多。记姓氏 i 在第 T 年的人口为 $pop_{i,T}$,则可得到以下公式:

① 笔者于 2011 年 11 月走访了该中心,以私人研究者的身份订制和购买了人口数居前 1500 位姓氏的一系列统计变量,包括人口、汉族人口、小学及以下学历人口,中学学历人口、大专学历人口、大学本科学历人口和研究生学历人口系列数据。姓氏数据是从该中心涵盖全体中国人口的身份信息的数据库中提取、计算和生成的。

② 在本研究中,将 1949 年后才开始使用的四个简化字姓氏,合并入原来的繁体字姓氏。其中,"付"并入"傅","代"并入"戴","闫"并入"阎","肖"并入"萧"。

$$pop_{i,T} = pop_{i,0}(1 + \bar{r}_{iT})^{T} \quad (2)$$

其中,$pop_{i,0}$为姓氏 i 在该姓氏诞生之时的期初人口(在此将姓氏 i 诞生之年标准化为第 0 年),而 $\bar{r}_{iT}$ 为姓氏 i 在第 0 年至第 T 年间的人口平均增长率(包括姓氏人口的自然增长,无姓或他姓民众采用姓氏 i,以及姓氏 i 改为他姓)。对方程两边同时取对数可得:

$$\ln pop_{i,T} = \ln pop_{i,0} + T\ln(1 + \bar{r}_{iT}) \quad (3)$$

从方程(3)可见,若给定期初姓氏人口 $pop_{i,0}$与姓氏人口的平均增长率 $\bar{r}_{iT}$,则期末姓氏人口 $pop_{i,T}$ 仅取决于该姓氏存在的时间 T。为了计算每个姓氏的 T,在理想状况下,我们希望知道姓氏诞生于哪一年,但这显然不可能,对于许多姓氏,甚至无法确定它们起源的具体朝代。为此,本文根据徐铁生编著的《中华姓氏源流大辞典》将姓氏起源时间划分为五个时期,分别为夏朝之前(即三皇五帝时期)、夏朝、商朝、周朝、周朝之后,并设置相应的虚拟变量 *prexia*、*xia*、*shang*、*zhou*、*postzhou*,取值均为 0 或 1。例如,王姓起源于商朝,则 $prexia = 0$,$xia = 0$,$shang = 1$,$zhou = 0$,$postzhou = 0$。

图 3 给出了姓氏诞生朝代的分布柱状图。从图 3 可见,在前 500 个姓氏中,有 232 个姓氏诞生于周朝(占 46.4%),这是姓氏诞生的黄金期。其次为周朝之后的朝代,以及周朝之前的商朝,分别诞生了 110 与 109 个姓氏(分别占 22.0% 与 21.8%)。而夏朝与夏朝之前,则仅分别诞生了 22 与 27 个姓氏(分别占 4.4% 与 5.4%),这是姓氏诞生的萌芽期。

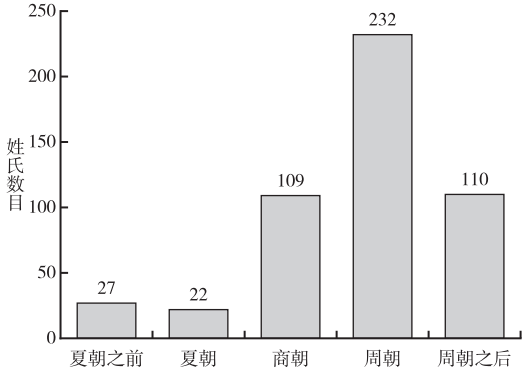

图 3　姓氏诞生朝代的分布

资料来源:据徐铁生编著的《中华姓氏源流大辞典》相关内容整理。

为了直观地考察姓氏历史久远度对于姓氏人口的影响,图 4 提供了区分姓氏诞生朝代的姓氏人口数对数箱形图。从图 4 可见,起源朝代越早的姓氏,其姓氏人口数对数的中位数(箱体中的横线)越大;而随着起源朝代离当今越近,此中位数依次递减。在图 4 中,也看到两个离群的极端值,分别为诞生于周朝的李姓与诞生于周朝之后的邱姓。其中,李姓虽然诞生于周朝(而非更为古老的夏、商或夏朝之前),却成为中国第二大姓,这或许在很大程度上得益于近 300 年的李唐王朝(详见下文讨论)。另一极端值为邱姓,虽然排名第 73 位,但仅诞生于清朝,因为"清雍正三年诏避孔子名讳,改丘姓为邱姓"。① 而未改姓的丘姓,本起源于西周(因姜太公封于营丘,子孙以丘为氏),排名第 237 位。

(三)姓氏是否曾为国姓及年限

在中国历史的长河中,有些姓氏建立过政权,在其统治期间则为"国姓"。国姓的生育率一般更高,而且可能更多人愿意采用国姓。为此,定义虚拟变量 *royal_dummy*,如果该姓曾建立过政权,取值

① 徐铁生:《中华姓氏源流大辞典》,第 440 页。

为 1;反之,则取值为 0。① 在前 500 个姓氏中,只有 33 个姓氏曾作为国姓,占 6.6%。为了直观考察国姓对于姓氏人口的影响,图 5 提供了区分是否国姓的姓氏人口数对数箱形图。从图 5 可见,国姓的姓氏人口分布(比如中位数)明显高于非国姓。特别地,在国姓组箱形图的底部存在一个极端值,为元姓。虽然元姓曾为北魏、东魏与西魏的国姓(合计 103 年),但人口仅 8.1 万,排名第 385 名。②

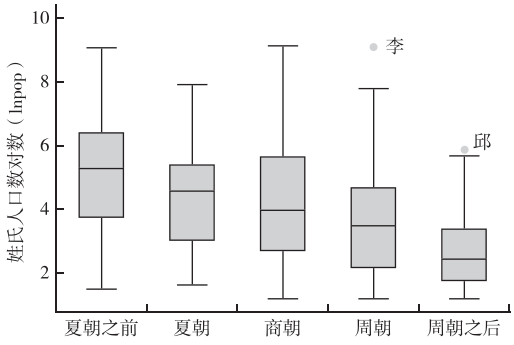

图 4　区分姓氏诞生朝代的姓氏人口数对数箱形图

资料来源:人口数据来自“全国公民身份证号码查询服务中心(NCIIC)”;姓氏诞生朝代信息来自徐铁生编著的《中华姓氏源流大辞典》。

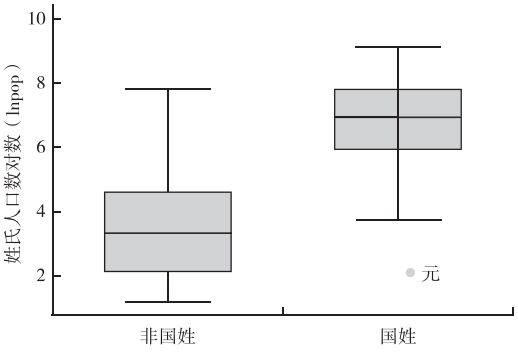

图 5　区分是否国姓的姓氏人口数对数箱形图

资料来源:人口数据来自“全国公民身份证号码查询服务中心(NCIIC)”;姓氏诞生朝代信息来自徐铁生编著的《中华姓氏源流大辞典》。

进一步,可将中国的历史政权分为统一政权与分裂(非统一)政权。在本文,我们将九个朝代视为统一政权,即秦、汉、晋、隋、唐(含武周)、宋、元、明、清。③ 由于统一政权的国姓之影响力可能大于分裂政权的国姓,故定义虚拟变量 *royal_u_dummy*,如果该姓曾建立过统一政权,取值为 1;反之,则取值为 0。在前 500 个姓氏中,只有 6 个姓氏曾作为统一政权的国姓(即刘、杨、李、武、赵、朱,以下简称统一国姓),仅占 1.2%。类似地,定义虚拟变量 *royal_d_dummy*,如果该姓曾建立过分裂(非统一)政权(以下简称分裂国姓),取值为 1;反之,则取值为 0。在前 500 个姓氏中,共有 31 个姓氏曾作为分裂政权的国姓,占 6.2%。

更细致地,可将姓氏分为以下四类,即非国姓、仅为分裂国姓、仅为统一国姓,以及同时为统一与

① 在计算历史政权时,不包括春秋战国时期的诸侯国。西周与东周也仅算一个朝代,即周朝。

② 或许因为元姓为少数民族(鲜卑族)统治者的国姓,故汉族民众较少采用。元姓本为起源于周朝的汉姓,北魏孝文帝时,改帝室拓跋氏为元氏,隋以后鲜卑族融入汉族,不复见于史书。

③ 历史学家对于应将哪些中国王朝视为统一王朝尚无定论,参见陈强:《自然灾害、民族多样性与国家规模:两千年中国历史的统一与分裂》,山东大学工作论文,2021 年。

分裂国姓。图 6 提供了这四个姓氏类别的分布柱状图。从图 6 可见,多达 467 个姓氏为非国姓(占 93.4%),27 个姓氏仅为分裂国姓(占 5.4%),2 个姓氏仅为统一国姓(即武、赵,占 0.4%),而 4 个姓氏同时建立过统一政权与分裂政权(即刘、杨、李、朱,占 0.8%)。

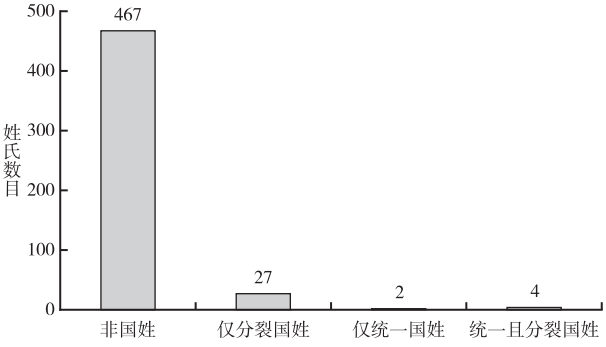

图 6　是否(统一/分裂)国姓的分布

资料来源:据方诗铭编著的《中国历史纪年表(新修订本)》(上海书店出版社 2013 年版)相关历代帝王姓氏内容整理。

针对以上的姓氏四个类别,可作姓氏人口数对数的箱形图,参见图 7。从图 7 可见,非国姓的姓氏人口数对数中位数最低,仅为分裂国姓或仅为统一国姓的姓氏人口数对数中位数的一半,而同时为统一与分裂国姓的姓氏人口数对数中位数最高(叠加了统一国姓与分裂国姓的效应之和)。另外,在仅为分裂国姓组,可再次看到"元姓"为位于箱形图底部的极端值。

