对于周边建筑的沉降预测的研究

灰色模型最合适

arima模型需要数据量比灰色模型更多，同时模型需要经常根据数据的变化进行更新，预测的精度较差

BP神经网络在数据量较小的情况下，表现不佳，同时对于前几周的数据预测比较符合，但是长期预测精度较差，也需要进行滚动预测，进行更新模型，BP神经网络模型的训练时间较长，不适合数据量较小的工程系统中的应用。但是从目前的结果可以看出，周边建筑物的沉降和时间具有很大的关联性，所以当训练的数据量达到一定的量级，BP神经网络预测无论在精度还是在速度上都会有很好的应用空间。

# 第五章 周边建筑物沉降预测典型智能算法的对比研究和实现

预测是指在现有信息的基础上，依照一定的规律和方法对未来系统的发展趋势进行推演，从而预先了解事情的发展过程以及结果。常见的预测方法主要包括时间序列模型预测，灰色系统预测模型，马尔科夫预测模型，支持向量机预测以及BP神经网络预测等。由于时间问题，本章选取时间序列模型，灰色预测模型，BP神经网络预测模型对周边建筑物沉降预测研究并用Python进行算法实现。

## 5.1 时间序列预测模型的研究与实现

### 5.1.1 时间序列预测模型简介

一般情况下，时间序列可以通过统计分析的方法，通过统计学中各种方法对时间序列的统计特性进行研究，从而发现其中的内在规律。在深基坑工程中，周边建筑物沉降的时间序列具有一定的内在规律，可以通过时间序列的分析方法进行研究。常用的时间序列模型有四种：自回归模型 AR(p)、移动平均模型 MA(q)、自回归移动平均模型 ARMA(p,q)、自回归差分移动平均模型 ARIMA(p,d,q)。前三种可以说是第四种模型的特殊形式。因此我们将以ARIMA(p,d,q)模型为例，对其进行研究，并应用到实际的周边建筑物沉降预测。

实际问题中，得到的时间序列大多为非平稳序列，可以对序列进行差分运算，继而生成平稳序列，对于这种差分平稳序列，进行ARIMA模型拟合，利用拟合的ARIMA模型进行未来情况的预测。

ARIMA(p,d,q)中，AR是"自回归"，p为自回归项数；MA为"滑动平均"，q为滑动平均项数，d为使之成为平稳序列所做的差分次数（阶数）。ARIMA(p,d,q)模型可以表示为：

其中L是滞后算子，。

### 5.1.2 建模流程



如上图所示，ARIMA模型的建模流程是首先进行平稳性检验，如若不通过，则进行差分运算，确保分析的时间序列是平稳序列，再进行白噪声检验，确保预测是有意义的；然后计算样本相关系数，进行模型识别和参数估计，确定回归项数p和滑动平均项数q，得到模型的参数，进行模型检验，如果合格进行模型的优化和序列的预测。

下面分别介绍这几个关键步骤

（1）平稳性检验

时间序列分析的前提条件是假定数据样本是平稳的，即序列的期望、方差、自协方差函数等都是不随时间而变化的，并且可以用时间平均代替总体的平均。因此对于样本数据的平稳性检验是建模的前提。平稳性检验的方法有以下几种：

①时序图判断法

时序图判断法是最简单也是最为直观的方法，根据平稳序列的均值、方差为常数的性质，来判断给定的时间序列是否为平稳序列。所谓时序图，就是在平面直角坐标系中，将所有的时间序列的值按照时间顺序绘制成折线图，观察期是否存在明显的趋势性，周期性以及波幅变化。如果时序图表现出：始终在固定的常数值上下波动、波动幅度随着时间变化幅度不增大、没有明显的周期性和趋势性，那我们认为该序列为平稳序列。

