# 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。

# 一、 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。

## 1.1 ARIMA(p, d, q)模型[1]

求和自回归移动平均模型（auto regressive integrated moving average，简记为$ARIMA$）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ARIMA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$ARIMA(p,d,q)$模型。它可以通过适当的$d$阶差分运算使得序列平稳。设$X=(x_1,x_2,\dots x_t,\dots x_n)$为一个等时间间隔为t的时间序列，ARIMA(p, d, q)模型具有如下结构：

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s)=0, s\neq t \\ E(x_t x_s)=0, \forall s<t \end{cases}$$

式中，$\nabla^d=(1-B)^d$; $\Phi(B)=1-\phi_1 B-\dots-\phi_p B^p$, 是平稳可逆$ARMA(p,q)$模型的自回归系数多项式；$\Theta(B)=1-\theta_1 B-\dots-\theta_q B^q$，是平稳可逆$ARMA(p,q)$模型的移动平滑系数多项式。$x_t$ 为非平稳序列，d 为差分运算的阶数，$\varepsilon_t$为零均值白噪声序列。

## 1.2 ARIMA(p, d, q) × $(P,D,Q)_S$模型[1]

ARIMA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$\mathrm{ARMA(p,q)}$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$\mathrm{ARMA(P,Q)}$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$\mathrm{ARMA(p,q)}$和$\mathrm{ARMA(P,Q)}$的乘积，也即乘积季节模型（$ARIMA(p,d,q)\times(P,D,Q)_S$）的完整结构如下：

$$\nabla^d\nabla_S^D x_t=\frac{\Theta(B)\theta_S(B)}{\Phi(B)\Phi_S(B)}\varepsilon_t$$

式中

$$\Theta(B)=1-\theta_1 B-\cdots-\theta_q B^q$$
$$\Phi(B)=1-\phi_1 B-\cdots-\phi_p B^p$$
$$\Theta_S(B)=1-\theta_1 B^S-\cdots-\theta_Q B^{QS}$$
$$\Phi_S(B)=1-\phi_1 B^S-\cdots-\phi_P B^{PS}$$

### 1.3 建模步骤

ARIMA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ARIMA模型时要求预处理后的时间序列是平稳的，所以在建立ARIMA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。