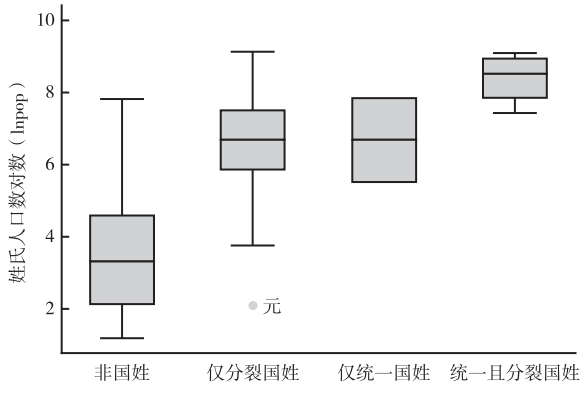

图 7　区分(统一/分裂)国姓的姓氏人口数对数箱形图

资料来源:据方诗铭编著的《中国历史纪年表(新修订本)》相关历代帝王姓氏内容整理。

当然,国姓的影响力也可能与其作为国姓的年限有关。为此,定义变量 *royal* 表示该姓氏所建立政权的存在时间(年数)。进一步,分别定义变量 *royal_u* 与 *royal_d* 为该姓氏所建立统一政权与分裂政权的存在时间(年数)。例如,杨姓既曾建立过统一政权隋朝(589—618,共 30 年),也曾建立过分裂政权,即五代十国的吴朝(902—937,共 36 年),故杨姓的 $royal\_dummy = 1$,$royal\_u\_dummy = 1$,$royal\_d\_dummy = 1$,$royal = 66$,$royal\_u = 30$,而 $royal\_d = 36$。①

为了直观地考察国姓年限与姓氏人口的关系,图 8 提供了姓氏人口数对数(*lnpop*)与国姓年限(*royal*)的散点图及线性拟合线。二者的相关系数为 0.27,且在 1% 水平上显著。在图 8 的最左边,国姓年限(*royal*)有大量的零值,即从未成为国姓的姓氏。在图 8 的右下方,有一个离群值为姬姓。

① 此处使用方诗铭编著的《中国历史纪年表(新修订本)》来识别中国的历史政权及其起讫年代。

姬姓起源于夏朝之前(相传出自黄帝之后),曾作为周朝的国姓达 790 年之久,但人口仅 42.3 万,排名第 216。这是因为,作为上古八大姓之一,从姬姓中分离出当今的许多姓氏,包括周、吴、郑、曹、魏等大姓。① 在图 8 的左上方,则为相反的情形,即虽然作为国姓的年限不长,但却人口众多,包括王姓(作为国姓仅 91 年,但人口排名第 1)、张姓(作为国姓仅 67 年,但人口排名第 3)、陈姓(作为国姓仅 45 年,但人口排名第 5)、杨姓(作为国姓仅 66 年,但人口排名第 6)等。

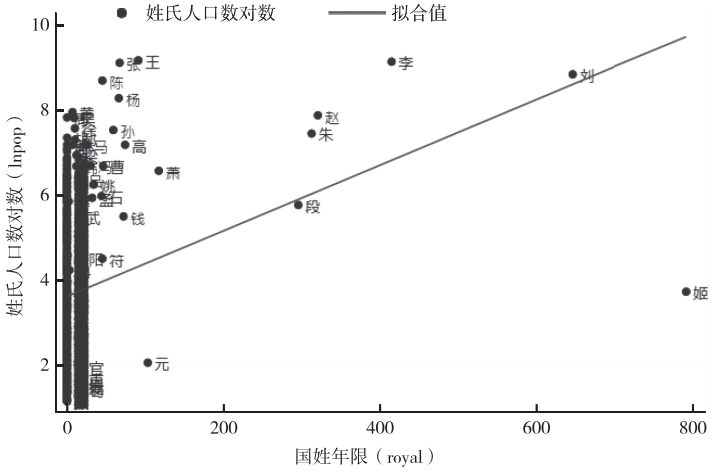

图 8　姓氏人口数对数与国姓年限的散点图

资料来源:人口数据来自“全国公民身份证号码查询服务中心(NCIIC)”;国姓年限据方诗铭编著的《中国历史纪年表(新修订本)》相关内容计算整理。

(四)姓氏是否为复姓

一个姓氏的复杂程度也可能影响民众对该姓氏的采用率。本文主要从两方面度量姓氏的复杂程度,即姓氏是否为复姓,以及姓氏笔画。显然,复姓比单姓的复杂程度更高。为此,定义虚拟变量 *compound*,如果该姓为复姓,取值为 1;反之,则取值为 0。在前 500 个姓氏中,只有 6 个复姓(仅占 1.2%),依次分别为欧阳(排名第 162)、上官(排名第 399)、皇甫(排名第 432)、令狐(排名第 447)、司徒(排名第 454),以及诸葛(排名第 465)。图 9 提供了区分单复姓的姓氏人口数对数箱形图。从图 9 可见,单姓的姓氏人口数对数中位数明显高于复姓。这意味着汉族民众似乎对于单姓有着较为强烈的偏好。即使复姓中人口第一的欧阳,也仅列第 162 位,而在图 9 的复姓组中,复姓欧阳表现为离群的极端值,其姓氏人口远超其他复姓。

(五)姓氏的笔画数

作为对姓氏复杂程度的另一度量,我们将姓氏繁体字写法的笔画数,记为变量 *stroke*。在具体计算时,先找出姓氏繁体字的写法,然后确认其笔画数。例如张姓,其繁体字写法为“張”,笔画数为 11,故张姓的 *stroke* 为 11。图 10 提供了有关姓氏笔画分布的直方图,以及相应的正态分布。从图 10 可见,姓氏笔画 *stroke* 的分布比较接近于正态分布,尽管它的分布有些右偏(右边有较长的尾巴)。图 11 提供了姓氏人口数对数与姓氏笔画的散点图。从图 11 可见,笔画最多的两个姓氏均为复姓,即诸葛与欧阳;而笔画最少的姓氏为丁与刁。然而,姓氏笔画(*stroke*)对于姓氏人口数对数(*lnpop*)几乎没有影响(相关系数仅为 0.051,不显著),尽管线性拟合线的斜率轻微为正。这意味着汉族民众似乎对于姓氏笔画没有明显的偏好。在下文的回归分析中,我们将姓氏笔画(*stroke*)作为控制变量。

① 据史料记载,西周初年分封诸侯时,姬姓子孙的封国最多,达到 53 个。但姬姓的王族后代多以被封之地为姓氏(即所谓“胙土命氏”),这在某种意义上使得姬姓成为孕育其他姓氏的“母姓”,姬姓人口本身反而不多了。另外,由于周朝的宗法制度比较森严,故大部分百姓没有姓氏,或者采用所在地区贵族的姓氏,而姬姓又是皇家独占的姓氏,故采用率不高。

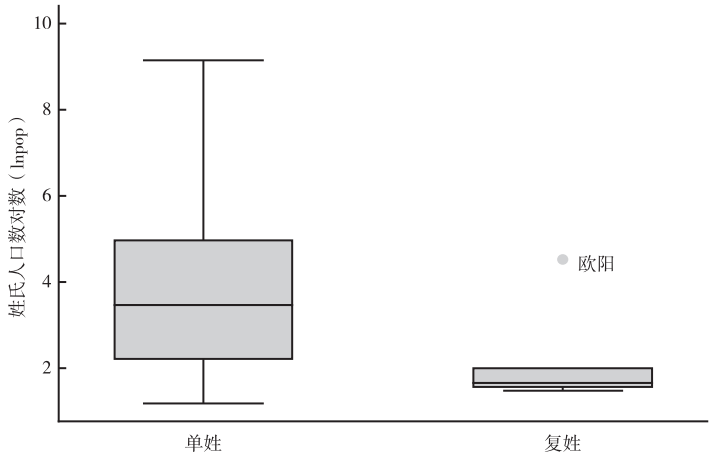

图 9　区分单复姓的姓氏人口数对数箱形图

资料来源:据“全国公民身份证号码查询服务中心(NCIIC)”人口数据整理。

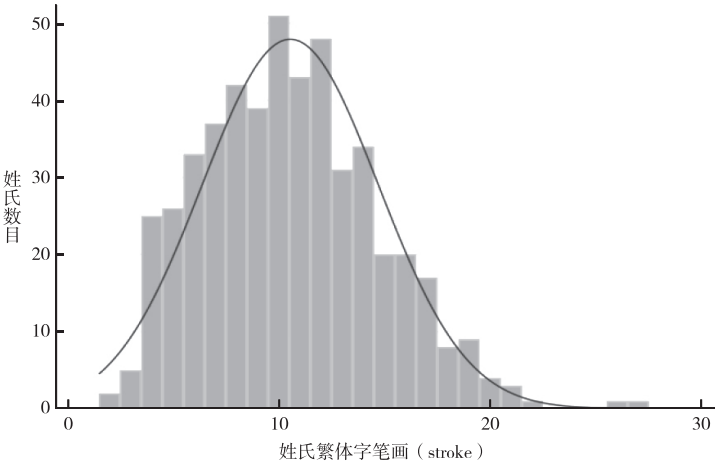

图 10　姓氏繁体字笔画的直方图

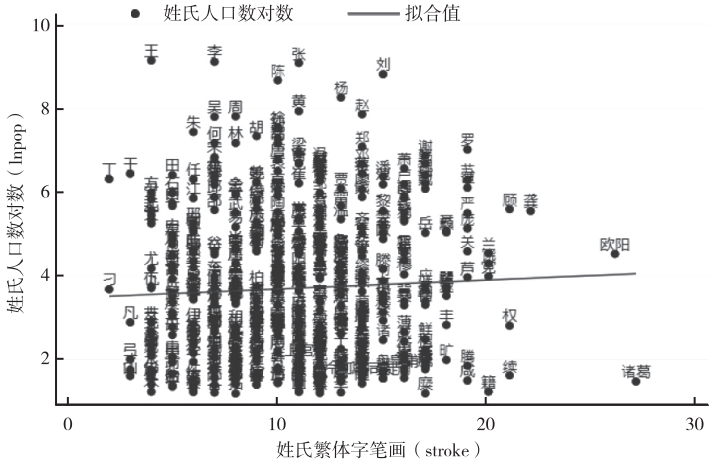

图 11　姓氏人口数对数与姓氏繁体字笔画的散点图

资料来源:据“全国公民身份证号码查询服务中心(NCIIC)”人口数据整理。
说明:图中汉字为简体字,但姓氏笔画根据其繁体字计算。

（六）姓氏的音调

姓氏的声音特性也可能影响姓氏人口。在汉语中，同样的拼音，赋予不同的音调，能够衍生出不同的汉字，这在世界语言中十分少见。具有某个音调的姓氏可能更为悦耳动听，导致采用此姓的人可能更多。为此，我们设置姓氏声调的相应虚拟变量 *tone1*（是否为第一声），*tone2*（是否为第二声），*tone3*（是否为第三声）以及 *tone4*（是否为第四声）。例如，李姓的读音为三声，则其 *tone1* = 0，*tone2* = 0，*tone3* = 1，而 *tone4* = 0。姓氏声调的分布柱状图参见图 12。从图 12 可见，最常见的姓氏声调为第二声（共 177 个姓氏），其次为第一声（共 131 个姓氏）与第四声（共 130 个姓氏），而第三声的姓氏最少（仅 62 个姓氏）。

