基于贝叶斯模型的江浙沪地区人口迁入驱动机制研究

吴雯静，叶绮霖

（南京大学地理与海洋科学学院，南京 210023）

摘要：改革开放后，中国放宽户籍政策并推进城镇化发展，人口由计划迁移转向自主迁移。江浙沪地区的迁入人口规模显著上升，研究其背后的驱动机制至关重要。基于“六普”中省际人口迁移数据，选取常住人口规模、人均GDP、就业人员平均工资、人均医疗床位数、大专以上学历人口及距离等变量，构建多元线性回归模型，并引入分层贝叶斯降低参数估计的不确定性。结果表明，经济因素为首要驱动力，其次是人口、教育和距离因素，而医疗因素影响较小。

关键词：贝叶斯模型；人口迁移；江浙沪地区；驱动机制

中图分类号：K901.3；C922 **文献标志码**：A **文章编号**：1671－1807(2025)14－0179－10

中国实施改革开放后，随着户籍政策的改革和城镇化进程的推进，人口迁移模式从原先的计划迁移转为以自发迁移为主[1]，到20世纪90年代，省际人口迁移规模快速扩大[2]，人口迁移表现为东南沿海集中化，逐步形成江浙沪地区、珠三角地区、京津冀地区等为主的人口流入区域[3]。人口迁移作为人口在空间上的再分布过程，推动着人力、经济等社会经济资源的流动，改善着生产要素的合理优化配置，有力地促进着国民经济增长和社会进步[4]。

江浙沪地区是中国经济发展活跃度最高，城市、产业分布最密集的区域，在国民经济发展中凸显出至关重要的影响力，同时也是主要的人口迁入聚集区域[5]。根据1990年第四次全国人口普查中数据统计显示，1985—1990年中国（未包括港澳台地区）省际迁移人口总规模为1 123.65万人，其中迁入江浙沪地区的人口数为179.25万人，占比为15.95%；根据2015年全国1%人口抽样调查资料中数据统计显示，2010—2015年中国（未包括港澳台地区）省际迁移人口总规模为5 327.63万人，其中迁入江浙沪地区的人口数为1 428.46万人，占比为26.81%，可见1985—2015年，江浙沪地区迁入人口规模和占比都表现为显著的上升趋势。同时，“人”作为推进区域经济一体化中的中坚力量，以劳动力占多数的人口迁移对区域经济发展起重大影响，这亟待江浙沪地区各省级政府制定有效的人口迁移政策来进行合理指引，并且研究江浙沪地区人口迁入机制能够为推动江浙沪地区发展更高质量区域一体化提供支持和借鉴。

江浙沪地区社会经济发展水平相当，彼此相互联系，在探讨其人口迁入驱动机制时，若将其视为独立个体进行分析，则有失全面性。因此，本文通过构建分层模型来体现其人口迁入驱动机制之间的内在联系，并且考虑到模型参数估计所存在的不确定性问题，引入贝叶斯理论来整合先验知识（如历史迁移数据），并通过后验分布量化参数的不确定性，从而提高建模的稳健性和精确性，从而得出更为准确的结果，这对协调区域发展和制定人口迁移政策都具有重要意义。

1 研究数据与方法

1.1 研究数据

研究对象为中国31个省份（因数据缺失，未包括港澳台地区）迁入江浙沪地区两省一市的人口数（省内人口迁移除外），所采用的各省级单元之间的人口迁移数据取自中国统计局公布的“全国按现住地和五年前常住地分的人口”，其中2005—2010年和2010—2015年的数据分别来自2010年第六次全国人口普查和2015年全国1%人口抽样调查资料。人口迁移在人口、社会、经济等因素综合作用下进行，影响因素存在多元性，选取各省份常住人口规模来表征人口因素，人均GDP和就业人员平均工资来表征经济因素，人均医疗床位数和拥有大专以上学历的人口数来表征社会因素以及各省份省会城

收稿日期：2025-02-25

基金项目：国家自然科学基金（42371435）

作者简介：吴雯静（2000—），女，江苏泰州人，硕士研究生，研究方向为空间数据分析；叶绮霖（2000—），女，广东佛山人，硕士研究生，研究方向为空间数据挖掘。

市间铁路交通时间来表征距离因素。为避免反向因果关系，社会经济因素数据分别源于中国统计局发布的《中国统计年鉴 2006》和《中国统计年鉴 2011》，距离因素数据来源于系列《全国铁路旅客列车时刻表》，其中人口、经济和社会因素数据为31个省份分别与江浙沪两省之间的差值数据（省内除外），从而将迁入地和迁出地相联系。各项数据均除以自身标准差来进行标准化处理，模型解释变量具体描述如表1所示。

表1 模型解释变量描述

影响因素	解释变量	解释变量名称
人口因素	常住人口规模/万人	Pop
经济因素	人均GDP/（元·人$^{-1}$）	GDP
	就业人员平均工资/元	Wage
社会因素	人均医疗床位数/（张·万人$^{-1}$）	Bed
	拥有大专以上学历的人口数/人	Edu
距离因素	省会城市间铁路交通时间/s	D

1.2 研究方法

1.2.1 空间自相关

地理学第一定律指出事物之间都存在关联，且越接近的事物关联越密切[6]。空间自相关分析包括全局和局部空间自相关，在于确定要素在空间上是否存在相关性及其相关性的程度。本文采用全局莫兰指数和热点分析（Getis's-Ord G_i^*）来分别分析江浙沪地区迁入人口的全局和局部空间自相关性。

1950年澳大利亚学者 Moran[7] 提出的全局莫兰指数，可用于衡量全局空间自相关程度，公式表达为

$$I=\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{1}$$

式中：x_i、x_j 为区域 i、j 的要素属性值；n 为要素总数；w_{ij} 为空间权重矩阵中第 i 行 j 列的元素，是区域 i 和 j 之间空间距离的一种衡量方法，可用于定义空间要素之间的邻近关系。全局莫兰指数 I 取值为[−1,1]，当 $I>0$ 时，表示该要素的高-高值之间相互聚集或者低-低值之间相互聚集，即体现空间正自相关；当 $I<0$ 时，表示该要素的高-低值之间相互聚集，即体现空间负自相关；当 $I=0$ 时，表示该要素之间彼此相互独立，即空间分布随机。

