Advances in Applied Mathematics 应用数学进展, 2022, 11(3), 1355-1366
Published Online March 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.113148

Hans汉斯

基于梁氏－克里曼信息流的江苏省人口流动预测

张克凡，吕广迎

南京信息工程大学，江苏 南京

收稿日期：2022年2月23日；录用日期：2022年3月21日；发布日期：2022年3月28日

摘　要

粮食产量作为一个牵动着基本民生福祉的问题，在影响人口流动方面也有着重要的作用。文章以江苏省为例，聚焦江苏省粮食生产比重最大的水稻，研究自1999年至2018年20年间，江苏省水稻产量和人口变化之间的联系。文章在研究过程中确定了将江苏省年均化肥使用量作为水稻产量的工具变量，将信息流的思想运用到因果分析上，利用梁氏－克里曼信息流计算出化肥使用量和江苏省人口流动之间的信息传递。结果表明，江苏省水稻产量和人口流动之间的因果关系是显著的，且江苏省水稻产量是人口流动的“因”。文章利用两种方法分别预测出江苏省未来5年和未来2年的非自然增长人口，均成上升趋势，表明政府需要采取相应的措施，应对大批人口涌入的情况。

关键词

梁氏－克里曼信息流，因果分析，人口流动，工具变量，粮食产量

Prediction of Population Flow in Jiangsu Province Based on Liang-Kleeman Information Flow

Kefan Zhang, Guangying Lv

Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Feb. 23rd, 2022; accepted: Mar. 21st, 2022; published: Mar. 28th, 2022

Abstract

As an issue that affects basic people’s livelihood and well-being, food production also plays an im-

文章引用: 张克凡, 吕广迎. 基于梁氏－克里曼信息流的江苏省人口流动预测[J]. 应用数学进展, 2022, 11(3): 1355-1366. DOI: 10.12677/aam.2022.113148

portant role in affecting population mobility. Taking Jiangsu Province as an example, this article focuses on rice, which accounts for the largest proportion of grain production in Jiangsu Province, and studies the relationship between rice production and population changes in Jiangsu Province during the 20 years from 1999 to 2018. In the process of research, the article determined that the average annual fertilizer use in Jiangsu Province was used as an instrumental variable for rice production, applied the idea of information flow to causal analysis, and used Liang-Kleeman information flow to calculate the information transfer between fertilizer use and population movements in Jiangsu Province. The results show that the causal relationship between rice production and population mobility in Jiangsu Province is significant, and that rice production in Jiangsu Province is the "cause" of population mobility. The article uses two methods to predict the unnatural population growth in Jiangsu Province in the next 5 years and the next 2 years, both of which are on the rise, indicating that the government needs to take corresponding measures to deal with the influx of large numbers of people.

Keywords

Liang-Kleeman Information Flow, Causal Analysis, Population Mobility, Instrumental Variable, Food Production

1. 引言

我国人口众多，并具有流动性大的特点，因此很容易影响社会的稳定性。近年来，人口流动的规模在不断增加，人口的流动不仅给地区间带来了积极的影响，例如更多的劳动力，更多的高技术人才以及更多的市场需求外，同时也会对社会福利、环境污染、治安管理等带来一定的负面影响。因此，通过统计学方法对地区间人口流动进行预测，不仅可以使人口流入地调整政策更合适地接纳大规模的新居民，也可以使人口流出地在一定程度上弥补由于人口流失带来的损失。

2. 方法

目前国内外对于人口流动原因的研究主要集中在环境，经济以及重大事件上，藏媛，郝枫构建异质劳动力跨期效用模型，匹配 CMDS-2017 与 286 个地级市统计数据，采用 IV-logit 模型考察空气质量变化对流动人口城市留居意愿强度的影响。研究发现空气质量改善会增强流动人口留居意愿强度，但提升能力边际递减[1]。宫湛秋，孙诚等利用基于信息流理论的因果分析方法，研究了 1880 年以来观测的 AMO 与北大西洋海表热通量间的因果关系。结果表明，在多年代际尺度上，从 AMO 到海表热通量的信息流要远大于二者相反方向的信息流，说明 AMO 是北大西洋海表热通量异常的因，海洋主导了海气间的热量交换[2]。

本文将以江苏省为例，聚焦江苏省粮食生产比重最大的水稻，研究自 1999 年至 2018 年 20 年间，江苏省水稻产量和人口变化之间的联系。由于在模型中，水稻产量作为本研究的主要研究对象，其很有可能是内生的，其不仅受到其他扰动项的影响，例如历史因素，产业结构等，在影响被解释变量的同时，也受到被解释变量的影响(人口的增加带来了更多的劳动力，更多粮食的需求，促进了水稻产量的提升)。因此，在实际研究中必须要构建合适的工具变量来进行分析。本文选取江苏省化肥使用量作为水稻产量

