Pure Mathematics 理论数学, 2023, 13(12), 3690-3706
Published Online December 2023 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2023.1312382

Hans汉斯

基于神经网络算法的区域碳排放量的建模分析

项兆佳

贵州大学大数据与信息工程学院，贵州 贵阳

收稿日期：2023年11月17日；录用日期：2023年12月18日；发布日期：2023年12月29日

摘　要

作为全球最大的发展中国家，中国的能源生产和碳排放位居首位。为实现碳达峰与碳中和目标，中国采取了多项应对气候变化的举措，取得明显效果。本文采用“华为杯”第二十届中国研究生数学建模竞赛D题中的数据，首先建立一个评估指标及体系，对碳排放量与经济、人口、能源消费量之间进行相关性分析，建立关联关系模型。在线性回归模型预测的基础上，引入XGboost和LSTM预测模型再次进行预测，分别得到预测结果。然后建立优化模型，以精度最小为目标函数，对三种预测方式进行加权，利用优化模型求解最优的权重，分别得到一个基于人口和经济变化的能源消费量预测模型，最终利用岭回归作为改进算法，构建碳排放量与人口、经济和能源消费量的多元线性回归的区域碳排放量预测模型。利用本文的算法，可以及时预测出区域碳排放量，使得中国能够更好、更快的走上绿色发展之路。

关键词

关联关系模型，碳排放量，线性回归，双碳目标

Modeling and Analysis of Regional Carbon Emissions Based on Neural Network Algorithm

Zhaojia Xiang

School of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

Received: Nov. 17th, 2023; accepted: Dec. 18th, 2023; published: Dec. 29th, 2023

Abstract

As the world’s largest developing country, China ranks first in terms of energy production and carbon emissions. In order to achieve the goals of carbon peak and carbon neutrality, China has

文章引用: 项兆佳. 基于神经网络算法的区域碳排放量的建模分析[J]. 理论数学, 2023, 13(12): 3690-3706.
DOI: 10.12677/pm.2023.1312382

taken a number of measures to combat climate change, and has achieved remarkable results. Based on the data in question D of the 20th China Graduate Mathematical Contest in Modeling, this paper first establishes an evaluation index and system, analyzes the correlation between carbon emissions and economy, population, and energy consumption, and establishes a correlation model. On the basis of linear regression model prediction, XGboost and LSTM prediction models were introduced to predict again, and the prediction results were obtained respectively. Then, the optimization model is established, the three prediction methods are weighted with the minimum precision as the objective function, and the optimal weights are used to solve the optimal weights, and an energy consumption prediction model based on population and economic changes is obtained respectively, and finally the ridge regression is used as an improved algorithm to construct a regional carbon emission prediction model with multiple linear regression of carbon emissions and population, economy and energy consumption. Using the algorithm in this paper, regional carbon emissions can be predicted in time, so that China can better and faster embark on the path of green development.

Keywords

Correlation Model, Carbon Emissions, Linear Regression, Dual Carbon Goals

1. 引言

国际社会普遍认为，二氧化碳过度排放是引起气候变化的主要因素。人类活动排放的二氧化碳等温室气体导致全球变暖，加剧气候系统的不稳定性，导致一些地区干旱、台风、高温热浪、寒潮、沙尘暴等极端天气频繁发生，强度增大，极端天气为我们的日常生产生活带来了诸多不便，全球生态平衡时刻遭到破坏。目前，全球范围内能源及产业发展低碳化的大趋势已经形成。我国近年来减排成效显著，2019年碳排放强度比2005年下降48.4%。2020年9月22日，习近平主席在第七十五届联合国大会一般性辩论上提出中国将提高国家自主贡献力度，采取更加有力的政策和措施，二氧化碳排放力争于2030年前达到峰值，努力争取2060年前实现碳中和[1]。“双碳”目标是着力解决资源环境约束突出问题、实现中华民族永续发展的必然选择，是构建人类命运共同体的庄严承诺。通俗来讲，碳达峰[2]指二氧化碳排放量在某一年达到了最大值，之后进入下降阶段；所谓的“碳中和”[3]，就是在一定时期内，由某一组织或某一群体所产生的二氧化碳，可以被自然或人为的手段所吸收或抵消，从而达到“零排放”的目的。

于中国而言，作为全球第二大经济体和最大的发展中国家，中国经济和社会各行各业正呈现蒸蒸日上的发展态势，而这背后需要庞大的资源和能源支撑，大量资源和能源消耗的同时也会带来二氧化碳排放的进一步增加[4]，在2020年，全球能源碳排放340亿吨，中国碳排放99亿吨，占全球碳排放的29% (如图1所示)，但是随着中国社会主义现代化建设的逐步完善，以及绿色低碳等创新技术的广泛应用，二氧化碳排放总量终将迎来下降的拐点，即“碳达峰”目标，因此，实施“双碳”目标刻不容缓。要想在2060年实现“零排放”，就需要解决“发展”与“减排”的矛盾。在这些问题中，最突出的问题是如何促进经济和社会的高质量发展。经济增长、碳排放量与能源消费量之间必然存在着关联关系[5]，要想做到节能减排，必须从提高能源使用效率、增加非化石能源消费比例着手，才能实现经济增长与碳排放的逆向转化。

本文根据“华为杯”第二十届中国研究生数学建模竞赛 D 题提供的数据，该数据假定位于中国东南沿海，地势平坦，水陆交通便利，人口密集，经济发达，科教资源丰富，但能源及生态碳汇资源相对匮乏，因此此数据能更准确的预测区域的碳排放量与各个指标之间的关系。该数据以 Excel 格式给出，时间范围从 2010 年至 2020 年，历史数据的基期是 2010 年，十二五时期为 2011~2015 年；十三五时期为 2016~2020 年。数据内容包括：人口、总产值(GDP)及三次产业与部门的产值分布、总能耗及三次产业与部门的能耗分布、总能耗及化石能源与非化石能源品种分布、能源消费的三次产业与部门的品种结构、碳排放量总量及产业与部门的分布、碳排放量的三次产业与部门的能源品种结构以及碳排放量相关各类能源的碳排放因子。首先对数据进行预处理操作，包括降维、异常值、缺失值检测等；利用 Python 寻找缺失值，对于存在异常的缺失值使用 KNN 算法进行填充处理。通过选定指标及构建体系，对碳排放以及人口、经济、能源消费量的现状进行分析，并建立关联关系模型，深入研究这些因素之间的相互关系。利用处理之后的数据进行进一步处理，建立回归预测模型。使用不同的预测模型(线性回归[6]、Xgboost [7]、LSTM [8]、灰色预测[9]、随机森林[10]等)对进行人口和 GDP 的预测，建立基于人口和经济变化的能源消费量预测模型。然后使用多元线性回归模型描述碳排放量与人口、GDP 和能源消费量之间的关系，以及碳排放量与各能源消费量部门和能源品种的关系。根据预测结果，可以确定最合适的路径和措施，以实现双碳目标，同时考虑全球碳排放趋势、新能源技术和新能源消费模式等因素，为制定更全面和可行的碳达峰和碳中和战略提供更多参考。

