

哪些因素对流动人口落户意愿的影响更大？ ——基于随机森林的变量重要性排序

黄造玉^{1,2}, 徐晓新¹, 李 升³

(1. 北京师范大学社会发展与公共政策学院, 北京 100875;

2. 科学技术部火炬高技术产业开发中心, 北京 100045;

3. 北京工业大学文法学院, 北京 100124)

摘要: 准确理解流动人口落户意愿的影响因素, 对于科学制定我国人口和城镇化政策具有重要意义。基于 2017 年全国流动人口动态监测数据, 以随机森林模型和平均不纯净度降低方法系统评估了影响流动人口落户意愿的多维因素。研究发现: 对落户意愿影响最大的是流动人口想融入流入地的希望和可以在流入地生活下去的能力。具体而言, 按重要性排序, 第一梯队是社会心理融合因素; 第二梯队是经济、住房和户籍因素; 第三梯队则是流动特征、个体、地区、工作、制度融合和社会关系融合因素。总体而言, 所有单一变量的影响都是非常有限的。这表明, 流动人口落户意愿的影响因素是高度多元的。因此, 推动流动人口落户和市民化需要系统综合的政策设计。

关键词: 落户意愿; 流动人口; 随机森林; 变量重要性; 主体功能区

中图分类号: C922

文献标识码: A

文章编号: 1005-0566(2023)04-0076-10

Which factors are more influential on floating population's intention for urban household registration and settlement: Variable importance ranking based on random forest

HUANG Zaoyu^{1,2}, XU Xiaoxin¹, LI Sheng³

(1. School of Social Development and Public Policy, Beijing Normal University, Beijing 100875, China;

2. Torch High Technology Industry Development Center, Ministry of Science & Technology, Beijing 100045, China;

3. Faculty of Humanities and Social Sciences, Beijing University of Technology, Beijing 100124, China)

Abstract: A comprehensive understanding of the influencing factors of the floating population's intention for urban household registration and settlement is of great significance for the decision-making of China's urbanization policies. Based on the 2017 China Migrants Dynamic Survey, using random forest model and mean decrease impurity (MDI) method, this paper finds that the most crucial factors of intention for urban household registration and settlement are the hope that the floating population want to integrate into cities and the ability to live in cities. But in general, the impact of any single variable is quite limited, which indicates that the influencing factors are highly diversified. It implies that a systematic and comprehensive policy design is needed to promote the floating population to register urban household and settlement.

收稿日期: 2022-06-22 修回日期: 2023-02-19

作者简介: 黄造玉(1995—), 女, 北京人, 科学技术部火炬高技术产业开发中心助理工程师, 研究方向为政策分析。通信作者: 徐晓新。

Key words: intention for urban household registration and settlement; floating population; random forest; variable importance; main-functional zones

促进有能力在城镇稳定就业生活的农业转移人口实现市民化,是新型城镇化的首要任务。理清流动人口城市落户意愿的影响因素及其重要性排序,对于优化政策设计具有重要意义。

关于流动人口城市落户意愿影响因素的研究,主要聚焦个体因素和地区因素。这些研究各自得出了颇有见地的结论,但彼此又存在诸多差异,相应的政策建议也各有侧重甚至相互矛盾(详见文献回顾)。这些研究发现的差异主要来自理论视角和数据来源等方面。首先,已有实证研究多是基于特定理论视角提出研究假设,再基于调查数据进行假设检验,分析方法以多元回归模型为主。尽管研究者也会将其他变量纳入分析模型,但多侧重于特定理论视角所关注的核心变量对落户意愿是否有影响以及影响大小,而不是综合评估多维变量的相对重要性。其次,有一些研究是基于特定地区的调查数据,地区差异使同一因素的影响有所不同,这在相对早期的研究中表现更为突出。鉴于此,基于国家卫生健康委员会2017年全国流动人口动态监测数据,本文尝试打破单一理论视角局限,采用随机森林模型和平均不纯净度降低方法将调查数据中所包含的尽可能多的变量都纳入模型统一进行重要性排序。相对于已有文献,本研究的优势在于,可系统地评估多维影响因素的重要性,为确定相关政策优先级提供实证支持。

一、理论和文献回顾

(一) 落户意愿及相关概念

关于落户意愿,不同研究所采用的概念有所差异,城市定居意愿、留城意愿、入户意愿和永久性迁移意愿等都可看作是对落户意愿问题的探讨^[1]。也有学者区分了“居留意愿”和“落户意愿”,并认为“有长期居留意愿的流动人口才可能有落户的需求”^[2]。根据流动人口的落户意愿和居留意愿,可划分出4种理想型(见表1)。新型城

镇化战略的方向是促进流动人口市民化,即要推动B、C和D3种类型向A转化^[3]。

表1 流动人口落户—居留意愿的四种类型

变量	落户意愿	
	是	否
居留意愿	是 A 制度永久—行为永久 (合法性永久迁移) 落户—定居	B 制度非永久—行为永久 (事实性永久迁移) 不落户—定居
	否 C 制度永久—行为非永久 落户—不定居	D 制度非永久—行为非永久 不落户—不定居