然而全局莫兰指数的缺陷是无法把都表现为空间正自相关关系的高-高聚集区域（“热点”区域）以及低-低聚集区域（“冷点”区域）相区分，基于此 Getis 和 Ord[8] 于1992年提出 Getis-Ord G_i^*，通过热点分析来有效解析要素在局部空间上聚类分布的特点，可以很好地反映在局部空间区域内热点和冷点的分布，其标准化公式为

$$G_i^*=\frac{\sum_{j=1}^{n}w_{ij}x_i-\bar{x}\sum_{j=1}^{n}w_{ij}}{\sqrt{\frac{\sum_{j=1}^{n}x_j^2}{n}-\bar{x}^2}\sqrt{\frac{\left[n\sum_{j=1}^{n}w_{ij}^2-\left(\sum_{j=1}^{n}w_{ij}\right)^2\right]}{n-1}}} \tag{2}$$

当 $G_i^*(z)$ 为正值且越大时，表示高-高聚集越紧密，在局部空间内形成热点；当 $G_i^*(z)$ 为负值且越小时，表示低-低聚集越紧密，在局部空间内形成冷点。

1.2.2 分层贝叶斯方法

贝叶斯理论综合样本信息、待估参数的先验信息以及总体信息，来求解模型参数值，从而得到参数的理论估计结果。贝叶斯统计的核心在于通过先验分布（prior distribution）、似然函数（likelihood function）和后验分布（posterior distribution）的联合作用，将数据信息与先验知识结合，更新参数估计。计算公式可表示为

$$P(\theta\mid y)\propto P(y\mid\theta)\cdot P(\theta) \tag{3}$$

式中：$P(\theta)$ 为参数的先验分布；$P(y\mid\theta)$ 为似然函数；$P(\theta\mid y)$ 为后验分布。

分层贝叶斯模型是贝叶斯模型的一种，分层贝叶斯方法的主要特征体现在当同时获取总体信息和部分参数的先验信息后，通过引入超参数构建多层次结构，允许参数在不同组间共享信息，解决先验分布对参数估计值的过甚影响问题，进而增强模型参数估计的稳健性。本文基于分层贝叶斯方法构建江浙沪地区迁入人口的多元线性回归模型。

第一层模型：构建江浙沪地区的多元线性回归模型，计算公式为

$$y_{in}\mid\alpha_i,\beta_i=\alpha_i+\beta_iX_{in}+\varepsilon_i,\ i=1,2,3 \tag{4}$$

式中：y 为江浙沪地区迁入人口数；i 为上海、江苏和浙江；n 为样本量；X 为解释变量；α_i 为截距项；β_i 为斜率；ε_i 为残差；其中，α_i、β_i、ε_i 为待估参数。

第二层模型：对江浙沪地区迁入人口的模型参数 β_i 做进一步建模，以体现其解释变量之间的内在联系，计算公式为

$$\beta_i\mid\mu_\beta,\boldsymbol{\Sigma}_\beta\sim\mathrm{MVN}(\mu_\beta,\boldsymbol{\Sigma}_\beta),\ i=1,2,3 \tag{5}$$

式中：β_i 服从多元正态分布，对应的均值为 μ_β；协方差矩阵为 $\boldsymbol{\Sigma}_\beta$；$\mu_\beta$ 和 $\boldsymbol{\Sigma}_\beta$ 为待估参数。

对贝叶斯模型中的待估参数选择无信息先验分布，分别为

$$\begin{cases}\alpha_i \sim N(0,10^4),\\ \varepsilon_i \sim N(0,10^4),\\ \beta_4 \sim N(0,10^4),\\ \mu_\beta \sim \mathrm{MVN}(0,10^4\boldsymbol{I}),\\ \boldsymbol{\Sigma}_\beta \sim \text{Inv-Wishart}(v_o,\boldsymbol{\Lambda}_o)\end{cases} \tag{6}$$

式中:$\boldsymbol{I}$ 为单位矩阵;v_o 为自由度,将自由度设定为比自变量个数多 1,以使其方差为均匀先验分布;$\boldsymbol{\Lambda}_o$ 为尺度矩阵。

后验分布函数可通过先验分布和似然函数相乘得到,进而得到贝叶斯理论参数估计值。江浙沪地区迁入人口的分层贝叶斯多元线性回归模型的后验分布函数定义为

$$p(\alpha_i,\beta_i,\varepsilon_i \mid y) \propto \prod_{i=1}^{3}\prod_{n=1}^{30} L[y_i(n) \mid \alpha_i + \beta_i X(n),\varepsilon_i]p(\alpha_i)p(\varepsilon_i)p(\beta_i \mid \mu_\beta,\boldsymbol{\Sigma}_\beta)\times p(\mu_\beta)p(\boldsymbol{\Sigma}_\beta \mid v_o,\boldsymbol{\Lambda}_o) \tag{7}$$

在分层贝叶斯方法中,对于已有先验分布的超参数如果无法确定,则可以设置超参数的第二个先验,此先验也被叫作超先验,而分层先验是由先验和超先验之间组合所决定的新的先验。确定分层先验的方法分为两步[9]。

第一步:设置未知参数 β 的先验分布为一个形式上已知的密度函数,$\beta \sim \pi_1(\beta \mid \lambda)$,其中,$\lambda$ 为超参数,取值范围为 Λ。

第二步:为超参数 λ 再设置一个超先验 $\pi_2(\lambda)$。

两步设置后所得分层先验的公式为

$$\pi(\beta) = \int_\Lambda \pi_1(\beta \mid \lambda)\pi_2(\lambda)\mathrm{d}\lambda \tag{8}$$

分层贝叶斯模型提供了考虑参数不确定性的后验概率分布估计,假设回归系数的向量服从同一多元正态分布,其未知参数被估计为分析的一部分。在江浙沪地区迁入人口研究中,分层贝叶斯模型通过捕捉三省份的异质性,避免了简单合可能的偏差,同时利用全局信息降低单个省份参数估计的方差,提升了模型稳健性。

