**基于灰色预测和BP神经网络的全球变暖趋势**

**及其驱动因素研究**

随着全球各地出现越来越多的极端高温天气，人类的生存环境已经受到了严重的威胁。这不由地引发我们深思，是不是全球变暖导致的极端高温现象的出现，因此需借助全球气温、自然灾害等历年数据集，找寻其内部联系，并需要对这些现象做出相对应的措施，改善自然生态环境，构建适宜人类生存的自然环境。

对于问题一，我们首先通过**Mann-Kendall突变**检验方法来对2022年3月全球气温数据进行检验，做出全球平均气温、热带、南温带、北温带平均气温突变检验图，可知突变点均不在2022年3月附近，可否定 2022年3月全球气温上升导致了比过去十年增幅更大的观点。再用**时间序列模型**描述过去，可知未来全球温度水平总体呈上升趋势，再用该模型进行预测，但检验结果拒绝该模型，于是我们引用**灰色预测模型**进行预测，得出时间响应方程，预测结果见附件，再进行准指数规律检验，由此证明模型的合理性。为了使预测结果更精确，我们再采用BP**神经网络模型**进行预测，其中数据采用全球平均气温观测点数据，即全球陆地平均气温，将年份作为输入量，平均气温作为输入量，其中选用Bayesian Regularization算法对模型进行训练，再将所需预测的年份用sim函数进行仿真，得到相应年份的平均气温预测值。最终，我们认为BP神经网络模型更准确，它预测2050年年均温将到达20.66摄氏度，2100年将到达21.17摄氏度。

对于问题二，我们将时间数据拆分为月份和年份，依据经纬度，我们将数据分为四组，即NE、NW、SE和SW，再依据他们1899年到2012的地区平均气温进行**斯皮尔曼相关系数**计算，由此分析全球温度、时间和位置之间的关系。为了使分析结果更准确，我们再采用**多元线性回归**的方法来探究相关因素之间的关系，我们将年份、月份、纬度和经度作为自变量，将平均温度作为自变量，使用最小二乘法对影响因素建立多元线性回归模型，所得结果见表5到表8。最终得出结论：全球气温受纬度影响最大，不同纬度的地区热量条件不同，导致对气温有较大的影响。其次是月份，不同的月份，太阳光照角度不同，导致太阳光照面积和时间不同，太阳光的不同，导致气温变化起伏不定。对于经度来说，对气温影响较小，可能的原因是不同经度所处的地域不同，有陆地和海洋地面的差异，升温条件不同。年份几步不影响全球平均温度水平。我们再采用可视化分析的方法，选择了三个典型的城市，即美国的COVID-19，澳大利亚的森林火灾以及印尼的火山爆发，筛选出他们的数据，并在网上查阅三国发生自然灾害的时间，首先用SPSS对数据进行统计性描述，发现印尼的年平均气温相对稳定，而美国和澳大利亚的年平均气温峰值和谷值相差较大。再用tableau绘制其折线图，再进行**可视化**描述，我们发现存在自然灾害的时候，三国气温均有较大的突变型转换，说明自然灾害对全球气温水平存在影响，美国COVID-19爆发时，气温骤降；澳大利亚发生森林火灾时，气温骤升；印尼发生火山爆发时，气温骤减。然后，我们再采用**主成分分析**法研究影响全球气温因素，通过网上资料查找，我们通过全球的CO2、O3、API、PM2.5、SO2、NO2、CO含量以及全球平均气温2015-2021年的数据，来进行主成分分析，得出 CO2浓度及O3浓度含量对全球温度影响最大。结合两个模型我们可知，影响气温的主要因素为人类活动、纬度位置和环境的突变（包括自然灾害的影响），减缓全球变暖举措见下文。

最后，我们编写了一份非技术性文章，对我们的模型进行解释，以及该模型对全球气温变化的发现和建议。

**关键词：Mann-Kendall突变、时间序列模型、灰色预测模型、BP神经网络模型、斯皮尔曼相关系数、多元线性回归、主成分分析**

目录

[1 问题重述 3](#_Toc120474510)

[1.1问题背景 3](#_Toc120474511)

[1.2问题要求 3](#_Toc120474512)

[2 问题分析 3](#_Toc120474513)

[2.1问题一的分析 3](#_Toc120474514)

[2.2问题二的分析 3](#_Toc120474515)

[3 模型假设 4](#_Toc120474516)

[4 符号说明 4](#_Toc120474517)

[5 数据的预处理 5](#_Toc120474518)

[5.1缺失值处理 5](#_Toc120474519)

[5.2数据标准化 6](#_Toc120474520)

[5.3各地区平均气温计算 6](#_Toc120474521)

[6 问题一模型的建立与求解 6](#_Toc120474522)

[6.1 Mann-Kendall突变分析气温增幅 6](#_Toc120474523)

[6.2预测模型分析未来全球温度水平 8](#_Toc120474524)

[6.2.1 时间序列模型描述过去并预测未来 8](#_Toc120474525)

[6.2.2 灰色预测模型优化预测结果及检验 11](#_Toc120474526)

[6.2.3模型的优化：BP神经网络预测 13](#_Toc120474527)

[6.3预测模型的论述和模型的评价 15](#_Toc120474528)

[6.3.1 预测模型的论述 15](#_Toc120474529)

[6.3.2 模型的评价 15](#_Toc120474530)

[7 问题二模型的建立与求解 15](#_Toc120474531)

[7.1分析影响温度变化的因素 15](#_Toc120474532)

[7.1.1斯皮尔曼相关系数分析温度变化因素 15](#_Toc120474533)

[7.1.2多元线性回归优化相关因素分析 18](#_Toc120474534)

[7.2自然灾害对全球气温影响 20](#_Toc120474535)

[7.2.1可视化分析自然灾害影响 20](#_Toc120474536)

[7.2.2主成分分析研究影响气温变化因素 22](#_Toc120474537)

[7.2.3 建议措施减缓全球变暖 24](#_Toc120474538)

[8 文章 25](#_Toc120474539)

[9模型的评价与改进 25](#_Toc120474540)

[9.1模型的优点 25](#_Toc120474541)

[9.2 模型的缺点 25](#_Toc120474542)

[9.3 模型的改进 25](#_Toc120474543)

[10 相关文献 25](#_Toc120474544)

# 1 问题重述

## 1.1问题背景

随着工业革命的兴起，越来越多的工厂被建立起来，随之而来的是更多的机器二氧化碳等其他温室气体的排放，地球的大气系统无法吸收完这些气体，温室气体在大气中的浓度不断累计与提升，导致的后果是：温室效应不断加剧，全球各地陆续出现温度破新高的危机。这样的情况已经严重威胁到了人类的生存环境，为此，我们应该借助多年来的累积的全球温度数据，利用数据去建立模型，去分析这些现象的背后，存在的全球变暖的真相，并急需对此采取行动，构建更好的自然生态环境。

## 1.2问题要求

基于上述背景下，借助题目所给数据及自行搜集的数据集对下列问题进行回答：

（1）根据附件数据及线上搜索数据来评判2022年3月全球温度上升的影响并建立数学模型描述过去数据及预测2050年与2100年的全球平均气温，求解何时全球平均气温能达到20℃，最后对所建立模型进行比较与评价。

（2）利用上述结果与所获数据集，构建数学模型分析全球气温、时间和地点之间的相关联系，接着搜集自然灾害的历史相关数据，并评判自然灾害是否会对全球气温产生影响。

（3）撰写一篇给组委会的非技术性文章，内容涵盖对数据处理、模型建立后的结果发现，以及针对全球变暖的情况给出针对未来的建议。

# 2 问题分析

## 2.1问题一的分析

我们采用Mann-Kendall突变检验方法来对2022年3月全球气温数据进行检验，做出全球平均气温、热带、南温带、北温带平均气温突变检验图，进而得出结论。我们再用SPSS作出时间序列图，发现数据规律，再用SPSS建立时间序列模型，进而得出预测结果，但由于检验情况较差，我们不采用该方法进行预测数据，仅用它描述过去全球温度水平情况。然后我们引用灰色预测模型进行预测，得出时间响应方程，作出全球平均气温拟合图，由此来分析全球平均气温水平，我们再对它进行准指数规律检验，作出数据光滑度分析图像，由此来证明模型的合理性。为了使预测结果更精确，我们再采用BP神经网络模型进行预测，我其中数据采用全球平均气温观测点数据，即全球陆地平均气温，将年份作为输入量，平均气温作为输入量，其中选用Bayesian Regularization算法对模型进行训练，再将所需预测的年份用sim函数进行仿真，得到相应年份的平均气温预测值。

## 2.2问题二的分析

我们将时间数据拆分为月份和年份，依据经纬度，我们将数据分为四组，即NE、NW、SE和SW，再依据他们1899年到2012的地区平均气温进行斯皮尔曼相关系数计算，由此分析全球温度、时间和位置之间的关系。为了使分析结果更准确，我们再采用多元线性回归的方法来探究相关因素之间的关系，我们将年份、月份、纬度和经度作为自变量，将平均温度作为自变量，使用最小二乘法对影响因素建立多元线性回归模型，根据构建系数结合斯皮尔曼相关系数得出最后结论。我们再采用可视化分析的方法，选择了三个典型的城市，即美国的COVID-19，澳大利亚的森林火灾以及印尼的火山爆发，筛选出他们的数据，并在网上查阅三国发生自然灾害的时间，首先用SPSS对数据进行统计性描述，再用tableau绘制其折线图，从而使得数据更直观，结合所查阅的三国发生自然灾害的时间，分析自然灾害和全球气温变化的原因。然后，我们再采用主成分分析法研究影响全球气温因素，通过网上资料查找，我们通过全球的CO2、O3、API、PM2.5、SO2、NO2、CO含量以及全球平均气温2015-2021年的数据，来进行主成分分析，得出对全球气温影响较大的因素。最后结合该问的两个模型，总结出对全球气温影响最大的因素并提出减缓全球变暖的举措。



图1 流程图

# 3 模型假设

1.假设未来的未知因素对全球气温变化影响较小，可忽略不计。

2.假设我们选取的因素对全球气温变化起着决定性因素的作用，其它因素对模型的影响忽略不计。

3.假设全球观测点平均气温数据均为真实有效。

# 4 符号说明

|  |  |
| --- | --- |
| **Symbol** | **Description** |
|  | Standardized variables |
|  | Standard deviation of data |
|  | Significance level |
|  | Development of gray number |
|  | Endogenous control of the number of ash |
|  | Statistical global average annual temperature data |
|  | The grey prediction model is then based on the cumulative global average annual temperature |

