基于 ARIMA 和 GBDT 的人口预测组合模型

摘 要

准确预测未来人口数量具有重要现实意义。本文针对人口中长期预测影响因素较复杂、可用历史数据较少、单一模型局限性等特点,构建了自回归综合移动平均模型（ARIMA）和梯度提升树回归模型（GBDT）组合预测方法。该模型将 ARIMA 模型和梯度提升树回归模型进行结合,充分发挥 ARIMA 模型对时间序列的精准预测性能，和集成学习模型对多元影响因素的准确提取、利用能力。我们使用该组合模型对 2005 后的中国人口数量进行了中短期和长期的预测，结果表明:与单一模型相比较,组合模型预测精度更高,相对误差低, 且预测结果更加稳定。

关键词： ARIMA 模型; 梯度提升树回归模型; 组合预测

一、 问题重述

1.1 问题背景

人口预测数据是国家制定人口、经济和社会发展等宏观发展战略规划中的最基础数据。随着近年来人口生育政策适度调整以及中国人口结构变动等因素的影响，中国的人口发展增速和结构各方面都变得越来越复杂，同时，人口与资源环境的关系、人口老龄化等问题都在不断深化，关于如何对未来人口变动趋势做出准确判断，已成为当下中国关注的重要问题。

1.2 问题回顾

为建立高准确率的人口预测模型，本文将主要解决以下两个问题：

(1)问题一：根据附录 1 和收集的数据，归纳中国人口特点，并将其划分为影响人口增长的直接因素和间接因素；

(2)问题二：分析普通人口增长模型的缺点与不足，并根据附录 2 和收集的数据，建立合适的人口预测模型对中国人口进行中短期和长期预测。

二、 问题分析

2.1 数据收集与处理

为更好地实现对人口影响因素的分析和建立更为精确的预测模型，我们在充分使用附录 1、2 的基础上，又收集了关于历年城镇化率、人口素质信息等其他数据。我们的数据来自国家统计局、联合国数据库(UN data)等相关网站，具体请见附录。

由于各类数据单位不同、数量级跨度较大，在使用之前我们对数据进行了最大值—最小值归一化操作，公式如下：

$$X = \frac{X - Xmin}{Xmax - Xmin}$$

2.2 问题一的分析

对于问题一，我们首先提取附件 1 中信息，将中国人口特点归纳为六大类，并对各个特点对于中国人口增长的影响进行描述性分析；然后我们使用收集到的衡量各指标的数据集，通过数据驱动的方法，使用 Spareman 系数和 SHAP 模型具体分析各指标的影响方向和程度。

