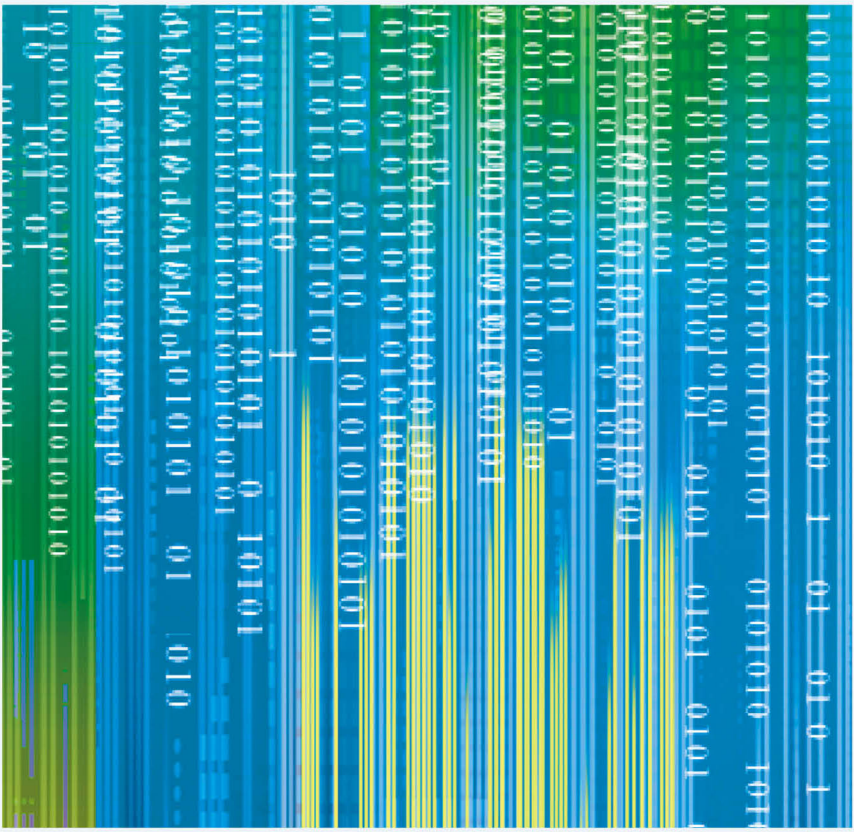
大数据与中国历史研究

第5辑



Big Data and the Study of Chinese History

付海晏 主编

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

集人文社科之思　刊专业学术之声

集 刊 名：大数据与中国历史研究
主　　编：付海晏
执行主编：薛　勤

BIG DATA AND THE STUDY OF CHINESE HISTORY

第5辑

集刊序列号：PIJ-2017-216
集刊主页：www.jikan.com.cn/ 大数据与中国历史研究
集刊投约稿平台：www.iedol.cn

大数据与中国历史研究

第⑤辑

付海晏 主编

Big Data and the Study of Chinese History

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

目　录

·研究动态·

·史料介绍·

中国历史研究中的数据库建设

中国历史官员量化数据库——清代（CGED-Q）的人名匹配与官员记录连接*

康文林　陈必佳

摘　要：本文介绍了利用中国历史官员量化数据库——清代（China Government Employee Dataset-Qing，CGED-Q）进行人名匹配和官员记录连接的方法。CGED-Q 包括缙绅录（Jinshenlu，JSL）和科举记录（Examina-tion Records，ER）两大部分，前者收录官、坊刻本文武官员季度名册，后者收录科举中式者记录名册。本文首先重点评估了原始史料中各项变量的多样性和识别不连贯记录的潜力，以此确定能够用于有效消歧的主要变量。民人官员的主要变量包括姓、名、籍贯省县，旗人官员的主要变量则包括名和旗分等。其次评估了可能有助于进行连接匹配的次要变量。最后，描述了主次变量记录匹配中各项问题的解决方法。

关键词：中国历史　人名匹配　精英　官员仕途

---

* 本文原刊 Campbell，Cameron and Bijia Chen，“Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing（CGED-Q），” *Historical Life Course Studies*，Vol. 12，2022：233-259。本文得到香港研究资助局一般研究基金项目 16602621、16601718 和 16600017（项目负责人：康文林）的支持。我们感谢李中清-康文林研究团队成员的反馈和建议，特别感谢董浩、李中清和倪志宏。我们还要感谢 Loretta Kim 分享有关满族人名习俗的知识，同时也感谢薛勤、陈俊以及其他使用该数据的研究者，他们向我们报告了他们发现的问题，这帮助我们改进了人名匹配连接的程序。韦圣彬完成了全文的初次翻译，侯玥然完成了最终校对，陈必佳、薛勤、高帅奇、陈俊、虞越协助修改了翻译。

一 引言

本文将介绍利用中国历史精英数据库开展大规模人名匹配的方法。匹配工作基于两个清代精英数据库“中国历史官员量化数据库——清代·缙绅录”(China Government Employee Dataset-Qing Jinshenlu，以下简称“CGED-Q-JSL”)和“中国历史官员量化数据库——清代·科举记录”(China Government Employee Dataset-Qing Examination Records，以下简称“CGED-Q-ER”)。首先，通过对CGED-Q-JSL中清朝1762—1911年各季文武官员名册的连接匹配，本文重建该数据库中各官员的仕途生涯。其次，通过匹配CGED-Q-JSL官员在CGED-Q-ER中的记录，本文补充了仕途经历之外各官员生年、科举成绩、出身等其他属性。以上两项连接匹配，使我们得以研究一个重要的社会学议题：家庭出身和“个人能力”(以科举成绩衡量)在官员的任命、晋升和离职过程中发挥了何种作用？同时，通过关联官员的生年和仕途生涯，我们也得以进一步研究清代官员的年龄结构以及任命、晋升和离职过程中的年龄动态。

本文提供的记录匹配方法，借鉴了既往学者们对CGED-Q-JSL中官员仕途生涯的一系列研究，包括对官员任命、晋升、离职的讨论，[①] 可视化平台搭建，[②] 数据库概述，[③] 以及学位论文，[④] 等等。既往基于CGED-Q-JSL的每

① 具体可参见陈必佳、康文林、李中清《清末新政前后旗人与宗室官员的官职变化初探——以〈缙绅录〉数据库为材料的分析》，《清史研究》2018年第4期；胡恒、陈必佳、康文林：《清代知府选任的空间与量化分析——以政区分等、〈缙绅录〉数据库为中心》，《新亚学报》2020年8月；胡存璐、胡恒、陈必佳、康文林：《清代州的政区分等与知州选任的量化分析》，《数字人文研究》2021年第1期；康文林：《清末科举停废对士人文官群体的影响——基于微观大数据的宏观新视角》，《社会科学辑刊》2020年第4期；薛勤、康文林：《清季改革视阈下吏部官员群体的人事递嬗与结构变迁(1898—1911)——以〈缙绅录〉数据库为中心》，《社会科学研究》2022年第2期。

② Wang, Y., Liang, H., Shu, X., Wang, J., Xu, K., Deng, Z., Campbell, C. D., Chen, B., Wu, Y., & Qu, H., “Interactive Visual Exploration of Longitudinal Historical Career Mobility Data.,” *IEEE Transactions on Visualization and Computer Graphics*, 2021, Early Access. doi: 10.1109/TVCG.2021.3067200.

③ Chen B., Campbell, C. D., Ren, Y., & Lee, J. Z., “Big Data for the Study of Qing Officialdom: The China Government Employee Database-Qing (CGED-Q),” *The Journal of Chinese History*, 4, Special Issue 2, 2002: 431-460. doi: 10.1017/jch.2020.15.

④ Chen, B., Origins and Career Patterns of the Qing Government Officials (1850-1912): Evidence from the China Government Employee Dataset-Qing (CGED-Q), Ph. D. dissertation, Hong Kong University of Science and Technology Division of Social Science, 2019.

一次研究，都不断暴露出数据库原始资料、抄录过程、人名匹配、记录连接程序等方面的新问题，因此我们必须解决这些问题。同时，随着数据集规模的扩张，我们改进了程序的可扩展性和运行速度。最终，正如后文所详述，我们采用了 STATA 中 dtalink 包[①]中的概率连接方法，为人名匹配和记录连接提出了解决方案。

本文主要内容是全面挖掘清代行政史料在记录官员姓名、籍贯和其他变量时出现的诸多问题，解释这些问题如何干扰人名匹配，并提供解决方案。希望我们的解决方案，对致力于基于其他中国历史资料开展人名匹配的研究者，以及 CGED-Q-JSL 的公开版本用户有所助益。[②] 本文提出的问题和提供的解决方案，同样适用于其他中国历史资料。匹配中的常见问题，主要有官员姓、名中的异体字、同音字、形似字问题，由行政区域变化引发的同地异名问题，等等。为了便于其他研究者利用中国历史资料开展人名匹配工作，我们已公开提供用于制作文内各汇总表的基础数据。[③]

本文包括六部分。第一部分介绍研究内容和主题。第二部分回顾中外学界关于历史人物姓名匹配的既往研究。第三和第四部分分别介绍本文使用的两个历史数据库——“中国历史官员量化数据库——清代·缙绅录”（CGED-Q-JSL）和“中国历史官员量化数据库——清代·科举记录”（CGED-Q-ER），详述数据库中可用于连接的变量，包括两个数据库均可用的主要变量，以及仅在 CGED-Q-JSL 可用的次要变量，并指出在连接过程中与主要变量相关的各项问题。第五部分介绍目前在 CGED-Q-JSL 和 CGED-Q-ER 中进行人名匹配和记录连接的方法。第六部分总结研究成果，并展望未来研究

---

① Kranker, K., DTALINK: Stata module to implement probabilistic record linkage, Statistical Software Components S458504, Boston College Department of Economics, 2018, revised 16 Feb. 2019, Retrieved from https://ideas.repec.org/c/boc/bocode/s458504.html.

② 我们已经公开 1850—1864 年以及 1900—1912 年的 CGED-Q-JSL 版本。数据库用户指南详见任玉雪、陈必佳、郝小雯、康文林、李中清《中国历史官员量化数据库——清代缙绅录 1900—1912 时段公开版用户指南》，DataSpace@HKUST V14.10.14711/dataset/E9GKRS，2019 年。数据库与相关文档可在 HKUST Dataspace 的李-康研究团队主页（https://doi.org/10.14711/dataset/E9GKRS）下载，也可以在 Harvard Dataverse 的李-康研究团队主页（https://doi.org/10.7910/DVN/GMQWVZ）下载。当数据库建设更完善时，我们也会公开 CGED-Q-ER 和 CGED-Q-JSL 其他时间段的数据。

③ 作为供其他研究人员用中国历史材料做人名连接的资源，我们已在 HKUST Dataspace 和 Harvard Dataverse 公布作为本文表 2 至表 8 的基础的完整数据。这些数据应该对需要解决姓名或籍贯地记录不一致问题的研究者有用，可以协助他们开发处理此类问题的方法。

中的改进方向。

二 背景

在目前欧洲及北美学界的人口、社会、经济史前沿研究中，大规模、自动化人名匹配及记录连接，是构建长时段历史“大数据”的一个关键性工具。人名匹配的常见应用，包括连接同一个体在不同时间点的人口普查记录，连接不同来源的出生、死亡、婚姻等记录。连接工作有时也涉及其他更专业化的资料，如税务记录、健康记录、退休及养老金记录等，以补充例行人口普查和户口登记记录未涉及的信息。连接后的数据不仅展现了个人生活史，有时也提供了跨越多世代的家族史。由于系列研究的开展，基于英文和其他拼音文字资料的大规模人名匹配连接方法已相对成熟。目前，已有大量文献详论人名匹配工作中面临的挑战，提供了各种解决方案，并开发了便捷易用的软件包。①

在美国、加拿大和欧洲学界，为了建设大规模、长时段的社会、经济史数据库，对人口普查、民事登记和其他行政数据的连接工作已开展了20余年。② 因此，关于英文及其他拼音文字个人记录的大规模人名匹配连接工作，已有大量相关文献。在一项早期研究中，明尼阿波利斯人口中心（Minneapolis Population Centre）将美国1860年、1870年、1900年人口抽样调查数据与1880年人口普查数据相连接，创建了一个统计上具有代表性的连续样本。③ 由此至今，匹配连接方法已取得可观的进步，如通过机器学习实现全自动化记录连接，④ 利用住址和人际关系信息提高连接成功

① 例如 Linkage Library，详见 https://www.icpsr.umich.edu/web/pages/about/linkage-library.html。

② 这些连接工作具体可参见 *Historical Methods Special Issues* 51（2）和53（4），以及 Sylvester, K., & Hacker, J. D., “Introduction to Special Issues on Historical Record Linking,” *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53（2）, 2020: 77-79, doi: 10.1080/01615440.2020.1707445。

③ Ruggles, S., “Linking Historical Censuses: A New Approach,” *History and Computing*, 14（1+2）, 2002: 213-224.