②样本自相关系数检验法

自相关系数检验法也是一种主观性较强的方法，基于平稳序列的短期相关性作为评判依据。根据平稳序列的短记忆性可以知道，自相关系数有两种表现形式：截尾和拖尾，表现为自相关系数随着延迟期数k的增加会很快降为在0的周围小幅度波动，或者迅速衰减为在0的周围小幅度波动。通过观察自相关系数图来判断序列的平稳性。

③分段检验法

分段检验法是利用参数检验，根据平稳序列的定义可以知道，均值和方差都与时间无关，并且自协方差函数只和时间间隔有关，所以对检验序列进行分段，通过判断各子序列的样本均值、方差、自协方差函数是否存在显著的差别，来检验样本序列是否为平稳序列。

④单位根检验法

单位根检验法是一个定量的检验方法，也就是给出了一些检验的统计量。后由Said和Dickey在1984年提出推广的检验方法，简称ADF，ADF检验简单的说就是判断时间序列是否存在单位根，如果存在说明不平稳，反之则为平稳序列。对应的模型为

其中，是确定性部分，比如常数项以及趋势项，差分是用于模型的误差项的建模，参数p的确定原则是使得误差项互不相关。平稳性检验的原假设和择备假设分别为

带有确定项的平稳序列也就是该序列去掉确定项之后变为平稳序列。检验的统计量为

其中为模型的最小二乘估计值，是 的标准差，对模型做简单的变化可以得到如下公式

其中。当原假设成立时，是平稳序列，。此时ADF检验变为对系数是否为0的普通t检验。

在实际的应用中，得到的显著性统计量，给定三个置信区间（10%，5%，1%），若小于这三个置信区间，则对用有（90%，95%，99%）的把握拒绝原假设，也就是同意备择假设，证明是平稳序列。

（2）平稳化方法

虽然平稳性序列有很多良好的统计特性，也是传统时间序列分析的基础，但是在实际应用中，所得到的时间序列通常都是非平稳序列，因此最为自然也是最为简单的方法就是通过随机序列差分法将非平稳序列转换为平稳序列。

定义相邻两个序列值之间的差值为一阶差分，记为

则k阶差分的定义为对k – 1阶差分序列再进行一次差分操作，即

根据Cramer分解定理的基础，方差齐次非平稳序列都可以分解成一个平稳序列和一个多项式决定的确定序列的和

其中是一个平稳序列，对序列进行d阶差分，相当于连续变量的d阶求导后

其中C为一个常量；为经过d阶差分后的平稳序列，可以利用传统的平稳序列分析方法进行处理了。又因为

因此，可以表示为

由此可见，d阶差分可以让非平稳序列变成平稳序列的实质就是利用过去d期的历史值作为自变量来解释当前序列值的变动，也就是利用自回归的方式提取确定性的信息。但是在实际的应用中，我们并不知道阶数d，因此每经过一次差分，都需要进行平稳性检验，一旦平稳就需要停止，防止过差分造成的序列信息的损失。

（3）白噪声检验

本文中时间序列分析的目的就是预测，也就是利用序列之间的相互关系，通过历史值对未来发展趋势做出估计。经过平稳性检测和差分后得到平稳序列，虽然序列平稳被认为是时间序列分析的前提，但是不是所有的平稳序列都具有这种分析的意义，例如白噪声序列就不存在这种预测意义，因为白噪声序列前后不存在相关性，历史值对未来趋势没有任何意义。

白噪声检验也被称为纯随机性检验，也就是检查一个时间序列的自相关系数是否满足

上式只是一个理论公式，在实际应用中，由于观测的序列的长度是有限的，导致了序列样本的自相关系数不可能完全满足上式的要求。白噪声检测方法很多，目前以下两种检测方法最为普遍：