的工具变量，分析化肥使用量和人口流动之间的因果关系。在目前大量的研究中，对于因果性的分析主要都是利用传统的格兰杰因果关系。但是格兰杰因果检验只是判断两个事件发生的先后顺序在统计学领域是否是显著的，不算真正的因果关系。本文运用的梁氏－克里曼信息流将信息传递的概念引入因果分析，将两个事件之间的信息传递可以定量的表达出来，既简洁，准确率也高。根据两者之间信息传递的绝对值便可以分辨出哪个是“因”，哪个是“果”。梁湘三提出，二维系统下两个事件之间的信息传递 $\hat{T}_{2\to1}$ 为：

$$\hat{T}_{2\to1}=\frac{C_{11}C_{12}C_{2,d1}-C_{12}^2C_{1,d1}}{C_{11}^2C_{22}-C_{11}C_{12}^2}$$

其中 C_{ij} 为样本协方差，$C_{i,dj}$ 是 x_i 和由 x_j 导出的级数之间的样本协方差[3]。

3. 数据和变量选取

为研究粮食产量和人口流动之间的关系，本文以江苏省为例，粮食产量选取江苏省粮食生产比重最大的水稻作为研究对象，研究自 1999 年至 2021 年 23 年间，江苏省年水稻产量和人口变化之间的联系。为了使模型的拟合程度更加高，本文还选取江苏省年均气温，年均降水量，年水稻种植面积以及年限额以上餐饮行业营业总收入作为其他解释变量加入模型。以上数据均来自中国统计年鉴(http://www.stats.gov.cn/tjsj/ndsj/)，其中人口数据为年末总人口减去自然增长的人口，使数据可以更好地体现人口流动的变化。

本文运用 stata 软件进行分析，在线性模型的基础上，模型的 R^2 值为 0.96，证明该模型拟合程度较好。在实际分析中，水稻产量作为本研究的主要研究对象，其很有可能是内生的。若模型存在内生解释变量，则需要利用外生的，和解释变量强相关的工具变量来代替解释变量，若工具变量和被解释变量存在因果关系，则根据相关性，就可以推导出解释变量和被解释变量间存在因果关系。本文选择江苏省年均化肥使用量作为水稻产量的工具变量，一方面是化肥使用量和水稻产量有着较强的相关性：化肥使用增加，水稻产量也会相应增加。另一方面，化肥使用和人口流动并没有直接的关系，因此可认为该工具变量是外生的，可以运用到本次研究中。文章所用到的变量标识和单位如表 1 所示。

Table 1. Variable Identification and Unit Name
表 1. 各变量标识和单位名称

标识	名称	单位
people	江苏省非自然增长年人口数	万人
rice	江苏省年水稻产量	万吨
income	江苏省年限额以上餐饮行业营业总收入	亿元
area	江苏省年水稻种植面积	千公顷
temp	江苏省年均气温	摄氏度
rain	江苏省年均降水量	毫米
fert	江苏省年均化肥使用量	万吨

4. 实例分析

4.1. 内生性检验

根据方差齐性的不同，内生性检验分为两种：在同方差情况下，运用 Hausman 检验；在异方差情况下，运用 Durbin-Wu-Hausman 检验(DWH 检验)。根据 WHITE 检验的结果，检验 p 值为 0.3474，大于 0.1，

因此可以认为模型不存在异方差，选取 Hausman 检验。Hausman 检验结果如图 1，根据图中结果可知，Hausman 检验的 p 值为 0.0327，小于阈值 0.05，因此可认为水稻产量作为解释变量是内生的，因此需要引入工具变量。

	Coefficients			
	(b) iv	(B) ols	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
rice	-1.300889	-.5618774	-.7390117	.3461028
income	2.164734	2.099222	.0655121	.0306814
area	.7427718	.0838258	.6589459	.3086055
temp	-58.6281	-23.29061	-35.33749	16.54968
rain	.2092569	.1625438	.0467131	.0218772
_cons	8418.887	8016.172	402.715	188.6043

b = consistent under Ho and Ha; obtained from ivregress
B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 4.56
Prob>chi2 = 0.0327
(V_b-V_B is not positive definite)

Figure 1. Test results of Hausman
图 1. Hausman 检验结果

4.2. 不可识别检验

不可识别检验是判断工具变量的个数是否小于内生解释变量的个数，其原假设 H_0 为：工具变量识别不足。在样本并非独立同分布的情况下，该检验需要利用 Kleibergen-Paap rk LM 统计量进行判断。根据图 2 显示结果，Kleibergen-Paap rk LM 统计量的 p 值为 0.0047，小于 0.05，拒绝原假设，证明工具变量的个数至少大于等于内生解释变量个数。