1.2.3 NUTS 算法

贝叶斯方法中计算后验分布往往较为复杂,贝叶斯模型主要是通过构建一个马尔可夫链来估计参数,而传统马尔可夫链蒙特卡洛(MCMC)算法(如 Metropolis-Hastings)在高维参数空间中易陷入局部最优或收敛缓慢,可基于 Stan 软件的 NUTS(No-U-Turn sampler)算法对模型参数进行估计。Stan 是一个基于贝叶斯理论建模的新型概率编程语言,拥有更广的应用平台,即便遇到复杂模型,计算速度依然很快,而且语言也相对更为灵活。Stan 使用的 NUTS 算法和 HMC(Hamiltonian Monte Carlo)算法,其原理模拟动力学中的跳点法(leap frog method)来更新参数位置,因此在复杂模型的条件下,基于 NUTS 算法和 HMC 算法的模型收敛所需的计算时间较常用的 Gibbs 算法和 Metropolis-Hastings 算法更短[10-11]。此外 Stan 允许使用不正确的先验。Gelman 等[12]将 Stan 软件用于其著作中阐述贝叶斯数据分析,Kruschke[13]也在其关于贝叶斯统计分析的著作中,详细讲解了 Stan 软件的教程。

Stan 内嵌的 NUTS 算法是 HMC 算法的拓展,也是近年来新兴的一种用于贝叶斯参数估计的 MCMC 算法。NUTS 算法在保留了 HMC 算法抑制随机游走行为能力的同时,消除了 HMC 算法对设定步长和跃迁步数参数的依赖,通过最小的人为干预来有效进行贝叶斯推理,并且它通过执行递归算法克服了时间可逆性的问题,在时间上向前或向后运行哈密顿模拟来保持可逆性,保证了能够收敛达到正确的分布[10]。NUTS 算法的优势在于无须手动调整参数,且对复杂模型的适应性更强,能够有效探索高维空间,确保参数估计的高效性。

在 NUTS 算法的核心结构中,对于每次取样,当既定约束条件都满足时,根据二项分布{-1,1}来抽取决定链走向的方向参数(-1 为向负方向构链,1 为向正方向构链),然后采用梯度二叉树方法来通过递归算法迭代生成一组候选参数,在一定条件下,根据已设定的概率来选择是否接受候选参数,重复上述步骤至不满足约束条件,停止更新参数,并保留此次取样中的一组"最优"参数,然后开始下一次取样[14]。图 1 为 NUTS 算法在一次取样中更新生成候选参数 θ 的过程。

2 研究结果

2.1 空间自相关分析结果

表 2 为 2005—2010 年江浙沪地区迁入人口的莫兰指数及显著性检验结果,可知全局莫兰指数均大于 0,表明其存在空间正自相关性,并且由于对应的 z-score均大于 2.58,对应的 P 也均小于 0.01,说明其

图 1 一次取样中参数更新的过程

通过 1%的显著性水平检验，因此江浙沪地区迁入人口在空间分布上呈现集聚效应且是显著的。

为了深入挖掘在局部空间上的集聚性特征，通过 Getis's-Ord G_i^* 进行热点分析（表 3），其中未标出的省份为不显著区域。结果显示江浙沪地区的热点区域，即高-高聚集的区域，主要集中于迁入地的邻近区域，冷点区域，即低-低值聚集的区域以及不显著区域则主要分布于西部地区和北部地区，说明江浙沪地区迁入人口与距离因素的联系仍然紧密。

表 2　2005—2010 年江浙沪地区迁入人口的全局莫兰指数及显著性检验

迁入省份	Moran's I	z-score	P
上海	0.346 6	3.885 4	0.000 1
江苏	0.191 3	2.714 7	0.006 6
浙江	0.291 9	2.826 2	0.004 7

表 3　2005—2010 年江浙沪地区迁入人口热点分析结果

迁入省份	热点区域	冷点区域
上海	陕西、河南、湖北、湖南、安徽、江西、江苏、浙江、福建	无
江苏	陕西、河南、湖北、湖南、江西、上海、浙江、福建	无
浙江	陕西、重庆、贵州、广西、河南、湖北、湖南、江西、福建、广东	内蒙古、吉林、辽宁

2.2　驱动机制

2.2.1　模型收敛性诊断

引入贝叶斯理论，对江浙沪地区迁入人口构建分层贝叶斯多元线性回归模型，以反映解释变量之间的内在联系。在贝叶斯统计中，通常采用 MCMC 方法来计算待估参数的后验分布，在此情况下马尔可夫链必须达到平稳收敛状态，进行 MCMC 算法必须解决的关键问题之一是何时达到收敛状态，因此可采用迹线图来进行收敛诊断，这是一种常用的图形收敛诊断方法，其显示每次迭代时每条马尔可夫链相对于迭代次数所得的抽样结果，可视化马尔可夫链在状态空间中的移动。当马尔可夫链的样本路径趋于平稳分布状态，只在一个较小固定区间范围内上下波动，认为链达到收敛状态，并且为了防止出现收敛于非最优结果的情况，通常选择多条马尔可夫链以观察其收敛区间是否一致[15]。

考虑 5 条相互独立的马尔可夫链，对每条链均选取不相同的初始值，且迭代 10 000 次，其中前 200 次迭代为"预热"，将每条链"预热"后所得的采样点轨迹进行可视化，如图 2～图 4 所示。图中可以看出 5 条链都收敛于某一固定区间内，没有明显的趋势，这说明参数都达到收敛状态，此时即为参