2.3 问题二的分析

对于问题二，我们基于问题一中分析结果和 2005 年之前的中国人口数据，选取对中国人

口增长影响程度较为显著的因素作为额外输入，建立 ARIMA—GBDT 组合模型，对 2005 年后的人口数量进行中短期和长期预测，并对预测结果进行评估。

三、 模型假设

为简化具体问题、方便模型建立，我们对模型做出以下合理假设：

（1）人口增长连续性：假设过去几十年的人口增长趋势将在未来继续，即：过去的人口数据可以反映未来的人口增长趋势与规律。

（2）出生率和死亡率的稳定性：假设未来的出生率和死亡率不发生突变或极端波动，按照已知的趋势缓慢变化。

（3）移民流动的一致性：假设未来的国际移民流入和流出人数将保持当前的趋势，不会有突然的大规模移民潮[i]。

（4）政策因素的稳定性：假设计划生育政策和其他有关人口控制的政策将持续执行，且效果不会发生剧烈变化。

（5）社会经济因素的稳定性：假设社会经济因素如教育水平、城镇化率等对人口增长的影响将继续按照当前的模式发展。

（6）自然灾害和疫情的忽略：模型中不特别考虑如大规模自然灾害或疫情等可能突发的事件，假设它们对总体人口趋势的影响是有限的。

（7）数据的准确性和完整性：由于数据来自官方网站，我们有理由假设用于模型建立的历史数据是准确和完整的，并且适合用于统计分析和模型预测。

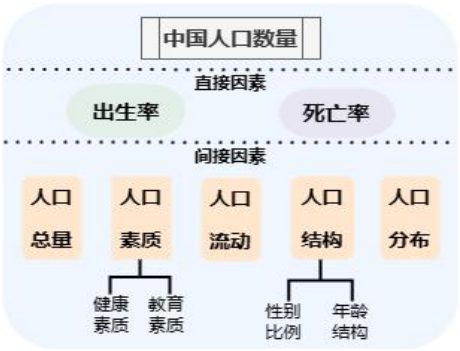
四、 符号说明

符号	说明	符号	说明
ADF Test	单位根检验	d	差分阶数
X_squared	卡方拟合度检验	q	移动平均项阶数
P_value	P 值	p	自回归项阶数
ACF	自相关函数	PACF	偏自相关函数

五、 问题一模型建立与求解

5.1 描述性分析

通过对附件 1 的总结，我们将影响中国人口增长的人口特点归纳为以下几类：



<1>生育率：生育率经历了从高到低的变化，目前低于更替水平。生育率直接决定了人口增长的速度。高生育率通常导致快速的人口增长，而低生育率则可能导致人口增长放缓甚至下降。

<2>死亡率：当前中国人口自然死亡率持续保持低位水平。死亡率直接影响人口增长，近年来中国自然死亡率保持稳定低位，对人口增长的影响较为稳定。

<3>人口总量：中国人口总量庞大，并在经历高速增长后进入增速趋缓时期，即将达到峰值。人口总量本身不直接决定增长速度，但是大规模人口可能对资源和环境造成压力，这间接影响生育政策和人们的生育选择，从而影响人口增长。

<4>人口素质：人口素质可大致分为健康素质和教育素质，见年来我国人口素质显著提高。人口素质将对人口增长产生较为复杂的间接影响：一方面，提高的教育水平通常会导致更低的生育率，因为受过良好教育的人倾向于推迟生育和减少子女数量。另一方面，高素质人口也能更好地适应和推动社会经济发展，间接影响人口结构和增长趋势。总体来说，高人口素质社会趋向于低人口增长。

<5>人口结构：包括年龄结构和性别比例，中国面临人口老龄化的挑战，以及出生人口性别比长期失衡的局面。年龄结构和性别比例间接影响人口增长，老龄化社会的生育率通常较低且自然死亡率较高；而失衡的性别比可能导致婚配困难，进而降低生育率和放缓人口增长。

<6>人口流动与分布：地区及城乡人口、劳动力分布不均，流动迁移人口规模庞大。人口的流动和分布间接地影响了人口增长和经济社会的发展。①劳动力集中的地区可能会有较低的生育率，因为城市居民往往面临更高的生活成本和更多的生活选择；②城乡差异可能导致资源分配不均，影响教育和医疗服务，进而影响人们的生育选择和总体健康水平；③这种分布也会影响国家的人力资源配置和经济发展策略，进而间接影响人口增长。

5.2 数据驱动分析

为进一步探究各因素对中国人口的影响大小和关联程度，我们收集了 2000—2005 年各指标的具体数据①（详见 data1.xlsx），通过计算 Spareman 相关性系数和建立 SHAP 模型对问题进行深入分析。

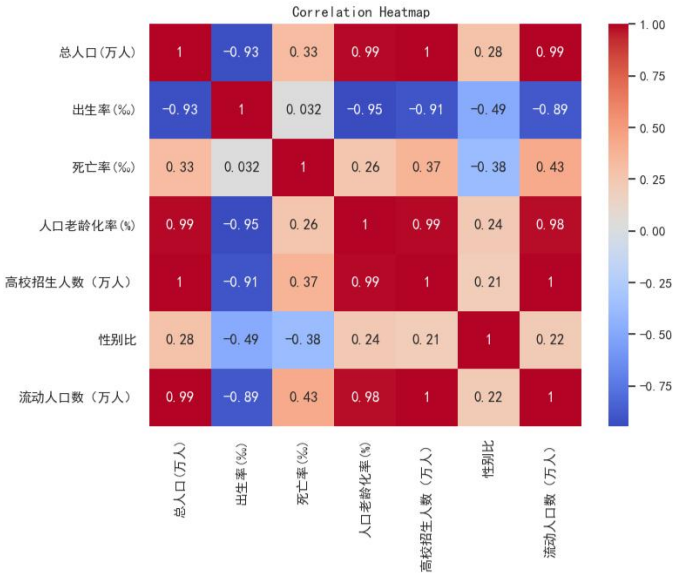
5.2.1 Spareman 系数分析

Spareman 相关性系数属于秩相关性系数的一种，通过评估两个变量之间的单调关系判断二者的相关性程度。其绝对值越接近 1，表明两变量相关性越强。其计算公式如下：

