厦门大学

本科毕业论文

(科研训练、毕业设计)

题　目：明清河南地区进士时空比较的统计分析

姓　　名：王正华

学　　院：人文学院

　　系：历史系

专　　业：历史学

年　　级：2011 级

学　　号：10320112202271

指导教师（校内）：林鹭　　职称：副教授

指导教师（校外）：　　职称：

2015 年 5 月 9 日

明清河南地区进士时空比较的统计分析

【摘要】 计量史学自从上世纪七十年代兴起以来，在经济史研究中运用颇多，但在社会史研究中尚不多见。笔者运用多元统计分析的理论知识、非负矩阵分解以及 spss 等相关软件，从数学和历史学的角度出发，对明清时期河南地区进士的时空分布进行具体的探讨，并具体分析影响进士数量的相关因素。

【关键词】 计量史学　进士　统计分析　非负矩阵分解

The Comparative Statistical Analysis from the Perspective of Time and Space of Scholars in Henan, Ming and Qing Dynasties

Abstract: Since the seventies of last century, the theory of measurement historiography has began to rise. And it has been applied so much in the research of economic history, but we just can see a little application in the study of social history. In this paper, the author has probed the distribution of time and space of scholars in Henan, Ming and Qing Dynasties specifically from the perspective of mathematics and history, which uses theoretical knowledge of multivariate statistical analysis, non-negative matrix factorization and some software like SPSS. And the author has analyzed the factors of affecting the quantity of scholars specifically.

Key words: theory of measurement; scholars; statistical analysis; non-negative matrix factorization

目录

引言

科举制度是中国传统社会中选拔人才的一项重要制度，同时也是社会各阶层人员流动的一个重要桥梁。而进士人数的多少以及时空分布，可以反映出当时当地的政治经济文化水平。另外，进士在历史的具体语境中对于当时当地而言掌握着举足轻重的话语权。目前，学界关于河南地区明清时期进士的具体问题研究尚不多见。因此，研究进士的时空分布以及影响进士数量的因素是有意义的。

在历史研究中运用数学方法由来已久，只是一直未能形成系统。由于年鉴派史学家和马克思主义史学家的实践，一般认为，从 19 世纪至 20 世纪，史学经历了由“叙述的历史”到“分析的历史”的转变[1]。也就是在这一过程中，计量史学逐渐发展壮大起来。在上世纪 50 年代时期，法国年鉴学派提出了“数学化的历史学”这一概念，主要是针对当时的社会科学遇到自然科学化的挑战。上世纪中叶到目前为止，随着计算机科学、信息理论和数学理论的不断发展，计量史学逐渐成为一个系统的概念。其主要内容是把计算机科学、信息理论和数学引用到历史领域。研究方法也不断多元化，数理统计中的多元相关的测定、随机变量、数学模型、回归系数分析、趋势推论、意义度量等方法日益被广泛地运用到历史研究中。

计量史学的关注点和研究内容也是丰富多彩的。目前看来，主要包括政治史（制度评估、政府表现等）、社会史（人口、社会流动、生活水平等）、经济史（劳动生产率、税收、GDP 等）、法制史（司法审判、诉讼类型等）、教育史（教育水平、教育机会等）以及文化史（文化符号、文化传播等）等方面。计量史学的研究成果也是相当可观的。上世纪初梁启超先生便提出了“历史统计学”的概念，到了 20 世纪二三十年代，历史统计学已经成为一种学术风尚。之后，梁方仲和傅衣凌先生在进行经济史研究时，将统计学的方法纳入其中。梁方仲在其成名作《中国历代户口、田地、田赋统计》[2]一书中利用大量的表格，对中国古代社会的田赋、户口和田地数量进行统计分析，从而证明官方的户口、田地统计数字与民间实际的户口、田地数字相差甚远。之后，美国的何炳棣教授的《明初以降人口及其相关问题》[3]依靠官方记载数字，对明清以来 6 个世纪的人口数据进行了具体的分析，阐述了移民、地区经济开发、人口发展等相关问题，在人口史

研究方面可以说是里程碑式的。由于学者专业的限制，这些研究中所运用到的数学方法大多是比较浅显的，并没有涉及到更为深奥的数学方法。时至今日，随着科学技术的发展和跨学科的交流，历史研究的方法论也不断深化发展，计量史学的研究也不断推进，涌现出一大批计量史学研究的最新成果。例如香港科技大学讲座教授龚启圣在研究清末义和团时期传教士逃亡路线与“东南互保”的关系时，引入了线性回归的数学模型进行研究。2010 年香港科技大学教授李伯重先生的著作《中国的早期近代经济——1820 年代华亭—娄县地区 GDP 研究》出版，书中的附录几占一半，都是对经济数据的统计分析[4]。而且，以厦门大学和中山大学为基地，在史学界形成了中国社会经济史学派，其深入民间，进行田野调查，收集族谱、碑刻、契约文书、唱本剧本等民间历史文献，并对其中的相关经济问题进行分析，用的也大多是统计分析的计量理论。代表人物有郑振满、陈支平、王日根、刘志伟、科大卫等新一辈学者。而在北方高校，由于受西方史学理论的影响，逐渐对计量史学重视。先后在清华大学和北京大学成立了计量史学研究中心，代表人物有陈志武、龙登高、周黎安等先生。就人口史的研究而言，所用的统计学的知识更为广泛。例如复旦大学教授葛剑雄主编的《中国人口发展史》[5]、《中国人口史》[6]以及《中国移民史》[7]、香港科技大学教授李中清的《人类的四分之一：马尔萨斯的神话与中国的现实》[8]、四川学者赵文林和谢淑君在上世纪八十年代所编纂的《中国人口史》[9]等等。