# 5 数据的预处理

## 5.1缺失值处理

在本次数据分析过程中，我们发现平均温度和平均温度不确定性列存在缺失数据，且缺失值仅存在为空值的情况，为了使得数据更准确，我们采用三次样条插值的方法对缺失数据进行处理，我们将每个城市分开进行计算，然后根据每个城市每年不同月份分别进行三次样条插值，由此可得出最后的缺失值，我们建立的构造函数如下：



同时满足以下条件：



通过这个模型，我们可以使数据更准确与完整，最后得出的数据见过见附件1。我们再选取其中一个的三次样条插值图形进行可视化展示，见下图

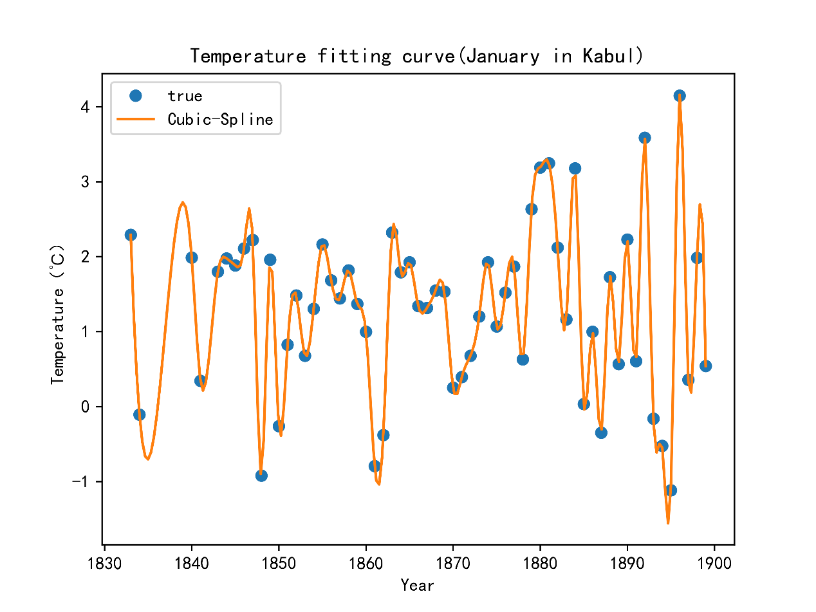


图2 三次样条插值图

## 5.2数据标准化

为消除量纲对数据进行运算时的影响，我们在之前删去无效数据的基础上，对数据的数据再进行数据标准化，使得数据运算时能更加准确。表格中每个样本数据按照公式进行处理：，其中表示标准化后的变量，表示该组数据的均值，表示数据的标准差。其中我们对其中的定性变量如城市和国家进行转换成定量变量处理，处理过程用该行数据除以总数据的频数进行定量数据转换。最后所得的标准化部分数据表如下图所示，完整标准化数据表见附件2**。**

图3

## 5.3各地区平均气温计算

通过搜索从网上找到2012至2020年的气温数据，对表格进行补充。同时将各个地区进行分类，分为北半球、南半球；热带、南温带、北温带及四个特殊典型国家：澳大利亚、印尼、美国和中国。并计算出每个地区1899至2020的每月平均气温。完整数据表见附件3

# 6 问题一模型的建立与求解

## 6.1 Mann-Kendall突变分析气温增幅

不同意2022年3月全球气温上升导致了比过去十年增幅更大。

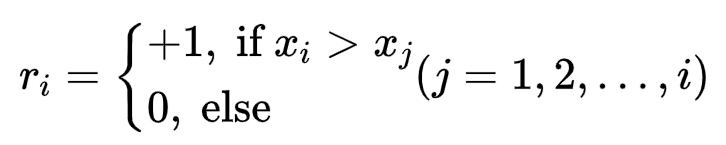
首先，分析得出，如果2022年3月全球气温上升导致了比过去十年间更大的上升，也就是说全球气温在2022年3月产生较大上升幅度，我们小组通过Mann-Kendall突变检验方法来对2022年3月全球气温数据进行检验，判断2022年3月是否为近十年全球气温突变点。

本小题我们采用Mann-Kendall突变检验方法，Mann-Kendall突变检验是一种非参数统计检验方法，有带你在于不仅计算便捷，而且可以明确突变开始的时刻点，并指出突变时间段区域[1]。

对于具有n各样本量的时间序列X，构造一个秩序列：

我们以2012年1月到2022年10月的全球平均气温数据构造秩序列。

其中



可见，秩序列是第个时刻数值大于时刻数值个数的累计数值。

在时间序列随机独立的假定下，定义统计量：

其中，，与是累计数的平均值与方差。在相互独立，且有相同连续分布列时，他们可由以下公式计算得出：

是标准正太分布，它是按时间序列顺序计算出的统计量序列，给定显著性水平α，查正态分布表，若，则表明序列存在明显的趋势变化。

将时间序列生成其对应的逆序列，重复上述计算过程，同时使得

。

首先我们计算顺序时间序列的秩序列，按照上述公式计算；

然后计算逆序时间序列的秩序列，按照上述公式计算；

并给定显著性水平，α=0.01，对于临界值为，将与两个统计量序列曲线与两条直线绘制在一个平面直角坐标系上。

分析绘制出的与曲线图，找出他们在临界线内的交叉点，其对应时刻即为突发开始时间。根据全球平均气温、热带、南温带、北温带平均气温突变检验图如下：

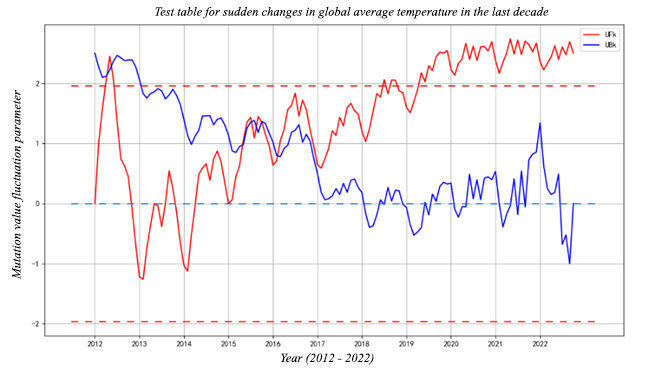


图4 Test table for sudden changes in global average temperature in the last decade

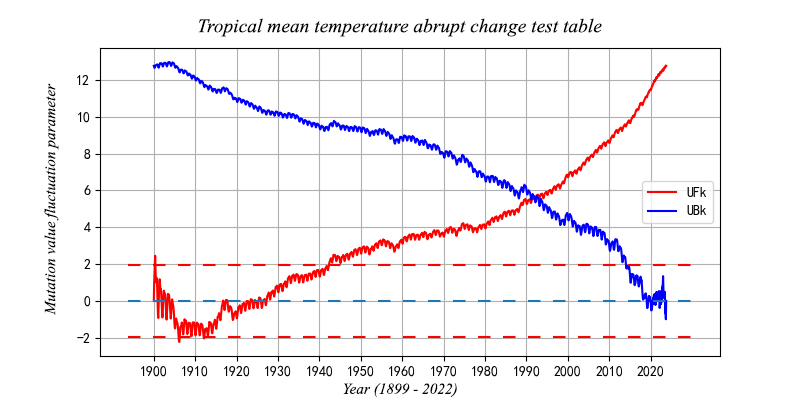


图5 Tropical mean temperature abrupt change test table

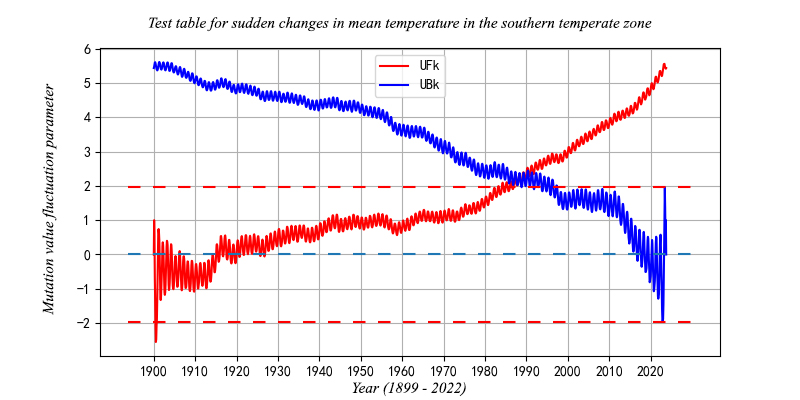


图6 Test table for sudden changes in mean temperature in the southern temperate zone

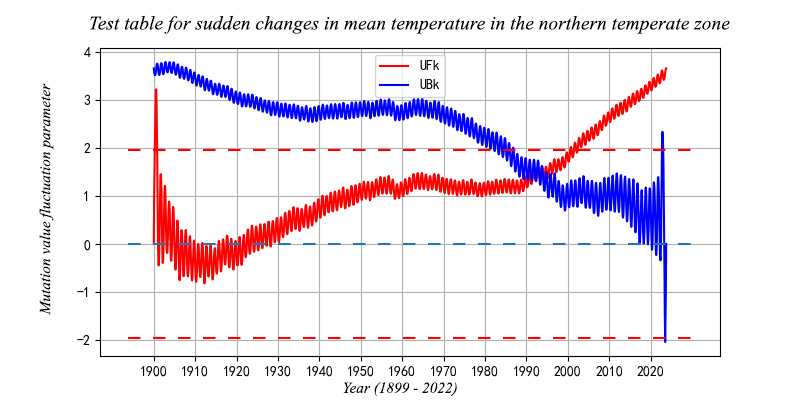


图7 Test table for sudden changes in mean temperature in the northern temperate zone

最终得出结论，突变点均不在2022年3月附近，所以本小组不同意2022年3月全球气温上升导致了比过去十年增幅更大的观点。

## 6.2预测模型分析未来全球温度水平

### 6.2.1 时间序列模型描述过去并预测未来

由于题目要求，描述过去的全球温度水平，所以我们根据数据预处理处理好数据残缺值后的数据进行统计，再结合近几年的全球温度水平和全球观测点气温水平，得出1882年到2015全球平均温度数据和观测点平均气温数据，具体数据见附件4和附件5，对该数据进行时间序列模型分析。我们首先进行时间序列图分析，所建立的图见下图：