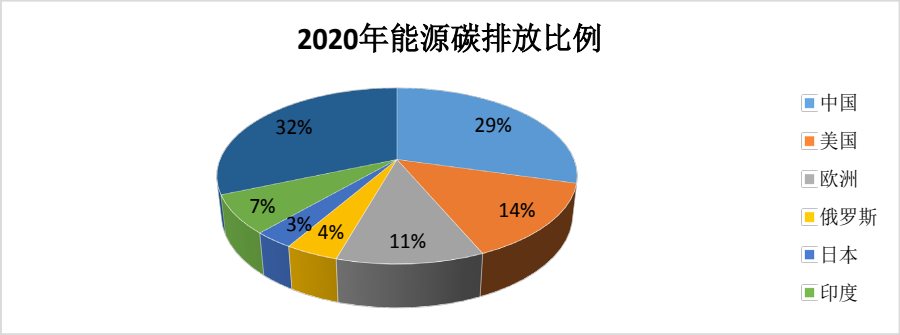

Figure 1. China’s carbon emissions from energy in 2020
图 1. 2020 年中国能源碳排放量比例

2. 建立线性关系

2.1. 数据预处理

首先利用 Python 对给出的数据进行异常缺失值的查找，经过查找得到数据中的异常值在表格中表示为“-”。以能源消费部门碳排放因子数据表格为例，数值栏中“-”表示该年该项能源未使用，无法计算实际的碳排放因子。在第三产业的建筑消费部门中煤炭的碳排放因子数据中存在数据缺失的情况，如表 1 所示，2012 年的数据被认定为异常缺失值。

Table 1. The abnormal missing value
表 1. 异常缺失值

年份	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
煤炭	2.66367346	2.66367346	-	2.65251185	2.66367346	2.62265793	2.66367346	2.66367346	2.66367346	2.66367346	2.7

为了保证结果的准确性，不能直接忽略该种缺失值。根据具体情况和需求，在本文的研究中，采用了插值填充方法。在插值填充方法中，基于 KNN 算法[11]的数值插补方法是一种有效的异常缺失值处理方法。这种方法可以利用已有数据的特征和相似性，通过寻找最近邻样本来预测并填补缺失值。相比于其他插值填充方法，基于 KNN 算法的数值插补方法更能保持数据的完整性，并有效降低数据分析和建模过程中的偏差。

调用 sklearn 库的 impute 模块中的 KNNImputer 函数可以实现该算法，将数据导入进行计算，得到部分结果如表 2 所示。

Table 2. The result of the processing
表 2. 处理后的结果

年份	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
煤炭	2.66367346	2.66367346	2.6620883	2.652511856	2.66367346	2.62265793	2.66367346	2.66367346	2.66367346	2.66367346	2.7

2.2. 构建指标评价体系

为了更加全面和准确地构建指标评价体系[12]，首先需要根据提供的数据和需要解决的问题，选择主要指标，即一级评价指标。这些一级评价指标包括区域经济、人口、能源消费量以及碳排放量。

根据给出的数据文件，对于人口指标，将常驻人口总量作为其二级指标；对于经济指标，选择 GDP 总量、第一产业(农林消费部门)、第二产业(总量、能源供应部门以及工业消费部门)、第三产业(总量、交通消费部门以及建筑消费部门)作为其二级指标；对于能源消费量，选择总量、化石能源消费和非化石能源消费作为二级指标；对于碳排放量，选择排放总量、与各部分碳排(农林消费部门、工业消费部门、第三产业总量、第三产业交通消费部门、第三产业建筑消费部门、居民生活消费)作为二级指标。

2.3. 现状分析

为了更加直观的观察分析某区域十二五(2011~2015 年)和十三五(2016~2020 年)期间的碳排放量状况，将给出的数据集进行拆分，利用 Excel 绘制十二五、十三五时间下变化趋势，如图 2 所示。

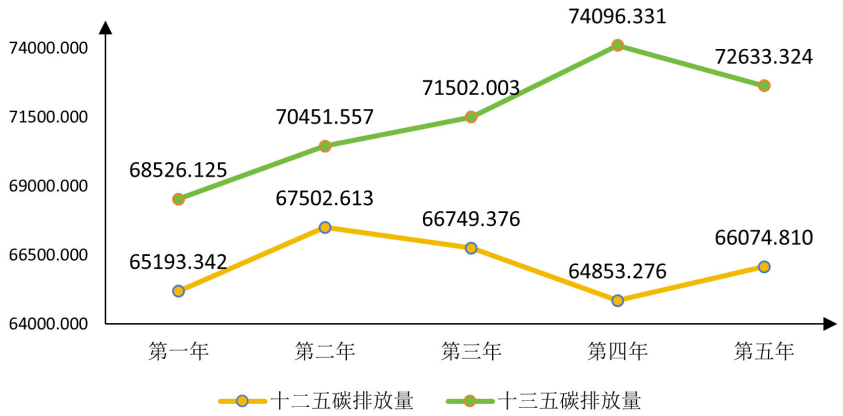

Figure 2. Trends in carbon emissions during the 12th and 13th five-year plans
图 2. 十二五、十三五期间碳排放量变化趋势

通过折线图，可以较为清晰的看出，十三五期间的碳排放量明显高于十二五期间。对于十二五期间而言，中间三年的碳排放量有明显下降趋势；十三五期间前面四年的碳排放量一直保持上升的趋势。为

了能直观的看出不同时期的增长率，本文计算了十二五、十三五期间碳排放量的增长率，绘制为柱状图如图 3 所示。

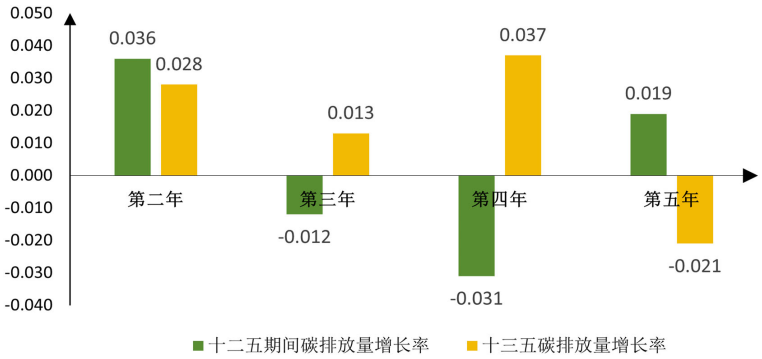

Figure 3. The growth rate during the twelfth and thirteenth five-year plans
图 3. 十二五、十三五期间增长率

