**基于季节ARIMA模型的世界平均气温序列分析**

姓名：李豪文

学号：2016061018

**摘要：**使用Python统计分析工具，对1960年到2012年的世界月平均气温序列进行统计分析，并对该动态数据进行建模和预测。采用差分平稳化方法对世界月平均气温序列进行分析，发现世界月平均气温序列具有一定的上升趋势和明显的以12个月为周期的季节性。因此建立季节ARIMA模型，并进行参数估计，优化模型。结果显示，季节ARIMA模型充分提取了世界月平均气温序列的趋势信息与季节。并对2011年1月到2012年12月的世界月平均气温进行预测，均方误差值为 0.08，拟合效果较好。

**关键词：**时间序列；ARIMA模型；世界月平均气温

**问题背景**

一些人认为气候变化是当前时代最大的威胁，他们认为全球[气候变暖](https://baike.baidu.com/item/%E6%B0%94%E5%80%99%E5%8F%98%E6%9A%96)是一种和自然有关的现象，是由于温室效应不断积累，导致地气系统吸收与发射的能量不平衡，能量不断在地气系统累积，从而导致温度上升，造成全球气候变暖。为检验这一说法，本文收集了1960年～2012年（共 53年）世界月平均气温[2]的动态数据，结合时间序列分析的方法来研究分析。

时间序列，又称为动态数据，是指随时间的顺序记录的一列有序数据。对时间序列进行观察、研究，寻找它的变化发展的规律，预测它将来的走势，就是时间序列分析[3]。时间序列建模的方法很多，求和自回归移动平均（ARIMA）模型就是其中一种。本文先对收集到的世界月平均气温的动态数据进行分析，观察其变化规律，建立季节模型；然后再对该序列进行模型识别、模型检验以及基于该模型的预测。在实际生产和生活中， 对于温度的研究和预测具有很大的意义。

1. **基本原理**

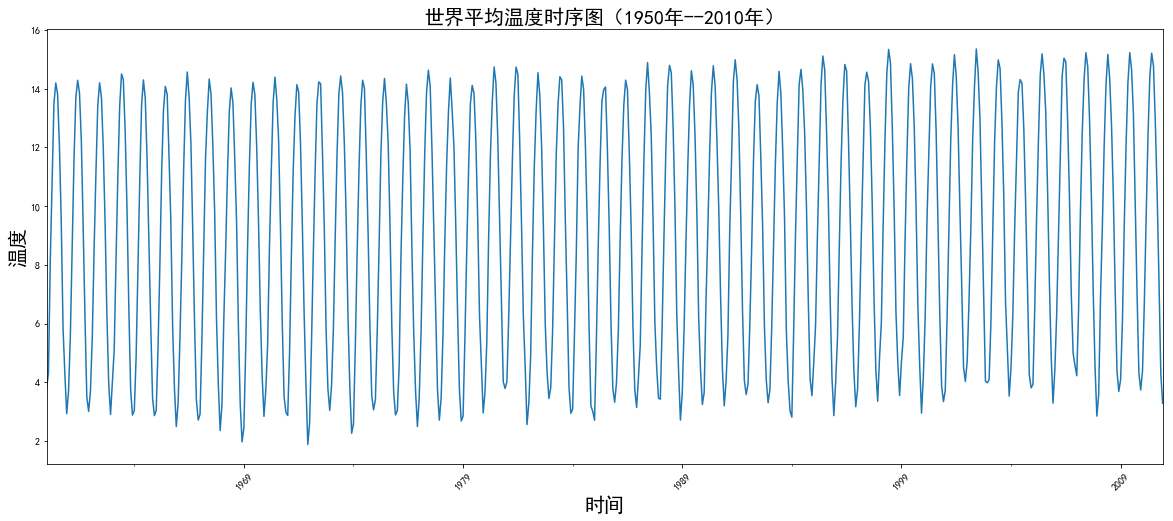
时间序列分析是处理动态数据的一种比较有效的时域分析方法，通过观察动态数据的变化规律，继而对数据进行控制并预测时间序列的未来变化趋势。

