郑州西亚斯学院

本科毕业论文（设计）

题　　目　基于 BP 神经网络的河南省总人口预测的研究

指导教师 _____ 职称 _____

学生姓名 张红飞 学号 2019102520136

专　　业 经济统计学 班级 1 班

院　（系） 商学院

完成时间 2022 年 3 月 22 日

目　录

基于 BP 神经网络的河南省总人口预测的研究

摘　要

本研究旨在使用 BP 神经网络模型来预测河南省的总人口变化。通过采用 BP 神经网络模型和传统方法常微分方程模型来预测已有的人口数据，并用预测数据与真实值作比较得出两种模型的准确率。结果表明，在预测准确率方面，BP 神经网络模型相对于传统方法具有更高的精度。此外，在分析河南省出生率、死亡率和自然增长率的变化情况时，本研究还探讨了导致人口下降的可能原因，为制定相关政策和规划提供了重要参考。通过本研究，我们可以了解到如何使用神经网络模型来预测总人口变化，并比较其与传统模型的优劣之处。这对于河南省优化人才结构和产业结构，促进经济发展具有重要意义。同时，本研究的结果还可以为其他省份或地区提供参考，以便更好地应对人口变化带来的挑战。

关键词　人口预测；BP 神经网络模型；常微分方程模型；人口变化；自然增长率

A STUDY OF TOTAL POPULATION PREDICTION IN HENAN PROVINCE BASED ON BP NEURAL NETWORK

ABSTRACT

The aim of this study is to use a BP neural network model to predict the total population change in Henan Province. Both the BP neural network model and the traditional method of ordinary differential equation model were used to predict the available population data, and the accuracy of the two models was compared with the predicted data and the true values. The results show that the BP neural network model is more accurate than the traditional method in terms of prediction accuracy. Furthermore, in analysing the changes in the birth rate, mortality rate and natural growth rate in Henan Province, this study also explores the possible causes of population decline, which provides an important reference for the formulation of relevant policies and plans. This study provides insights into how neural network models can be used to predict total population change and compare their strengths and weaknesses with traditional models. This is important for optimising the talent structure and industrial structure and promoting economic development in Henan Province. At the same time, the results of this study can also provide a reference for other provinces or regions to better cope with the challenges posed by demographic changes.

KEY WORDS Population projections; BP neural network models; Ordinary differential equation models; Population change; Natural growth rates

一、 绪论

（一）研究背景及意义

随着二十大的召开，中国正迈入现代化建设新征程，河南省要想做好现代化建设并引领中部崛起，就必须认识到河南的优势是什么，而加快河南经济建设优势在于人口，潜力也在于人口。如何发挥人口优势，将人口大省转为人才强省和经济强省，就需要对河南未来人口走势进行预测，并根据预测趋势制定相关政策来优化人才结构和产业结构，促进经济发展。

人口问题始终是制约河南省经济发展的重要因素之一。无论是对当前河南省经济发展状况的认识，还是对未来河南省发展的预测，人口问题的研究都具有十分重要的意义。想要获得准确的人口预测就需要考虑合适的预测方法。本研究挑选 BP 神经网络模型和传统方法常微分方程模型来预测现有的人口统计数据，并用预测数据与真实值作比较得出两种模型的准确率。并用 BP 神经网络模型预测未来几年的人口变化。本研究能够比较传统模型预测与 BP 神经网络何者更有优势，使人口预测在神经网络方法选择上提供理论支持。

（二）文献综述

国内有关的研究如下：

贾楠（2012）应用神经网络系统来预测人口，分别采用三种不同的神经网络方法进行预测，比较哪种方法预与实际人口数量更接近，并将这三种神经网络的预测人口结果进行了比较，同时对BP神经网络进行了改进并做了进一步的研究[1]。刘晓峰（2007）以 2005 年河南省 1%人口抽样调查分年龄人口状况、分年龄死亡人口状况、育龄妇女分年龄分孩次的生育状况以及 2000 年河南省第五次人口普查数据的有关资料为基础，结合制约河南人口发展变化的主要因素，采用多因素分析的动态人口预测模型。对未来人口的出生、死亡和年龄构成变化进行预测[2]。

周美旭（2015）提出一种组合模型，利用 GM(1,1)模型、BP 神经网络模型相组合对江西人口预测进行建模，并利用 GA 算法对 BP 神经网络做优化;利用这种创新的

组合模型进行人口的预测建模，结果表明无论是在拟合精度上还是在预测精度上都要优于其他的单一的模型[3]。陆文珺,柳炳祥（2016）基于 BP 神经网络的时间序列预测方法对我国人口总数进行预测，使用该方法的预测结果与实际误差很小，精确度较高，模型简单易行[4]。

国外有关的研究如下：

随着神经网络技术的不断发展，人口预测领域的研究也取得了丰硕的成果。

Gerland 和 Raftery(2014)通过神经网络模型对全球人口增长趋势进行了预测。作者们指出，本世纪人口稳定的可能性不大，这将对资源分配和国际政治产生深远影响[5]。Raftery 和 Alkema(2014)在这篇文章中采用贝叶斯方法结合神经网络技术，对联合国人口预测进行了改进。这一方法提高了预测准确性，为全球各国政策制定提供了更为可靠的人口数据支持[6]。Şahinarslan F V，Tekin A T，Çebi F(2021)的研究应用机器学习算法进行人口预测，使用了不同的机器学习算法，包括极端梯度提升、CatBoost、线性回归、岭回归、Holt-Winters、指数、自回归移动平均（ARIMA）和预言家预测模型。研究使用了 1960 年至 2017 年 262 个不同国家的 1595 个不同人口统计指标进行模型训练[7]。

（三）研究内容及方法

本研究旨在探讨如何使用 BP 神经网络模型来预测河南省的人口变化，并比较其与传统模型的优劣之处。人口问题一直是制约河南省经济发展的重要因素之一，因此准确预测人口变化对于优化人才结构和产业结构，促进经济发展具有重要意义。为了实现这一目标，本研究首先对河南省的人口变化趋势进行了分析，并提出了基于常微分方程模型的预测方法。然后，我们引入了 BP 神经网络模型，并将其与传统模型进行比较和评估。最后，我们使用 BP 神经网络模型来预测未来几年河南省的人口变化，并提出了相应的建议。

以下是本研究中使用的具体方法：

（1） 马尔萨斯模型：马尔萨斯模型是一种基于人口增长率和资源增长率之间关系的数学模型。该模型假设人口增长率呈指数增长，而资源增长率呈线性增长。通过对历史数据进行拟合，可以预测未来人口变化趋势。

（2） Logistic 模型：Logistic 模型是一种基于 S 形曲线的数学模型。该模型假设

人口增长率随着人口数量的增加而逐渐减缓，并最终趋于稳定。通过对历史数据进行拟合，可以预测未来人口变化趋势。

（3） 改进 Logistic 模型：改进 Logistic 模型在传统 Logistic 模型的基础上引入了时间因素和外部影响因素，如政策、经济、社会和自然因素等。通过对历史数据进行拟合，并考虑这些外部因素的影响，可以提高预测准确度。

（4） 基于长短时记忆网络（LSTM）的预测方法：LSTM 是一种特殊类型的神经网络，能够有效地处理序列数据中存在的长期依赖关系。通过对历史数据进行训练，LSTM 模型可以预测未来人口变化趋势。

二、 人口预测的相关知识

（一）关于人口的一些基本概念

人口预测作为一种为社会经济发展规划提供关键信息的有效手段，旨在帮助我们了解发展过程中可能面临的问题，并据此制定合理的政策。这一预测方法的历史可追溯至 1696 年，当时英国知名社会学家 G•金运用初步的数学方法对英国未来 600 年的人口发展进行了预测。尽管受限于当时的社会生产力和生产关系，他的预测结果与实际人口数量存在较大差异，然而，他的预测思想为后世研究人口预测提供了宝贵的启示[1]。

在现代，人口预测已经成为一个较为成熟和精确的科学方法。通过收集和分析大量数据，以及运用先进的统计学和计算机技术，我们可以预测出未来的人口数量、年龄结构、性别比例等各项指标。这些预测结果对于规划国家和地区的经济发展、教育、医疗、社保等各个方面都具有非常重要的参考价值。

总的来说，人口预测不仅是一个重要的研究领域，也是一个应用广泛的工具。通过不断地改进和完善，我们可以更好地利用人口预测的信息，为社会和经济的发展做出更加明智的决策。由于人口预测是一项较为复杂的工程，涉及概念较多，本文只介绍预测中所使用的一些主要概念：