资料来源:基于参考文献[3],略有修改。

(二) 流动人口落户意愿的影响因素

个体和地区层面的诸多因素都会对流动人口的落户意愿产生影响。相关理论框架主要有社会融合理论、人力资本模型、新劳动力理论、迁移理论、空间区位理论等(不同理论框架的核心观点及其变量选择参见表2),这些理论框架均被实证研究证明具有一定解释力。从研究方法来看,多以量化研究为主,通常是选定少数几个或一个关键的影响因素后,将其他因素作为控制变量,着重分析所选定的因素与落户意愿之间的关系或其影响落户意愿的机制。所采用的模型主要有logistic回归、probit回归和倾向值匹配等。

在对落户意愿的影响因素重要性的认识上,已有研究得出了颇具启发性的结论,但彼此间不乏矛盾之处。如有的研究认为,经济因素最为重要^[4-5];有的研究发现,经济因素的核心变量——收入,对落户意愿并没有显著影响^[6-7]。有的研究认为,较之经济因素,住房因素对落户意愿的影响更大^[8];有的研究则认为,从长期来看,工作稳定性比住房稳定性更为重要^[9]。在经济、住房和工作因素之外,户籍(土地权益)^[10]、子女教育^[11]、公共服务获得(社会保险)^[12]等也是已有研究中关注到的重要影响因素。在这些研究中发现,多维影响因素对落户意愿的相对重要性仍不甚清晰,而相应地,决策者难以基于不同视角的实证研究所提出的建议确定政策优先级。这正是本文试图作出贡献的空间。

表 2 流动人口落户意愿的主要理论框架和相关变量

理论框架	机制	归因方式	变量
个体层面			
社会融合理论	社会融合的正向回报机制	客观社会融合程度高,对流入地的认同度和归属感高的流动人口更倾向于落户	社会融合程度:流入地制度融合(建立居民健康档案、办理个人社会保障卡);流入地社会关系融合(参加社区党团活动、和本地人来往最多);流入地社会心理融合(认为自己是本地人、认为本地人不排斥外地人)
人力资本模型	人力资本的正向经济回报机制	迁移能力较强,人力资本较高的流动人口更倾向于落户	个体特征:性别、年龄、婚姻状况、受教育程度、收入、健康自评;流动特征:外出(流动)经历、流动时间等
新劳动力迁移理论	家庭资源的正向激励机制	在流入地家庭成员较多,家庭资产水平更高的流动人口更倾向于落户	家庭特征:本地家庭人口数、家庭收支情况、住房状况等
地区层面			
制度差异	城乡户籍资源禀赋激励机制	原户籍资源禀赋较高的流动人口更倾向于不落户	流出地禀赋:承包地面积、承包地收入、宅基地面积、村集体分红等
空间区位理论	区位差异	流入地区位条件更优越,流动人口更倾向于落户	流入地区位特征:如位于东部或沿海地区、属于优化开发区、重点开发区或限制开发区

二、研究设计

(一)数据与变量说明

1. 数据来源

本文使用 2017 年全国流动人口动态监测数据,调查对象为在本地居住一个月及以上,非本区(县、市)户籍的流动人口。该数据在一定程度上反映了全国流动人口的整体情况,并且基本包括了以往相关研究中涉及到的影响流动人口落户意愿的各类因素。如无特殊说明,下文的数据均为加权后的结果。

2. 变量说明

(1)因变量说明

本文兼顾居留意愿和落户意愿,认为愿意将户口迁入流入地并有在流入地长期居留意愿的流动人口,是符合政策导向、“真正”愿意落户的群体。落户意愿以对“如果您符合本地落户条件,您是否愿意把户口迁入本地?”的回答来刻画,“愿意”编码为 1,“不愿意”和“没想好”编码为 0。居留意愿则以对“如果您打算留在本地,您预计自己将在本地留多久?”的回答来刻画。流动人口预计自己在流入地居留 10 年以上(包括“定居”)编码为 1,在流入地居留 10 年及以下(包括“没想好”)编码为 0。“落户—居留”是本文希望识别出的符合政策导向的流动人口落户意愿类别,在下文中简称为“落户”,其他类型则统称为“不落户”。

(2)自变量说明

为系统评估多维因素对落户意愿的相对重要

表 3 研究样本落户意愿和居留意愿分布

落户意愿	居留意愿		
	不落户	落户	总计
不居留	80 965 (48.0%)	31 944 (18.9%)	112 909 (66.9%)
居留	16 624 (9.9%)	39 199 (23.2%)	55 823 (33.1%)
总计	97 589 (57.8%)	71 143 (42.2%)	168 732 (100.0%)

