基于征信大数据分析的中国劳动力人口迁徙研究

姚前，谢华美，司恩哲，景志刚，胡青青
中国人民银行征信中心，北京 100031

摘要
人口迁徙数据是反映城镇化建设的重要数据，可以为我国当前推进国家新型城镇化建设提供决策支持。中国人民银行征信中心依法收集的征信信息及时地反映了信息主体在迁徙过程中发生经济行为的时间和地点，在确定人口迁徙轨迹方面具有其他数据源难以比拟的优势。对3.9亿信息主体、48.8亿条征信记录进行分析挖掘，得出改革开放以来全国劳动力人口迁徙的新特点，并利用Logistic模型预测未来青壮年劳动力人口的迁徙趋势，得出相关结论。

关键词
人口迁徙；征信系统；Logistic模型

中图分类号：TP311.132　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016055

Research of China's labor force migration based on big data analysis of credit information

YAO Qian, XIE Huamei, SI Enzhe, JING Zhigang, HU Qingqing
Credit Reference Center, the People's Bank of China, Beijing 100031, China

Abstract
The data of population migration is an important data source to reflect urbanization and can provide support for government to promote new urbanization construction. The Credit Reference Center, the People's Bank of China collects data from financial institutions which recorded time and place for economic behaviors during migration. The authentic and reliable data source give the Credit Reference Center an incomparable advantage to identify individual's immigration trajectory. A comprehensive analysis was conducted based on 4.88 billion personal credit records of 390 million citizens from the database of Credit Reference Center. Several new characteristics of labor force migration since the economic reform and open up were concluded based on the analysis, furthermore, a Logistic model was applied to predict future tendency of labor force migration.

Key words
population migration, credit reference system, Logistic model

1 引言

在一个国家的经济发展进程中，人口迁徙与人口城镇化一般是同步进行的。人口迁徙对于城镇化建设具有重要的影响。近年来，我国经济社会发展进入了新常态，城镇化进程作为重要经济增长点也进入深入发展的新时期，因此掌握人口迁徙的情况可以为我国当前推进国家新型城镇化建设提供部分决策支持[1]。然而传统的人口调查涉及大量的入户调查，这些调查当前正面临前所未有的挑战，成本高昂且无响应率极高。自从“计算社会科学”的概念[2]于2009年由包括美国哈佛大学教授拉泽尔[3]在内的15名顶级学者在《Science》杂志上正式提出后，大数据技术逐渐在社会科学领域发挥重要作用。基于大数据分析的人口迁徙研究正成为当前使用大数据参与社会研究和调查的热点。

自1978年改革开放以来，我国取得了巨大的成就：经济实现了持续快速增长，综合国力进一步增强；人民生活水平显著提高，总体上进入了小康水平[4]。同时，以青壮年为主的劳动力人口出于对美好生活的向往[5]，逐步开始大规模跨省迁徙。在迁徙过程中，不可避免地会与金融机构发生经济行为，如申领信用卡、贷款买房、缴纳公积金及社保等，根据个人征信系统接口规范，这些行为的记录数据都可以被中国人民银行征信中心依法采集。征信中心收集的信息记录及时地反映了信息主体发生业务的时间和地点，在确定人口迁徙轨迹方面具有其他数据源难以比拟的优势。本文通过对3.9亿个信息主体、48.8亿条征信记录进行分析挖掘，得出了改革开放以来全国劳动力人口迁徙的新特点，并预测了未来人口迁徙的趋势。

根据联合国《多种语言人口学辞典》的定义，人口迁徙是指“人口在两个地区之间的地理流动或者空间流动，这种流动通常会涉及永久性居住地的变更”。根据此定义可以判断描述人口迁徙的数据涉及永久性居住地的变更，在时间跨度上应该较长，在地点跨度方面至少跨越了行政界限。因此，在此次人口迁徙研究中，选取年作为时间跨度的基本单位，地点跨度上以跨省为判断标准。

2 数据描述和研究方法

2.1 迁徙数据生成原理

根据人口迁徙的定义，选取身份信息、初始地点、迁入地点、迁出地点、迁徙时间作为数据项，在统计维度上，选取了年龄、性别、区域作为研究维度。通过处理征信数据库内的个人身份数据，根据身份证号码生成规则可以获得其初始地点、年龄、性别等数据。根据个人信贷、社保、公积金等业务数据，可以获得个人在发生上述业务时的居住地信息。通过对比业务发生地点与初始地点，可以判断个人是否发生了迁徙行为。例如自然人A，初始居住地为山东，在2005年发生了一笔房贷业务，业务发生地点为上海，据此认为A 2005年的居住地点为上海，发生了从山东到上海的迁徙行为。原理如**图1**所示。

2.2 数据描述

本次研究不仅使用个人征信中的个人信贷、个人社保、个人公积金数据，还使用企业征信中的机构信用码、企业高管、企业贷款卡等数据。这些数据源基本反映了自然人的购房地点、工作地点或开设公司

