

基于LSTM神经网络的人口预测研究

朱挺欣

江西科技学院理学教学部, 江西 南昌

收稿日期: 2026年1月6日; 录用日期: 2026年1月28日; 发布日期: 2026年2月9日

摘要

本文对人口数据进行整理, 探究全国总人口数与相关影响因素之间的分布关系。我国人口数量的影响因素有自然资源、社会环境、其他因素(男女占比、人口老龄化问题、相关政策等)。本文提出LSTM神经网络预测模型, 以我国人口统计数据为研究样本, 整合多项关键指标, 构建LSTM预测模型, 并将该模型与传统灰色预测GM(1, 1)模型、阻滞增长模型开展对比实验。实验结果表明, 所构建的LSTM神经网络模型预测精度更高, 其结果通过预测2020年我国总人口数据与已知2020年我国总人口数据进行对比得到, 同时预测2030年我国总人口数据, 该研究成果为人口动态预测提供了可靠的技术方法。

关键词

人口预测, LSTM神经网络, 灰色预测GM(1, 1)模型, 阻滞增长模型

Research on Population Prediction Based on LSTM Neural Network

Tingxin Zhu

School of Sciences, Jiangxi University of Technology, Nanchang Jiangxi

Received: January 6, 2026; accepted: January 28, 2026; published: February 9, 2026

Abstract

This paper organizes population data to investigate the distributional relationships between the total national population and its relevant influencing factors. The factors affecting China's population size include natural resources, the social environment, and other elements (such as the gender ratio, population aging, and relevant policies). This paper proposes an LSTM neural network-based prediction model, utilizing China's population statistics as the research sample. By integrating multiple key indicators, an LSTM prediction model is constructed, and comparative experiments are conducted between this model and traditional forecasting approaches, namely the Grey Prediction GM(1, 1) model

and the Logistic Growth Model. The experimental results indicate that the constructed LSTM neural network model achieves higher prediction accuracy. This conclusion is validated by comparing the model's predicted total population data for China in 2020 with the known actual data for the same year. Furthermore, the model is used to forecast China's total population for 2030. The findings of this study provide a reliable technical approach for dynamic population prediction.

Keywords

Population Prediction, LSTM Neural Network, Grey Prediction GM(1, 1) Model, Logistic Growth Model

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,我国人口发展呈现老龄化等特征,出生人口变动、相关政策(如 1982 年将计划生育政策定为我国的基本国策、2016 年 1 月 1 日我国正式实施两孩生育政策、2021 年 5 月 31 日正式实施三孩生育政策)等因素直接影响人口。人口预测有助于调控人口发展节奏,精准的人口预测是科学调控人口发展节奏、实现人口长期均衡发展的重要支撑,其价值不仅体现在对人口规模变动趋势的前瞻性研判,更在于为国家发展规划制定、区域就业布局优化提供精准的数据锚点。人口总量的增减波动与结构演化,深刻关联劳动力供给质量、公共服务资源配置效率,直接影响社会发展战略的顶层设计方向,更对国民经济调控的重点领域与政策力度产生系统性影响[1]。从实践来看,基于人口预测成果,可提前预判劳动年龄人口变动趋势以优化产业结构布局,预判老龄化进程以完善养老保障体系,预判区域人口流动方向以统筹公共服务资源配置,从而实现人口与经济社会、资源环境的协调可持续发展[2]。目前我国处于全面建设社会主义现代化强国的关键阶段,人口作为关系中华民族伟大复兴的重要因素,其发展趋势直接影响现代化建设的质量与进程。尽管我国经济社会发展取得显著成就,但人口发展已从增量阶段转向减量阶段,面临着老龄化程度持续加深、人口基数庞大且结构严重失衡等多重挑战,这些问题不仅对社会保障体系、劳动力供给、经济活力提升形成持续压力,更对人口与经济社会协调可持续发展提出了严峻考验。数据显示,我国 65 岁及以上人口占比已进入中度老龄化区间,2024 年占比达 15.66%,与 0~14 岁人口占比(15.81%)已非常接近,预计“十五五”时期将进一步超过少儿人口占比,迈向重度老龄化阶段人口总量负增长的长期趋势已逐步显现。因此开展人口相关研究,深入剖析人口数量、结构、素质、分布的动态变化规律,精准预测人口问题对经济发展、社会治理、公共服务等领域的深远影响,不仅能够为落实积极应对人口老龄化国家战略、优化生育支持政策体系提供坚实的理论和数据支撑,更能为以高质量发展支撑中国式现代化提供科学的决策依据[3]。精准的人口预测能够更好地提升公共资源配置效率,推动人口与经济社会协调发展,因此成为人口学与数据科学交叉领域的研究焦点[4]。