4.3. 弱工具变量检验

该项检验主要是研究选取的工具变量与内生解释变量的相关性强度。在弱相关的情况下，使用工具变量进行估计的结果会相较于 OLS 的结果相差较大。在非独立同分布的情况下，该检验需要利用 Kleibergen-Paap rk WaldF 统计量进行判断。根据图 2 显示结果，Kleibergen-Paap rk WaldF 统计量值为 15.383，大于显著性 15%的临界值，因此可以推断该模型不存在弱工具变量，选用化肥使用量作为工具变量和水稻产量之间的相关性较强。

4.4. 过度识别检验

该检验主要是研究所有的工具变量是否全是外生的。在非独立同分布情况下，该检验利用 HansenJ 统计量进行判断。由图 2 结果“(equation exactly identified)”可以看出，此时工具变量个数和内生解释变量个数恰好相等，模型恰足确认，因此可以认为化肥使用量作为工具变量是外生的。

4.5. 因果分析

在本阶段的研究中，文章使用梁氏－克里曼信息流来证明化肥使用量和人口变化之间的因果关系。梁氏－克里曼信息流运用的前提条件为：两条研究的时间序列对象必须是平稳的，且时间间隔相同。根

据单位根检验可知，肥使用量和人口变化序列并非是平稳的，带有时间趋势，分别对两序列做一阶差分处理，由图 3 和图 4 单位根检验 p 值可以看出，差分运算后的序列平稳，因此可以进行梁氏－克里曼信息流的求解。

```
Underidentification test (Kleibergen-Paap rk LM statistic):              8.004
                                                    Chi-sq(1) P-val =    0.0047
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):               21.223
                         (Kleibergen-Paap rk Wald F statistic):         15.383
Stock-Yogo weak ID test critical values: 10% maximal IV size            16.38
                                         15% maximal IV size             8.96
                                         20% maximal IV size             6.66
                                         25% maximal IV size             5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.000
                                                    (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         rice
Included instruments: income rain temp area
Excluded instruments: fert
```

Figure 2. Test results of instrumental variables
图 2. 工具变量检验结果

```
Dickey-Fuller test for unit root                   Number of obs   =        37

                                 ---------- Interpolated Dickey-Fuller ---------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)            -5.345           -3.668            -2.966           -2.616
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

Figure 3. Unit root test of population change
图 3. 人口变化的单位根检验

```
Dickey-Fuller test for unit root                   Number of obs   =        37