性,本文尽可能全面地将变量纳入分析之中。以相关文献为基础,本文将作为影响因素的自变量分为个体层面和地区层面两大类。个体层面包括个体因素、经济因素、工作因素、住房因素、流动特征因素、户籍因素、制度融合、社会关系融合、社会心理融合和生活情况 10 类。地区层面则由流动人口户籍地地理位置、流入地所在的主体功能区和“两横三纵”城市化战略划分的 21 个城市化地区表示。以县级行政区为基础单元的主体功能区规划为认识流动人口问题提供了一个空间上和政策上的总体性框架,它在一定程度上可视为衡量地区情况的综合指标。主体功能区人口政策中关于流动人口本地化的要求、新型城镇化的政策目标,以及支持流动人口市民化的“人地钱挂钩”政策等,都旨在促进流动人口在流入城市落户并长期稳定发展。下文将重点关注城市化地区(优化开发区、重点开发区和限制开发区)的流动人口样本,去掉流入地为禁止开发区和不便分类区域的样本,保留样本数为 168 732。变量操作化及描述性统计结果详见表 4。

表4 自变量操作与描述性统计结果

变量	频数	频率/%	变量	频数	频率/%
个体因素			住房因素		
性别*			目前住房性质*		
1 = 男性	86 728	51.4	0 = 非正规居所	781	0.5
年龄			1 = 单位/雇主房	20 116	11.9
(均值 37.174 6; 标准差 10.933 8)			2 = 公租房	1 094	0.6
婚姻状况*			3 = 自购房	31 805	18.8
0 = 未婚	26 126	15.5	4 = 借住房	1 864	1.1
1 = 已婚	138 363	82.0	5 = 就业场所	3 566	2.1
2 = 离婚	2 684	1.6	6 = 自建房	3 153	1.9
3 = 丧偶	1 559	0.9	7 = 整租	85 376	50.6
受教育程度			8 = 合租	20 977	12.4
0 = 未上过小学	4 500	2.7	家庭月住房支出		
1 = 小学	24 233	14.4	(均值 5.023 1; 标准差 2.936 8)		
2 = 初中	74 306	44.0	流动因素		
3 = 高中/中专	37 033	21.9	本次流动范围*		
4 = 大学专科	16 653	9.9	0 = 跨省	110 833	65.7
5 = 大学本科及以上	12 007	7.1	1 = 省内跨市	43 127	25.6
健康状况			2 = 市内跨县	14 772	8.8
0 = 健康	139 917	82.9	本次流动时间		
1 = 基本健康	25 165	14.9	(均值 7.418 0; 标准差 6.143 6)		
2 = 不健康	3 650	2.2	本次流动原因*		
政治身份*			0 = 务工工作	115 193	68.3
0 = 中共党员	8 320	4.9	1 = 经商	32 322	19.2
1 = 共青团员	8 536	5.1	2 = 家庭流动	13 196	7.8
2 = 其他	151 876	90.0	3 = 照顾家人	4 899	2.9
经济因素			4 = 拆迁搬家	559	0.3
家庭月总支出			5 = 投靠亲友	998	0.6
(均值 8.115 8; 标准差 0.681 3)			6 = 养老	734	0.4
家庭月总收入			7 = 其他	832	0.5
(均值 8.841 5; 标准差 0.620 6)			总流动城市数		
月收入			(均值 2.161 3; 标准差 1.951 7)		
(均值 7.032 6; 标准差 3.037 2)			父母流动经历*		
月收入变化*			0 = 记不清	3 189	1.9
0 = 不适用	27 631	16.4	1 = 父母均有	27 701	16.4
1 = 减少	39 311	23.3	2 = 父亲或母亲有	11 328	6.7
2 = 基本不变	76 328	45.2	3 = 父母均没有	126 095	74.7
3 = 增加	25 462	15.1	4 = 出生就流动	419	0.2
工作因素			户口性质*		
周工作小时数			0 = 农业户口	134 138	79.5
(均值 48.378 6; 标准差 25.808 5)			1 = 非农业户口	21 085	12.5
单位性质*			2 = 居民户口/其他	13 509	8.0
0 = 无工作	24 341	14.4	户籍因素		
1 = 机关、事业单位	3 176	1.9	承包地耕种情况*		
2 = 国有及国有控股企业	5 274	3.1	0 = 没有承包地	97 111	57.6
3 = 集体企业	1 317	0.8	1 = 自己/家人耕种	35 889	21.3
4 = 股份/联营企业	6 118	3.6	2 = 他人耕种	18 255	10.8
5 = 个体工商户	48 822	28.9	3 = 转租	10 266	6.1
6 = 私营企业	51 117	30.3	4 = 撂荒	5 570	3.3
7 = 港澳台独资企业	5 128	3.0	5 = 种树	969	0.6
8 = 外商独资企业	3 650	2.2	6 = 有承包地 - 其他	672	0.4
9 = 中外合资企业	2 452	1.5	户籍地是否有宅基地*		
10 = 社团/民办组织	489	0.3	1 = 有	104 014	61.6
11 = 其他	2 496	1.5	户籍地是否有集体分红*		
12 = 无单位	14 355	8.5	1 = 有	3 947	2.3

表 4 自变量操作与描述性统计结果(续)

