第一章　绪论

一、研究背景和意义

人口是可持续发展诸多因素中的关键性要素之一。人口要素是分析人口政策、劳动力供求、城镇化，以及健康、教育、交通、住房、社保等民生问题的基础依据，是重大基础设施建设、重要资源开发利用、生态建设和环境保护、公共事业发展及相关重要产业发展的重要基础，是经济社会的基础变量和动力源泉。大规模的人口迁移流动已经成为当前我国人口发展的常态化特征，对经济、社会、环境、资源产生重要影响。21 世纪，美国高科技产业和中国城镇化是影响世界最大的两个事件。随着我国社会形态由传统农业社会向现代工业社会转变，基本制度框架由城乡二元结构向城乡一体化转变，城镇化将伴随我国现代化的全过程，人口城镇化将成为我国基本国情。

随着工业化、城镇化和社会转型的不断加快，人口流动迁移将进一步深入发展，我国正经历着人类历史上最大规模的人口迁移流动。回顾我国流动人口发展历程，可以发现，人口大规模流动迁移是在改革开放背景下产生的最显著的人口现象。1958 年，我国第一部户籍制度《中华人民共和国户口登记条例》颁布，确立了一套严格的户口管理制度，人们不能随意流动和迁移。20 世纪 50 年代后期至 80 年代初期，由于实行严格的计划经济管理，加上严格的户口管理制度，全国流动人口数量很少。1982 年第三次全国人口普查数据显示，流动人口数量为 657 万人，占全国总人口的 0.66%。80 年代中期以后，由于国家相关政策的调整，流动人口经历了一个迅速增长的过程。1984 年 10 月，国务院出台了《关于农民进入集镇落户问题的通知》，标志着国家在一定程度上放松了对农业人口进入中小城镇的控制，并由此带来对整个人口流动控制的松动。1987 年全国流动人口猛增到 1810 万。1988 年出现了“百万民工下广东”的民工潮现象。1992 年邓小平南巡讲话后，人口流动目的地逐渐突破小城镇而大量涌入城市，流动人口的发展进入新的高峰期。[①] 据 1990 年第四次全国人口普查，流动人口达到 2135 万，占全国总人口的

① 段成荣等．中国流动人口研究．北京：中国人口出版社，2012

1.89%。1995 年流动人口达到7073 万,占5.86%。2000 年超过1 亿。进入21 世纪,流动人口继续保持快速增长的势头,2005 年为1.47 亿人。2010 年第六次全国人口普查数据表明,流动人口达到了2.21 亿。国家统计局最新公布数据显示,2014 年全国流动人口总量达2.53 亿,占总人口的18.5%,相当于不到6 个中国人中就有1 人在流动,规模庞大的流动人口为国家发展做出了特殊贡献。据专家测算,改革开放以来到2012 年左右,由于人口流动与生产要素等结合,流动人口因素对全国 GDP 的贡献率在23%左右,人口流动对 GDP 贡献显著。①

但是,目前缺乏对人口迁移流动的产生原因以及影响因素的定量分析,缺乏对大规模人口迁移流动对各区域经济社会影响的分析,不利于引导人口合理流动,不利于合理配置公共资源,不利于保障流动人口合法权益。而且,人口的自由流动使得区域人口已经由封闭人口转变为开放人口,影响区域人口规模变动也由人口自然变动转变为机械迁移变动。流动人口是区域人口变动的重要因素,因此探索人口迁移流动变动趋势、流向分布以及迁移流动影响因素,构建人口迁移流动预测和分析模型平台,以便对人口迁移流动进行深入分析和科学预判,成为经济社会发展规划亟待解决的问题。

二、研究目标和研究过程

针对我国人口净增数量快速下降、年龄结构进入快速变化时期、人口预期寿命不断提高、老年人口数量快速增长、人口流动规模巨大且呈继续扩大之势、人口受教育程度显著提高等状况,以及这些状况对经济社会环境资源的影响,中国人口与发展研究中心开展了科技部"十二五"重点科技支撑计划项目"人口与发展数学模型与综合决策支持系统"研究。"人口与发展数学模型与综合决策支持系统"项目分为四个一级子课题,分别是:子课题一"人口政策模型建构研究与信息整合开发";子课题二"多状态人口分析预测关键技术与模型研究";子课题三"多区域人口预测关键技术研究";子课题四"人口与发展综合决策系统集成与应用示范"。本课题为该项目的第三子课题。根据项目设计,本子课题的总体研究目标为:应用多区域人口预测技术,探索我国人口迁移流动的规律和趋势,发展人口迁移的建模技术,在国内实现人口多区域迁移流动且动态平衡的预测方案和软件系统。具体目标为:汇集国内外先进的多区域人口分析和预测的方法与模型,结合我国实际情况,构建中国人口迁移流动多维(数量、素质、结构、经济社会状态、产业结

① 段成荣等. 中国流动人口研究. 北京:中国人口出版社,2012

构、时间等维度)“推力—拉力”模型,基于 Agent 思想的多区域人口迁移流动模型,建立多区域人口迁移流动多维矩阵,支持预测全国、各省市人口变动趋势,分析人口迁移流动的机制及其对经济社会发展的影响,为国家人口和发展的重大决策提供支持。