（1） 总人口：指一定时点、一定地区范围内的有生命的个人总和。总人口通常使用常住人口的口径[8]。

（2） 常住人口：本地常住人口是指实际经常居住在某地区半年以上的人口。在

人口普查和抽样调查规定下，主要包括以下三类人口：第一，在本地居住，且户口也在本地的人口；第二，户口在外地，但在本地居住半年以上的人口，或者离开户口所在地半年以上，但在调查时在本地居住的人口；第三，在调查时居住在本地，但在任何地方都没有登记常住户口，例如持有户口迁移证、出生证、退伍证、劳改劳教释放证等尚未办理常住户口的人，即“口袋户口”人口。

（3）出生率：指在一定时期内(通常为一年)一定地区的出生人数与同期内平均人数(或期中人数)之比，用千分率表示。本资料中的出生率指年出生率，其计算公式为:

$$出生率 = 年出生人数/年平均人数 \times 1000\% \quad (2\text{-}1)$$

（4） 死亡率：指在一定时期内(通常为一年)一定地区的死亡人数与同期平均人数(或期中人数)之比一般用千分率表示。计算公式为:

$$死亡率 = 年死亡人数/年平均人数 \times 1000\% \quad (2\text{-}2)$$

（5）人口自然增长率：指在一定时期内(通常为一年)人口自然增加数(出生人数减死亡人数) 与该时期内平均人数(或期中人数)之比，一般用千分率表示。计算公式为:

$$人口自然增长率=(本年出生人数-本年死亡人数)/年平均人数 \times 1000\% = 人口出生率 - 人口死亡率 \quad (2\text{-}3)$$

（6） 性别比：总人口中男性人数与女性人数之比。通常用每 100 个女性人口相应有多少男性人口表示。其计算公式为:

$$性别比 = 男性人口数/女性人口数 \times 100\% \quad (2\text{-}4)$$

（二）人口预测的意义

人口预测在众多领域具有重要的意义。首先，它可以帮助揭示未来人口规模、结构与分布的变化趋势，使政策制定者和研究人员更好地了解和预测未来的人口现象。其次，人口预测为政府和企业在社会经济规划与政策制定中提供关键的参考信息，以满足不同领域如教育、卫生、就业、养老和住房的需求，实现资源的合理分配与利用。此外，通过研究人口预测，人们可以深入探讨人口变动的原因及其与社会、经济、环境等因素之间的相互关系，从而提高人口学、经济学、地理学等相关学科的理论深度和实证研究水平。最后，人口预测也有助于政府和国际组织制定应对全球性挑战的策

略，如气候变化、粮食安全、人口迁徙和疫情防控等，为提前采取有效措施、降低潜在风险和实现全球可持续发展提供重要依据。

（三）人口预测的传统方法

1、 一元线性回归法

一元线性回归法是一种通过使用线性回归模型来预测人口增长趋势的方法。它是一种基于历史数据和趋势来预测未来人口增长的技术。一元线性回归是一种统计分析方法，用于确定两个变量之间的线性关系，其中一个变量是自变量（也称为预测变量），另一个变量是因变量（也称为响应变量）。在预测人口增长的情境中，自变量可以是时间或年份，而因变量则是人口数量。通过使用一元线性回归模型，可以确定时间或年份与人口数量之间的线性关系。该模型可以使用历史数据来拟合一条直线，以描述这种关系，并且可以根据这条直线来预测未来的人口增长。需要注意的是，使用一元线性回归法预测人口增长具有一定的局限性。例如，该方法假设人口增长趋势是线性的，但实际上人口增长可能受到多种因素的影响，可能会发生突变或非线性变化。

2、 常微分方程模型

常微分方程模型在人口预测中具有广泛的应用，通过建立描述人口数量或密度随时间变化的方程，以揭示人口动态的内在规律。其中，马尔萨斯模型和 Logistic 模型是人口预测领域的两个典型模型。马尔萨斯模型假设人口增长呈指数形式，即人口数量随时间按比例增长。该模型以简单易懂的形式反映了人口增长过程，但并未考虑资源限制和人口密度对增长速率的影响。Logistic 模型则是对马尔萨斯模型的扩展，考虑了环境容量对人口增长的限制。该模型表明，在环境容量限制下，人口增长将呈现出 S 型曲线，即人口数量在一定阶段内呈指数增长，随后增速减缓，最终趋向于稳定。通过运用常微分方程模型，研究人员可以描述和预测人口动态，为人口预测提供了一种有力的理论工具。然而，需要注意的是，实际人口预测过程中可能受到多种因素影响，因此在应用这些模型时，需要结合实际情况进行修正和调整。

3、 灰色系统法

预测人口的灰色系统法是一种基于灰色系统理论的预测方法，通过建立一个基于小样本的灰色模型，对人口数量进行预测。灰色系统理论认为，系统的运动规律有时

会受到难以预测的因素的影响，因此不能使用传统的统计方法进行预测。相对地，灰色系统理论是基于少量数据而发展的一种预测方法，它可以在数据量不足的情况下对系统进行预测。在预测人口时，首先需要将历史数据进行序列化处理，以便分析其规律。然后，根据序列化的历史数据，建立灰色模型，对未来的人口数量进行预测。

灰色系统法的优点是可以在小样本数据的情况下进行预测，并且不需要假设数据的分布。同时，该方法可以较好地处理非线性问题和非稳态问题。需要注意的是，在使用灰色系统法进行人口预测时，需要仔细考虑所选的模型和方法的适用性，并结合其他因素进行分析和预测。该方法也存在一定的局限性，需要满足数据序列的平稳性和相对变化率的恒定性等假设条件，否则可能会影响预测结果的准确性。

4、人口发展方程

人口发展方程是一种预测人口变化的动态模型，它考虑了出生率、死亡率、迁移率等因素，用数学公式预测未来人口的变化趋势。然而，人口发展方程方法存在一些不足，比如人口发展方程方法基于一些假设前提，如未来的出生率、死亡率和迁移率与历史数据呈现相同趋势。然而，在实际情况中，这些假设并不一定成立，从而影响预测结果的准确性。而且人口发展方程方法主要考虑长期因素对人口数量变化的影响，如经济、社会、文化和政治等因素，而难以预测一些突发事件，如自然灾害、战争等对人口数量变化的影响，从而影响预测结果的准确性。

接下来，我们会用人口预测的传统方法的常微分方程模型的两个经典模型——马尔萨斯模型和 Logistic 模型进行人口预测与基于反向传播算法的 LSTM 模型进行比较，并用 LSTM 模型预测河南省未来十年的人口。

三、 河南省人口现状

（一）横向分析

横向分析主要关注同一时点上不同地区或群体之间的人口特征差异。在河南省的背景下，横向分析可以比较同一时间内不同人口群体（例如：不同年龄段、性别、职业等）的人口规模、结构与分布等特点。这有助于理解河南省各地区或群体在人口特征方面的异同，并为地区间合作与政策制定提供依据。

根据河南省统计局发布的 2022 年河南省国民经济和社会发展统计公报显示，2022 年年末，我省总人口为 9872 万人，同 2021 年相比下降了 11 万人，出生人口数为 73.3 万人，出生率为 7.42%，死亡人口数为 74.1 万人，死亡率为 7.50%，自然增长率为-0.08%。自然增长率为负与 2022 年疫情和国家政策调节放开相关，由于 65 岁以上人口免疫力较弱且携带多种并发症，老年人属于高危群体，同 2021 年相比 65 岁以上人口下降 62.263 万人，比重上升了 0.60%，说明我省老龄化趋势也并未减弱。

2021 年底，全省共有家庭户 3365 万户，户籍人口 11533 万人。全省人口中，男性为 5947 万人，占 51.57%，女性为 5586 万人，占 48.43%，性别比为（以女性为 100，男性对女性的比例）为 106.5。

全省人口中，根据第七次人口普查，大学文化程度的人口为 1167 万人，占 11.74%，高中文化程度为 1514 万人，占 15.24%，初中文化程度为 3728 万人，占 37.52%，小学文化程度为 2440 万人，占 24.55%，文盲和半文盲为 223 万人，占 10.95%。与第六次人口普查相比，全省人口具有大学程度的由 602 万人上升为 1167 万人，具有高中程度的由 1242 万人上升为 1514 万人，初中程度的由 3993 万人下降为 3728 万人，小学程度的由 2267 万人上升到 2440 万人。从以上数据可以看出我省人口的受教育水平正在稳步增长。