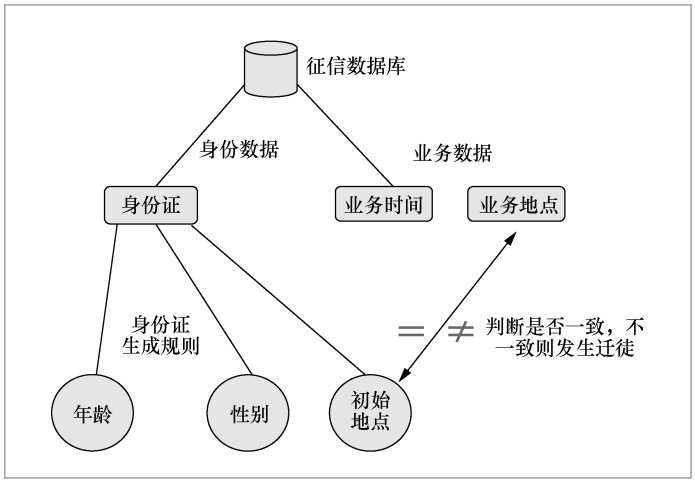


图1 迁徙数据生成原理

的地点，这些地点信息基本涉及了永久居住地址的变更，符合人口迁徙的定义。数据源情况见**表1**。

此次研究共涉及近50亿条业务数据，需要提取近200亿个数据字段，数据处理量大，数据预处理任务繁重。数据源方面，涉及信贷、公积金、社保、企业贷款等多种业务数据和不同的数据类型，呈现多样性的特点。根据征信中心数据采集规定，目前征信中心数据库每天需要接收相关机构数据近1亿条，数据更新速度快，数据增量大。数据质量方面，多数数据采集自金融机构、政府机关等，数据源可靠，数据质量和数据精度相对较高。综上所述，本次研究使用的数据呈现数据量大、数据源种类多、数据更新速度快和数据准确度高的大数据“4V”特点。

表1 人口迁徙数据源及数据量

数据源	数据量/亿条
个人公积金数据	30.7
个人信贷数据	12.2
个人社保及低保数据	5.0
机构信用码高管信息	0.5
企业高管及法人代表	0.3
企业贷款卡	0.1
总计	48.8

2.3 数据预处理

预处理主要包含两部分：筛选身份数据和数据去重。

本次研究涉及多个业务数据源，全部业务均包含自然人的身份数据，然而这些身份数据呈现多样化的特点。除了身份证号码外，还有护照号码、军官证号码、社保号码等信息。根据研究需求，必须确认这些身份信息所有人的初始地点、年龄等信息，然而此类信息在部分数据源中有大量的缺失。为了保证数据质量以及解决缺失值的问题，本次研究对身份信息进行了筛选，选取18位的身份证号码为唯一的身份数据的自然人，并根据身份证号码生成规则，判断出自然人的性别、年龄、迁徙初始地点等基本信息。这种处理方式保证了数据反映的自然人相关信息的质量，并解决了部分数据源缺失的问题。

数据去重方面，由于涉及近50亿条记录以及大量的历史数据，本次研究不可避免地在数据清洗中面临大量重复数据的去重处理问题。例如在公积金数据库中，自然人每个月均产生上缴公积金的业务数据，数据由公积金中心上报之后，一年中共产生12笔业务数据，由于本次研究的统计口径是年，因此会面临较多的重复数据。面对大量数据的去重问题，本次研究采用了在分区基础上建立子分区的优化方案，将去重范围缩小至子分区，提高去重效率。例如将公积金数据先按照年份进行分区，之后在此基础上建立地域即省份的子分区。数据迁移完成之后就可以在子分区的基础上去重。这种数据处理方法有效缩短了本次大批量重复数据的清洗时间。

2.4 数据加工及绘制人口迁徙路径

在完成数据预处理工作后，得到包含自然人ID、性别、年龄、初始地点、业务地点、业务发生时间的数据项。根据迁徙数据生成原理，得到当年迁徙地点数据之后，为了较为完整地反映每个自然人的迁徙过程，专门绘制了自然人迁徙路径数据。具体包括以下数据字段：自然人ID、性别、年龄、初始地点、年份（1978—2014年）。其中，自然人ID是用来判断自然人的数据标识，性别、年龄、初始地点均由身份证号码生成规则判断而来。需要注意的是在年份数据段中记录的是自然人该年所处的地点，例如1978年记录为河北，1979年记录为北京，一直记录到2014年。通过这种处理方法，最终得到自然人1978—2014年每年所在地的数据，绘制迁徙路径，可以完整地反映每个自然人的全部迁徙过程和轨迹，为发现自然人多次迁徙行为和回迁行为提供了数据支持。