通过柱状图，可以清晰的观察到，在十三五计划末期碳排放增长率已经开始降低，而十二五末期碳排放增长率开始上升。

为了更加直观的看各个指标之间的关系，绘制指标之间的散点图观察变化趋势。如图 4、图 5 所示。

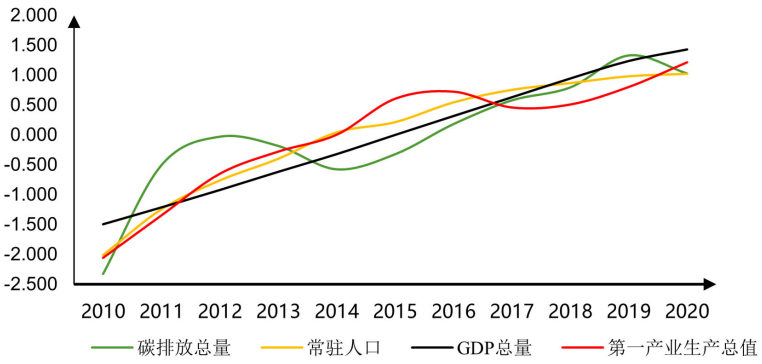

Figure 4. Carbon emissions
图 4. 碳排放量相关图

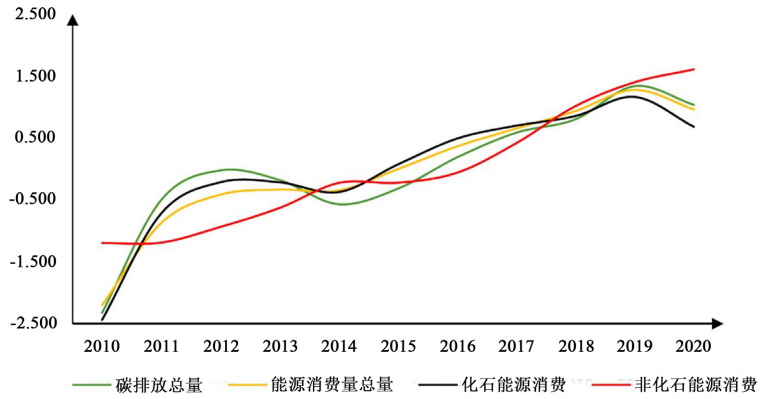

Figure 5. Carbon emissions and energy consumption
图 5. 碳排放量与能源消费相关图

通过散点图，可以看出大部分指标都与碳排放量呈现一定的正相关。为了更加客观地判定结果，这里可以建立相关性分析模型，引入 person 相关性，判断指标之间的关系。其中样本协方差、样本标准差和样本皮尔逊相关系数的公式分别如(1)、(2)、(3)、(4)所示。

$$S_{xy}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(Y_i-\bar{Y}\right)}{n-1} \tag{1}$$

$$S_x=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1}} \tag{2}$$

$$S_y=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}{n-1}} \tag{3}$$

$$r_{xy}=\frac{S_{xy}}{S_X S_Y} \tag{4}$$

利用相关系数研究大部分指标都与碳排放量的相关关系，得到相关性分析结果如表 3 所示，这里首先展示二级指标之间的相关性。

Table 3. Correlation of secondary indicators
表 3. 二级指标相关性

二级指标	GDP 总量	第一产业生产总值	第二产业生产总值	能源供应部门	工业消费部门
GDP 总量	1	0.918	0.997	0.985	0.999
第一产业生产总值	0.918	1	0.918	0.9	0.918
第二产业生产总值	0.997	0.918	1	0.986	1
能源供应部门	0.985	0.9	0.986	1	0.985
工业消费部门	0.999	0.918	1	0.985	1

为了更加直观的看出指标间的相关性，这里利用 origin 绘制了矩阵热力图，如图 6 所示。

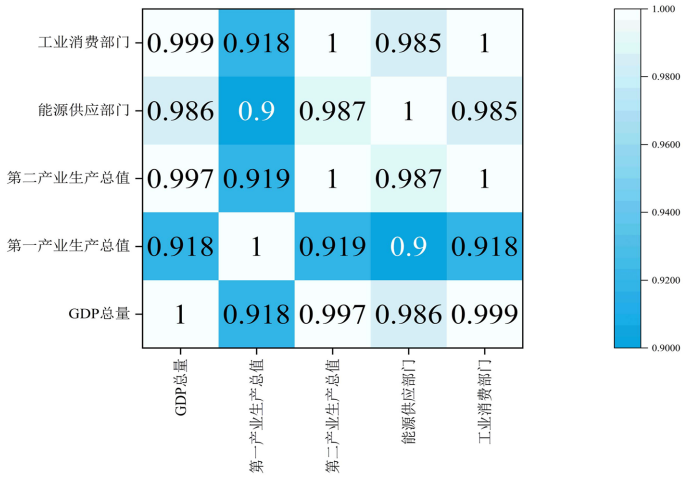

Figure 6. Matrix thermodynamic diagram
图 6. 矩阵热力图

通过相关性分析图表的可视化结果，本文可以看出经济指标与能源消费量之间具有较好的相关性。同时对几个多维一级指标进行降维处理，将结果导入关联分析模型，得到下面的结果，如表 4 所示。

Table 4. Results of correlation analysis
表 4. 相关性分析结果

指标	常驻人口	经济	能源消费量	碳排放
常驻人口	1 (0.000***)	0.970 (0.000***)	−0.352 (0.290)	−0.92 (0.000***)
经济	0.970 (0.000***)	1 (0.000***)	−0.159 (0.643)	−0.967 (0.000***)
能源消费量	−0.352 (0.290)	−0.159 (0.643)	1 (0.000***)	0.048 (0.891)
碳排放	−0.92 (0.000***)	−0.967 (0.000***)	0.048 (0.891)	1 (0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平。

通过上述分析，本文可以得出以下结论：

1) 对于这四项一级指标大部分都具有较好的相关性；

2) 除了能源消费量与其他三个指标的相关性不太明显。

3. 建立预测模型

通过对数据集的分析，常驻人口总量与年份之间有着一定的线性关系，所以选择线性回归模型进行预测，而常驻人口的变化量通常还与经济，能源消费量有关，因此为了得到更好的预测结果，可以将经济变化指标相关因素，能源消费量作为常驻人口预测的特征。能源消费指标可以考虑煤炭消费量、油品消费量等，这些特征呈非线性关系。由于线性回归模型无法捕捉到序列中的非线性因素，而神经网络算法在处理非线性问题具有独特的优势，因此本文准备用另外两个机器学习模型(LSTM, XGboost)分别进行预测。LSTM 时间序列[13]数据通常具有长期依赖性，意味着当前时间步的值受到之前时间步的值的影响，选择了多个特征进行模型预测。XGBoost 模型[14]可以通过评估特征在决策树中的分裂贡献度或特征覆盖度来度量特征的重要性，再将各预测模型通过加权方法，得到最终的预测结果。