④ Abramitzky, R., Mill, R., & Pérez, S., “Linking Individuals across Historical Sources: A Fully Automated Approach,” *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53（2）, 2020: 94-111, doi: 10.1080/01615440.2018.1543034.

率，等等。①

连接英文和其他拼音文字书写的姓名时，存在一些关键问题。其中，拼写错误、姓名变更、使用变体、年龄或生日等其他变量记录不一致等，可能导致漏连（false negatives），即应被关联的记录未被关联；广泛存在的重名现象，则可能导致错连（false positives），即不应被关联的记录被关联。人们在不同时间、地点对姓名的书写差异，普查员或其他官方记录中对姓名的登记差异，都可能引发拼写错误。国际移民可能会被移民官根据原名音译重起新名或自行改名，妇女婚后常随夫姓，由此引发姓名变更。原名与缩写、昵称有时间杂出现，例如不同场合下分别写 Bill 和 William。在许多欧洲社区中，重姓、重名现象广泛存在，这使判断同姓名者是否为同一人颇具难度。

处理中文姓名时所遇问题，与前述问题大不相同。中文姓并不多样，2020 年，中国前五大姓人口占总人口的 30.8%，前一百大姓人口占总人口的 85.8%。② 但中文名相当多样，因为名通常由两个汉字组成，而这两个汉字有成千上万种可能组合。现实中，名的多样性与不同时代、不同阶层的取名习惯有关。在清代及 20 世纪上半叶，精英男性的名十分多样，因为显贵之家为示其博学，常以具备文学、历史或哲学意涵的生字为子孙取名。由于具有政治或爱国意义的单字姓名广泛流行，1960—1980 年代出生人口姓名多样性较低。③

---

① Akgün, Ö., Dearle, A., Kirby, G., Garrett, E., Dalton, T., Christen, P., Dibben C., & Williamson, L., "Linking Scottish Vital Event Records Using Family Groups," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53 (2), 2020: 130-146, doi: 10.1080/01615440.2019.1571466; Helgertz, J., Price, J., Wellington, J., Thompson, K., Ruggles, S., & Fitch, C., "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 55 (1), 2022: 12-29, doi: 10.1080/01615440.2021.1985027.

② 参见公安部《2019 年全国姓名报告》和《2020 年全国姓名报告》。

③ 参见 Bao, H., Cai, H., Jing, Y., & Wang, J., "Novel Evidence for the Increasing Prevalence of Unique Names in China: A Reply to Ogihara," *Frontiers in Psychology*, 12, 2021: 731244, doi: 10.3389/fpsyg.2021.731244; Cai, H., Xi, Z., Yi, F., Liu, Y., & Jing, Y., "Increasing Need for Uniqueness in Contemporary China: Empirical Evidence," *Frontiers in Psychology*, 9, 2018, Article 554, doi: 10.3389/fpsyg.2018.00554; Chua, I., "What Can We Tell from the Evolution of Han Chinese Names?" *Kontinentalist*, 2021, Downloaded April 8, 2022, Retrieved from https://kontinentalist.com/stories/a-cultural-history-of-han-chinese-names-for-girls-and-boys-in-china。基于 Han-Wu-Shang（Bruce）Bao 在 https://github.com/psychbruce/ChineseNames 分享的 Chinese Name Database（1930-2008），Chua（2021）概述了中国当代命名习惯，并展示了 20 世纪不同类型姓名的流行度的描述性统计。

人物记录连接程序的开发，具有重要意义。目前学界已有众多致力于开发中国历史人物传记数据库的项目，其中典型例子如中国历代人物传记资料库（China Biographical Database）、[①] 中国近代人物传记数据库（Modern China Historical Database），[②] 以及李-康团队的各数据库项目。[③] 这些数据库是开展中国历史社会群体，尤其是精英群体人物志研究[④]的基础。数据库的创建者们进行了数据消歧（disambiguation）工作，以评估两个或多个来源中姓名及其他属性相同的记录是否为同一人，然后对数据库中的每个人赋予唯一标识符。这些工作与本文对CGED-Q的匹配连接工作类似，同时它们也涉及更广泛的非结构化文本（如报刊、通史、方志等）中的人名。[⑤] 适用于拼音文字、基于发音的姓名连接方法并不适用于中文姓名，因为汉语中同音异义现象非常普遍，发音相同的姓名实际可能完全不同，对形似字的误读也可能引发歧义。

---

① Fuller, M. A., *The China Biographical Database User's Guide*, Revised Version 3.3, May 26, 2021, Retrieved from https://projects.iq.harvard.edu/cbdb/supporting-documents; Tsui, L. K., & Wang, H., "Harvesting Big Biographical Data for Chinese History: The China Biographical Database (CBDB)," *Journal of Chinese History*, 4 (2), July, 2020: 505-511, doi: 10.1017/jch.2020.21.

② Armand, C., Guo W., Henriot, C., Hu, Y., & Van den Bosch, N., Modern China Biographical Database (MCBD) User Manual, ENP-China, Aix-Marseille University, 2022, Retrieved from https://bookdown.enpchina.eu/mcbd_usermanual/.

③ Campbell, C. D., & Lee, J. Z., "Historical Chinese Microdata: 40 Years of Dataset Construction by the Lee-Campbell Research Group," *Historical Life Course Studies*, 9, Special Issue 4, 2020: 130-157, doi: 10.51964/hlcs9303.

④ Stone, L., "Prosopography," *Daedalus*, 100 (1), 1971: 46-79.

⑤ 详见Campbell & Lee (2020) 和陈必佳和康文林《从一种到多种史料：理解清代官员仕途的新方法》，未发表，2022。这两篇文章简要地描述过CGED-Q中的连接工作。运用两个公开的中国多代人口数据库—辽宁和双城（CMGPD-LN和CMGPD-SC），我们也对清代辽东和双城的八旗人丁做过人名匹配和记录连接，并展示了我们的方法和结果。详见 Lee, J. Z., Campbell, C. D., & Chen S., *China Multi-Generational Panel Dataset*, *Liaoning* (*CMGPD-LN*) *1749-1909*, *User Guide*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2010; Wang, H., Chen, S., Dong, H., Noellert, M. Campbell, C. D., & Lee, J. Z., *China Multi-Generational Panel Dataset*, *Shuangcheng* (*CMGPD-SC*) *1866-1914*, *User Guide*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2013, doi: 10.3886/ICPSR35292.v9, Retrieved from https://www.icpsr.umich.edu/web/DSDR/st-udies/35292; Lee, J. Z., & Campbell, C. D., *Fate and Fortune in Rural China. Social Organization and Population Behaviour in Liaoning*, *1774-1873*, Cambridge, UK: Cambridge University Press, 1997, pp. 223-237 Appendix A。根据与中国历代人物传记资料库和中国近代人物传记数据库负责人的交流，我们得知他们尚未有描述记录连接和消除歧义程序的学术发表。

下文所述关于中文姓名匹配及消歧的各项研究，主要涉及当代非结构化文本（如网页）中的姓名，而非与 CGED-Q 相似的结构化记录。此处述及这些研究，是因为它们可能有助于后续将 CGED-Q 中的官员记录与非结构化文本中的官员信息相匹配。陈松与王宏甦曾评估中文文本中姓名消歧中的问题，认为单字名比双字名更具挑战性，在姓氏与单字名的组合为常用词时尤难消除歧义。① 例如，“高峰”二字可以是姓“高”名“峰”，也可以意指真正的高峰。② Han 等及 Fan 等的文章介绍了基于聚类、利用与上下文中共现词语进行姓名消歧的方法。③ 以上研究虽与本研究涉及的结构化数据库人名匹配有所区别，但足以体现既往研究者在非结构化史料（如报刊、书籍、论文）中提取姓名、开展姓名消歧的努力。

此外，部分研究讨论了中文文本作者姓名消歧工作。Han 等人介绍了中文出版物作者姓名消歧的案例，设计了基于合作者姓名、作者机构和“语义指纹”的消歧技术。④ Kim 等人认为，中文作者的姓名与其音译名的共现，对英文出版物的中国作者姓名消歧大有帮助。⑤ Yin 等人采用有监督机器学习方法，在人工标注数据集基础上，尝试对 1985—2016 年中国专利发明人姓名进行了消歧。⑥

---

① Chen，Y.，& Huang，C.，“Exploring personal name disambiguation from name understanding，” 2010 4th International Universal Communication Symposium，2010：345－349，doi：10.1109/IUCS.2010.5666185.

② 文本分词同样重要，因为在没有单词间空格的情况下，一个词的最后一个字符和紧随其后的词的第一个字符可能会被误认为是一个名字。

③ Han，W.，Xu，X.，& Zhao，T.，“Study on Chinese Person Name Disambiguation Based on Multi-Stage Strategy，” *Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*，2011，10.1109/FSKD.2011.6019646；Fan，C.，& Li，Y.，“Chinese Personal Name Disambiguation Based on Clustering，” *Wireless Communications and Mobile Computing*，2011，3790176.10.1155/2021/3790176.

④ Han，H.，Yao，C.，Fu，Y.，Yu，Y.，Zhang，Y.，& Xu，S.，“Semantic Fingerprints-based Author Name Disambiguation in Chinese Documents，1879－1896，” *Scientometrics*，111(3)，2017，10.1007/s11192－017－2338－6.

⑤ Kim J.，Kim，J.，& Kim，J.，“Effect of Chinese Characters on Machine Learning for Chinese Author Name Disambiguation：A Counterfactual Evaluation，” *Journal of Information Science*，2021，Online First，doi：10.1177%2F01655515211018171.

⑥ Yin，D.，Motohashi，K.，& Dang，J.，“Large-scale Name Disambiguation of Chinese Patent Inventors（1985－2016），” *Scientometrics*，122，2020：765－790，doi：10.1007/s11192－019－03310－w.

另有一些研究试图首先设法评估汉字单字读音和外形的相似性，继而评估字符串的相似性。如 Liu 等人曾提出一种根据发音、字形和字义对相似字进行编码，进而比较字符间相似度的方法。[①] 陈鸣等人设计了一种反映汉字发音和字形的“声形码”，并以此码为衡量两字相似与否的基础。[②] Xu 等人进一步将单字“声形码”与字符串“Dice 形似度”结合应用。[③] 以上方法解决了下文所述的另一个挑战：由于原始材料中或抄录时的错误，在数据集中，同一人的姓名在各条记录中可能有所不同，有时被误录为同音字，有时被误录为形似字。

三 中国历史官员量化数据库——清代·缙绅录

中国历史官员量化数据库——清代·缙绅录（CGED-Q-JSL）使用清代按季更新印行的文武官员名册——缙绅录构建。此前，研究团队已发表数篇文章详述该数据库的史料来源和构建方法，[④] 此处仅阐述部分关键细节。清代缙绅录分为文缙绅录（官刻本多以《爵秩全览》为名，坊刻本多以《缙绅全书》为名）和武缙绅录（多以《中枢备览》为名）。官刻本的文武缙绅录分别由吏部和兵部负责印行。[⑤] 清廷刊印官刻本缙绅录，是为了记录官职及在任官员。19 世纪，民间书坊开始印行并发售坊刻本缙绅录，并在其中收入由书坊收集的更为详细的官员信息。[⑥] 出于查找官缺、访寻入仕亲友同年近况等目的，时人争相购买坊刻本缙绅录。

---

① Liu M., Rus, V., Liao, Qi., & Liu, L., “Encoding and Ranking Similar Chinese Characters,” *Journal of Information Science and Engineering*, 33, 2017: 1195-1211, doi: 10.6688%2fJISE.2017.33.5.6.

② 陈鸣、杜庆治、邵玉斌、龙华：《基于音形码的汉字相似度比对算法》，《信息技术》2018 年第 11 期。

③ Xu, S., Zheng, M., & Li, X., “String Comparators for Chinese-Characters-Based Record Linkages,” *IEEE Access*, 9, 2020: 3735-3743, doi: 10.1109/ACCESS.2020.3047927.