图表

描述已自动生成

图8 时间序列图

我们根据图可以发现，数据大致为平稳序列，数据围绕着均值上下波动，无明显的趋势和季节性，我们可以发现，过去的全球温度水平总体呈上升趋势，中间波荡起伏，说明全球温度普遍升高，全球气候变暖现象显著。其中，1882年到1909年全球气温变化不明显，呈略微下降趋势，说明20世纪前工业化水平不高，碳排放少，气候变化不显著。而1975年到2015年上升趋势最大，表现在各个国家高速发展，碳排放增多，导致气候变暖，气候变暖现象显著。于是，我们再用SPSS建立专家建模器，由此得出最佳时间序列模型，我们根据全球平均温度为“年”单位，进行时间序列模型分析，得出最后的预测结果，见附件6。然后我们可以得到它的统计表格，如表1所示。

表1 模型统计表

表格

描述已自动生成

我们发现显著性为0.125，说明该预测结论接受了原假设，对于预测效果存在较大的误差，然后我们所得的样本ACF和PACF图见下图：

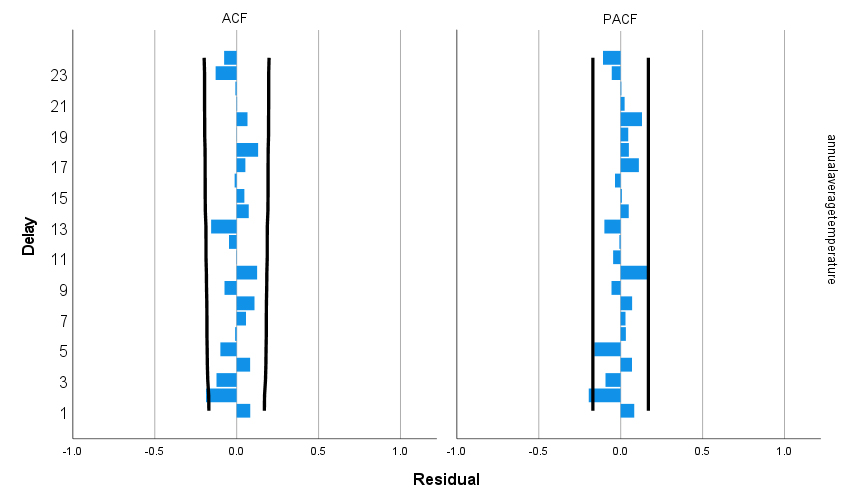


图9 ACF和PACF图

其中ACF是一个完整的自相关函数，可为我们提供具有滞后值的任何序列的自相关值。它描述了该序列的当前值与其过去的值之间的相关程度，ACF在寻找相关性时会考虑所有成分，它描述了一个观测值和另一个观测值之间的自相关，包括直接和间接的相关性信息。PACF 是部分自相关函数或者偏自相关函数。基本上，它不是找到像ACF这样的滞后与当前的相关性，而是找到残差。因此，如果残差中有任何可以由下一个滞后建模的隐藏信息，我们可能会获得良好的相关性，并且在建模时我们会将下一个滞后作为特征[2]。由所得出的值我们可以知道，它的相关性较好，只有部分区域超出接受的范围，说明结果较为准确，但仍然存在一定的误差。我们所得出的拟合以及预测结果见下图：

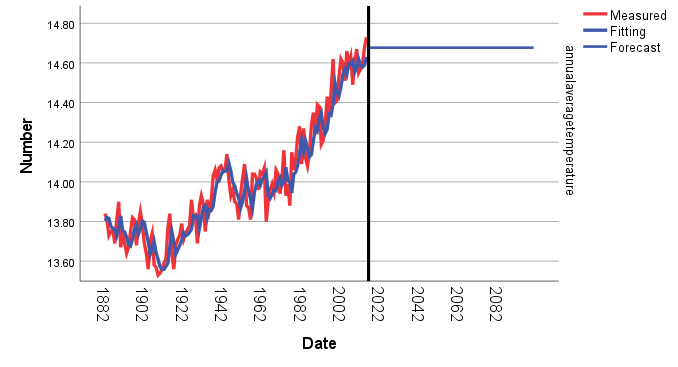


图10时间序列分析拟合图

由所得图像我们可知，拟合结果较为近似，说明明显拟合效果较好，但预测结果不尽人意，原因是时间序列模型不适合长期数据的预测，且操作预测结果较为复杂，于是我们仅用时间序列分析模型来描述过去的全球温度水平，之后全球温度水平的预测方法见下文。

### 6.2.2 灰色预测模型优化预测结果及检验

由于时间序列模型所给出的预测结果存在较大的误差，我们再采用灰色预测模型对预测结果进行优化。首先我们根据题目所给出的国家气温表以及自己搜集的近几年全球平均气温数据建立数据表，具体数据见表2。

表2 全球平均气温表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 年份 | 1882 | 1883 | 1884 | … | 2025 |
| 序号 | 1 | 2 | 3 | … | 134 |
|  | 13.84 | 13.81 | 13.73 | … | 14.73 |
|  | 13.84 | 27.65 | 41.38 | … | 1876.59 |

其中，

凭借此数据我们可建立矩阵如下：



我们可利用最小二乘法得到参数a,b的估计值为



其中，表示发展灰数，表示内生控制灰数。由Matlab代码我们可以解得=0.00049885，=13.5394。

由此，我们可将值代入，推出时间响应方程为：



解得未来全球平均温度拟合图像如下图所示：

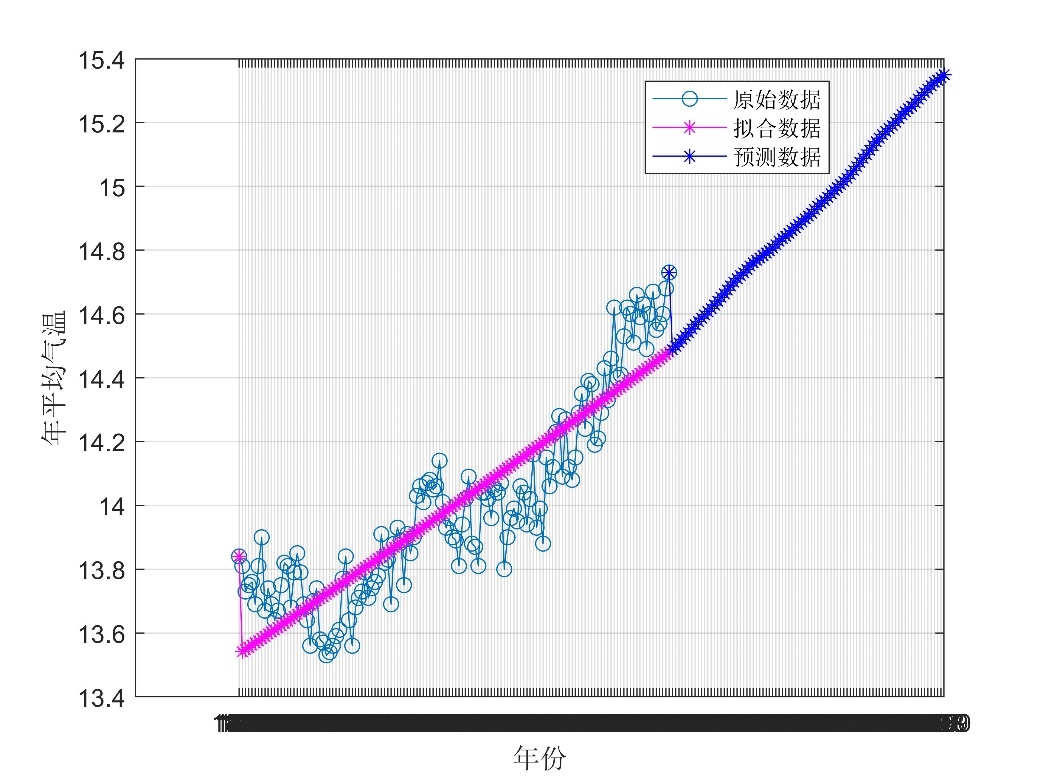


图11 灰色预测拟合图

所得未来全球观测点平均温度预测值见附件7，为了校验模型预测值的精确度，我们对模型进行准指数规律检验，检验过程如下：

(1)数据具有准指数规律是使用灰色系统建模的理论基础。

(2)累加r次的序列为：,定义级比。

(3)如果,且区间长度,则称累加r次后的序列具有准指数规律。

(4)具体到GM(1,1)模型中，我们只需要判断累加一次后的序列是否具有准指数规律。

(5)根据上述公式：序列的级比为：，定义为原始序列的光滑比，注意到，假设为非负序列（生活中的常见的时间序列几乎都满足非负性），那么随着增加，最终会逐渐接近0，因此要使得具有具有准指数规律，即，区间长度，只需要保证即可，此时序列的级比。

我们将上表数据通过Matlab运算可得准指数规律检验的两个指标如下：

指标1：光滑比小于0.5的数据占比为50%，

指标2：除去前两个时期外，光滑比小于0.5的数据占比为100% 。

参考标准：指标1一般要大于50%, 指标2要大于90%，原始数据的光滑度如图12，经比较，上述数据通过了准指数规律检验，故该问题可以建立灰色预测模型。

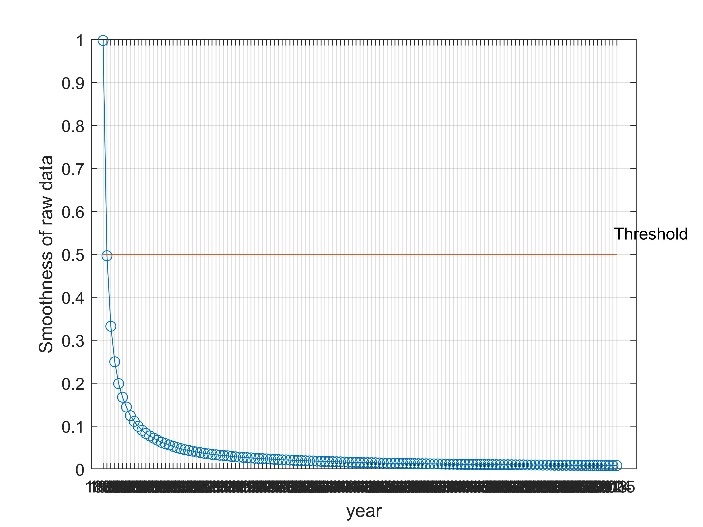


图12 指标平滑度检验图

### 6.2.3模型的优化：BP神经网络预测

由于上述方法求得的结果精确度不高，因此我们运用BP神经网络来对模型进行优化。我们运用BP神经网络模型可以实现通过对数据集的训练得出更为准确的预测结果。是将我们所拥有的数据分成训练集、验证集和测试集，只要运用原始数据训练好神经网络模型，并将我们需要预测的数据集代入运行求解便可直接运用模型得出想要的预测数据[3]。