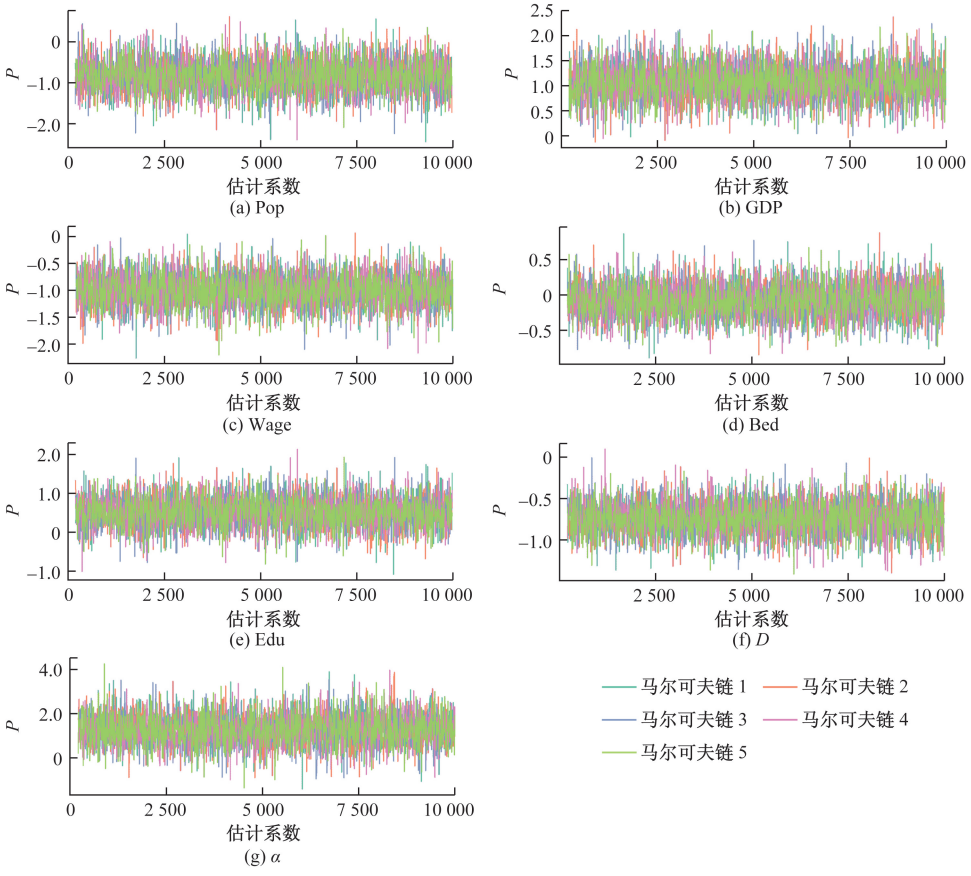

图 2　上海市迁入人口贝叶斯多元线性模型参数收敛迹

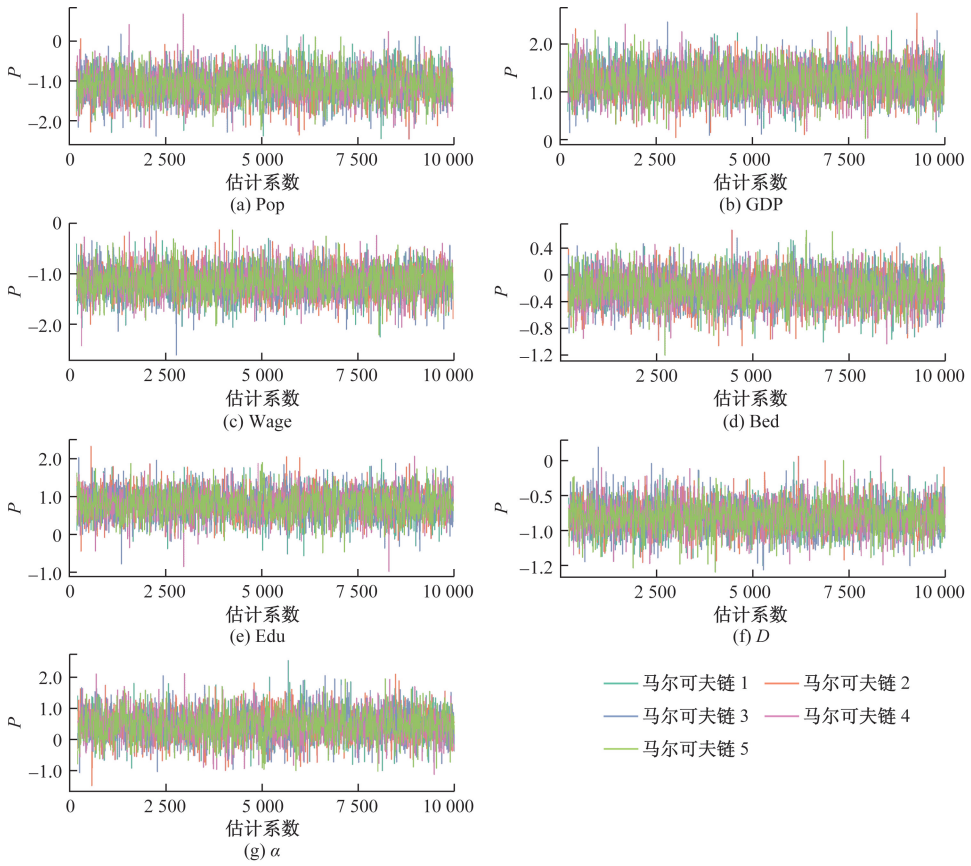

图 3　江苏省迁入人口贝叶斯多元线性模型参数收敛迹

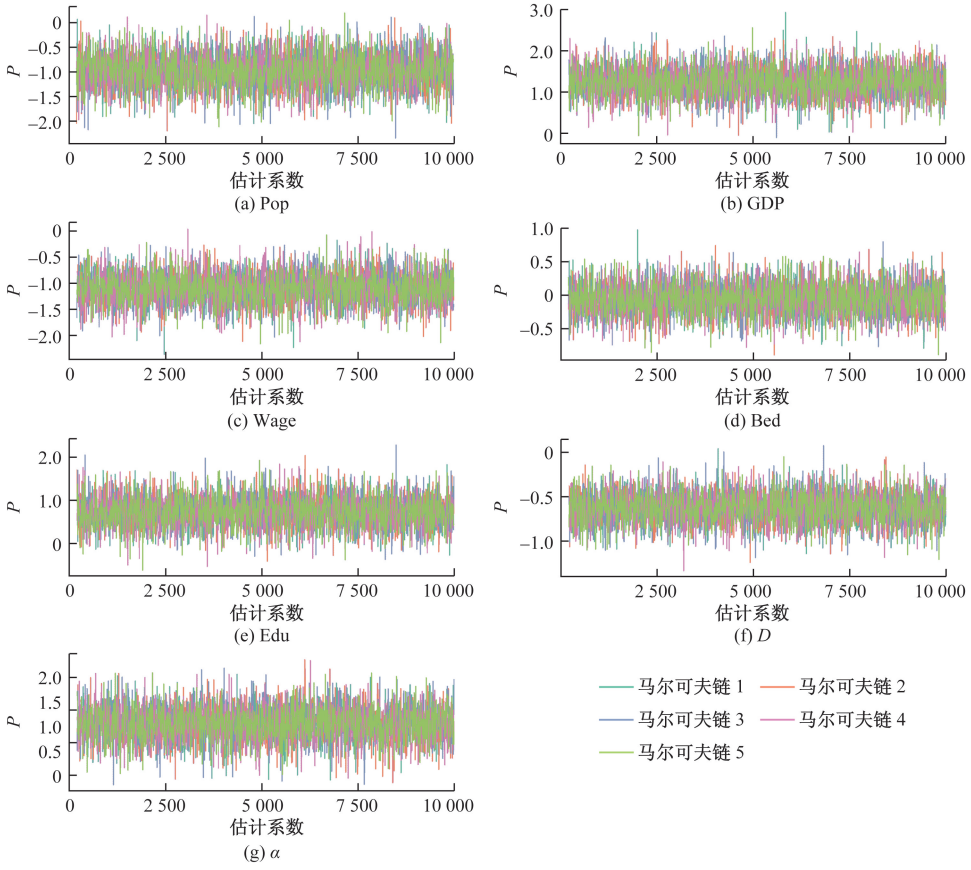

图 4　浙江省迁入人口贝叶斯多元线性模型参数收敛迹

数的平稳分布，所得参数估计值具有可信度。

2.2.2　贝叶斯模型拟合优度检验

基于 2010 年第六次全国人口普查中省际人口迁移数据，构建江浙沪地区迁入人口的贝叶斯多元线性回归模型。在模型检验方面，选用 KS 检验方法来判断贝叶斯模型残差是否服从正态分析，从而对模型的拟合优度进行检验。表 4 为迁入人口贝叶斯多元线性回归模型残差的 KS 检验，根据所得的 P 均大于 0.05 的结果可知，贝叶斯多元线性回归模型的残差均服从正态分布，即贝叶斯模型的拟合优度良好。

表 4　检验贝叶斯模型拟合优度的 KS 检验

KS 检验	上海	江苏	浙江
D 统计量	0.142 8	0.137 9	0.098 1
P	0.527 2	0.571 0	0.907 7

2.2.3　模型参数结果

在模型检验之后，基于所得参数估计值对模型结果进行分析。通过生成 5 条相互独立的马尔可夫链，每条链选取的初始值均不相同且分别迭代 10 000 次，其中前 200 次为“预热”样本，以 10 为间隔进行分组取样，最终每条链产生 1 000 个样本，即每个参数有 5×1 000 个样本估计值。