当然，计量史学也正遭受到很多挑战。例如是否要将历史完全科学化？计量史学研究的问题选择应为如何？运用计量史学方法得到的结论在多大程度上是可靠的？职业的历史学家能否同时成为专业的自然科学家？如果不能，能否通过其他方式（例如信息化）来兼顾前者的修养和后者的技能？无论怎样，就目前史学发展的趋势来看，计量史学很有可能成为史学研究的下一个汇聚点之一，而其或多或少对于史学的发展都起着推动作用。

另外，目前各学科出现不断细化和专业化的倾向，跨学科的研究变得是必须的。笔者正是在如此的学术研究倾向下，试图利用多元统计分析的相关知识，对明清两代河南地区进士的时空分布和影响进士数量的因素进行分析。

一、数据的来源及研究方法

1.数据来源及处理

本文关于河南地区在明清两代进士的统计，主要依据是《皇明进士登科考》[10]、《明代登科录汇编》[11]、《明清进士题名碑录索引》[12]、顺治《河南通志》[13]、雍正《河南通志》[14]以及各个地方大量的地方志史料。在具体的区域划分上，由于明清两代之间行政区划多有变革。笔者依照“山川形变、犬牙交错”的基本原则，选定明初洪武年间对于河南地区的行政区划作为底板，将河南地区划分为9个区域，分别为开封府、河南府、归德府、汝宁府、南阳府、怀庆府、卫辉府、彰德府以及汝州。其具体包括的下级单位如下表：

开封府：	祥符	陈留	杞县	通许	太康	洧川	鄢陵	禹（钧）州	密县
	扶沟	中牟	阳武	郑州	荥阳	荥泽	河阴	汜水	原武
	陈州	商水	西华	项城	沈丘	仪封	新郑	封丘	延津
	兰阳	许州	许县	临颍	襄城	郾城	长葛	尉氏	共城
	淮宁	虞氏	应城	宣武卫	河南郡牧所		河南仪卫司		
河南府：	洛阳	偃师	巩县	孟津	宜（伊）阳	永宁	新安	渑（沔）池	卢氏
	灵宝	阌乡	登封	嵩县	陕州	阳城	河南卫	弘农卫	河南卫
归德府：	商丘	宁陵	鹿邑	夏邑	永城	虞城	睢州	考城	柘城
	归德卫	睢阳卫							
汝宁府：	汝阳	真阳	上蔡	新蔡	西平	确山	遂平	信阳	光山
	固始	息县	商城	罗山	光州	商州	正阳		
南阳府：	南阳	镇平	唐县	泌阳	桐柏	南召	邓州	内乡	裕州
	舞阳	叶县	新野	淅川	新城				
怀庆府：	河内	济源	修武	孟县	温县	武陟			
卫辉府：	汲县	胙城	新乡	获嘉	淇县	辉县	丰城	濬（浚）县	
彰德府：	安阳	临漳	汤阴	林县	磁州	武安	涉县	滑县	内黄
	临县	陟县							
汝州：	鲁山	郏县	宝丰	伊阳	汝州				

（明代曾在全国范围内设立卫所，属于独立的军事管理机构，上表中将存在于宣武卫、河南郡牧所以及河南仪卫司中的军籍进士划入开封府内，将弘农卫、河南卫中的军籍进士划入河南府内，将归德卫、睢阳卫中的军籍进士划入归德府内）

时间分段的问题，笔者以不同的皇帝来进行划分，主要是考虑到政策的变动，明清两代共 27 朝，本研究便将时间划分为 27 个时间段，即从明洪武至清光绪。

2. 多元线性回归与非负矩阵分解基本原理

研究方法上，笔者首先是对所得的数据进行描述性分析，即通过图表的形式来观察具体的分布和波动情况。另外，对不同地区进士数量变化的均值、方差、标准、偏度以及峭度等基本的数据信息进行处理。

另外，对进士数量、对应人口数量、书院分布情况以及前代进士数量（前两朝进士之和）数据进行抽样处理，利用 spss 软件构建以进士数量为因变量的多元线性回归模型，估计回归参数，对所得回归方程进行显著性检验，求解各个因变量之间的偏相关系数等，具体分析影响进士数量的因素。

一个因变量多个自变量的回归模型的基本理论如下[15]：

设随机变量 y 与一般变量 $x_1, x_2, \cdots, x_p$ 的线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

其中，$\beta_0, \beta_1, ..., \beta_p$ 是 $p+1$ 个未知参数，β_0 称为回归系数，$\beta_1, ..., \beta_p$ 称为回归系数。y 称为被解释变量（因变量），而 $x_1, x_2, \cdots, x_p$ 是 p 个可以精确测量并可控制的一般变量，称为解释变量（自变量）。ε 是随机误差，我们对其假定

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{var}(\varepsilon) = 0 \end{cases}$$

这样，$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ 为理论回归方程。

在一个实际问题中，我们获得 n 组观测数据 $(x_{i1}, x_{i2}, ...x_{ip}; y_i) i = 1, 2, ..., n,$ 则线性回归模型可表示为：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_p x_{2p} + \varepsilon_2 \\ ..... \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_p x_{np} + \varepsilon_n \end{cases}$$

写成矩阵形式为：

$$y = X\beta + \varepsilon$$

其中，

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},\quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},\quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

矩阵 X 是一 $n \times (p+1)$ 矩阵，称 X 为回归设计矩阵或资料矩阵。

为了有效地进行模型的参数估计，对上述回归模型有一些基本假定：

1. 解释变量 $x_1, x_2, \cdots, x_p$ 是确定性变量，不是随机变量，且要求 $rk(X) = p+1 < n$。这里的 $rk(X) = p+1 < n$,表明设计矩阵 X 中的自变量列之间不相关，样本容量的个数应大于解释变量的个数， X 是一满秩矩阵。

2. 随机误差项具有 0 均值和等方差，即

$$\begin{cases} E(\varepsilon_i) = 0, i = 1, 2, \cdots, n \\ \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases} (i, j = 1, 2, \cdots, n) \end{cases}$$