3 研究结果及数据展示

为全面系统地阐述人口迁徙状况，以下从宏观阐述、热点省市迁徙情况和不同群体迁徙偏好3个方面来得出人口迁徙的相关结论。

3.1 人口迁徙概况

本次研究共涉及征信系统3.9亿个自然人的征信数据，监测到发生迁徙行为的共有1.2亿人，迁徙次数共计1.9亿人次。

从历年人口迁徙数量来看，监测到的每年迁徙人数从2005年开始逐年增长，2008年出现显著增长，2009年出现回落，之后快速回升，并逐年增长，如**图2**所示。

迁徙人群中男女比例约为6∶4，年龄结构方面以青壮年为主，“80后”成为迁徙主力。迁徙人群中男性人数为7 000万人，女性人数为5 000万人。人群中“80后”比例最大，为44.1%，“60后”、“70后”和“90后”所占比例依次为14.1%、28.5%和7.8%，如**图3**所示。

从各省人口迁徙状况来看，大多数省份处于人口净流出状态，人口净流入的省份主要集中在“北上广”等经济发达地区，截至2014年底，人口净流出最多的省份依次为河南、湖南、四川、湖北、安徽，详情如**图4**所示。

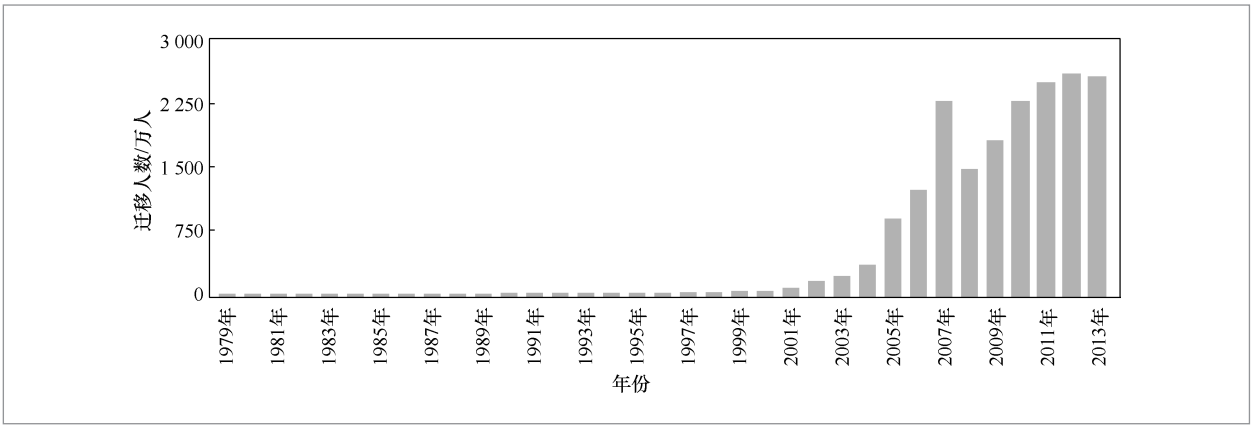


图 2 历年人口迁徙趋势

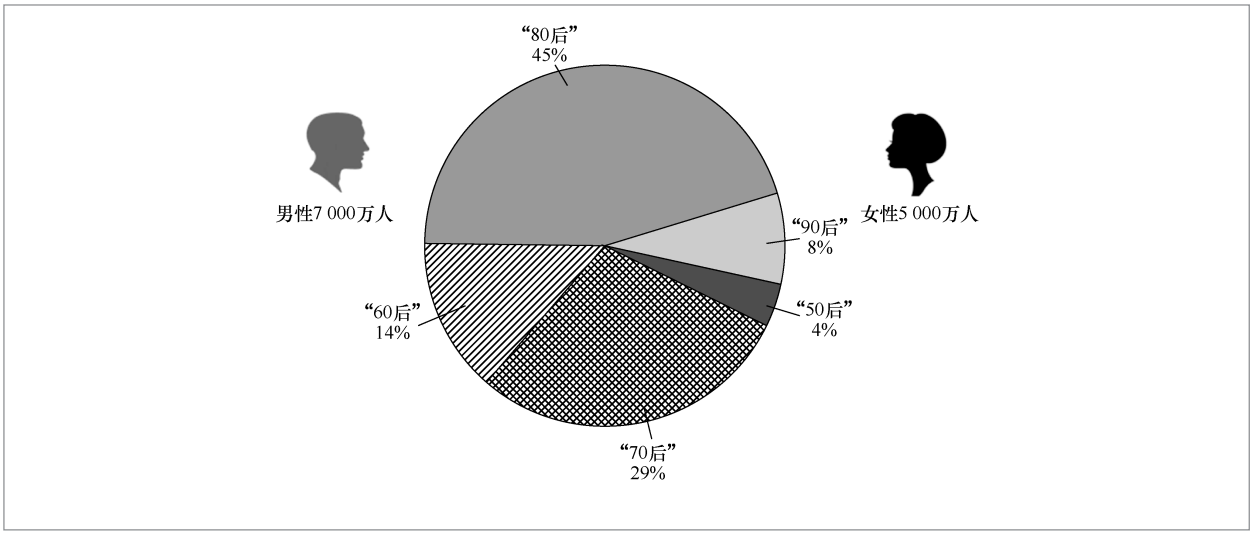


图 3 迁徙人口性别年龄概况