变量	频数	频率/%	变量	频数	频率/%
就业身份*			生活情况		
0 = 无工作	24 341	14. 4	是否有本地生活困难*		
1 = 有固定雇主的雇员	85 960	50. 9	1 = 有	92 768	55. 0
2 = 无固定雇主的雇员	9 509	5. 6	本地困难—生意难做*		
3 = 雇主	8 907	5. 3	1 = 有	38 943	23. 1
4 = 自营劳动者	38 158	22. 6	本地困难—找稳定工作*		
5 = 其他	1 857	1. 1	1 = 有	31 826	18. 9
制度融合			本地困难—买不起房		
本地参加医疗保险*			1 = 有	61 181	36. 3
1 = 是	55 016	32. 6	本地困难—歧视*		
户籍地参加医疗保险*			1 = 有	11 045	6. 5
1 = 是	112 401	66. 6	本地困难—子女上学*		
其他地区参加医疗保险*			1 = 有	30 165	32. 5
1 = 是	1 475	0. 9	本地困难—收入低*		
是否办理暂住证/居住证*			1 = 有	64 516	38. 2
1 = 是	122 373	72. 5	本地困难—生活不习惯*		
是否建立居民健康档案*			1 = 有	8 857	5. 2
1 = 是	41 216	24. 4	是否有老家生活困难*		
社会关系融合			1 = 有	68 673	40. 7
我喜欢我现在居住的地方			老家困难—老人赡养*		
0 = 完全不同意	1 417	0. 8	1 = 有	48 004	28. 5
1 = 不同意	3 363	2. 0	老家困难—子女照看*		
2 = 基本同意	96 132	57. 0	1 = 有	17 885	10. 6
3 = 完全同意	67 821	40. 2	老家困难—子女教育费用*		
我关注我现在居住地方的变化			1 = 有	13 664	8. 1
0 = 完全不同意	1 681	1. 0	老家困难—配偶生活*		
1 = 不同意	7 693	4. 6	1 = 有	3 541	2. 1
2 = 基本同意	95 522	56. 6	老家困难—缺钱治病*		
3 = 完全同意	63 835	37. 8	1 = 有	20 997	12. 4
我很愿意融入本地人当中,成为其中一员			老家困难—缺劳动力*		
0 = 完全不同意	1 869	1. 1	1 = 有	22 322	13. 2
1 = 不同意	13 313	7. 9	地区层面		
2 = 基本同意	91 575	54. 3	城市化地区*		
3 = 完全同意	61 975	36. 7	0 = 环渤海地区	14 658	8. 7
我觉得本地人愿意接受我成为其中一员			1 = 长江三角洲地区	41 626	24. 7
0 = 完全不同意	2 080	1. 2	2 = 珠江三角洲地区	43 461	25. 8
1 = 不同意	15 861	9. 4	3 = 冀中南地区	691	0. 4
2 = 基本同意	100 951	59. 8	4 = 太原城市群	1 602	0. 9
3 = 完全同意	49 840	29. 5	5 = 呼包鄂榆地区	3 601	2. 1
我感觉本地人看不起外地人			6 = 哈长地区	851	0. 5
0 = 完全不同意	24. 6	24. 6	7 = 东陇海地区	1 744	1. 0
1 = 不同意	55. 3	55. 3	8 = 江淮地区	1 639	1. 0
2 = 基本同意	17. 3	17. 3	9 = 海峡西岸经济区	17 242	10. 2
3 = 完全同意	2. 8	2. 8	10 = 中原经济区	1 506	0. 9
按照老家的风俗习惯办事对我比较重要			11 = 长江中游地区	5 264	3. 1
0 = 完全不同意	15 164	9. 0	12 = 北部湾地区	2 024	1. 2
1 = 不同意	60 291	35. 7	13 = 成渝地区	6 797	4. 0
2 = 基本同意	72 049	42. 7	14 = 黔中地区	3 891	2. 3
3 = 完全同意	21 228	12. 6	15 = 滇中地区	1 793	1. 1
我的卫生习惯与本地市民存在较大差别			16 = 藏中南地区	155	0. 1
0 = 完全不同意	41 349	24. 5	17 = 关中天水地区	890	0. 5
1 = 不同意	93 903	55. 7	18 = 兰州—西宁地区	1 074	0. 6
2 = 基本同意	28 683	17. 0	19 = 宁夏沿黄经济区	241	0. 1
3 = 完全同意	4 797	2. 8	20 = 天山北坡地区	2 096	1. 2

表 4 自变量操作与描述性统计结果(续)