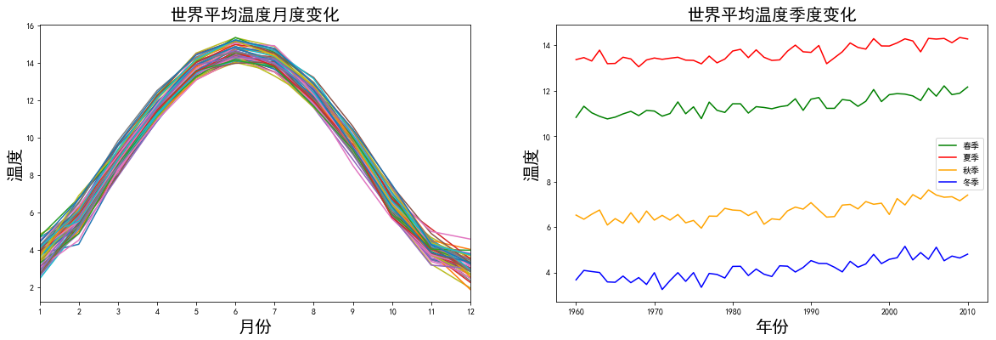
* 1. **模型[1]**  
      求和自回归移动平均模型（auto regressive integrated moving average，简记为）拟合的是差分平稳序列，实际上是差分运算与ARMA模型的结合。差分运算具有强大的确定性信息提取能力，许多非平稳序列差分后会显示出平稳序列的性质，对差分平稳序列可以使用模型进行拟合。对于含有非季节性的时间序列进行建模，可以使用模型。它可以通过适当的阶差分运算使得序列平稳。设为一个等时间间隔为的时间序列，模型具有如下结构：

式中，是平稳可逆模型的自回归系数多项式；, 是平稳可逆模型的移动平滑系数多项式。 为非平稳序列，d为差分运算的阶数，为零均值白噪声序列。

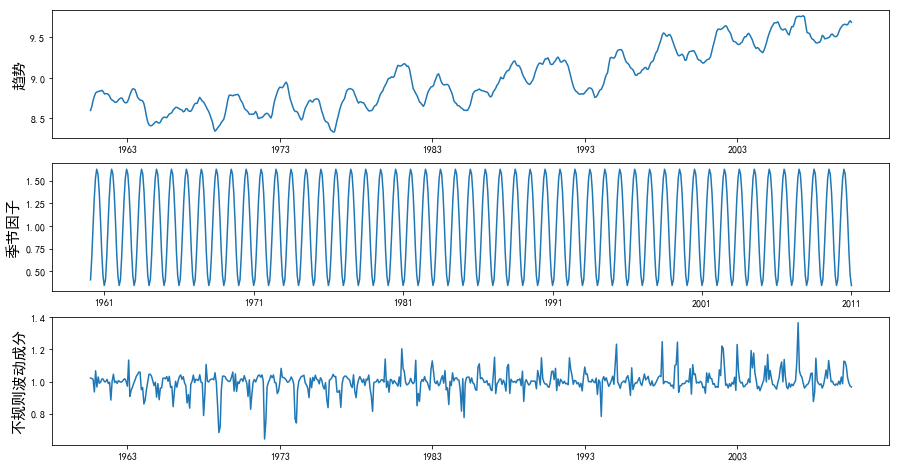
* 1. **模型[1]**  
      模型可以对具有季节效应的序列建模。季节性是指在某一个固定的时间间隔上，序列的某种特征重复出现，温度的月平均数据序列具有较为明显的周期性变化。关于季节性周期的设定，一般根据序列的实际背景设定。一般月度资料的时间序列的季节周期为12个月；季度资料的时间序列的季节周期为4个季。  
      对于含有明显季节因素的时间序列（如本文的世界月平均温度），ARIMA模型也可以对其建模。简单的季节模型通常通过以周期为步长的差分运算就可以将该时间序列中的季节信息提取充分，使该时序变成平稳序列，它的残差序列也是一平稳序列。而常见情况是，序列的季节效应、长期趋势效应和随机波动之间存在复杂的交互影响关系，简单的ARIMA模型并不足以提取其中的相关关系，这时通常考虑乘积季节模型。当序列具有短期相关性时，通常可以使用低阶模型提取。当序列具有季节效应，季节效应本身还具有相关性时，季节相关性可以使用以周期步长为单位的模型提取。由于短期相关性和季节相关性之间具有乘积关系，所以拟合模型实际是和的乘积，也即乘积季节模型()的完整结构如下：  
     式中
  2. **建模步骤**  
      模型建模一般分为3个阶段：模型识别、参数估计和模型检验。建立模型时要求预处理后的时间序列是平稳的，所以在建立模型之前必须对时间序列进行平稳性检验，以保证处理好的数据具有平稳性。对处理后时间序列用上述3个步骤反复调用，直到模型通过检验为止。
  3. **AIC准则[1]**  
      Akaike的信息准则（Akaike Information Criteion）是适用性较广泛的准则，即AIC准则，可用来确定优模型。。AIC准则是拟合精度和参数个数的加权函数，定义如下：  
     AIC准则同时体现了残差不相关原则和模型的简介原则相结合，且能够排除建模者的主观因素。使AIC函数达到最小模型则被认为是最优模型。
  4. **Ljung-Box检验[1]**  
      Ljung-Box test是对随机性的检验,或者说是对时间序列是否存在滞后相关的一种统计检验。对于滞后相关的检验，我们常常采用的方法还包括计算ACF和PCAF并观察其图像，但是无论是ACF还是PACF都仅仅考虑是否存在某一特定滞后阶数的相关。LB检验则是基于一系列滞后阶数，判断序列总体的相关性或者说随机性是否存在。  
      时间序列中一个最基本的模型就是高斯白噪声序列。而对于ARIMA模型，其残差被假定为高斯白噪声序列，所以当我们用ARIMA模型去拟合数据时，拟合后我们要对残差的估计序列进行LB检验，判断其是否是高斯白噪声，如果不是，那么就说明ARIMA模型也许并不是一个适合样本的模型。