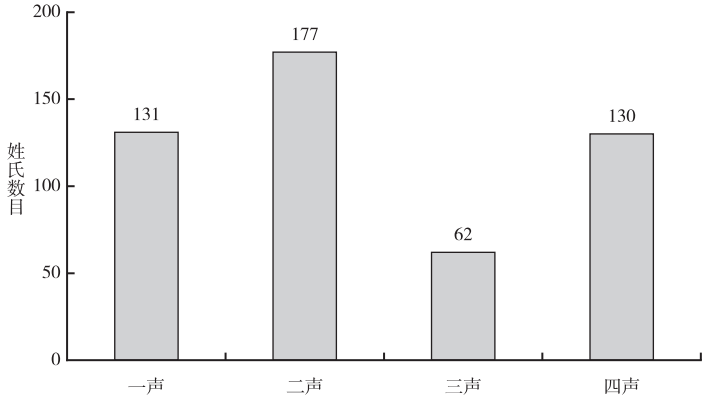

图 12　姓氏声调的分布

为了直观地考察姓氏声调对于姓氏人口的影响，图 13 提供了区分姓氏声调的姓氏人口数对数箱形图。从图 13 可见，不同声调姓氏的姓氏人口数对数中位数比较接近。因此，声调对于姓氏人口似乎也没有影响。另外，在第一声的姓氏中，张姓人口最多，且表现为离群的极端值。在下文的回归分析中，我们将声调（*tone1*，*tone2*，*tone3*）作为控制变量。

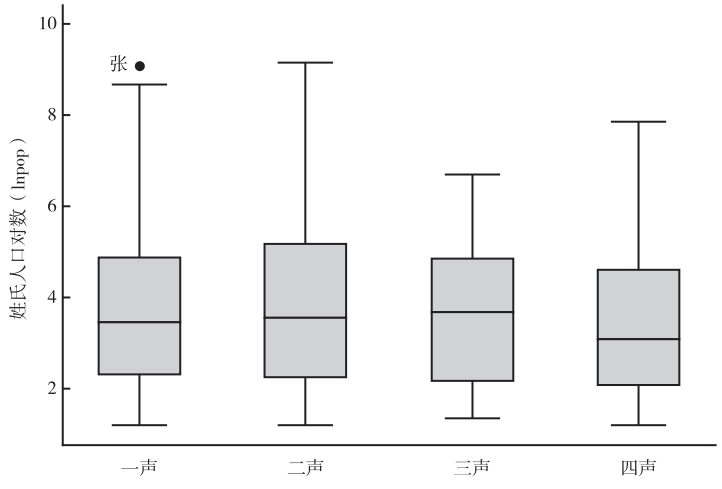

图 13　区分姓氏声调的姓氏人口数对数箱形图

以上介绍了本文所使用的主要变量。表 1 提供了这些变量的基本统计特征。其中，所有变量的观测值均为 500。另外，在第七部分探讨作用机制时，还将引入三个额外的变量（即姓氏起源数目、少数民族姓氏人口、姓氏人口的地理集中度），分别作为姓氏采用率与人口迁移率的代理变量（详见第七部分）。

表 1 **主要变量的统计特征**

变量名称	变量描述	观测值	均值	标准差	最小值	最大值
pop	汉族姓氏人口(万)	500	244.72	872.07	3.28	9290.23
lnpop	汉族姓氏人口的对数	500	3.70	1.77	1.19	9.14
rank	姓氏排名	500	250.5	144.48	1	500
prexia	姓氏是否起源于夏朝之前	500	0.054	0.23	0	1
xia	姓氏是否起源于夏朝	500	0.044	0.21	0	1
shang	姓氏是否起源于商朝	500	0.22	0.41	0	1
zhou	姓氏是否起源于周朝	500	0.46	0.50	0	1
postzhou	姓氏是否起源于周朝之后	500	0.22	0.41	0	1
royal	姓氏作为国姓的时间(年)	500	7.66	55.51	0	790
royal_u	姓氏作为统一政权国姓的时间(年)	500	2.72	29.77	0	426
royal_d	姓氏作为分裂政权国姓的时间(年)	500	4.95	40.75	0	790
royal_dummy	姓氏是否曾为国姓	500	0.07	0.25	0	1
royal_u_dummy	姓氏是否曾为统一政权的国姓	500	0.01	0.11	0	1
royal_d_dummy	姓氏是否曾为非统一政权的国姓	500	0.06	0.24	0	1
compound	姓氏是否为复姓	500	0.01	0.11	0	1
stroke	姓氏繁体笔划数	500	10.55	4.16	2	27
tone1	姓氏是否为一声	500	0.26	0.44	0	1
tone2	姓氏是否为二声	500	0.35	0.48	0	1
tone3	姓氏是否为三声	500	0.12	0.33	0	1
tone4	姓氏是否为四声	500	0.26	0.44	0	1

资料来源:据“全国公民身份证号码查询服务中心(NCIIC)”、方诗铭编著的《中国历史纪年表(新修订本)》、徐铁生编著的《中华姓氏源流大辞典》相关信息内容整理。

三、中国姓氏人口的齐普夫定律

1932 年,哈佛大学语言学家齐普夫在研究英文单词出现频率时,发现如果把单词出现频率按由大到小的顺序排列,则每个单词出现的频率与其频率排名存在反比关系(二者乘积为常数),这被称为“齐普夫定律”。① 它表明在英语单词中,只有极少数的词被经常使用,而绝大多数词很少使用。事实上,包括汉语在内的许多国家语言都有此特点。这个定律后来在很多领域得到验证,包括城市人口、公司规模、网站热度等。② 更一般地,如果离散型随机变量 Y 服从以下“齐普夫分布”,则称变量 Y 满足齐普夫定律:

$$P(Y = k) = \frac{c}{k^{\gamma}}(k = 1,2,\cdots,N;\gamma > 0) \tag{4}$$

其中,k 为排名,而 $P(Y=k)$ 为随机事件“$Y=k$”的发生概率,与 k 的 γ 次幂呈反比,即服从所谓“幂规律”(powerlaw)。参数 $\gamma>0$,而常数 $c=(\sum_{k=1}^{N}k^{-\gamma})^{-1}$ 以保证所有概率之和为 1。如果 $\gamma=1$,则为齐普夫最初所发现的反比关系。将方程两边同时取对数可得:

$$\ln P(Y = k) = \ln c - \gamma \ln k(k = 1,2,\cdots,N) \tag{5}$$

从方程(5)可知,如果将发生频率(或频数)与频率排名都取对数,则二者存在线性的回归关系。因此,为了验证中国姓氏人口是否符合齐普夫定律,我们把 *lnpop* 对 *lnrank* 进行线性回归,所得结果

① 在齐普夫之前,已经有人注意到齐普夫定律的现象。但齐普夫是第一个专门研究,并试图解释此现象的人。George Kingsley Zipf, *Human Behavior and the Principle of Least Effort*, Cambridge: Addison-Wesley Press,1949.

② Xavier Gabaix, “Power Laws in Economics: An Introduction,” *Journal of Economic Perspectives*, Vol. 30, No. 1,2016, pp. 185 – 206.

如下：

$$\widehat{\ln pop} = 12.76 - 1.73 \ln rank$$
$$(0.32)(0.058) \qquad R^2 = 0.918 \qquad (6)$$

其中，括号中为稳健标准误。变量 *lnrank* 的系数估计值为 -1.73（即 $\hat{\gamma} = 1.73$），且在 1% 水平上显著。此回归的拟合优度达到 91.8%。

更直观地，图 14 提供了 *lnpop* 与 *lnrank* 的散点图，以及二者的线性拟合线，也称为“齐普夫图”。从图 14 可见，中国姓氏人口大致服从齐普夫定律，但对于线性拟合线也有明显的偏离。特别地，排名前三的姓氏人口过于相互接近（参见图 2），其中 2012 年汉族王姓人口为 9290.23 万（占总人口 7.59%），李姓人口为 8997.9 万（占总人口 7.35%），而张姓人口为 8762.07 万（占总人口 7.16%）。[①] 在通常满足齐普夫定律的数据中，第 1 名的数量比第 2 名大很多（甚至多达 2 倍），而第 1 名的数量也比第 3 名大很多（甚至多达 3 倍）。由此可见，中国姓氏人口的集中度，高于齐普夫定律的一般预测，导致同姓率较高。

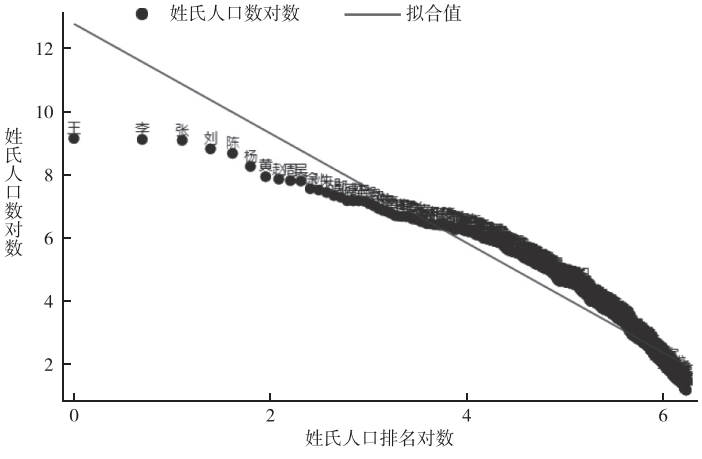

图 14　姓氏人口排名对数与姓氏人口数对数的散点图

资料来源：据“全国公民身份证号码查询服务中心（NCIIC）”相关人口数据整理。

关于数据为何服从齐普夫定律，目前还没有公认的解释。[②] 一种有影响力的假说为“吸引力偏好”，比如，富者容易变得更富，而大城市容易变得更大。以城市人口为例，则人们在选择居住城市时，一般更倾向于首选大城市。类似地，中国姓氏人口大致服从齐普夫定律，这意味着中国人在选择姓氏时，也更倾向于首选大姓。在常人印象中，祖传的姓氏似乎一成不变，很难更改。事实上，在中国历史上，改姓经常发生，而改姓原因则包括避祸[③]、避仇、避讳、避嫌、帝王赐姓[④]、少数民族改为汉姓[⑤]，以及入赘、过继、收养、随母亲姓等。个体在改姓时，则面临姓氏选择问题，此时“吸引力偏好”即可能起作用。例如，帝王赐姓，几乎无一例外都赐予大姓。

① 此处的总人口针对本研究的样本而言，即 2012 年汉族前 500 名姓氏的人口总和。

② 例如，白胜玑（Seung Ki Beak）等试图从人口动力学的角度，解释人口数据服从齐普夫定律的起源。Seung Ki Beak, Hoang Anh Tuan Kiet and Beom Jun Kim, “Family Name Distributions: Master Equation Approach,” *Physical Review E*, Vol. 76, No. 4, 2007, pp. 46 - 113.