表 5 为江浙沪地区两省一市迁入人口的贝叶斯多元线性回归模型的参数估计值，其中包括中位值和 95%置信区间。贝叶斯模型通过后验分布得到的各因素影响的概率范围，为政策制定提供更全面的风险参考。

图 5～图 7 为由生成的 5 条马尔可夫链所得到的所有参数的后验分布图。由表 5 和图 5～图 7 可以得出以下结论。

表 5　江浙沪地区迁入人口贝叶斯多元线性回归模型的参数估计值

变量	上海			江苏			浙江		
	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
Pop	−1.57	−0.84	−0.11	−1.84	−1.09	−0.36	−1.66	−0.96	−0.30
GDP	0.42	1.07	1.75	0.53	1.22	1.88	0.52	1.23	1.91
Wage	−1.56	−0.99	−0.43	−1.75	−1.16	−0.51	−1.65	−1.08	−0.50
Bed	−0.51	−0.07	0.38	−0.68	−0.20	0.29	−0.50	−0.06	0.40
Edu	−0.22	0.53	1.28	0.05	0.78	1.53	0.08	0.74	1.44
D	−1.13	−0.74	−0.35	−0.96	−0.63	−0.28	−0.96	−0.62	−0.28
α	−0.07	1.29	2.65	−0.49	0.48	1.45	0.35	1.05	1.74