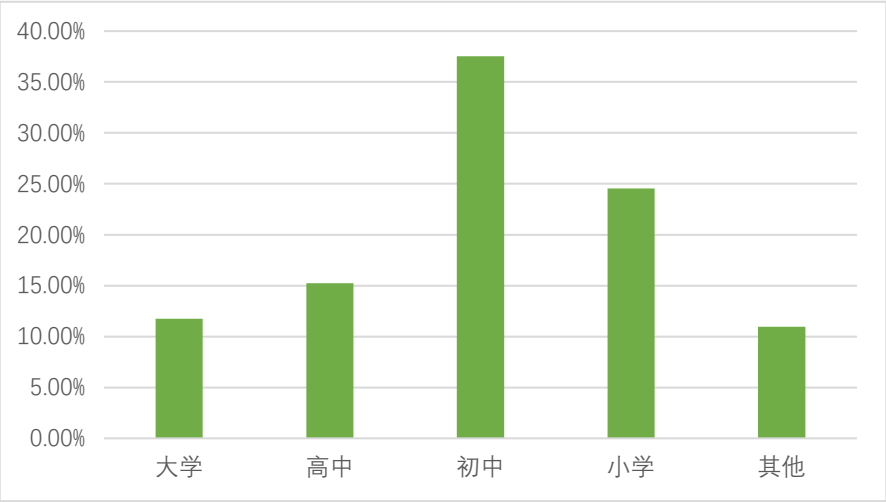

图 3-1 第七次人口普查受教育程度情况

在 2022 年年末全省人口中，0-15 岁的人口为 2266 万人，占 23.00%；16-59 岁

的人口为 5744 万人，占 58.20%；60 岁及以上的人口为 1862 万人，占 18.90%。人口结构相对合理，属于成年型人口结构，但仍要警惕人口老龄化趋势，人口老龄化会对地区经济形成冲击。同 2021 年普查数据相比，65 岁及以上人口的比重上升了 0.60%。从图 3-2 可以看出，目前我省人口，0-15 岁人口较少，60 岁以上人口最少，可以看出我省人口年龄结构较合理，其未来发展趋势还需要进一步分析。

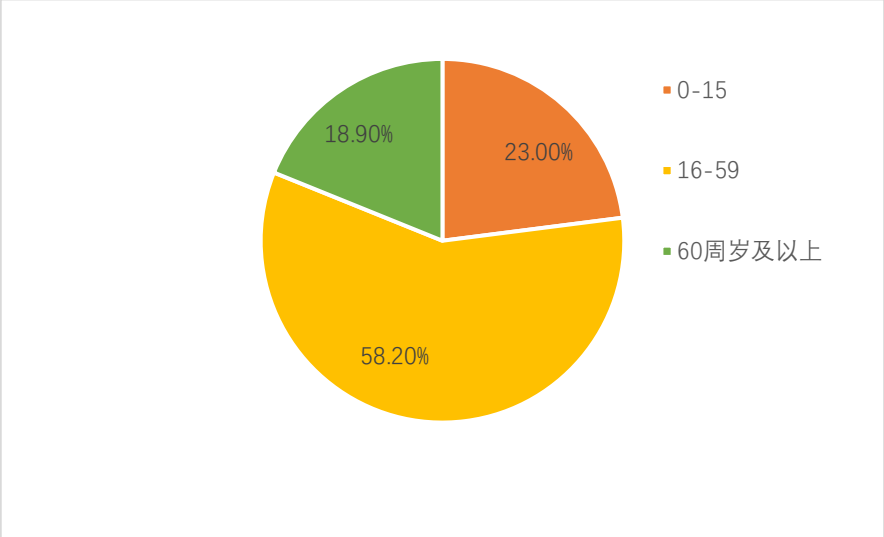

图 3-2 2022 年各年龄段所占比重

（二）纵向分析

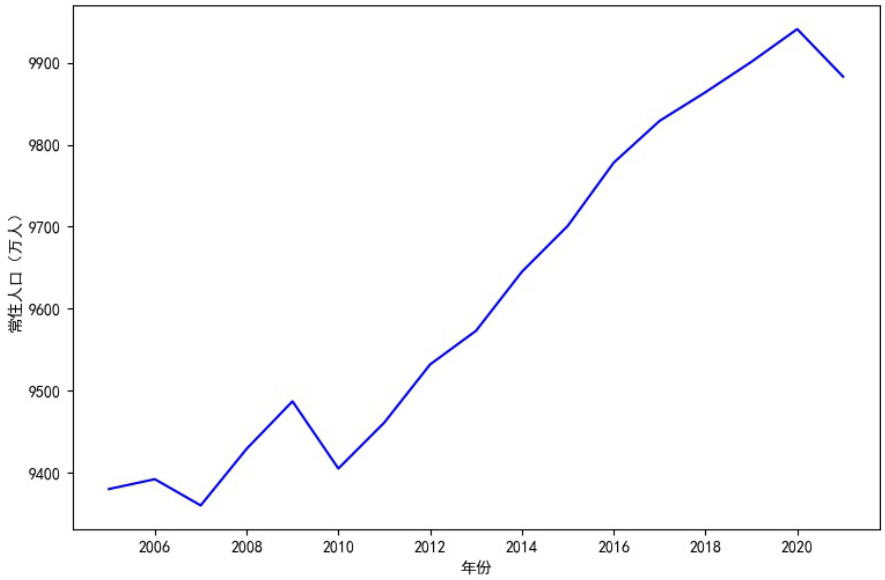

图 3-3 2022 年统计年鉴总人口数据

纵向分析主要关注某一地区或群体在一段时间内的人口特征变化。在河南省的背景下，纵向分析可以研究过去几十年内河南省整体生育率、死亡率和迁移率等关键人口变量的变化。这有助于揭示河南省人口发展的历史趋势与未来预测，为政策调整和资源配置提供参考。

根据 2022 年统计年鉴和 2022 年年河南省国民经济和社会发展统计公报的数据，由图 3-3 可知，从 2005 年年底到 2022 年底我省总人口出现三次峰值，第一次峰值出现在 2006 年，人口数达到 9392 万人；第二次峰值出现在 2009 年，人口数达到 9487 万人；第三次峰值出现在 2020 年，人口数达到 9941 万人。2021 年比 2020 年减少 58 万人，省统计局方面分析由于 2021 年全国疫情逐渐好转，出省人口大量增加[9]。

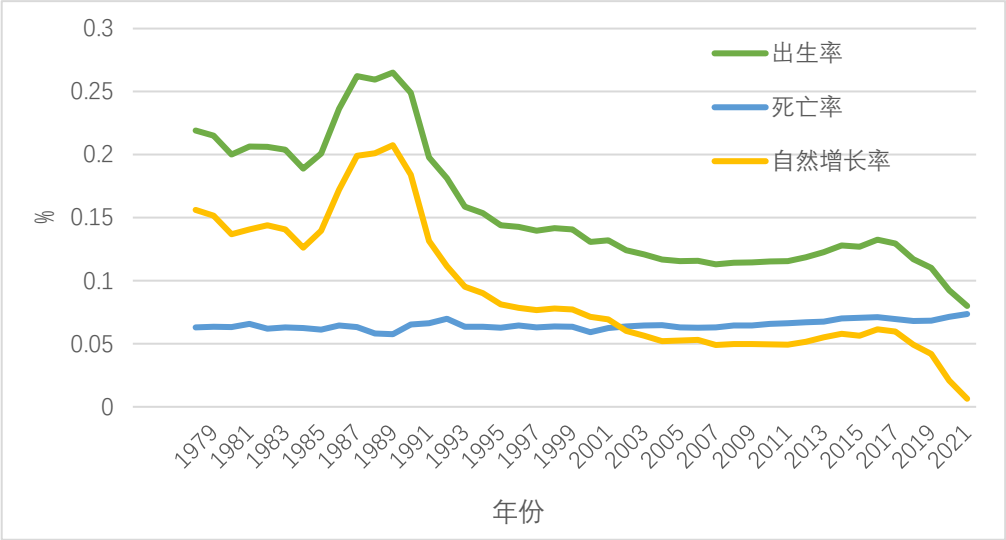

图 3-4 河南省出生率、死亡率、自然增长率变化情况

由图 3-4 河南省出生率、死亡率、自然增长率变化情况可知，由于 1982 年左右国家依据人口增长趋势调整政策，加强计划生育政策执行，严格控制人口，1989 年河南省人口自然增长率达到峰值，在此之后，河南省人口自然增长率逐渐下降并趋于平缓。且随着基础设施的改善和教育水平的提高，河南省居民的生育观念逐渐发生改变。越来越多的人选择少生育或不生育，这也可能是导致河南省人口下降的一个原因。

四、 基于常微分方程模型的预测方法

（一）马尔萨斯模型

1. 模型介绍

马尔萨斯模型是一个简单的指数增长模型，其函数与函数增长率成正比。该模型得名于托马斯•罗伯特•马尔萨斯，马尔萨斯是 18 世纪末至 19 世纪初的英国经济学家和社会学家。他在 1798 年发表了名为《人口原理》的著作，该书警告，若人口数量以指数形式增长而粮食生产以线性增长的情况下，除非出生率下降，否则二者的差距将导致粮食匮乏和饥荒。该模型主要探讨人口增长与资源之间的关系，强调人口增长对经济和环境的影响。