变量	频数	频率/%	变量	频数	频率/%
我觉得我已经是本地人了			主体功能区*		
0 = 完全不同意	7 559	4.5	0 = 优化开发区	102 313	60.6
1 = 不同意	51 757	30.7	1 = 重点开发区	55 790	33.1
2 = 基本同意	79 516	47.1	2 = 限制开发区	10 629	6.3
3 = 完全同意	29 900	17.7	户籍位置		
本地社会交往			0 = 农村	131 697	78.1
0 = 很少与人来往	38 360	22.7	1 = 乡镇	17 766	10.5
1 = 同乡本地户口	4 680	2.8	2 = 县城	10 547	6.3
2 = 同乡老家户口	58 686	34.8	3 = 地级市	5 964	3.5
3 = 同乡其他户口	3 815	2.3	4 = 省会城市	1 996	1.2
4 = 其他本地人	35 984	21.3	5 = 直辖市	761	0.5
5 = 其他外地人	27 207	16.1			
没回老家时间					
0 = 大于等于 1 年	26 012	15.4			
1 = 小于 1 年	142 720	84.6			

注:标“*”的变量按哑变量处理,如变量只有两个类别赋值(1和0,含义为“男/女”“有/无”“是/否”等),则仅报告赋值为1的情况。收入和支出变量为对原始变量加1后求自然对数(ln)。

(二)模型选择、变量重要性评分和分析工具

在数据的置信度有一定保证,且样本量较充足的情况下,本文使用随机森林(Random Forests)进行变量重要性排序。相对于传统假设—检验模型,随机森林的优势在于它是非参数的,允许数据包含非线性关系和交互作用,输入相关数据的数量多有助于模型训练。本文采用 CART(Classification And Regression Tree)算法^[13],分裂准则为基尼指数。在随机森林超参数的确定过程中,使用正确率(accuracy)作为评价模型性能的指标,超参数包括基学习器数量、最大特征数量、树的深度和叶子节点最小样本数。在随机森林模型的性能评价上,使用准确率(precision)、召回率(recall)和 F1 分数(F1 score)3 个指标。

在变量重要性评分方面,本文采用平均不纯净度降低方法(Mean Decrease Impurity, MDI),统计每个自变量在随机森林中每棵决策树上的贡献大小(以基尼指数衡量),加总后取平均值,最后比较所有自变量贡献平均值的大小。自变量在模型中为正确识别样本类别提供的信息越多,重要性评分越高,排序越靠前。

编码和描述性统计使用 SPSS 和 EXCEL,随机

森林和变量重要性排序的实现语言为 Python。本文以全国流动人口数据建立模型 1;考虑到落户意愿的影响因素在不同地区可能有所差异,分别以优化开发区、重点开发区和限制开发区的流动人口建立模型 2 至模型 4。

三、实证分析结果

所有变量的重要性评分经过归一化后和为 1,本文报告重要性在约前 20.0% 的变量,并认为这些变量在影响流动人口落户意愿中发挥了更大作用^①。限于篇幅,仅以模型 1 为例汇报详细结果,并报告模型 1 至模型 4 的汇总结果。

(一)流动人口落户意愿影响因素重要性排序

在表 5(模型 1)呈现的全国流动人口落户意愿影响因素重要性排序中,前 28 个变量提供了约 85.0% 的分类所需信息。总体来看,个体因素的影响超过地区因素。其中,社会心理融合因素最为突出,尤其是流动人口对“我很愿意融入本地人当中,成为其中一员”的判断,贡献了约 12.9% 的信息。衡量流动人口在流入地基本生活情况的变量,如月收支情况、月住房支出、住房性质等,也有较大影响。户籍因素中的户口性质、是否有宅基地和是否有承包地重要性排序也相对靠前。

① 在超参数确定后,根据变量重要性排序每次去掉 5.0% 最不重要的变量重新训练模型发现,基本上保留前 20.0% 的变量时模型性能较好。

表 5 流动人口落户意愿影响因素重要性排序

序号	变量	重要性评分	序号	变量	重要性评分
1	我很愿意融入本地人当中,成为其中一员	0.128 930	15	户籍地参加医疗保险	0.024 373
2	住房性质—自购房	0.081 650	16	家庭平均月收入	0.022 265
3	家庭平均月支出	0.056 405	17	年龄	0.021 565
4	月住房支出	0.052 906	18	户口性质—非农业户口	0.021 500
5	月工资收入	0.048 307	19	是否有宅基地	0.020 161
6	本次流动时间	0.040 930	20	本地参加医疗保险	0.017 545
7	户口性质—农业户口	0.035 476	21	主体功能区—重点开发区	0.015 034
8	本地人愿意接受我成为其中一员	0.035 280	22	住房性质—自建房	0.013 042
9	主体功能区—优化开发区	0.034 658	23	流动范围—跨省	0.011 438
10	我喜欢我现在居住的地方	0.032 544	24	按照老家的风俗习惯办事对我比较重要	0.010 534
11	我关注我现在居住地方的变化	0.027 873	25	是否有承包地	0.010 490
12	周工作小时数	0.025 065	26	户籍所在地	0.010 065
13	我觉得我已经是本地人了	0.025 060	27	本地社会交往—其他本地人	0.007 461
14	受教育程度	0.024 678	28	是否办理居住证/居住证	0.006 384

表 6 呈现了对模型 1 的性能评价^②。总体来看,模型 1 能较好实现对全国流动人口落户意愿的分类,正确率为 86.0%,识别“落户”类别的准确率达 94.0%。