这个假定常称为 Gauss-Markov 条件。 $E(\varepsilon_i) = 0$, 即假设观测值没有系统误差，随机误差 ε_i 的平均值为零。随机误差项 ε_i 的协方差假定表明随机误差项在不同的样本点之间是不相关的（在正态假定下即为独立的），不存在序列相关，并且有相同的精度。

3. 正态分布的假定条件为：

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \cdots, n \\ \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{相互独立} \end{cases}$$

对于多元线性回归的矩阵形式，这个条件便可表示为：

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

由上述假定和多元正态分布的性质可知，随机向量 y 遵从 n 维正态分布，回归模型的期望向量

$$E(y) = X\beta$$
$$\mathrm{var}(y) = \sigma^2 I_n$$

因此， $y \sim N(X\beta, \sigma^2 I_n)$ 。

非负矩阵分解基本原理[16, 17]：

非负矩阵分解算法的基本原理是对于任意给定的一个非负矩阵 V ，非负矩阵分解算法能够找到一个非负矩阵 W 和一个非负矩阵 H ，满足

$$V_{n\times m} \approx W_{n\times r} H_{r\times m}$$

一般，选取的 r 值要远小于 n 和 m 。非负矩阵分解将一个非负的矩阵近似分解为两个非负矩阵的乘积。由于分解前后的矩阵中仅包含非负的元素，因此，原矩阵 V 中的列向量可以解释为对基矩阵 W 中所有列向量（成为基向量）的加权和，而权重系数为系数矩阵 H 中对应列向量中的元素。非负矩阵的求解过程实际上是一个优化过程，即通过迭代使 V 和 WH 之间的重构误差最小。算法的关键步骤是目标函数的设定和迭代规则的选择。实验中，采用的目标函数和算法如下：

目标函数：最小化 $\| V - WH \|_F^2$ 。迭代法则：

$$H_{kj} := H_{kj} \frac{(W^T V)_{kj}}{(W^T WH)_{kj}}$$

$$W_{ik} := W_{ik} \frac{(VH^T)_{ik}}{(WHH^T)_{ik}}$$

其中 $(V)_{ij}$ 表示取 V 的第 (i, j) 元。

非负矩阵分解算法基于基向量组合的表示形式具有很直观的解释，表达了原始数据基于部分的表现形式，体现了局部组成整体的思想，所获得的数据是原始数据加性的，非负的组合，算法所得到的非负基向量组 W 具有一定的线性无关性和稀疏性，能有力表达原始数据的特征及结构。选取合适的 r 值，则可获取原始数据的特征，且其非负约束符合人文学科研究的条件，因此，可以被引入应用于人文学科的研究。

二、明清河南地区进士数量变化的时空分析

据笔者根据相关史料统计，明清两代河南地区共中进士 2971 名，明代进士共 1430 名，清代进士共 1541 名。具体数字详见下表：

地点 时间	开封府	河南府	归德府	汝宁府	南阳府	怀庆府	卫辉府	彰德府	汝州	总计
洪武	15	5	5	5	0	4	2	4	2	42
建文	4	0	0	0	0	0	0	0	0	4
永乐	31	6	8	8	6	3	3	3	2	70
洪熙	0	0	0	0	0	0	0	0	0	0
宣德	17	0	1	3	4	0	0	4	1	30
正统	13	3	2	13	4	0	3	0	1	39
景泰	12	4	0	3	0	2	2	2	1	26
天顺	15	4	3	7	2	2	2	2	0	37
成化	56	15	14	34	14	6	8	13	4	164
弘治	23	5	7	16	3	2	10	5	1	72
正德	33	11	4	24	10	2	2	5	2	93
嘉靖	83	25	22	66	18	9	19	16	3	261
隆庆	17	5	7	9	3	0	0	3	1	45
万历	96	45	65	97	26	12	12	13	8	374
泰昌	0	0	0	0	0	0	0	0	0	0
天启	9	6	5	9	1	2	3	1	1	37
崇祯	39	17	25	31	4	3	7	9	1	136
顺治	76	25	44	24	4	33	14	22	5	247
康熙	79	20	81	22	9	31	16	21	5	284
雍正	24	15	15	4	3	6	4	3	0	74
乾隆	71	22	45	44	7	33	10	11	7	250
嘉庆	37	7	14	48	2	4	4	6	4	126
道光	52	13	18	35	9	12	6	7	6	158
咸丰	28	11	6	18	4	5	9	8	3	92
同治	29	4	9	27	9	12	0	11	1	102
光绪	66	8	14	61	18	14	10	17	0	208
宣统	0	0	0	0	0	0	0	0	0	0
总计	925	276	414	608	160	197	146	186	59	2971

表 1

1.地区分析

根据上图以时间为横坐标，各地区进士数量所占比例为纵坐标，绘制柱形图如下：

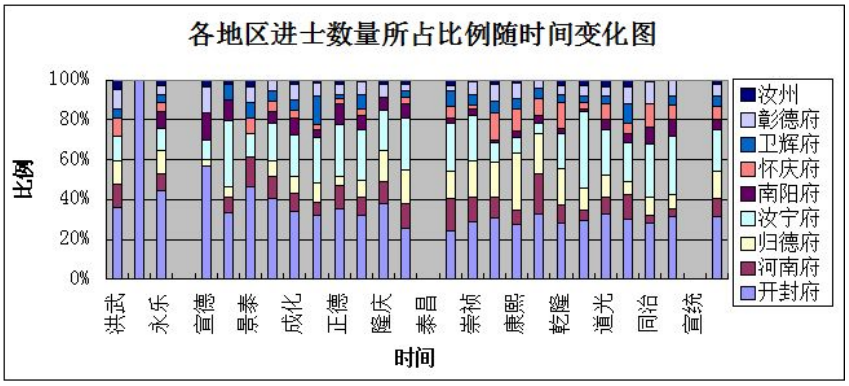

图 1

从上图可以看出，洪熙、泰昌和宣统年间，各地均无中进士。洪熙帝在位仅 8 个月，而泰昌帝在位仅 1 个月，在此期间并无科举进行，自然无进士出现。而宣统帝 1909 年到 1912 年、1917 年 7 月 1 日到 1917 年 7 月 12 日先后两次在位，但是科举考试在 1905 年便被清王朝废除，因而宣统年间也无进士出现。