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

其中 di 为 Xi 和 Yi 间的等级差。

对于归一化处理好的数据，我们使用 pandas 库中.corr()方法计算各因素间斯皮尔曼系数，并绘制相关性热力图如下：



由图可知，总人口数与出生率呈显著负相关，与人口老龄化率和人口素质呈显著正相关，而与死亡率和性别比相关性较小，这是由于死亡率和性别比长期恒定、随人口变化较小，其对人口增长的影响也趋于稳定态势。同时，通过热力图还可以发现一些有价值的现象，如出生率与老龄化率呈显著负相关（-0.95），这侧面支持了老龄化主要通过降低出生率对人口增长起负作用。

5.2.2 SHAP 模型分析

相关性系数只能较为粗略的反应变量间关联程度，但我们希望精确了解各自变量（影响因素）对因变量（人口增长）的“贡献值”大小。对此，我们首先构建机器学习模型，用来拟合自变量和因变量间的关系，然后通过 SHapley Additive exPlanations 模型对各因素的影响值进行拆解分析。

SHAP 是 Python 开发的一个"模型解释"包，可以解释任何机器学习模型的输出。在合

① 人口素质可通过受教育水平反应，使用当年普通高校招生人数进行衡量。

作博弈论的启发下 SHAP 构建一个加性的解释模

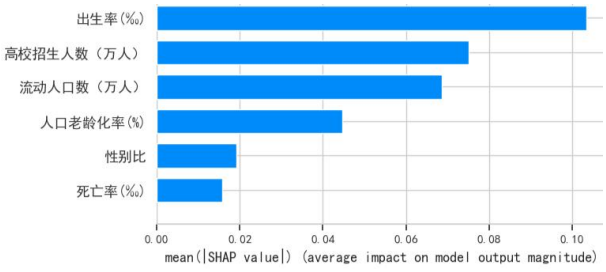
型，所有的特征都视为“贡献者”。对于每个预测样本，模型都产生一个预测值，SHAP value 就是该样本中每个特征所分配到的数值。假设第 i 个样本为 xi，第 i 个样本的第 j 个特征为 xij，模型对该样本的预测值为 yi，整个模型的基线（通常是所有样本的目标变量的均值），为 y_base，那么 SHAP value 服从以下等式：

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{ik})$$

首先，我们尝试构建多种机器学习模型，分别计算其拟合的平均百分比误差和 r2，以择优选择模型，结果如下图所示。

模型	R^2	平均百分比误差
XGBOOST 模型	0.99997	0.1186
决策树模型	1.00000	0.0000
随机森林模型	0.94039	7.6525
ADABOOST 模型	1.00000	0.0000
GBDT 模型	0.99999	0.0001

根据模型评估结果，我们选择拟合效果最佳的决策树模型作为最终方案，建立 SHAP 解释器对影响因素进行“贡献分析”。各影响因素的 SHAP Value 如下图所示：



由图可知，出生率对人口数量的变化影响最为显著，其次人口素质，而性别比和死亡率对人口变化影响较小，约只有出生率的 1/6，这与 Spareman 系数的分析结果相一致。

具体来看，在考虑到各变量间的相互影响下，人口数量每增加单位 1（已归一化），出生率的贡献值高达 0.6264，而流动人口贡献值为 0.1891，老龄化率贡献值为 0.1429。

六、 问题二模型建立与求解

6.1 经典模型的弊端

关于人口增长预测问题，存在这许多经典数学模型，如 Malthus 模型、Logistic 模型等，但由于模型自身的局限性和问题的复杂性，经典问题难以胜任中国人口准确预测的任务。

Malthus 模型，也称为指数增长模型，由托马斯·马尔萨斯提出。该模型的基本假设是，在没有资源限制的情况下，人口会以恒定的比率增长。在该问题下，此模型有以下主要弊端：