在对国内外理论、模型综合分析和研究的基础上,课题组确定了重点分析和研究解释型模型,如“推力—拉力”模型、流动人口多维动因模型、重力模型、基于 Agent 思想的多区域人口迁移流动模型等;确定了利用时间序列模型对流动人口规模进行复合预测。对模型的研究重点放在模型应用上。结合我国流动人口的特点和规律,研究模型的假设前提、参数、适用性等问题。

1. 人口迁移流动理论和模型研究

系统研究国内外区域人口分布和人口分析预测技术相关文献,开展人口迁移流动理论和模型研究,如人口迁移的重力模型、多区域人口预测数学模型、基于 Agent 的多区域人口迁移等综述研究。

2. 收集和研究人口迁移流动模型

从人口迁移流动宏观模型、人口迁移流动微观模型、数学模型、地理模型、区域人口承载力等领域收集模型并进行分析研究。

3. 数据收集和分析

收集全国 31 个省份的 1990 年、2000 年、2010 年三次全国人口普查省际迁移矩阵数据;1987 年、1995 年、2005 年三次全国 1% 人口抽样调查省际迁移矩阵数据。并对数据进行标准化处理,包括抽样比的加权与统计口径的统一。分省份、分年份收集影响人口迁移流动的社会经济指标,主要包括常住人口数、户籍人口数、农业人口数、非农业人口数、城镇化率、国民(内)生产总值、分三次产业国民(内)生产总值、分三次产业从业人员数、城镇登记失业人口数、城镇登记失业率、城镇居民家庭年人均可支配收入、农村居民家庭年人均纯收入、城镇居民家庭平均每人年现金消费支出、农村居民家庭平均每人现金消费支出、平均受教育年限等,计算获得各省会城市之间距离矩阵数、省际相邻性指数;分析数据间的相关性、多重共线性;搜集各地政府部门的城镇化规划数据,作为区域预测的先验信息。

4. 确立人口迁移流动解释和预测模型,并进行试算、优化

通过系统梳理和研究,确立重点研究领域:人口迁移时间序列模型、迁移流动多维动因分析模型、基于 Agent 思想的多区域人口迁移流动模型、联合国城乡人口增长差异模型(URGD 模型)以及改正的城乡人口增长差异模型。

5. 模型平台建设

在收集到的105个人口迁移流动相关的模型基础上，筛选出40余个模型进行开发，构建以时间序列模型、流动人口URGD模型等多区域人口迁移流动预测模型和推力—拉力模型、流动人口多维动因模型、基于Agent思想的多区域人口迁移流动模型等人口迁移流动解释模型为主体，以人口迁移流动数学模型等拓展模型为辅助的综合模型平台。

三、主要内容

本书是“多区域人口迁移流动分析和预测模型”子课题成员集体完成的主要成果。通过两年多的研究和系统开发，构建了人口迁移流动的多维动因模型、时间序列模型、城乡差别增长（URGD）模型，以及基于Agent思想的迁移流动情景分析模型，并建成可视化平台，为中国人口迁移流动预测和分析提供基础性工具。

1. 人口迁移流动时间序列预测模型

时间序列是指将某指标按时间先后顺序排列而形成的序列，时间序列方法基于随机过程理论和数理统计学，利用过去数据资料对未来的发展趋势和水平进行预测以用于解决实际问题。受数据的限制，人口迁移时间序列模型采用一次指数平滑模型、二次指数平滑模型（布朗单一参数线性指数平滑模型）和三次指数平滑模型（布朗三次指数平滑模型）作为预测中备选的模型，并以人口普查、小普查中的按现住地和五年前常住地省内省际迁移人口的数据进行案例分析，首先选取模型，将三个模型的算法代入；选取2010年作为预测检验年，之前年份的数据作为试验数据；选取测试准则，残差绝对值最小、预测精度最高、5年内预测残差最小；输出结果。结果显示，当参数选择$a=0.5$时，二次指数平滑模型的相对误差最小；三次指数平滑的平均绝对误差最小。经过案例应用，可认为选取的三种时间序列模型能够拟合并预测出相应的流入或流出人口数。

2. 人口迁移流动多维动因分析模型

依次构建了三个模型。（1）初步模型—人口迁移重力模型。在美国社会学家Zipf重力模型基础上构建中国人口迁移重力实证模型，研究的解释程度为46.5%。（2）基础模型—多维动因模型。在M. P. Todaro的城乡人口迁移模型和Lowry动因模型基础上构建中国人口迁移多维动因实证模型，通过对迁入地和迁出地的失业率、城镇家庭人均可支配收入、农村居民人均纯收入、三次产业就业比例、三次产业产值比例、GDP、人口数、城镇化率以及省会城市距离、政策等18个变

量对人口迁移量的影响试算及共线性进行分析，筛选出迁入地和迁出地的GDP、人口数以及距离等5变量构建中国人口迁移多维动因模型，总模型的解释程度达66.4%。从分时代看，基于2010年数据基础上的模型解释程度达74.6%；从分省份看，流入省份为北京、广东、贵州、浙江等地的模型解释程度超过80%，流出省份为广西、安徽、广东等地的模型解释程度超过80%。(3)深化模型—推拉模型。在艾维李的推拉理论模型基础上，通过熵值理论将理论模型转化为实证模型，对中国人口迁移流动数据进行了分析。模型的解释程度为60.6%。