2. 模型基本假设

（1）没有重大或者毁灭性的自然灾害和疾病发生；

（2）人口数量的变化是封闭的，即人口数量的增加与减少只取决于人口中个体的生育和死亡，且每一个体都具有同样的生育能力与死亡率。

（3）所有人口的年龄上限为 100 岁；

（4）跨省人口迁移的影响基本上可以忽略，省内迁入和迁出人口的数量基本保持平衡。

3. 模型的建立及预测

取 2022 年统计年鉴的数据为基础，并将这些数据用 python 的 SciPy 库的子模块 optimize 进行曲线拟合，记 2005 年$x_0=9380$（万人）为初始人口数量，t 时刻到$t+\Delta t$ 时刻人口的增量为$x(t+\Delta t)-x(t)=rx(t)\Delta t$，于是得

$$\begin{cases}\dfrac{dx}{dt}=rx,\\ x(0)=x_0,\end{cases} \tag{4-1}$$

其解为

$$x(t)=x_0e^{rt}. \tag{4-2}$$

表 4-1 总人口实际数据表

年份	2005	2006	2007	2008	2009	2010	2011	2012	2013
总人口（万人）	9380	9392	9360	9429	9487	9405	9461	9532	9573
年份	2014	2015	2016	2017	2018	2019	2020	2021	2022
总人口（万人）	9645	9701	9778	9829	9864	9901	9941	9883	9872

数据来源：2022 年《河南省统计年鉴》，2022 年河南省国民经济和社会发展统计公报

表 4-1 总人口实际数据表 2005 年到 2021 年的实际人口总量的数据，通过 2005-2021 年数据拟合得到人口增长率 r 为 0.003436（保留 6 位小数），另外利用 python 可以得到预测数据与实际数据的对比结果（如 图 4-1 马尔萨斯人口模型拟合曲线图 图 4-1 马尔萨斯人口模型拟合曲线图），实线表示马尔萨斯模型预测的河南省人口数据，圆点表示河南省的实际人口数据。以 2022 年为例，真实总人口数据

为 9872 万人，而马尔萨斯模型预测的 2022 年总人口为 9944.16 万人，与真实数据相差较大。

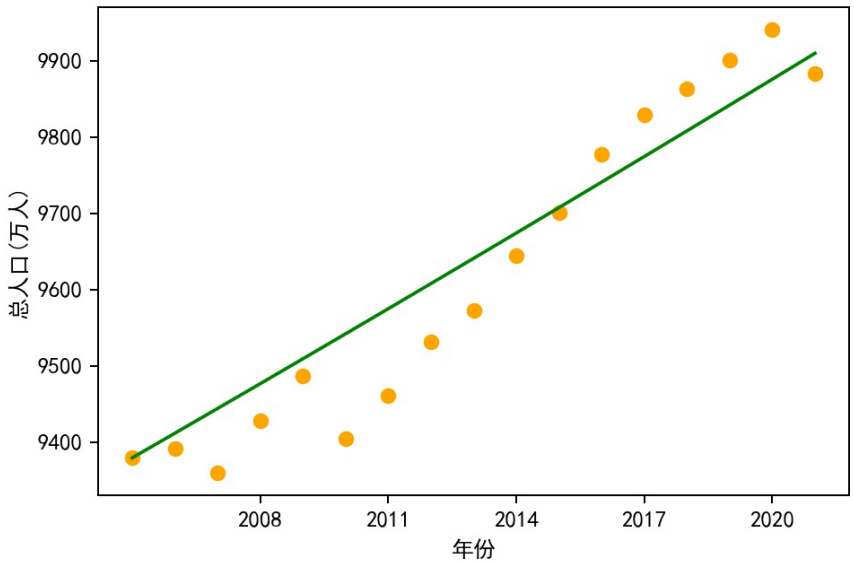

图 4-1 马尔萨斯人口模型拟合曲线图

（二）Logistic 模型

1. 模型介绍

Logistic 人口预测模型又称阻滞增长模型，是一种基于 Logistic 函数的数学模型，用于描述人口增长的动态过程。Logistic 函数是一种 S 形曲线，可以反映人口增长受到环境容量的限制，即人口不能无限制地增长，而是会趋向于一个稳定的平衡状态。Logistic 人口预测模型的微分方程为：

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \tag{4-3}$$

解这个微分方程可以得到一般公式：

$$f(t) = \frac{K}{1 + \left(\frac{K}{x_0} - 1\right)e^{-r(t-2005)}} \tag{4-4}$$

其中，N 是人口数量，t 是时间，r 是人口增长率，K 是环境承载力，K 指的是在特定环境条件下，一个生态系统所能支持的某种生物种群的最大稳定数量。在 Logistic 模型中，当种群数量接近 K 时，增长速度会减缓，最终趋于稳定。K 值在 logistic 模型中决定了种群数量的上限。该模型可以通过拟合历史数据来估计 r 和 K 的值，从而对未来的人口变化进行预测。Logistic 人口预测模型具有一定的理论和实际意义，可以帮助我们分析人口问题，制定合理的人口政策。

2. 模型基本假设

（1）人口增长率与人口规模成正比，也就是说，人口越多，增长越快；

（2）人口增长率受到环境容量的限制，也就是说，当人口接近环境容量时，增长率会下降；

（3）环境容量是一个常数，不随时间变化；

（4）人口增长率和环境容量之间的关系可以用一个 S 形曲线来描述。

3. 模型的建立及预测

通过表 4-1 总人口实际数据表的数据拟合 logistic 模型，得到人口增长率 r 为 0.003436，环境承载力 K 为 568080901.09，利用 python 拟合可以得到 Logistic 模型和实际数据对比的结果（如图 4-2），实线表示 Logistic 模型预测的河南省人口数据，圆点表示河南省的实际人口数据。以 2022 年为例，真实总人口数据为 9872 万人，而马尔萨斯模型预测的 2022 年总人口为 9944.16 万人，由于数据量较小和数据点之间的差异较小，数据没有显著的波动或趋势变化，人口数量在不同年份间的增长速度相对稳定，变化幅度不大。这导致了在这个特定数据集上，马尔萨斯模型和 logistic 模型的拟合效果非常接近。

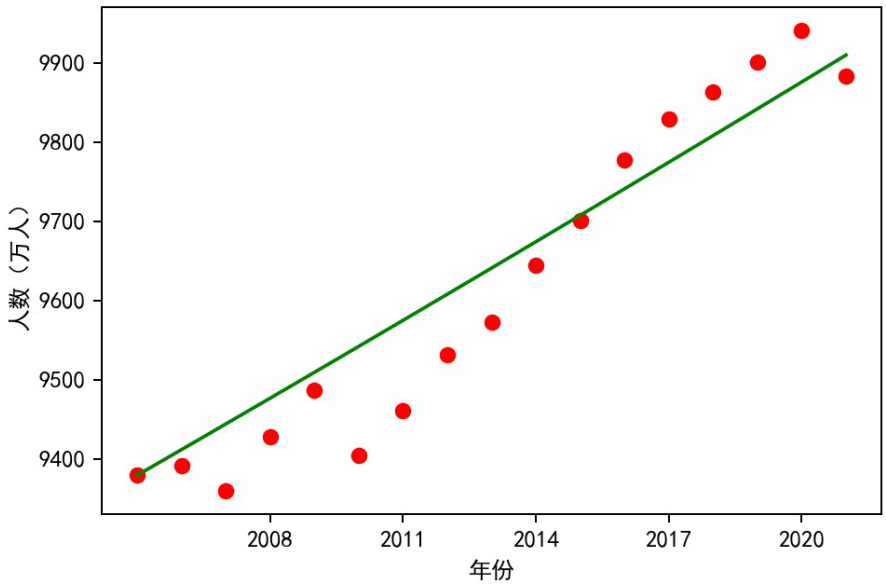

图 4-2 Logistic 人口模型拟合图

（三）改进的 Logistic 模型

1. 模型的建立

对于初始的 Logistic 模型，拟合效果与马尔萨斯模型相似，可以通过改进模型参数使 Logistic 模型拟合效果更好。在这里，为了使 Logistic 模型拟合效果更好，可以引入一个新的参数 C，C 是一个常数，参数 C 允许更灵活地拟合不遵循简单 sigmoid 曲线的数据，增加了这个参数后，曲线的形状也会受到影响，因为 C 参数会影响曲线的斜率和拐点。一般来说，C 参数越大，曲线越陡峭，拐点越靠后；C 参数越小，曲线越平缓，拐点越靠前。因此，这个参数可以用来调节曲线的灵敏度和适应性，从而更准确地预测人口数量的增长趋势。

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right) \tag{4-5}$$

解这个微分方程可以得到 Logistic 模型的一般公式：

$$f(t) = \frac{K}{1 + C \cdot e^{-r(t-2005)}} \tag{4-6}$$

在这个函数中，t 是时间变量，K 和 r 是我们需要通过拟合实际数据来确定的

参数。K 代表人口的最大可能值，即环境承载量；r 描述了人口生长的速率，反映了人口增长的快慢。接下来，我们使用实际的人口数据 x 和对应的年份 t，利用非线性最小二乘法对模型参数 r、K 和 C 进行估计。在拟合过程中，我们设置了参数的初始值和上限，并将最大迭代次数设为足够大的值，以确保收敛。通过拟合得到的参数值，我们可以计算出逻辑生长函数在不同年份的人口预测值。