目前人口预测技术主要划分为传统统计方法与机器学习方法两大类。其中传统技术以统计学为基础,可解释性强、数据需求低,适用于基础性的人口预测,主要包括通过线性、指数等函数拟合历史趋势外推的趋势外推法(长期易受政策等因素影响)和按年龄性别分队列预测死亡、生育等事件的队列要素法(能精准刻画人口结构,但对数据质量要求高),现代数据驱动技术则可整合多源数据、捕捉非线性特征以提升精度,可融合多因素以及通过多层神经网络捕捉深层特征、适配长时序数据与多因素交互场景

的深度学习模型。LSTM 神经网络作为循环神经网络(RNN)的改进形态,通过引入门控机制能够有效克服传统 RNN 的长期依赖问题,在时序数据预测领域呈现出独特的应用优势[5]。基于此,本文以全国人口预测为核心目标,构建 LSTM 预测模型,整合多维度人口影响因子,通过对比实验验证模型的有效性,为人口发展规划提供精准、可靠的预测技术方案。

2. 相关方法概述

2.1. 灰色预测 GM(1, 1)模型

本文选取传统灰色预测 GM(1, 1)模型[6]作为对照模型,该模型为人口预测领域的经典技术方案。设 $X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\}$ 为原始序列,对其进行一次累加得到:

$$X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n)\}$$

其中 $x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i) (k=1, 2, \dots, n)$, 称 $X^{(1)}(k)$ 为 $X^{(0)}(k)$ 的第一次累加。

可得 $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$ 为白化微分方程, $x^{(0)}(k) + az^{(1)}(k) = b$ 为差分, a 表示发展因子, b 表示灰色因子。由最小二乘原理可得:

$$[a, b]^T = (B^T B)^{-1} B^T Y_n$$

其中:

$$Y_n = [x^{(0)}(2), x^{(0)}(3), \dots, x^{(0)}(n)]^T$$

$$B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix}$$

其中, $z^{(1)}(k+1) = \frac{1}{2} [x^{(1)}(k) + x^{(1)}(k+1)], k=1, 2, \dots, n-1$

方程 $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$ 的解为:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a} \right) \cdot e^{-ak} + \frac{b}{a}$$

还原值为:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) = (1 - e^a) \left(x^{(0)}(1) - \frac{b}{a} \right) \cdot e^{-ak}$$

其中 $k=1, 2, \dots, n-1$ 。用最小二乘法求解灰参数 \hat{a} , 则 $\hat{a} = (B^T B)^{-1} B^T Y_n$ 将灰参数 \hat{a} 代入 $\frac{dx^{(1)}}{dt} + ax^{(1)} = u$, 并对其进行求解。将上述结果累减还原, 即可得到预测值。

2.2. 阻滞增长模型

阻滞增长模型[7]由于资源、环境等因素对人口增长的阻滞作用,当人口增长到一定数量后,增长率会下降,假设人口的增长率为 x 的减函数,如设 $r(x) = r(1 - x/x_m)$, 其中 r 为固有增长率(x)很小时, x_m

为人口容量, 可得如下微分方程:

$$\begin{cases} \frac{dx}{dt} = rx \left(1 - \frac{x}{x_m} \right) \\ x(0) = x_0 \end{cases}$$

阻滞作用在对人口增长率 r 的影响上, 使得 r 随着人口数 x 的增加而下降。若将 r 表示为 x 的函数 $r(x)$, $r(x) = r - sx (r > 0, s > 0)$ 这里 r 称固有增长率, 表示人口很少时的增长率。为确定系数 s , 引入自然资源和环境条件所能容纳的最大人口数据 X_m , 称人口容量, 当 $x = X_m$ 时人口不再增长, 即此时增长率为 0。