3.1. 线性回归预测

线性回归(Linear Regression)是确定两种或两种以上变量之间所存在的定量关系的一种统计方法，线性回归的数据集的形式为多个属性 X 与一个对应的 Y，目的是求解 X 与 Y 之间的线性映射关系，优化求解参数的目标是降低预测值与 Y 之间的差别。它可以通过数据学习得到一个通过自变量的线性组合来进行预测因变量的函数，形式如下式所示：

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n w x_n + b \tag{5}$$

大多数所见到的均为向量形式，如下：

$$w^{\mathrm{T}} x + y = b \tag{6}$$

其损失函数是单个样例的误差，具体形式为：

$$\left| h_\theta \left(x^i \right) - y^i \right| \tag{7}$$

代价函数是对数据集整体的误差描述，也就是损失函数的总和的平均，具体形式为：

$$\frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2 \tag{8}$$

目标函数为了避免过拟合或者进行一些特殊目的的约束，一般在代价函数的后面加入正则化项：$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$ + 正则化项

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2 + \text{正则化项} \tag{9}$$

为了对常驻人口进行预测，建立线性回归模型。建立线性回归模型，进行相关分析，如公式(10)所示：

$$\begin{cases} x_{i\,\text{人口}} = \chi_0 + \chi_1 t_i \\ \xi \sim N\left(0,\sigma^2\right) \end{cases} \tag{10}$$

其中，ξ服从正态分布，χ_1为自变量系数，χ_0为常数项系数。$X_{i\,\text{人口}}$表示第 i 年该地的常驻人口数量，t_i表示第 i 年。最终得到表 5 中的结果，如下所示。

Table 5. Linear regression results
表 5. 线性回归结果

	非标准化系数		标准化系数	t	VIF	R^2	F
	B	标准误差	Beta				
常数	108334.380	11712.370	-	−9.25	-	0.918	F = 99.12 P = 0.000***
常驻人口	57.868	5.812	0.957	9.958	1		

注：***、**、*分别代表 1%、5%、10%的显著性水平。

因此，得到的结果为公式(11)所示。

$$x_{i\,\text{人口}} = -108334.380 + 57.868t_i \tag{11}$$

为了更加直观的观察二者的关系式，绘制了拟合示意如图 7 所示。

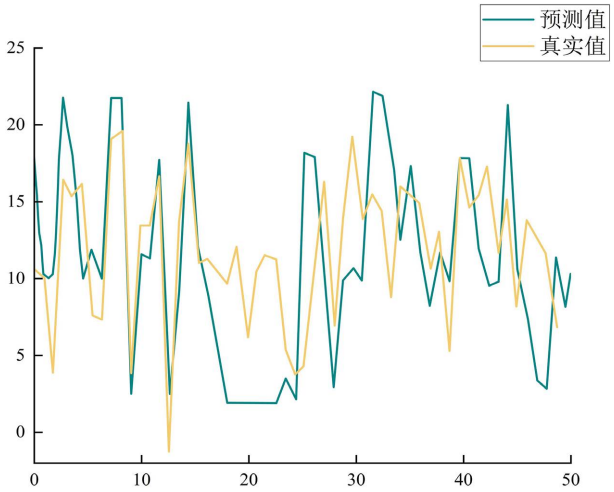

Figure 7. Fitting effect diagram
图 7. 拟合效果图

3.2. XGBoost 模型预测

XGBoost 是 Extreme Gradient Boosting 的缩写，称为随机梯度提升算法，是决策树算法的集合。

XGBoost 算法的思想是将许多弱分类器集成在一起，形成一个强分类器(个体学习器间存在强依赖关系，必须串行生成的序列化方法)。XGBoost 的基本组成元素是决策树，这些组成 XGBoost 的决策树之间是有先后顺序的：后一棵决策树的生成会考虑前一棵决策树的预测结果，即将前一棵决策树的偏差考虑在内，使得先前决策树做错的训练样本在后续受到更多的关注，然后基于调整后的样本分布来训练下一棵决策树。优点如下：1) 精度高。XGBoost 对损失函数进行了二阶泰勒展开，一方面为了增加精度，另一方面也为了能够自定义损失函数，二阶泰勒展开可以近似许多损失函数。2) 灵活性强。XGBoost 不仅支持 CART，还支持线性分类器；XGBoost 还支持自定义损失函数，只要损失函数有一二阶导数。3) 防止过拟合。XGBoost 在目标函数中加入了正则项，用于惩罚过大的模型复杂度，有助于降低模型方差，防止过拟合。

本文将建立 XGboost 模型，对常驻人口的数据进行预测。XGboost 是由多个个基础模型组成的一个加法模型，在训练出一颗树的基础上再训练下一颗树，预测它与真实分布之间的差距，通过不断训练用来弥补差距的树，最终用树的组合实现对真实分布的模拟。

假设本文第 t 次迭代要训练的树模型是 $f_t(x_i)$，则有：

$$y_t^{(t)} = \sum_{k=1}^{t} f(x_i) = y_i^{(t-1)} + f_t(x_i) \tag{12}$$

其中 $y_t^{(t)}$ 是第 t 次迭代后样本的预测结果，$y_i^{(t-1)}$ 是前 $t-1$ 的预测结果，$f_t(x_i)$ 是第 t 棵树的模型。训练该模型通常是定义一个目标函数去优化它，而 XGboost 的目标函数包含损失函数和正则项两部分，损失函数代表模型拟合的程度，正则项被用来控制模型的复杂程度。其目标函数为如公式(13)所示：

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, y_i^{(t)} + f_t(x_i)\right) + \Omega(f_t) \tag{13}$$

XGboost 的正则项是一个惩罚机制，叶子节点的数量越多，惩罚力度越大，从而限制他们的数量。惩罚项如公式(14)所示：

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{14}$$

其中，γT 为惩罚力度，$\frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ 为惩罚项，γ 为叶子节点 T 的系数。

3.3. LSTM 模型预测

LSTM，长短时记忆网络(Long Short Term Memory Network, LSTM)，是一种改进之后的循环神经网络，可以解决 RNN 无法处理长距离的依赖的问题，目前比较流行。LSTM (Long Short-Term Memory)通过门控机制和细胞状态的记忆能力，能够有效地捕捉和利用序列中的长期依赖关系(LSTM 对时间序列的预测其结构如图 8 所示)，在 t 时刻，LSTM 的输入有三个：当前时刻网络的输入值 x_t、上一时刻 LSTM 的输出值 h_{t-1}、以及上一时刻的单元状态 C_{t-1}；门的输出是 0 到 1 之间的实数向量，当门输出为 0 时，任何向量与之相乘都会得到 0 向量，这就相当于什么都不能通过；输出为 1 时，任何向量与之相乘都不会有任何改变，这就相当于什么都可以通过。