2. 预测

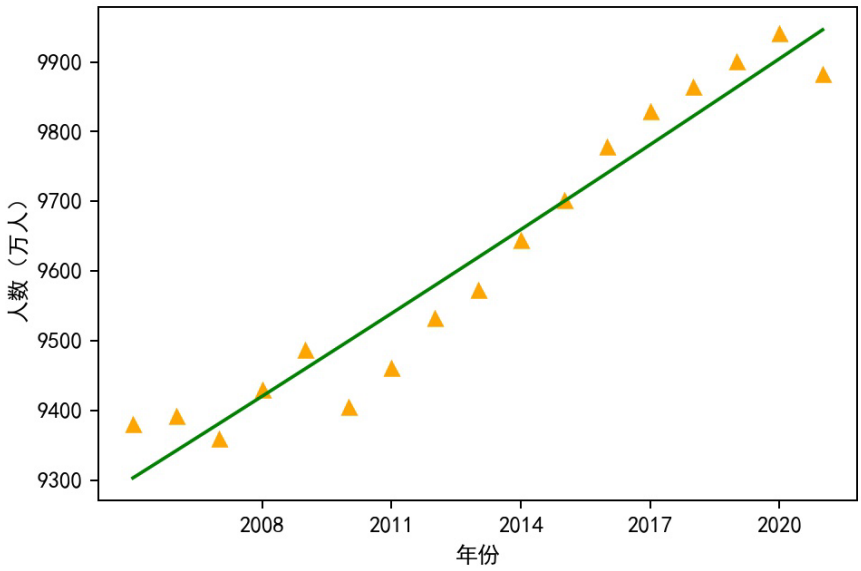

图 4-3 改进的 Logistic 人口模型拟合图

通过 2005-2021 年的人口数据拟合，于是我们得到改进的 Logistic 人口模型拟合图（如图 4-3），实线表示改进的 Logistic 人口模型预测的河南省人口数据，三角形表示河南省的实际人口数据。以 2022 年为例，真实总人口数据为 9872 万人，而 Logistic 人口模型预测的 2022 年总人口为 9904.67 万人，可以看出改进后的模型明显比未改进的模型更接近真实值。

五、 反向传播算法和 LSTM 神经网络模型的介绍

（一）反向传播

反向传播（Backpropagation）是一种常用于训练神经网络的优化算法。其基本思想是通过计算网络输出与实际标签之间的误差，并将误差通过网络反向传播，更新网络权重以最小化误差。

假设我们有一个有n层的神经网络，其中第i层的输出为$a^{(i)}$，$W^{(i)}$表示第i层到第$i+1$层的权重矩阵，$b^{(i)}$为第$i+1$层的偏置向量，$z^{(i+1)}$为第$i+1$层的输入。我们的目标是最小化损失函数$J(W,b)$。其中，W表示所有权重参数，b表示所有偏置参数。

首先，我们通过正向传播计算网络的输出，并计算出损失函数$J(W,b)$。接下来，我们通过反向传播来计算每个参数对损失函数的影响，并更新这些参数以最小化损失函数。

反向传播的推导过程可以通过链式法则来实现。我们假设L表示最终输出层的误差，即$L=\frac{\partial J}{\partial a^{(n)}}$。我们的目标是计算$L$对每个权重和偏置的偏导数，即$\frac{\partial J}{\partial W^{(i)}}$和$\frac{\partial J}{\partial b^{(i)}}$。

我们可以根据链式法则，将L的导数分解为每一层的导数，即：

$$\begin{aligned}\frac{\partial J}{\partial W^{(i)}}&=\frac{\partial J}{\partial z^{(i+1)}}\frac{\partial z^{(i+1)}}{\partial W^{(i)}}=\delta^{(i+1)}(a^{(i)})^{T}\\\frac{\partial J}{\partial b^{(i)}}&=\frac{\partial J}{\partial z^{(i+1)}}\frac{\partial z^{(i+1)}}{\partial b^{(i)}}=\delta^{(i+1)}\end{aligned}\tag{5-1}$$

其中，$\delta^{(i+1)}$表示第$i+1$层的误差，可以通过以下公式计算：

$$\delta^{(i+1)}=\frac{\partial J}{\partial z^{(i+1)}}=\frac{\partial J}{\partial a^{(i+1)}}\frac{\partial a^{(i+1)}}{\partial z^{(i+1)}}=g^{(i+1)}(z^{(i+1)})\odot\frac{\partial J}{\partial a^{(i+1)}}\tag{5-2}$$

其中，$g^{(i+1)}(z^{(i+1)})$表示第$i+1$层的激活函数的导数，$\odot$表示对应元素相乘。这个公式的含义是，第$i+1$层的误差是由输出误差L和第$i+1$层的激活函数的导数相乘得到的。可以看出，误差是从输出层开始计算，沿着神经网络向前传播，通过链式法则逐层计算得到的。

最后，我们可以通过梯度下降法来更新网络的参数，以最小化损失函数。具体地，我们可以使用以下公式更新权重和偏置：

$$W^{(i)} \leftarrow W^{(i)} - \alpha \frac{\partial J}{\partial W^{(i)}}$$
$$b^{(i)} \leftarrow b^{(i)} - \alpha \frac{\partial J}{\partial b^{(i)}} \tag{5-3}$$

其中，α是学习率，控制每次参数更新的步长。

总之，反向传播是一种有效的神经网络训练算法，通过计算误差的反向传播，逐层更新网络的权重和偏置，以最小化损失函数。

（二）LSTM 模型介绍

LSTM（长短时记忆网络）是一种深度学习中的循环神经网络（RNN）。与传统的 RNN 相比，LSTM 引入了“门”的概念，通过控制门的开关来控制信息的流动，从而解决了传统 RNN 在长序列训练中的梯度消失或爆炸问题。

LSTM 中的关键部件是“记忆单元”，它可以记住一段时间内的信息，并根据输入数据和上一时刻的状态决定哪些信息需要被遗忘或保留。同时，LSTM 还有输入门、输出门和遗忘门等“门”的组件，它们可以控制信息的流动，保证模型具有更好的记忆和泛化能力。

下面介绍 LSTM 的具体公式。

首先，我们定义 LSTM 的输入为x_t，输出为h_t，上一时刻的状态为c_{t-1}，当前时刻的状态为c_t。LSTM 的状态更新公式如下：

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
\tag{5-4}
$$

其中，σ 是 sigmoid 函数，$\odot$ 表示逐元素相乘。f_t、i_t 和 o_t 分别是遗忘门、输入门和输出门的值。$\tilde{c}_t$ 是当前时刻的候选状态，c_t 是当前时刻的状态，h_t 是当前时刻的输出。

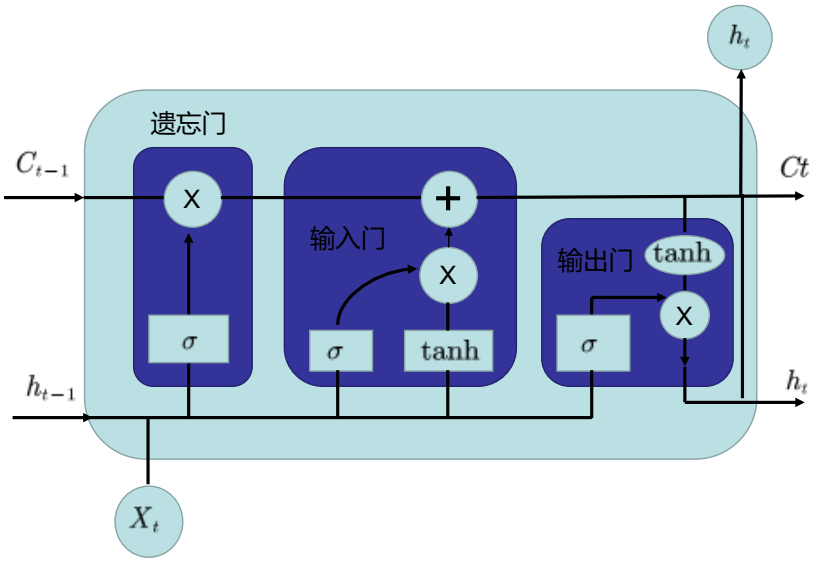

图 5-1　LSTM 网络的循环单元结构

遗忘门的作用是控制上一时刻状态 c_{t-1} 中哪些信息需要被遗忘，哪些信息需要被保留。输入门的作用是控制当前时刻输入 x_t 中哪些信息需要被加入状态 c_t 中。输出门的作用是控制当前时刻的输出 h_t 中哪些信息需要被保留。