表 6 模型 1 性能评价

变量		准确率	召回率	F1 分数	样本量
落户意愿	落户	0.94	0.47	0.63	38 821
	不落户	0.86	0.99	0.92	124 968
平均值		0.90	0.73	0.77	163 789
加权平均值		0.88	0.87	0.85	163 789
正确率		0.86			163 789

表 7 汇总了模型 1 至模型 4 的结果。在不同主体功能区中,影响落户意愿的主要因素是基本稳定的:①个体因素中,年龄和受教育程度影响较

大;②经济因素中,个人月收入和家庭月收支影响较大;③工作因素中,工作时长影响较大;④住房因素中,家庭月住房支出,以及住房性质的“自购房”和“自建房”类别影响较大;⑤流动因素中,本次流动时间和流动范围发挥了主要作用;⑥户籍因素中,户口性质,是否有宅基地和承包地影响较大;⑦与社会融合相关的变量几乎都进入了重要性排序的前 20.0%,其中,社会心理融合重要性评分最高,本地交往人员情况中,仅“其他本地人”类别影响较大;⑧地区因素中,主体功能区中的“优化开发区”和“重点开发区”类别对落户意愿的影响较大,流入地所在的城市群类别并未对落户意愿的分类提供太多信息。

表 7 不同地区流动人口落户意愿影响因素重要性排序

全国		优化开发区		重点开发区		限制开发区	
影响因素	重要性评分	影响因素	重要性评分	影响因素	重要性评分	影响因素	重要性评分
社会心理融合	0.260 221	社会心理融合	0.304 943	社会心理融合	0.248 405	社会心理融合	0.138 221
住房因素	0.147 598	住房因素	0.240 696	户籍因素	0.103 006	经济因素	0.129 021
经济因素	0.126 977	经济因素	0.118 635	住房因素	0.097 247	住房因素	0.099 630
户籍因素	0.087 627	户籍因素	0.062 036	经济因素	0.093 773	流动因素	0.084 189
地区因素	0.059 757	流动因素	0.055 383	个体因素	0.063 297	个体因素	0.075 191
流动因素	0.052 368	个体因素	0.044 462	流动因素	0.059 510	工作因素	0.045 086
制度融合	0.048 302	制度融合	0.024 161	制度融合	0.055 887	制度融合	0.030 564
个体因素	0.046 243	工作因素	0.016 989	工作因素	0.029 735	社会关系融合	0.028 779
工作因素	0.025 065	地区因素	0.010 620	社会关系融合	0.023 337	地区因素	0.011 673
社会关系融合	0.007 461	社会关系融合	0.009 044	地区因素	0.010 951	户籍因素	0.000 000

(二)模型结果讨论

基于上文分析结果,本部分主要关注:①变量

重要性排序结果是稳定的吗? ②社会心理融合因素这一表征主观态度变量的影响是否可被研究中

② 模型性能评价中的样本量为非加权结果。下文相同。

其他客观因素所解释?

1. 变量重要性排序的稳定性

基于 MDI 的变量重要性排序会偏向于选择那些具有较多类别的变量^[14]。当存在相关变量时,一旦某个变量被使用,其相关变量的重要性评分会大大降低。为验证排序是稳定且有效的,本文使用平均正确率降低 (Mean Decrease Accuracy,

MDA) 方法进行比较分析。如表 8 所示,不同算法对模型 1 呈现出的排序结果基本相同^③,这表明本文使用 MDI 方法对变量重要性的排序是有效的。对其他模型的验证也得到了类似结果。这从侧面说明本文中随机森林的超参数设置是较为合适的,模型很好地控制了基学习器之间的相关性。