我们将data3数据中的年份这列数据作为输入量，将全球年平均气温作为输出量（如下图13所示），并取神经隐藏层数为10对神经网络进行训练。

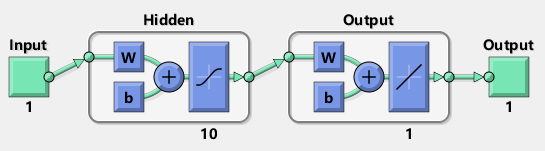


图13 神经元网络图

我们令X为1880年到2015年136个年份数据，Y为1880年到2015年的全球年平均气温的数据，new\_x为2016年到2100年全球的年平均气温的数据。我们用Matlab的神经网络拟合工具箱对数据先进行训练，Matlab按照我们所给的比例70%的训练集、15%的验证集和15%的测试集来随机抽取样本，我们选用Bayesian Regularization方法来训练，因为此算法有执行时修改参数达到结合贝叶斯正则化算法及其避免过拟合的优点。

我们经训练后得出了神经网络训练模型，如下图14所示。我们由图可知，经训练32次后，神经网络模型的均方误差MSE达到最小，故最佳模型对应于最小的MSE，即第三十二次的训练模型。

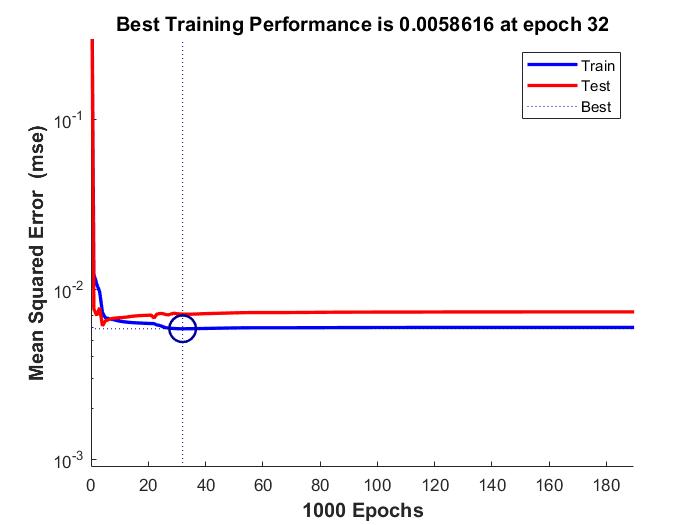


图14 不同训练阶段的MSE变化曲线图

我们将该模型对初始数据训练后所得的拟合值对真实值回归，得出了如下图15所示结果，按照拟合优度越高，对于拟合效果越好的原则，我们可以看出拟合的效果比较好。

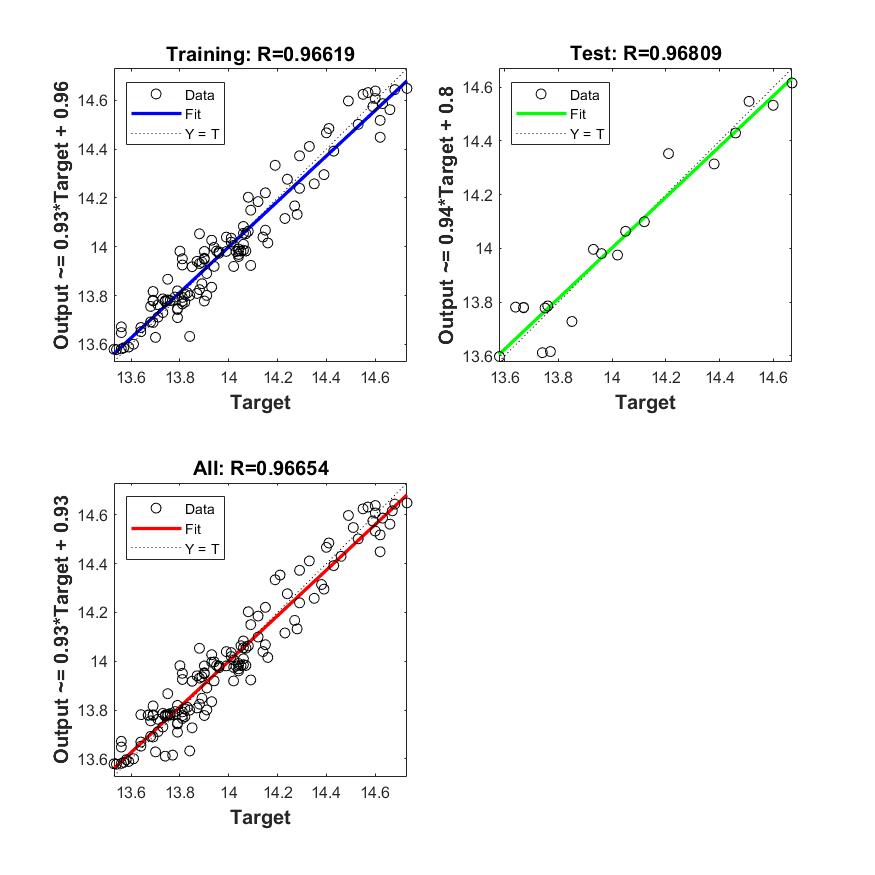


图15 拟合回归曲线图

我们将所要预测的年份数据集所对应的数据运用sim函数代入该训练模型进行仿真求解，得出的预测结果，见附件8。

表3 BP神经网络算法的预测结果

## 6.3预测模型的论述和模型的评价

### 6.3.1 预测模型的论述

我们依据灰色预测模型所得出的结果为：2050年全球平均气温将到达14.84℃，2100年全球平均气温将到达15.35℃，我们预计到2400年左右全球平均气温将到达20℃的高温。我们采用BP神经网络模型得出的结果为：2050年全球平均气温将到达20.66℃，2100年全球平均气温将到达21.17℃。结果如下表4所示。

表4 预测模型概览图

|  |  |  |
| --- | --- | --- |
| Models | Gray prediction model | Neural Network Model |
| Year 2050 | 14.84℃ | 20.66℃ |
| Year 2100 | 15.35℃ | 21.17℃ |
| Year of reaching 20°C | Year 2400 |  |

### 6.3.2 模型的评价

时间序列模型描述过去功能较为完善，但针对于数据的预测，它不适用于对长期连续数据进行预测且当外界遇到巨大变化时，存在较大的偏差。对于灰色预测模型，在处理较少的特征值数据，不需要数据的样本空间足够大，就能解决历史数据少、序列的完整性以及可靠性低的问题，能将无规律的原始数据进行生成得到规律较强的生成序列，但是它只适用于中短期的预测，只适合近似于指数增长的预测，所以在该题还存在误差。我们认为BP神经网络预测模型更准确，它具有强大的非线性映射能力，由于受未来影响因素较多，它适合求解内部机制复杂的问题，且它在在训练时，能够通过学习提取输出、输出数据间的“合理规则”，并自适应的将学习内容记忆于网络的权值中，具有高度自学习和自适应的能力。故我们认为BP神经网络模型更准确。

# 7 问题二模型的建立与求解

## 7.1分析影响温度变化的因素

### 7.1.1斯皮尔曼相关系数分析温度变化因素

为了更好地探究全球温度变化和各个因素之间之间的关系，我们采用Spearman相关系数，分别探究各个变量之间的关系，其中为了更好地区分各个因素对全球温度变化的影响，我们根据纬度，分别分为北半球和南半球，根据经度划分规矩（即20度西经到160度东经） 我们将其划分为东半球和西半球，对于温度的变化我们采取各个地区的平均温度作为判定标准，我们将数据分为四组，分别计算其斯皮尔曼相关系数，即北半球东半球、北半球西半球、南半球东半球和南半球西半球地区。

由于Spearman相关系数是用来描述两个定序数据集的相关性，于是我们先将变量转化为定序数据序列，即对指标评分大小进行排序，以此计算Spearman相关系数，计算公式如下：，其中表示两个指标间评分差。由于样本数>30，我们采用统计量。令，，由此计算检验值，以此判断数据显著性，根据最后所得到的结果，发现数据显著性良好，说明该模型较为准确，具体斯皮尔曼相关系数热力图如下图所示：

图片包含 Teams

描述已自动生成

图16北半球东半球相关系数热力图

图片包含 Teams

描述已自动生成

图17 北半球西半球相关系数热力图

图片包含 Teams

描述已自动生成

图18 南半球东半球相关系数热力图

图片包含 树状图

描述已自动生成

图19 南半球西半球相关系数热力图

根据上面四图的结果我们可知，总体来说，地区所处的纬度对气温影响最大，其次是经度和月份，而年份与全球气温的关系近乎无。对于北半球地区来说，无论东西半球，纬度对气温影响较大，其次是经度，而月份和年份对气温影响相对来说较小。对于南半球地区来说，东西半球影响程度略有不同，对于南半球的东半球地区，纬度影响最大，其次是月份，而年份和经度对气温影响较小；而对于南半球的西半球地区来说，区别最大，不同于任何半球，影响因素最大的是它所处的经度，其次再是纬度，然后月份和年份对气温影响均较小。

### 7.1.2多元线性回归优化相关因素分析

为了使关系结果更加准确，我们采用多元线性回归的方法来进行更精确的分析。多元线性回归的模型方程：

，b0 is a constant term，b1, b2 , ... bk are the regression coefficients.