④ 参见 Chen et al.（2020）；任玉雪、陈必佳、郝小雯、康文林、李中清：《清代缙绅录量化数据库与官僚群体》，《清史研究》2016 年第 4 期；任玉雪等（2019）。

⑤ 我们将从缙绅录和《中枢备览》官员名册构建的数据集统称为中国历史官员量化数据库——清代·缙绅录。因为大多数情况下，《中枢备览》是同一季缙绅的一部分，当代的图书馆和档案馆也把《中枢备览》编为缙绅录的部分。只有少量的《中枢备览》不是缙绅录合辑的一部分，而是独立的。

⑥ 关于商业出版和官方出版的缙绅录的内容差异，参见 Chen et al.（2020）。

截至本文写作时，CGED-Q-JSL 已收录 275 季文官缙绅录和 75 季武官缙绅录，共有记录 4433600 条。每季文缙绅录含 13000—15000 个文官官职及在任官员信息；每季武缙绅录约含 8000 个武官职官及在任官员信息。CGED-Q-JSL 所涉时段为 1762—1912 年，其中 1830 年以前的记录较为分散，1830 年后的记录相对完整。1830—1911 年，基本每年至少收录有一季缙绅录，并在数年中收录全部四季各版。武缙绅录本身版本则较为分散，可用版本的时间间隔也较长。

在 CGED-Q-JSL 中，78.9%的官员是民人，其余几乎都是旗人。[①] 大多数旗人官员为满洲八旗或蒙古八旗子弟，亦有 16.4%为汉人，即汉军八旗子弟。旗人在清代官僚系统中享有特权，拥有独立的铨选和晋升途径，并在某些职位上享有固定官缺。因此，尽管旗人仅占清代人口的 2%—4%，[②] 却占文官总数的五分之一，京官人数的三分之二，盛京官员的九成。[③]

为了按时间纵向连接 CGED-Q-JSL 中的官员记录，重构官员仕途生涯，本研究将官员信息分为主要变量和次要变量。主要变量，即官员基础、恒定的个人信息，这些基础信息存在于所有或绝大多数记录中，并有较大概率存在于其他相关史料来源中。其中，最核心的主要变量是姓名。次要变量，即 CGED-Q-JSL 中特有的变量，这些变量不一定存在于其他史料来源，也不一定存在于所有版本的文武缙绅录中。次要变量的内容也可能随时间而变，例如官员在不同季缙绅录中的官职。次要变量可用于辅助判断根据主要变量生成的记录连接是否正确，但不足以独立支持 CGED-Q-JSL 内部或 CGED-Q-JSL 与 CGED-Q-ER 之间的人名匹配及记录连接工作。

因为数据库中旗人官员大多无姓，民人官员则基本都有姓，因此二者可用于匹配连接的主要变量有所不同，在实践中，我们将两类人员分别处理。有姓官员记录占总体记录的 80.2%，包含所有民人官员和三分之一的

---

① 我们所称的“民人”也可能在 1950 年后被划分为少数民族，但是缙绅录没有让我们区分他们的民族身份的信息。

② Elliott，M.，Campbell，C.，& Lee，J.（2016），“A Demographic Estimate of the Population of the Qing Banners.，” *Études chinoises*，XXXV-1，9-40.

③ Chen *et al.*（2020），p. 454.

汉军旗人官员。[①] 这些官员的基础恒定信息不仅包括姓和名，还包括籍贯。籍贯通常为省及州县，下文中还将讨论一些更复杂的情况。无姓官员包括所有满军旗人、蒙古旗人及三分之二的汉军旗人官员。[②] 原则上，名和旗分是无姓官员仅有的稳定变量，即用于匹配连接旗人官员记录的主要变量。

在一季缙绅录中，有姓官员和无姓官员用于标识身份的主要变量不同。对于有姓官员，姓、名、籍贯省、籍贯县的组合通常是唯一的。如果这些信息在每一季缙绅录中都能被准确一致地记录下来，原则上就足以支持记录连接。表1统计了每季缙绅录中主要变量组合重复出现的次数。95.0%的有姓官员的姓名组合在当季版本中是唯一的。也即是说，95.0%的记录在当季版本中没有重名现象。4.4%的记录在当季中版本仅有一条与其重名的记录。若将姓、名、籍贯组合在一起，98.1%的有姓官员记录在当季版本中唯一。进一步的分析表明，当季版本中大多数重复记录实际上属于同一位官员。如果同一官员同时担任不止一职，每个职位都会被单独记录，引发重复。

表1　缙绅录官员的主要变量重复情况（1760—1912）

同季版本内重复次数*	有姓官员（%）		无姓官员（%）	
	姓和名	加上籍贯后	名	加上旗分后
1	95.0	98.1	64.0	88.0
2	4.4	1.7	19.9	9.9
3	0.5	0.2	8.2	1.4
4	0.1	0.0	4.0	0.4
大于5	0.01	0.0	3.9	0.4
总计	100	100	100	100
记录数（条）	2817156	2817156	784502	784502

* 重复次数指的是在同一季的版本中，具有特定组合的主要变量的记录总数。

① 只有三分之一有汉军旗人身份记录的官员有姓，其余汉军旗人没有姓，我们推断他们已有满洲名。详见 Campbell，C.D.，Lee，J.Z.，& Elliott，M.，“Identity Construction and Reconstruction：Naming and Manchu Ethnicity in Northeast China，17491909，”（*Historical Methods*，35，3（Summer），2002：101-116）对东北的汉人转用满洲名的讨论。

② 满洲旗人和蒙古旗人分别占所有旗人的71.4%和12.2%。

对于无姓官员，仅凭名不足以确保记录连接的可靠性。只有三分之二的无姓记录在当季版本中无重名，剩余三分之一至少有两条同名记录。若使用旗分和名的组合，88%的记录在当季版本中唯一，12%的记录在当季版本中至少有一条重复。进一步的分析表明，这些重复既包含同一官员同时任多职的情况，也包含同名但非同一官员的情况。

上述统计结果说明，姓名匹配和记录连接方法必须视官员是否有姓而定。由于有姓官员的姓、名、籍贯省县组合基本唯一，不同官员记录被错连的情况应十分罕见。因此，连接有姓官员记录的主要任务是避免漏连现象，即因为姓名或其他变量前后记录不一致而连接失败。无姓官员则因缺少姓与籍贯地信息，“错连”风险较高，纵使两条记录名与旗分均一致，也很可能指向不同官员。

下文将详细介绍用于人名匹配和记录连接的主要和次要变量，重点阐述这些变量的同质或异质性，并评估它们在连接工作中的可用性。对变量的讨论分为两部分：一是有姓官员记录中可用的变量，二是无姓官员记录中可用的变量。

（一）有姓官员可用变量

1. 姓

由于少数大姓占据了大多数有姓记录，用姓作为连接主要变量的效用有限。表 2 为 CGED-Q-JSL 中 100 个最常见姓的累计百分比。最常见的前五大姓——王、张、李、陈和刘约占全部有姓官员记录的四分之一，前十大姓约占 38.3%，前二十大姓约占一半，前两百大姓约占 95.1%。虽然 CGED-Q-JSL 出现了 1626 个不同的姓，但正如表 2 所示，由于异体字、形似字的存在，姓的实际数量会更少。

表 2　CGED-Q-JSL 中前 100 个常见姓的累计频率（1760—1912）

	1—20		21—40		41—60		61—80		81—100	
	姓氏	%	姓氏	%	姓氏	%	姓氏	%	姓氏	%
1	王	6.6	林	51.6	蔡	65.4	魏	75.0	薛	81.5
2	張	12.7	謝	52.4	韓	65.9	戴	75.4	廖	81.8

续表

	1—20		21—40		41—60		61—80		81—100	
	姓氏	%	姓氏	%	姓氏	%	姓氏	%	姓氏	%
3	李	18.7	郭	53.3	唐	66.5	盧	75.7	白	82.0
4	陳	23.5	高	54.1	鄧	67.1	田	76.1	嚴	82.3
5	劉	27.7	許	54.9	蔣	67.6	崔	76.5	萬	82.6
6	楊	30.6	馮	55.6	方	68.2	夏	76.8	施	82.8
7	周	32.8	吳	56.4	孔	68.7	熊	77.2	賈	83.1
8	吳	34.7	羅	57.1	蕭	69.3	陶	77.5	洪	83.3
9	徐	36.5	梁	57.8	袁	69.8	秦	77.8	雷	83.6
10	趙	38.3	姚	58.5	曾	70.3	俞	78.2	邱	83.8
11	朱	40.0	葉	59.2	董	70.8	江	78.5	姜	84.1
12	孫	41.5	程	59.9	章	71.3	譚	78.8	孟	84.3
13	胡	42.9	余	60.5	傅	71.7	鄒	79.2	賀	84.5
14	馬	44.2	宋	61.1	錢	72.2	史	79.5	毛	84.8
15	沈	45.4	潘	61.7	顧	72.6	于	79.8	侯	85.0
16	黃	46.6	丁	62.4	范	73.0	鍾	80.1	尹	85.2
17	何	47.8	彭	63.0	杜	73.4	龔	80.4	武	85.4
18	鄭	48.8	陸	63.6	蘇	73.8	邵	80.7	郝	85.6
19	黃	49.7	曹	64.2	任	74.2	石	80.9	葛	85.8
20	汪	50.7	金	64.8	呂	74.6	湯	81.2	倪	85.9

注：基于 CGED-Q JSL 中 3244484 个可辨认的姓计算。

用姓连接记录会产生一个问题：相邻季版本中，同一人的姓被记录为不同形似字。数据库中，共计有 1559380 对版本相邻、间隔时间不超过一年的记录，它们的双字名、籍贯省县、官职、科名[①]均相同，几乎可以确定每一对实际指代同一名官员。其中，20055 对（1.3%）记录姓氏写法存在差异。表 3 列出了姓氏差异字符对的累计频率。最常见的姓氏差异字符对（黄、黃）占比 22.4%，前 20 个最常见的差异字符对占比近三分之二，前 100 个占比 79.2%。

① 通过科考或捐纳获得科名，民人就此获得任官资格。科名是我们下面会介绍的次要变量。

表 3 CGED-Q-JSL 相邻版本中前 100 个最常见的写法不一致的姓氏对的累计频率

	1—20		21—40		41—60		61—80		81—100	
	姓氏对	%	姓氏对	%	姓氏对	%	姓氏对	%	姓氏对	%
1	黄 黃	22.4	衛 衛	63.7	鄧 鄭	70.7	盧 虞	74.6	蔣 薛	77.2
2	吳 吳	35.7	闞 關	64.2	曹 曾	71.0	丁 于	74.7	翰 韓	77.3
3	高 高	41.4	孫 馮	64.7	章 童	71.2	関 關	74.9	向 尚	77.4
4	吕 呂	44.8	張 章	65.1	杜 林	71.5	馮 馮	75.0	俞 喻	77.5
5	段 段	47.2	柳 柳	65.6	余 徐	71.7	葉 蔡	75.1	褚 諸	77.6
6	錢 錢	49.3	劉 陳	66.0	徐 涂	71.9	曾 魯	75.3	徐 許	77.7
7	宋 朱	51.3	甯 甯	66.4	全 金	72.1	張 陳	75.4	束 束	77.8
8	閆 閻	52.8	程 陳	66.8	鄔 鄢	72.4	董 黃	75.5	寇 寇	78.0
9	汪 王	54.1	楊 陽	67.1	董 黃	72.6	刑 邢	75.7	樂 欒	78.1
10	凌 淩	55.3	余 金	67.5	員 員	72.8	宋 宗	75.8	張 楊	78.2
11	賴 賴	56.5	楊 湯	67.8	于 王	72.9	萬 黃	75.9	苑 范	78.3
12	余 俞	57.5	毛 王	68.1	李 陳	73.1	強 强	76.1	郭 鄧	78.4
13	龐 龐	58.4	佘 余	68.4	吳 呂	73.3	王 黃	76.2	娄 婁	78.5
14	温 溫	59.2	竇 竇	68.8	晉 晋	73.5	曹 曺	76.3	柏 栢	78.6
15	馬 馮	59.9	季 李	69.1	曹 賈	73.6	潘 王	76.4	丁 李	78.7
16	涂 涂	60.6	嵇 稽	69.4	童 董	73.8	湛 諶	76.6	褚 褚	78.8
17	顏 顔	61.2	龍 龔	69.6	劉 鄧	74.0	杜 樊	76.7	範 范	78.9
18	関 關	61.9	侯 候	69.9	邊 邊	74.1	唐 康	76.8	廉 廉	79.0
19	江 汪	62.5	朱 李	70.2	瞿 翟	74.3	寇 寇	76.9	呂 吳	79.1
20	鍾 鐘	63.1	陳 陸	70.5	宫 宮	74.4	荆 荊	77.0	孫 張	79.2