就各地区进士数量所占比例随时间变化来看，各地区进士数量的多少大致从多到少排列为：开封府、汝宁府、归德府、河南府、南阳府、怀庆府、卫辉府、彰德府、汝州。这个排序和最后的进士总数在各地区的分布大致相同。开封府作为明清两代河南的省治所在，是政治经济中心，且其曾是古都，文化重镇，其对于教育事业十分重视，当地文风盛行，因而开封地区的进士数量最为居多。而汝州地区进士数量比重最为少的缘由很大一部分是由于其行政区划所占土地面积和人口在河南地区是最少的，因而进士数量较少。

2. 时段分析

反过来思考，我们以各地区为横轴，以进士数量在各时段的比重为纵轴得下图：

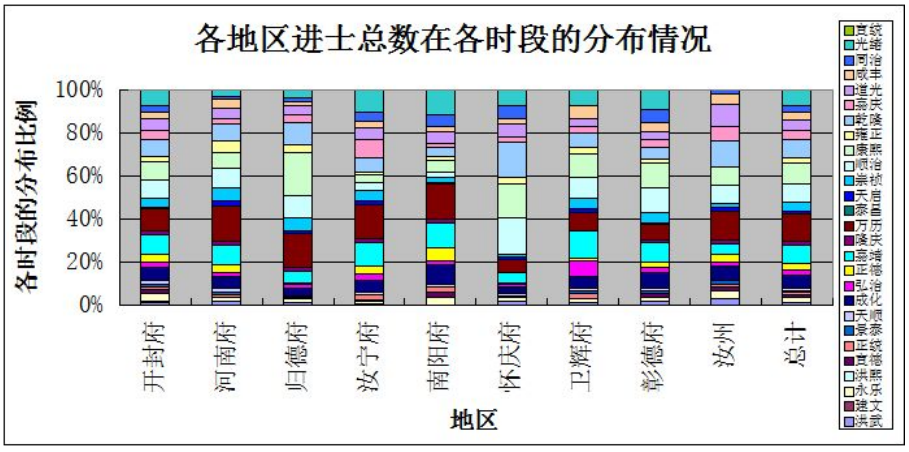

图 2

从上图可以看出，中进士较多的几个时期是成化、嘉靖、万历、顺治、康熙、乾隆和光绪帝。成化帝在位 23 年，嘉靖帝在位 45 年，万历帝则在位 48 年，顺治帝在位 18 年，康熙帝则在位 61 年，乾隆帝在位 60 年，光绪帝在位 34 年。这些帝王时期所中进士总数也是相对较多的。这些帝王在位时间在众多帝王中不可谓不久，因而也就可以解释为何会出现如此多的进士了。另外，除了光绪帝外，另外几个帝王时期都是处于治世或者盛世时期，政治昌明，经济发展，对于文化教育方面也更加重视。

从各时段的比例来看，明清两代，时段分布比例上均出现了两边少中间多的情况。王朝初期，往往刚刚经历战火，重在与民休息，发展农业，因而科举事业并不发达。王朝中期，进入治世或者盛世时期，文化事业便也随之发达起来。王朝末期，内忧外患，王朝面临灭亡之际，科举也便随之衰落，虽然会出现一些破例特招之进士，但始终是少数。由此可见，在一定意义上，科举的兴衰和王朝的兴衰是同步的。

另外，我们利用非负矩阵分解的方法来探讨就各地而言，进士数量在各时段所占比重的情况。将表 1 作为矩阵，然后利用 MATLAB 软件应用 MU 算法对其进行矩阵分解，r 取 4，迭代 500 次，结果如下：

W：

0.023429984	0.16681988	0.442178474	0.163231746
0	1.96E-36	0.13453674	0.005454182
0.066629078	0.023906587	0.961425586	0.369228717
0	0	0	0
1.24E-33	6.88E-13	0.670146129	0.013638762
0.495091682	3.89E-23	0.301452809	0.102754744
1.18E-27	2.89E-28	0.478840111	0.05324322
0.161294713	1.64E-05	0.467787461	0.153765535
1.005545352	0.024529362	1.625175204	0.640890146
0.463424348	0.055922242	0.643902891	0.27259897
0.802385707	1.84E-24	0.951336885	0.26808205
2.217201678	0.005980145	2.142834125	1.043140293
0.299328464	0.001612892	0.330979914	0.339451688
3.99609114	0.744206368	0.256832508	2.777878948
0	0	0	0
0.343797548	0.024049655	0.079088485	0.272492385
1.145240013	0.264479968	0.328729931	1.07873269

0.001685672	1.941856761	1.64653415	0.970474743
0.158824782	2.904666081	0.271228232	1.954600687
0.0003587	0.133943626	0.37790744	0.790596785
1.141780795	2.040007825	0.929537596	0.902591798
1.86259314	0.402161279	0.459318847	0.309252566
1.061520529	0.547077846	1.185516436	0.510550856
0.473993834	0.000124517	0.89967043	0.340071779
0.855226333	0.636028713	0.691041308	1.90E-12
1.997613976	0.70690446	1.879751314	3.95E-06
0	0	0	0

H：

7.886196417	1.461472077	1.098978872	22.26622445	4.004969334	2.22E-10	0.007156223	0.528792156	0.603573775
12.20256527	5.72E-05	14.57918946	5.305032419	4.27E-10	12.02096627	1.388254573	4.403192663	1.099642723
22.21366196	4.205397591	3.56E-05	7.025249431	3.41281263	3.982722649	4.399972007	5.651783282	0.775228437
18.03978677	13.08450106	18.14437478	0.969435486	2.332639416	0.132784629	4.327766512	2.299360123	1.451772815