通过对数据集的分析，人口常驻总量随年份变化形成时间序列。时间序列数据通常表现出长期依赖性，即当前时间步的值受之前时间步的值的影响。传统的机器学习方法(如线性模型)难以捕捉这种长期依赖关系。因此，本文使用 LSTM 来捕捉这些非线性关系，同时选择使用多特征进行预测，从而提高预测的准确性。

为了消除不同场景由于不同因素造成的影响，采用最大/最小归一化方法对样本数据进行归一化处理，公式如下：

$$x=\frac{X-\min(X)}{\max(X)-\min(X)} \tag{15}$$

式中：X 为原始样本数据；$\min(X)$ 为原始样本数据的最小值；$\max(X)$ 样本数据的最大值；x 为归一化处理后的样本数据。

然后，本文需要选择网络输入与输出变量。具体地，选用三个核心指标的数据作为样本数据，确定网络输入变量的个数。

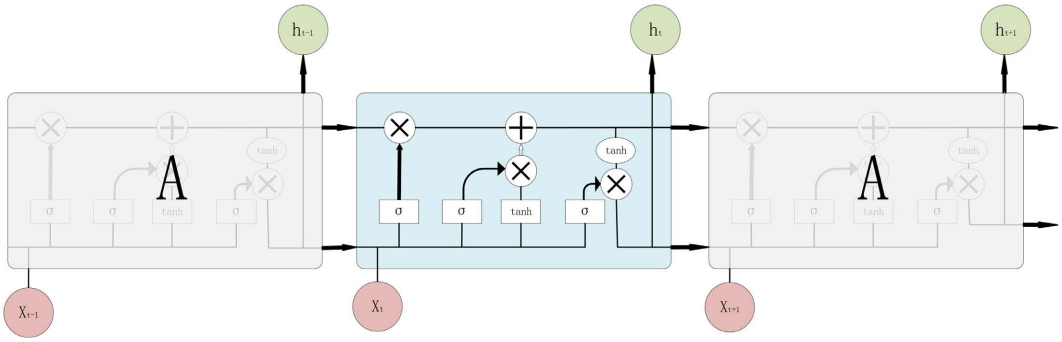

Figure 8. A schematic of the LSTM model
图 8. LSTM 模型示意图

其中，输入门，它决定了当前时刻网络的输入 x_t 有多少保存到单元状态 C_t：

$$I_t=\sigma\left(X_tW_{xi}+H_{t-1}W_{hi}+b_i\right) \tag{16}$$

遗忘门，它决定了上一时刻的单元状态 C_{t-1} 有多少保留到当前时刻 C_t：

$$F_t=\sigma\left(X_tW_{xf}+H_{t-1}W_{hf}+b_f\right) \tag{17}$$

输出门，控制单元状态 C_t 有多少输出到 LSTM 的当前输出值 h_t：

$$0_t=\sigma\left(X_tW_{xo}+H_{t-1}W_{ho}+b_o\right) \tag{18}$$

候选记忆细胞计算：

$$\overline{C_t}=\tanh\left(X_tW_{xc}+H_{t-1}W_{hc}+b_o\right) \tag{19}$$

将数据导入 Matlab，利用编写的程序进行预测。其训练进度如图 9 所示：

4. 建立区域碳排放量预测模型

4.1. 加权预测模型

对于基于人口的能源消费量预测模型，需要考虑如何减小预测值与真实值的差距，提高该模型的预测精度，因此选择建立一个加权预测模型。该加权模型选择线性回归模型、XGboost 预测模型和 LSTM 神经网络分别建立不同的子模型，然后对预测结果构建加权优化模型，以提高模型精度。

设 $y_i\left(i=1,2,\cdots,m\right)$ 是同一个问题通过 m 种预测方法(每种方法都通过各自的检验)所得到的预测结果，y'_{ij} 是第 $i\left(i=1,2,\cdots,m\right)$ 种预测方法对原始数据 $y_i\left(i=1,2,\cdots,m\right)$ 的模拟值。则这 m 种方法的加权预测值为：

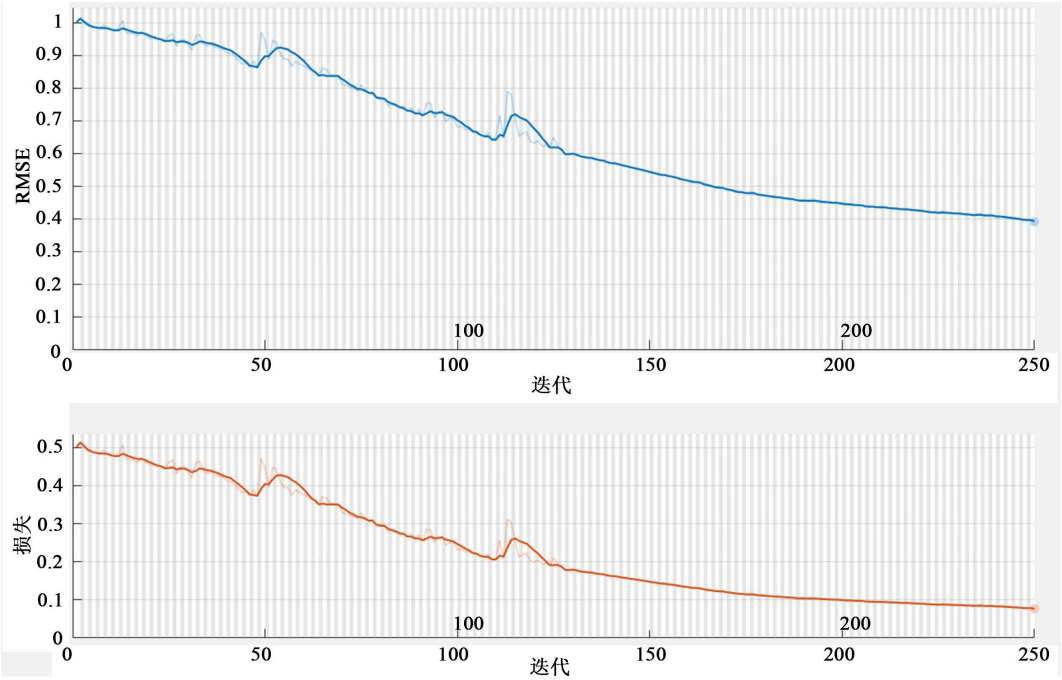

Figure 9. Training schedule
图 9. 训练进度

$$J = x_1\hat{y}_1 + x_2\hat{y}_2 + \cdots + x_m\hat{y}_m \tag{20}$$

其中，$\sum_{i=1}^{m} x_i = 1, x_i \geq 0, (i = 1,2,\cdots,m)$

$e_{ij} = y_{ij} - y_j$ 表示第 i 种预测方法与第 j 个历史数据的模拟值之间的误差，记：

$$E_j = \begin{pmatrix} e_{1j}^2 & e_{1j}e_{2j} & \cdots & e_{1j}e_{mj} \\ e_{1j}e_{2j} & e_{2j}^2 & \cdots & e_{2j}e_{mj} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1j}e_{mj} & e_{2j}e_{mj} & \cdots & e_{mj}^2 \end{pmatrix} \tag{21}$$

$$\begin{aligned} L &= \sum_{j=1}^{n}\left(\hat{J}_j - y_j\right)^2 = \sum_{j=1}^{n}\left(x_1\hat{y}_{1j} + x_2\hat{y}_{2j} + \cdots + x_m\hat{y}_{mj} - y_j\right)^2 \\ &= \sum_{j=1}^{n}\left(x_1e_{1j} + x_2e_{2j} + \cdots + x_me_{mj}\right)^2 = \sum_{j=1}^{n} x^{\mathrm{T}}E_j x \end{aligned} \tag{22}$$