注：在双字名、籍贯省县、科名及官职完全相同、版本相邻且间隔不超过一年的 1559380 对记录中，有 20055 对（1.3%）姓氏写法不一致。

表 3 的统计结果还揭示了可能导致漏连现象的两个问题。第一是有些姓存在异体写法。表 3 列举了四个最常见的姓氏差异字符对："黄"和"黃"、"吳"和"吳"、"高"和"高"，以及"吕"和"呂"。在 Unicode 标准中，这些字符被认为是同一字符的不同表示方式，因此这个问题可以得到直接解决。第二个问题更具挑战性——在某些不同版本之间，姓氏字符被看起

来形似但实际完全不同的另一个字符取代。例如表3中第5项（“叚”和“段”）、第7项（“宋”和“朱”）、第9项（“汪”和“王”）以及第15项（“馬”和“馮”）。这些问题或因不同季版本缙绅录中记录本就前后不一，或由史料录入人员转录错误导致。表3中，也有一些差异字符对由明显不同的字符组成，例如第24项“張”和“章”及第28项“程”和“陳”。这类差异所涉姓氏多为常见姓，虽然它们可能是不同官员的记录，但也可能是数据录入过程中转录错误所致。

2. 名

在有姓官员记录中，人名是所有主要变量中最多样的，因此对身份识别、匹配连接实现最有帮助。双字名的官员记录占全部记录的85%，单字名的官员占15%。整个数据集中，共计有102648个不同的人名，其中98745个为双字名，3090个为单字名。如表4所示，双字名非常多样化，最常见的100个双字名仅占全部记录的5.7%，前200个占9%，前1000个占23%，前10000个仅占61%。双字名的多样性，反映了可供取名的汉字数量之巨：在CGED-Q-JSL中，至少有5764个字出现在双字人名中。①

表4 CGED-Q JSL中前100个最常见的有姓官员双字名累计频率

	1—20		21—40		41—60		61—80		81—100	
	名字	%	名字	%	名字	%	名字	%	名字	%
1	汝霖	0.1	樹棠	1.7	瑞麟	2.9	祖培	3.9	錫麟	4.9
2	文炳	0.2	炳文	1.8	桂芳	3.0	繼昌	4.0	登雲	4.9
3	得勝	0.3	雲龍	1.8	殿元	3.0	沛霖	4.0	文彬	4.9
4	占魁	0.4	桂林	1.9	玉麟	3.1	祖蔭	4.1	安邦	5.0
5	兆麟	0.5	占鰲	2.0	國泰	3.1	鴻鈞	4.1	錫疇	5.0
6	作霖	0.6	逢春	2.0	維藩	3.2	其昌	4.2	建勳	5.1
7	廷棟	0.7	廷桂	2.1	恩培	3.2	鵬飛	4.2	鴻恩	5.1
8	秉鈞	0.8	鳳翔	2.2	紹曾	3.3	炳章	4.3	毓麟	5.2
9	承恩	0.9	步雲	2.2	文蔚	3.3	炳南	4.3	玉堂	5.2

① 我们制作了一个完整表格，包含数据库中所有双字名使用过至少一次的汉字，该表格可在Harvard和HKUST Dataverses下载。

续表

	1—20		21—40		41—60		61—80		81—100	
	名字	%	名字	%	名字	%	名字	%	名字	%
10	慶雲	1.0	國楨	2.3	殿魁	3.4	國祥	4.4	樹森	5.2
11	世昌	1.0	煥章	2.3	桂森	3.4	長庚	4.4	念祖	5.3
12	步瀛	1.1	文藻	2.4	國華	3.5	定邦	4.4	桂芬	5.3
13	兆熊	1.2	長春	2.5	光祖	3.5	振邦	4.5	學海	5.4
14	培元	1.2	登瀛	2.5	國瑞	3.6	萬春	4.5	連陞	5.4
15	文光	1.3	慶元	2.6	廷珍	3.6	慶恩	4.6	家駒	5.4
16	維翰	1.4	維城	2.6	世榮	3.7	永清	4.6	錫祺	5.5
17	樹勳	1.4	恩榮	2.7	恩溥	3.7	永清	4.7	文治	5.5
18	文煥	1.5	錫齡	2.7	維新	3.8	廷杰	4.7	濟川	5.6
19	錫恩	1.6	國棟	2.8	春華	3.8	榮光	4.8	占春	5.6
20	振聲	1.6	壽昌	2.8	遇春	3.9	廷楨	4.8	鶴年	5.7

注：基于 CGED-Q-JSL 中 2718433 条可辨认姓和双字名记的记录计算。

与姓一样，名中的汉字在不同季版本中也存在不一致现象，若不解决，同样可能导致漏连。表 5 呈现了双字名中差异字符对的累积百分比。[①] 在相距不超过一年的两个相邻版本缙绅录中，姓、籍贯、职位、科名均相同的 1539198 对记录中，有 4.35%（66994 对）在双字名中的一个字上存在差异。名的差异字符对远多于姓，最常见的差异字符对（“清”和“清”）仅占所有差异字符对的 3.7%，前 20 个最常见的差异字符对约占五分之一（20.3%），前 100 个占 39.2%。

表 5　CGED-Q JSL 相邻版本中最常见的前 100 个双字名中差异字符对的累计频率

	1—20		21—40		41—60		61—80		81—100	
	字符对	%	字符对	%	字符对	%	字符对	%	字符对	%
1	清　清	3.7	鳴　鳴	20.8	覲　覿	27.8	元　光	32.7	得　德	36.4
2	勳　勳	5.7	增　曾	21.2	之　芝	28.1	堯　尧	32.9	台　臺	36.6

① 将连接限定在双字名之间并要求至少有一个字相同，显著提高了在其他所有变量匹配的情况下两条记录指向同一个人的可能性。

续表

	1—20		21—40		41—60		61—80		81—100	
	字符对	%	字符对	%	字符对	%	字符对	%	字符对	%
3	齡　齡	7.0	曾　會	21.6	穀　穀	28.4	峯　峰	33.1	宜　宣	36.7
4	鳳　鳳	8.3	遠　遠	22.0	春　椿	28.6	榮　荣	33.3	城　成	36.9
5	壽　壽	9.5	延　廷	22.4	壁　璧	28.9	世　士	33.5	捷　捷	37.0
6	寶　寶	10.7	廉　廉	22.8	緒　緒	29.2	寬　寬	33.7	嘉　家	37.2
7	晉　晋	11.7	懷　懷	23.2	燮　燮	29.4	顯　顯	33.9	彝　彝	37.3
8	煥　煥	12.7	燿　耀	23.6	凌　淩	29.7	為　爲	34.1	日　曰	37.5
9	賓　賨	13.6	愼　慎	24.0	瀚　翰	29.9	惟　維	34.3	如　汝	37.6
10	彥　彦	14.4	濂　濂	24.3	繩　繩	30.1	甲　申	34.5	連　運	37.8
11	恆　恒	15.2	熙　熙	24.7	保　葆	30.4	煇　輝	34.7	宗　崇	37.9
12	傅　傳	15.9	献　獻	25.0	崧　松	30.6	昭　照	34.8	誠　誡	38.1
13	青　青	16.7	瀾　瀾	25.4	均　鈞	30.9	恩　榮	35.0	弼　弼	38.2
14	思　恩	17.3	棻　芬	25.7	柱　桂	31.1	瑞　端	35.2	煒　燮	38.4
15	鍾　鐘	17.9	蕃　藩	26.0	聯　聯	31.3	祿　禄	35.4	柏　栢	38.5
16	鎮　鎮	18.5	高　高	26.3	豊　豐	31.6	繩　繩	35.6	彝　彝	38.6
17	庭　廷	19.0	啓　啟	26.7	方　芳	31.8	丙　炳	35.7	讓　讓	38.8
18	熙　熙	19.4	樹　澍	27.0	廸　迪	32.0	璋　章	35.9	鰲　鼇	38.9
19	達　達	19.9	先　光	27.3	祐　祜	32.3	堂　棠	36.1	萼　萼	39.1
20	聯　聯	20.3	聯　聯	27.6	舉　舉	32.5	清　青	36.2	逵　達	39.2

注：在姓、双字名中的一个字、籍贯省县、科名及官职完全相同、版本相邻且间隔不超过一年的1539198对记录中，有66994对名中有一个字不同。

与上文中姓存在的问题一样，相邻版本中名的差异，最常因异体字引发，表5中前7个差异字符对皆是因此产生。如“清”和“清”，实际上都是“清”。名中也有形似字引发的差异，如表5中第12、14、22和39对的“傅”和“傳”、“思”和“恩”、“增”和“曾”、“先”和“光”。和姓一样，这些问题既可能是原始史料差异所致，也可能是转录错误所致。

单字名的连接多样性相对较低。如表6所示，最常见的前10个单字名占所有单字名记录的6.6%，前100个占37%，前200个占54%，前500个占78%。单字名差异字符对的模式与表5中的双字名差异字符对类似，在此

因篇幅所限不作罗列。大多数字符差异仍由异体字和形似字引起，但也有一些字符明显不同——它们仍有可能是来自同一县且同职位的不同官员，不应被错误地连接在一起。因此，我们在连接单字名官员记录时，采用了不同的标准，即在评估候选连接结果时，更严格地要求其他变量的匹配相似度。

表 6　CGED-Q-JSL 中最常见的前 100 个有姓官员单字名的累计频率

	1—20		21—40		41—60		61—80		81—100	
	汉字	%	汉字	%	汉字	%	汉字	%	汉字	%
1	鈞	0.9	芳	12.0	煜	20.2	璋	26.9	溶	32.5
2	榮	1.7	煦	12.4	勳	20.6	煥	27.2	琦	32.7
3	鑑	2.4	淦	12.9	潤	21.0	桐	27.5	瑛	33.0
4	炳	3.0	源	13.4	濤	21.3	鎔	27.8	坦	33.2
5	鏞	3.7	澐	13.8	鵬	21.7	玉	28.1	超	33.5
6	鈺	4.3	浩	14.3	鴻	22.0	筠	28.4	鎬	33.7
7	瀛	4.9	培	14.7	沅	22.4	治	28.7	貴	34.0
8	湘	5.5	棠	15.1	椿	22.7	傑	29.0	鐸	34.2
9	楷	6.0	謙	15.6	釗	23.0	榕	29.3	翰	34.5
10	堃	6.6	溥	16.0	均	23.4	坤	29.5	芬	34.7
11	震	7.2	泰	16.4	琳	23.7	增	29.8	藻	34.9
12	杰	7.7	燦	16.8	雲	24.0	濬	30.1	模	35.2
13	銘	8.2	斌	17.2	華	24.3	燾	30.4	炘	35.4
14	彬	8.6	澍	17.6	瑞	24.7	煌	30.6	棟	35.7
15	霖	9.1	熙	18.0	鉞	25.0	琛	30.9	寅	35.9
16	森	9.6	英	18.4	林	25.3	焜	31.2	濟	36.1
17	俊	10.1	煒	18.7	元	25.6	桂	31.4	淳	36.3
18	楨	10.6	瀚	19.1	灝	25.9	蘭	31.7	濂	36.6
19	鼎	11.0	照	19.5	珍	26.3	銑	32.0	塏	36.8
20	銓	11.5	錦	19.9	璜	26.6	麟	32.2	錕	37.0

注：基于 CGED-Q-JSL 中 514417 条可辨认姓和双字名的记录计算。

除了上述问题外，CGED-Q-JSL 中的名是相对稳定且正确的，它们也是

官员在家谱和其他资料记录（如 CGED-Q-ER）中的人名，而非表字或号。本团队向家谱数据库的开发团队共享了 CGED-Q-JSL，据他们尝试，CGED-Q-JSL 中的官员名可成功与家谱中男性成员名连接。CGED-Q-JSL 数据库的用户也报告称，他们用家谱或其他记录中的名，成功在搜索页面找到了其先祖或其他人物。[①] 至于更名问题，由于团队尚未系统研究官员改名现象，除了前文所述原始史料和转录问题外，我们尚未明确获知 CGED-Q-JSL 中是否存在改名案例。[②]

3. 籍贯

在缙绅录数据库中，文官缙绅录和武官缙绅录的籍贯详细程度不同。缙绅录中的籍贯，是官员首次参加科举考试的地点，在多数情况下是家族居住地，但也存在例外。在 CGED-Q-JSL 的有姓文官记录中，有 95%记录了籍贯县或可用于推断籍贯省的当前任职省份。[③] 有姓武官记录中，有 13%记录了籍贯省县，84%仅记录籍贯省，3%仅记录籍贯县。