我们根据题中所给的数据，选取 1899 年 1 月 到 2012 年 12月的地区平均温度数据以及其他相关数据。为了使数据更准确，我们采取和斯皮尔曼系数一样的分组方法，分为四组数据，分别对其进行多元线性回归。其中假设 x1为年份，x2为月份，x3为纬度，x4为经度，x1、x2、x3、x4均为自变量；y是地区平均气温为因变量。然后使用Python来进行多元线性预测。

我们使用最小二乘法对影响因素建立多元线性回归模型[4]，使用 Python 对四个因素建立最小二乘法的模型，分别获得四组数据最小二乘法的线性回归模型，线性回归分析结果表见下四表：

表5 北半球东半球线性回归分析表

表格

描述已自动生成

表6 北半球西半球线性回归分析表

表格

描述已自动生成

表7 南半球东半球线性回归分析表

表格

描述已自动生成

表8 南半球西半球线性回归分析表

表格

描述已自动生成

我们可以发现，从F检验的结果分析可以得到，显著性P值大致为0，水平上呈现显著性, 拒绝回归系数为0的原假设，且对于它的变量共线性表现，VIF均小于10，因此模型没有多重共线性问题，因此我们认为该模型构建较为合理，具体线性表达式如下：

北半球东半球：

北半球西半球：

南半球东半球：

南半球西半球：

根据上述公式我们可以发现，总体来说，全球气温受纬度影响最大，不同纬度的地区热量条件不同，导致对气温有较大的影响。其次是月份，不同的月份，太阳光照角度不同，导致太阳光照面积和时间不同，太阳光的不同，导致气温变化起伏不定。对于经度来说，对气温影响较小，可能的原因是不同经度所处的地域不同，有陆地和海洋地面的差异，升温条件不同。最后，年份对气温的影响微乎其微，可能存在的影响可能是每年突发的自然灾害导致全球气温有略微的变化。

## 7.2自然灾害对全球气温影响

### 7.2.1可视化分析自然灾害影响

由于要研究自然灾害对全球气温的影响，我们选择了三个典型的城市，对他们进行分析，从而得出结论。三个城市包括，美国的COVID-19，澳大利亚的森林火灾以及印尼的火山爆发，我们查阅资料，所得到三个城市分别爆发这些灾害的时间段，并统计在word中，具体统计结果见附件9。然后我们根据题目所给的数据，分别筛选出美国、澳大利亚、印尼这三个国家的数据，对它们从1899的平均气温进行可视化分析，为了使结果更直观，我们首先对三个国家的数据用SPSS进行统计性描述，结果如表9所示：

表9 统计性描述表

表格

描述已自动生成

由该数据我们可以发现，印尼的年平均气温相对稳定，而美国和澳大利亚的年平均气温峰值和谷值相差较大，说明有其他的因素对这两个城市的年平均气温造成较大的影响。为了使数据更直观和清晰，我们根据查阅到的信息，对三个城市的年平均气温分别进行折线图展示，展示结果如图20、21和22所示：