3. 基于Agent思想的多区域人口迁移流动模型

该模型主要借鉴智能体建模思想，在省级层面上的每个迁移周期，迁出人口根据其他各省区对迁出地的吸引力大小，选择最终要迁入的省区。吸引力方程采用他人有关研究成果，模型因子包括迁出地和迁入地的人口规模、迁入地城镇人均可支配收入、迁出地农村人均纯收入、迁出地和迁入地之间的距离，以及迁出地和迁入地之间的相邻性指数。以2000年“五普”数据和2010年“六普”数据代入模型运算，并与他人的计算结果进行比对，结论基本一致，证明该模型是稳定的、可靠的，分别能够解释71%和77%的省际迁移人口变动。

4. 联合国城乡人口增长差异模型(URGD模型)

联合国法是联合国用来预测世界各国城镇化水平时常用的一种方法。它的关键是根据已知的两个代表年份的城镇人口和乡村人口，求取城乡人口增长率差。假设城乡人口增长率在预测期保持不变，则向外推可求得预测期末的城镇人口比重，向内推则可以估测代表年份之间各年的城镇人口比重。使用该软件向前可以估计出两次人口普查年之间每一次的城镇化水平及城镇人口比重，向后可以预测某年的城镇化水平及城镇人口比重。在一个国家或地区人口迁移流动相对较弱的情况，联合国城乡增长差异模型可以较好地预测封闭人口的城镇化水平。本研究把这种情况称为模型Ⅰ。如果要对一个国家内有着频繁人口迁移流动的多个区域进行分区域的城镇化水平评估或预测，则需要考虑迁移流动人口对城镇化率的影响。采用两次城乡增长差异模型进行处理，第一次进行区域人口占全国人口比重预测，考虑迁移流动可能造成的影响。第一次使用城乡增长差异模型时将预测区域选作城镇或乡村，国内其他区域相应作为乡村或城镇，根据两次普查时点人口进行一次区域人口预测；第二次进行区域内城镇化预测。两次使用联合国城乡增长差异模型，本研究把这种情况称为模型Ⅱ。模型Ⅰ广泛应用于国内外一个国家不同时期的城镇化率评估和预测。模型Ⅱ用于预测我国分省城镇化率，与多数省份的规划目标接近。

通过多区域人口迁移流动模型的研究和开发，首次构建了多层次的中国人口

迁移流动预测和解释模型体系。基于随机过程基础上的人口迁移流动时间序列模型能够对短期区域人口迁移流动进行较为准确的预测;多维动因模型能够量化分析社会、经济、政策等变量对人口迁移流动的影响作用;应用熵值理论综合评价分析方法,把人口迁移流动推力—拉力理论模型转化为实证模型,对影响中国人口迁移的因素进行了综合分析处理;借鉴 Agent 智能体建模思想,构建省际人口迁移流动情景分析模型;改进联合国城乡人口差别增长(URGD)模型,评估和预测多区域城乡人口增长情况。

第二章　时间序列预测模型

一、文献综述

时间序列模型是一种利用过去数据资料对未来的发展趋势和水平进行预测的模型。时间序列分析常用在国民经济宏观控制、区域综合发展规划、天气预测等多个重要领域的预测中。近年来，随着国民经济的快速发展，大规模的人口迁移流动成为当前我国的一种基本国情，对流动人口的流量和流向方面分析研究和预测的需求日益突出。在此背景下，利用历史数据开展对未来的预测成为必然选择，时间序列模型在预测方面的优势使其成为此类研究不可或缺的选项。在多区域研究中，时间序列模型作为重要的研究方向，为完成多区域人口迁移流动预测提供了重要的实现手段。

时间序列分析是概率统计学科下的一个应用较强的分支学科，它在金融、市场、电子商务、天气预报、地质水文、数据挖掘等众多领域有着广泛的应用。传统的时间序列模型包括移动平滑模型、指数平滑模型、趋势外推模型、季节变动预测模型、AR 模型、MA 模型、ARMA 模型、ARIMA 模型等。1927 年数学家耶尔(Yule)为了预测市场变化的规律，提出了自回归概念，这标志着时间序列分析方法的诞生。1931 年瓦尔格(Walker)在自回归模型的基础上，建立了移动平均模型和自回归移动平均模型。20 世纪 40 年代，Norbort Wiener 和 Andei Kolmogonor 等人对时间序列的参数拟合及推断过程做出了重要推动，促进了该方法在工程领域上的应用。

20 世纪 70 年代，G. P. Box 和 G. M. Jenkins 发表了专著《时间序列分析预测和控制》，对平稳时间序列数据提出了 ARMA 模型以及一整套建模、估计、检验和控制的方法，该项工作奠定了时间序列分析在各领域应用的基础。此后不同领域的学者们开始不断完善时间序列分析的理论并拓展时间序列分析应用的新领域。1982 年 ARCH 模型建立，时间序列模型获得了快速发展，并在众多应用领域获得了显著成功。从时间序列分析法产生直至 20 世纪 70 年代末，这期间几乎所有的时间序列模型都是线性的。直到 20 世纪 70 年代后期，人们越来越清楚地看到线