1 忽略了当今我国资源限制、环境承载力等因素影响。

2 忽略了社会经济因素影响。比如：城镇化率提高、医疗条件改善、计划生育等。

3 过于简化。在日新月异的当代中国，无法准确预测长期人口动态。

Logistic 模型，又称为 S 曲线增长模型，由皮埃尔-弗朗索瓦提出，是对 Malthus 模型的扩展。该模型考虑到了环境承载力对人口增长的限制。在该问题下，此模型有以下主要弊端：

1 忽略了环境承载力的复杂性。认为环境承载力是静态的不变的，也忽略了气候、资源可用性等自然条件的变化性。

2 忽略了社会因素影响。尽管 Logistic 模型比 Malthus 模型更加复杂，它仍然忽视了计划生育、技术进步等因素影响。

综上所述，Malthus 模型和 Logistic 模

型都不适用于现代社会的人口增长模型，因为它们忽略了许多影响人口增长的因素，并且过于简化了人口增长的模式。在现代社会，人口增长是一个复杂的现象，需要更综合、更动态的模型来进行准确预测。

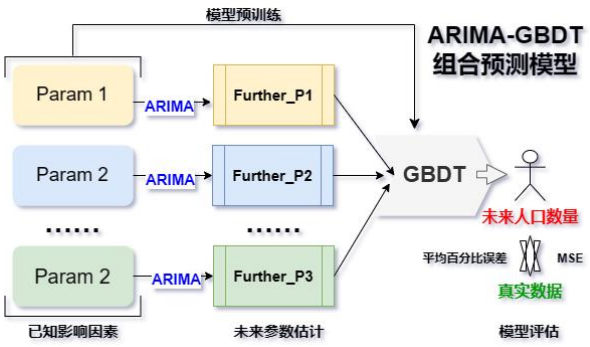
6.2 ARIMA-GDBT 组合模型

建立能够准确预测中国长、短期人口数量的数学模型难点主要有两个：

①对时间序列数据的准确预测：人口数量本质为一组时间序列，因此模型必须具备对时序数据的良好预测能力。

②对多影响因素的有效提取与利用：人口数量的变化受多种因素影响，因此模型必须能够从复杂的影响因素中提取有效数据，并判定多种影响因素的总和作用。

单一的数学模型很难同时具备这两种能力，对此，我们提出了 ARIMA-GDBT 组合预测模型，结合 ARIMA 对时间序列的精准预测能力和梯度提升树回归(GDBT)的复杂变量综合预测能力，实现对中国人口的准确预测，其构建思路如下：



6.3.1 平稳性检验

数据的平稳性是建立 ARIMA 模型的前提，否则需对训练数据进行差分处理。本文通过计算 1965~2005 年各组数据的 ADF 值，对数据平稳性进行检验，以“人口出生率(‰)”数据序列预测结果为例：

```
Results of ADF Test:
Test Statistic                  -1.028023
p-value                          0.742907
#Lags Used                      10.000000
Number of Observations Used     30.000000
Critical Value (1%)             -3.669920
Critical Value (5%)             -2.964071
Critical Value (10%)            -2.621171
dtype: float64
```

分析结果现实，检测值大于 5%临界水平，数据不平稳，指对数据进行差分处理，在进行二阶差分后数据变得平滑。其他数据类似。

6.3.2 纯随机性检验

如下表所示，原序列在各阶数下 LB 统计量的 P 值均小于 5%临界水平，因此该序列拒绝原假设，即：我国 1965~2005 年人口总数二阶差分序列为非白噪声序列，可直接拟合模型。