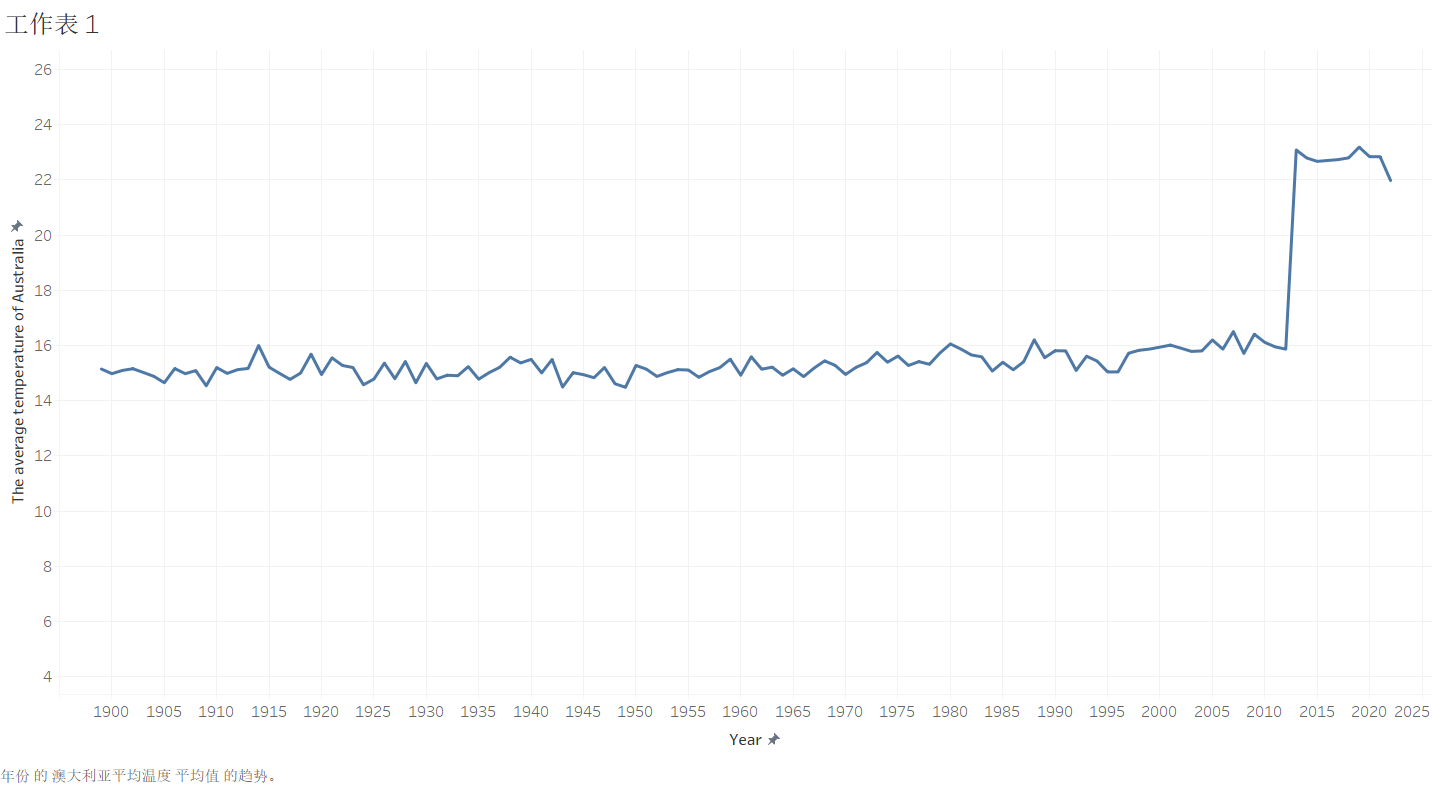


图20 澳大利亚年平均气温折线图

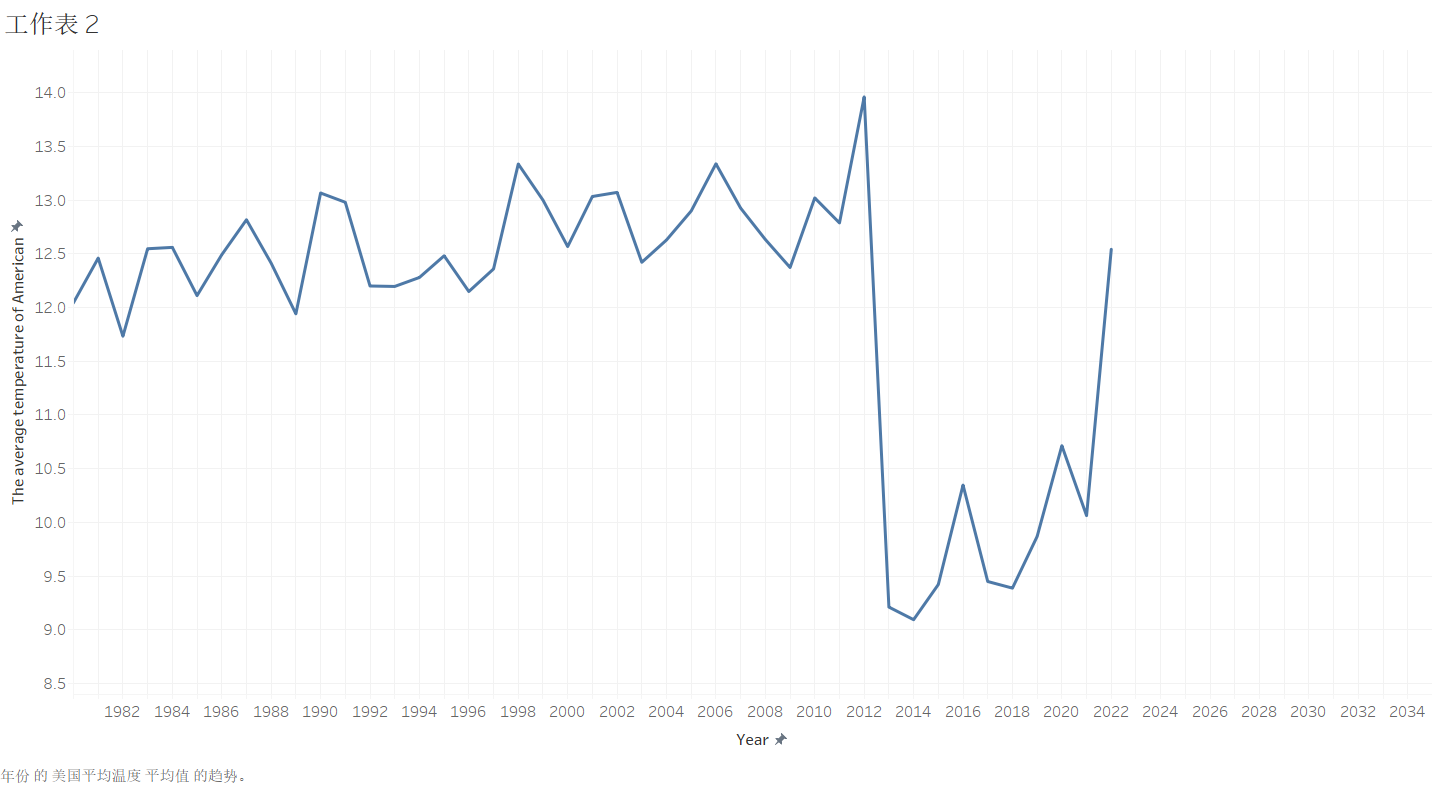


图21 美国年平均气温折线图

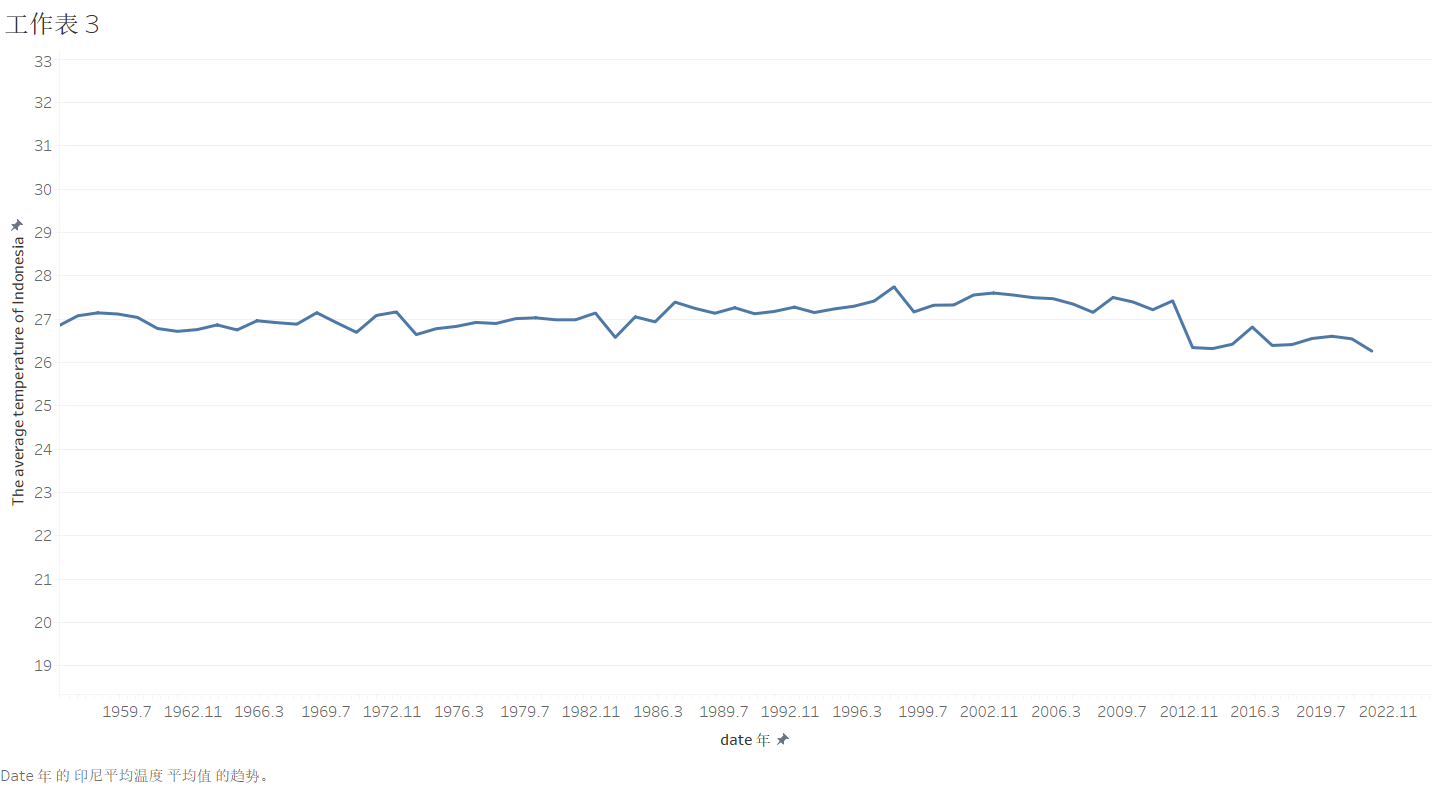


图22 印尼年平均气温折线图

根据三张可视化表格结果前面搜集的数据我们发现：对于美国，2020年出现新冠疫情，导致到2021年气温骤降，可能的原因是疫情导致美国工业化水平停滞不前，减少了工业碳排放，以及人们生活的限制，导致了气温骤降，而2021年到2022年气温的回转，正是疫情好转后，人们生活回到了正规，才使美国年均温渐渐恢复正常，说明疫情对全球气温变化存在影响。对于澳大利亚，在2019年到2020年发生了重大的森林火灾，而我们发现，在这两年澳大利亚的气温急剧上升，森林火灾的发生是全球气温上升的信号，因为气温上升，天气变的高温干燥，导致森林更容易引发火灾，并且火灾时导致局部温度上升，也是气温上升的一个原因。对于印度尼西亚，我们发现它在1963到1964、1982、1983、2012和2014年有火山爆发，而这些时间我们可以发现印尼的气温急剧下降，因为火山爆发会释放出大量炽热的火山灰和气体，进入大气中会阻挡太阳辐射，导致气温降低。并且火山灰颗粒的体积和重量较大，在大气中会滞留，而且还会影响大气环流系统，导致全球气温下降。综上，我们可以发现，自然灾害会影响全球气温变化，且自然灾害发生时影响较大。

### 7.2.2主成分分析研究影响气温变化因素

地球表面温度是由太阳辐射照射到地球表面量和地表反射回去的红外辐射量决定的。大气中的二氧化碳水蒸气、臭氧、一氧化碳、二氧化硫等气体可以使太阳短波辐射轻易通过，但可以吸收地表反射的长波辐射。所以二氧化碳等30多种气体也被称为温室气体，会造成温室效应。下面我们对二氧化碳等气体对气温的影响展开研究。

我们采用PCA主成分分析法研究CO2、O3、API、PM2.5、SO2、NO2、CO七个数据对大陆观测点平均温度影响大小。

我们将n个样本、p个指标，构成大小为n\*p的样本矩阵x：

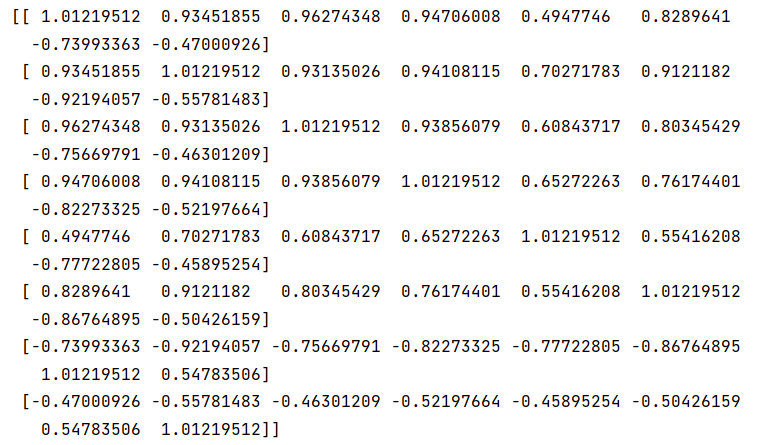
首先我们对数据进行标准化处理：

按列计算均值和标准差，计算得到标准化数据，原本样本矩阵经过标准化变为：

计算标准化样本的协方差矩阵：

其中：

得到协方差矩阵如下：



计算R的特征值及特征向量，对特征值进行降序排序，结果如下图：

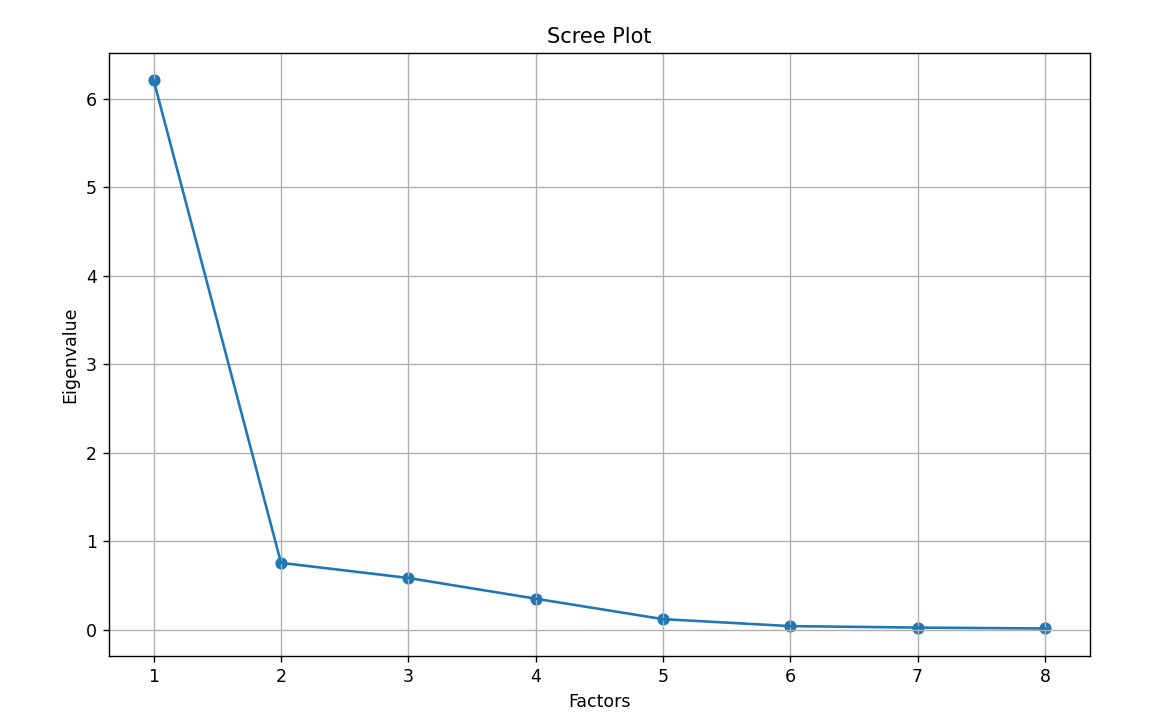
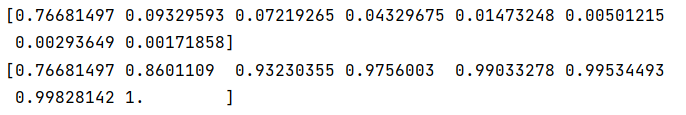