性的时间序列分析存在诸多的局限性,这便要求学者们在原来线性的基础上有所突破,以"非线性"的眼光来对待时间序列。汤家豪对非线性时间序列分析做出了开创性工作,他将有关非线性时间序列分析的研究与动力系统科学的模型连接,并于20世纪70年代末提出了门限自回归模型,该模型在20世纪80年代初获得了系统性的发展,至今仍然获得了广泛的应用。Litterman、Sargent和Sim等人在20世纪80年代初提出了向量自回归模型,用于替代联立方程结构模型,该模型的提出提高了经济预测的准确性。向量自回归模型是自回归模型的拓展,该模型中,在同一样本期间内的多个内生变量可作为它们过去值的线性函数。

恩格尔(Engle)于1982年提出了自回归条件异方差模型,该模型假设误差项的无条件方差为常数,条件方差随时间变化而波动。该假设创新性地将误差项的异方差模型化。在建立了ARCH模型后,恩格尔还提出了一种判断误差项的条件方差是否为常数的实用检验方法。Bollerslev于1986年提出了广义自回归条件异方差模型,该模型需要的滞后阶数较小,具有与ARMA模型类似的结构。均值广义自回归条件异方差模型是GARCH模型的一个拓展,该模型多用于描述预期风险与收益密切相关的金融资产分析。分整型广义自回归条件异方差模型在金融资产领域取得了较多应用,很好地反映了金融资产的异方差性和长期记忆的变动性。自从GARCH模型发展起来以后,GARCH-M模型、FIGARCH模型、TARCH模型、GJR-GARCH模型、SW-ARCH模型、FIEGARCH模型、SW-GARCH模型等十余项模型都属于GARCH模型的拓展。

Grange与Joyeux等学者把分数维差分噪声模型与ARMA模型结合后,创建了分整自回归移动平均模型(ARF-IMA)。ARFIMA模型克服了ARMA模型只能描述较短时间序列的缺陷,扩展了长期时间序列的记忆性。Granger和Lee于1990年将ARFIMA模型拓展为VARFIMA。近年来,ARFIMA模型又发展到了与ARCH族模型相结合,形成了ARFIMA-GARCH、ARFIMA-ARCH、ARFIMA-FIGARCH、ARFIMA-TARCH及ARFIMA-EGARCH等一系列联合预测模型。VARFIMA模型在参数估计方面遇到的问题较为容易解决,但是随着变量维数的增加,也面临着有效降维的问题。徐正国等学者提出了FIVAR模型代替VARFIMA模型,解决参数难以估计的问题;郭名媛、张世英提出了正交ARFIMA模型,对协方差矩阵的建模问题进行了研究。①

安鸿志、朱力行、陈敏研究了条件方差为非常数的回归和自回归模型的平稳性、遍历性和检验方法,首次给出了完全对立的假设检验方法,关于回归或自回归

① 张美英,何杰. 时间序列预测模型研究综述[J]. 数学的实践与认识,2011,41(18):189~195

的非线性检验问题，具有重要的意义。在非线性自回归模型的平稳性、遍历性和高阶矩等方面取得了系列成果，并获得了这些性质的最弱条件。

姚琦伟提出了一般随机系统对初始条件敏感性的度量及估计方法。在高维模型领域，提出用复系数线性模型近似高维非线性回归函数的新方法，以克服高维模型中样本量较少的问题。在时间序列模型的最大似然估计方法的研究中，他建立了金融风险管理中经常应用的 ARCH 和 GARCH 模型为最大似然估计的极限理论。提出了基于 boostrap 的新的估计方法以及稳健统计方法，解决了重尾部分布模型建模中的问题。此外，他还首次建立了在空间域上空间 ARMA 过程的最大似然估计理论，并对时间序列的最大似然估计理论首次给出了一个完整的时域上的证明。

韩明涛针对大型数据库的海量数据分析，提出一种进行时间序列模式挖掘的算法，为用户的决策支持和趋势预测提供依据，该算法适用于挖掘超过用户给定支持度和置信度的时间序列。吴煲宁等提出了一种基于模糊集合的数据挖掘时间序列模式算法，该算法可满足商业销售的智能分析需求。翁颖钧、朱仲英通过计算时序数据之间的最短弯曲路径来获得序列的匹配，进而提出了基于动态时间弯曲技术的相似搜索算法。该算法有很高的精度和对振幅差异、噪声和线性漂移有强的解释性，通过基于不同距离测度的聚类分析对比，结果表明该算法具有良好的应用价值。吕安民等利用分形理论中的 R/S 分析，研究了某些时间序列所具有的分形特征，发现了具有分形特征的时间序列模式的方法，进而预测时间序列未来的发展趋势。

江东等利用气象卫星 NOAA AVHRR 资料，分析了 NDVI 时间曲线的波动与农作物生长发育阶段及农作物长势的响应规律。李本纲等以地理信息系统为技术支撑，以数字图像处理和标准主成分分析为核心，开发了一种处理多年气象观测数据的新方法，该方法适用于处理空间分布广、时间序列长的多类型气象观测数据。李宏等将时间序列模式既用于具有时间关系的购买行为的分析，以揭示购买行为后面一种序列关系信息，又用于其他有时间关联的事件分析。国家重点基础研究发展规划项目《我国生存环境演变和北方干旱化趋势预测》，利用非线性时间序列分析的数学方法对我国多时空尺度历史和现代环境资料进行定量处理，应用处理非线性时间序列的数学方法对各种资料作了信号分析处理，揭示了环境要素中存在的年代际和世纪时间尺度的变化特征及其在空间上的信号传播规律。该分析对可能的物理机制为极端事件和环境预测提供了一定的依据。