                                 ---------- Interpolated Dickey-Fuller ---------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
------------------------------------------------------------------------------
 Z(t)            -4.560           -4.270            -3.552           -3.211
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0012
```

Figure 4. Unit root test of fertilizer usage
图 4. 化肥使用量的单位根检验

对两条时间序列进行梁氏－克里曼信息流的求解。根据二维线性模型系统下，x_1 到 x_2 的信息流率的最大似然估计 $\hat{T}_{2\to1}$ 的公式：

$$\hat{T}_{2\to1}=\frac{C_{11}C_{12}C_{2,d1}-C_{12}^2C_{1,d1}}{C_{11}^2C_{22}-C_{11}C_{12}^2}$$

可以得出化肥使用量对人口流动的信息传递为 0.04934。为探究该因果关系是否是显著的，因此还需

对结果进行显著性检验[4]。文章基于三种不同的显著性水平 90%，95%和 99%，分别计算三种不同的显著性水平下的置信区间。若 0 不包含在置信区间中时，即可认为该因果关系是显著的。三种不同的置信区间如表 2 所示。从表中信息可以看出在 90%和 95%的显著性水平下，化肥使用量和人口流动之间存在着显著的因果关系。相反的，从表 3 可以看出，人口流动对化肥使用量信息传递的置信区间均包含 0，因此可以判定该因果关系不显著。对比两者可以推导出，化肥使用量是人口流动的“因”，根据工具变量的强相关，进一步可以推导出江苏省水稻产量和江苏省人口流动有着因果关系。在实际运用中，决策部门可以聚焦于当地水稻产量的变化，若水稻产量即将出现大幅上涨或下跌的趋势，那么就应该做好相应的准备，面对可能出现的大规模的人口流动。

Table 2. Confidence interval of fertilizer usage on the population mobility flow information
表 2. 化肥使用量对人口流动信息传递的置信区间

显著性水平	置信区间
90%	[0.01284, 0.08584]
95%	[0.00598, 0.0927]
99%	[−0.0073, 0.10598]

Table 3. Confidence intervals of population mobility for information transmission of fertilizer use
表 3. 人口流动对化肥使用量信息传递的置信区间

显著性水平	置信区间
90%	[−0.04966, 0.01494]
95%	[−0.05573, 0.02101]