表 2

我们将W中的每一列视为虚拟的四个村庄，每一行为每个时期。可以看出，开封府的特征可以由村庄 3 和村庄 4 表现出来，河南府的特征可以由村庄 4 来表现出来，归德府的特征可以由村庄 2 和村庄 4 来表现出来，汝宁府的特征可以由村庄 1 来表现，南阳府的特征可由村庄 1 和村庄 3 来表现，怀庆府的特征可由村庄 2 来表现，卫辉府的特征可由村庄 3 和村庄 4 来表现，彰德府的特征可由村庄 2 和村庄 3 来表现，汝州的特征可由村庄 2 和村庄 4 来表现。而就 W 而言，村庄 1 中，嘉靖和万历年间的数据特征较为明显；村庄 2 中，顺治、康熙和乾隆年间的数据特征较为明显；村庄 3 中，成化、嘉靖、顺治、道光和光绪年间的数据特征较为明显；村庄 4 中，嘉靖、万历、崇祯和康熙年间的数据特征较为明显。结合两者分析，可以看出，开封府进士数量特征较为明显的是成化、嘉靖、万历、崇祯、顺治、康熙、道光和光绪年间；河南府的进士数量特征较为明显的是嘉靖、万历、崇祯和康熙年间；归德府的进士数量特征较为明显的是嘉靖、万历、崇祯、顺治、康熙和乾隆年间；汝宁府的进士数量特征较为明显的是嘉靖和万历年间；南阳府的进士数量特征较为明显的是成化、嘉靖、万历、顺治、道光和光绪年间；怀庆府的进士数量特征较为明显的是顺治、康熙和乾隆年间；卫辉府的进士数量特征较为明显的是成化、嘉靖、万历、崇祯、顺治、康熙、道光和光绪年间；彰德府的进士数量特征较为明显的是成化、嘉靖、顺治、康熙、

乾隆、道光和光绪年间；汝州的进士数量特征较为明显的是嘉靖、万历、崇祯、顺治、康熙和乾隆年间。上述各地区的进士数量特征表现明显的时段即是相对于其他时段而言进士数量较多的时期，将其与表 1 中的数据进行对照，基本一致，由此可见，分解的结果相对较好。

另外，将表现四个村庄数据特征较好的时期汇总在一起，便是：成化、嘉靖、万历、崇祯、顺治、康熙、乾隆、道光和光绪。这和图 2 所呈现出的状况基本一致。其中崇祯年间的进士数量较多，主要是由于当时明王朝处于内忧外患之际，为了应对各种危机，特招了很多进士。而道光年间的进士数量较多，也主要是针对开封、南阳、卫辉和彰德四府而言，且其特征性在表现时并非十分突出。另外，道光末年，鸦片战争爆发，中国战败，中国面临着从未有过的内忧外患，当时也对进士进行了大量特招以应付清王朝所面临的危机。

这是从进士数量上来看，从具体的数量变化波动来分析，我们又可以得到其他的结论。以时间为横轴，各地对应的进士数量为纵轴画折线图和平滑曲线图分别如下：

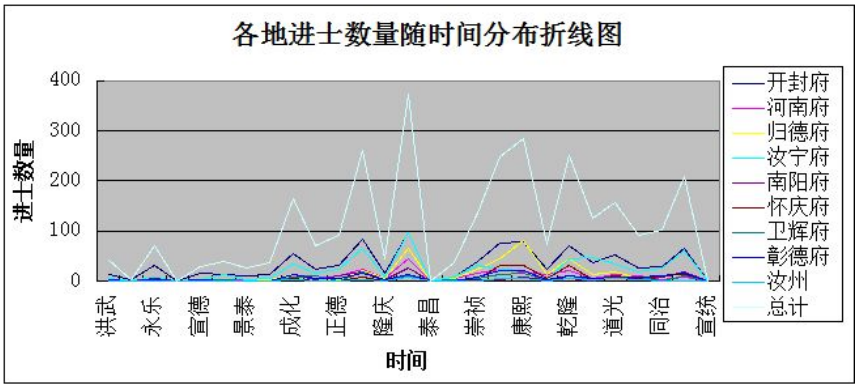

图 3

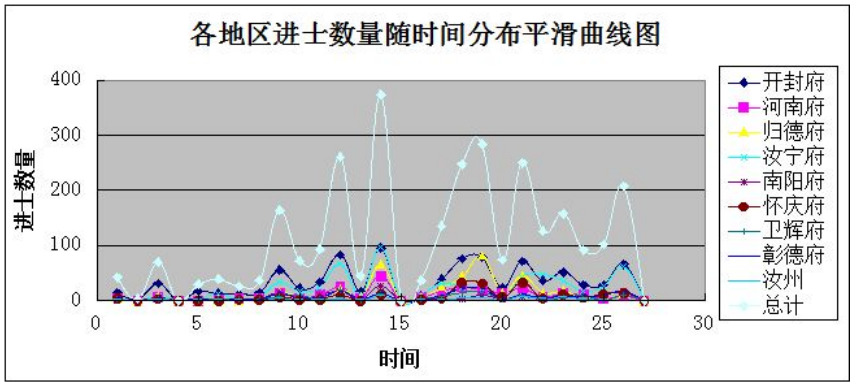

图 4

从图 3 和图 4 所反映的升降趋势来看，各府的进士数量随时间变化的升降趋势大致呈现出一致的状况。这证明了对于不同的区域，影响进士数量的因素中存在某些共同。而且上图也证明了在明中期和清中期时分进士的数量达到最高点。