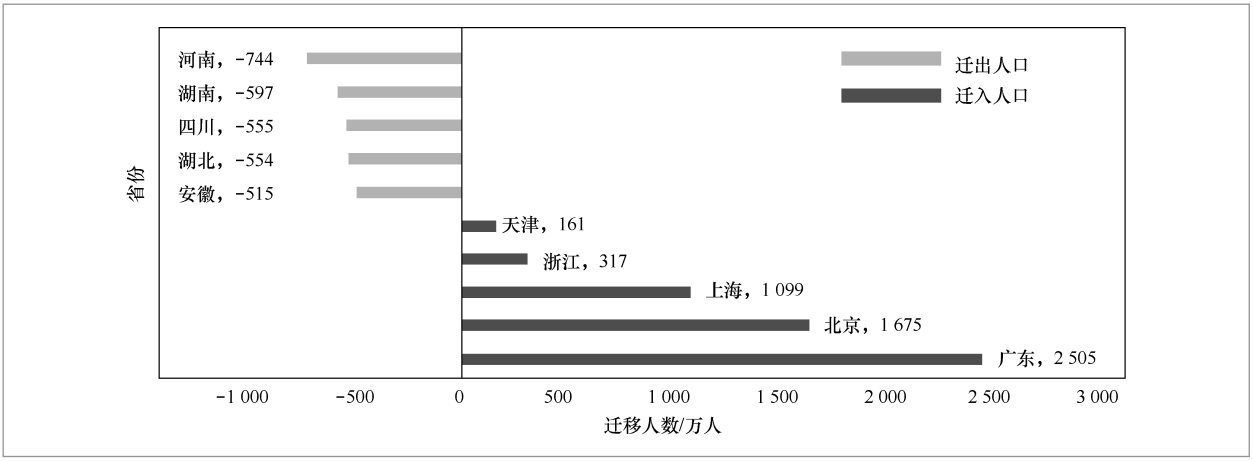


图 4 各省人口流失概况

3.2 热点省市迁徙状况

最近，特大城市人口控制和逃离“北上广”成为了社会和媒体关注的热点，因此针对北京、上海、广东的人口迁徙状况，展开了相关研究。

一线城市依然处于人口净流入状态，但逃离“北上广”的迹象已初露端倪。从一线城市历年人口迁徙数量来看，“北上广”自2001年以来一直处于人口净流入的状态，迁入人口在2008年增速最大，达到了1 300万人，在2009年经济危机期间明显减少。值得注意的是“北上广”的人口迁入规模在2011—2014年间，增速逐渐放缓。而与此同时，人口迁出规模逐年增长。按照此趋势，“北上广”将逐渐实现迁入人口和迁出人口的平衡，如**图5**所示。

从迁入“北上广”的人口构成来看，主要集中在热点区域的周边省市。自改革开放以来，人口净流入最多的3个省份是广东、北京和上海，分别为2 505万人、1 675万人、1 099万人，见**表2**。其中，北京的迁入人口

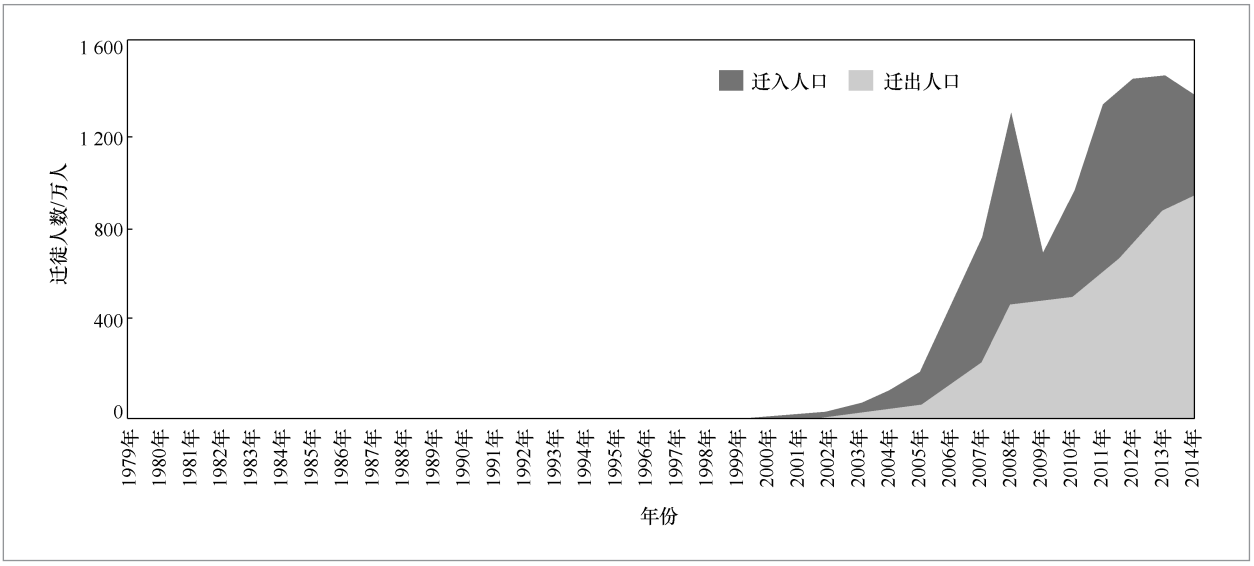


图 5 “北上广”历年人口迁徙情况

中，最多的是河北人，达到270万人，其次是河南人和山东人，分别是160万人和140万人。上海的迁入人口中，居前的是江苏人134万人、安徽人109万人和山东人86万人。广东也有类似的情况，都是以周边省市为主。