指标名称	延迟 1 阶	延迟 2 阶
X-squared	6.4627	6.788
p-value	0.01102	0.03357

6.3.3 ARIMA(p, d, q)定阶

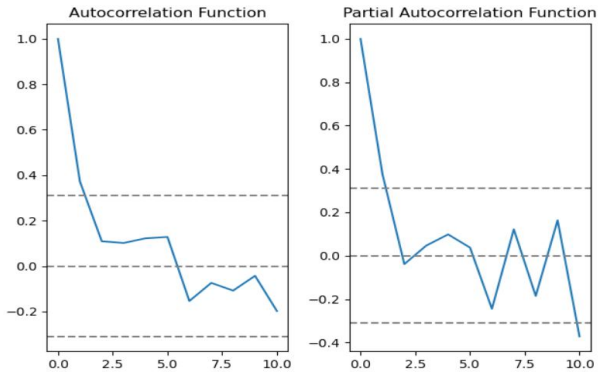
建立预测准确的 ARIMA 模型需要对 p,d,q 三大参数进行定阶：

p（自回归项阶数）：可以通过 PACF 图来确定。在 PACF 图中，p 是第一个显著大于置信区间（通常用蓝线表示）的滞后阶数。

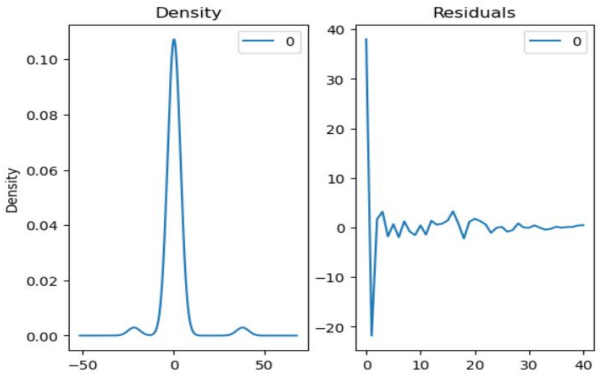
d（差分阶数）：这是为了使时间序列平稳所需的差分次数。根据 ADF 检验结果确定。

q（移动平均项阶数）：可以通过 ACF 图来确定。在 ACF 图中，q 是第一个显著大于置信区间的滞后阶数[ii]。

下图为中国 1965~2005 年“人口出生率(‰)”二阶差分后序列自相关和偏自相关图。

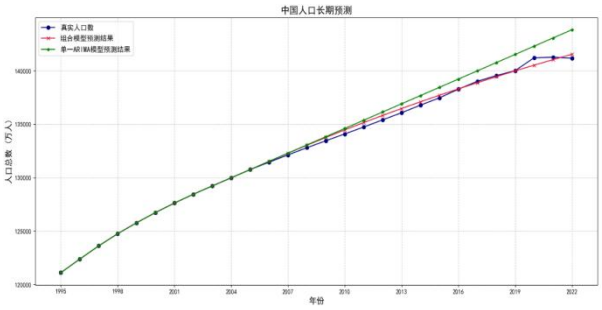


由图分析可知，p 可以为 1，q 可以为 1，d 为 2。其他数据定阶方式类似。最后对定阶后的模型进行误差-残差分析（如下图所示），可见在选定的模型参数下，误差、残差分布接近白噪声，说明模型拟合良好，不存在系统误差。



6.3.3 预测结果分析

在通过 ARIMA 得到未来参数估计后，我们将预测的影响因素带入训练好的 GBDT 模型，分别输出 2006~2015 和 2006~2023 年中国人口预测值。同时，为证明组合模型的优越性，我们设置了对照试验：使用单一 ARIMA 模型进行预测，并在同一张图中绘制预测曲线与真实值进行比较。如下图所示：



由图可知，与单一 ARIMA 模型相比，组合模型对真实情况预测更加准确。为更精确衡量误差，我们进一步计算了预测值与真实值的平均百分比误差，如下表所示：

平均百分比误差	单一 ARIMA 模型	组合模型
中短期预测	-0.404%	-0.157%
长期预测	-0.771%	-0.098%

与单一模型性比，组合模型体现出显著的优势。特别是在长期预测中，平均百分比误差约为 -0.098%，而单一 ARIMA 模型为 -0.771%。这意味着：在 14 亿人口规模下，组合模型预测平均误差不足 150 万人，模型预测效果十分精确。

综上所述，ARIMA-GBDT 组合模型在中短期（2006-2015）和长期（2006-2023）预测中均表现优异，充分体现了模型的优越性和实用价值。

七、 参考文献

i 陈霞, 肖岚. Logistic 模型的改进与中国人口预测[J]. 成都电子科技大学学报, 2020, 35(2): 239-243.

ii 徐翔燕, 侯瑞环. 基于 GM(1,1)-SVM 组合模型的中长期人口预测研究[J]. 计算机科学, 2020, 47(S1): 485-487+493.