陈平首次将时间—频率分析引入经济学，在新兴交叉学科——复杂系统科学和非线性经济动力学的研究中居于世界前沿，该研究从美国股票价格指数中发现

经济色混沌现象,并将经济诊断方法引入经济周期的分析预测。王卫宁等以2002年上证指数的高频数据,重构了2002年上证指数时间序列的奇怪吸引子,计算其关联维数,分析了价格波动的非线性特征,并求出其Lyapunov指数为正,确认了上证指数时间序列的混沌行为。张屹山等采用时间序列的谱分析方法,对我国工业生产、投资、消费、外贸、物价、财政、金融等主要月度经济指标的增长率周期波动进行了测定和分析,认为20世纪80年代以来,我国向市场经济体制转轨过程中出现的周期波动与以往存在不同的新特征,即产生了7~9年为主的中周期波动。此外,还存在一个2~3年的作用相对较弱的短周期波动。潘文卿等利用中国1978~2001年28个省区的数据,采用面板数据的模型方法,对中国改革开放以来的资本配置效率及其与中国金融发展的相关性进行了时间序列分析与横截面数据分析。发现随着改革的深入,资本配置效率总体上呈现上升趋势,但波动性很大,且资本配置效率呈东、中、西梯度递减特征。认为国有银行的信贷行为抑制了资本配置效率的提高,而非国有银行金融机构的信贷与投资行为对资本配置效率有促进作用。钱争鸣对金融市场不确定性的探讨和实证分析,分析了我国金融市场的有效性,测度金融市场的系统风险,该分析寻求最优动态无风险策略,协助政府制定和完善金融政策。

邹小平等分析了数字散斑时间序列相关方法中由于参考平面沿Z轴方向平移而引起的散斑场的平移和缩放现象,在计算机模拟后,获得了决定参考平面间距选取的因素。该结果有助于确定数字散斑时间序列相关系统的设计参数,以及系统校准过程中参考平面间距的确定原则,为系统设计提供了理论依据和技术方案。孙枫等利用混沌序列的遍历性,对混沌序列的遍历性和置换网络的时间复杂度做了分析,并给出了一种分组密码置换网络的设计。模拟结果显示,混沌分组密码置换网络具有复杂性高、抗破译性强的优点,增强了信息系统的安全性。金友渔研究了多变量时间序列的弱信号进行高精度复原的分解模型和算法,运用该算法经过数字仿真后发现,该分解模型和算法对Logistic混沌或非混沌序列弱信号具有高精度的复原能力。邓自立等提出了一种正向和反向两种固定区间稳态Kalman平滑新算法,并给出了保证算法最优性的最优初值公式。

方兆本教授领导的课题组在公共卫生领域应用了时间序列分析。该分析基于全球公开发表的有关SARS确诊病例、疑似病例和死亡病例的实际数据,建立起有关流行病学的空间统计模型,并根据空间流行病学规律获得了可靠的统计预测结果。

周春光等利用商务流通中的经济时间序列数据,设计开发了一套经济时间序列预测系统。在该系统中采用了多种预测模型,该系统包括了指数平滑算法、AR

算法、Holt－Winter 算法、回归分析算法等多种统计学算法，此外，该系统还包含神经网络算法，通过采用动态学习的 BP 神经网络进行训练预测，取得了较好的实用效果。①

二、时间序列预测模型方法及原理

（一）一次指数平滑

由于加权移动平均法所需权数较难确定，因此指数平滑法对此加以改进。一次指数平滑也称作简单指数平滑，给出时间序列 N 期的资料：

$$Y_1, Y_2, Y_3, Y_4, \cdots, Y_T, Y_{T+1}, Y_{T+2}, \cdots, Y_N$$

在指定预测移动步长 T 之后，预测值 S 算法为：

表 2－1　指定预测移动步长 T 后的预测值 S 算法

时间	预测值计算公式
$T+1$	$S_{T+1} = \alpha Y_T + (1-\alpha) S_T$
$T+2$	$S_{T+2} = \alpha Y_{T+1} + (1-\alpha) S_{T+1}$
$T+3$	$S_{T+3} = \alpha Y_{T+2} + (1-\alpha) S_{T+2}$
…	…

其中：$0 < \alpha < 1$，S_T 是 T 时刻的一次指数平滑值。

一次指数平滑模型输出相应的计算数据，模型结果展示如图 2－1。

α= □0.05 ☑0.1 ☑0.3 □0.5 □0.7 自定义

t	初始值	t	实际值	一次平滑a1	一次平滑a2
1987	86.94			a=0.1	a=0.3
1988	103.3	1987	86.94		
1989	104.88	1988	103.3		
1990	98.18	1989	104.88	88.58	91.85
1991	96.4	1990	98.18	90.21	95.76
1992	112.65	1991	96.4	91	96.48
1993	114.22	1992	112.65	91.54	96.46
1994	113.44	1993	114.22	93.65	101.32
1995	109.94	1994	113.44	95.71	105.19
1996	117.66	1995	109.94	97.48	107.66
1997	130.24	1996	117.66	98.73	108.35
1998	128.27	1997	130.24	100.62	111.14
		1998	128.27	103.58	116.87
		1999		106.05	120.29