从“北上广”迁出人口的去向来看，其目的地依然以“北上广”为主。例如，上海的迁出人口中，去广东的人最多，达33万人，其次是21万人去了北京；同样广东的迁出人口中，去北京的人最多，达59万人，其次是51万人去了上海；北京的迁出人口情况类似。

长期以来，“北上广”等一线城市是人们的向往之地，特别是许多大学毕业生把工作生活在一线城市作为自己的首选，但在一线城市打拼数年后，受生活成本高和就业压力大的影响，不少人重新选择到二、三线城市发展，这个群体虽然还不算庞大，但和多年来奔向“北上广”的潮流形成鲜明对比。而这批人大多选择浙江、江苏、广西、湖南等南方省市作为新起点，见**表3**。

表 2 大省市迁入人口数量及比例

省市	人口数量/万人	所占比例
广东	2 505	42%
北京	1 675	28%
上海	1 099	18%
其他省份	696	12%
合计	5 975	100%

表 3 迁出“北上广”去往部分南方省市的人口数量

省市	人口数量/万人				
	浙江	江苏	广西	湖南	福建
北京	1	1.4	0.4	1.6	0.5
上海	3.3	5.2	0.3	0.3	0.7
广东	6	3.5	5.5	3.1	3.2
总计	10.4	10.2	6.2	5	4.4

3.3 人群迁徙偏好

本节主要分析迁徙人群的偏好，通过将迁徙人群细分，可以更加深入地了解不同人群发生迁徙行为的原因，更加全面地了解人口迁徙的状况。

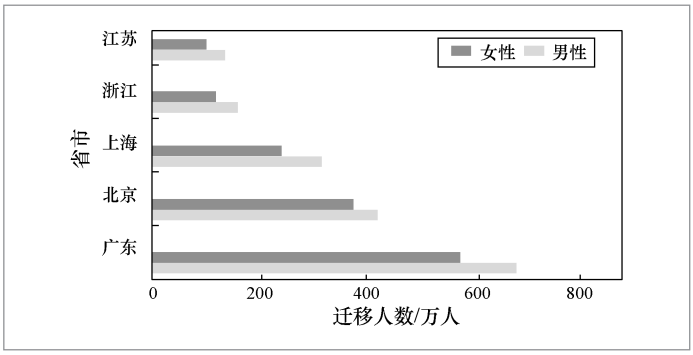


图 6 男女迁徙偏好

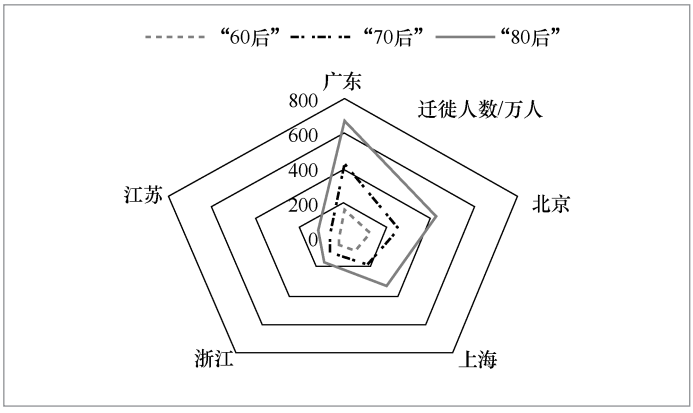


图 7 不同年龄人群迁徙偏好

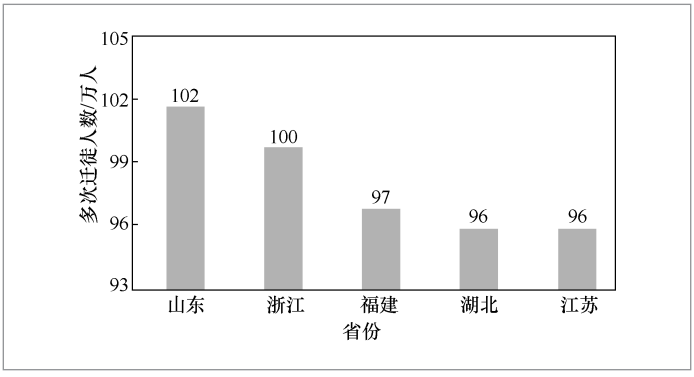


图 8 多次迁徙人群概况

迁徙人群选择迁徙城市的主要动因可能与经济有关，不存在明显的性别、年龄差异。如图6、图7所示，无论男女老少，迁徙人数最多的省份均是广东、北京、上海、浙江、江苏，并且各个省份的迁徙比例基本一致。这说明迁徙人群选择迁徙目的地不存在明显的性别和年龄差异，最影响迁徙的因素可能与经济发展前景相关。