1. **实例应用**  
    本文收集了1960年～2012年（共 53年）世界月平均气温的动态数据，以下将对该时间序列进行分析。
   1. 数据特征描述



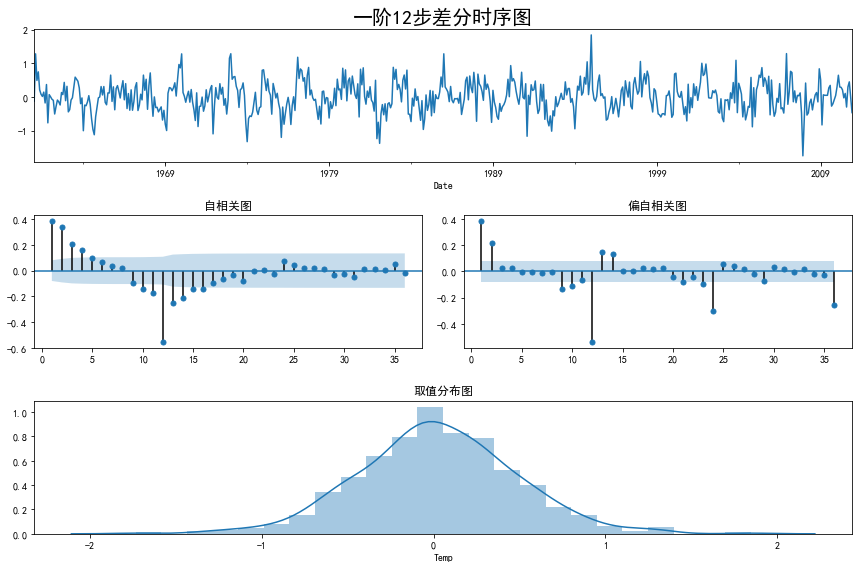
**图2.1.1：**世界平均气温时序图  
  
**图2.1.2：**世界平均气温的季节性

从图2.1.1，图2.1.2可以看到，该时间序列明显具的季节性，并且可以观察到序列有一定的上涨趋势，因此该序列为非平稳序列。较低的温度在12月和1月左右，较高的温度在6月和7月之间。  
 为了验证序列是否有上升趋势，绘制该时间序列各成分的分解，如图2.1.3。可以看到该序列有一个比较明显的有规律的波动：季节波动。并且可以看出近50年来气候确实有上升的趋势。世界平均气温总体上变化幅度不大，从1960年以来上升了1度。但是从水文的角度来看，这1度的上升带来的改变是明显的。这里也侧面证实了人们对全球变暖的担心。



**图2.1.3：**时间序列各成分的分解

* 1. **差分平稳化**  
      综合2.1的结果，世界月平均气温序列有一定的上升趋势性，且具有很强的周期性，蕴含着以年为固定周期的周期性。通过消除数据的周期性，计算其自相关函数和偏自相关函数，以尝试用季节性ARIMA模型拟合时间序列。对序列进行12步的周期差分，希望提取趋势效应和季节效应。差分后序列时序图如图2.1.1所示。