因此，本文将建立一个优化模型，将误差最小化作为目标函数，并且加上权重和为 1 的约束条件。然后，本文可以使用多目标粒子群优化算法[15]来求解该规划模型，相比直接使用 lingo 或 matlab 进行求解，这种方法可以减少计算量。

通过本文的规划模型，每个粒子确定每个粒子个体的最优解并从这些个体最优解找到一个全局最优值，这里涉及到了本文粒子的位置和距离公式，公式如下所示：

$$z_{ij} = z_{ij} + d_{1ij} \times \mathrm{rand}(\,) \times \left(\mathrm{pbest}_{ij} - x_{ij}\right) + d_{2ij} \times \mathrm{rand}(\,) \times \left(\mathrm{pbest}_{ij} - x_{ij}\right) \tag{23}$$

$$x_{ij} = x_{ij} + z_{ij} \tag{24}$$

$$z_{ij} = \omega \times z_{ij} + d_{1ij} \times \mathrm{rand}(\,) \times \left(\mathrm{pbest}_{ij} - x_{ij}\right) + d_{2ij} \times rand(\,) \times \left(\mathrm{pbest}_{ij} - x_{ij}\right) \tag{25}$$

其中，rand( )为 0 到 1 之间的随机数。ω 为惯性权重因子，其值非负。当 ω 的数值较大时，该算法全局寻优能力强，局部寻优能力弱；当 ω 的数值较小时，该算法全局寻优能力弱，局部寻优能力强。ω 的引入，可以极大地提升了粒子群优化算法的性能，使其能够根据不同的搜索问题调整全局和局部搜索的能力。同时在粒子群优化搜索的过程中对 ω 进行动态化处理，这样可以获得更好的寻优结果，因此本文采用线性递减权值策略[16]，如公式(26)所示：

$$\omega^{(t)} = (\omega_{\text{ini}} - \omega_{\text{end}})(G_k - g)/G_k + \omega_{\text{end}} \tag{26}$$

其中，G_k 为最大迭代次数，ω_{ini} 为初始惯性权值，ω_{end} 迭代到最大进化数时的惯性权值。

为了方便更好的理解该算法，绘制了思维导图，如图 10 所示：

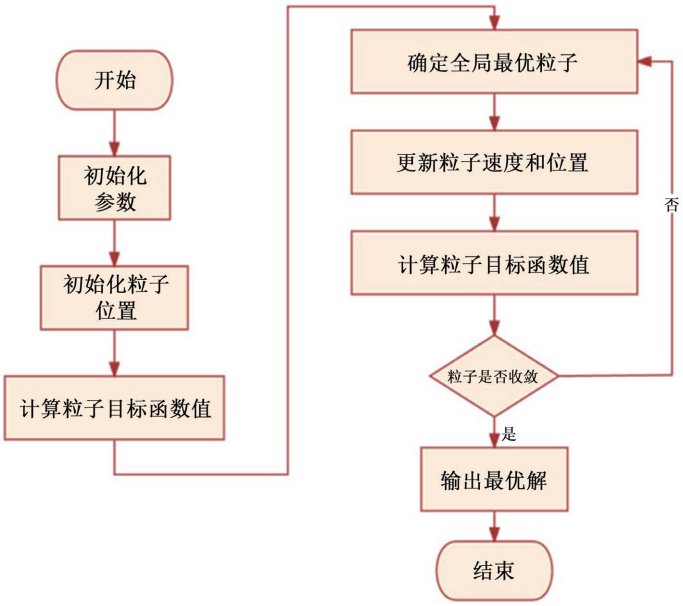

Figure 10. Mind map
图 10. 思维导图

最终，得到预测结果如表 6 所示。

Table 6. Energy consumption forecast based on population forecast
表 6. 基于人口预测的能源消费量预测结果

年份	2021	2022	2023	2024	2025
人口	8591.622	8643.531	8695.755	8748.298	8801.153
人口－能源消费量	32889.18175	33515.43602	34145.46640	34779.30894	35416.97579
年份	2056	2057	2058	2059	2060
人口	7951.225784	7830.246394	7694.467128	7546.078095	7385.96665
人口－能源消费量	25163.45385	23703.9588	22065.91743	20275.75213	18344.16744

4.2. 模型评估

在模型评估中，可以选择以下几种方法：平均绝对误差(MAE) [17]、平均绝对误差(MAE) [17]、决定

系数(R-squared) [18]以及平均绝对百分比误差(MAPE)。本文选择 MAPE (Mean Absolute Percentage Error, 平均绝对百分比误差)方法对模型进行评估，该方法有助于理解和比较预测结果与真实结果之间的平均偏差，更好地理解模型的表现和可靠性，并在业务决策中提供可靠的参考。具体而言，对于某一组样本数据，将计算其所有预测值与真实值之间的相对误差，然后取平均值并将其转换为百分比形式，即得到了该组数据的 MAPE 值。例如，如果 MAPE 为 5，则表示该组数据的预测结果平均偏离真实结果的 5%。通过计算不同数据组的 MAPE 值，本文可以比较不同模型在预测能力方面的优劣，并选择表现最佳的模型作为最终选择。

MAPE 公式为：

$$\mathrm{MAPE}=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{\mathrm{actual}(t)-\mathrm{forecast}(t)}{\mathrm{actual}(t)}\right|\times 100\% \tag{27}$$

利用给出的数据对 2021~2060 年的人口数据进行预测，得到预测结果，作为 $\mathrm{forecast}(t)$，而将数据中的当年数据作为 $\mathrm{actual}(t)$。

最终，本文得到基于人口的能源消费量的四种模型模型的 MAPE 进行汇总得到结果如表 7 所示。

Table 7. Results of model evaluation
表 7. 模型评估结果

方法	线性回归	XGboost	LSTM	集成模型
第一段预测 MAPE 值	8.5	4.52	6.3	4.15
第一段预测 MAPE 值	15.1	29.29	8.4	7.79

