孙路强,刘磊,卞真付,等.利用时间序列算法实现测震数据存储服务器磁盘容量预测[J].地震工程学报,2018,40(增刊):170-173.doi:10.3969/j.issn.1000－0844.2018.Supp.170
SUN Luqiang,LIU Lei,BIAN Zhenfu,et al.Disk Capacity Prediction for the Seismic Data Storage Server Using the Time Series Algorithm[J].China Earthquake Engineering Journal,2018,40(Supp):170-173.doi:10.3969/j.issn.1000－0844.2018.Supp.170

利用时间序列算法实现测震数据存储服务器磁盘容量预测

孙路强，刘　磊，卞真付，许　贺
（天津市地震局，天津 300201）

摘要：天津市地震局通过“九五”、“十五”数字地震观测网络的建设，实现了测震台站数据网络化传输，将波形数据实时存储于 AWS 数据存储服务器中，用户通过人机交互系统从存储数据中调取波形数据进行分析并提交分析结果。测震系统在日常运行时会对存储服务器端软硬件造成一定程度的负荷，影响系统的整体应用效能。针对这一问题，利用 Python 时间序列算法实现对 AWS 数据存储磁盘容量进行预测，判断存储服务器的负载率情况，避免应用系统由于存储容量耗尽造成的系统瘫痪。根据该算法同时可监控台站实时数据量，生成台站数据流量变化曲线，从而成为一种评估台基噪声情况的新方法。

关键词：Python；时间序列；负载；测震；预测
中图分类号：P315　　**文献标志码**：A　　**文章编号**：1000－0844(2018)增刊－0170－05
DOI:10.3969/j.issn.1000－0844.2018.Supp.170

Disk Capacity Prediction for the Seismic Data Storage Server Using the Time Series Algorithm

SUN Luqiang, LIU Lei, BIAN Zhenfu, XU He
(*Earthquake Agency of Tianjin Municipality*, *Tianjin* 300201, *China*)

Abstract: The Earthquake Agency of Tianjin Municipality realized a seismic data transmission network by constructing a digital seismological observation network. The seismic waveform data is real-time stored in the AWS data server, and users can call the data from the AWS server by a man-machine interface system. Therefore, the seismic system operation generates load in the server hardware and software, thus affecting the system overall efficiency. In this study, we use a Python-based algorithm of time series to predict the disk capacity of the AWS server and estimate the load rate of the storage server, thus avoiding the occurrence of system paralysis because of storage capacity depletion. The real-time data of stations can also be monitored by the algorithm, and variation curves of station data can be accordingly generated.

Keywords: Python; time series; load; seismic; prediction

收稿日期:2017-05-18
基金项目:中国地震局监测、预测、科研“三结合”课题(160201)
第一作者简介:孙路强(1985－),男,天津静海人,工程师,主要从事地震速报、编目、系统运行和软件开发工作。
E-mail:sunluqiang163@163.com。

0　引言

随着计算机相关软件的开发，数学知识不再是空谈理论。时间序列分析主要建立在数理统计基础之上，是一种动态数据处理的统计方法，它承认事物发展的延续性，在考虑过去数据的情况下同时考虑事物发展的随机性，以预测未来事物的发展。

作为存储测震波形数据的数据存储服务器，其出现突发故障造成系统瘫痪的可能性很小，通常是一个逐渐变化的过程。由于系统长时间的使用，台站实时数据不断地写入存储服务器中，存储空间逐渐变小，最终因存储空间不足造成系统故障。因此在不考虑人为因素的影响时服务器存储空间变化与时间是相关联的，且历史存储数据对未来存储空间的发展存在影响，可利用时间序列分析方法预测服务器磁盘使用空间，并提醒测震台网人员提前采取措施，以保障测震系统正常运行。

1　基于 Python 的时间序列分析过程

1.1　容量预测建模流程

首先从 AWS 数据存储服务器磁盘中选择抽取历史数据和每天定时抽取数据，并对抽取数据进行周期性分析以及数据清洗、属性变换等操作，使抽样数据成为建模数据。采用时间序列分析法对抽取的数据进行模型构建，从而达到预测存储磁盘使用情况的目的。测震数据存储容量建模流程见图 1。