99%	[−0.06747, 0.03275]

5. 人口预测

5.1. 基于 ARIMA 模型的人口流动预测

在明确了江苏省水稻产量对江苏省的人口流动有着显著的因果关系之后，本文的下一个目标便是根据水稻产量的变化来预测未来江苏省人口的流动。在本环节，由于模型存在内生变量，所以本文选用两阶段最小二乘对模型进行回归。在得到回归模型之后，本文选取 1990 年至 2013 年江苏省年均气温、江苏省年水稻种植面积、江苏省年均降水量和江苏省年水稻产量，利用 ARIMA 模型对该 4 个指标进行拟合，并计算出 2014 年至 2018 年各指标的预测值，根据回归模型得到未来 5 年江苏省人口流动的预测值，并与实际数据进行比对，若预测值和实际值较为相似，即可认为该方法准确度较高。

根据 2sls 回归得到的结果，模型拟合程度的 p 值为 0.00，小于显著性水平 0.05，因此可以认为该模型显著性较高，拟合程度较好。模型表现形式如下：

$$people = -0.9 * rice + 0.18 * rain - 0.35 * area + 2.22 * income - 31.66 * temp + 8133.31$$

得到关于江苏省流动人口的模型之后，文章将利用 ARIMA 模型对 1990 年至 2013 年江苏省年均气温、江苏省年水稻种植面积、江苏省年均降水量和江苏省年水稻产量进行拟合，并得到 2014 年至 2018 年各指标的预测值。由于针对不同变量建模过程较为相似，文章以江苏省年水稻产量为例，进行 ARIMA 模型拟合。

首先，文章对江苏省年水稻产量进行 ADF 单位根检验，若时间序列是非平稳的，则需要进行差分处

理，使其变为平稳序列，再进行模型拟合。根据计算可得 ADF 统计量为−3.426482，p 值为 0.010089，小于显著性水平 0.05，拒绝原假设，因此该序列是平稳的。经 LB 检验，LB 检验统计量显著性水平为 0.00，小于显著性水平 0.05，因此该时间序列并非白噪声序列。做出该平稳序列的自相关和偏自相关图，结果如图 5 所示。由图中看出，江苏省年水稻产量的自相关图 1 阶拖尾，偏自相关图 1 阶拖尾，根据 ARIMA 模型定阶准则，满足要求的模型有 ARIMA (1, 0, 1)，ARIMA (4, 0, 1)，ARIMA (5, 0, 1)，ARIMA (8, 0, 1) 和 ARIMA (9, 0, 1)。分别计算出该三个模型的 AIC 值，结果如表 4。根据 AIC 准则，选取 AIC 值最小的模型，因此对于江苏省年水稻产量选取 ARIMA (4, 0, 1)模型。根据模型预测出未来 5 年江苏省水稻年产量，2014 至 2018 年水稻产量为分别为 1790.38 万吨、1825.06 万吨、1836.22 万吨、1828.57 万吨、1809.69 万吨。

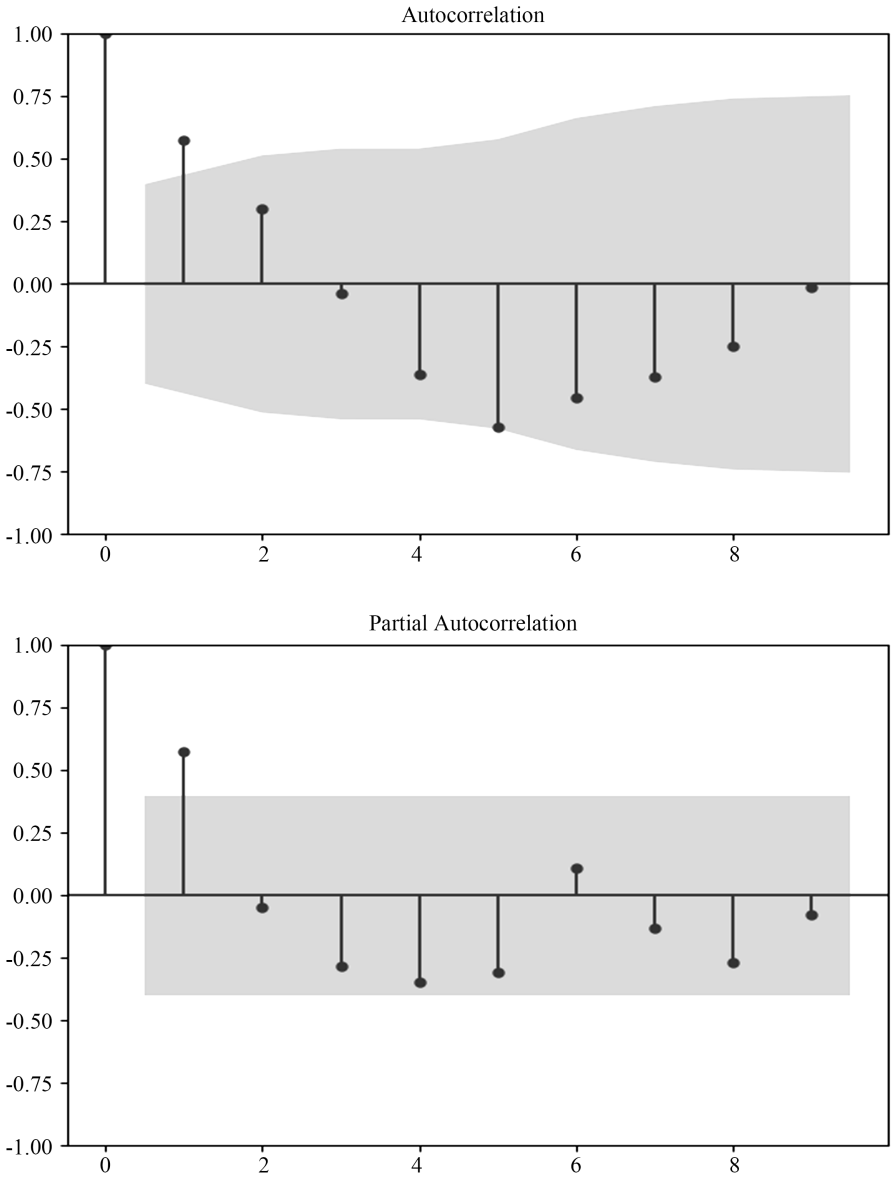

Figure 5. Autocorrelation and partial autocorrelation of annual rice output in Jiangsu Province