③ 如春秋时期陈国的陈完公子，为避祸逃往齐国，并改姓田；其后代取得了齐国的政权，即所谓“田氏代齐”。又比如，司马迁的两个儿子为了避祸而分别改姓冯与同。

④ 如郑和本姓马，因明成祖赐姓而改姓郑，才有以后的“郑和下西洋”之说。

⑤ 如北魏鲜卑皇室拓跋氏在孝文帝改革时改为汉姓元。

另一方面,中国姓氏人口显然还受到其他特殊因素的影响,比如王朝的国姓。以姓氏人口排名第二的李姓为例,其姓氏诞生于周朝,相对而言并不古老。在所有诞生于周朝的姓氏中,李姓之所以能异军突起,成为离群的极端值(参见图 4),显然与近 300 年李唐王朝的强盛有关。陈意浓认为,"首先李姓在唐朝获得了比其他姓氏更为优越的地位与优渥的生存环境;其次唐朝功臣中被赐李姓的不计其数;以及大唐地域的开拓,使很多异族人融入汉姓,当时首选之姓也必然是以李姓居多。"类似地,对于中国第四大姓的刘姓,陈意浓认为,刘姓成为中国的大姓之一,最主要的还是应该归功于历史上长达 400 年的强盛的刘姓汉朝的统治。① 一个合理的猜想是,由于国姓等特殊因素的影响,使得中国姓氏人口虽大致服从齐普夫定律,但也产生了明显的偏离。

四、姓氏历史久远度对姓氏人口的影响

由于姓氏历史久远度(以姓氏诞生朝代为度量)可视为外生变量,故我们首先集中考察姓氏诞生朝代对于姓氏人口的影响。基准回归方程为:

$$\ln pop_i = \beta_0 + \beta_1 prexia_i + \beta_2 xia_i + \beta_3 shang_i + \beta_4 zhou_i + \beta_5 compound + \beta_6 stroke_i + \beta_7 tone1_i + \beta_8 tone2_i + \beta_9 tone3_i + \varepsilon_i \quad (7)$$

其中,被解释变量为姓氏人口数对数(*lnpop*),核心解释变量为姓氏诞生朝代(包括虚拟变量 *prexia*,*xia*,*shang* 与 *zhou*,以未包含的 *postzhou* 为参照系),而控制变量包括是否复姓(*compound*),姓氏笔画(*stroke*),以及声调变量(包括虚拟变量 *tone1*、*tone2*、*tone3*,以未包含的 *tone4* 为参照系)。考虑到国姓变量可能的内生性,故方程(7)并未包括国姓变量,而专门在第五与第六部分考察。显然,由于核心变量"姓氏诞生朝代"为前定变量,而控制变量皆为姓氏本身的固有特征(是否复姓、笔画、声调),故方程(7)中的所有解释变量均可视为外生变量。因此,可使用普通最小二乘法(OLS)一致地估计线性方程(7),②回归结果参见表 2。

表 2　　姓氏诞生朝代对姓氏人口的影响

	被解释变量:*lnpop*			
	(1)	(2)	(3)	(4)
	全样本	全样本	前半样本	前半样本
prexia	2.472***	2.534***	1.295***	1.292***
	(0.379)	(0.373)	(0.374)	(0.368)
xia	1.785***	1.773***	1.075***	1.052***
	(0.388)	(0.386)	(0.345)	(0.345)
shang	1.575***	1.583***	1.100***	1.084***
	(0.215)	(0.215)	(0.232)	(0.232)
zhou	0.986***	0.991***	0.595***	0.592***
	(0.154)	(0.154)	(0.198)	(0.197)
compound	-1.259***	-1.042***	-0.470	-0.512***
	(0.368)	(0.392)	(0.322)	(0.106)
stroke	0.0270		0.00462	
	(0.0180)		(0.0174)	
tone1	0.111		-0.0463	
	(0.198)		(0.217)	

① 陈意浓编著:《中国人姓氏源流分析和归类》,第 61、68 页。

② 在表 2 第(1)列的线性回归之后,我们进行了 Ramsey RESET 检验,结果未发现遗漏非线性项。

续表 2

	被解释变量:*lnpop*			
	(1)	(2)	(3)	(4)
	全样本	全样本	前半样本	前半样本
tone2	0.291		0.171	
	(0.191)		(0.192)	
tone3	0.0955		−0.276	−0.341*
	(0.222)		(0.206)	(0.174)
_cons	2.269***	2.688***	4.322***	4.441***
	(0.238)	(0.111)	(0.290)	(0.174)
样本容量	500	500	250	250
R^2	0.160	0.152	0.099	0.093

说明:括号中为稳健标准误。* 表示 $p<0.1$, ** 表示 $p<0.05$, *** 表示 $p<0.01$。所有回归表格均相同,后文不再重复说明。

表 2 第(1)列汇报了使用全样本与全部解释变量的估计结果。其中,姓氏诞生朝代的虚拟变量(*prexia*,*xia*,*shang*,*zhou*)均在 1% 水平上显著为正,而回归系数则呈现递减的趋势。具体而言,虚拟变量 *prexia*(姓氏起源于夏朝之前)的回归系数为 2.472。这意味着,在给定其他控制变量的情况下,诞生于夏朝之前(*prexia* = 1)的姓氏人口是起源于周朝之后(*postzhou* = 1)的姓氏人口的 $e^{2.472}=11.85$ 倍,因为:

$$\widehat{\ln pop}(prexia=1)-\widehat{\ln pop}(postzhou=1)=\ln\left(\frac{\widehat{pop}(prexia=1)}{\widehat{pop}(postzhou=1)}\right)=2.472 \tag{8}$$

显然,这是一个在经济意义上很显著的效应。进一步,虚拟变量 *xia*(姓氏起源于夏朝)、*shang*(姓氏起源于商朝)与 *zhou*(姓氏起源于周朝)的回归系数分别为 1.785、1.575 与 0.986,这意味着,在给定其他控制变量的情况下,诞生于夏朝、商朝或周朝的姓氏人口数分别为起源于周朝之后姓氏人口的 $e^{1.785}=5.96$ 倍、$e^{1.575}=4.83$ 倍与 $e^{0.986}=2.68$ 倍。更直观地,可画图展示姓氏诞生朝代对于姓氏人口数对数(*lnpop*)的边际效应(回归系数),以及相应的 95% 置信区间,参见图 15。

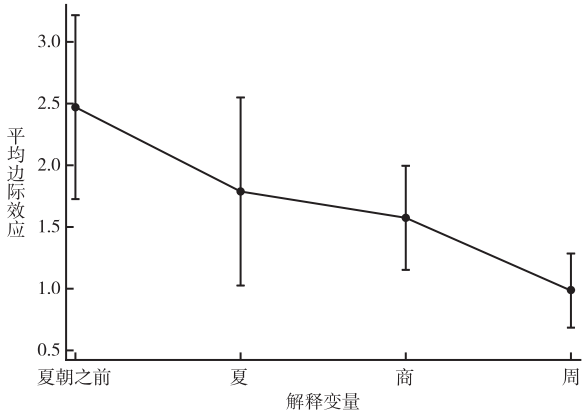

图 15　姓氏诞生朝代对于姓氏人口数对数的边际效应

在控制变量中,虚拟变量 *compound*(是否复姓)在 1% 水平上显著为负,其系数估计值为 −1.259。这意味着,在给定其他变量的情况下,复姓的姓氏人口平均仅为单姓人口的 $e^{-1.259}=28.3\%$。显然,复姓的人口劣势很明显。其他控制变量则均不显著。

表 2 第(2)列汇报了依次去掉不显著变量的回归结果(每次均去掉最不显著的变量,直至所有变量至少在 10% 水平上显著),所得结果与第(1)列类似。作为一个稳健性检验,我们使用排名前 250

位的姓氏子样本,重复第(1)列的回归。所得结果汇报于第(3)列,在性质上依然与第(1)列类似,只是姓氏诞生朝代变量的回归系数变小,而变量 *compound* 变得不显著(可能因为在前 250 个姓氏中只有 1 个复姓,即欧阳,排名第 162 位)。第(4)列依次去掉第(3)列回归的不显著变量,所得结果与第(3)列类似,但变量 *compound* 重新变得显著。

在表 2 中,没有汇报排名后 250 位的姓氏子样本回归结果,因为在后 250 个姓氏中,只有极少数姓氏起源于夏朝或夏朝之前,导致朝代诞生变量变得不显著。具体来说,在后 250 个姓氏中,只有 4 个姓氏起源于夏朝之前(分别为阚、屠、危、咸);而起源于夏朝的姓氏也只有 7 个(分别为封、奚、雍、宿、扈、过、巢)。这些古老的姓氏,之所以未能在人口规模上发扬光大,可能是因为姓氏的汉字本身比较生僻(比如阚、奚),或不太讨喜(比如屠、危)。然而,作为姓氏的汉字,对于生僻或褒贬程度并无客观的度量,故本文未包括这方面的变量,而将其放入回归方程的扰动项中。

五、国姓对姓氏人口的影响

从本节开始,我们考虑国姓对于姓氏人口的影响。首先考察国姓年限(*royal*)对于姓氏人口的作用,其基准回归方程为:

$$\begin{aligned}\ln pop_i = {} & \beta_0 + \beta_1 royal_i + \beta_2 prexia_i + \beta_3 xia_i + \beta_4 shang_i + \beta_5 zhou_i \\ & + \beta_6 compound + \beta_7 stroke_i + \beta_8 tone1_i + \beta_9 tone2_i + \beta_{10} tone3_i + \varepsilon_i \end{aligned} \tag{9}$$

与上节的方程(7)相比,方程(9)增加了关键变量 *royal*(姓氏作为国姓的年限)。若将中国历史政权区分为统一政权与分裂政权(非统一政权),则可进一步将国姓年限细分为统一国姓年限(*royal_u*)与分裂国姓年限(*royal_d*),由此可得如下回归方程:

$$\begin{aligned}\ln pop_i = {} & \beta_0 + \beta_1 royal\_u_i + \beta_2 royal\_d_i + \beta_3 prexia_i + \beta_4 xia_i + \beta_5 shang_i + \beta_6 zhou_i \\ & + \beta_7 compound + \beta_8 stroke_i + \beta_9 tone1_i + \beta_{10} tone2_i + \beta_{11} tone3_i + \varepsilon_i \end{aligned} \tag{10}$$

由于人口众多的大姓在概率意义上也更有机会成为国姓(如假设通过掷骰子选择国姓),故可能存在从姓氏人口到国姓的逆向因果关系,从而导致国姓年限变量(*royal*,*royal_u*,*royal_d*)以及相应的虚拟变量(*royal_dummy*,*royal_u_dummy*,*royal_d_dummy*)为内生变量。然而,究竟哪个姓氏成为一个朝代的国姓,毕竟具有很强的随机性。例如,在样本中,统一王朝的国姓只有 6 个(即刘、杨、李、武、赵、朱),并不包括人口排名分别为第 1、第 3 与第 5 的王姓、张姓与陈姓。反之,排名仅第 91 名的武姓成为武周的统一国姓(持续 15 年),显然与武则天异军突起的偶然性有关。另一方面,由国姓所带来的姓氏人口增长则更为具体而直接,包括皇族的繁衍、帝王赐予功臣国姓、少数民族改国姓(后融入汉族)等。再以明朝的国姓(即朱姓)为例,根据袁义达、张诚的推算,宋朝时期的朱姓人口大约 110 万,约占当时全国人口的 1.4%。① 因此,朱姓成为明朝国姓的概率其实很低,具有极强的偶然性。而在成为国姓之后,明朝的朱姓人口快速增长,达到大约 186 万,约占当时全国人口的 2%。进一步,在朱姓偶然成为国姓之后,其作为国姓的持续年限(变量 *royal*),显然不依赖于朱姓人口的数量,因为毕竟朱姓人口仅占全国人口约 2%,而且朱姓人口也未必全部支持中央政权。因此,可以合理地认为,在国姓与姓氏人口的双向因果关系中,较强的作用方向为从国姓到姓氏人口数量的因果效应,而从姓氏人口数量到国姓的反方向作用则比较微弱(带有较强的偶然性)。尽管如此,依然不可否认国姓变量的潜在内生性。为此,我们在第六部分使用国姓组与非国姓组的倾向得分匹配,作为稳健性检验,并缓解可能的内生性。最后,作为对于姓氏人口的首次历史计量分析,本文带有探索性,故即使仅定量分析国姓与姓氏人口的相关性,也依然较有价值。