LSTM 的参数包括 W_f、W_i、W_c、W_o、b_f、b_i、b_c、b_o 等。这些参数可以通过反向传播算法来训练，使得 LSTM 可以适应各种不同的任务。

六、 基于 LSTM 神经网络的人口预测

（一）LSTM 模型的建立

1. 数据集介绍

本文使用的数据集来自于 2022 年河南省统计年鉴。数据集包括年份、常住人口、人口密度、城镇化率、性别比、男性人口、女性人口、户籍人口等变量。选择这些变量的原因如下：

首先，常住人口是本研究的预测目标，是影响河南省人口变化的重要因素之一。人口密度和城镇化率则是常住人口变化的主要驱动力之一，能够反映城市化和人口流动的情况。性别比、男性人口和女性人口是人口结构的重要组成部分，对于了解河南省人口的基本情况和未来趋势具有重要意义。户籍人口则可以反映出城乡人口分布情况，是了解城市化进程和农村人口流动的重要指标。

其次，本文选择 2005 年到 2021 年的数据，是因为 1978 年到 2004 年常住人口存在缺失值，无法满足本研究的需求。因此，本文选择了 2005 年到 2021 年的完整数据集进行研究。

在使用 LSTM 模型进行人口预测之前，本文使用特征选择方法对数据集进行了筛选和处理，选择了上述变量作为输入特征。这些变量具有代表性和可解释性，并且在进行预测时具有重要的预测能力。

2. 数据预处理

在利用长短时记忆神经网络（LSTM）对河南省的人口数据进行预测之前。首先，我们对原始数据进行预处理，包括时间格式处理、缺失值处理和数据标准化。预处理后的数据集作为模型训练和测试的基础。

在模型构建之前，我们对预处理后的数据进行了相关性分析。相关性分析能够揭示数据集中各变量之间的关联程度，有助于了解哪些变量对目标变量河南省总人口（使用的是常住人口的口径）具有较强的预测能力。在本研究中，我们使用了皮尔逊相关系数作为衡量变量之间关联程度的指标，我们计算了数据集中所有变量两两之间的皮尔逊相关系数，并将结果以热力图的形式进行可视化（如图 6-1）。皮尔逊相关系数的取值范围为 [-1, 1]，其中 -1 表示完全负相关，1 表示完全正相关，0 表示无关。相关系数的绝对值越大，说明两个变量之间的相关程度越高。热力图中，蓝色表示正相关，红色表示负相关，颜色的深浅代表相关程度的大小。

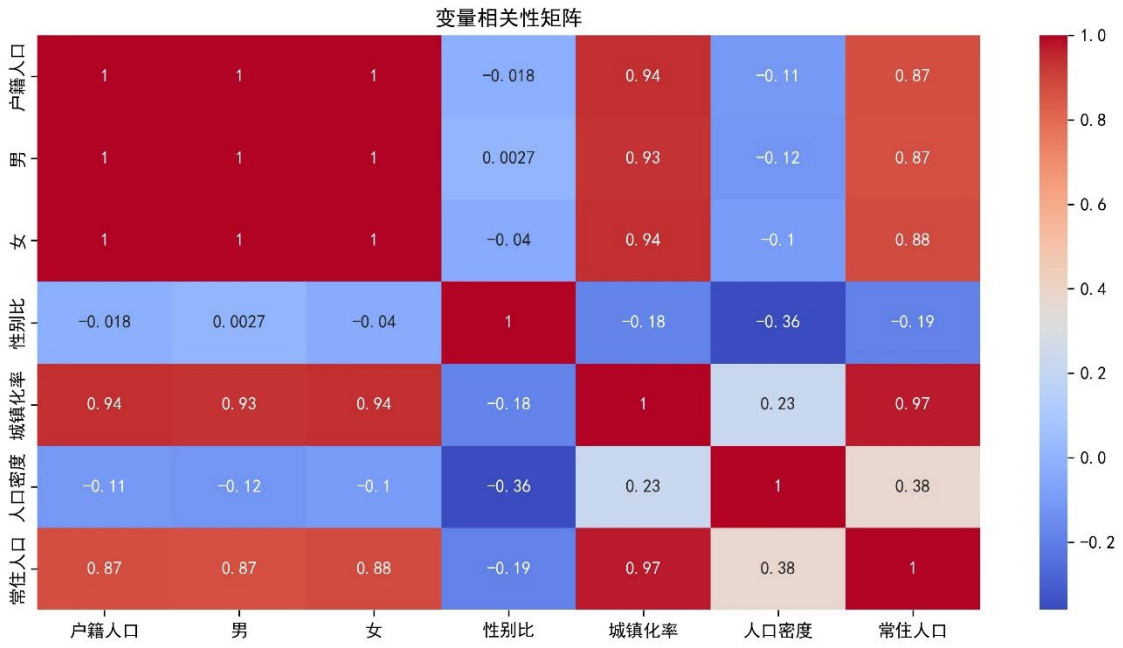

图 6-1　人口预测相关性矩阵热力图

通过观察相关性矩阵，我们可以发现与目标变量相关性较高的特征变量。其中，与目标变量总人口（使用常住人口的口径）呈正相关关系的变量有城镇化率、户籍人口数、男性人数、女性人数、人口密度，城镇化率相关性最高为 0.97，人口密度最低为 0.38，性别比与常住人口呈负相关关系，相关系数为-0.19，相关性较弱。由于人口密度和性别比与目标变量相关性较低，可解释性较低，要用于 LSTM 模型的构建并进行人口预测需要剔除人口密度和性别比这两个特征变量，最后保留 6 个特征变量，它们这些变量在后续的模型训练中具有较高的可解释性和较高的预测价值。

3. 创建模型

在数据预处理和相关性分析的基础上，我们开始构建基于长短时记忆神经网络（LSTM）的河南省人口预测模型。本研究的模型构建主要包括以下几个步骤：

（1） 划分训练集和测试集：为了评估模型在未知数据上的泛化能力，我们将预处理后的数据集按照 8:2 的比例划分为训练集和测试集。训练集用于模型的训练，而测试集用于评估模型的预测性能。在本研究中，我们采用了前 80% 的数据作为训练集，后 20% 的数据作为测试集。

（2） 构建训练集：为了便于 LSTM 模型的训练，我们需要将训练数据和测试数据转换为特定的格式。本研究中，我们采用了滑动窗口法来构建训练集，即使用前 look_back 个时刻的数据来预测下一个时刻的常住人口。look_back 参数代表滑动窗口的大小，其值可以根据实际情况进行调整。在本研究中，我们设置 look_back 为 2，表示使用前两年的数据来预测下一年的人口。

（3）创建 LSTM 模型：我们搭建了一个包含两层 LSTM 和两层 Dropout 层的神经网络。模型结构如下：

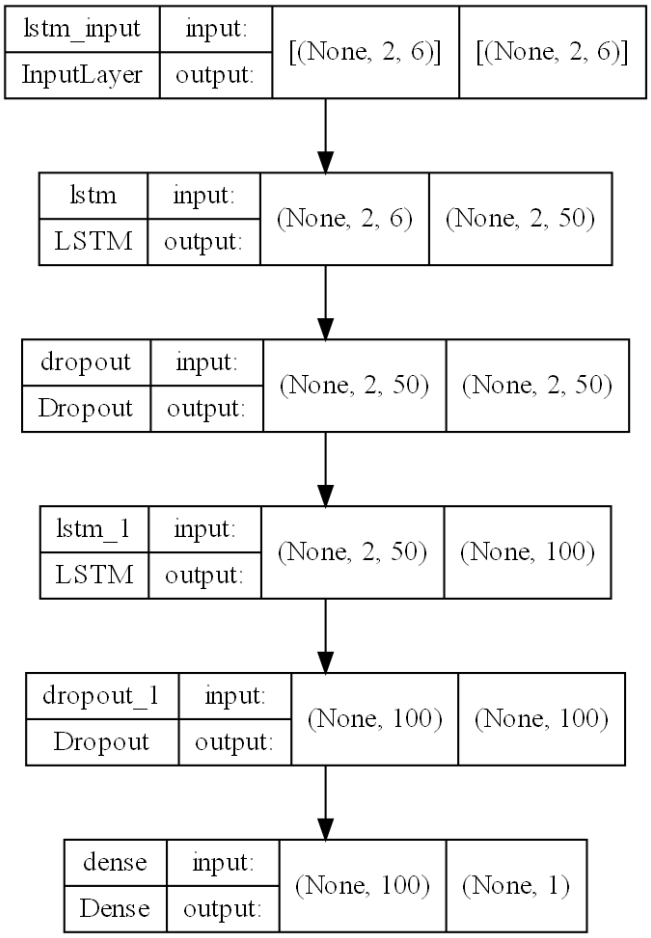

图 6-2　河南省总人口预测 LSTM 模型结构图

第一，输入层（InputLayer）：将形状为（None，2，6）的输入数据传递给第一层 LSTM。在这里，″None″代表批量大小（batch_size），可以在训练时根据需要指定；″2″代表滑动窗口大小（look_back），本例中表示使用前两年的数据来预测下一年的人口；″6″代表输入特征数量。输入层的输出形状同样为（None，2，6）。