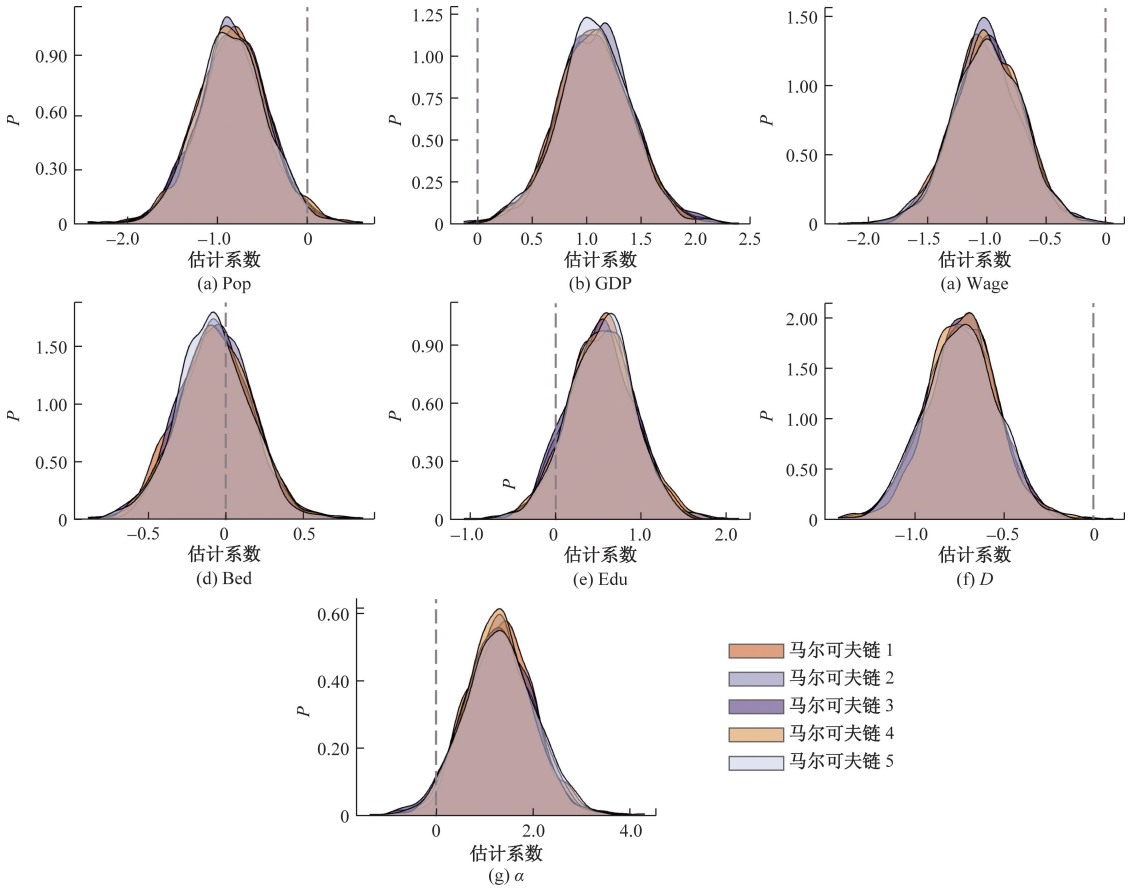

图 5　上海市迁入人口贝叶斯多元线性模型参数后验分布

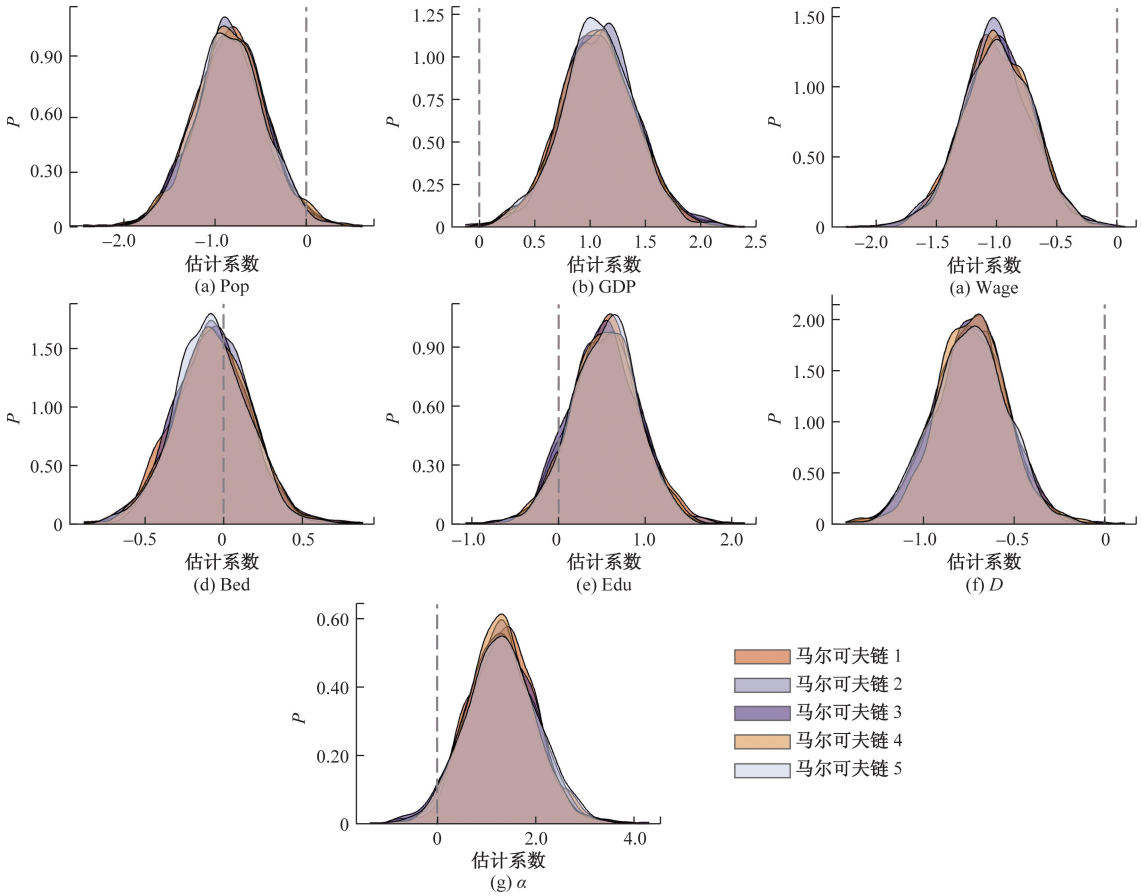

图 6　江苏省迁入人口贝叶斯多元线性模型参数后验分布

(1)上海、江苏和浙江迁入人口模型参数估计值显示三者人口迁入的驱动机制存在相似性,自变量按影响程度大小综合排序为人均 GDP、就业人员平均工资、常住人口规模、拥有大专以上学历的人口数、省会城市间距离以及人均医疗床位数。

(2)对于江浙沪地区而言,经济追求是人口迁入江浙沪地区的首要原因。迁入地与迁出地之间人均 GDP 的差值对人口迁入江浙沪地区表现出显著的促进作用,反映“经济梯度效应”。迁出地人均 GDP 较江浙沪地区越低,人口迁入意愿越强。这一现象与“推-拉理论”一致,即欠发达地区(推力)与发达地区(拉力)共同驱动迁移。江浙沪是中国经济发达程度最高的地区之一,就业机会更多,吸引人口迁入。

(3)人口因素对人口迁入江浙沪地区的影响仅次于经济因素,迁出省份的人口规模越大,迁移储量也越多,并且由于人口压力可能引起更多问题,如就业机会少、竞争激烈、基础设施不能满足人们需要等,从而会促进人口的外流。

(4)在社会因素方面,拥有大专以上学历的人口数少的地区更倾向于迁入大专以上人口数多的地区,大专以上学历人口数(Edu)的正向影响表明,江浙沪地区的高技能岗位需求与教育资源集聚形成“人才磁场”,如上海市高校密集带来的“教育移民”效应。现今社会对于教育的要求越来越高,并且随着义务教育的普及和社会竞争的压力,人们也往往会继续追求更高的教育。而人均医疗床位数的影响程度在所有解释变量中最低,这可能与现代医疗卫生事业的进步和医疗保险覆盖人群的扩大相关,一方面国家大力建设遍及城乡的医疗卫生服务体系,有力地改善着当地医疗卫生条件;另一方面,医改政策的实施逐步解决了异地就医问题,使得各地的医疗资源能够满足当地需求。

(5)迁出地与迁入地之间的距离与江浙沪迁入人口数之间为负相关关系,表明“迁移成本”仍对迁移流产生一定影响,距离迁入地越近,迁移人口越多,这一现象也可从江浙沪地区迁入人口空间分布图中看出。因此,尽管中国交通事业发展