经统计，山东、江苏、浙江和福建出现多次迁徙行为的人数最多，并且倾向于在迁徙之后重返故土。在本次研究中，分析了多次迁徙（迁徙次数大于或等于3次）和迁徙后返回初始地点的迁徙行为，这些行为主要集中在东部经济较为发达的江苏、山东、浙江、福建等省份，如图8、图9所示。

4 利用Logistic模型预测未来青壮年劳动力人口迁徙趋势

青壮年劳动力人口迁徙问题是影响中国发展的重要问题，预测未来人口迁徙的发展趋势对国家的宏观政策调整有重要的参考意义。分析历年人口迁徙趋势（如图2所示）发现3个特点：一是单调递增性；二是增长有限性；三是形状为S形，增长速度递减，逐渐趋于零增长。这3个特点完全符合Logistic模型的前提条件[6]，因此本文采用该模型进行拟合，并给出可靠的预测值。

4.1 Logistic模型原理

Logistic模型（阻滞增长模型）原理[7]考虑到自然资源、环境条件等因素对人口增长的阻滞作用，人口增长率r随着人口数量x的增加而下降。若将r表示为x的函数$r(x)$，为减函数，即：

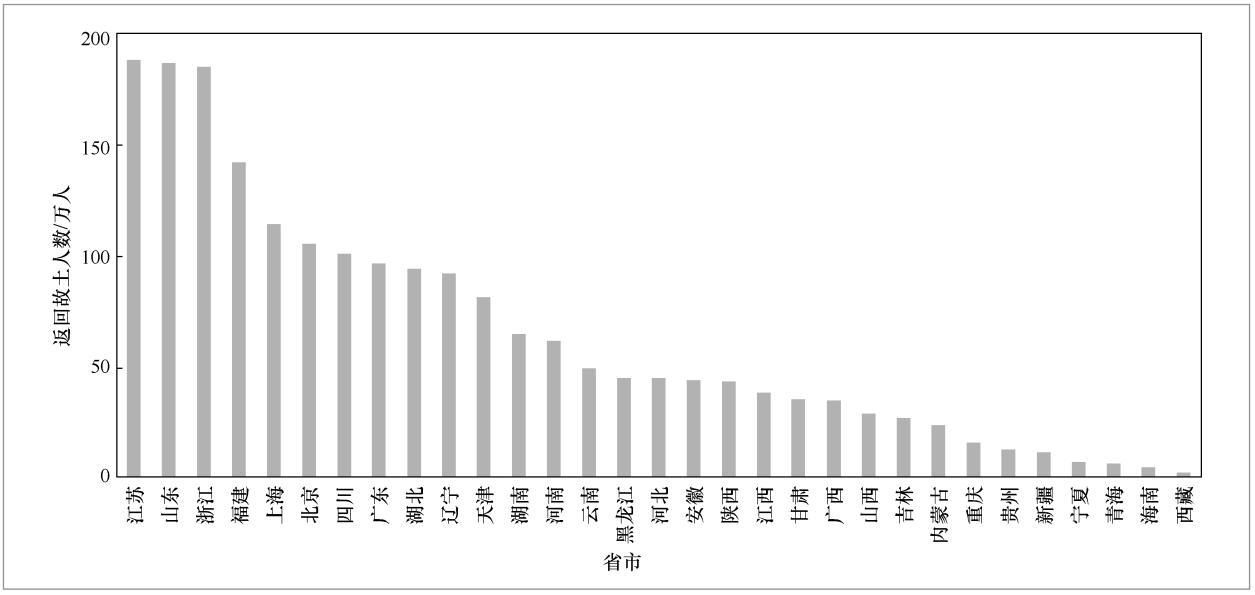


图 9 迁徙之后返回故土的人群概况

$$\begin{cases}\dfrac{dx}{dt}=x\cdot r(x)\\ x(0)=x_0\end{cases} \tag{1}$$

对$r(x)$的一个最简单的假设是：$r(x)$为x的线性函数，k为线性函数的斜率，因为这是减函数，所以$k>0$,且截距$r>0$，即：

$$r(x)=r-k\cdot x \tag{2}$$

设自然资源和环境条件所能容纳的最大人口数量为x_m，当$x=x_m$时，人口不再增长，即增长率$r(x_m)=0$，代入式(2)得$k=\dfrac{r}{x_m}$，于是由式(2)推出：

$$r(x)=r-\frac{r\cdot x}{x_m}=r\left(1-\frac{x}{x_m}\right) \tag{3}$$

把式(3)代入式(1)得：

$$\begin{cases}\dfrac{dx}{dt}=r\cdot x\cdot\left(1-\dfrac{x}{x_m}\right)\\ x(0)=x_0\end{cases} \tag{4}$$