① 袁义达、张诚:《中国姓氏:群体遗传和人口分布》,第 233 页。

表 3 汇报了国姓年限影响姓氏人口数量的 OLS 回归结果。其中,第(1)列汇报了对方程(9)的估计结果。国姓年限变量(*royal*)仅在 10% 水平上显著为正,而系数估计值为 0.00606。从图 8 可知,上古八大姓之一的姬姓为离群的极端值(国姓年限高达 790 年,但姓氏人口仅排名第 216),可能对回归系数与显著性有很大影响。鉴于姬姓的极端特殊性(参见第二部分有关国姓年限的讨论),表 3 第(2)列去掉了姬姓的观测值(样本容量变为 499),这使国姓年限变量(*royal*)变得在 1% 水平上显著为正。进一步,变量 *royal* 的系数估计值增加到 0.0109。这意味,国姓年限每增加一年,则姓氏人口平均将增加 1.09%,这是一个经济意义比较显著的效应。无论在第(1)列或第(2)列,姓氏诞生朝代变量(*prexia*,*xia*,*shang*,*zhou*)均在 1% 水平上显著为正,且相应的系数估计值依次递减,而复姓变量(*compound*)在 1% 水平上显著为负,其余控制变量则不显著。

表 3　　**国姓年限对姓氏人口数量的影响**

	被解释变量:*lnpop*			
	(1)	(2)	(3)	(4)
	国姓年限	国姓年限(去掉姬姓)	统一/分裂国姓年限	统一/分裂国姓年限(去掉姬姓)
royal	0.00606 *	0.0109 ***		
	(0.00314)	(0.00241)		
royal_u			0.0113 ***	0.00658 *
			(0.00233)	(0.00351)
royal_d			0.00293	0.0181 ***
			(0.00345)	(0.00654)
prexia	2.038 ***	2.040 ***	2.032 ***	2.048 ***
	(0.360)	(0.345)	(0.360)	(0.341)
xia	1.773 ***	1.753 ***	1.777 ***	1.735 ***
	(0.384)	(0.382)	(0.387)	(0.378)
shang	1.557 ***	1.538 ***	1.563 ***	1.519 ***
	(0.212)	(0.209)	(0.213)	(0.207)
zhou	0.950 ***	0.919 ***	0.948 ***	0.904 ***
	(0.152)	(0.151)	(0.152)	(0.151)
compound	−1.241 ***	−1.232 ***	−1.239 ***	−1.228 ***
	(0.368)	(0.368)	(0.369)	(0.367)
stroke	0.0284	0.0281	0.0276	0.0289 *
	(0.0175)	(0.0175)	(0.0175)	(0.0174)
tone1	0.0892	0.131	0.111	0.126
	(0.194)	(0.191)	(0.193)	(0.191)
tone2	0.280	0.257	0.279	0.245
	(0.185)	(0.183)	(0.185)	(0.185)
tone3	0.124	0.145	0.128	0.152
	(0.222)	(0.221)	(0.222)	(0.221)
_cons	2.259 ***	2.257 ***	2.261 ***	2.252 ***
	(0.234)	(0.233)	(0.235)	(0.233)
样本容量	500	499	500	499
R^2	0.193	0.228	0.203	0.235

表 3 第(3)列汇报了对方程(10)的估计结果,将国姓年限进一步细分为统一国姓年限(*royal_u*)与分裂国姓年限(*royal_d*)。结果显示,统一国姓年限(*royal_u*)在 1% 水平上显著为正,其回归系数为 0.0113。这意味着,每增加一年作为统一国姓的年限,则姓氏人口平均将增加 1.13%。然而,分裂国姓年限(*royal_d*)并不显著,且系数估计值仅为 0.00293。当然,由于姬姓被作为分裂国姓,故此结

果也可能因为姬姓的极端值所致。为此,第(4)列重复第(3)列的回归,但去掉了姬姓的离群观测值。结果显示,分裂国姓年限(*royal_d*)变得在1%水平上显著为正,且系数估计值大幅上升至0.0181。然而,去掉姬姓之后,统一国姓年限(*royal_u*)变得仅在10%水平上显著为正。这或许是因为统一国姓的观测值太少,仅有6个(刘、杨、李、武、赵、朱),导致对于统一国姓的效应估计不够准确。另外,在6个统一国姓中,有4个大姓同时也是分裂国姓(刘、杨、李、朱),这使得在回归结果中,不易区分统一国姓的单独贡献。

进一步,我们将国姓设为虚拟变量,考察是否国姓(*royal_dummy*),以及是否统一国姓(*royal_u_dummy*)与是否分裂国姓(*royal_d_dummy*)对于姓氏人口数量的影响,回归结果汇报于表4。表4第(1)列以虚拟变量 *royal_dummy* 替换方程(9)中的国姓年限变量 *royal*。其中,国姓虚拟变量 *royal_dummy* 在1%水平上显著为正,其系数估计值为2.844。这意味着,在给定其他变量的情况下,作为国姓的姓氏人口数平均为非国姓姓氏人口的 $e^{2.844}=17.18$ 倍,这是一个经济意义上非常显著的效应。姓氏诞生朝代变量(*prexia*,*xia*,*shang*,*zhou*)仍在1%水平上显著为正,且系数估计值依次递减。复姓变量 *compound* 依然在1%水平上显著为负。姓氏笔画(*stroke*)在10%水平上显著为正,这意味着民众或许偏好笔画较多的姓氏(但王姓显然是反例)。所有声调变量(*tone1*,*tone2*,*tone3*)则均不显著。第(2)列依次去掉了第(1)列回归中的不显著变量,所得结果与第(1)列类似。

表4　　是否国姓对姓氏人口数量的影响

	被解释变量:*lnpop*			
	(1)	(2)	(3)	(4)
	国姓	国姓	统一/分裂国姓	统一/分裂国姓
royal_dummy	2.844***	2.862***		
	(0.286)	(0.285)		
royal_u_dummy			2.090***	2.091***
			(0.390)	(0.386)
royal_d_dummy			2.600***	2.616***
			(0.295)	(0.292)
prexia	1.757***	1.761***	1.765***	1.767***
	(0.335)	(0.334)	(0.332)	(0.330)
xia	1.517***	1.507***	1.539***	1.529***
	(0.342)	(0.340)	(0.343)	(0.341)
shang	1.302***	1.303***	1.313***	1.314***
	(0.199)	(0.198)	(0.198)	(0.197)
zhou	0.830***	0.830***	0.826***	0.825***
	(0.149)	(0.148)	(0.149)	(0.148)
compound	-1.212***	-1.214***	-1.206***	-1.210***
	(0.366)	(0.366)	(0.366)	(0.366)
stroke	0.0334*	0.0335*	0.0334*	0.0334*
	(0.0172)	(0.0171)	(0.0171)	(0.0170)
tone1	0.0181		0.00809	
	(0.183)		(0.184)	
tone2	0.0961		0.0754	
	(0.174)		(0.175)	
tone3	0.0339		0.0441	
	(0.224)		(0.222)	
_cons	2.296***	2.337***	2.305***	2.338***
	(0.228)	(0.203)	(0.228)	(0.202)
样本容量	500	500	500	500
R^2	0.309	0.309	0.317	0.317

表 4 第(3)列以虚拟变量统一国姓 *royal_u_dummy* 与分裂国姓 *royal_d_dummy* 分别替代方程(10)中的统一国姓年限 *royal_u* 与分裂国姓年限 *royal_d*。其中,*royal_u_dummy* 与 *royal_d_dummy* 均在 1% 水平上显著为正,而系数估计值分别为 2.09 与 2.60。统一国姓虚拟变量 *royal_u_dummy* 的回归系数反而小于分裂国姓虚拟变量 *royal_d_dummy* 的回归系数,这似乎有些意外。然而,在样本中,统一国姓只有 6 个(刘、杨、李、武、赵、朱),而分裂国姓则有 31 个,故统一国姓虚拟变量 *royal_u_dummy* 的回归系数估计得更不精确,其标准误(0.390)也大于分裂国姓虚拟变量 *royal_d_dummy* 的标准误(0.295)。另外,在 6 个统一国姓中,有 4 个同时也是分裂国姓(刘、杨、李、朱),这使得回归结果不易区分统一国姓的单独贡献。进一步,在剩余的两个纯粹统一国姓中(武、赵),武姓(武周)作为国姓仅维持了 15 年,而姓氏人口排名 91,这无疑也降低了统一国姓的效应。第(3)列其余变量的显著性与系数大小均与第(1)列类似,第(4)列依次去掉了第(3)列回归中的不显著变量,所得结果仍然类似。

六、国姓组与非国姓组的倾向得分匹配

作为稳健性检验,也为了部分缓解国姓变量(*royal_dummy*)可能的内生性,本节将曾是国姓的样本(*royal_dummy* = 1,简称"国姓组")作为处理组,而将不是国姓的样本(*royal_dummy* = 0,简称"非国姓组")作为控制组,进行倾向得分匹配。为此,先将处理变量是否国姓(*royal_dummy*)对前定变量(*prexia*,*xia*,*shang*,*zhou*;*stroke*;*tone1*,*tone2*,*tone3*)进行 Logit 回归,[①]估计倾向得分,然后针对倾向得分使用近邻法,将处理组的每个姓氏与控制组的姓氏进行匹配。匹配之后,处理组与控制组的姓氏人口数对数的平均差距即为参与者平均处理效应(Average Treatment Effects on the Treated,简记 ATT 或 ATET)。首先,汇报第一阶段 Logit 回归的估计结果(样本容量为 500):

$$\widehat{royal\_dummy} = \Lambda(-18.79 + 17.16prexia + 16.08xia + 16.04shang + 15.44zhou - 0.04stroke + 0.92tone1 + 1.45tone2 + 0.69tone3)$$

$$(0.77)^{***}(0.47)^{***}(0.81)^{***}(0.36)^{***}(0.33)^{***}(0.041)(0.68)(0.65)^{**}(0.81)$$

其中,$\Lambda(\cdot)$为逻辑分布的累积分布函数。结果显示,姓氏诞生朝代变量(*prexia*,*xia*,*shang*,*zhou*)均在 1% 水平上显著为正,声调变量 *tone2* 在 5% 水平上显著为正,而其余协变量则不显著。此 Logit 回归的准 R^2 达到 0.145。另外,使用 Stata 16 的官方命令 teffects psmatch,[②]分别将处理组的每个姓氏与控制组倾向得分最接近的 1—5 个姓氏进行 K 近邻匹配,所得结果参见表 5。

表 5　　**参与者平均处理效应**

	被解释变量:*lnpop*				
	(1)	(2)	(3)	(4)	(5)
	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
royal_dummy	2.90***	2.92***	3.00***	3.13***	3.16***
	(0.42)	(0.33)	(0.33)	(0.33)	(0.31)

说明:括号中为 AI 标准误。

在表 5 的括号中为基于阿巴迪(Abadie)和因本斯(Imbens)的研究提出的 AI(Abadie-Imbens)标准误,这是目前倾向得分匹配唯一正确的标准误。[③] 其中,第(1)列为 1 对 1 匹配,第(2)列为 1 对 2

① 其中,复姓变量 *compound* 未作为协变量放入倾向得分匹配的第一阶段回归,因为"*compound* = 1"(复姓)可以完美地预测倾向得分为 0。

② 使用 Stata 官方命令的最大好处在于,可获得正确的 AI 标准误(详见正文),而流行的非官方命令 psmatch2 所提供的标准误并不正确。

③ Abadie, Alberto and Guido Imbens, "Matching on the Estimated Propensity Score," *Econometrica*, Vol. 84, No. 2, 2016, pp. 781 – 801.