### 1.4 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：

$$\mathrm{AIC}=-2\ln(\text{模型的极大似然函数值})+2(\text{模型中未知参数个数})$$

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。

### 1.5 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。

## 二、 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。

### 2.1数据特征描述

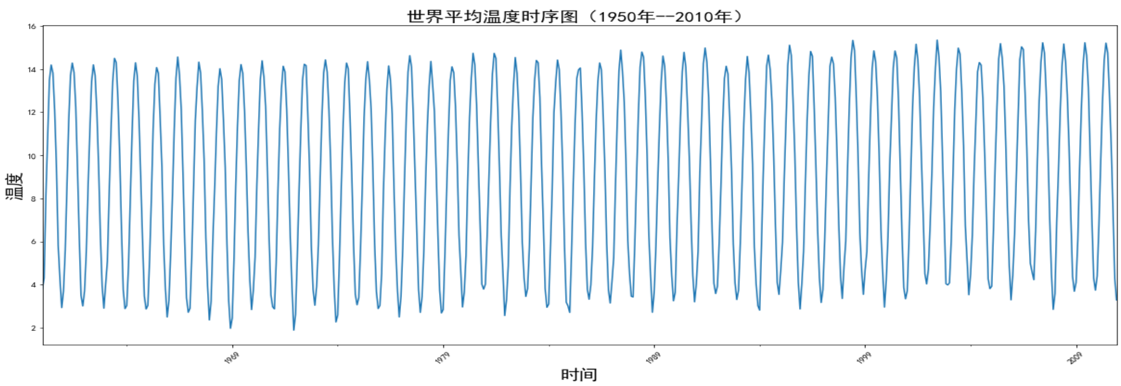


**图 2.1.1：**世界平均气温时序图

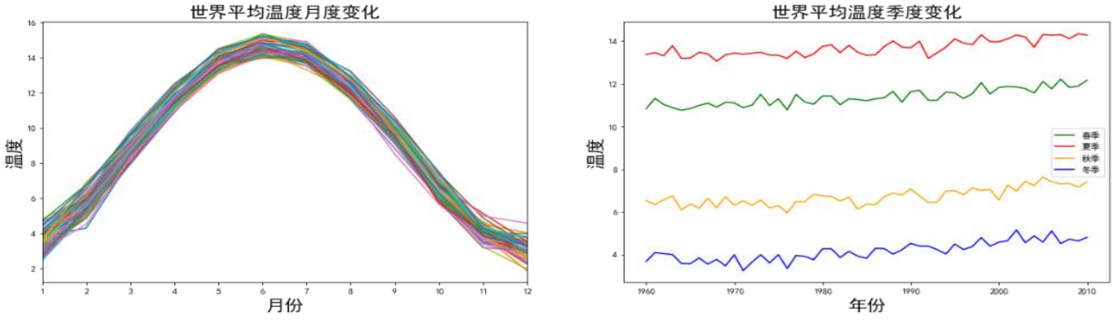


**图 2.1.2：**世界平均气温的季节性

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。

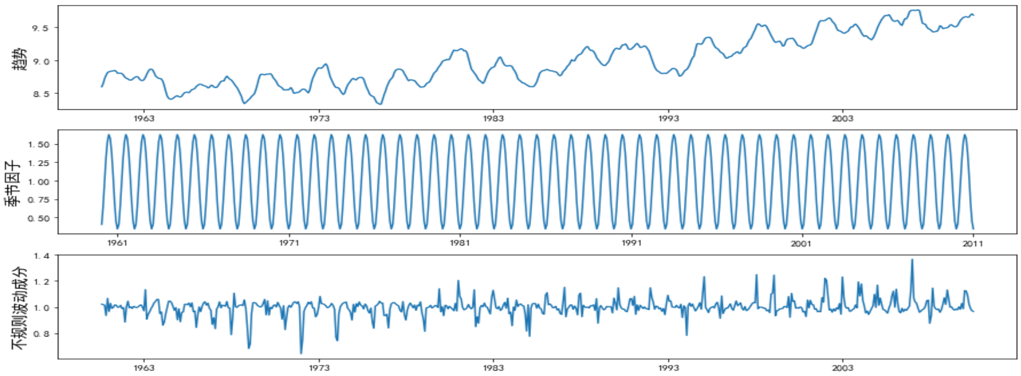


**图 2.1.3**：时间序列各成分的分解

## 2.2 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。

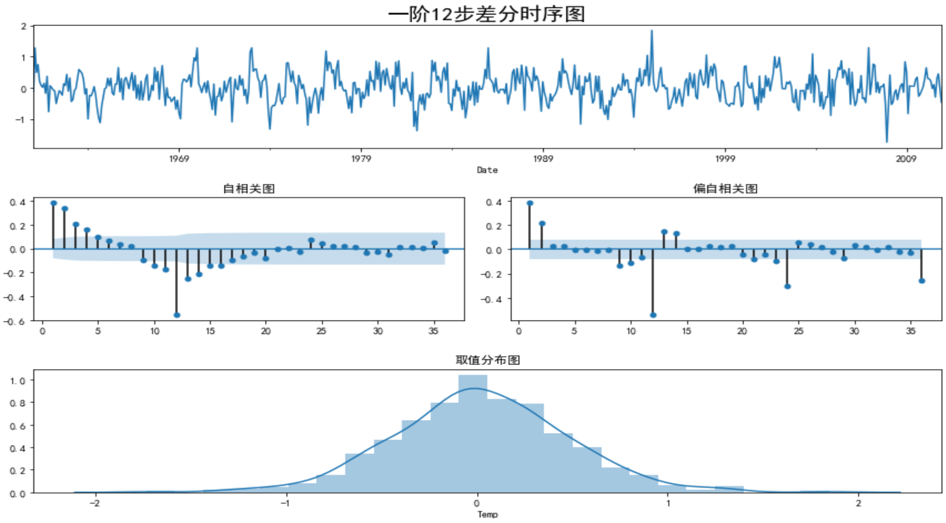


**图 2.2.1**：1 阶 12 步差分序列结果

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 | |

## 2.3 **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 $ARMA(p,q)$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$p \le 2, q \le 3$.