经过推导，得到：

$$x=\frac{x_m}{1-\dfrac{x_0-x_m}{x_0}e^{-r\cdot t}}=\frac{x_m}{1+\left(\dfrac{x_m}{x_0}-1\right)e^{-r\cdot t}} \tag{5}$$

为了计算方便，可以将式(5)进行简化，得：

$$x=\frac{A}{1+m\cdot e^{n\cdot t}} \tag{6}$$

进一步简化，得：

$$x=\frac{A}{1+e^{B(C-t)}} \tag{7}$$

其中，A是该阶段人口数量的极限值，B是增长率，C是达到极大值数量之半的年份。

假设在时间点t_1、t_2、t_3的人口数量分别为x_1、x_2、x_3，系数的计算结果如下：

$$A=\frac{2x_1x_2x_3-x_2^2(x_1+x_3)}{x_1x_3-x_2^2}\ (\text{当}t_1-t_2=t_3-t_2\text{时}) \tag{8}$$

$$A=\frac{r_1\sqrt{x_2x_3}-r_2\sqrt{x_1x_2}}{r_1-r_2}\ (\text{当}t_2-t_1\neq t_3-t_2\text{时}) \tag{9}$$

$$B=\frac{1}{t_2-t_1}\ln\left(\frac{x_2(A-x_1)}{x_1(A-x_2)}\right) \tag{10}$$

$$C=\frac{1}{B}\ln\left(\frac{A-x_1}{x_1}\right)+t_1 \tag{11}$$

其中：

$$r_1=\frac{1}{t_2-t_1}\ln\left(\frac{x_2}{x_1}\right) \tag{12}$$

$$r_2=\frac{1}{t_3-t_2}\ln\left(\frac{x_3}{x_2}\right) \tag{13}$$

4.2 Logistic模型在劳动力人口迁徙中的应用

从1979—2014年，共有36个有效统

计值，t_1、t_2、t_3的取值范围分别为：t_1取值范围为[1979，2012]；t_2取值范围为[t_1+1，2013]；t_3=2014。

多次循环后，取最小值时的参数和残差平方，得t_1=2000、t_2=2007、t_3=2014，残差平方R^2=0.969 9。预测至2020年的结果如图10所示，见表4。

表 4 青壮年劳动力人口迁徙预测值

年份	迁徙人口预测值/人
2015年	26 541 553
2016年	26 639 890
2017年	26 693 626
2018年	26 722 914
2019年	26 738 853
2020年	26 747 522

表 5 优化后青壮年劳动力人口迁徙预测值

年份	迁徙人口预测值/人
2015年	26 819 310
2016年	27 103 773
2017年	27 279 000
2018年	27 386 226
2019年	27 451 574
2020年	27 491 302

4.3 去掉异常点后的预测效果

根据历年人口迁徙趋势（如图2所示）可以发现一个异常点：2008年人口迁徙有一个突增。主要原因是2008年我国首次举办奥运会，奥运经济吸引了劳动力人口大批迁徙。去掉该异常点后，重新对人口迁徙进行预测，经过多次循环后，取为最小值时的参数和残差平方，得t_1=1980、t_2=1997、t_3=2014，R^2=0.990 1。可见，去掉这个异常点后，预测效果更好，更能反映实际情况，预测至2020年的结果如图11所示，见表5。