文官籍贯虽不及名多样，但也同样丰富。表 7 展示了 CGED-Q-JSL 中有姓官员记录中前 100 个高频籍贯地的累计百分比。在多数情况下，籍贯地是官员取得生员资格的省府州县。在成为生员后，官员有资格参加乡试或捐官。由于籍贯地通常仅记县不记府，下文中仅提及州县级行政单位。CGED-Q-JSL 中共计有 10156 种不同省份与州县的组合。这一数字大于任何特定时间的州县实数，下文将详论其原因。

表 7　CGED-Q-JSL 中有姓文官前 100 个高频籍贯地的累计频率

	1—20		21—40		41—60		61—80		81—100	
	省县	%	省县	%	省县	%	省县	%	省县	%
1	順天大興	5.2	湖北漢陽	23.4	四川華陽	31.6	山西平定	37.5	江西建昌	41.9
2	順天宛平	7.5	江蘇吳縣	23.9	廣東順德	31.9	四川重慶	37.7	山東歷城	42.1
3	浙江山陰	9.6	湖南善化	24.3	陝西同州	32.2	江西新建	38.0	浙江餘姚	42.3

① 缙绅录数据库搜索页面见 http://vis.cse.ust.hk/searchjsl/。

② 如果我们在以后发现官员改名的系统性证据，目前 CGED-Q-JSL 的内部连接程序及其和 CGED-Q-ER 之间的跨数据库连接程序，都应做相应调整。

③ 我们也可以根据官员当前任职的省份来推断官员的籍贯省，因为如果官员的任职省份和籍贯省份相同，缙绅录就不会特别注明官员的籍贯信息。

续表

	1—20		21—40		41—60		61—80		81—100	
	省县	%	省县	%	省县	%	省县	%	省县	%
4	浙江會稽	11.1	貴州貴陽	24.8	直隸河間	32.5	山西介休	38.2	江西南豐	42.5
5	湖南長沙	12.2	山東濟南	25.2	廣東嘉應	32.9	浙江慈谿	38.5	湖南湘潭	42.7
6	浙江仁和	13.2	福建閩縣	25.7	江蘇元和	33.2	安徽合肥	38.7	直隸清苑	42.9
7	直隸天津	14.2	河南開封	26.1	安徽涇縣	33.5	廣東番禺	38.9	陝西長安	43.1
8	浙江錢塘	15.1	陝西西安	26.5	雲南昆明	33.8	直隸永平	39.2	河南固始	43.3
9	四川成都	15.9	廣西桂林	26.9	廣東肇慶	34.1	江蘇金匱	39.4	安徽太平	43.5
10	浙江山陰	16.7	浙江紹興	27.3	安徽歙縣	34.4	安徽甯國	39.6	安徽懷甯	43.7
11	廣東廣州	17.4	浙江蕭山	27.8	山西汾州	34.7	江蘇無錫	39.8	江蘇常州	43.9
12	浙江歸安	18.1	福建侯官	28.2	江蘇長洲	35.0	直隸保定	40.1	江蘇蘇州	44.0
13	安徽桐城	18.8	河南祥符	28.6	河南光州	35.3	江蘇常熟	40.3	浙江秀水	44.2
14	江西南昌	19.5	廣西臨桂	29.0	浙江烏程	35.6	江西南城	40.5	山東武定	44.4
15	福建福州	20.1	浙江杭州	29.4	山西太原	35.9	安徽婺源	40.7	廣東香山	44.6
16	江蘇上元	20.8	浙江嘉興	29.8	廣東南海	36.1	山東諸城	40.9	湖北黃州	44.7
17	江蘇陽湖	21.3	湖北武昌	30.1	江蘇吳縣	36.4	浙江上虞	41.1	河南南陽	44.9
18	江蘇武進	21.9	貴州貴筑	30.5	山東萊州	36.7	江西吉安	41.3	江蘇儀徵	45.1
19	順天通州	22.5	江蘇江甯	30.9	雲南臨安	37.0	直隸順天	41.5	江西新城	45.3
20	湖北江夏	23.0	江蘇丹徒	31.2	山東登州	37.2	貴州遵義	41.7	河南衛輝	45.4

注：基于 CGED-Q-JSL 中有姓且有籍贯省县的 2615955 条文官记录。由于《中枢备览》的武官记录几乎都没有籍贯地，本表不统计武的籍贯地。

最常见的 10 个籍贯地占全部有籍贯记录的 16.7%，前 100 个籍贯地占 45.4%。最常见的两个籍贯地是顺天的大兴和宛平，即京官子弟参加科举（顺天乡试）的报考地点。在这种情境及其他少数情况下，官员的原籍实际是其他省县。[①] 然而，只要这些官员的籍贯省县在各版本中都以在顺天的报考地为准且前后一致，匹配连接工作就不会受到干扰。除顺天以外，常见的籍贯县多位于浙江——传统上科举中式者、捐纳者、入仕者的重要来源地。其他常见籍贯省县还有排名第五位的湖南长沙、第七位的直隶天津、

① 数据库中明确登记寄籍的记录很少，对于记录连接并无太大助益。有籍贯省县的官员记录中，仅有 13533 条（占 0.39%）记录明确有寄籍。

第九位的四川成都等。

数据库中出现的省县组合数量超过任何特定时期的县总实数，主要有两个原因。连接记录时，也需格外注意这两个因素。首先，即使籍贯县没有变化，籍贯省在不同版本的缙绅录中也可能发生变化。[①] 在1789985对姓名、官职及科名相同的相邻版本记录中，有0.1%（1941对）记录的籍贯省发生了变化。这些变化或由省界重划引起，但多数情况下是缙绅录出版或数据录入时的错误所致。在相同官员的不同记录中，籍贯县相同时，有几组邻近省份出现频繁互替，包括广东和广西，浙江、江苏、江西和安徽，湖北和湖南，山东和山西，顺天和直隶，陕西和甘肃。[②]

其次，不同版本的缙绅录中，县名可能被写为不同字符。在1581616个版本相邻、姓名、籍贯省、科名、官职相同的记录对中，有3.6%（57066对）的记录县名不同。这些差异几乎都由异体字引起，例如表7中位列第三和第十的籍贯地分别是“浙江山陰”和“浙江山陰”，实际上是同一个县，但县名在原始缙绅录中即存在异体写法。又如位列第二十二和第五十七的籍贯地分别是“江苏吳縣”和“江苏吳縣”，实际上也是同一个县，但“吳”字存在异体写法。其他例子包括浙江钱塘（“錢塘”和“錢塘”）和直隶清苑（“淸苑”和“淸苑”）等。[③]

姓、名和籍贯的差异问题累积，将严重影响官员记录的连接。结合计算上述主要变量的差异率，可以估计同一官员的四个主要变量在两个相邻版本中，至少有一个出现差异的概率。假设四个主要变量的差异概率相互独立，则相邻版本中两条相同记录出现差异的总体概率为1-（1-0.035）（1-0.001）（1-0.0434）（1-0.0128）=0.0896，即8.96%。假设一名官员有5年（20季）仕途生涯，那么在这20季中，至少有一对记录出现差异的

① 《中枢备览》有其独特的复杂性。在18世纪末19世纪初的一些《中枢备览》中，武官的籍贯是“湖广”，即湖南与广东的组合。我们将慈利、祁阳、衡阳、道州四县划归为湖南。同样的，在《中枢备览》和部分缙绅录中，江苏、浙江、安徽和江西的一些县会被列为属于江南。

② 我们比较了间隔不到三年的相邻版本中姓、名、籍贯县、科名、官职完全相同的记录。我们发现36条记录中，尽管官员的籍贯县是临桂，籍贯省却被列为广东，而只有1条记录中籍贯省列为广西；25条记录中，官员籍贯县是昌平，在1条记录中归属顺天，其他记录中归属直隶；21条记录中，官员籍贯县是丹徒，在1条记录中归属江苏，其他记录则归属江西；19条记录中，官员籍贯县是汉阳，在1条记录中归属湖北，其他记录则归属湖南。所属地区在顺天和直隶之间相互切换的县有保定、武清、宁河和宛平。

③ 我们已经在相同的表格公布网站上制作了一份县记录的字符差异对列表。

概率为 3.2%［1 -（1-0.0896）^19］。换言之，在假设四个主要变量差异概率相互独立的情况下，几乎可以肯定，任何一名入仕若干年的官员，都至少有一条记录无法与其他记录完全匹配。如果没有采取恰当措施处理这些差异，许多（甚至是大多数）官员的记录都将被错误地拆分为两名或多名官员的记录。后文将展示经本文所述连接方法处理后得到的官员职业生涯表，该表将证实上述差异确实普遍存在。

4. 次要变量

当有姓官员两条记录的主要变量几乎相同但不完全一致时，次要变量可以用于判断连接的准确性。次要变量可用于证实候选匹配（candidate match），但由于这些次要变量不存在于所有版本的缙绅录中，或各版本记录形式及内容有变，仅凭次要变量不足以证伪候选匹配。坊刻本缙绅录常包含比官刻本更多的细节信息。对于有姓官员，可用的次要变量有科名、捐纳功名、官职、字、号及爵位等。

次要变量中，最重要的是科名中 84.2%的有姓官员记录包含通过科考或捐纳取得的科名，这些科名让他们有资格入仕。对于有进士或举人功名的官员，数据库中没有直接记录其科名，而是记录其中式年份（干支）。举人和进士科名分别在乡试和会试中取得，根据已知的乡会试开科年份，可以用考试干支年份推断对应官员是否为举人或进士出身。若将以此法推断的科名计入统计，则 CGED-Q-JSL 数据库中有 93.2%的有姓官员记录有科名信息。原始记录中的科名有数百种，但其中 89.3%属于以下五类科名之一：①进士，即会试中式者；②举人，即乡试中式者；③通过科举获得科名的正途贡生；④通过捐纳获得科名的异途贡生；⑤通过捐纳获得科名的监生。[①] 在 1405138 对版本相邻，且姓、名、籍贯、官职相同的记录对中，只有 7.5%（106007 对）的科名在两个版本之间不同。[②]

官职对于证实候选匹配也有重要作用。缙绅录记录了京官的所属机构，

---

① 具体参见 Chen B.，Campbell，C. D.，Ren，Y.，& Lee，J. Z.，“Big Data for the Study of Qing Officialdom：The China Government Employee Database-Qing（CGED-Q），” *The Journal of Chinese History*，4，Special Issue 2，2020：431-460，doi：10.1017/jch.2020.15。缙绅录中的少量文官和大量武官有武举科名，小部分文官是荫生，即凭借上代余荫而取得监生资格者。Chen *et al.*（2020）详述了这类科名的统计数据和时间变化趋势。

② 如果科名有变化，通常是向上变动，即在任期间官员通过科举或捐纳获得了更高的科名。最常见的是从监生变为举人，共有 1022 例；举人转变为进士有 633 例。

和外官的所在省、府（州）、县。在主要变量完全相同的所有记录对中，官职名称在相邻版本中发生变化的频率为7.3%。这种变化或因官员官职有变，或因官职名称记录有变。若将官职与所在地或所属机构结合计算，官职记录在相邻版本中发生变化的频率为12.6%——同样既可能由实际变化引起，也可能由各版本记录不一致引起。官职记录具有高度异质性：85%的有姓官员记录，职位与所在地或所属机构的组合在当季版本中唯一。此外，我们还将官职按品级划分，并进一步划分为高、中、低和未入流，以便辅助评估同姓名记录是否为同一官员。①

其他一些只存在于少数官员记录中的变量，也可用于辅助确认姓名匹配正确与否。11.7%的有姓官员的记录中记载了表字或号。是否记录表字或号也随季节版本不同而有所差异：275季文官缙绅录中，有74季完全没有记录表字或号。在CGED-Q-ER中，也仅有零星记录包含表字或号，这使这个变量在跨库连接方面的作用有限。文缙绅录均记载爵位，但仅有约0.5%的文官有爵位。在CGED-Q-JSL中，铨选年份对匹配连接工作可能有所帮助，但仅有60.2%的有姓官员记录包含铨选年份，有57季文缙绅录不含铨选年份，CGED-Q-ER中则完全没有铨选年份。