①Q统计量

Barlett定理指出，如果是观测样本是白噪声序列，那么当观测量N足够大时，观测样本的自相关系数独立，并且满足

基于Barlett定理，Box和Pierce构造出了Q统计量：

其中为自由度为m的卡方分布，N为观测期数，m为指定的延迟期数。

原假设

责备假设

在原假设的条件下，给定检测水平α，查找自由度m的分布表，得到临界值，即

当，或者时，有理由拒绝原假设，也就说该序列不是白噪声序列，否则接受原假设，该序列为白噪声序列

需要注意的是，Q统计量的构造基础是观测序列样本量N足够大，一般在实际应用中，Q统计量用于观测值较多的序列，一般时观测效果比较好

②LB统计量

鉴于Q统计量的缺陷，Ljung和Box又推导出适用于观测数据较少的LB统计量：

检验过程类似Q统计量检验法的过程，比较两个公式可以发现，当N较大时，两者近似相等。

（4）模型识别与定阶

通过平稳性检验和白噪声检验，经过d阶差分后得到平稳非白噪声序列，可以利用AR、MA、ARMA等三类模型对平稳序列进行拟合，根据序列的自相关系数和偏自相关系数的性质可以知道，分为三种情况：自相关系数拖尾并且偏自相关系数截尾，则序列用AR（p）模型建模；若自相关系数截尾并且偏自相关系数截尾，则序列用MA（q）模型建模；若自相关系数和偏自相关系数都拖尾，则序列用ARMA（p，q）模型建模。

确定模型后，需要对模型进行阶数p，q的定阶。常见的做法是，事先选择多组（m，n）值，并通过之后介绍的参数估计的方法建立对应的ARMA（m，n）模型，同时经过模型检验后得到机组通过检验的参数，再利用AIC、BIC或者FPE准则选出一组最好的阶数的值。下面以AIC定阶为例，简要介绍一下定阶过程：

①考虑样本长度为N的序列，分别确定（p，q）的上界和，记作

②对于所有的，计算AIC函数：

其中是p = m, q = n时模型误差方差的估计值

③使得AIC最小的那组(m,n)的值则是阶数(p,q)的AIC定阶

BIC准则和FPE准则和AIC定阶过程类似，不同的是定阶函数的不同，BIC函数为：

（5）模型的参数估计

确定了差分系数d，得到了最佳的(p,q)值，则ARMA模型为

剩下来的就是估计ARMA模型中的参数，目前主要由矩估计，最小二乘估计以及最大似然估计等方法。其中最小二乘估计使用最为广泛，下面简要介绍一下最小二乘估计的方法。

最小二乘估计的核心思想是对于给定的(p,q)的值，如何能够确定ARMA模型中p+q个参数，使得残差的平方和达到最小。假设,给定样本，则ARMA(p,q)模型的残差平方和

使得残差平方和最小的参数即为最小二乘估计β，但是式中也是未知的，所以无法直接估计得到参数β。

目前有两种处理方法，一种是条件最小二乘估计法，假设过去没有观测到的数据序列全为0，则

系数是逆函数，可以通过迭代得出，具体算法不作详细解释，此时残差平方和转换为

通过迭代的方法，求解得到参数β的估计值，也就是参数β的最小二乘估计值。

但是该方法需要迭代计算，迭代计算的过程比较复杂，还有一种方法是自回归逼近的方法，该方法的核心思想是观测数据可以用AR(m)模型拟合，因此残差平方和中的估计值可以先通过AR(m)模型拟合数据得到，再估计参数β。

（6）模型的检验

模型检验包括有效性检验和参数显著性检验。

统计模型只是对观测数据的真实的过程的一种近似，在模型拟合之后，还需要对模型的有效性进行检验，检验拟合的模型是否充分提取了观测序列中的信息，因此模型检验包括残差的白噪声检验，白噪声检验的过程同步骤（3），不同的是对残差进行检验。

不仅如此，模型检验还包括参数的显著性检验，检验模型中的参数是否都显著为零，检验后可以删除那些不显著的参数，来简化模型。

（7）预测

经过上述步骤后，模型参数都已经知道了，只有在无限观测值的情况下，才能做到精确地预测，预测公式如下

实际应用中，往往只能得到有限的样本量，根据有限的样本，可以做k步最佳预测的近似。已知模型的参数，因此Green系数可以推导得出。然而上式中随机误差的估计值仍然无法得出。常见的近似的方法就是假设开始观测之前的误差都是零，即