通过单个模型及加权预测模型在人口总量的预测的精度对比，图 11 中可以看出，不同模型预测时的表现的具有一定的差异性，也存在个别异常时间点在单个模型预测中效果不佳。首先对 3 个模型的预测情况，线性回归模型的预测效果相对较弱，且对数据波动非常敏感。在对比 XGBoost 和 LSTM 的预测效果时，发现它们的预测结果相对接近，而且都具有较好的稳定性。具体来说，LSTM 在预测精度方面表现更出色，但同时 XGBoost 模型的预测精度相对较差。因此本文结合 3 个不同的机器学习模型进行加权对最终结果进行预测，即保持了模型预测稳定的特点，同时也能提高模型的精度，组合模型保持了良好的预测精度，预测结果的精度对比单一模型有很大的提高。

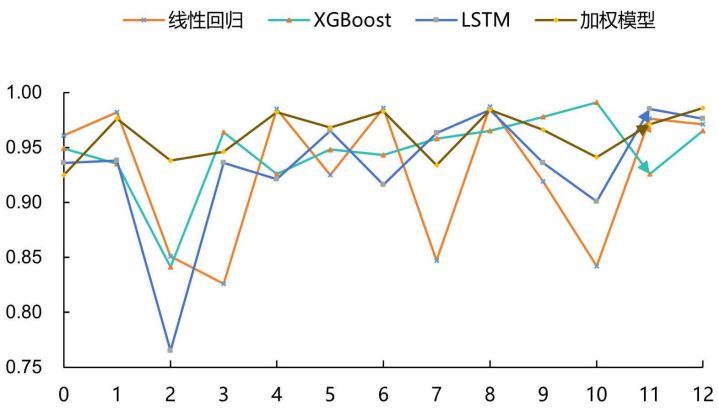

Figure 11. Comparison of the accuracy of different models
图 11. 不同模型的精度对比图

4.3. 基于经济变化的能源消费预测模型

建立与上一节相同的线性回归预测模型，结果如表 8。

Table 8. Linear regression model results
表 8. 线性回归模型结果

	非标准化系数		标准系数	t	P	VIF	R^2	F
	B	标准误差	Beta					
常数	19513.995	1171.040	-	16.665	0.000***	-	0.888	F = 69.864
GDP	0.148	0.018	0.941	8.358	0.000***	1		P = 0.000***

注：***、**、*分别代表 1%、5%、10%的显著性水平。

得到结果，如下所示:

$$y_{i\text{能源}} = 19513.995 + 0.148x_{i\text{人GDP}} \tag{28}$$

根据关系式建立相同的预测模型，得到结果如表 9 所示。

Table 9. Forecast results of energy consumption based on economic changes
表 9. 基于经济变化的能源消费量的预测结果

年份	2021	2022	2023	2024	2025
经济	94782.85650	99656.16244	104529.4684	109402.7744	114276.0803
经济－能源消费量	33257.50720	33964.13656	34670.76590	35377.39528	36084.02465
年份	2056	2057	2058	2059	2060
经济	265348.5646	270221.8708	275095.1767	279968.4827	284841.7884
经济－能源消费量	57989.53490	58696.16422	59402.7936	60109.42296	60816.05234

4.4. 区域碳排放量预测模型的建立

由于多元线性回归在区域碳排放量与人口、GDP 和能源消费量之间的建模效果不理想，本文考虑使用岭回归作为一种改进方法。

岭回归(Ridge Regression)是一种线性回归的扩展方法，旨在应对线性回归中可能出现的多重共线性(multicollinearity)问题。多重共线性指的是在回归分析中，自变量之间存在高度相关性，这可能会导致模型的不稳定性和不准确性。在岭回归中，模型会保留所有的特征变了，但是会减小特征变量的系数值，让特征变量对预测结果的影响变小，岭回归[19]是通过改变其 alpha 参数来控制减小特征变量系数的程度。而这种通过保留全部特征变量，只是降低特征变量的系数值来避免过拟合的方法，称之为 L2 正则化。通过引入正则化项，岭回归的优化目标在于最小化损失函数。由于正则化项的存在，模型在估计系数时会更倾向于将它们缩小，从而减少多重共线性的影响。这有助于提高模型的稳定性和泛化性能。相比于普通线性回归，岭回归能够有效解决多元线性回归中的过拟合问题，并提高模型的稳定性和可靠性。通过压缩模型对特征的依赖程度，岭回归可以更准确地估计变量之间的关系，避免了模型对训练数据中的异常过度敏感的情况。

其目标函数为：

$$\left\| \arg\min_{w} X * w - Y \right\|^2 + \lambda \|w\|^2 \tag{29}$$

岭回归的主要优点是可以稳定模型的参数估计，减少参数估计的方差，从而提高模型的泛化性能。然而，与传统的线性回归相比，岭回归引入了正则化项，可能会导致参数估计的偏差增加。因此，在选择合适的正则化强度参数 α 时需要进行调参。

同样，区域碳排放量预测中采用岭回归模型[19] [20]，有望提高模型的预测精度和稳定性。

对于二范数的岭回归和优化函数，分别如公式(30)、(31)所示：

$$\min_{\theta} \frac{1}{2} \sum_{i=1}^{m} \left(h_{\theta}\left(x^{(i)}\right) - y^{(i)} \right)^2 + \lambda \|\theta\|^2 \tag{30}$$

$$\theta_j = \theta_j - \alpha \sum_{i=1}^{m} \left(h_{\theta}\left(x^{(i)}\right) - y^{(i)} \right) x_j^{(i)} - 2\lambda\theta_j \tag{31}$$

从最小二乘的角度来看，当特征之间存在高度相关性时，会导致矩阵的奇异性，进而使得无法求解回归系数。通过引入 L2 范数正则化项，也被称为岭回归，可以对模型的参数进行约束，使得矩阵保持可逆性。具体而言，L2 范数正则化项的加入，会在目标函数中加入参数向量的平方和，并乘以一个调节参数 λ。这样，当 λ 足够大时，正则化项的影响会压制原始的最小二乘解，从而使得矩阵能够满秩，即可逆。同时有效减小过拟合的风险，提高了模型的鲁棒性和泛化能力。所涉及公式如下所示：

$$\begin{aligned} &\nabla_{\theta} J(\theta) = XX^{\top}\theta - XY + \lambda\theta = 0 \\ &\Rightarrow \theta = \left(XX^{\top} + \lambda I\right)^{-1} XY \end{aligned} \tag{32}$$