图 5. 江苏省年水稻产量自相关和偏自相关图

Table 4. AIC value of each model of annual rice production in Jiangsu Province
表 4. 江苏省年水稻产量各模型 AIC 值

拟合模型	AIC
ARIMA (4, 0, 1)	296.002
ARIMA (5, 0, 1)	300.598
ARIMA (8, 0, 1)	296.823
ARIMA (9, 0, 1)	297.835
ARIMA (1, 0, 1)	300.473

分别对江苏省年均气温、江苏省年水稻种植面积、江苏省年均降水量和江苏省限额以上餐饮行业年收入进行 ARIMA 模型拟合并预测，得到结果如表 5。根据该四个变量的预测值计算出 2014 年至 2018 年江苏省非自然增长人口数的预测值，并与真实值进行比较，结果如图 6。通过对预测值和真实值的对比，发现预测值和真实曲线拟合程度较高，因此该种方法的准确率较好。据此预测出 2022 至 2026 年的江苏省非自然增长人口，如表 6 及图 7 所示。

Table 5. Predicted values of various variables from 2014 to 2018
表 5. 各变量 2014 年至 2018 年预测值

变量名称	模型	预测值				
		2014	2015	2016	2017	2018
水稻产量	ARIMA (4, 0, 1)	1790.38	1825.06	1836.22	1828.57	1809.69
年均气温	ARIMA (8, 1, 1)	16.87	15.25	15.6	16.34	16.3
年均降水量	ARIMA (1, 0, 0)	1029.11	1071.79	1063.94	1065.39	1065.21
种植面积	ARIMA (1, 3, 1)	2104.98	2086.24	2106.11	2149.86	2198.22
限额以上餐饮收入	ARIMA (5, 2, 1)	363.6	370.9	381.41	404.56	457.41

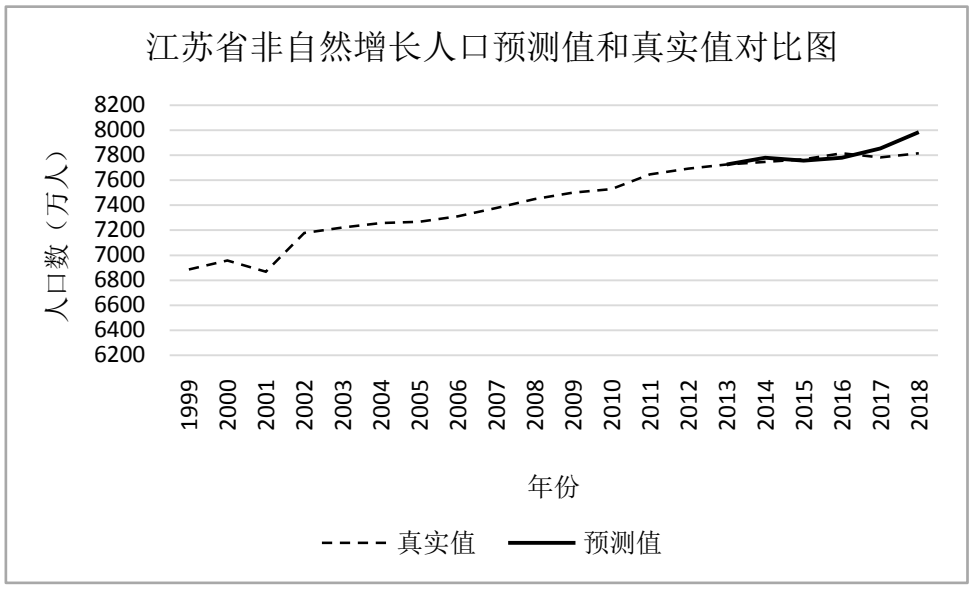

Figure 6. Comparison of predicted and true values of unnatural population growth in Jiangsu Province
图 6. 江苏省非自然增长人口预测值和真实值对比图

Table 6. The predicted value of unnatural population growth in Jiangsu Province in the next five years
表 6. 未来 5 年江苏省非自然增长人口预测值

年份	江苏省非自然增长人口预测值
2022	8431.6778
2023	8457.2622
2024	8619.2578
2025	8607.6518
2026	8775.885

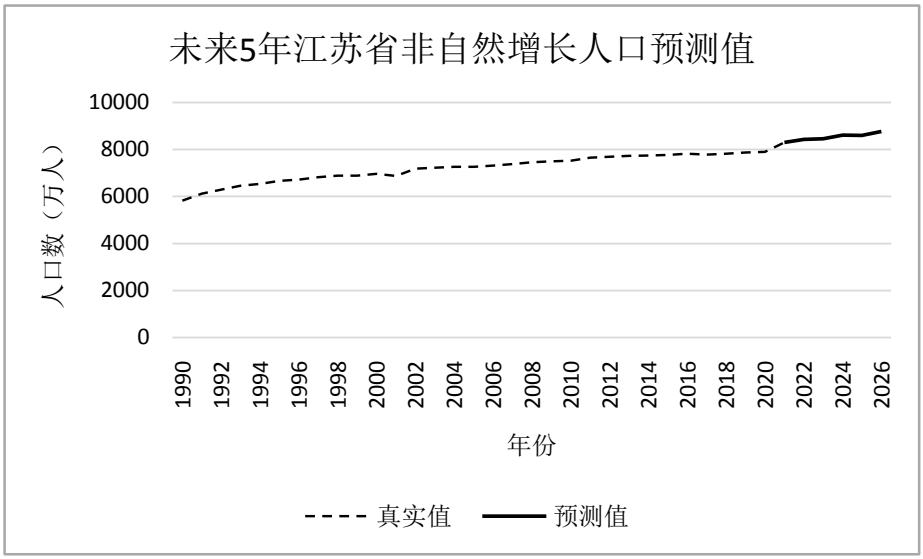