从而根据预测公式，取k = 1以及公式

可以迭代算出随机误差，从而可以得到第k步的预测值。

（8）差分还原

得出模型的预测值，经过以上分析可知，这个预测值是差分运算后的序列的预测值，要得到原序列的预测值，需要进行差分逆运算。

### 5.1.3 Arima算法步骤的改进优化

按照普通的ARIMA模型建模分析的方法，无法适应本系统中深层水平位移的实际预测，因此做出以下两点改进优化。

（1）数据预处理

在深基坑工程监测的过程中，得到的深层水平位移的序列值往往不是等时间间隔的，考虑这种实际情况，一般常用的是牛顿分段线性内插法进行处理。但是实际上，常规情况下，内插点距离上次测量的时间和下次测量的时间是不同的，时间不同，前后两次的沉降量对于本次沉降量的影响也必然是不同的，所以本文采用2次距离权内插法进行改进。







根据上述公式进行处理，生成新的等间隔时间序列。

（2）滚动建模预测

易知，预测的步长越长，未知的信息也就越多，从而根据模型预测的估计值的精度也就越差，如果按照原有的模型，随着时间的发展，在原有的观测值的基础上会增添越来越多的新的观测值，新的观测值所带来的的新的信息也就没有起到作用。因此为了能够更好的利用新的观测值，提高预测的精度，并适应实际的系统需求，本文采用滚动修正预测的方法进行预测。

具体方法为模型的步长固定，假定为d，则有序列步长可以根据以往实际的观测情况进行事先研究确定，预测d + 1的值，每次有新的观测值，则按照序列进行建模，预测d+2的值，按照此方法进行固定步长的滚动预测。

### 5.1.4 Arima算法编程实现

原有建模方法停留在研究阶段，无法应用于本文的实际预测系统中，因此从实用性的角度出发，增加自动搜索最佳差分次数，自动定阶。综合上述的基本流程以及改进优化，按照下图的流程图编写ARIMA算法预测模块。

c

预测模块接口方法如下：

def arima\_model(time\_series,beginIndex, d):

其中time\_series为数组类型，记录了建筑物沉降时间序列经过等间距处理后的序列值；beginIndex为建模所需要的历史数据的开始序号；d为模型建立所需要观测值的个数；接口方法的返回值为下一个时间点的预测值。

大致图建模流程进行编程，实现自动预测的功能。

（1）平稳性判别以及非平稳序列的平稳化

本系统采用ADF单位根检验的方法，由上述方法分析可知，判断给定的序列是否存在单位根，如果存在说明不稳定，反之则为稳定序列。而判断是否存在单位根，则根据t检验，得到显著统计量，判断是否小于某个指定的置信区间，本系统选用5%的置信区间。核心代码如下：

from statsmodels.graphics.tsaplots import acf,pacf,plot\_acf,plot\_pacf

from statsmodels.tsa.arima\_model import ARMA

def test\_stationary(time\_series):

df\_test = sm.tsa.stattools.adfuller(time\_ series, )

return df\_test < 0.05

对于非平稳的序列，本系统采用差分的方法进行平稳化，核心代码如下

def difference (time\_series):

time\_ series = time\_series(1)

time\_series= time\_series.dropna(how=any)

time\_series.plot(figsize=(8,6))

plt.show()