利用 SPSSPRO 进行求解，并利用 SPSSPRO 绘制了岭迹图如图 12 所示。同时，模型的拟合优度 R^2 为 0.983，模型表现为较为优秀。

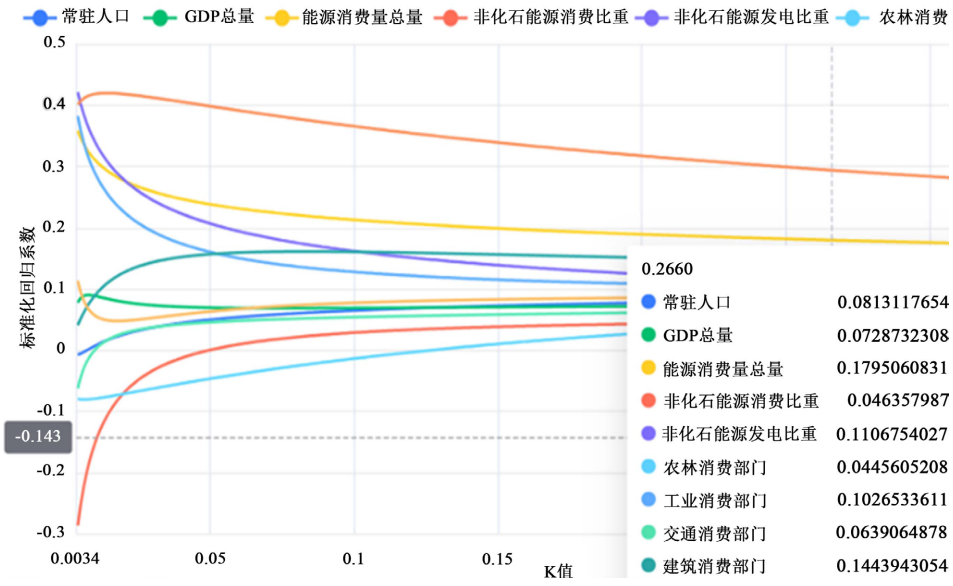

Figure 12. Ridge trace map
图 12. 岭迹图

建立得到，碳排放量与人口、GDP 和能源消费量预测相关联关系式为：

$$\begin{aligned} W_{碳排} = {} & 8512.373 + 1.736x_{i\,人口} + 0.21x_{i\,GDP} + 0.387x_{i\,能耗总量} + 9721.063x_{i\,非化石能源消费} \\ & + 89650.488x_{i\,非化石能源发电} + 1.123x_{i\,农林消费} + 0.363x_{i\,工业消费} + 0.736x_{i\,交通消费} \\ & + 4.632x_{i\,建筑消费} + 1.128x_{i\,居民生活} + 2.308x_{i\,能源供应} \end{aligned} \tag{33}$$

最终，根据关联式采用加权预测模型，得到结果如表 10 所示。

Table 10. The end result
表 10. 最终结果

年份	2021	2022	2023	2024	2025
碳排放量	78473.418	76147.757	77751.046	74155.611	75245.457
年份	2056	2057	2058	2059	2060
碳排放量	67078.84463	66974.95138	66731.62992	66415.67405	65589.38461

5. 结论

本文结合实际背景，考虑了现实世界中各种复杂的因素和不确定性，包括建立数学模型、数据预处理、多种模型预测方法的比较和分析、情景设计、路径规划等多个方面，成功地解决了碳达峰和碳中和目标相关的复杂问题，为推进碳减排和促进可持续发展做出了重要贡献。本文不仅使用了多元线性回归模型，还引入了 LSTM、XGboost 等多种方法，以提高预测精度，并采用加权平均的方法综合考虑各模型的结果，这有助于减小预测值与真实值之间的误差。总而言之，本文考虑了问题的多个方面，采用了多种方法和技术，以全面、系统地解决问题。

参考文献

[1] 舒印彪, 张丽英, 张运洲, 等. 我国电力碳达峰, 碳中和路径研究[J]. 中国工程科学, 2021, 23(6): 1-14.

[2] 胡鞍钢. 中国实现 2030 年前碳达峰目标及主要途径[J]. 北京工业大学学报(社会科学版), 2021, 21(3): 1-15.

[3] 庄贵阳, 魏鸣昕. 城市引领碳达峰, 碳中和的理论和路径[J]. 中国人口·资源与环境, 2021, 31(9): 114-121.

[4] 马冰, 贾凌霄, 于洋, 等. 地球科学与碳中和: 现状与发展方向[J]. 中国地质, 2021, 48(2): 347-358.

[5] 郭春丽, 易信. “双碳”目标下的中国经济增长: 影响机制, 趋势特征及对策建议[J]. 经济学家, 2022, 1(7): 24-33.

[6] 王惠文, 孟洁. 多元线性回归的预测建模方法[J]. 北京航空航天大学学报, 2007, 33(4): 500-504.

[7] 李进强, 喇磊. 基于 Xgboost 算法的国际期货涨跌预测分析[J]. 金融, 2018, 8(5): 211-220.

[8] 张驰, 郭媛, 黎明. 人工神经网络模型发展及应用综述[J]. 计算机工程与应用, 2021, 57(11): 57-69.

[9] 吉培荣, 黄巍松, 胡翔勇. 灰色预测模型特性的研究[J]. 系统工程理论与实践, 2001, 21(9): 105-108, 139.

[10] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[11] 耿丽娟, 李星毅. 用于大数据分类的 KNN 算法研究[J]. 计算机应用研究, 2014, 31(5): 1342-1344.

[12] 朱守先, 梁本凡. 中国城市低碳发展评价综合指标构建与应用[J]. 城市发展研究, 2012(9): 93-98.

[13] 白盛楠, 申晓留. 基于 LSTM 循环神经网络的 $PM_{2.5}$ 预测[J]. 计算机应用与软件, 2019, 36(1): 67-70.

[14] 杨梦晨, 陈旭栋, 蔡鹏, 等. 早期时间序列分类方法研究综述[J]. 华东师范大学学报(自然科学版), 2021(5): 115.

[15] 薛洪波, 伦淑娴. 粒子群算法在多目标优化中的应用综述[J]. 渤海大学学报: 自然科学版, 2009, 30(3): 265-269.

[16] 陈贵敏, 贾建援, 韩琪西. 粒子群优化算法的惯性权值递减策略研究[J]. 西安交通大学学报, 2006, 40(1): 53-56, 61.

[17] 钟进, 李宗航. 基于趋势性时间序列的全国碳排放量预测研究[J]. 运筹与模糊学, 2023, 13(4): 3870-3881.

[18] 赵晋芳, 罗天娥, 曾平, 等. 决定系数的 Bootstrap 可信区间估计[J]. 中国卫生统计, 2014, 31(1): 49-52.

[19] 杨楠. 岭回归分析在解决多重共线性问题中的独特作用[J]. 统计与决策, 2004(3): 14-15.

[20] 狄乾斌, 侯智文, 陈小龙. 基于夜间灯光数据的中国海岛县碳排放时空分异及影响因素研究[J]. 地理与地理信息科学, 2022, 38(6): 23-28.