Figure 7. The predicted value of unnatural population growth in Jiangsu province in the next five years
图 7. 未来 5 年江苏省非自然增长人口预测值

5.2. 基于滞后因果关系的人口流动预测

除了利用各变量每一年的数据进行 2sls 拟合，并建立各自变量的 ARIMA 模型，计算出各变量的预测值，从而通过线性模型得到江苏省非自然增长人口的预测值这一种方法外，本文还提出了另外一种方法。本文将所有因变量江苏省非自然增长人口滞后若干期，探究在人口数滞后的条件下，水稻产量是否会和人口数之间存在着显著的因果关系。若存在，便可以通过水稻产量的变化，预测出江苏省人口流动的变化。此外，对所有自变量与滞后的因变量这一新的数据集进行两阶段最小二乘回归，得到关于滞后人口数量的线性模型，便可以根据不同年份的数据，计算出未来各年，江苏省非自然增长人口的预测值。经过滞后不同的期数，重新进行工具变量的选取和梁氏－克里曼信息流的计算，本文得出结果：在人口数滞后 2 年，选取江苏省年化肥使用量作为江苏省年水稻产量的工具变量的条件下，化肥施用量和非自然增长人口数之间存在着显著的因果关系。工具变量检验及化肥施用量对滞后 2 年江苏非自然增长人口的梁氏－克里曼信息流如图 8 和表 7 所示。

不可识别检验中，Kleibergen-Paap rk LM 统计量的 p 值为 0.0055，小于 0.05，拒绝原假设，证明工具变量的个数至少大于等于内生解释变量个数。在弱工具变量检验中，Kleibergen-Paap rk Wald F 统计量为 11.374，大于显著性水平 15%的临界值，因此该模型中不存在工具变量。由于该模型中工具变量与内

生变量的数量相同，所以必定会通过过度拟合检验。因此，在该模型中，使用江苏省年化肥使用量作为江苏省年水稻产量的工具变量是合适的。计算江苏省年化肥使用量和江苏省年非自然增长人口之间的梁氏-克里曼信息流置信区间，结果如表 7 所示。从表 7 数据可以看出，在显著性 90%和 95%的水平下，化肥使用量对人口流动信息传递因果关系的置信区间均不包含 0，因此可以认为该因果关系是在 90%和 95%的显著性水平下是显著的。根据工具变量和内生变量的强相关性，可以认为江苏省年水稻产量和滞后 2 年的江苏省非自然增长人口之间具有显著的因果关系，且水稻产量是人口流动的“因”。

根据两阶段最小二乘法建立各自变量和滞后 2 年的江苏省非自然增长人口之间的线性模型，得到模型如下：

$$people = -0.77 * rice + 52.92 * temp + 0.46 * area + 0.03 * rian + 2.06 * income + 6577.25$$

该线性模型的 R^2 为 0.9241，p 值为 0.00，小于显著性水平 0.05，说明该模型拟合程度较好。根据 2020 和 2021 年江苏省年水稻产量，江苏省年水稻种植面积，江苏省年均降水，江苏省年均气温和江苏省限额以上餐饮企业年收入的数据，可以计算出 2022 和 2023 年江苏省非自然增长的人口数量，分别为 8333.812 万人和 8574.089 万人，相较于之前的数据，江苏省非自然流动人口也呈上升趋势。