2.3. LSTM 神经网络模型

LSTM 神经网络模型, 优点在于通过门控机制解决传统问题, 适配人口时序数据的长周期演化特征 [5]。其主要结构由输入门、遗忘门和输出门三大门控单元及细胞状态组成, 本文采用的 LSTM 神经网络预测的核心原理是借助输入门、遗忘门与输出门三大门控单元构建记忆机制, 精准捕捉人口时序数据的长期依赖关系, 通过筛选并保留历史数据中的关键信息(如历年出生率、死亡率等多维度特征)、刺激人口政策, 实现对人口变化趋势的精准拟合与预测。

我们采用 MATLAB 软件进行求解, 首先整理 1953~2023 年全国总人口及多维度影响因素(人均 GDP、育龄妇女占比、生育政策量化指标等)的数据, 其中育龄妇女占比是育龄妇女数量与总人口数量的比值, 国际通用育龄妇女年龄范围: 15~49 周岁 [8]。直接采用 1953~2023 年的普查人口数据 [6], 直接从普查公报中提取分年龄、分性别的人口数据, 汇总 15~49 岁女性人数。政策量化通过政策是否实施转化为离散数值变量, 明确政策实施的关键时间节点, 如二孩政策以 2016 年为界, 三孩政策以 2021 年为界。其次, 针对年度时间序列数据赋值, 对每个年份, 政策实施前赋值为 0, 代表政策未落地、无相关影响; 政策实施当年及之后年份赋值为 1, 代表政策已实施。

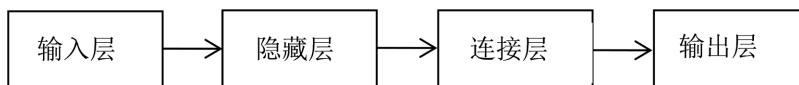
经数据清洗、min-max 归一化消除量纲差异后, 以时间步长 5 构建“连续 5 年特征数据 - 第 6 年总人口”的监督学习样本, 划分训练集与测试集。

min-max 归一化公式:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

其中, X 为原始数据, X' 为归一化后的数据(取值范围[0, 1]), X_{\max}, X_{\min} 分别为原始数据序列的最大值和最小值。

用 MATLAB 的 lstmLayer、fullyConnectedLayer 等函数搭建模型, 网络结构从输入到输出依次为:



其中, 输入层(输入维度 = (5, 6), 对应时间步长 = 5、特征数 = 6); LSTM 隐藏层(双层结构, 第一层 128 个神经元, 第二层 64 个神经元, 采用 tanh 激活函数, 设置 dropout 率 = 0.3 抑制过拟合); 全连接层(两层结构, 分别含 128、64 个神经元, 采用 relu 激活函数实现特征整合); 输出层(1 个神经元, 线性激活函数, 输出归一化后的总人口预测值)。

模型训练参数配置: 选用 Adam 优化器(初始学习率 = 0.0005), 损失函数设为均方误差, 迭代次数取 200, 批次大小取 8, 通过 earlyStoppingCallback 函数引入早停机制, 当测试集损失连续 20 轮无下降时终

止训练，保障模型泛化能力。

将特征数据(如 2025~2029 年)整理为维度为(5, 6)的输入序列，采用归一化逆运算公式

$$X = X' \times (X_{\max} - X_{\min}) + X_{\min}$$

其中， X 为最终还原后的实际数值， X' 为 LSTM 模型输出的归一化预测值(取值范围[0, 1])， X_{\max} 为原始数据序列中的最大值， X_{\min} 为原始数据序列中的最小值。

3. 模型构建与实现

3.1. 数据处理

本文选取七次人口普查中的全国人口进行数据处理七次人口普查的男性人数、女性人数、出生人数、死亡人数整理如下表 1 所示：

Table 1. Data from the seven national population censuses of China

表 1. 我国七次人口普查数据

年份	总人数(万人)	男性人数(万人)	女性人数(万人)	出生率(%)	死亡率(%)
1953	60,194	2431	2457	37	14
1964	72,037	35,652	33,806	30	10
1982	103,188	51,943	48,874	20.91	6.36
1990	116,002	58,495	54,873	21.06	6.67
2000	124,261	64,028	60,234	14.03	6.45
2010	133,281	68,685	65,287	11.90	7.11
2020	141,178	72,334	68,844	8.52	7.07