从具体的运算来看，我们用均值、方差、标准差、变异系数、偏度和峭度来具体分析各个区域进士数量的分散程度。

样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$

样本标准差 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$

变异系数 $V=\frac{S}{\bar{X}}$

偏度系数 $V_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^3}{S^3(n-1)}$

峭度系数 $V_2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^4}{s^4(n-1)}$

通过计算得下表：

	均值	方差	标准差	变异系数	偏度系数	峭度系数
开封府	34	799.123	28.269	0.825	0.678	2.210
河南府	10	107.872	10.386	1.016	1.530	5.469
归德府	15	423.231	20.573	1.341	1.878	5.706
汝宁府	23	570.182	23.878	1.060	1.406	4.606
南阳府	6	42.917	6.551	1.106	1.463	4.561
怀庆府	7	99.140	9.957	1.365	1.722	4.762
卫辉府	5	29.712	5.451	1.008	0.871	2.712
彰德府	7	42.949	6.554	0.951	0.867	2.648
汝州	2	5.387	2.321	1.062	1.030	2.936

表 3

由于各个地方的进士数量随着时间变化的均值是不同的，因此我们不能直接比较其标准差来分析数据的离散程度。从标准差来看，各地的标准差存在很大差异。但从变异系数 V，我们可以看出，所有的 V 值都接近于 1，各地随时间分布

的进士数量的分散程度大致是相同的。

从偏度系数来看，V_1的值都大于零，因此各地在明清时代的前中期进士数量都呈陡然上升的趋势，之后稍微缓慢下降。将V_2的值和 3 进行比较，可以发现，开封府、卫辉府、彰德府和汝州的进士数量变化并不太陡峭，而河南府、归德府、汝宁府、南阳府和怀庆府的进士数量变化十分陡峭。

三、进士数量影响因素的回归分析

1. 建立线性回归模型

上述内容证明了在影响进士数量的因素中，各地区是存在某些共通点的。笔者大致列出土地（衡量经济因素）、人口、书院（文化事业）、前代进士（文风）、政策等因素。但是，明清两代的政府土地统计状况严重不可靠，甚至出现从开国初到国灭，土地数字未变的状况，因而无法对土地数字进行运用。而且，政策因素也无法进行量化。最后，笔者选定人口、书院数量和前代进士数量三个因素进行回归分析，其中，前代进士数量主要是前两代的进士数量。另外，数据也主要是进行抽样选定的，时间选定为洪武、永乐、成化、嘉靖、乾隆、嘉庆、道光和光绪。人口数据的来源主要是曹树基先生的《中国人口史》以及梁方仲先生的《中国历代户口、田地、田赋统计》。得到的数据如下表：

进士数量	人口数量（万）	书院数量	前代进士数量（两代）
15	118.3339	0	10
5	52.8567	0	3
4	13.2015	0	2
2	10.0714	0	1
4	19.669	0	2
0	11.6977	0	0
5	18.3123	0	3
2	12.9362	0	1
31	113.3722	0	19
6	52.3952	1	5
3	14.4013	0	4

3	16.2488	0	2
3	28.0098	0	4
6	10.9633	0	0
8	20.8592	0	5
2	13.1441	0	2
56	204.6702	4	27
15	73.8227	3	8
13	32.0096	0	4
8	20.2618	1	4
6	43.53	1	4
14	32.2023	5	2
34	44.7375	1	10
4	25.2734	2	1
83	203.8542	16	56
25	78.099	6	16
16	34.1104	2	10
19	19.5647	1	12
9	42.1896	2	4
18	38.8433	11	13
66	52.9103	8	40
3	33.7517	7	3
71	581.9	57	103
22	192.1	47	35
11	108.6	6	24
10	148.8	14	20
33	150.8	21	37
7	353.4	22	12
44	433.9	22	26
7	69.7	7	5
45	275.8	11	96
37	693.7	57	95
7	224.8	49	37
6	136.8	6	14
4	175	14	14
4	180.3	22	39
2	421.3	25	10
48	506	24	48
4	83.1	8	7
14	328.8	11	60

52	779.1	66	108
13	249.9	50	29
7	159.7	6	17
6	196.2	15	14
12	204.7	24	37
9	472.4	26	9
35	556.1	26	92
6	92.6	13	11
18	366.4	13	59
66	544.6	77	57
8	204.3	55	15
17	117.6	6	19
10	135.2	20	9
14	152.5	30	17
18	518.1	27	13
61	600.1	31	45
0	77.6	13	4
14	271.8	13	15

表 4

以进士数量为因变量 y，人口数量、书院数量以及前代进士数量分别为因变量 x_1、x_2 和 x_3。通过 spss 软件计算增广相关阵以及自变量的相关阵。输出结果如下：

相关性					
		进士数量	人口数量	书院数量	前代进士数量
进士数量	Pearson 相关性	1	.568**	.453**	.708**
	显著性（双侧）		.000	.000	.000
	N	68	68	68	68
人口数量	Pearson 相关性	.568**	1	.774**	.772**
	显著性（双侧）	.000		.000	.000
	N	68	68	68	68
书院数量	Pearson 相关性	.453**	.774**	1	.654**
	显著性（双侧）	.000	.000		.000
	N	68	68	68	68
前代进士数量	Pearson 相关性	.708**	.772**	.654**	1
	显著性（双侧）	.000	.000	.000	
	N	68	68	68	68
**. 在 .01 水平（双侧）上显著相关。					

从相关阵看出，y 与 x_1, x_2, x_3 的相关系数分别是 0.568,0.453 和 0.708，说明所选自变量是与 y 存在线性相关的，用 y 与自变量做多元线性回归是合适的。y 与 x_2 的相关系数 $r_{y2} = 0.453$ 偏小，这说明书院数量对进士数量无特别显著影响。那么在回归方程中是否还应该包含 x_2？仅凭简单相关系数的大小是不能决定变量的取舍的，在初步建模时还是应该包含 x_2 在内的。用 spss 软件对原始数据进行回归分析，结果如下：