匹配,以此类推。无论使用 1 对 1 乃至 1 对 5 的倾向得分匹配,参与者平均处理效应(ATT)均在 1% 水平上显著为正,且数值稳定在 3 附近(接近于表 4 回归分析的结果)。这意味着,在给定协变量(*prexia*,*xia*,*shang*,*zhou*;*stroke*;*tone1*,*tone2*,*tone3*)的情况下,作为国姓的姓氏人口数平均为非国姓姓氏人口的 $e^3 = 20.09$ 倍,这是一个经济意义上很显著的效应。

当然,以上结果是否可靠,还取决于倾向得分匹配的前提条件是否满足。为此,下面进行诊断性检验。首先,考察倾向得分匹配的"重叠假定"或"匹配假定"是否成立,结果参见图 16。从图 16 可见,非国姓组倾向得分的概率密度大量集中在 0 附近,而国姓组倾向得分的概率密度则集中在 0.1 附近,但二者的概率分布均呈现出多峰形态。① 尽管如此,非国姓组与国姓组倾向得分的概率分布依然有很多交叠部分,故满足重叠假定。

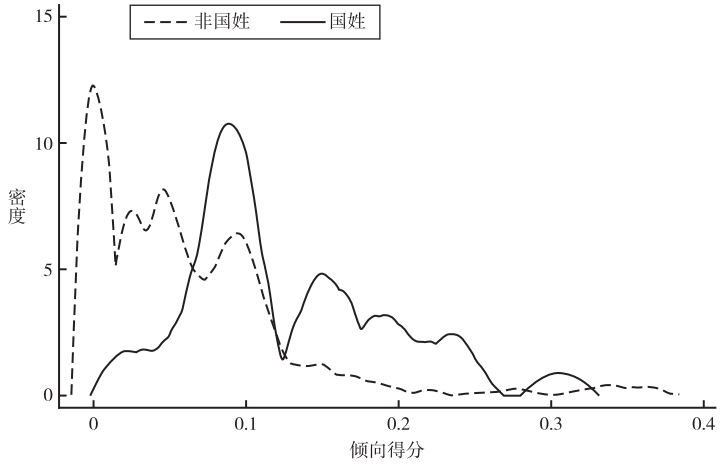

图 16　国姓组与非国姓倾向得分的概率密度图

其次,考察匹配前后国姓组与非国姓组倾向得分的箱形"平衡图",参见图 17。从图 17 可见,在匹配前(即图中左列"原始样本"),国姓组倾向得分的中位数明显高于非国姓组,而在匹配后(即图中右列"匹配样本"),国姓组与非国姓组倾向得分的中位数几乎没有区别。

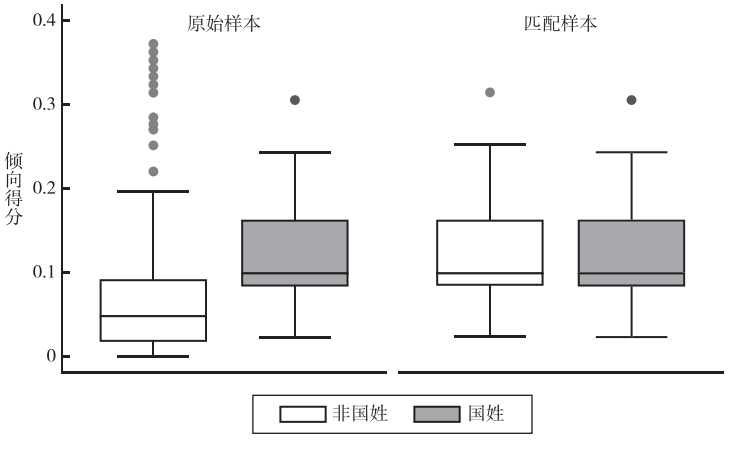

图 17　匹配前后国姓组与非国姓组倾向得分的箱形图

进一步,考察匹配前后国姓组与非国姓组倾向得分的核密度平衡图,参见图 18。从图 18 可见,在匹配之前(即图中左列"原始样本"),国姓组倾向得分的核密度图明显分布在非国姓组核密度的右边(倾

① 在进行核密度估计时,使用了二次核。使用其他核函数的估计结果类似,在此从略。

向得分的高峰右移约 0.1),而在匹配之后(即图中右列“匹配样本”),国姓组与非国姓组倾向得分的核密度图几乎重合。这表明,经过倾向得分匹配之后,国姓组与非国姓组的倾向得分之分布基本无区别。由此可见,本部分所使用的倾向得分匹配通过了所有常规的诊断性检验,故所得结果较为可信。

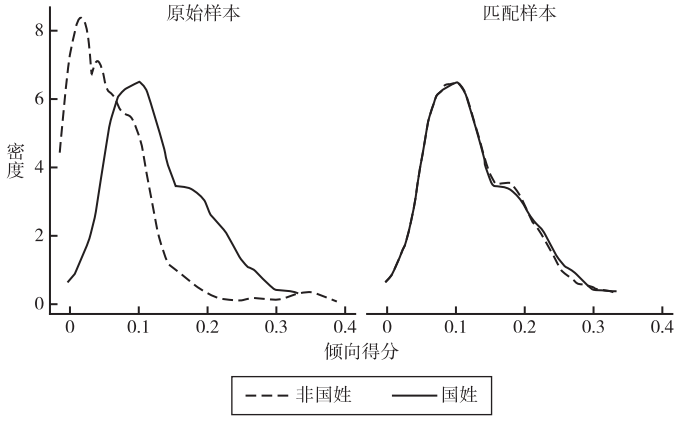

图 18　匹配前后国姓组与非国姓组倾向得分的核密度图

七、作用机制探讨

本部分探讨国姓以及姓氏诞生朝代对于姓氏人口的作用机制,着重于姓氏采用率与人口迁移率两个方面。如果一个姓氏为一王朝的国姓(无论统一或分裂政权),则该姓氏通常更容易被其他姓氏民众或少数民族所采用,或者由皇帝赐姓。进一步,作为国姓,由于王族常被分封或派遣到全国各地,故其后代更容易开枝散叶,这无疑会增大国姓的人口迁移率,降低国姓人口的地理集中度。另一方面,如果一个姓氏诞生于更早的朝代,则显然有更多机会被无姓或他姓民众所采用,也同样更可能迁移到全国各地,从而降低姓氏人口的地理集中度。对于姓氏采用率,我们使用“姓氏起源数目”与“少数民族姓氏人口数”作为代理变量。对于姓氏人口的迁移率,则使用“姓氏人口的地理集中度”作为其(反向)代理变量。在下文中,我们统称这三个变量为“机制变量”。

(一)姓氏起源数目

徐铁生编著的《中华姓氏源流大辞典》记载了每个姓氏的不同来源。计算每个姓氏有记载的起源数目,即可得到变量“姓氏起源数目”,[①]记为 *num_origin*。例如,刘姓有 55 个有记载的起源,故其 *num_origin* 为 55。显然,对于一个姓氏而言,除了最早的姓氏起源为原创,其他姓氏起源一般可视为“姓氏采用”。因此,姓氏起源数目可作为姓氏采用率的一个代理变量。图 19 提供了姓氏起源数目的直方图。从图 19 可见,姓氏起源数目的分布呈现出明显的右偏,在右边有一个较长的尾巴。不同姓氏的起源数目差别很大。例如,复姓司徒只有一个起源,即周朝的官职“司徒”,而李姓则有多达 97 个起源。显然,起源比较单一的姓氏,其姓氏采用率较低;反之,起源较多的姓氏,则其姓氏采用率较高。

进一步,通过散点图(参见图 20),考察姓氏人口数对数与姓氏起源数目之间的关系。从图 20 可见,姓氏人口数对数与姓氏起源数目高度正相关(相关系数达 0.81,且在 1% 水平上显著)。姓氏起源数目最多者为李姓(多达 97 个起源),其次为张姓(83 个起源)、王姓(71 个起源)、陈姓(65 个起源)等超级大姓。

① 在样本中,有两个姓氏(即“负”与“圣”)未记载姓氏起源,我们将这两个姓氏的起源数目设为 1。

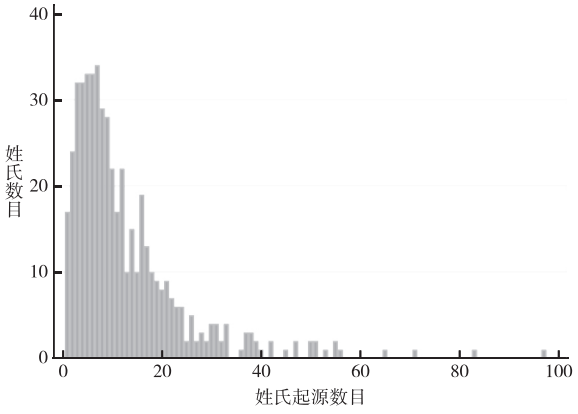

图 19　姓氏起源数目的直方图

资料来源:据徐铁生编著的《中华姓氏源流大辞典》相关内容整理。

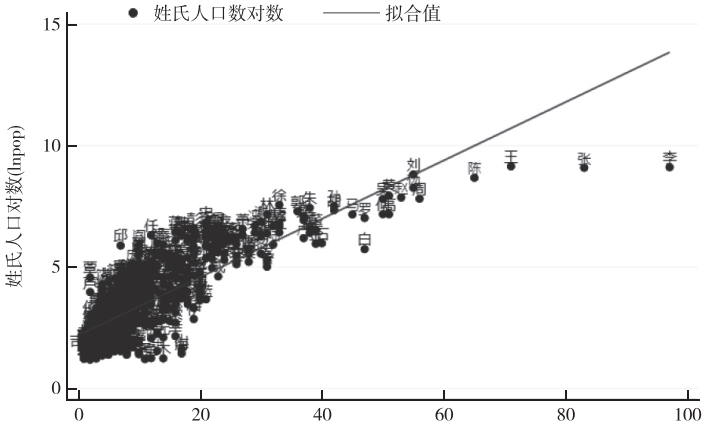

图 20　姓氏人口数对数与姓氏起源数目的散点图

资料来源:人口数据来自"全国公民身份证号码查询服务中心(NCIIC)";姓氏起源数目据徐铁生编著的《中华姓氏源流大辞典》相关内容整理。

(二)少数民族姓氏人口

在中国历史上,少数民族经常采用汉姓。有些少数民族逐渐融入汉族,但也有些少数民族依然保持了其独特的民族身份。因此,可以使用少数民族的姓氏人口数作为姓氏采用率的另一代理变量。例如,2012 年汉族的李姓人口为 9290.23 万,而少数民族的李姓人口也高达 551.986 万。假设历史上汉族民众采用李姓的偏好,以及后来融入汉族的少数民族采用李姓的偏好,均与保持民族身份的少数民族采用李姓的偏好较为相关,则可用少数民族姓氏人口数作为姓氏采用率的较好代理变量。