3.2. 实验结果

(1) 误差分析

分别采用灰色预测 GM(1, 1)模型、阻滞型模型与 LSTM 神经网络模型进行训练，开展预测实验，预测 2020 年全国总人口数与真实数据进行对比，寻找最佳预测方法，三种模型的预测结果见表 2 所示。

Table 2. Predicted results of the national total population for 2020

表 2. 2020 年全国总人口预测结果

模型	灰色 GM(1, 1)预测模型	阻滞型人口预测模型	LSTM 神经网络预测模型	真实数据
总人数(万人)	141,910	138,500	140,895	141,177

由表 2 数据可知，LSTM 神经网络预测模型在评价指标上均优于两种对照模型。这一结果表明，LSTM 模型能够更精准地捕捉人口数据的时间依赖关系与多因素耦合特征，预测精度得到提升。

(2) 预测 2030 年全国人口总数

三种模型的预测结果见表 3 所示：

Table 3. Predicted results of the national total population for 2030

表 3. 2030 年全国总人口预测结果

模型	灰色 GM(1, 1)预测模型	阻滞型人口预测模型	LSTM 神经网络预测模型
总人数(万人)	150,656	132,667	130,073

LSTM 神经网络预测模型与传统灰色预测 GM(1, 1)模型、阻滞增长模型开展对比实验。实验结果表明, 所构建的 LSTM 神经网络模型预测精度更高、误差较小, 其结果通过预测 2020 年我国总人口数据与已知 2020 年我国总人口数据进行对比得到。这一对比结果清晰说明, 该模型对人口数据的时间关联特性以及多维度影响因素有更强的预测能力, 可大幅度提高预测的精度, 并用于 2030 年全国总人数预测。

4. 结论

本文采用 LSTM 神经网络模型探究了育龄妇女占比及生育政策等因素对总人口变化的影响, 并与传统灰色预测 GM(1, 1)模型、阻滞增长模型进行对比, 同时完成对 2030 年全国总人口的预测。对比实验结果表明, 所构建的 LSTM 神经网络模型预测精度更高、误差更小。人口负增长是我国目前及未来一段时期的客观发展趋势, 该模型得出的人口预测结果与当前我国人口负增长、老龄化加剧等社会国情契合, 为后续社会发展调整提供了更多的数据支撑。但 2030 年总人口预测值依然存在局限性, 如难以模拟多因素协同作用下的人口变化过程、未能充分捕捉政策对生育意愿的直接作用、时间步长设计未充分考虑人口政策的滞后效应、模型对历史数据中人口结构失衡惯性的过度拟合加剧了预测值的下行偏差。可为后续开展更高精度的人口预测研究提供参考, 结合预测趋势与国情, 建议完善育儿服务与养老服务, 减轻家庭生育与养老负担, 同时稳妥实施渐进式延迟退休政策, 激活老年人力资源潜能, 替代传统人口红利, 努力培育积极婚育文化, 提供婚育保障, 助力生育水平稳定在适度区间。

参考文献

- [1] 中华人民共和国国家发展和改革委员会. 国家人口发展规划(2016-2030 年) [Z]. 2016.
- [2] 国务院. 国家人口发展规划(2016-2030 年) [Z/OL].
https://www.gov.cn/zhengce/content/2017-01/25/content_5163309.htm, 2017-01-25.
- [3] 王广州. 中国人口预测方法的探索与思考[J]. 中国人口科学, 2020(2): 2-13.
- [4] 中共中央 国务院. 关于优化生育政策促进人口长期均衡发展的决定[Z]. 2021.
- [5] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780.
<https://doi.org/10.1162/neco.1997.9.8.1735>
- [6] 国家统计局. 中国统计年鉴[M]. 北京: 中国统计出版社, 2023.
- [7] 姜启源. 数学模型[M]. 北京: 高等教育出版社, 1987.
- [8] 《人口研究》编辑部. 人口统计指标规范[J]. 人口研究, 1982, 6(4): 55-58.