**图2.2.1：**1阶12步差分序列结果

时序图显示差分后序列类似平稳，再对差分后序列进行单位根检验，如表2.2.1。结果表明ADF值小于0.0001，明显小于显著性水平，故认为该差分后序列平稳。如果样本的自相关、偏自相关图既不截尾也不拖尾，在相应于周期S的整倍数点上，自相关（或偏自相关）函数出现绝对值相当大的峰值并呈现振荡变化，就可以判断数据序列适用于乘积季节模型。

**表2.2.1：**ADF检验结果

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Test Statistic | Prob |
| ADF Value |  | -8.5188 | <0.0001 |
| Critical Value | 1% level | -3.4416 |  |
| 5% level | -2.8665 |  |
| 10% level | -2.5694 |  |

* 1. **模型识别与参数估计**  
      首先考虑1阶12步差分之后序列12阶以内的自相关系数和偏自相关系数的特征，以确定短期相关模型。如图2.2.1中自相关系数图和偏自相关图显示12阶以内的自相关系数和偏自相关系数均不截尾，所以尝试使用模型提取差分后序列短期自相关信息。并且根据自相关图和偏自相关图，选取.  
      再考虑季节自相关特征，这时考察延迟12、24阶等以周期长度为单位的自相关系数和偏自相关系数的特征。可以看到延迟12阶自相关系数显著非零，但是延迟24阶以后自相关系数落入2倍标准差范围内。而偏自相关图显示延迟12阶和延迟24阶等偏自相关系数都显著非零，并有递减趋势。因此，可以认为季节自相关特征是自相关系数截尾，偏自相关系数拖尾，这时以12步为周期的模型提取差分后序列的季节自相关信息。  
      综上，确定要拟合的模型为. 然后根据AIC准测选取最好的模型，不同阶数下拟合模型的样本决定系数值如下表2.3.1。根据AIC准则，最终确定的模型为. **表2.3.1：**模型系数

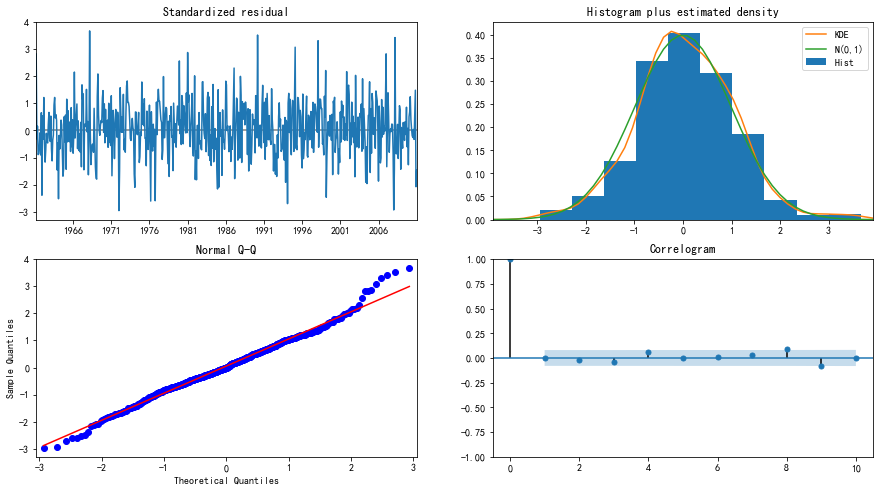
|  |  |  |  |
| --- | --- | --- | --- |
| 模型参数(p,q) | AIC | 模型参数(p,q) | AIC |
| (1,1) | 290.60 | **(2,1)** | 268.65 |
| (1,2) | 300.59 | **(2,2)** | 270.65 |
| (1,3) | 271.71 | **(2,3)** | 269.08 |
| (1,4) | 270.81 | **(2,4)** | 272.87 |