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$ARMA(0,1)_{12}$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$ARIMA(p,1,q)\times(0,1,1)_{12}$. 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$ARIMA(2,1,1)\times(0,1,1)_{12}$.

**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 |

## 2.4 **模型检验及预测**

**表 2.4.1**：参数显著性检验

| **待估计参数** | $\boldsymbol{\theta_1}$ | $\boldsymbol{\theta_{12}}$ | $\boldsymbol{\phi_1}$ | $\boldsymbol{\phi_2}$ |
|---|---|---|---|---|
| **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$ARMA(2,1,1)\times(0,1,1)_{12}$拟合数据得到的参数显著，确定模型的口径为：

$$\nabla\nabla_{12}x_t=\frac{1-0.9778B}{1+0.2877B+0.2112B^2}(1-0.9887B^{12})\varepsilon_t$$

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 |
|---|---|---|---|---|
| $Q_{LB}$ | 3.3198 | 15.0151 | 19.6835 | 25.8945 |
| Prob(>Q) | 0.7677 | 0.24061 | 0.35091 | 0.3585 |

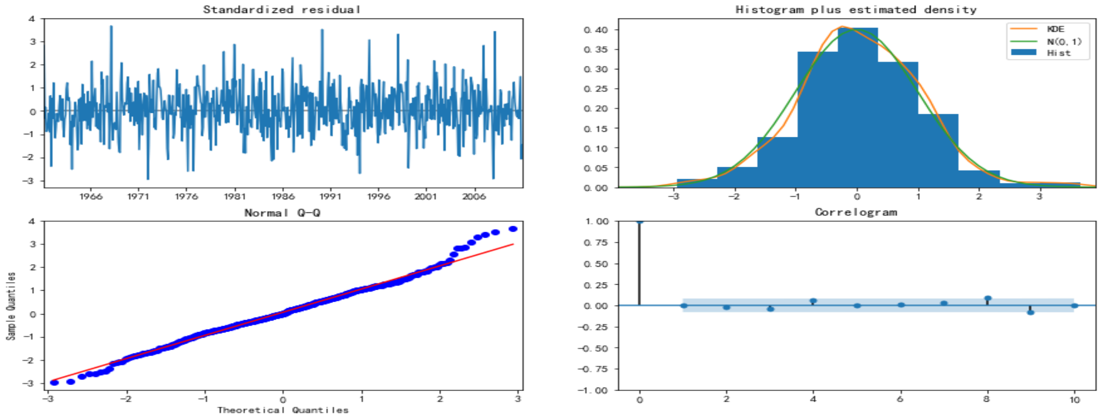


**图 2.4.1**：残差白噪声检验

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。

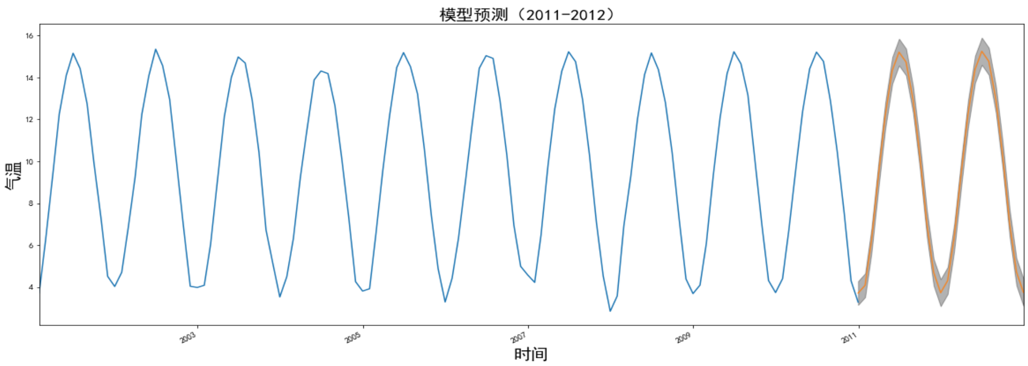


**图 2.4.2**：模型预测

## 三、 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：

$$\nabla\nabla_{12}x_t=\frac{1-0.9778B}{1+0.2877B+0.2112B^2}(1-0.9887B^{12})\varepsilon_t$$

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 |