表 8 流动人口落户意愿影响因素重要性排序比较

平均不纯净度降低 (MDI)			平均正确率降低 (MDA)		
序号*	变量	重要性评分	序号*	变量	正确率变化
1	我很愿意融入本地人当中,成为其中一员	0.128 930	1	我很愿意融入本地人当中,成为其中一员	0.016 252
2	住房性质—自购房	0.081 650	2	主体功能区—优化开发区	0.008 320
3	家庭平均月支出	0.056 405	3	住房性质—自建房	0.005 995
4	月住房支出	0.052 906	4	住房性质—自购房	0.004 953
5	月工资收入	0.048 307	5	是否有宅基地	0.004 018
6	本次流动时间	0.040 930	6	我喜欢我现在居住的地方	0.003 831
7	户口性质—农业户口	0.035 476	7	户籍地参加医疗保险	0.003 124
8	本地人愿意接受我成为其中一员	0.035 280	8	我觉得我已经是本地人了	0.002 821
9	主体功能区—优化开发区	0.034 658	9	本次流动时间	0.002 426
10	我喜欢我现在居住的地方	0.032 544	10	月住房支出	0.002 219
11	我关注我现在居住地方的变化	0.027 873	11	本地参加医疗保险	0.001 921
12	周工作小时数	0.025 065	12	户口性质—农业户口	0.001 895
13	我觉得我已经是本地人了	0.025 060	13	受教育程度	0.001 800
14	受教育程度	0.024 678	14	家庭平均月收入	0.001 749
15	户籍地参加医疗保险	0.024 373	15	我关注我现在居住地方的变化	0.001 708
16	家庭平均月收入	0.022 265	16	是否有承包地	0.001 567
17(29)	年龄	0.021 565	17	主体功能区—重点开发区	0.001 420
18	户口性质—非农业户口	0.021 500	18	家庭平均月支出	0.001 375
19	是否有宅基地	0.020 161	19	周工作小时数	0.001 268
20	本地参加医疗保险	0.017 545	20	流动范围—跨省	0.001 188
21	主体功能区—重点开发区	0.015 034	21	户口性质—非农业户口	0.001 158
22	住房性质—自建房	0.013 042	22(31)	城市化地区—长三角	0.001 092
23	流动范围—跨省	0.011 438	23	月工资收入	0.000 829
24	按照老家的风俗习惯办事对我比较重要	0.010 534	24(34)	没回老家时间	0.000 778
25	是否有承包地	0.010 490	25	本地人愿意接受我成为其中一员	0.000 758
26(36)	户籍所在地	0.010 065	26(47)	流动范围—市内跨县	0.000 748
27(39)	本地社会交往—其他本地人	0.007 461	27	是否办理居住证/居住证	0.000 743
28	是否办理居住证/居住证	0.006 384	28	按照老家的风俗习惯办事对我比较重要	0.000 682

注: * 括号中的序号表示在另一个方法中该变量的排序。未标记括号则说明在两个方法中,该变量的重要性排序均在前 28 位。

2. 社会心理融合因素的影响

本文分析结果显示,社会心理融合因素是影响落户意愿最重要的变量,而已有研究表明,社会融合因素和本文中分析的其他客观因素在理论和实证上都是相关的。鉴于此,有必要进一步认识社会心理融合因素的重要性。

以 MDI 进行变量重要性排序时,如果一个变量尽管重要但与其他变量高度相关,那么在模型训练时去掉这个变量,相关变量会补充损失的信息,使模型性能未必明显下降。基于 MDI 的这一特性,本文以模型 1 为例,去掉测量社会心理融合的 8 个变量建立随机森林模型 1-1,观察模型性

③ 比较 MDI 和 MDA 两种方法中进入一定重要性范围的变量名单是有意义的,但对两种变量重要性排序做一一对照比较的意义可能不是很大。首先,如表 8 所示,重排变量值带来的模型准确率变化不大,且每个变量之间的差异更小,模型中细微的改变都可能带来 MDA 中变量重要性排序的变化;其次,受算力限制,本文中的准确率变化值是 5 次随机选择样本计算的平均值,增加随机采样次数求平均值可获得更稳定的排序结果,但变量排序所处的范围不会变化。

能和变量重要性排序的变化。与模型 1 相比,模型 1-1 性能明显下降(指标见表 9),变量重要性排序则变化不大(结果从略)。这表明模型 1 的变量重要性排序结果是较为稳定的,社会心理融合因素表征了一些本文所分析的其他客观变量无法替代的内容。

表 9 模型 1-1 性能评价

变量		准确率	召回率	F1 分数	样本量
落户意愿	落户	0.88	0.39	0.54	38 821
	不落户	0.84	0.98	0.91	124 968
平均值		0.86	0.69	0.72	163 789
加权平均值		0.85	0.84	0.82	163 789
正确率		0.84			163 789

四、结论与讨论

本文以主体功能区规划为框架划分流动人口流入地,基于 2017 年流动人口动态监测数据,以随机森林模型和平均不纯净度降低方法,对流动人口落户意愿影响因素的重要性进行了排序,并以平均正确率降低方法对排序稳定性进行了验证。

(1)高重要性的影响因素涉及范围很广。在纳入分析的 100 多个变量中,重要性评分最高的前 20.0% 涉及个体因素、经济因素、工作因素、住房因素、流动因素、户籍因素、社会融合因素和地区因素等各个方面。

(2)对落户意愿影响最大的是流动人口想融入流入地的希望和可以在流入地生活下去的能力。在落户意愿上,流动人口是积极的决策者,相比于限制性因素,如在流入地和老家面临的各类困难,流动人口融入流入地的希望和能力为落户意愿的分类提供了更多信息。在融入的希望方面,愿意融入流入地的态度“我很愿意融入本地人当中,成为其中一员”、对流入地的好感“我喜欢我现在居住的地方”、对本地人的积极判断“我觉得本地人愿意接受我为其中一员”等在不同主体功能区中都是为流动人口落户意愿分类提供较多信息的变量;在融入的能力方面,经济实力(月收支情况),稳定的住房(自购房和自建房的“自有”住房性质),以及表征获取收入和住房能力的个体因素(年龄和受教育程度)是高重要性评分的变量。