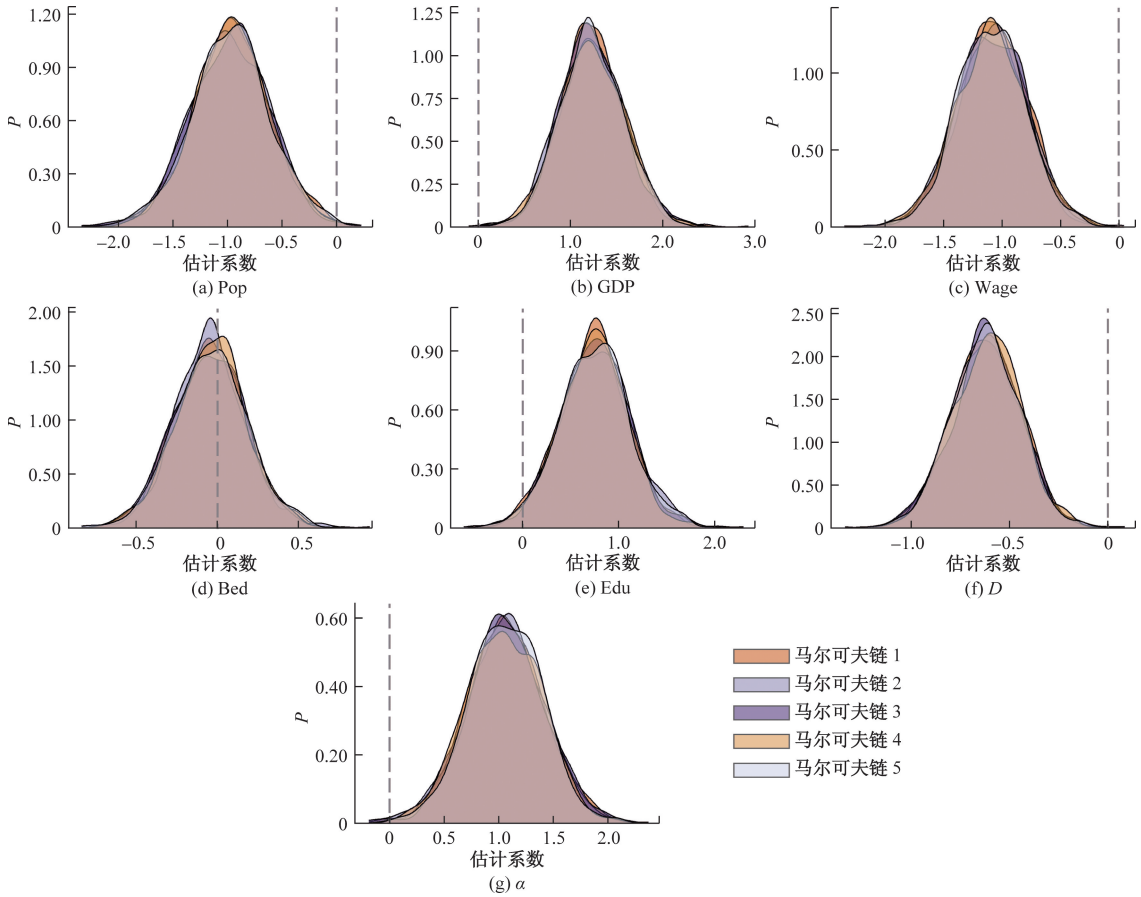

图 7　浙江省迁入人口贝叶斯多元线性模型参数后验分布

快速，但空间距离对于省际人口迁移的影响依然不可忽视。

2.2.4　贝叶斯模型预测检验

基于 2015 年全国 1%人口抽样调查资料中省际人口迁移的实际数据来对江浙沪地区迁入人口的贝叶斯多元线性回归模型的预测性进行检验。图 8 为 2010—2015 年江浙沪地区迁入人口真实值和模型预测值及其 95%置信区间，其中，横坐标表示不同迁出省份，分别为北京、天津、河北、山西、内蒙古、辽宁、吉林、黑龙江、上海、江苏、浙江、安徽、福建、江西、山东、河南、湖北、湖南、广东、广西、海南、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏和新疆（不包括迁入地省份），如 1 代表北京，30 代表新疆等。结果显示大部分省份人口迁入江浙沪地区的真实值落入模拟结果的 95%置信区间内，并且 30 个省份（除了自身省份）迁入江浙沪地区的模拟人口数值变化趋势与真实迁入人口变化趋势相当。

根据 2010—2015 年江浙沪地区迁入人口数的模型预测结果，江苏和浙江迁入人口预测结果不及上海，但通过计算均方根误差（root mean squared error，RMSE）发现，江苏省和浙江省迁入人口的分层贝叶斯多元线性回归模型的 RMSE 分别是 0.909 1 和 0.836 9，都小于其简单多元线性回归模型的 RMSE，分别为 0.934 0 和 0.850 1，因此表明基于贝叶斯理论建模能够提高模型的稳健性和精确性。

3　结论

首先对江浙沪地区迁入人口的空间自相关性进行分析，并基于分层贝叶斯模型来捕捉江浙沪三省份的异质性，从而研究 2005—2010 年江浙沪地区人口迁入的驱动机制，以及基于 2015 年全国 1%人口抽样调查资料中的人口迁移数据进行模型预测性检验，得到以下结论。