（二）无姓官员可用变量

1. 名

CGED-Q-JSL中共有26727个无姓官员的名。理论上讲这些都应为旗人官员，大多数是满洲八旗，有些是蒙古八旗。其中，84.1%的名由两个汉字组成，11.2%由三个汉字组成，不到1%字由四个及以上的汉字组成。如表8所示，前100个高频名占全部记录的8.6%，高于有姓官员前100个高频名的占比（6.6%）。有无姓官员名的主要区别在于，无姓官员的名多样性更低，频率分布尾部更短。无姓官员前200个高频名占13%，前1000个占36%，前10000个占92%。相较而言，有姓官员记录中，前10000个高频名仅占64%。需要指出的是，虽然无姓官员总量较少，但统计的分布形态仍然有效。

---

① 官员同时担任两个或两个以上官职时，这些官职通常属于同一品级或相邻品级。如果不同季节版本中记录的官职有变，变化范围也通常限于相同或相邻品级。由高品降为低品的情况很罕见。

表 8 CGED-Q-JSL 中无姓官员前 100 个高频名的累计频率

	1—20		21—40		41—60		61—80		81—100	
	名字	%	名字	%	名字	%	名字	%	名字	%
1	文光	0.2	錫麟	2.4	恩壽	4.1	恒安	5.7	明安	7.0
2	祥麟	0.3	松林	2.5	德興	4.2	恒昌	5.7	慶昌	7.1
3	玉山	0.4	桂森	2.6	祥安	4.3	慶雲	5.8	崇勳	7.1
4	英俊	0.6	瑞麟	2.7	文海	4.4	玉崑	5.9	文溥	7.2
5	文英	0.7	松齡	2.8	延齡	4.5	奎文	5.9	桂斌	7.3
6	文明	0.8	文治	2.9	吉昌	4.5	恩慶	6.0	恩承	7.3
7	長春	0.9	恩光	3.0	崇福	4.6	祥瑞	6.1	定保	7.4
8	慶安	1.0	鍾秀	3.0	恩榮	4.7	祥泰	6.2	清安	7.4
9	慶福	1.2	榮慶	3.1	玉衡	4.8	榮桂	6.2	長慶	7.5
10	毓秀	1.3	常明	3.2	松壽	4.8	文成	6.3	文斌	7.6
11	奎英	1.4	松秀	3.3	文桂	4.9	文惠	6.4	桂昌	7.6
12	恩霖	1.5	文貴	3.4	榮昌	5.0	雙福	6.4	全福	7.7
13	扎拉芬	1.6	慶恩	3.5	榮恩	5.1	佛爾國春	6.5	英奎	7.7
14	英秀	1.7	榮安	3.6	景福	5.1	德馨	6.6	慶祥	7.8
15	慶麟	1.8	崇禧	3.6	景昌	5.2	春慶	6.6	托克托布	7.9
16	德祿	1.9	文瑞	3.7	吉順	5.3	恩明	6.7	英麟	7.9
17	慶瑞	2.0	興奎	3.8	恩隆	5.4	麟祥	6.7	文敬	8.0
18	崇恩	2.1	文麟	3.9	德麟	5.4	桂芬	6.8	常興	8.0
19	桂林	2.2	文秀	4.0	榮光	5.5	德克精額	6.9	松年	8.1
20	文興	2.3	桂芳	4.1	恩綸	5.6	文俊	6.9	全順	8.2

注：基于 CGED-Q-JSL 中 811580 条无姓官员记录计算。

数据库中无姓官员的名通常是满文或蒙古文名字的汉语音译。相同的旗人官员名可能存在不同音译版本。例如，最常见的名 Qing'an，有时音译作“慶安”，有时作“清安”和异体“淸安”。第二常见的名为 Xilin，常见译法有錫麟”、“錫霖”、“熙麟”和“西林”。如果不论声调，汉文译名的多样性将有所降低，仅有 14560 个不同的名。其中，前 100 个高频名占 11.8%，前 200 个占 19.4%，前 1000 个占约一半，前 10000 个占 99.0%。

在 CGED-Q-JSL 中，同一满人或蒙古官员的汉译名很少发生改变。尽管

在入仕之初，同名的满人或蒙古官员可能取不同音译汉名，但其名一旦选定，后续就几乎不会改名。在560559对名的无声调发音、旗分、官职完全相同，相距不超过一年的无姓官员记录中，仅有2.3%（13128对）汉文名发生了变化。且进一步的检查证实，许多变化是以异体字的形式出现。

2. 旗分

旗人官员的旗分信息足以辅助证实候选连接，但由于旗分偶尔有变，因此不足以证伪候选连接。每名旗人子弟隶属于正黄、镶黄、正白、镶白、正红、镶红、正蓝、镶蓝八旗之一。[①] 在88734个汉字名、官职所在地或所属机构和官职均相同，版本相邻的旗人记录对中，有24.4%（21634对）的旗分发生了变化。其中，超过四分之一是在同色正旗和镶旗之间变动，大部分变化发生在笔帖式、员外郎和主事三种官职中。目前，我们尚未详考官员旗分变动过程，需要在清史专家帮助下进一步深入研究。

3. 旗人的次要变量

无姓官员记录中，同季版本中同职现象较为普遍。表9展示了无姓官员官职和所在地（或京官所属机构）的组合统计结果。平均而言，每个版本中仅有16.7%的官职唯一，超过四分之三的官职在同一版本中出现了5次及以上，其中笔帖式、员外郎和主事最多。即使将官职所在地或所属机构与官职组合统计，也仅有不到三分之一的组合在同季中唯一，仍有超过一半的官职在同一版本中有5条及以上相同记录，其中中央各部院衙门笔帖式最多。

表9　缙绅录中无姓官员的官职和所在地（或所属机构）组合的重复概率

在一个版本中的重复次数	官职（%）	官职+所在地（或所属机构）（%）
1	16.7	31.1
2	2.5	8.5
3	1.2	4.0
4	1.4	3.8
5	78.2	52.7
总计（%）	100	100
记录数量（条）	784502	784502

① 上三旗是镶黄旗、正黄旗、正白旗，下五旗是正红旗、镶白旗、镶红旗、正蓝旗、镶蓝旗。

无姓官员还有其他可作为次要变量的信息，但这些信息仅存在于少数记录中，例如官员是否为宗亲或觉罗。无姓官员中，宗亲及觉罗占比7.4%；有姓官员中，宗亲及觉罗占比1.7%。整个清朝皇室共计有3656名男性成员，该群体在官员中的占比，远超其在全国人口中的占比。此外，缙绅录中记录了约三分之一（35.8%）旗人文官的科举及捐纳科名，这些记录主要集中于19世纪后期各版本中。另外，11.6%旗人官员记录有表字或号，7.5%有铨选年。

四　中国历史官员量化数据库——清代·科举记录

中国历史官员量化数据库——清代·科举记录（CGED-Q-ER），是由零散的、不同科次中式者名册抄录而来的。CGED-Q-ER最重要的史料来源为同年齿录。清代乡试、会试中榜者为了加强联谊和社交，常自行编纂同榜中式者人名录及履历名册。科举时代，同榜录取者通常互称“同年”，因此这些名册大多以“同年齿录”命名。每本同年齿录罗列同榜录取者姓、名、籍贯省县，以及已入仕者的当前官职。除本人信息外，同年齿录中还记录父亲、祖父及外祖父的姓名和科名。多数同年齿录还记录年龄，间或罗列其他亲属信息。CGED-Q-ER主要收录进士和举人同年齿录，同时也涉及其他层次中式者，如贡生的同年齿录。目前，数据库中共有5724条进士记录，26870条举人记录，以及11990条其他记录。

除同年齿录外，CGED-Q-ER数据库中也抄录有乡会试中榜者的官方记录，即乡试录和进士题名录，但官方记录所含信息较略。乡试录记载了该科乡试中榜者的姓、名、籍贯县、科考名次和年龄，根据考籍，也可推断中榜者籍贯省。乡试录与同年齿录存在部分信息重合。进士题名录记载了进士的姓、名、中式年、籍贯省县，以及会试和殿试中的科考名次。由于同年齿录已记载大量10世纪进士信息，且更详细，因此进士题名录主要用于未被同年齿录记载的进士信息。

CGED-Q-ER主要涉及两项匹配连接任务。其一是连接同一中式者在同年齿录、乡试录、进士题名录中的记录，以便对同一科次的不同记录去重。一般而言，已有同年齿录的某场乡试，如果同时也有《乡试录》，或者被涵

盖于同一年另外出版的、汇集多省乡试的同年齿录中，就会出现记录重复。在 CGED-Q-ER 内部，不同等级科名之间也可连接，即将举人的记录与其后来成为进士的记录相连——这使我们得以研究中举者的个人和家庭情况对其中进士概率的影响。其二是将 CGED-Q-ER 中中式者的个人和家庭信息，与其在 CGED-Q-JSL 的官职记录相关联——这使我们得以研究中式者个人特征，如家庭背景和家庭成员科举表现等，对其入仕和晋升机会的影响。

CGED-Q-JSL 和 CGED-Q-ER 中记录的姓、名、籍贯省县、中式年、科名等，可用于完成以上两项匹配连接任务。使用 CGED-Q-ER 中的姓、名、籍贯省县开展连接所产生的问题，与上文所述使用 CGED-Q-JSL 时面临的问题相似。姓、名、籍贯省县组合对绝大多数中式者来说是独一无二的，因此无须像上文讨论 CGED-Q-JSL 时那样详尽分析。在不同种史料中，仍存在异体字现象。下文描述的 CGED-Q-JSL 官员记录匹配连接方法，也适用于 CGED-Q-ER。在 CGED-Q-ER 中，中式年可辅助限定匹配范围，排除那些先成为进士后成为举人，或成为举人久于十年之后才成为进士的错误匹配。

五 连接

官员记录的连接工作分为四步。第一步，标准化处理关键变量。第二步，开展简单的确定性连接（deterministic linkage），匹配多个主要、次要变量完全一致的记录，形成可确定为同一官员记录的记录组，并将每个记录组的第一条记录用于后续连接。第三步，利用 STATA 的概率连接包 dtalink[①] 确定用于“分组”（blocking）的变量。当且仅当记录对在被选定的变量上有相同记录时，它们才会被挑选出来，成为候选记录对（candidate pairs of records）。这个步骤会排除大量明显不匹配的记录对，例如姓名都不同的记录对，大幅减少完成匹配连接工作的时长。第四步，再次利用 dtalink 进行概率连接。程序将对上述被挑选出的候选记录对进行概率评分，当评分高于一定阈值时，所有与候选记录对处于相同记录组的记录，都将被分配同一个唯一标识符，也即被连接在一起。

① Kranker, K. (2018).

（一）准备

用于连接匹配的数据集中，主要变量和次要变量都经过标准化处理。为了避免由不同版本中姓名字符差异导致的漏连，我们创建了姓和名的标准化版本。首先，Unicode 编码标准本身包含同一字符的不同版本，可用于处理姓名中的异体字，例如表 6 中的“清”和“清”、“勲”和“勳”等。[①] 经这一步骤转换后的姓名版本，称“CV 版本”（Consolidated Variants，异体字合并版本）。在异体字合并版本基础上，我们进行了第二轮合并：将名中的形似字（但非异体字）归类，例如表 5 中的傅”和“傳”、“思”和“恩”、“增”和“曾”、“先”和“光”等。[②] 经此步骤转换后的姓名版本，称“SC 版本”（Similar Characters，形似字合并版本）。最终，数据库中的每条记录均包含名的原始版本、CV 版本及 SC 版本。

姓的 CV 版本和 SC 版本生成过程大体上与名相同。但在制作姓的 SC 版本时，我们手动检查表 3 的统计结果后，放弃了对一部分常见差异字符对的替换。因为正如前文所述，即使其他变量都相同，仍没有足够证据认为这些异姓官员确实是同一人。最终，我们选定了 12 组有极高概率在出版或抄录过程中出错的形似字，用以制作姓的 SC 版本。[③] 采取这种更保守的做法，是因为姓的字符多样性较低，当其余变量相同而只有姓不同时，不是同一人的可能性更高。未来，研究团队将继续尝试改进形似字处理方案。

籍贯省、县也均经过标准化处理。为了解决县名中出现的同音字误替，我们对县名进行了两个版本的拼音编码（详见表 10）。第一个版本（PY 版本）包含声调，第二个版本（PY TL 版本）不包含声调。此外，为了解决县所属省份不一致的问题，我们创建了一个新的籍贯省版本（C 版本），将安徽、江苏、江西和浙江合并为“江南”，湖南、广东和广西合并为“湖

① Unicode 也被用于将错误录入的简体字转换为繁体字。参见 Unicode 汉字数据库报告，https://unicode.org/reports/tr38/。本研究使用的 Unicode 汉字数据库下载自 https://www.unicode.org/Public/UCD/latest/ucd/Unihan.zip。