输入／移去的变量 [b]

模型	输入的变量	移去的变量	方法
1	前代进士数量, 书院数量, 人口数量	.	输入

a. 已输入所有请求的变量。
b. 因变量: 进士数量

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.710[a]	.504	.481	14.19456

a. 预测变量: (常量), 前代进士数量, 书院数量, 人口数量。

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	13112.395	3	4370.798	21.693	.000[a]
	残差	12895.075	64	201.486		
	总计	26007.471	67			

a. 预测变量: (常量), 前代进士数量, 书院数量, 人口数量。
b. 因变量: 进士数量

系数 [a]

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	6.185	2.396		2.581	.012
	人口数量	.010	.017	.093	.556	.580
	书院数量	-.066	.151	-.062	-.439	.662
	前代进士数量	.497	.103	.677	4.841	.000

a. 因变量: 进士数量

表 5

进行回归诊断：

1. 回归方程为

$$\hat{y} = 6.185 + 0.01x_1 - 0.066x_2 + 0.497x_3$$

2. 复相关系数 R=0.710，决定系数 R^2=0.504，由决定系数可以看出回归方程是显著的。

3. 方差分析表，$F = 21.693 > F_{0.05}(3,64)$，$p$值=0. 000，表明回归方程很显著，说明 x_1, x_2, x_3 整体上对 y 有显著的线性影响。

4. 回归系数的显著性检验。自变量 x_1, x_2 对 y 的影响不如 x_3 更为显著。

通过对自变量分别为 $(x_1),(x_1,x_2),(x_1,x_3),(x_2),(x_2,x_3),(x_3)$ 的模型进行回归分析，发现单独以 x_3 为自变量进行回归分析的效果最好，结果如下：

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.708[a]	.502	.494	14.01411

a. 预测变量: (常量), 前代进士数量。

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	13045.385	1	13045.385	66.424	.000[a]
	残差	12962.085	66	196.395		
	总计	26007.471	67			

a. 预测变量: (常量), 前代进士数量。

b. 因变量: 进士数量

系数[a]

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	6.405	2.223		2.881	.005
	前代进士数量	.520	.064	.708	8.150	.000

a. 因变量: 进士数量

表 6

对其进行回归诊断，发现 x_3 对 y 具有显著的影响。也就是说前代的进士数量

对后来的进士数量具有明显的影响，而人口数量和书院数量对其影响较小。那么，为何会出现这种情况呢？

2.模型结果的历史分析

首先，人口基数大，确实会增加中进士的概率。但是，中国传统社会可以说是一个农耕社会，人口的增加，带来的更多的是负担，吃饭问题亟待解决。因而对于普通民户而言，人口的增加，更多的是需要开垦土地，发展农业。对于科举入仕，考取功名反而没有那么重视。另外，人口的增加并不意味着文化人的增加，更多的人还是选择“男耕女织”的生活方式。

再者，就是书院数量的问题。书院的多少和进士数量之间的关系呈现出一种不显著的状态，这确实是有悖于常识。笔者认为这其中有两个因素。其一是有可能二者的因果关系出现了互置，即是进士数量影响了之后的书院的多少，而书院的多少再影响进士数量。其二便是书院一般都是由官方进行建设的，其规模是有限的，从书院的数量存在很多 0 因子以及变化很小可以看出这点。而对于普通百姓而言，读书的途径更多的是选择私塾教育，先考过童试，才有资格继续参加国家举行的乡试、会试和殿试。由于私塾大多存在于民间，不见于官方文献记载之中，因而对于私塾数量的统计很难做到。

最后，便是为何前代的进士数量对于后来的进士数量影响如此之大呢？第一个方面便是对于地方风气的影响。前代学子考中进士，便会激励后来学子沿其道路前进。例如位于河南进士最多之地开封府的祥符县内，进士数量一直居于各县之首，前代学子对后代学子的影响不能忽略。另外，前代进士考中功名，进入仕途，对于普通百姓和学子的激励作用更为明显。科举入仕，便成为当时耕读世家不错的选择。例如当时归德府地区的沈鲤，嘉靖四十四年中进士，拜东阁大学士，官拜礼部尚书。在明中期，曾有人云：“满朝文武半江西，小小归德四尚书”。可见，进士的地域性极为明显，这也正证明了前代进士对于后代进士的激励作用。另外一点，便是家族势力的影响。能够安心读书，去考取功名的，往往家境殷实，在地方上便表现为地方大族。而这些大族中往往出现兄弟、父子同时或者先后中进士的情况。例如河南孟津的王无咎、王无忝两兄弟，分别在顺治 3 年和康熙 9 年考取进士；而河南洛阳中护卫的温如春和温如璋两兄弟分别在明嘉靖 32 年和嘉靖 35 年考取进士；河南嵩县的董相及其两子董遂、董选分别在明正德 6 年、

明嘉靖 29 年和明隆庆 5 年考中进士。这些例子在数据当中大量存在，这便证明了前代进士为何对后代进士数量产生如此大的影响。

另外，还有一个问题很需要注意的，便是数据源的问题。一来，由于主观或者客观的原因，历史中的因素有很多我们无法量化，包括经济、政策等方面，因而对于自变量的数据收集并不全面，有很多方面无法进行处理，这也是造成模拟效果不佳的一个重要原因。二来，便是对于本身收集到的数据的处理问题，很多数据本身存在一些问题，不仅是技术上的处理，更多是历史的处理，如果处理不当，有可能会出现很多错误。