（1）江浙沪地区迁入人口呈现显著的空间集聚效应，且热点区域主要集中于江浙沪地区的邻近省份，表现出距离因素对人口迁入江浙沪地区的重要影响。

（2）模型结果显示，经济因素是人口迁入江浙沪地区的主要原因，其次是人口因素，迁出地人口

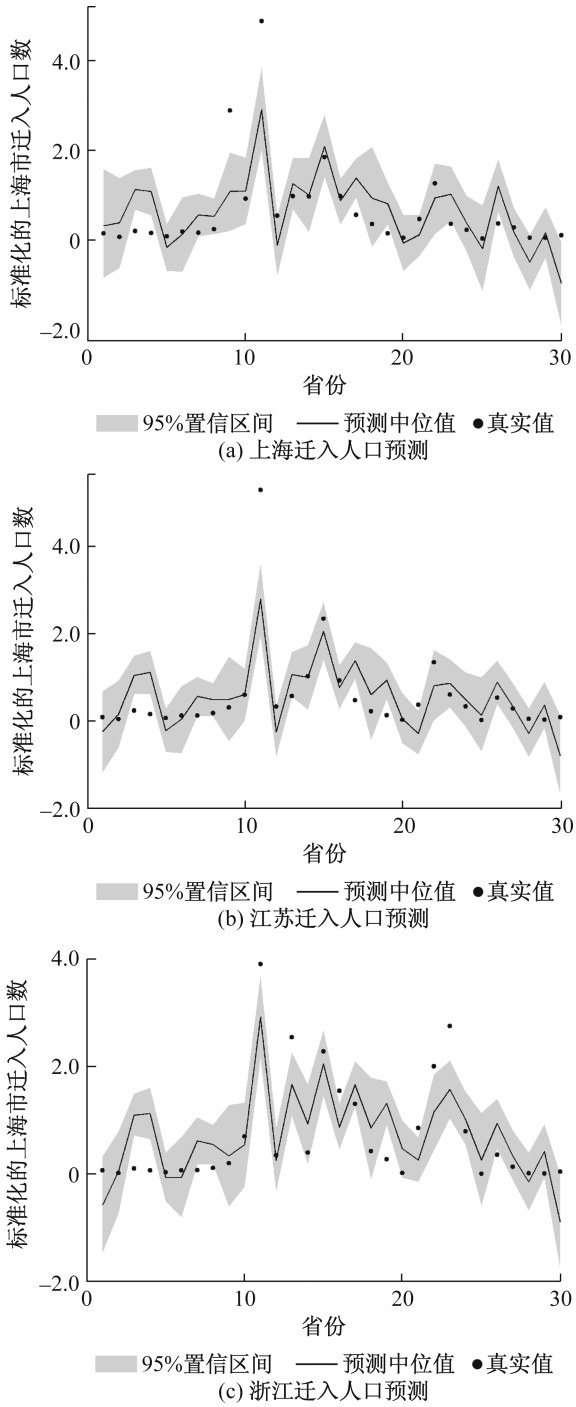

图 8　2010—2015 年江浙沪地区迁入人口真实值和预测中位值及其 95%置信区间

规模大促进人口向江浙沪地区的流入。江浙沪地区高的教育水平也在一定程度上吸引外省人口的迁入,而医疗资源对迁移影响较弱。因此,可以优先聚焦经济与教育领域,通过税收优惠吸引企业落户以提升就业机会,或建立长三角高校联盟,简化跨省学历认证流程,促进高学历人才流动,同时推进交通网络建设,降低迁移成本,促进邻近省份人口流动,从而强化人口迁入的“拉力”。

(3)在模型检验方面,KS 检验结果显示贝叶斯多元线性回归模型的拟合优度良好。此外基于 2015 年全国 1%人口抽样调查资料中人口迁移数据进行模型预测,结果显示贝叶斯模型预测结果良好,大部分省份迁入江浙沪地区的真实人口值都位于 95%置信区间内,且分层贝叶斯多元线性回归模型的 RMSE 小于简单多元线性回归模型的 RMSE,表明在建模中引入分层贝叶斯理论可以提高模型的稳健性和精确性。此外,若后续研究获得历史迁移数据或专家经验,可通过调整模型中的先验分布进一步提升模型精度。

这为分析地区迁入人口的驱动机制提供了一定的参考与借鉴,但是研究仍然存在一定的局限性。在对江浙沪地区人口迁入的驱动机制研究方面只关注社会经济方面因素,而江浙沪地区迁入人口存在空间正自相关,因此后续研究可在构建模型中加入空间要素,从而统筹优化江浙沪地区人口迁入的驱动机制分析。

参考文献

[1] 王桂新. 新中国人口迁移 70 年:机制、过程与展[J]. 中国人口科学, 2019(5): 2-14.

[2] 段成荣, 谢东虹, 吕利丹. 中国人口的迁移转变[J]. 人口研究, 2019, 43(2): 12-20.

[3] 柯文前, 朱宇, 陈晨, 等. 1995—2015 年中国人口迁移的时空变化特征[J]. 地理学报, 2022(2): 411-425.

[4] 李小萌, 李思涵, 史毅, 等. 中国省际人口迁移模式的复杂性探索[J]. 北京师范大学学报(自然科学版), 2023, 59(5): 785-795.

[5] 牛少凤, 于新东, 贾凡梅. 长三角区域经济一体化测度与经济发展效应分析[J]. 科技和产业, 2022, 22(5): 204-207.

[6] TOBLER W R. A computer movie simulating urban growth in the detroit region[C]//International Geographical Union, Commission on Quantitative Methods. Worcester, America: Clark University, 1970: 1-20.

[7] MORAN P. Notes on continuous stochastic phenomena [J]. Biometrika, 1950, 37(1): 17-23.

[8] GETIS A, ORD J K. The analysis of spatial association by the use of distance statistics[J]. Geographical Analysis, 1992, 24(3): 189-206.

[9] 杨世娟, 汪建均. 基于分层贝叶斯模型的稳健参数设计[J]. 系统工程与电子技术, 2019, 41(10): 2293-2303.

[10] HOFFMAN M D, GELMAN A. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo[J]. Journal of Machine Learning Research, 2014, 15(1): 1593-1623.

[11] 刘晋，汪秀琴，李天萍，等. 贝叶斯统计分析的新工具——Stan[J]. 中国卫生统计，2019，36(3)：462-465.
[12] GELMAN A，CARLIN J B，STERN H S，et al. Bayesian data analysis[M]. Boca Raton：CRC Press，2014.
[13] KRUSCHKE J. Doing Bayesian data analysis：a tutorial with R，JAGS and Stan[M]. Cambridge：Academic Press，2014.
[14] 何立杰，何洪林，任小丽，等. 基于贝叶斯机器学习的生态模型参数优化方法研究[J]. 地球信息科学学报，2017，19(10)：1270-1278.
[15] ROY V. Convergence diagnostics for Markov Chain Monte Carlo[J]. Annual Review of Statistics and Its Application，2020，7(1)：387-412.

A Bayesian Analysis of the Driving Mechanism of the Migrant Population in the Jiangsu-Zhejiang-Shanghai Region

WU Wenjing，YE Qilin
(School of Geography and Ocean Science，Nanjing University，Nanjing 210023，China)

Abstract：Since the reform and opening-up policy was implemented in 1978，China has loosed its household registration policy and rapidly advanced the construction of urbanization. The pattern of population migration has shifted from planned migration to autonomous migration. The Jiangsu-Zhejiang-Shanghai region has experienced a significant increase in the scale of the in-migration population，thence it is highly meaningful to research its driving mechanisms. Based on the in-migration data in the sixth population census，a multivariate linear regression model was constructed by the independent variables，such as the resident population，per capita GDP，per capita employment wage，per capita medical beds，population with a college degree or above and distance. Furthermore，the hierarchical Bayesian theory was introduced to reduce the indeterminacy of parameter estimation. The results show that economic factors serve as the primary driving force. Population factors，the education factor and distance factor also play important roles，while the medical factor has a relatively weak effect.

Keywords：Bayesian model；migrant population；Jiangsu-Zhejiang-Shanghai region；driving mechanism