② 生成 SC 版本的详细过程如下：首先使用 CV 版本生成一份类似表 5 的常见差异字符对列表，然后手动评估并标记形似字符对，最后将每组形似字映射到其中一个字符上。

③ 这 12 个字符组是：（1）宋，朱，宗；（2）叚，段；（3）王，汪，江；（4）馬，馮，憑；（5）柳，栁；（6）季，李；（7）龍，龔；（8）余，徐，涂；（9）湛，諶；（10）冠，寇；（11）樂，欒；（12）褚，諸。

广”。对于原始缙绅录中明确登记寄籍的记录，我们使用寄籍地作为籍贯省县。

表10　CGED-Q-JSL 和 CGED-Q-ER 连接类型及其对应的分组方式

连接类型		分组方式
CGED-Q-JSL 内部连接		
1	双字名有姓官员	姓（SC）+名（SC） 或 姓（PY）+名（PY）
2	单字名有姓官员	姓（SC）+名（SC）
3	无姓官员	姓（SC）+八旗旗分 或 姓（PY）+八旗旗分+宗室或觉罗+爵位+职位
CGED-Q-ER 内部连接		
4	所有记录	姓（SC）+名（SC）
CGED-Q-ER 和 CGED-Q-JSL 跨数据库连接		
5	双字名有姓者或无姓者	姓（SC）+名（SC） 或 姓（PY NT）+名（PY NT）
6	单字名有姓者	姓（SC）+名（SC）

注：为节省篇幅，本表中“形似字合并版本”（SC 版本）、“有声调拼音版本”（PY 版本）、“无声调拼音版本”（PY NT 版本）皆用英文缩写指代。

（二）确定性连接

我们将版本间隔不超过一年、数个主次变量完全相同的记录归入同一记录组，并将记录组中的第一条该组的摘要记录。将记录归入同一记录组的标准非常严格，足以杜绝错连情况的发生。[①] 利用确定性连接创建记录组

① 对于有姓官员，同一组内的记录必须保证姓、名的 CV 版本，籍贯省的 C 版本，以及籍贯县的 PY 版本均相同。对于无姓的旗人官员，同一组内的记录必须保证名的 CV 版本、旗分、官职、官职所在地或所属机构均相同。我们在分组时要求旗人官员官职记录也一致，是由于同季缙绅录中常有旗人官员名和旗分组合不唯一的现象出现（参见表1）。

的过程简单直接，此处不再赘述。在此步骤后，需要连接的记录组数量比原始记录数量少了一个数量级，这使后续第二和第三阶段所需时间大大减少。

（三）分组

以所用变量类型和错连、漏连风险差异为标准，针对 CGED-Q-JSL 和 CGED-Q-ER 的数据连接工作可分为六种类型。首先是仅针对 CGED-Q-JSL，以重构官员仕途生涯为目标的内部连接，包括三种类型：连接有姓、单字名官员；连接有姓、双字名官员；连接无姓官员。之所以将单字名和双字名的有姓官员区分开来，是因为我们在对比表 4 和表 5 时发现，连接单字名官员记录时错连风险更高，需要对其他变量的匹配程度有更严格的要求。相对而言，由于姓和双字名的组合更有可能唯一，连接双字名的有姓官员时，可相对放宽其他变量的匹配程度标准。无姓官员可用作记录连接的主要变量仅有名和旗分，而名和旗分的组合几乎不可能唯一，因此匹配无姓官员记录时，必须更加重视次要变量。其次是仅针对 CGED-Q-ER 的内部连接。CGED-Q-ER 记录总数相对较少，在进行单字名有姓中式者连接时，错连的概率也较小，因此针对该数据库的所有内部连接都采取相同处理方式——这是第四种类型的连接。GED-Q-JSL 和 CGED-Q-ER 之间的跨数据库连接，按照有姓官员的名是单字还是双字，可以分为另外两种类型，即第五和第六类。

表 10 总结了六种连接类型中用于将官员记录分组的变量。过严的分组标准会导致漏连，过松的分组标准又会增加后续连接用时。在每类连接中，我们都尝试在二者间寻找平衡。通常情况下，我们尽可能使用宽松的分组标准，同时尽量防止明显不可能匹配的记录对进入匹配度评分阶段。因此，多数情况下我们采用名的 SC 版本（形似字合并版本）而非 CV 版本（异体字合并版本）来分组，然后利用基于其他变量计算的匹配度，评估在 SC 版本上匹配但在 CV 版本上不匹配的记录对。

不同类型的 CGED-Q-JSL 内部连接工作，对应不同的分组标准。对于第一种类型，即双字名有姓官员，分组时以姓和名的 SC 版本及 PY 版本分组。不同组的摘要记录，如果姓和名的 SC 版本及 PY 版本均相同，就会成为一组候选匹配，以供基于其他变量（包括姓和名的 CV 版本）进一步评估匹配

度。未采用姓名CV版本分组是因为该版本过于严格，可能会遗漏能用PY版本匹配到的结果。对于第二种类型，即单字名有姓官员，只有姓和名的SC版本均相同时，记录对才会成为候选匹配——这种类型中PY版本会造成较多错连，故不采用。

对于第三种类型，即无姓官员，我们以名的SC版本和旗分分组，或以名的PY版本、旗分、宗亲及觉罗、爵位及官职的组合开展分组。换言之，即使一对记录中名的SC版本不相匹配，当且仅当名的PY版本和其他一系列变量都完全匹配时，它们仍然可以成为候选匹配以供评分。在基于PY版本的匹配中，我们严格要求多个次要变量完全相同，因为选用PY版本会大大增加需要检查的记录对数量。

对于第四种类型，即CGED-Q-ER的内部连接，姓和名的SC版本已足够成为分组依据。此处我们没有对单字名有姓官员采取单独的分组方法，而是在评估候选匹配时采用更严格的评分标准。[①] 对于第五种类型，即CGED-Q-ER和CGED-Q JSL之间的双字名有姓者连接，使用姓和名的SC版本或PY TL版本进行候选配对。[②] 对于第六种类型，即CGED-Q-ER和CGED-Q JSL之间的单字名有姓者连接，仅使用名的SC版本配对。

（四）概率连接

概率连接方法应用广泛，这在众多专业图书、网络资源中均有详述，此处仅简述其基本概念。概率匹配可在分组处理后的数据集中筛选所有可能的匹配记录对，然后依照用户指定标准对每个记录对进行匹配度评分。在评分时，用户除了可以指定用于匹配的变量外，还可以指定随每个变量异同程度增减的具体匹配分值。对于数值型变量，用户还可指定判定匹配的阈值范围，若两条记录在该变量上的差异位于阈值范围内，则增加匹配分值。通过计算候选匹配对的匹配分，我们可以筛选出匹配度最高且达到用户自定义及格分的候选匹配对，作为最终匹配结果。

在数据分组基础上，我们使用上述方法，依据各主要、次要变量对所

---

① 我们没有为CGED-Q-ER中的旗人单独设置分组规则，因为旗人在CGED-Q-ER中占比很低（总体仅占1.2%）。

② 由于CGED-Q-ER中旗人数量较少且少有重名，在这种类型的连接中将旗人并入有姓官员中。

有候选匹配对进行了评分。表 11 和表 12 总结了六种类型的连接中各主要变量异同引发的具体匹配度分变化值。如果候选匹配对满足标题行指定的条件，匹配分就会增加（表中的“+”列）；否则就会减少（表中的“-”列）。表 11 和表 12 还包含确认连接候选匹配对时采用的及格分。为了平衡漏连和错连风险，我们根据变量具体匹配情况，分别调整了每种连接类型中的匹配分增减额度。当需要连接的记录数量较大时，特别是在进行 CGED-Q-JSL 内部连接时，我们设定的评分标准更严格，要求记录间主要变量具有更高相似度。当错连风险较低（需要连接的记录较少）时，我们设定的评分标准更宽松，如 CGED-Q-ER 的内部连接。

表 11　针对 CGED-Q-JSL 内部连接的匹配度评分标准

	CGED-Q-JSL					
	类型一 双字名有姓		类型二 单字名有姓		类型三 无姓	
	+	−	+	−	+	−
主要变量						
姓（CV）+名（CV）+县（PY）						
姓（SC）+名（SC）+县（SC）						
姓（SC）+名（SC）+省（C）						
姓（CV）+名（CV）	100	0	100	0		
姓（CV）+名（SC）						
姓（SC）+名（PY）						
名（CV）					50	0
名（SC）					50	0
籍贯省（C）	100	−400	100	−400		
籍贯县（原始版本）						
籍贯县（PY）	200	−100	200	−100		
江南、湖广的籍贯县（PY）	0	−200	0	−200		
八旗旗分					50	−100
次要变量						
字号	300	0	300	0	200	0

续表

	CGED-Q-JSL					
	类型一 双字名有姓		类型二 单字名有姓		类型三 无姓	
	+	−	+	−	+	−
次要变量						
是否宗室或觉罗					100	0
爵位					100	0
官职						
省	25	0	25	0		
中央部院或府州厅	25	0	25	0		
所属机构或所在县	25	0	25	0		
官职名称	25	0	25	0		
完整官职	100	0	100	0		
品级						
相同					0	−25
相差小于2					0	−50
相差小于3					0	−400
科名（含捐纳）						
原始记录	50	0	50	0	50	0
科名等级（含捐纳）	0	−100	0	−100	0	−100
江南、湖广科名等级（含捐纳）	0	−100	0	100		
科考年份						
相同					50	0
相隔小于5年					0	−50
相隔小于10年		−100		−100	0	−100
相隔小于20年		−200		−200		
相隔小于30年						
相隔小于40年		−500		−500	0	−400

续表

	CGED-Q-JSL					
	类型一 双字名有姓		类型二 单字名有姓		类型三 无姓	
	+	−	+	−	+	−
上一条相邻记录的姓名	50	0	50	0	50	0
下一条相邻记录的姓名	50	0	50	0	50	0
及格分		100		100		150

注：为节省篇幅，本表中“形似字合并版本”（SC 版本）、“有声调拼音版本”（PY 版本）、“无声调拼音版本”（PY NT 版本）、“省合并版本”（C 版本）皆用英文缩写。下表同。

表 12　针对 CGED-Q-JSL 内部连接、CGED-Q-JSL 与 CGED-Q-ER 跨库连接的匹配度评分标准

	CGED-Q-ER		CGED-Q-JSL to CGED-Q-ER			
	类型四 所有记录		类型五 双字名有姓 或无姓		类型六 单字名有姓	
	+	−	+	−	+	−
首要变量						
姓（CV）+名（CV）+县（PY）			500	0	500	0
姓（SC）+名（SC）+县（SC）	300	0	200	0	200	0
姓（SC）+名（SC）+省（C）	100	0	200	0		
姓（CV）+名（CV）			150	0	200	0
姓（CV）+名（SC）			100	0	150	0
姓（SC）+名（PY）			50	0		
名（CV）						
名（SC）						
籍贯省（C）		−200	0	−200		
籍贯县（原始版本）			100	0		
籍贯县（PY）	0	−200	100	0		
科考年份						
相同						
相隔小于 5 年						

续表

	CGED-Q-ER		CGED-Q-JSL to CGED-Q-ER			
	类型四 所有记录		类型五 双字名有姓 或无姓		类型六 单字名有姓	
	+	−	+	−	+	−
科考年份						
相隔小于10年	0	−100	100	0	100	0
相隔小于20年	0	−300				
相隔小于30年						
相隔小于40年			0	−500	0	−500
及格分		100		200		200

表11及表12中的匹配度评分标准，均经过反复迭代调整。每一轮连接结束后，我们都会复查连接结果：在数据中搜索官职、科名等次要变量完全匹配，大多数主变量匹配，而未能顺利与任一官员记录组连接的零散记录，定位漏连情况；通过检查各记录组内记录是否为同一人，定位错连情况。复查有助于厘清异体字使用状况，并启发我们创建了姓名的CV和SC版本。迭代过程中，我们还增加了字号、完整官职等次要变量完全相同时匹配分的增额，因为这些变量很难偶然完全一致。省份名称变化问题，也是在复查时发现的。

识别错连时，我们研究了同记录组中至少在一个主要变量出现差异的情况。我们由此发现连接单字名官员记录时需采用更严格的标准，此外，当籍贯省和科名等本应固定不变的主变量有差异时，匹配分值应有更大降幅。使用数据库开展特定类别官员委任及晋升问题的研究者，也向我们反映了他们发现的问题，以供调查参考。①