第二，第一层 LSTM（LSTM_50）：包含 50 个神经元，接收来自输入层的数据。输入形状为（None，2，6），输出形状为（None，2，50），其中″50″代表第一层 LSTM 的神经元数量。

第三，第一层 Dropout（Dropout_1）：设置 20% 的 Dropout 率，用于防止模型过拟合。输入形状为（None，2，50），输出形状为（None，2，50）。

第四，第二层 LSTM（LSTM_100）：包含 100 个神经元。输入形状为（None，2，50），输出形状为（None，100），其中″100″代表第二层 LSTM 的神经元数量。注意，第二层 LSTM 的输出形状不再包含滑动窗口维度（look_back），因为该层的输出是为全连接层准备的，需要将时间步展平。

第五，第二层 Dropout（Dropout_2）：设置 20% 的 Dropout 率，用于防止模型过拟合。输入形状为（None，100），输出形状为（None，100）。

第六，输出层（Dense_1）：全连接层，用于输出预测的总人口（使用的是常住人口作为口径）。输入形状为（None，100），输出形状为（None，1），其中″1″代表输出的预测结果维度，即预测的人口数量。

（4） 编译和训练模型：在搭建好网络结构之后，我们需要对模型进行编译。本研究中，我们选择了均方误差（mean_squared_error）作为损失函数，采用了自适应矩估计（Adam）作为优化器。然后，我们将处理好的训练数据输入到模型中，进行训练。本研究中，我们设置了 100 个训练轮次（epochs）和 1 的批量大小（batch_size）。

通过以上步骤，我们成功地构建了一个基于 LSTM 的河南省人口预测模型。在模型训练完成后，我们对其预测性能进行了评估，并将预测结果与实际人口数据进行了对比。

（5） 反标准化预测值：我们首先对原始数据进行了预处理，包括将年份转换为时间格式、设为索引、删除含空值的行等操作。然后，我们利用 MinMaxScaler 对数据进行了标准化处理，将数据值缩放到［0，1］区间内，以便提高模型的训练效率和收敛速度。在完成模型训练和预测后，我们需要将预测值从标准化后的值还原为原始值，以便与实际数据进行对比分析。这一过程称为“反标准化”。

为了进行反标准化，我们将预测结果与对应的输入特征拼接在一起，形成一个完整的数据矩阵，然后使用 MinMaxScaler 的 inverse_transform 方法将其还原为原始数据范围。这样，我们便得到了训练集和测试集上的反标准化预测值，可以与实际人口数据进行对比。

（6）调节参数：根据训练结果手动调节了一些参数，以寻找最佳的训练策略。以下是尝试过的一些参数组合及相应的结果：

第一， batch_size=5, look_back=2, epoch=150：训练集上的 MAE 为 26.68，测试集上的 MAE 为 93.52。

第二， batch_size=4, look_back=2, epoch=150：训练集上的 MAE 为 28.82，测试集上的 MAE 为 35.66。

第三， batch_size=3, look_back=2, epoch=150：训练集上的 MAE 为 20.06，测试集上的 MAE 为 64.94。

第四， batch_size=4, look_back=2, epoch=155：训练集上的 MAE 为 19.50，测试集上的 MAE 为 60.48。

根据这些结果，我们选择了参数组合：batch_size=4, look_back=2, epoch=150，因为它在测试集上的平均绝对误差（MAE）最低，达到了 35.66。这表明，经过多次手动调整参数和训练策略，我们构建的基于 LSTM 的深度学习模型在河南省人口预测任务上取得了较好的效果。

（6）真实值与预测值对比：我们将数据集划分为训练集和测试集，并使用滑动窗口法构建训练集和测试集。模型的训练过程中，我们设置了 look_back 参数为 2，意味着模型将使用前两年的数据来预测下一年的人口。在模型训练过程中，我们通过损失函数曲线图（如图 6-4）观察到模型的损失值逐渐减小并收敛到一个较稳定的水平。根据损失函数曲线图，我们选择训练 150 轮以使模型获得较好的收敛效果。经过训练后，我们得到了训练集上的平均绝对误差（MAE）为 28.82，测试集上的平均绝对误差（MAE）为 35.66。这些结果表明，LSTM 模型具有较好的预测能力。为了进一步评估模型性能，我们将预测值与实际值进行对比，并绘制了折线图（如图 6-3 ）。从对比折线图中，我们可以观察到预测值和真实值在整个时间序列上的走势。整体上，LSTM 模型的预测值能够较好地反映真实值的变化趋势。这表明，经过调整参数和训练策略，我们构建的基于 LSTM 的深度学习模型在河南省人口预测任务上取得了令人满意的效果。

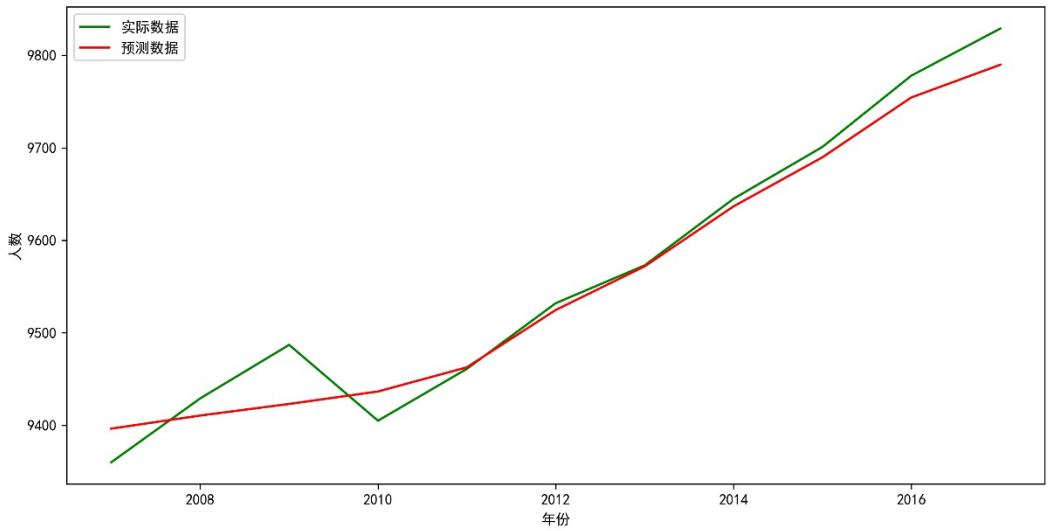

图 6-3　LSTM 河南省人口预测对比折线图

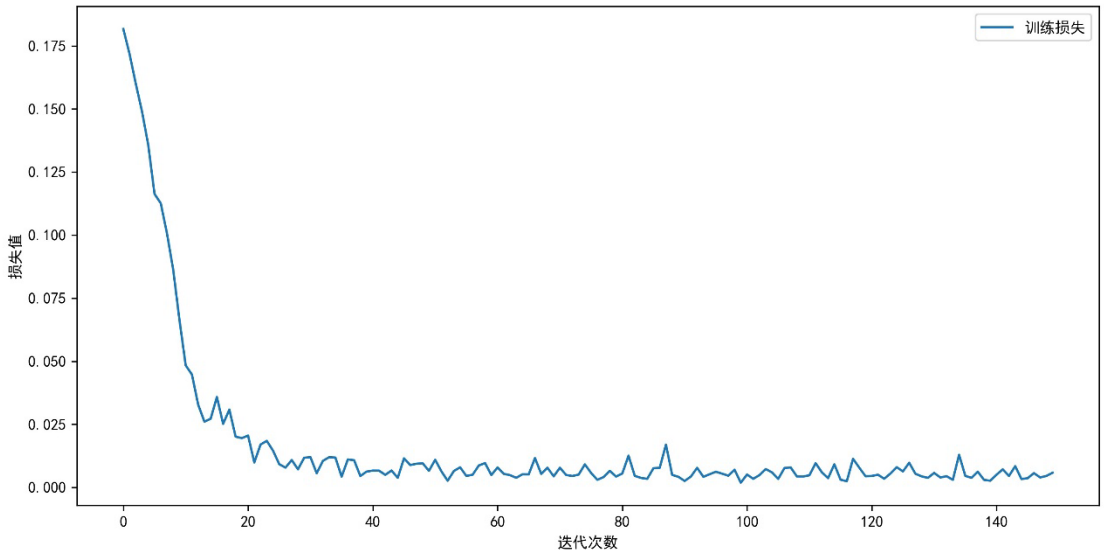

图 6-4　损失曲线图

（二）基于 LSTM 模型人口的预测

根据我们基于长短时记忆网络（LSTM）模型的预测结果，从 2022 年至 2031 年间，河南省的常住人口呈现出波动变化的趋势。具体来说，预测数据如下：

表 6-1 未来十年人口预测表

年份	总人口（万人）
2022	9837.763
2023	9791.276
2024	9758.769
2025	9754.108
2026	9774.08
2027	9837.784
2028	9794.806
2029	9775.141
2030	9758.103
2031	9747.953