图 2－1　一次指数平滑模型结果展示

① 李锐，向书坚．我国时间序列分析研究工作综述[J]．统计教育，2006(7)：6～8

(二)二次指数平滑

双重指数平滑,是对一次指数平滑值再进行一次平滑,当时序有趋势存在时,一次和二次指数平滑都落后于实际值,布朗单一参数线性指数平滑较好地解决了这一问题。

由于加权移动平均法所需权数较难确定,因此指数平滑法对此加以改进。给出时间序列 N 期的资料:

$$Y_1, Y_2, Y_3, Y_4, \cdots, Y_t, \cdots, Y_N$$

指定超前期数 m 和权数 α 后,t 为任意时刻,则 $t+m$ 期的预测值为:

$$F_{t+m} = \alpha_t + b_t m$$

其中:

$$\alpha_t = 2S_t^{(1)} - S_t^{(2)}$$

$$b_t = \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)})$$

式中 $S_t^{(1)}$ 为 t 时刻的一次指数平滑值,$S_t^{(2)}$ 为 t 时刻的二次指数平滑值:

$$S_t^{(1)} = \alpha Yt + (1 - \alpha) S_{t-1}^{(1)}$$

$$S_t^{(2)} = \alpha S_t^{(1)} + (1 - \alpha)$$

当 $t=1$ 时,通常采用

$$S_0^{(1)} = S_0^{(2)} = Y_1$$

二次指数平滑模型输出相应的计算数据,模型结果展示如图 2-2。

α= □0.05 ☑0.1 ☑0.3 □0.5 □0.7 自定义 [ ] m= [3]

t	初始值	t	实际值	一次平滑a1	一次平滑a2	二次平滑a1	二次平滑a2
1987	86.94			a=0.1	a=0.3	a=0.1	a=0.3
1988	103.3	1987	86.94				
1989	104.88	1988	103.3				
1990	98.18	1989	104.88	88.58	91.85		
1991	96.4	1990	98.18	90.21	95.76		
1992	112.65	1991	96.4	91	96.48		
1993	114.22	1992	112.65	91.54	96.46	90.54	99.7
1994	113.44	1993	114.22	93.65	101.32	93.93	107.51
1995	109.94	1994	113.44	95.71	105.19	95.31	105.87
1996	117.66	1995	109.94	97.48	107.66	96.07	102.99
1997	130.24	1996	117.66	98.73	108.35	100.26	113.66
1998	128.27	1997	130.24	100.62	111.14	104.12	120.02
		1998	128.27	103.58	116.87	107.18	122.01
		1999		106.05	120.29	108.95	119.48

图 2-2 二次指数平滑模型结果展示

(三)三次指数平滑

对二次平滑值再进行一次平滑,并用以估计二次多项式参数的一种方

法。[①] 布朗三次指数平滑是对二次平滑值再进行一次平滑。给出时间序列 N 期的资料：

$$Y_1, Y_2, Y_3, Y_4, \cdots, Y_t, \cdots, Y_N$$

指定超前期数 m 和权数 α 后，t 为任意时刻，则 $t+m$ 期的预测值为：

$$F_{t+m} = a_t + b_t m + \frac{1}{2} c_t m^2$$

其中：

$$a_t = 3S_t^{(1)} - 3S_t^{(2)} + S_t^{(3)}$$

$$b_t = \frac{\alpha}{2(1-\alpha)^2}[(6-5\alpha)S_t^{(1)} - (10-8\alpha)S_t^{(2)} + (4-3\alpha)S_t^{(3)}]$$

$$c_t = \frac{\alpha^2}{(1-\alpha)^2}[S_t^{(1)} - 2S_t^{(2)} + S_t^{(3)}]$$

式中：

$$S_t^{(1)} = \alpha Y_t + (1-\alpha)S_{t-1}^{(1)}$$

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha)S_{t-1}^{(2)}$$

$$S_t^{(3)} = \alpha S_t^{(2)} + (1-\alpha)S_{t-1}^{(3)}$$

当 $t=1$ 时，通常采用

$$S_1^{(1)} = S_1^{(2)} = S_1^{(3)} = Y_1$$

三次指数平滑模型（布朗三次指数平滑模型）输出相应的计算数据，模型结果展示如图 2-3。

α= □0.05 ☑0.1 ☑0.3 □0.5 □0.7 自定义 [ ]　m= [3]

t	初始值	t	实际值	一次平滑a1	一次平滑a2	二次平滑a1	二次平滑a2	三次平滑a1	三次平滑a2
1987	86.94			a=0.1	a=0.3	a=0.1	a=0.3	a=0.1	a=0.3
1988	103.3	1987	86.94						
1989	104.88	1988	103.3						
1990	98.18	1989	104.88	88.58	91.85				
1991	96.4	1990	98.18	90.21	95.76				
1992	112.65	1991	96.4	91	96.48				
1993	114.22	1992	112.65	91.54	96.46	90.54	99.7	92.85	110.94
1994	113.44	1993	114.22	93.65	101.32	93.93	107.51	98.07	120.96
1995	109.94	1994	113.44	95.71	105.19	95.31	105.87	99.73	111.91
1996	117.66	1995	109.94	97.48	107.66	96.07	102.99	100.3	103.12
1997	130.24	1996	117.66	98.73	108.35	100.26	113.66	106.51	122.07
1998	128.27	1997	130.24	100.62	111.14	104.12	120.02	111.87	129.47
		1998	128.27	103.58	116.87	107.18	122.01	115.67	127.93
		1999		106.05	120.29	108.95	119.48	117.21	119.03