```
Underidentification test (Kleibergen-Paap rk LM statistic):              7.712
                                                      Chi-sq(1) P-val =    0.0055
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):               13.855
                         (Kleibergen-Paap rk Wald F statistic):         11.374
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.000
                                                    (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         rice
Included instruments: temp area rain income
Excluded instruments: fert
```

Figure 8. Instrumental variable test with a two-year lag
图 8. 滞后两年的工具变量检验

Table 7. Confidence interval of fertilizer use on the information transmission of population flow with a lag of two years
表 7. 化肥使用量对滞后两年人口流动信息传递的置信区间

显著性水平	置信区间
90%	[0.02407, 0.17859]
95%	[0.00956, 0.1931]
99%	[−0.01853, 0.22119]

6. 结语

本文基于中国统计年鉴上江苏省非自然增长人口数据，利用江苏省化肥使用量作为江苏省水稻产量的工具变量，对数据进行因果分析。经由不可识别检验、过度识别检验和弱工具变量检验，选取化肥使用量作为水稻产量的工具变量是合理的。经过计算化肥使用量和江苏省非自然增长人口数之间的梁氏－

克里曼信息流，可得，江苏省化肥使用量和江苏省非自然人口变化之间存在显著的因果关系，且化肥施用量是人口流动的“因”。根据工具变量的强相关，进一步可以推导出江苏省水稻产量和江苏省人口流动有着因果关系。

在对江苏省非自然增长人口的预测阶段，文章采用了两种不同的方法进行实践。第一种方法是根据历年江苏省非自然增长人口数和江苏省年均气温、江苏省年水稻种植面积、江苏省年均降水量和江苏省年水稻产量数据，通过 2sls 方法建立起工具变量回归模型。得到人口数和各自变量的回归模型后，通过时间序列分析法，对江苏省水稻产量等自变量数据进行模型拟合，分别对各自变量建立 ARIMA 模型，并通过模型对未来 5 年数据进行预测。得到各自变量的预测值后，通过先前构建的工具变量回归模型，计算出未来 5 年江苏省人口数。经过将预测值和真实值进行对比后，该种方法准确率较高，可以用于预测江苏省未来的流动人口。文章选取 1999 年至 2021 年江苏省各变量数据，得到 2022 年至 2026 年江苏省非自然增长人口，并且人口数呈持续上升的趋势。

第二种预测方法则是将江苏省人口数据滞后若干年，分别计算江苏省化肥使用量和流动人口之间的梁氏－克里曼信息流。如果滞后某一年数的人口数量与化肥使用量存在显著的因果关系的话，便可以根据当前数据，得到未来某一年的江苏省流动人口数。对滞后的数据集进行工具变量检验可得，使用江苏省年化肥使用量作为江苏省流动人口的工具变量依然合适，因此，经过反复试验，文章得出结果：在江苏省非自然增长人口数滞后两年时，与江苏省年化肥使用量之间存在显著的因果关系，所以可以使用该结论对未来江苏省流动人口进行预测。文章对江苏省非自然增长人口和各自变量建立 2sls 模型，模型 R^2 值为 0.9241，证明该模型准确度较高。利用 2020 年和 2021 年各自变量的数据，根据 2sls 模型可计算出 2022 和 2023 年的非自然增长人口数，分别为 8333.812 万人和 8574.089 万人，相较于 2020 年的 7902.144 万人和 2021 年的 8300.6784 万人，也是呈持续上升的趋势。

对比两种预测方法，文章更倾向于使用通过研究滞后的流动人口数与化肥使用量之间的因果关系，通过建立 2sls 模型来计算出流动人口预测值这一方法。在第一种方法当中，需要对每个自变量建立 ARIMA 模型，计算出各自变量预测值后利用 2sls 模型进行预测，建立的 ARIMA 模型数量较高，误差较大，虽然与真实值对比结果较好，但是稳定性不如第二种方法。此外，文章还利用多元 LSTM 神经网络和随机森林对江苏省未来的流动人口进行预测，但是由于本研究内容的数据集规模较小，使用神经网络和随机森林的误差较大，特别是验证误差保持持续上升的趋势，因此，使用神经网络和随机森林的准确率较低，利用滞后因变量这一方法更适用于本研究内容。

人口的流动不仅给地区带来了新鲜的劳动力，降低了劳动力成本，也带来了更多的物品需求，带动了地区经济发展。高教育人才的流动，会对流入地的技术科技能力带来积极的影响，增强该地区的竞争能力，同时也会对流出地造成不可忽视的损失，在高新技术方面也会有一定的落后。因此，在实际运用中，决策部门可以聚焦于当地水稻产量的变化，若水稻产量即将出现大幅上涨或下跌的趋势，那么就应该做好相应的准备，面对可能出现的大规模的人口流动。此外，本文为应对人口流动的研究提供了一个全新的思路。以往为应对人口流动而提出的一些措施，大多集中在环境条件领域，例如空气污染、水污染等，又或者是教育和产业结构方面，对于例如粮食、服装产业等也与民生息息相关的领域关注较少。因此，本文提出了一个思路，可以将目光更多转移到这些领域，例如粮食、餐饮业、服装业等，分析其与人口流动之间的联系，为人口流动的研究提出建议。

本文同样也存在着一些的不足之处，由于影响人口流动的因素是多样化的，仅从粮食产量这一点来解释人口流动说服力不足。此外，即使在预测过程中，通过两种方法建立的回归模型里，江苏省年水稻产量的回归系数均为负值，也不能断定水稻产量和人口流动是负相关的，还需要结合更多的变量，研究更多的地区进行综合讨论。因此在之后的研究中，希望能将更多变量引入模型中，建立多变量之间的因

果模型，对人口流动提供更准确，更客观的预测。此外，由于本研究所涉及的样本数量较少，导致神经网络、随机森林等机器学习方法不能有效准确地运用，这还需要引入更多样本，找到合适的模型，进行进一步的验证与分析。

文章中只对江苏省非自增长人口和江苏省年水稻产量这二者之间的因果关系进行分析，并且只考虑到农业对人口流动的影响这一方面，对问题的研究不够全面。因此，在后续的研究中，希望能建立起多元时间序列之间的因果分析，不仅仅只关注农业，也将餐饮业，服装业等领域引入模型，并且根据不同的时间区间，计算出每一个时间点上，各自的梁氏－克里曼信息流，行成一个随时间变化的，动态的因果网络，并研究更准确，更合适的预测方法，对未来的人口流动进行便捷而有效的预测。

基金项目

国家自然科学基金委员会面上项目(12171247)。

参考文献

[1] 藏媛, 郝枫. 空气质量对流动人口城市留居意愿强度的影响[J/OL]. 软科学, 2020(10): 48-61.

[2] 宫湛秋, 孙诚, 李建平, 冯娟, 谢飞, 杨韵, 薛佳庆. 基于信息流理论的因果分析在辨析大西洋多年代际振荡物理机制中的应用[J]. 大气科学, 2019, 43(5): 1081-1094.

[3] Liang, X.S. (2016) Information Flow and Causality as Rigorous Notions *ab initio*. *Physical Review E*, **94**, 13-40. https://doi.org/10.1103/PhysRevE.94.052201

[4] Liang, X.S. (2014) Unraveling the Cause-Effect Relation between Time Series. *Physical Review E*, **90**, 105-116. https://doi.org/10.1103/PhysRevE.90.052150