四、结论

笔者通过运用多元统计分析的相关知识和 spss 软件等，对明清两代河南地区进士数量的时空分布以及影响进士数量的因素进行了具体的探讨。得出如下结论：

1.就各地区进士数量所占比例随时间变化来看，各地区进士数量的多少大致从多到少排列为：开封府、汝宁府、归德府、河南府、南阳府、怀庆府、卫辉府、彰德府、汝州。

2.中进士较多的几个时期是成化、嘉靖、万历、顺治、康熙、乾隆和光绪帝。

3.各地中进士较多的时期分别为：

开封府	成化、嘉靖、万历、崇祯、顺治、康熙、道光、光绪
河南府	嘉靖、万历、崇祯、康熙
归德府	嘉靖、万历、崇祯、顺治、康熙、乾隆
汝宁府	嘉靖、万历
南阳府	成化、嘉靖、万历、顺治、道光、光绪
怀庆府	顺治、康熙、乾隆
卫辉府	成化、嘉靖、万历、崇祯、顺治、康熙、道光、光绪
彰德府	成化、嘉靖、顺治、康熙、乾隆、道光、光绪
汝州	嘉靖、万历、崇祯、顺治、康熙、乾隆

4.科举的兴衰和王朝的兴衰是同步的。

5.对于不同的区域，影响进士数量的因素中存在某些共通。

6.各地随时间分布的进士数量的分散程度大致是相同的。

7.前代的进士数量对后来的进士数量具有明显的影响，而人口数量和书院数量对其影响相对较小。文化和家族因素对于进士数量的影响更大。

致谢语：

怀着惴惴不安的心情，论文最终定稿了。文章水平虽然不如数学学院同学的高，但自身已经尽力而为，问心无愧。

在论文写作期间，林鹭导师在很多方面给予了我悉心的指导和帮助。从最初的选题到最终的定稿，导师都付出大量的时间和精力，每周都会专门挤出时间就论文的问题和我交流。由于本人能力有限，对文章中的一些问题经常难以理解，但导师总会耐心地给我讲解，直到我彻底明白。同时，导师乐观豁达的生活观和严谨的治学态度亦使我受益匪浅，真正的可以说是我的一位良师益友。在此，谨向导师致以深深的谢意，我会永远铭记这段岁月。

感谢我的大学同学林婵娟、步凡、刘伟强在文章内容上对我的帮助。也感谢我的同乡王晓栋同学每天都陪我一起去图书馆查阅资料，撰写论文。这段友情将成为我人生中最美好的回忆之一。

最后，我要感谢我的父母多年来对我学业的支持。他们的理解和支持是我在人生和学业上不断追求前进的最大动力。

王正华

2015年5月9日

参考文献：

【1】王旭东.20 世纪历史学传统嬗变和方法论的计量化[J].甘肃社会科学，2013 年第 5 期，68—70.

【2】梁方仲.中国历代户口、田地、田赋统计[M].北京：中华书局，2008 年 11 月.

【3】（美）何炳棣著，葛剑雄译.明初以降人口及其相关问题[M].上海：生活·读书·新知三联书店，2000 年 11 月.

【4】李伯重.中国的早期近代经济：1820 年代华亭—娄县地区 GDP 研究[M].北京：中华书局，2010 年.

【5】葛剑雄.中国人口发展史[M].福建：福建人民出版社，1991 年 6 月.

【6】葛剑雄.中国人口史[M].上海：复旦大学出版社，2002 年.

【7】葛剑雄：中国移民史[M].福建：福建人民出版社，1997 年 7 月.

【8】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[M].上海：三联书店，2000 年 1 月.

【9】赵文林/谢淑君.中国人口史[M].北京：人民出版社，1988 年.

【10】（明）俞宪.皇明进士登科考[M].台湾：学生书局，1969 年 12 月.

【11】屈万里.明代登科录汇编[M].台湾：学生书局，1969 年.

【12】朱保炯/谢沛霖.明清进士题名碑录索引[M].上海：上海古籍出版社，1979 年 10 月.

【13】（清）沈荃/贾汉复纂修.顺治河南通志[M].江苏·上海·四川：凤凰出版社·上海书店·巴蜀书社，2011 年 8 月.

【14】（清）田文镜修/张灏纂.雍正河南通志[M].1914 年河南教育司印历次补修本.

【15】何晓群.应用多元统计分析[M].北京：中国统计出版社.

【16】D.D. Lee, H.S. Seung. Algorithms for non-negative matrix factorization [J], Advances in Neural, Information Processing Systems, 3 (2001), 556 - 562.

【17】D.D. Lee, H.S. Seung. Learning the parts of objects by non-negative matrix factorization [J], Nature, 401 (1999), 788 - 791.

【18】何晓群/刘文卿.应用回归分析（第二版）[M].北京：中国人民大学出版社，2007 年 7 月.

【19】马金生/李宏.中国大陆“计量史学”现状的本土化反思[J].广播电视大学学报（哲学社会科学版），2009 年第 2 期，87—91.

【20】王瀛培.计量史学研究综述——数学统计、计算机与历史研究的结合[J]. 池州学院学报，2011 年第 1 期，107—109.

【21】王洪瑞/吴宏岐.明代河南书院的地域分布[J].中国历史地理论丛，2002 年第 4 辑，86—102.

【22】王洪瑞.清代河南书院的地域分布特征[J].史学月刊，2004 年第 10 期，96—105.

【23】刘维湘/郑南宁/游屈波.非负矩阵分解及其在模式识别中的应用[J].科学通报，2006 年第 3 期，241—250.

【24】关永强.从历史主义到计量方法：美国经济史学的形成与转变（1870—1960）[J].世界历史，2014 年第 4 期，114—123.

【25】瞿宁武.计量经济史学评介[J].中国经济史研究，1992 年第 2 期，147—154.

【26】王爱云.计量史学方法在当代中国史研究中的运用[J].当代中国史研究，2013 年第 6 期，94—102.

【27】温长松.试述抽样调查方法在历史研究中的应用[J].沈阳大学学报，2006 年第 1 期，40—41.

【28】袁山.统计分析工具在历史研究中的应用[J].中国青年政治学院学报，2002 年第 4 期，88—93.