我们从全国公民身份证号码查询服务中心获取了 2012 年的全国姓氏人口数与汉族姓氏人口数,将二者相减,即可得到少数民族的姓氏人口数,记为 *pop_minority*,并记其对数为 *lnpop_minority*。图 21 提供了汉族姓氏人口数对数与少数民族姓氏人口数对数的散点图。显然,二者高度正相关(相关系数达 0.81,且在 1% 水平上显著)。从图 21 的右上方可见,汉族的大姓一般也是少数民族的大姓。这意味着,汉族大姓的姓氏采用率一般较高。

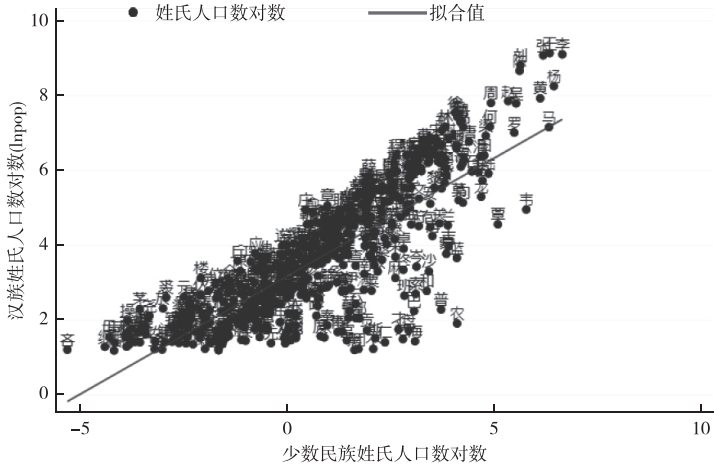

图 21　汉族姓氏人口数对数与少数民族姓氏人口数对数的散点图

资料来源:据“全国公民身份证号码查询服务中心(NCIIC)”相关人口数据整理。

(三)姓氏人口的地理集中度

如果一个姓氏的人口主要居住在某个局部区域(地理集中度较高),则该姓氏人口的增长可能受到该区域资源的限制(正如马尔萨斯人口学说所主张)。另一方面,姓氏人口的地理分布越集中,则在面临战乱时,该姓氏人口所承受的风险也越高。反之,如果一个姓氏更积极地参与跨区域的人口移民(地理集中度较低),则该姓氏人口更可能开枝散叶,且不易受战乱冲击。① 历史证据表明,国姓人口(特别是宗室人口)在面对生存危机时具有更高的迁移率。可能的原因是国姓人口应对生存风险的资源更多,能力更强。②

为此,我们从 2005 年全国 1% 人口抽样调查数据获得地级市层面的汉族姓氏人口,③并通过赫芬达尔指数(HHI),计算姓氏 i 的地理集中度变量(hhi)如下:

$$hhi_i = \sum_j (pop\_share_{ij})^2 \tag{11}$$

其中,pop_share_{ij}为姓氏 i 在第 j 个地级市的汉族人口占该姓氏全国汉族总人口的比重。姓氏人口的地理集中度变量 hhi 的直方图,参见图 22。从图 22 可见,姓氏人口的地理集中度呈现右偏的分布,在右边有一个较长的尾巴。大多数姓氏的地理集中度接近于 0(分布较为分散),而少数姓氏的地理集中度较高(分布较为集中)。进一步,考察姓氏人口数对数与地理集中度的散点图,参见图 23。

从图 23 可见,姓氏人口数对数($lnpop$)与姓氏人口地理集中度(hhi)呈现明显的负相关关系(相关系数为 −0.56,且在 1% 水平上显著)。特别地,一些较为罕见的姓氏有很高的地理集中度。例如,地理集中度最高的植姓(姓氏人口排名第 380 名),其地理集中度高达 0.473。这意味着,从中国随机选取两位植姓居民,则二人同住一个地级市的概率高达 47.3%。查阅徐铁生编著的《中

① 事实上,姓氏人口的地理集中度也可能包含姓氏采用率的因素。比如,若一个姓氏被不同地区的多个人群所采用,也将降低该姓氏人口的地理集中度。在本文中,我们主要将地理集中度视为姓氏人口迁移率的代理变量。

② 根据吴松弟相关著作中所记载的 1477 个靖康之变后南迁移民的资料,笔者计算了其姓氏分布。其中,赵姓移民共 542 名,占总移民的 36.7%,且宗室人口占 80% 以上,作为对比,赵姓在宋代人口中的比重为 5.65%。参见吴松弟:《中国移民史》第 4 卷《辽宋金元时期》,福建人民出版社 1997 年版。

③ 分区域的详细姓氏人口数据,仅在 2005 年人口普查数据中才有。尽管如此,也存在一些缺失值,故变量 hhi 的观测值为 409。特别地,由于姬姓的 hhi 值也缺失,故在本部分的回归中,不需要特别删除姬姓的极端值。

华姓氏源流大辞典》可知,植姓只有 2 个起源(即古代百粤土著,以及广东怀集永固镇植姓),至今人口仍大部分居住在广东省。表 6 给出了以上三个机制变量(*num_origin*,*pop_minority*,*hhi*)的基本统计特征。

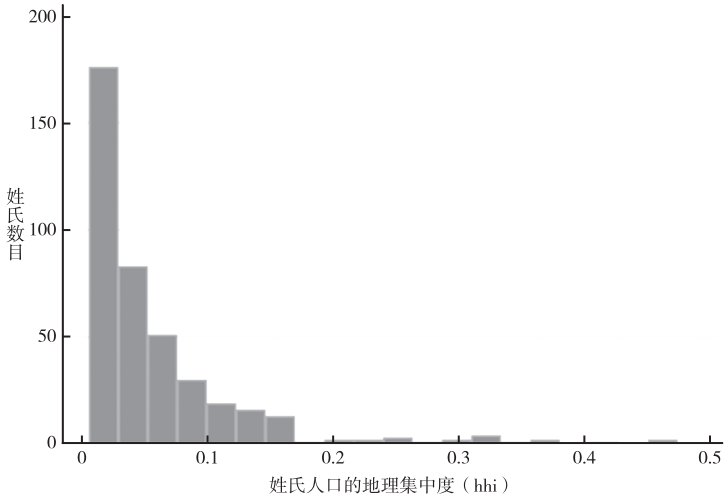

图 22　姓氏人口地理集中度的直方图

资料来源:据"2005 年 1% 人口抽样调查"相关数据整理。

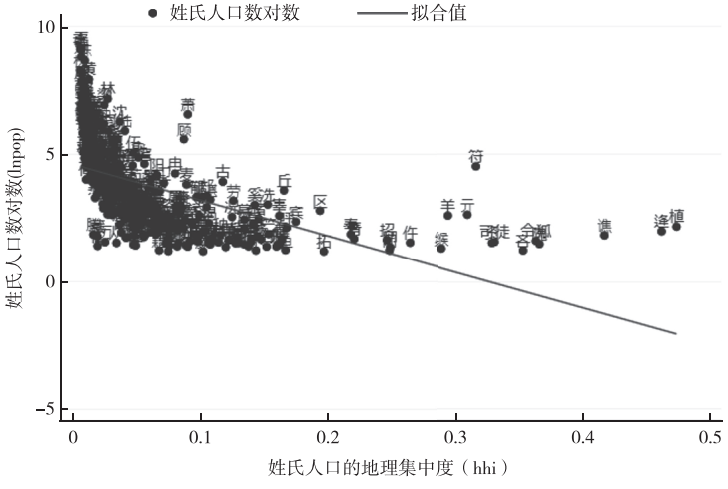

图 23　姓氏人口数对数与地理集中度的散点图

资料来源:人口数据来自"全国公民身份证号码查询服务中心(NCIIC)";姓氏人口地理集中度据"2005 年 1% 人口抽样调查"相关数据整理。

表 6　　有关机制变量的统计特征

变量名称	变量描述	观测值	均值	标准差	最小值	最大值
num_origin	姓氏起源个数	500	12.70	11.91	1	97
pop_minority	少数民族姓氏人口(万)	500	20.62	69.81	0.0049	749.45
hhi	姓氏人口的地理集中度	409	0.061	0.071	0.0063	0.473

资料来源:姓氏起源个数来自徐铁生编著的《中华姓氏源流大辞典》;少数民族姓氏人口数量来自"全国公民身份证号码查询服务中心(NCIIC)";姓氏人口地理集中度据"2005 年 1% 人口抽样调查"相关数据整理。

作为机制变量，*num_origin*、*lnpop_minority*、*hhi* 必须与核心的国姓变量（*royal*，*royal_u*，*royal_d*，*royal_dummy*，*royal_u_dummy*，*royal_d_dummy*）较为相关。为此，表 7 汇报了国姓变量与这些机制变量之间的相关系数。

表 7　　**国姓变量与机制变量的相关系数**

变量	*royal*	*royal_u*	*royal_d*	*royal_dummy*	*royal_u_dummy*	*royal_d_dummy*
num_origin	0.389***	0.360***	0.266***	0.574***	0.372***	0.559***
lnpop_minority	0.210***	0.186***	0.151***	0.383***	0.203***	0.374***
hhi	−0.091*	−0.076	−0.082*	−0.122**	−0.113*	−0.113**

从表 7 可见，所有国姓变量均与 *num_origin* 及 *lnpop_minority* 高度正相关，并在 1% 水平上显著。所有国姓变量均与 *hhi* 负相关，其中，*royal_dummy*、*royal_d_dummy* 与 *hhi* 在 5% 水平上显著负相关，而 *royal*、*royal_d*、*royal_u_dummy* 则在 10% 水平上与 *hhi* 显著负相关。这些结果表明，机制变量与国姓变量具有较强的相关性，故国姓变量可能通过这些机制变量而影响被解释变量 *lnpop*。

类似地，表 8 汇报了机制变量与姓氏诞生朝代的相关系数。不妨以周朝为参照系，将周朝之前视为古老姓氏，而将周朝之后视为年轻姓氏。从表 8 可见，虚拟变量 *zhou* 本身与三个机制变量的相关性均不显著。然而，虚拟变量 *prexia*、*xia*、*shang* 基本与姓氏起源数目（*num_origin*）及少数民族姓氏人口数对数（*lnpop_minority*）显著正相关，而与姓氏人口的地理集中度（*hhi*）显著负相关。这意味着，古代姓氏的采用率一般更高，而地理集中度则通常更低。虚拟变量 *postzhou* 则与 *num_origin*、*lnpop_minority* 显著负相关，而与 *hhi* 显著正相关。这意味着，年轻姓氏的采用率一般更低，而地理集中度则通常更高。

表 8　　**姓氏诞生朝代与机制变量的相关系数**

变量	*prexia*	*xia*	*shang*	*zhou*	*postzhou*
num_origin	0.200***	0.060	0.197***	−0.042	−0.284***
lnpop_minority	0.157***	0.090**	0.160***	−0.024	−0.261***
hhi	−0.097**	−0.047	−0.127***	−0.064	2.77***