从预测折线图（如图 6-5 ）可以观察到，河南省未来十年的常住人口变化表现出一定的波动性。在 2023 年和 2025 年，人口出现了略微的下降，而在 2027 年出现了一次人口峰值，之后又呈现下降趋势。整体而言，从 2022 年到 2031 年，河南省的常住人口预计将从 9837.763 万人下降至 9747.953 万人，减少约 89.81 万人。这表明在未来十年里，河南省的人口规模将缓慢下降。

需要注意的是，这些预测结果基于 LSTM 模型，并受到模型假设和训练数据的限制。实际的人口变化可能会受到政策、经济、社会和自然因素等多种因素的影响。因此，在解读和应用这些预测结果时，应谨慎对待并考虑其他相关因素。

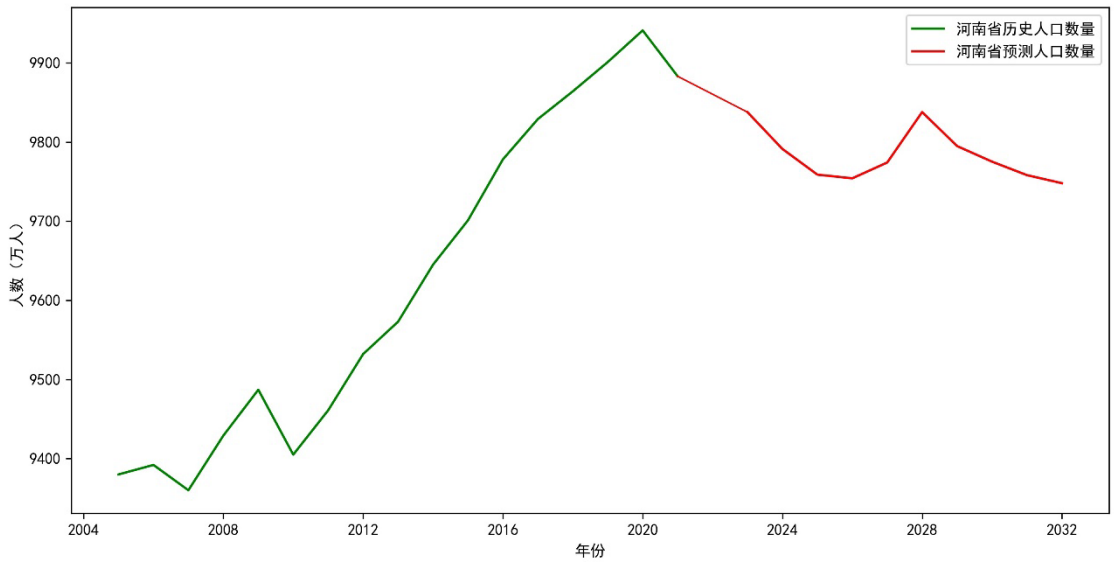

图 6-5 河南省未来十年总人口预测折线图

结论

1. 结论

在研究河南省未来人口变化的预测中，本文尝试了多种模型来评估预测效果。首先，我们使用了基于马尔萨斯模型的指数预测方法，其平均绝对误差（MAE）为 53.31。其次，我们采用了 Logistic 模型进行预测，得到的平均绝对误差为 53.31，接着，我们对 Logistic 模型进行了改进，得到的平均绝对误差降低至 42.90，最后，我们采用了基于长短时记忆网络（LSTM）的预测方法，得到的平均绝对误差为 28.82（训练集）和 35.66（测试集）。

根据平均绝对误差指标，我们可以看出 LSTM 模型的预测效果优于其他三个模型。LSTM 模型是一种基于深度学习的时间序列预测方法，可以捕捉数据中的复杂模式和非线性关系，因此在人口预测问题上表现出较好的性能。相比之下，马尔萨斯模型和 Logistic 模型是基于简单的数学公式进行预测的，可能无法充分捕捉到数据中的变化规律。

综上，通过对比分析四种预测模型，我们可以得出结论：基于长短时记忆网络（LSTM）的预测方法在预测河南省未来人口变化上具有较好的性能和准确性。在未来研究中，我们可以继续优化和改进 LSTM 模型，以提高预测准确度和稳定性。

2. 建议

BP 神经网络模型在人口预测方面具有更高的准确性和可靠性，相比传统的微分方程模型，BP 神经网络模型更适合用于河南省总人口预测。此外，本研究还发现河南省未来的人口趋势将呈现出缓慢减少的态势。因此，我们建议政府应该加强对人口问题的关注，并制定相应的政策来优化人才结构和产业结构，促进河南省经济发展。同时，我们也建议在未来的研究中进一步探索 BP 神经网络模型在其他领域中的应用。

致　谢

在论文完成之际，首先，我要感谢×××领我走进了机器学习的大门，从此便对机器学习、深度学习和人工智能产生了浓厚的兴趣。尤其是，在今年 OPEN AI 公司公布了他们的大型自然语言模型 GPT4，我仿佛看到了一种趋势，大模型的趋势还有人工智能再一次兴起的趋势。这意味着我们未来的工作还有生活都会发生翻天覆地的巨大变化，而我们这一代人恰逢这个十字路口，尽管学过一些机器学习和深度学习的知识，虽然知道 GPT4 用了什么原理，不会像普通人见到魔法一样，但是还是惊叹于科技发展和人工智领域迭代的速度，我们以后的学习思维也要发生转换，思考学什么重要，以及怎么学，尤其是在这个信息和知识爆炸的年代。但是这些深度学习领域的奇迹更多诞生于美国，随便举一个深度学习模型的例子都是外国人的名字，而且 ChatGPT 推出后还限制中国人使用，钱学森先生曾说，外国人能搞的东西，我们中国人也能搞。每每想到便觉遗憾和气愤，虽然我没有什么天赋和能耐，但是我希望能以后继续深入学习这个领域，如果未来努力达到了，并且运气好，希望能在这个领域深耕。

当然除了对这个领域的好奇外，还有就是从×××身上学到了严谨治学和待人和善的品质。最后还要衷心感谢帮助过我的老师和同学，大学结束意味着人生新的起点的开始，望未来可期。

参考文献

[1] 贾楠. 基于某些人工神经网络的人口预测的研究[D]. 中北大学, 2012.

[2] 刘晓峰. 河南省未来人口发展趋势[J]. 河南教育学院学报(哲学社会科学版), 2007(04):43-48.

[3] 周美旭. 基于灰色 GA-BP 神经网络的江西人口预测[D]. 景德镇陶瓷学院, 2015.

[4] 陆文珺, 柳炳祥. 一种基于 BP 神经网络的人口总数预测方法[J]. 中国管理信息化, 2016, 19(20):144-145.

[5] Gerland P, Raftery A E, Ševčíková H, et al. World population stabilization unlikely this century[J]. Science, 2014, 346(6206): 234-237.

[6] Raftery A E, Alkema L, Gerland P. Bayesian population projections for the United Nations[J]. Statistical science: a review journal of the Institute of Mathematical Statistics, 2014, 29(1): 58.

[7] Şahinarslan F V, Tekin A T, Çebi F. Application of machine learning algorithms for population forecasting[J]. International Journal of Data Science, 2021, 6(4): 257-270.

[8] http://lwzb.stats.gov.cn/pub/lwzb/ekp/jcbjzs/tjjczs/201707/t20170717_3860.html

[9] 李鹏. 2021 年河南人口发展报告公布[N]. 河南日报, 2022.

[10] 马晓星. 基于 BP 神经网络的人口普查收入预测[J]. 现代计算机, 2021(04):38-41.

[11] 罗万春. 基于 BP 神经网络的重庆市人口预测[J]. 黑龙江科学, 2022.

[12] 刘天麒. 基于 BP 神经网络的广东省第三产业就业人口数量预测研究[J]. 无线互联科技, 2017(19):136-138.

[13] 孙文渊. 基于 BP 神经网络模型下预测吉林省 GDP[D]. 延边大学, 2015.

[14] 李国成, 吴涛, 徐沈. 灰色人工神经网络人口总量预测模型及应用[J]. 计算机工程与应用, 2009, 45(16):215-218.

[15] 丁惠敏. 基于循环神经网络的人口死亡率预测研究[D]. 南开大学, 2022.

[16] 周博军, 王旺, 黄俊达, 陈超凡, 陈震, 邓玉新, 马福海. 基于 BP 神经网络对中国体育彩票销售金额的预测[J]. 体育教育学刊, 2022.