图 2-3　三次指数平滑模型结果展示

① 易丹辉. 统计预测——方法与应用[M]. 北京：中国统计出版社，2001：95～268

(四)ARIMA 模型

ARIMA 模型指将非平稳时间序列转化为平稳时间序列,因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归。ARIMA 模型是由美国学者 George Box 和英国统计学家 Gwilym Jenkins 发明的一种时间序列预测方法。它将预测对象随时间变化形成的序列看作一个随机序列。其中,单个序列值的出现具有不确定性,但整个序列的变化,却呈现一定的规律性。它的基本思想是,这一串随时间变化而又相互关联的数字序列,可以用相应的数学模型加以近似描述。通过对相应数学模型的分析研究,能更本质地认识这些动态数据的内在结构和复杂特性,从而达到在最小方差意义下的最佳预测。ARIMA 模型是一种精度较高的短期预测方法,在建模过程中,需要注意时间序列的随机性、平稳性和季节性特性。随机性和平稳性可以通过序列的自相关图进行判断,在通过差分变换消除序列的趋势性后,可以通过时序曲线图直观地识别出季节性。不同的 ARIMA 模型可以通过 AIC 准则进行识别,该准则可以在模型参数极大似然估计的基础上,对模型的阶数和相应参数给出一种最佳估计。一般来说,AIC 达到最小的那一组阶数为理想阶数,因此 AIC 准则又称作最小信息准则。①

根据原序列是否平稳及回归中所含部分的不同,包括移动平均过程(MA)、自回归过程(AR)、自回归移动平均过程(ARMA)以及 ARIMA 过程。其表达式为:

$$\mu_t = c + \psi_1\mu_{t-1} + \psi_2\mu_{t-2} + \Lambda + \psi_p\mu_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \Lambda + \theta_q\varepsilon_{t-q};$$

其中,$t = 1,2,\cdots,T$。

(五)Logistic 模型

Logistic 模型曲线主要用来描述在环境资源受限制的情况下,生物种群的增长规律。在生物种群实际增长中,增长率受当前现有种群规模、资源和环境等因素的影响,当种群规模较小时,资源和环境相对宽松,增长会快些,当种群规模很大时,资源和环境条件不足,会减缓种群的增长。

Logistic 模型用于描述流动人口增长规律,其优点在于此模型考虑了自然资源、环境条件等因素对人口连续增长的阻滞作用,能够较好地描述流动人口的增长规律,比较符合实际情况。

Logistic 模型算法:

① 吴喜之. 统计学:从数据到结论[M]. 北京:中国统计出版社,2011:237~259

设 $p(t)$ 表示在 t 时刻种群的大小，于是 Logistic 竞争的生物模型为：

$$\begin{cases} p = ap - bp^2 \\ p(t_0) = p_0 \end{cases}$$

解微分方程得到：

$$p(t) = \frac{ap_0}{(a - bp_0)e^{-a(t-t_0)}}$$

（六）灰色系统模型

白色系统指系统内部特征完全已知，黑色系统指系统内部信息完全未知的，而灰色系统是介于白色系统和黑色系统之间的一种系统，其内部部分信息已知、部分信息未知或不确定。灰色系统理论是研究解决灰色系统分析、建模、预测、决策和控制等的理论，灰色预测是对灰色系统所做的预测，灰色预测模型具有所需建模信息少、运算方便、建模精度高等特点，在各种预测领域有着广泛的应用，是处理小样本预测问题的有效工具。

流动人口增长是由环境、社会、经济等诸多因素影响和制约的共同结果，如此众多的因素不可能通过几个指标就能表达清楚，它们对流动人口的潜在而复杂的影响更是无法精确计算。这反映出人口系统具有明显的灰色性，适宜采用灰色模型去发掘和认识原始时间序列综合灰色量所包含的内在规律。

灰色 GM(1,1)模型算法：

设原始序列为：

$$x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\}$$

这是一组信息不完全的灰色量，具有很大的随机性，将其进行生成处理，以提供更多的有用信息。下面选用累加生成，则 m 次累加生成的结果为：

$$x^{(m)} = \{x^{(m)}(1), x^{(m)}(2), \cdots, x^{(m)}(n)\}$$

式中

$$x^{(m)}(k) = \sum_{i=1}^{k} x^{(m-1)}(i) \quad (k = 1, 2, \cdots, n)$$

一般通过一次累加生成就能使数据呈现一定的规律，若规律不够，可增加累加生成的次数。

用线性动态模型对 $X^{(1)}$ 拟合和逼近，其白化形式的微分方程为：

$$\frac{dx^{(1)}(k)}{d(k)} + ax^{(1)}(k) = b$$

白化形式微分方程的离散解为：

$$x^{(1)}(k+1)=\left(x^{(0)}(1)-\frac{b}{a}\right)e^{-ak}-b/a,(k=0,1,2,\cdots,n-1)$$