同时本系统提供自动搜索最佳差分次数，能够通过最少差分次数，使得非平稳时间序列转换为平稳时间序列，返回最少差分次数，核心代码如下

def stabilization(time\_series):

d = 0

while(test\_stationary(time\_ series)):

difference(time\_ series)

d++

return d

（2）白噪声检测

由于观测值较少，通过以上方法的分析，本系统采用LB的统计量进行序列的白噪声检测，对于所有的LB统计量的值，如果大于显著水平，则认为是白噪声，否则该序列不是白噪声，核心代码如下：

import statsmodels.api as sm

def test\_white\_noise(time\_ series):

r,rac,Q = sm.tsa.acf(time\_series\_diff, qstat=True)

prac = pacf(time\_series\_diff,method='ywmle')

table\_data = np.c\_[range(1,len(r)), r[1:],rac,prac[1:len(rac)+1],Q]

table = pd.DataFrame(table\_data, columns=['lag', "AC","Q", "PAC", "Prob(>Q)"])

print(table)

for i in range(len(table)):

if(table[i+1][5] > 0.05)

return true

return false

（3）模型识别与定阶

按照时间序列模型的识别规则，根据自相关系数以及偏自相关系数的拖尾与截尾性，确定模型。确定模型后，本系统借助BIC准则进行反复测试，确定最优的阶数p和q，核心代码如下：

def get\_p\_q(time\_series)

order\_p,order\_q,bic=[],[],[]

model\_order=pd.DataFrame()

for p in range(4):

for q in range(4):

arma\_model=sm.tsa.ARIMA(time\_series.dropna(),(p,q)).fit()

order\_p.append(p)

order\_q.append(q)

bic.append(arma\_model.bic)

print('The BIC of ARMA(%s,%s) is %s'%(p,q,arma\_model.bic))

model\_order['p']=order\_p

model\_order['q']=order\_q

model\_order['BIC']=bic

P=list(model\_order['p'][model\_order['BIC']==model\_order['BIC'].min()])

Q=list(model\_order['q'][model\_order['BIC']==model\_order['BIC'].min()])

P\_Q = []

P\_Q.append(P[0])

P\_Q.append(Q[0])

return P\_Q

（4）模型检验

模型检验包括有效性检验和参数显著性检验。由上述步骤的方法分析可知，其中有效性检验为检验残差的白噪声检验，同样采用白噪声检验的代码进行检验，不同的是，传入的参数不是原序列，而是残差序列。对于参数显著性检验，核心代码如下：

arma\_mod = ARMA(time\_series,(p,d,q)).fit(disp=-1,method='mle')

summary = (arma\_mod.summary2(alpha=.05, float\_format="%.8f"))

print(summary)

（5）预测和差分还原

通过了模型检验后，利用该模型进行预测下一个序列的值，并进行差分还原，得到真正的预测值，核心代码如下：

arma\_model = sm.tsa.ARMA(time\_series,(p,d,q)).fit(disp=-1,maxiter=100)

predict\_data = arma\_model.predict(start=endIndex - beginIndex - 1, end=endIndex - beginIndex + num, dynamic = False)

real\_ predict\_data = data[-1] + predict\_data

## 5.2 灰色预测模型的研究与实现

### 5.2.1 灰色预测模型简介

白色系统是结构内部的信息是完全透明的，黑色系统是指结构内部的系统是完全未知的，介于这两者之间的系统为灰色系统。由灰色系统理论可知，灰色系统能够更加准确地从参数、结构、信息的内部去发掘一个系统的本质规律。灰色预测模型是灰色系统理论的一个最为重要的内容之一。GM(1,1)是常用的灰色预测模型，灰色预测模型主要是针对现实中大量的存在灰色不确定性的预测问题，用少量的有效数据和灰色不确定性的数据，通过序列的累加，揭示事务的未来的发展趋势，从而达到提前防范的作用。

### 5.2.2 建模步骤



由上图可知，在建立灰色模型之前，需要对原始数据进行处理，通过累加生成数据序列，经过灰色作用处理，生成灰数，另外如果原序列是非等时间间隔的序列，通过5.1.3节中介绍的方法，进行变化，生成等时间间隔的数据，以满足灰色模型的数据需求。经过变量参数的求解，进行求解得到模型后，进行验证，如果通过验证则进行序列的预测，否则重新构造及优化模型。下面简单介绍几个重要的步骤：