图23 Eigenvalue Scree Plot

计算主成分贡献率及累计贡献率，结果如下：



绘制相关系数热力图如下：

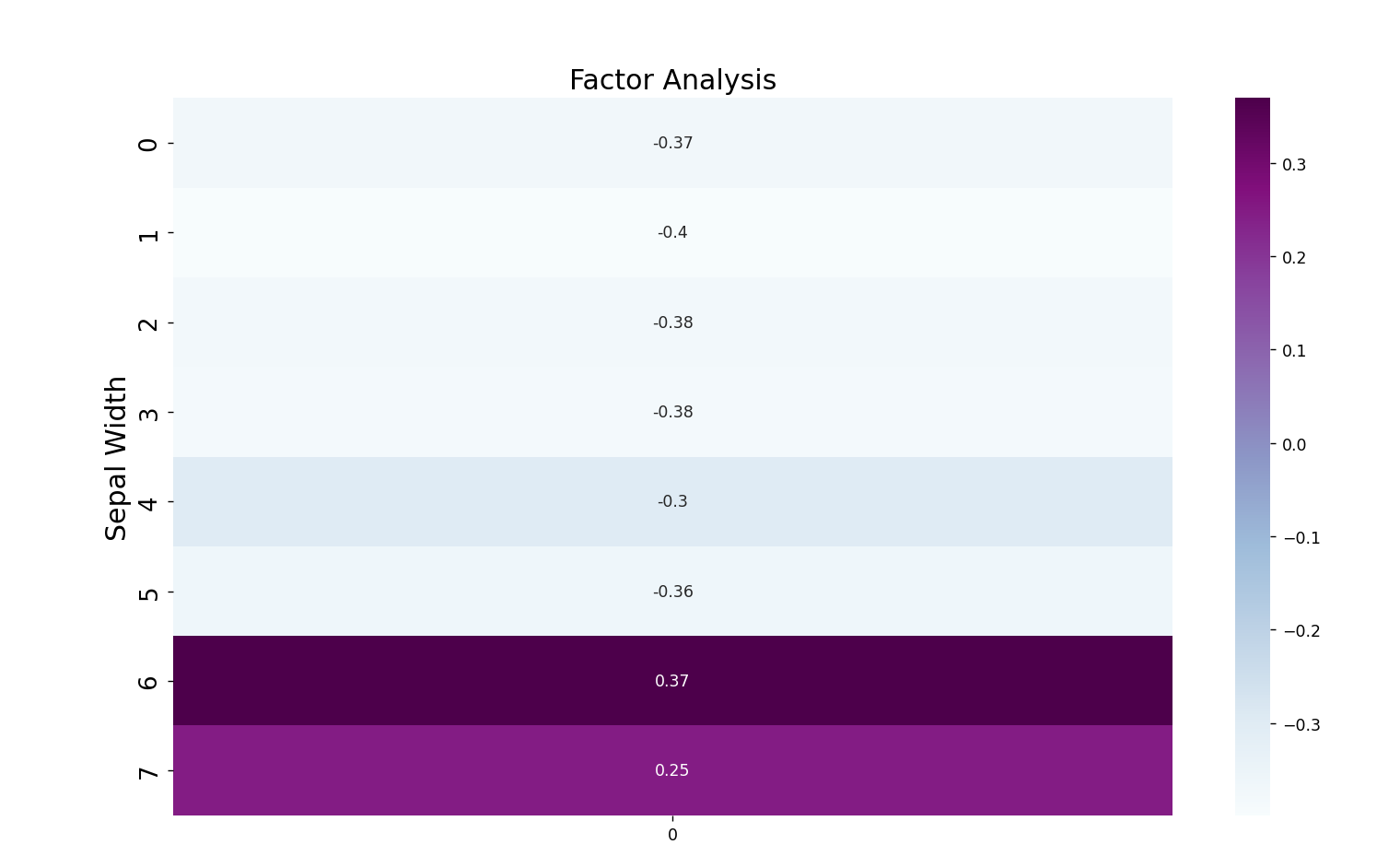


图24 相关系数占比图

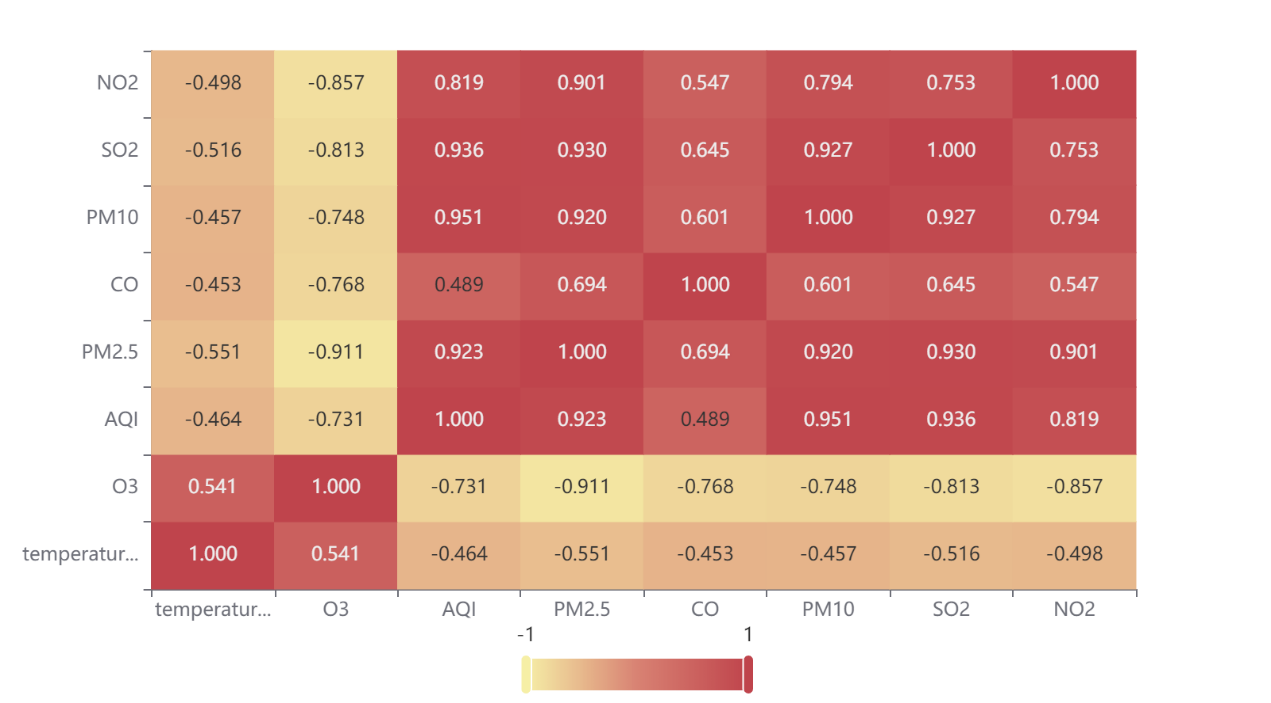


图25 相关系数热力图

通过以上分析，得出CO2浓度及O3浓度对全球平均气温影响最大。

### 7.2.3 建议措施减缓全球变暖

综合以上分析，我们认为，影响气温的主要因素主要有三个，第一个是人为原因，包括人口的激增和人口的一系列工业化活动等，因为人类活动，导致气体如CO2、氧氮化物等排放量激增，导致全球气温变暖。第二个是纬度因素，大气的根本热量来源是太阳辐射，在低纬度地区，太阳辐射多，气温高，在高纬度地区，太阳辐射少，气温低，所以纬度会影响全球气温变化。第三个是环境的突变，如自然灾害火山爆发、森林火灾和疫情等因素，都会导致环境的突变，土地遭到破坏，气象条件收到改变，人为活动受到影响，这些都导致了全球气温的变化。

因此我们提出以下建议：

1. 严格控制人类活动，如汽车出行、工厂排放废气等会产生大量CO2或氧氮化物气体的活动。

2. 减少化石燃料的使用。

3. 植树造林，防风固土。

4. 保护自然环境，减少极端自然灾害的发生。

# 8 文章

# 9模型的评价与改进

## 9.1模型的优点

1.采用BP神经网络对数据进行预测，使结果更准确。

2.对于相关因素相关性分析，我们不仅采用了斯皮尔曼相关系数，还引用了多元线性回归，通过回归方程更好地反映因素之间的关系。

3.用时间序列模型描述过去，用可视化分析形容自然灾害的影响，使论文论述更为直观。

## 9.2 模型的缺点

1.建立模型时，忽略了未来部分因素对自然变化的影响。

2.研究自然灾害对全球气温水平的影响时，找的多为典型案例，缺乏普遍性案例。

## 9.3 模型的改进

1.查阅资料上可以做的更全面，特别是自然灾害可利用普遍性案例，这样能更好地求出它与全球平均气温的相关性。

2.建立预测模型时，可考虑加入未来部分对未来全球平均气温造成影响的因素，增加预测的准确性。

# 10 相关文献

1. 朱瑞杰. 基于Mann-Kendall的咸阳市冬季气温变化趋势及突变分析[J]. 咸阳师范学院学报, 2020, 35(6):8.
2. Alex Tech Bolg，【时间序列】怎么理解ACF 和PACF，<https://blog.csdn.net/qq_41103204/article/details/105810742，2022年11月26>日
3. 何胜甲. 基于GA-BP神经网络模型的深基坑坡顶位移预测研究[J]. 工程技术研究, 2022, 7(15):4.
4. 李冰箫, 张世伟, 郑舒宇,等. 基于多元线性回归与ARIMA组合模型的水电功率预测研究[J]. 科学技术创新, 2022(33):4.

**文章**

全球变暖所引发的问题已经以不同的方式越来越多地出现在全球各地，我们应该对此予以高度关注。我们通过10年来的温度数据分析气温增幅的程度，已经发现全球变暖的趋势已越来越明显，气温增幅已不容忽视。借助19世纪末期至今的观测点的平均气温数据，我们借助预测模型，发现陆地观测点的气温在2050年的气温会达到20.66℃，在2100年会达到21.17℃，按照这样的趋势对于我们每个人来说，只会感觉一年会比一年热。那么我们必须找出全球变暖的原因，并对此采取相应的措施。

于是，我们先分析所搜集数据中温度变化的具体因素，我们发现纬度、月份的不同，对温度变化的体现也不尽相同，但是这些因素却起着较大作用。在夏季炎热、秋季干燥的适合，非常容易引起诸如火山喷发、森林火灾等自然灾害，之后自然灾害所造成的破坏又会反过来加剧气温的升高，形成恶性循环。故对此我们急需采取措施改善现状。

对此，我们团队决定从以下几个方面给出建议：

1. 倡导低碳生活、提高节能意识、减少碳足迹

目前,人们对低碳经济的了解程度不高,高碳的消费理念依然深入人心。首先树立低碳观念和节能意识,通过电视、报纸、网络等媒体进行宣传,营造浓厚的社会环保氛围,提升大家的环保意识。

2.开展植树造林、增加森林碳汇

首先保护现有森林资源,扩大森林保护面积,减少森林的破坏。其次,发动广大民众积极植树,拓宽森林植被面积。通过增加改良草地、治理退化、沙化和碱化草地,治理荒漠化土地等,提高绿化面积。