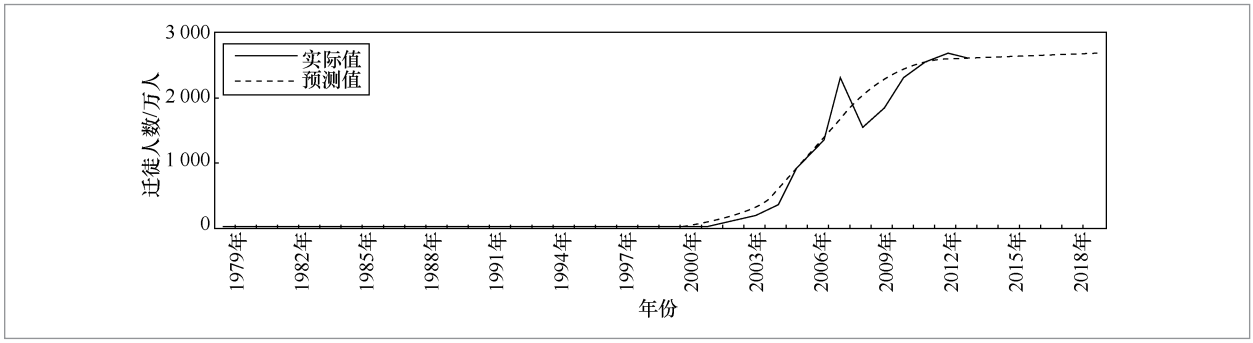


图 10 人口迁徙预测趋势

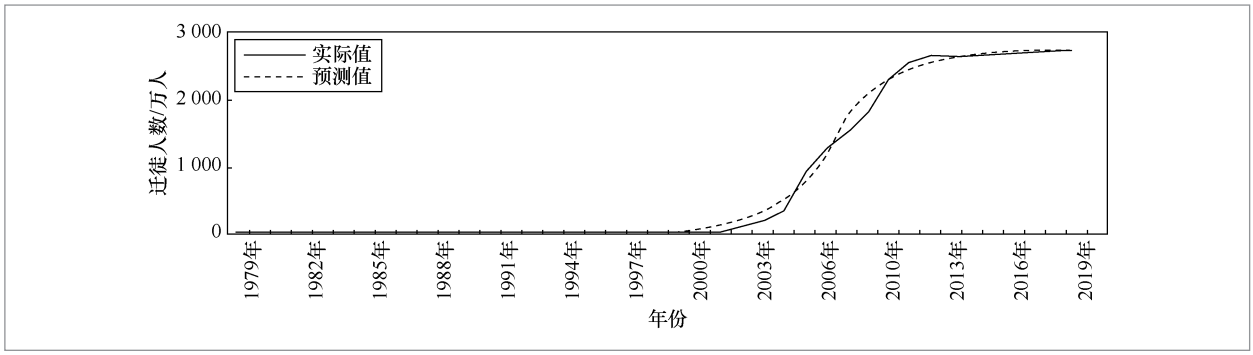


图 11 优化后人口迁徙趋势预测

5 结束语

通过对征信中心个人征信数据的进一步挖掘以及使用Logistic模型对迁徙人口数量进行预测，得出有关青壮年劳动人口迁徙特点的相关结论如下。

● 截至2020年，青壮年劳动力人口迁徙速度放缓，但每年迁徙人口仍在高位，总数达2 700万人。因此，国家应该堵疏结合，既要放开户籍管理，鼓励自由迁徙，促进资源和人口合理配置，又要促进全国经济均衡发展，合理引导劳动力往人口密低度的区域迁徙。

● 青壮年劳动力人口净流出的省份，劳动力急剧减少，高端人才大量流失，税收大幅度下滑，基础实施落后，发展缺乏后劲，更没有足够的财力改善民生，应该尽快制定优惠政策，跨越式发展，吸引劳动力就地安置，同时放开二胎政策，鼓励生育，产生劳动力后备军。

● 青壮年劳动力人口净流入的省份，劳动力人口日趋饱和，应该调整产业结构，提供更多就业机会，防止出现不稳定因素，同时制定相应政策，妥善解决迁入人口的子女入学、买房置业等实际问题。

致谢

中国人民银行征信中心数据部高健、邓林慧、李状君、徐方林、熊欣等同事对本研究工作给予了大量帮助，特此感谢。

参考文献:

[1] 王向明. 人口迁移和流动对人口城镇化进程的影响[J]. 人口与经济, 1988(2): 19-24.
WANG X M. The influence of migration and population flow on the process of urbanlization[J]. Population & Economics, 1988(2): 19-24.

[2] MEYER B D, MOK C, SULLIVAN J X. Household surveys in crisis[J]. Journal of Economic Perspectives, 2015, 29(4): 199-226.

[3] LAZER D, PENTLAN A, ADAMIC L, et al. Life in the network: the coming age of computational social science[J]. Science, 2009, 323(5915): 721-723.

[4] 邓振芳. 浅析中国改革开放以来取得的巨大成就及其对世界的影响[J]. 学理论, 2010(20): 6-7.
DENG Z F. The great achievement made by China since the reform and open up and its impact on the world [J]. Theory Learning, 2010(20): 6-7.

[5] 蔡建明, 王国霞, 杨振山. 我国人口迁移趋势及空间格局演变[J]. 人口研究, 2007, 31(5): 9-19.
CAI J M, WANG G X, YANG Z S. Future trends and spatial patterns of migration in China[J]. Population Research, 2007, 31(5): 9-19.

[6] 王学保, 蔡果兰. Logistic模型的参数估计及人口预测[J]. 北京工商大学学报, 2009, 27(6): 75-78.
WANG X B, CAI G L. Parameter evaluation of Logistic model and population prediction[J]. Journal of Beijing Technology and Business University , 2009, 27(6): 75-78.

[7] 黄润龙, 帅友良. 人口增长的Logistic模型及其应用研究[J]. 南京人口管理干部学院学报, 2000, 16(3): 25-27.
HUANG R L, SHUAI Y L. Applied research on Logistic model of population growth[J]. Journal of Nanjing College for Population Programme Management, 2000, 16(3): 25-27.

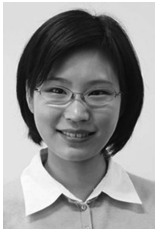
作者简介

姚前(1970-),男,中国人民银行征信中心副主任、高级工程师,主要研究方向为分布式系统和计算机安全。

谢华美(1976-),男,中国人民银行征信中心数据部副总经理,主要研究方向为数据挖掘。

司恩哲(1985-),男,就职于中国人民银行征信中心数据部,主要研究方向为数据挖掘。

景志刚(1977-),男,就职于中国人民银行征信中心数据部,主要研究方向为数据挖掘。



胡青青(1984-),女,就职于中国人民银行征信中心数据部,主要研究方向为数据挖掘。

收稿日期:2016-02-17