（1）原始数据序列处理

假设某个监测点有n个原始观测数据的序列：



其中；

经过一次累加操作（AGO-1）后得到新的序列：



其中



称为GM(1,1)模型的原始形式。

经过邻均值等权操作后得到新的序列，也称之为背景值：



称为GM(1,1)模型的基本形式

（3）白化方程构建和变量参数求解

GM（1,1）模型的白化方程，也称之影子方程，为：



（4）模型选择与验证

（5）序列预测

### 5.2.3 灰色预测算法的改进优化

（1）改进的新陈代谢GM（1,1）模型

建模的目的是预测，随着事务的发展，一般来说，老的数据信息的意义将越来越低，因此，为了提高后期的模拟预测精度，需要陆续删除老信息，补充新信息，特别是当量变积累到一定程度，发生质变的时候，去掉无法反应系统目前状况的老数据就显得尤为重要了。

因此本文采用新陈代谢GM（1,1）模型对周边建筑物的沉降预测，但是样本数量的选择盲目性较大，目前也没有足够的理论可以参考。本文采用自适应增长的方式进行样本数量的选择。灰色预测模型对于原始数据并没有太大的样本量的要求，只要观测的数据量大于4个，即可建立模型。具体如下：

假设有观测序列，样本量为k，k初始值为，同时最小为，的选择可以通过事先研究，进行提前的配置，利用灰色预测的建模步骤对下一个序列值进行预测，当观测到实际值时，进行自适应增长，方法如下：给定一个误差可接受度，如果则表示目前的样本数量可接受，下一次预测的样本数量加1；反之，如果则表示目前的样本量不可接受，下一次的样本数量减1，变为k-1，但是不可无限减小，最小样本数量限定为4。其中，误差可接受度的选择根据实际设计时对于沉降的监测变化量的预警值确定。

（2）误差修正（修正系数，）

### 5.2.4 灰色预测算法编程实现

## 5.3 BP神经网络预测模型的研究与实现

5.3.1 BP神经网络预测算法的基本原理

5.3.2 建模步骤

5.3.3 BP神经网络预测算法的改进优化

奇偶交替

5.3.4 BP神经网络算法编程实现

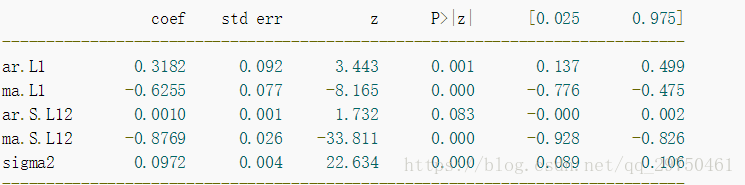
## 5.4 实验对比分析

为验证几种预测算法在周边建筑物沉降上预测的可靠性，选取某地铁站Z70监测点观测到的建筑物沉降原始数据进行建模分析，观测数据以7天作为一个周期

5.4.1 ARIMA算法实验结果分析

以某地铁站Z70监测点观测到的建筑物沉降数据为例，说明ARIMA预测模块的编程实现，根据多次实验发现序列的步长d定为20最为合适，经过数据预处理，生成等间隔的时间序列。以周为单位，根据第42周到61周的数据建立ARIMA模型，预测第62周的沉降量。

结果：



   由SARIMAX的输出产生的SARIMAX返回大量的信息，但是我们将注意力集中在系数表上。 coef列显示每个特征的重量（即重要性）以及每个特征如何影响时间序列。 P>|z| 列通知我们每个特征重量的意义。 这里，每个重量的p值都低于或接近0.05 ，所以在我们的模型中保留所有权重是合理的。

5.4.2 灰色预测算法实验结果分析

5.4.3 BP神经网络实验结果分析

5.4.4 算法对比分析

选取某地铁施工周边建筑物沉降的部分数据进行模型的实验对比分析