(3)社会心理融合因素、经济因素和住房因素在不同主体功能区中都是对落户意愿影响较大的

因素。按照变量重要性评分从高到低排序,排在第一梯队的是社会心理融合因素,第二梯队的是经济因素、住房因素和户籍因素,第三梯队的是流动因素、个体因素、地区因素、工作因素、制度融合和社会关系融合。尽管在不同主体功能区中落户意愿影响因素的重要性排序不尽相同,但 3 个梯队的相对排序是基本稳定的。相对于已有基于特定理论视角的研究,本文提供了多维影响因素相对重要性的综合排序,为确定政策优先级提供了数据基础。

研究表明,社会心理融合因素不仅是为落户意愿分类提供信息最多的变量,而且它所表征的部分信息是独立于客观因素的,而此前的研究多将其视为中介变量。研究发现,纳入模型的变量中鲜有变量对落户意愿分类是完全没有贡献的,而单个变量的改变对模型整体分类正确率的影响都是微小的。这表明,影响流动人口落户意愿的因素是非常多元的,推动流动人口落户需要系统综合的政策“组合拳”。这一点呼应了已有研究得出的单纯降低落户门槛的边际落户效应递减,“降低落户门槛+综合配套改革”的边际落户效应更为积极持续^[15],“户籍制度改革需要农村土地制度、城市住房保障制度等协同推进”^[16]等研究发现。

将基于随机森林的变量重要性测度方面应用到流动人口落户意愿的影响因素重要性排序上,是本文在研究方法上的探索和创新,此前这一方法多应用在生物信息学等“硬科学”上。在这一方法的运用上,还有一些问题值得进一步思考和探讨。如从模型评价结果来看,针对不同主体功能区所构建模型的性能存在明显差异。在重点开发区(模型 3)中,尽管模型对流动人口落户意愿识别的正确率为 82.0%,但在类别“落户”上的准确率为 76.0%,召回率为 21.0%,F1 分数为 0.33。这意味着,在重点开发区中可能还存在其他关键变量未纳入模型。因此,在基于标准问卷进行全国性流动人口调查基础上,在特定区域开展更深入的流动人口落户意愿相关访谈以识别被忽视的流动人口落户意愿的影响因素对于优化政策设计仍

是非常必要的。另外需要指出的是,本文使用的数据并非是“自然”数据,而是已结构化的数据,这制约了对可能更“有效”的变量的发现能力。这也是本文研究的主要局限所在。

参考文献:

- [1]肖璐,马成成. 农民工城市落户影响因素的文献梳理、述评与展望[J]. 江苏大学学报(社会科学版),2019,21(2):86-92.
- [2]邱红,周文剑. 流动人口的落户意愿及影响因素分析[J]. 人口学刊,2019,41(5):91-100.
- [3]李飞,杜云素. 城镇定居、户籍价值与农民工积分落户——基于中山市积分落户入围人员的调查[J]. 农业经济问题,2016,37(8):82-92.
- [4]欧阳慧,胡杰成,刘保奎,等. 如何增强农民工在城镇的落户意愿?——基于对农民工分区域分群体的调查[J]. 城市发展研究,2019,26(6):52-60.
- [5]王春兰,丁金宏. 流动人口城市居留意愿的影响因素分析[J]. 南方人口,2007(1):22-29.
- [6]聂伟,风笑天. 就业质量、社会交往与农民工入户意愿——基于珠三角和长三角的农民工调查[J]. 农业经济问题,2016,37(6):34-42.
- [7]赵翌,郝明松,悦中山. 制度与非制度因素对农民工落户城镇意愿的影响[J]. 西北农林科技大学学报(社会科学版),2016,16(4):88-95.
- [8]张新,周绍杰,姚金伟. 居留决策、落户意愿与社会融合度——基于城乡流动人口的实证研究[J]. 人文杂志,2018(4):39-48.
- [9]邢朝国. 中国农业转移人口的城市落户意愿——以城市落户能力为分析框架[J]. 调研世界,2020(6):49-54.
- [10]张翼. 农民工“进城落户”意愿与中国近期城镇化道路的选择[J]. 中国人口科学,2011(2):14-26.
- [11]张龙. 农民工市民化意愿的影响因素研究[J]. 调研世界,2014(9):40-43.
- [12]杨义武,林万龙,张莉琴. 地方公共品供给与人口迁移——来自地级及以上城市的经验证据[J]. 中国人口科学,2017(2):93-103.
- [13]BREIMAN L, FRIEDMAN J, OLSHEN R, et al. Classification and regression trees[M]. Boca Raton: CRC Press, 2017:14-54.
- [14]STROBL C, BOULESTEIX A-L, ZEILEIS A, et al. Bias in random forest variable importance measures: illustrations, sources and a solution[J]. BMC bioinformatics, 2007, 8(1):25.
- [15]张义博,刘敏. 户籍制度改革的边际落户效应[J]. 宏观经济管理,2018(9):28-36.
- [16]苏红键. 中国流动人口城市落户意愿及其影响因素研究[J]. 中国人口科学,2020(6):66-77.

(本文责编:王延芳)