按 $x^{(0)}(k+1)=x^{(1)}(k+1)-x^{(1)}(k)$，累减生成还原，计算后得到拟合值。

三、时间序列预测模型建模

（一）收集整理数据

运用时间序列模型进行多区域流动人口预测的基础在于数据的收集与整理，通过对不同年份分省数据的收集，并利用线性方法对其中的缺失值进行插补，建立一套用于预测流动人口规模的基础信息数据库。

（二）设定先验值

1. 背景

利用时间序列方法对省际迁移流动人口进行预测，中短期预测精度较高，长期预测会出现较大的偏差。为克服这一问题，避免发生流动人口预测值不合逻辑、不合常理的情况，利用辅助变量对其进行极限控制，考虑到流动人口伴随着中国城市化进程出现、发展，因此考虑使用城市化率作为辅助变量对流动人口在某一时点的总量进行极限控制。

2. 数据来源

流动人口数据来源于普查数据、小普查数据，中间年份的数据利用了线性插值方法进行内插。2004 年及以前年份城市化率数据来自各省统计年鉴及新中国 60 年统计资料汇编，2005 年及以后年份数据来自《中国统计年鉴》。各省份未来规划的城市化率数据来自各地区 2011 ~ 2013 年政府工作报告，以及“十二五”规划纲要。

3. 描述性分析

（1）散点图

见图 2 -4 ~ 图 2 -6。

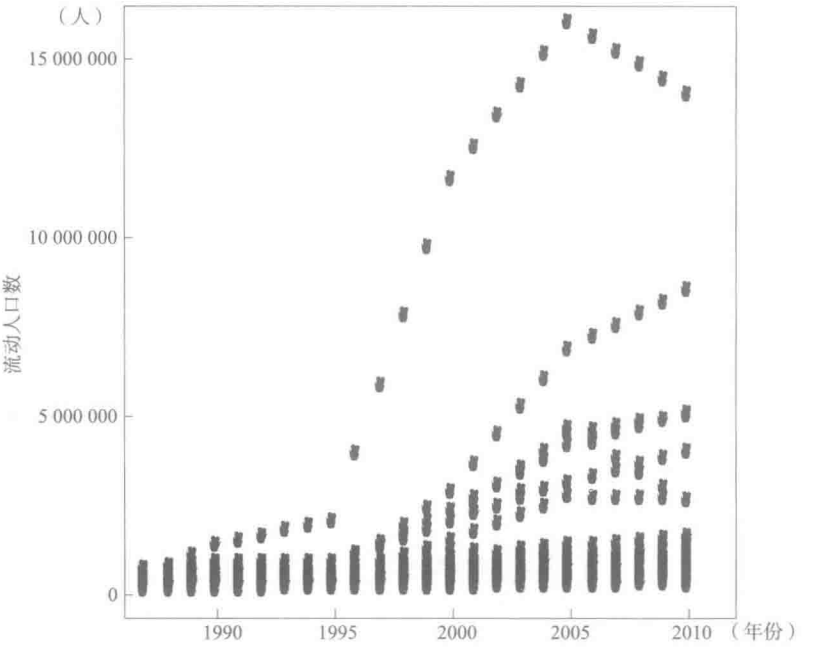

图 2－4　时间与流动人口散点图

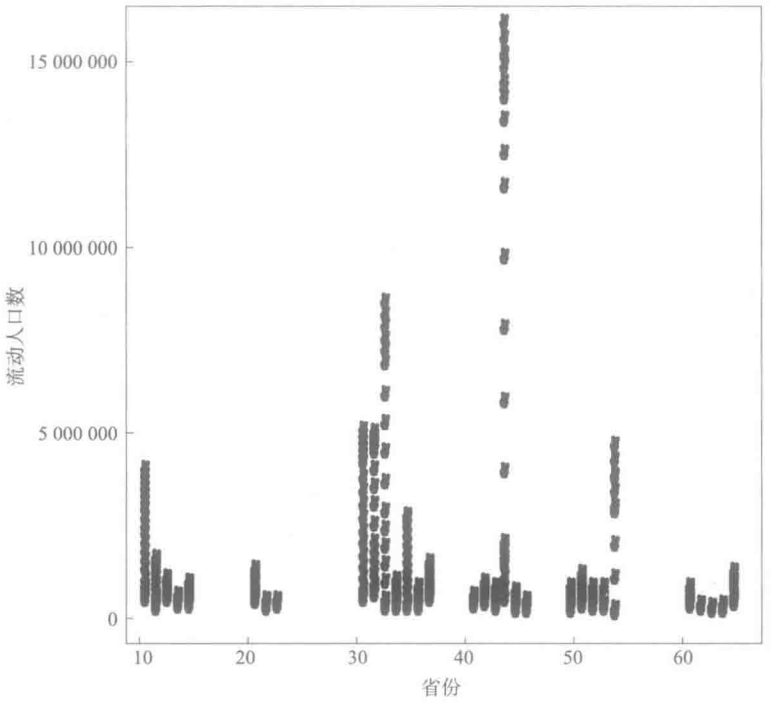

图 2－5　省份与流动人口散点图