3.加强科技创新、节约传统能源、开发新能源、优化能源结构

大力促进节能技术的开发与研究,注重与国际合作与交流,提高传统能源的利用效率。极开发我国的绿色能源:太阳能、风能、水能、核能、潮汐能、地热能等。减少煤炭在能源消费结构中的比例,进一步优化能源结构,促进能源结构多样化。

4.大力推进减排重点工程

落实减排任务和责任,实行强制措施和惩罚机制,在推进减排重点工程中,设立专项资金和技术支持,保证各项任务顺利完成。

Appendix

Appendix 1:

Description: A list of documents supporting materials

1. Annex 1: Data table after processing missing values by cubic spline interpolation
2. Annex 2: Data table after data standardization
3. Annex 3: Average temperature data table by region
4. Annex 4: Global average temperature data table
5. Annex 5: Observation Point Average Temperature Data Table
6. Annex 6: Time Series Model Forecast Results
7. Annex 7: Grey projections of future global average temperature at observations
8. Annex 8: BP neural network predictions of future global average temperature at observation points
9. Annex 9: Timing of natural disasters in three cities

Appendix 2:

Description: List of specific code files

1. Attachment 1: Python implements cubic spline interpolation code
2. Annex 2: Python implements standardized data code
3. Annex 3: MATLAB calculates the average temperature codes for each region
4. Annex 4: Python implements the Mann-Kendall mutation analysis code
5. Annex 5: Time Series Model and Grey Forecast Code
6. Annex 6: BP Neural Network Prediction Code
7. Annex 7: Spearman correlation analysis code
8. Annex 8: Multiple Linear Regression Optimization Model Code
9. Annex 9: Principal Component Analysis code

Due to the huge amount of code, only some of the code is shown below, please see the attachment for the detailed code.

附录

附录1：

介绍：支撑材料的文件列表

1. 附件1：通过三次样条插值处理缺失值后的数据表
2. 附件2：数据标准化后的数据表
3. 附件3：各地区平均气温数据表
4. 附件4：全球平均气温数据表
5. 附件5：观测点平均气温数据表
6. 附件6：时间序列模型预测结果
7. 附件7：灰色预测得出的未来全球观测点平均温度预测值
8. 附件8：BP神经网络得出的未来全球观测点平均温度预测值
9. 附件9：三个城市自然灾害发生时间

附录2：

介绍：具体代码文件列表

1. 附件1：python实现三次样条插值代码
2. 附件2：python实现标准化数据代码
3. 附件3：matlab计算各地区平均气温代码
4. 附件4：python实现Mann-Kendall突变分析代码
5. 附件5：时间序列模型及灰色预测代码
6. 附件6：BP神经网络预测代码
7. 附件7：斯皮尔曼相关分析代码
8. 附件8：多元线性回归优化模型代码
9. 附件9：主成分分析

因代码量庞大，以下只展示部分代码，详细代码请见附件。

Standardize data codes：

**from** sklearn **import** preprocessingdf = preprocessing.scale(df)  
print(df)

Cubic spline interpolation code：

**def** calculateEquationParameters(x):  
 *# parameter为二维数组，用来存放参数，sizeOfInterval是用来存放区间的个数* parameter = []  
 sizeOfInterval = len(x) - 1;  
 i = 1  
 *# 首先输入方程两边相邻节点处函数值相等的方程为2n-2个方程* **while** i < len(x) - 1:  
 data = init(sizeOfInterval \* 4)  
 data[(i - 1) \* 4] = x[i] \* x[i] \* x[i]  
 data[(i - 1) \* 4 + 1] = x[i] \* x[i]  
 data[(i - 1) \* 4 + 2] = x[i]  
 data[(i - 1) \* 4 + 3] = 1  
 data1 = init(sizeOfInterval \* 4)  
 data1[i \* 4] = x[i] \* x[i] \* x[i]  
 data1[i \* 4 + 1] = x[i] \* x[i]  
 data1[i \* 4 + 2] = x[i]  
 data1[i \* 4 + 3] = 1  
 temp = data[2:]  
 parameter.append(temp)  
 temp = data1[2:]  
 parameter.append(temp)  
 i += 1  
 *# 输入端点处的函数值。为两个方程, 加上前面的2n - 2个方程，一共2n个方程* data = init(sizeOfInterval \* 4 - 2)  
 data[0] = x[0]  
 data[1] = 1  
 parameter.append(data)  
 data = init(sizeOfInterval \* 4)  
 data[(sizeOfInterval - 1) \* 4] = x[-1] \* x[-1] \* x[-1]  
 data[(sizeOfInterval - 1) \* 4 + 1] = x[-1] \* x[-1]  
 data[(sizeOfInterval - 1) \* 4 + 2] = x[-1]  
 data[(sizeOfInterval - 1) \* 4 + 3] = 1  
 temp = data[2:]  
 parameter.append(temp)

Python implements the Mann-Kendall mutation analysis code：

**def** mktest(inputdata):  
 **import** numpy **as** np  
 inputdata = np.array(inputdata)  
 n = inputdata.shape[0]  
 Sk = np.zeros(n)  
 UFk = np.zeros(n)  
 r = 0  
 **for** i **in** range(1, n):  
 **for** j **in** range(i):  
 **if** inputdata[i] > inputdata[j]:  
 r = r + 1  
 Sk[i] = r  
 E = (i + 1) \* i / 4  
 Var = (i + 1) \* i \* (2 \* (i + 1) + 5) / 72  
 UFk[i] = (Sk[i] - E) / np.sqrt(Var)  
  
 Sk2 = np.zeros(n)  
 UBk = np.zeros(n)  
 inputdataT = inputdata[::-1]  
 r = 0  
 **for** i **in** range(1, n):  
 **for** j **in** range(i):  
 **if** inputdataT[i] > inputdataT[j]:  
 r = r + 1  
 Sk2[i] = r  
 E = (i + 1) \* (i / 4)  
 Var = (i + 1) \* i \* (2 \* (i + 1) + 5) / 72  
 UBk[i] = -(Sk2[i] - E) / np.sqrt(Var)  
 UBk2 = UBk[::-1]  
 **return** UFk, UBk2

Time Series Model and Grey Forecast Code：

*# 1、数据均值化处理*x\_mean = x.mean(axis=1)  
**for** i **in** range(x.index.size):  
 x.iloc[i, :] = x.iloc[i, :] / x\_mean[i]  
  
*# 1、数据差值化处理*x = (x - x.min()) / (x.max() - x.min())  
x = x.T  
  
*# 1、数据初值化处理*x\_mean = x.mean(axis=1)  
**for** i **in** range(x.index.size):  
 x.iloc[i, :] = x.iloc[i, :] / x.iloc[i, 0]  
  
*# 2、提取参考队列和比较队列*ck = x.iloc[0, :]  
print(**" 参考队列："**, ck)  
cp = x.iloc[1:, :]  
print(**" 参考队列："**, cp)  
  
*# 比较队列与参考队列相减*t = pd.DataFrame()  
**for** j **in** range(cp.index.size):  
 temp = pd.Series(cp.iloc[j, :] - ck)  
 t = t.append(temp, ignore\_index=**True**)  
  
*# 求最大差和最小差*mmax = t.abs().max().max()  
mmin = t.abs().min().min()  
rho = 0.4  
  
*# 3、求关联系数*ksi = ((mmin + rho \* mmax) / (abs(t) + rho \* mmax))  
  
*# 4、求关联度*r = ksi.sum(axis=1) / ksi.columns.size  
  
*# 5、关联度排序，得到结果*result = r.sort\_values(ascending=**False**)

Spearman correlation analysis code：

myfont = FontProperties(fname=**r'C:\Windows\Fonts\simhei.ttf'**, size=40)  
sns.set(font=myfont.get\_name(), color\_codes=**True**)  
*# corr = df.corr(method='pearson') # 使用皮尔逊系数计算列与列的相关性  
# corr = df.corr(method='kendall') # 肯德尔秩相关系数*data\_corr = data.corr(method=**'spearman'**) *# 斯皮尔曼秩相关系数  
# data\_corr = data.corr(method='pearson') # 使用皮尔逊系数计算列与列的相关性*plt.figure(figsize=(20, 15)) *# figsize可以规定热力图大小*fig = sns.heatmap(data\_corr, annot=**True**, fmt=**'.2g'**, annot\_kws={**'fontsize'**: 20}) *# annot为热力图上显示数据；fmt='.2g'为数据保留两位有效数字*print(fig)  
fig.get\_figure().savefig(location + **'\_S.png'**) *# 保留图片*

Principal Component Analysis code：

*# 求解系数相关矩阵*covX = np.around(np.corrcoef(df.T), decimals=3)  
*# print(covX)  
  
# 求解特征值和特征向量*featValue, featVec = np.linalg.eig(covX.T) *# 求解系数相关矩阵的特征值和特征向量  
# print(featValue, featVec)*

*# 去中心化*data\_adjust = df - avgs  
*# print(data\_adjust)  
  
# 协方差阵*covX = np.cov(data\_adjust.T) *# 计算协方差矩阵  
# print('协方差矩阵')  
# print(covX)  
  
# 计算协方差阵的特征值和特征向量*featValue, featVec = np.linalg.eig(covX) *# 求解协方差矩阵的特征值和特征向量*

*# 绘制散点图和折线图  
# 同样的数据绘制散点图和折线图*plt.scatter(range(1, df.shape[1] + 1), featValue)  
plt.plot(range(1, df.shape[1] + 1), featValue)  
  
*# 显示图的标题和xy轴的名字  
# 最好使用英文，中文可能乱码*plt.title(**"Scree Plot"**)  
plt.xlabel(**"Factors"**)  
plt.ylabel(**"Eigenvalue"**)  
  
plt.grid() *# 显示网格*plt.show() *# 显示图形  
  
# 求特征值的贡献度*gx = featValue / np.sum(featValue)  
  
*# print(gx)  
  
# 求特征值的累计贡献度*lg = np.cumsum(gx)  
*# print(lg)  
  
# 选出主成分*k = [i **for** i **in** range(len(lg)) **if** lg[i] < 0.85]  
k = list(k)  
*# print(k)  
  
# 选出主成分对应的特征向量矩阵*selectVec = np.matrix(featVec.T[k]).T  
selectVe = selectVec \* (-1)  
*# print(selectVec)  
  
# 主成分得分*finalData = np.dot(data\_adjust, selectVec)  
print(finalData)