在CGED-Q-JSL的内部连接（类型一至类型三）中，当名、官职、籍贯县等变量一致时，匹配分的增额最大，因为这些变量丰富度极高，恰巧一

① 具体可参见陈必佳、康文林、李中清（2018），胡恒、陈必佳、康文林（2020），胡存璐，胡恒，陈必佳，康文林（2021），以及薛勤、康文林（2022）。这些研究对记录连接结果的分析，都引导我们发现新问题并进一步完善连接程序。

致的概率最低。尽管针对单字名和双字名记录的分组方法有所不同，但针对二者的匹配度评分方式是相同的。在字号、完整职位等次要变量一致时大幅增加匹配分，有助于抵消籍贯省县名前后记录不一致的影响。由于各版本缙绅录中罗列官职的顺序一致，我们在上下两条相邻记录中官员姓名一致时，也会增加匹配分。当籍贯省、科名等级之类容易错误匹配的变量一致时，仅会小幅增加匹配分。

对于无姓官员，职位品级相差太大的候选匹配也会被减分，因为这有助于降低高级官员的记录被错误连接到拥有相同名字的低级官员的风险。对于更多变化的变量，如详细的科考科名或捐纳科名，匹配不一致时减分较少。由于字号很多样且经常不被记录，我们不对这个变量的不匹配施以惩罚。同样的，我们也不惩罚不匹配的职位，因为当官员晋升或被重新分配时，完整的职位称号或者组成完整职位的部分称号很可能发生变化，而且即使官员没有晋升或被调整职位，不同版本也可能以不同的方式记录职位。

当籍贯省、旗分、科名等级这类本应固定不变、多样性较低的变量不一致时，匹配分将大幅降低。如果籍贯县不一致，匹配分也会降低——当该县位于湖广或江南时，匹配分降幅更大。[①] 候选匹配记录对相隔时间较远时，我们也会扣除匹配分，如果相隔很远以至于几乎不可能为同一官的记录，我们会采用大量扣分的方式排除该候选匹配。无姓官员记录匹配中，当官职品级相差过大时，匹配分将降低——这有助于降低高级官员记录被错误连接到重名低级官员记录的风险。多样性较高的变量（如科举及捐纳科名等）不一致时，匹配度降分相对较少。由于字号记录较少，且字号可能会改变，在字号不匹配时不扣除匹配分。完整官职不匹配时，同样不减匹配分，因为当官员官职升转时，其完整官职名即会整体或局部发生变化，甚至没有升转时，不同季缙绅录中记录的完整官职也可能有所差别。

CGED-Q-ER 的内部连接（类型四）中，用于进行概率连接的变量有姓

① 由于籍贯地可能因政区重划而改变，我们在籍贯县不匹配时设定的匹配度减分额度，足以被其他变量匹配时的加分抵消。换言之，如果其他变量皆相同，候选匹配不会因为籍贯县不匹配被排除。

的SC版本、名的SC版本、籍贯省县和中式年份等。当姓的SC版本、名的SC版本、籍贯省县的组合一致时，匹配分增加，当县的PY版本或省的C版本不一致时，匹配分大幅降低。这里之所以使用名的SC版本，是因为CGED-Q-ER包含的总人数远低于CGED-Q-JSL，错连风险也相应较小。中式年份差别不大时，匹配分仅小幅降低——因为它们可能分别是同一人中举人和中进士的年份。但当中式年份差别过大时，匹配分会大幅降低。

CGED-Q-JSL和CGED-Q-ER之间的跨数据连接（类型五和类型六）中，主要使用姓、名、籍贯省县、CGED-Q-ER的中式年，以及CGED-Q-JSL的版本年等变量。分组时所用的姓和名，都是SC版本。概率连接中，如果姓和名的CV版本一致，将大幅提高匹配分；如果SC版本一致，也将小幅提高匹配分。匹配籍贯时，CGED-Q-JSL中的原籍，以及应试时的寄籍（通常在顺天）省县都参与匹配。另外，匹配时要求中式年和缙绅录首条记录年份之差，不超过30年。

（五）结果

为了说明上述方法如何在提高连接速度的同时降低错连和漏连概率，下文将进一步讨论CGED-Q-JSL内部连接结果（即类型一至类型三）。之所以详述这三种类型的连接结果，是因为它们最复杂、最具挑战性，且涉及最多主要变量和次要变量。按照上一节阐述的方法，我们对CGED-Q-JSL中的4108586条记录进行连接，最终得到326315组记录——每组即为一名官员的仕途生涯历史。表13展示了三类CGED-Q-JSL内部连接的初始待连接记录数、确定性连接后筛选出的记录组数、分组后筛选出的候选匹配对数，以及概率连接后的最终官员数。如表13所示，基于主要变量和部分次要变量的确定性连接，已能有效降低待连接记录数。在第一类连接中，确定性连接使待连接记录数减少了88.6%，从2767108条降至315015条。经历分组步骤后，最终进入评分环节的候选匹配对数量较为适中。第一类连接中，候选匹配对数量低于记录组数，这是因为在分组时已有许多记录组直接被确认为同一名官员，且该组中首条记录没有候选匹配对。第三类连接中候选匹配对数量要大得多，因为这一类别中仅有名和旗分可用于分组，而名和旗分的多样性远低于有姓官员的姓、名和籍贯省县。

表 13 内部连接三种类型的最终匹配结果

	类型一 双字名有姓	类型二 单字名有姓	类型三 无姓
初始待连接记录数	2767108	527570	813908
确定性连接筛选出的记录组数	315015	76885	171449
分组后筛选出的候选匹配对数	199263	46231	398353
概率连接产生的官员数量	218946	45965	64940

当主要变量的标准化版本有差异而次要变量一致时，概率连接可以通过协调匹配分，减少漏连现象的发生。与之相对，如果我们要求所有主变量的原始版本严格匹配（不使用概率连接），连接后的记录组就将确定属于同一位官员。表 14 为经概率连接匹配后，各类最终结果中主要变量原始版本严格一致的记录比例。据该表可得，单字名有姓官员中，28%（100-72）的记录姓、名或籍贯地不完全一致；双字名有姓官员中，29.9%（100-70.1）的记录姓、名或籍贯地不完全一致；无姓官员中，13.9%（100-86.1）名或籍贯地不完全一致。总体而言，如果既不对变量进行标准化处理也不使用概率连接，而仅在各变量原始版本上严格连接，总共可得到 453375 名官员的仕途生涯记录——在概率连接时被匹配在一起的记录将分属于不同官员。换言之，官员总数将被夸大 38%。在 CGED-Q-ER 内部，以及 CGED-Q-JSL 和 CGED-Q-ER 之间应用概率连接同样有效：能与进士记录连接的举人数大幅增加，举人和进士与 CGED-Q-JS 的连接数也大幅增加。

表 14 各类连接结果中姓、名、籍贯县或旗分（原始版本）的组合数

连接结果中主要属性（原始版本）组合数（个）	类型一双字名有姓官员（%）	类型二单字名有姓官员（%）	类型三无姓官员（%）	总计（%）
1	70.1	72.0	86.1	73.5
2	20.9	20.0	11.7	19.0
3	5.7	5.0	1.7	4.8
4	2.1	1.8	0.4	1.7
5	1.3	1.4	0.1	1.1
总计（%）	100	100	100	100
官员总数（个）	218946	45965	64940	329851

六 结语

本文不会对人名匹配和记录连接方法（尤其是无姓者的匹配连接方法）做出定论。手动检查连接结果后，我们相信目前对有姓官员的连接结果，已充分平衡了漏连和错连风险并接近最优——进一步合并存在显著差异的官员记录，将增加错连风险。今后我们还将进一步优化连接有姓官员记录的方法，比如小幅度精炼形似字列表，以及改进对籍贯省、籍贯县等变量的处理等。然而，在对旗人官员的连接上，由于名与旗分的组合缺乏多样性，我们怀疑目前仍有较大错连风险。

本文关于人名匹配和记录连接的解决方案，以及关于官员姓名的描述性分析，应该对其他开展中国历史数据库大规模记录连接的研究团队有所帮助。本文讨论的问题和连接方法，主要适用于连接从名册等史料中提取的高度结构化数据。致力于非结构化数据源（如报纸文章）研究的学者，或许会对本文关于姓名记录的多样性和潜在不一致性的分析有兴趣。需要格外注意的是，在不同数据源中，个人姓名的字符很可能被异体字或形似字代替。

团队正在努力构建、连接和分析民国时期（1911—1949）个人数据库。目前，民国大学生数据库的研究工作已经取得了重要进展，[①] 其他民国官员、专家和其他社会精英数据库项目也在同步进行。尽管基于清代史料的人名匹配和记录连接方法也可扩展应用于民国史料，但在处理民国时期的数据时，也存在其他特定问题，例如取名习惯，家谱中对名、字、号的记录习惯，籍贯地记录习惯，等等，均已发生演变。

（康文林，香港科技大学人文社会科学院署理院长、社会科学部讲座教授、华中师范大学历史文化学院特聘教授；
陈必佳，独立学者）

---

① Ren, B., Chen, L., & Lee, J. Z., "Meritocracy and the Making of the Chinese Academe, 1912-1952," *The China Quarterly*, 244, 2020: 842-968, doi: 10.1017/S0305741020001289.

图书在版编目(CIP)数据

大数据与中国历史研究．第5辑／付海晏主编．
北京：社会科学文献出版社，2025.3.--ISBN 978-7-
5228-4884-6

Ⅰ.K207

中国国家版本馆CIP数据核字第2025R2X680号

大数据与中国历史研究(第5辑)

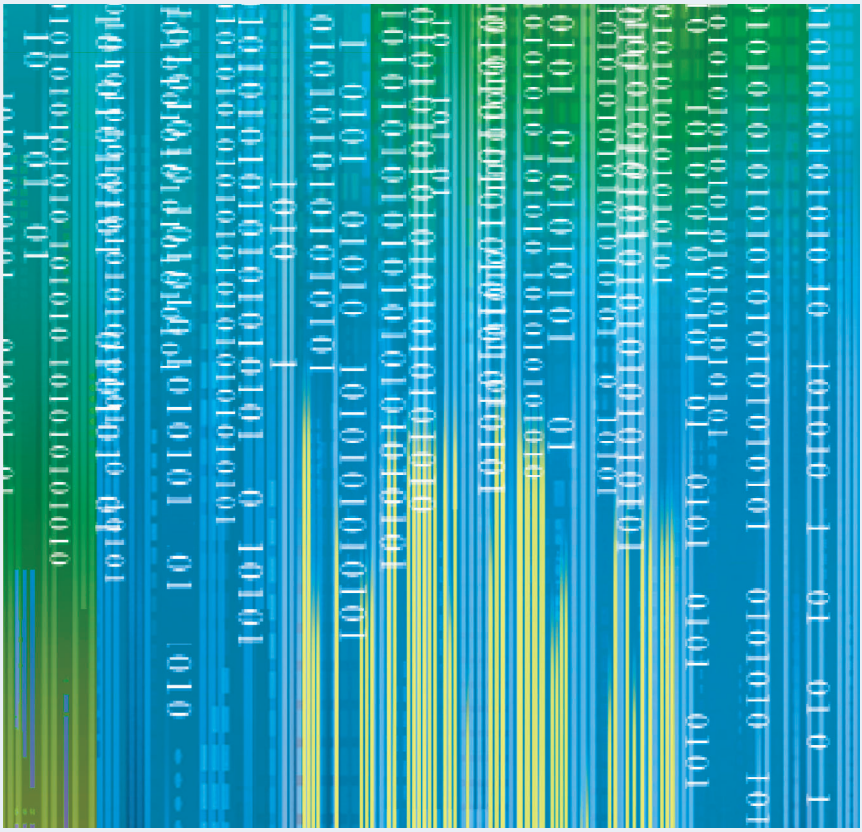
主　　编／付海晏

出 版 人／冀祥德
责任编辑／邵璐璐
责任印制／岳　阳

出　　版／社会科学文献出版社·历史学分社（010）59367256
地址：北京市北三环中路甲29号院华龙大厦　邮编：100029
网址：www.ssap.com.cn
发　　行／社会科学文献出版社（010）59367028
印　　装／唐山玺诚印务有限公司

规　　格／开 本：787mm×1092mm　1/16
印 张：14.25　字 数：229千字
版　　次／2025年3月第1版　2025年3月第1次印刷
书　　号／ISBN 978-7-5228-4884-6
定　　价／98.00元

读者服务电话：4008918866



Big Data and the Study of Chinese History

ISBN 978-7-5228-4884-6

出版社官方微信

www.ssap.com.cn

定价：98.00元