* 1. **模型检验及预测  
      表2.4.1：**参数显著性检验

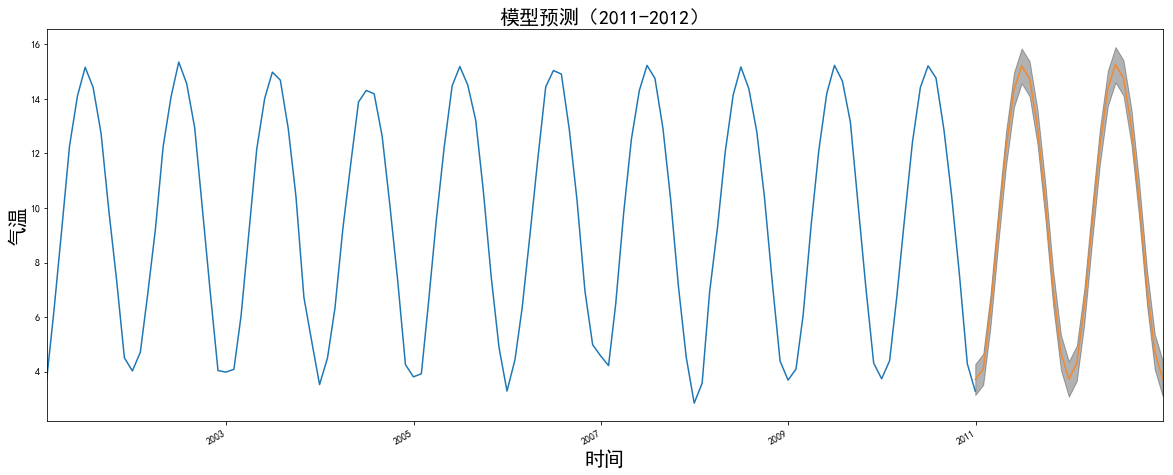
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 待估计参数 |  |  |  |  |
| 系数 | 0.9778 | 0.9887 | 0.2877 | 0.2112 |
| t统计量 | 74.763 | 13.801 | 6.869 | 5.023 |
| P值 | <0.0001 | <0.0001 | <0.0001 | <0.0001 |

结果表明使用模型为拟合数据得到的参数显著，确定模型的口径为：  
 模型是否合适，需要对其拟合优度进行检验，典型方法是对观测值和模型拟合值的残差序列进行白噪声检验分析。如果残差序列不是白噪声序列，则说明还有信息包含在相关的残差序列中未被提取，模型其他参数不能完全代表建模对象的统计性质，即所建模型不是最终模型。  
 **表2.4.1：**残差白噪声检验

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 延迟期数 | 6 | 12 | 18 | 24 |
|  | 3.3198 | 15.0151 | 19.6835 | 25.8945 |
| Prob(>Q) | 0.7677 | 0.24061 | 0.35091 | 0.3585 |

  
**图2.4.1：**残差白噪声检验

由图2.4.1可以看到，残差序列的自相关图显示自相关系数显著为零，而且由表2.4.1知残差白噪声检验Q统计量的P值都远远小于0.05。因此，可认为残差序列为白噪声序列，模型提取的信息充分完整。  
 用此模型对世界月平均气温数据序列进行拟合并对2011年1月至12月的月平均气温进行预测，具体数据如附表1。预测结果如图2.4.2所示，预测值与实际值基本吻合，表明模型的选择是正确的，平均绝对误差为0.08，拟合效果较好。



**图2.4.2：**模型预测

1. **结论**

世界月平均气温序列是一非纯随机性时间序列，相关性分析显示其具有明显的以年为周期的周期性。本文对世界月平均气温的动态数据进行了建模与预报，拟合效果较好。模型的口径为：  
这进一步说明世界月平均气温序列存在着一个上升趋势。但是影响气温变化的因素有很多，如降水的变化、生态环境的因素（植被覆盖）等。可以尝试将影响气温的诸多因素作为输入变量，温度作为输出变量进行多变量的时序分析，以便拟合更优的模型。

**参考文献：**

[1] 王燕． 应用时间序列分析[M]. 北京：中国人民大学出版社，2016．1

[2] Climate Change: Earth Surface Temperature Data

<https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data>

[3] 刘 佳， 赵慧文， 刘光荣. 基于SAS的非平稳时间序列分析及实证研究[J]. 汕头大学学报. 2010.02

[4] 方启东, 温鑫, 蒋佳静, 等. 基于时间序列分析的股价预测[J]. 宿州学院学报. 2010.08

[5] Python statsmodels工具包 <http://www.statsmodels.org/stable/index.html>

**附录：**

**附表1：**世界月平均气候预测值

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 预测值 | 1月 | 2月 | 3月 | 4月 | 5月 | 6月 | 7月 | 8月 | 9月 | 10月 | 11月 | 12月 |
| 2011年 | 4.08 | 6.21 | 9.34 | 12.18 | 14.33 | 15.21 | 14.74 | 12.95 | 10.28 | 7.04 | 4.71 | 3.73 |
| 2012年 | 4.31 | 6.37 | 9.44 | 12.25 | 14.39 | 15.25 | 14.77 | 12.98 | 10.31 | 7.06 | 4.73 | 3.75 |