将以上三个机制变量，分别加入第五部分的回归分析中，所得估计结果参见表 9。表 9 第（1）列的结果，对应于表 3 第（1）列，但要加上三个机制变量。与我们的预期相一致，代表姓氏采用率的机制变量 *num_origin* 与 *lnpop_minority*，均在 1% 水平显著为正；代表人口迁移率的机制变量 *hhi*，则在 1% 水平上显著为负。加入机制变量之后，国姓年限 *royal* 变得很不显著，且在数值上接近于 0。这意味着，国姓变量 *royal* 很可能通过这些机制变量起作用，故在控制机制变量之后，国姓变量 *royal* 在统计上不再显著。进一步，姓氏诞生朝代变量 *xia*、*shang*、*zhou* 也不再显著，尽管 *prexia* 依然在 5% 水平上显著，但其系数估计值大幅降至 0.477，而在表 3 第（1）列中为 2.038。这说明，在相当程度上，姓氏诞生朝代也通过机制变量起作用。另外，在控制机制变量之后，复姓变量 *compound* 也变得不显著。显然，*compound* 也通过姓氏采用率起作用（复姓的采用率更低）。

表 9 第（2）列将第（1）列中的国姓年限 *royal* 进一步细分为统一国姓年限 *royal_u* 与分裂国姓年限 *royal_d*，所得结果在性质上类似于第（1）列。其与表 3 第（4）列相比，表 9 第（2）列中的统一国姓年限 *royal_u* 不再显著，而分裂国姓年限 *royal_d* 变得在 5% 水平上显著为正，但其系数估计值则降至 0.00276，而在表 3 第（4）列中为 0.0181。

表 9 第（3）列将第（1）列中的国姓年限 *royal*，替换为相应的虚拟变量 *royal_dummy*。其中，机制变量均在 1% 水平上显著。其与表 4 第（1）列相比，虽然表 9 第（3）列的国姓变量依然在 1% 水平上显著为正，但其系数估计值则大幅下降为 0.594，而在表 4 第（1）列中则为 2.844。这意味着，在相当程度上，国姓变量 *royal_dummy* 通过机制变量起作用。进一步，姓氏诞生朝代变量 *xia*、*shang*、*zhou* 均不再显著，尽管 *prexia* 依然在 5% 水平上显著，但其系数估计值大幅降至 0.450，而在表 4 第（1）列中

为1.757。表9第(4)列将第(3)列中的国姓变量 *royal_dummy*,进一步细分为统一国姓 *royal_u_dummy* 与分裂国姓 *royal_d_dummy*,所得结果类似于第(3)列。

综上所述,在加入代表姓氏采用率与人口迁移率的三个机制变量后,国姓变量与姓氏诞生年代变量要么失去统计显著性,要么经济显著性大幅下降。这意味着,一方面,在相当程度上,国姓变量与姓氏诞生年代变量通过这些机制变量而起作用;另一方面,这也验证了在国姓变量与姓氏人口数量的双向因果关系中,因果关系主要从国姓导致姓氏人口,而非反方向。

表9　　国姓影响姓氏人口的作用机制

	被解释变量:*lnpop*			
	(1)	(2)	(3)	(4)
	国姓年限	统一/分裂国姓年限	是否国姓	是否统一/分裂国姓
num_origin	0.0627***	0.0622***	0.0548***	0.0558***
	(0.00715)	(0.00718)	(0.00762)	(0.00746)
lnpop_minority	0.313***	0.312***	0.315***	0.313***
	(0.0409)	(0.0406)	(0.0397)	(0.0397)
hhi	−4.185***	−4.227***	−4.463***	−4.459***
	(0.991)	(0.986)	(0.982)	(0.986)
royal	0.000100			
	(0.000961)			
royal_u		−0.00134		
		(0.00153)		
royal_d		0.00276**		
		(0.00136)		
royal_dummy			0.594***	
			(0.166)	
royal_u_dummy				−0.133
				(0.370)
royal_d_dummy				0.605***
				(0.169)
prexia	0.477**	0.484**	0.450**	0.478**
	(0.204)	(0.203)	(0.197)	(0.198)
xia	0.271	0.269	0.259	0.254
	(0.172)	(0.171)	(0.172)	(0.172)
shang	0.126	0.123	0.126	0.127
	(0.138)	(0.138)	(0.137)	(0.137)
zhou	0.0809	0.0776	0.0799	0.0768
	(0.117)	(0.117)	(0.116)	(0.116)
compound	0.417	0.420	0.417	0.419
	(0.294)	(0.296)	(0.306)	(0.307)
stroke	0.0164	0.0164	0.0164	0.0167
	(0.0109)	(0.0109)	(0.0109)	(0.0109)
tone1	−0.127	−0.127	−0.125	−0.134
	(0.112)	(0.112)	(0.112)	(0.112)
tone2	−0.230**	−0.229**	−0.237**	−0.246**
	(0.111)	(0.111)	(0.110)	(0.110)
tone3	−0.138	−0.134	−0.155	−0.148
	(0.136)	(0.136)	(0.137)	(0.138)
_cons	2.769***	2.775***	2.855***	2.849***
	(0.193)	(0.193)	(0.191)	(0.192)
样本数量	409	409	409	409
R^2	0.779	0.780	0.783	0.783

进一步的问题是,在以上三个机制变量中,究竟哪个变量对于被解释变量的作用更大? 在原则上,可通过回归系数来比较这三个机制变量(姓氏起源个数 *num_origin*,少数民族姓氏人口数对数 *lnpop_minority*,以及姓氏人口的地理集中度 *hhi*)的作用力度。但实践障碍在于,这三个变量的量纲完全不同(第 1 个变量为非负整数,第 2 个变量为万人的对数,而第三个变量为取值介于 0 与 1 之间的概率)。

为了统一量纲,我们考察这三个变量分别变化一个标准差对于被解释变量 *lnpop*(姓氏人口数对数)的影响。以表 9 第(1)列为例,三个机制变量的回归系数,标准差以及标准化的回归系数(回归系数乘以标准差)参见表 10。

表 10　　**三个机制变量的边际效应比较**

变量名称	回归系数	标准差	标准化的回归系数
num_origin	0.0627	11.91	0.747
lnpop_minority	0.313	2.275	0.712
hhi	-4.185	0.071	-0.297

说明:此表中的回归系数来自表 9 的第(1)列。

从表 10 可见,作为姓氏采用率的代理变量,姓氏起源个数(*num_origin*)与少数民族姓氏人口数对数(*lnpop_minority*)的标准化系数,均远超姓氏人口的地理集中度(*hhi*,作为姓氏人口迁移率的代理变量)的标准化系数绝对值,且为后者的两倍以上。这表明,姓氏采用率对于姓氏人口的作用明显高于姓氏人口迁移率的作用。

八、结论

中国的姓氏文化源远流长,但对于姓氏人口的定量研究还十分缺乏。本文使用 2012 年中国汉族人口排名前 500 位的姓氏数据(已占汉族总人口 99.8%),通过描述性统计、回归分析与倾向得分匹配,揭示了中国姓氏人口的一些典型特征与决定因素。首先,中国姓氏人口的分布大致服从齐普夫定律,但也有明显偏差,人口集中于大姓,但集中度高于该定律的预测。其次,姓氏诞生朝代越久远,姓氏作为国姓时间越长,则平均而言姓氏人口越多。这些实证结果通过了一系列稳健性检验,包括控制姓氏的笔画、声调、是否复姓,使用子样本区分统一与分裂政权的国姓,以及针对国姓组与非国姓组进行倾向得分匹配。进一步,我们发现姓氏采用率(以姓氏起源数目与少数民族姓氏人口数为代理变量)与人口迁移率(以姓氏人口的地理集中度为代理变量)是驱动以上结果的两个作用机制。

基于以上认识,对于中国人口集中于少数大姓而同姓率远高于欧美国家的原因,我们可以作一些对比和分析。首先,由于中国姓氏起源非常早,加上历史上的王朝更替并没有造成语言文字甚至文化的断裂,这使得更古老的姓氏有更多机会被民众采用,从而成为大姓。中国历史上的少数民族,不管是作为征服者还是被征服者,大都放弃了原有的语言文字和姓氏,而采用了汉族的语言文字和姓氏。相比而言,欧洲历史上的"蛮族入侵"带来了语言文字甚至文化的断裂。"蛮族"带来了新的语言文字和姓氏,而没有采用原住民的语言文字和姓氏,所以大部分姓氏的历史只能追溯到中世纪,因此大姓没有足够的时间去积累人口上的优势。① 其次,在中国历史上,作为国姓的姓氏,在绵延数百年的朝代中得以发扬光大(譬如刘汉、李唐),使得姓氏人口更为集中。相比而言,"国姓效应"在欧洲几乎不存在。在贵族分封制度下,社会等级森严,姓氏作为社会地位和身份的标识,其所有权和使用权具有排他性。因此,像都铎(Tudor)、兰开斯特(Lancaster)、哈布斯堡(Habsburg)这样的王族姓氏

① Gregory Clark, et al., *The Son Also Rises: 1,000 Years of Social Mobility*.

不可能被平民大量采用,而成为大姓的姓氏。一部分欧洲姓氏跟职业相关,比如史密斯(Smith)、克拉克、梅森(Mason),而另一部分欧洲姓氏则源于地名。而中国的社会流动机制和西方不同,很多国姓家族本身来自平民(比如刘邦、朱元璋),所谓"王侯将相宁有种乎",也没有任何制度障碍阻止平民采用国姓。总的来说,更高的姓氏集中度,可能是中国历史上政治稳定性高、文化延续性强、社会流动的制度性障碍少的结果。①

当然,本研究也存在不足。如,国姓变量的内生性,虽然较为轻微,但依然可能存在。另外,不同姓氏的固有特征仍有待进一步挖掘,比如姓氏作为汉字的词义褒贬及生僻程度,也可能通过影响姓氏的采用率而作用于姓氏人口。这些均有待于在未来的研究中进一步完善。

An Econometric and Historical Analysis of Chinese Surname Population: Typical Characteristic, Decisive Factors and Mechanism of Action

Chen Qiang, *Liu Chunyu*, *Hao Yu*

Abstract: China has a great concentration of surname population. Based on the 2012 Chinese census data of top 500 surname population (about 99.8% of the Han population), this paper reveals several important stylized facts. First, the surname distribution follows the Zipf's law, but with a substantial deviation. Second, a larger surname population is associated with greater antiquity of the surname and longer duration as a royal surname. Furthermore, we find that adoption rate of surname (proxied by the number of surname origins and minority surname population) and migration rate of surname population (proxied by geographic concentration of surname populations) are two important mechanisms. It appears that the greater political stability, stronger cultural persistence, and fewer obstacles to social mobility have contributed to the great concentration of surname population in China.

Keywords: Surname Population, Zipf's Law, Geographic Concentration, Surname Antiquity, Emperors' Surname

(责任编辑:马　烈)

① 也有可能中国文化(乃至东亚文化)更为强调从众(而不鼓励个性),故民众在选择姓氏时倾向于选择大姓或国姓。这一点在韩国也表现明显:韩国第一大姓金姓占全国人口超过五分之一,金姓与第二大姓李姓的人口之和占全国人口超过三分之一,而前四大姓(金、李、朴、崔)的人口之和已几乎占全国人口的一半。