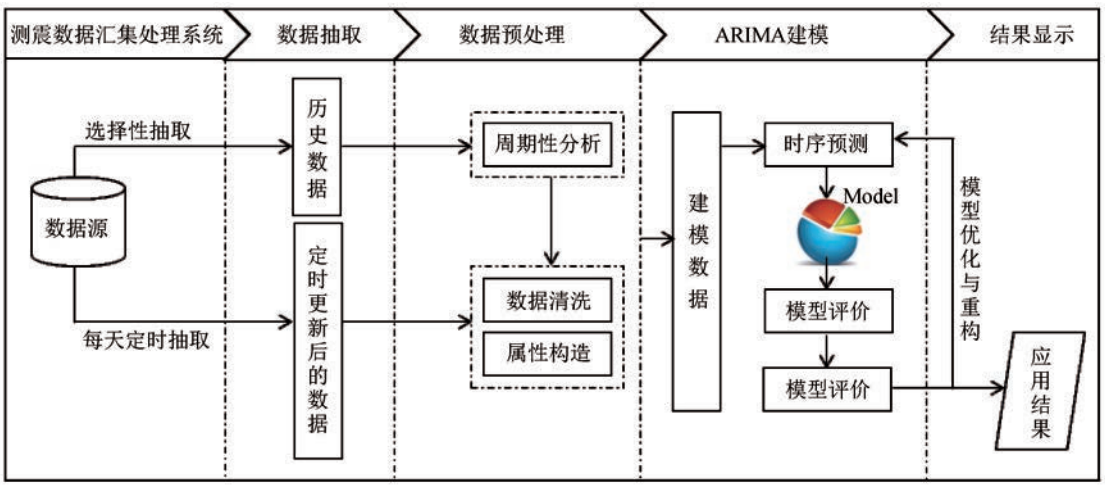


图 1　测震数据存储容量建模流程图

1.2　数据初级分析

为了能够建立合理的数据模型，应进行数据平稳性分析[1]。通过时序图研究测震台网连续波形数据变化规律，以服务器光纤盘已使用大小为标准，以天为单位，进行周期性分析。60 天内其时序如图 2 所示。

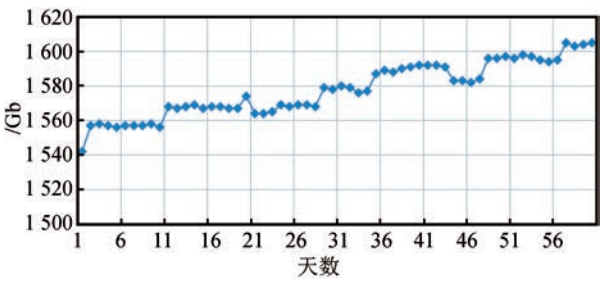


图 2　/root/emc_data 磁盘使用空间时序图

通过磁盘空间使用时序图可以发现磁盘的消耗情况是不具备周期性的，但其整体呈增长的态势，可认定数据变化是非平稳的。

1.3　数据预处理

测震系统在运行过程中，台站实时数据存储在规定的目录下，监控系统每天定时统计磁盘容量信息，磁盘空间在不扩容的情况下固定不变，磁盘原始数据中会存在重复数据，因此需要采用数据清洗的方法剔除磁盘中的重复数据。数据清洗的目的是删除重复信息、纠正存在的错误，并提供数据一致性。数据清洗是一个反复的过程，清洗过程中应避免将有用数据过滤掉，应经确认后再进行过滤规则的验证[2-3]。磁盘经过数据清洗后，其相关属性信息以数据记录的形式存在数据中，存储格式列于表 1。服务器磁盘信息可通过表中的 NAME、DISGET、ENTITY 三个属性进行分区，并将三属性进行合并，构造成新属性值。属性变化后性能表列于表 2。合并属性的 Python 代码如下：

表 1 原始性能表

SYSTEM	NAME	DISGET	DESCRIPTION	ENTITY	VALUE	COLLECTTIME
测震系统	CZTW_AWS	164	磁盘已使用大小	root\emc_data\waveform	1542	2016/11/15

表 2 属性变换后的性能表

SYSTEM	CZTW_AWS:164:root\emc_data\waveform	COLLECTTIME
测震系统	1542	2016/11/15

```
#属性变换
import pandas as pd
#参数初始化
discfile='../emc_data/waveform/discsize.xls'        #磁盘相关属性记录
transformeddata='../tmp/discsize_process.xls'       #预处理后磁盘信息记录
data=pd.read_excel(discfile)
data=data[data['DISGET']==164].copy()               #只保留 DISGET 为 164 的数据
data_group=data.groupby('COLLECTTIME')              #以时间分组
def attr_trans(x):                                  #属性变换函数
  result=pd.Series(index=['SYSTEM','CZTW_AWS:164:root\emc_data\waveform','COLLECTTIME'])
  result['SYSTEM']=x['SYSTEM'].iloc[0]
  result['COLLECTTIME']=X['COLLECTTIME'].ILOC[0]
  result['CZTW_AWS:164:root\emc_data\waveform']=x['VALUE'].iloc[0]
return result
data_processed=data_group.apply(attr_trans)         #分组处理
data_processed.to_excel(transformeddata,index=False)
```

2 模型构建

通过预处理后的数据可分为两部分,包括用于建模的样本数据和模型验证数据。在构建磁盘空间预测模型之前,首先对时间序列值平稳性进行检测,对不平稳的模型值序列进行差分处理直至数据平稳[4]。对于平稳数据再次进行白噪声检验,用于判断序列中的有用信息是否已被提取完毕,并只剩下随机干扰数据。通过白噪声检验后,确定测震数据存储磁盘空间预测模型为 ARIMA 模型,并利用模型识别方法进一步确认该模型的 p、q 参数[5]。磁盘空间预测建模见图 3。

2.1 平稳性检验

对预测模型进行平稳性检验的目的是确定原始数据中是否存在随机趋势或确定趋势,避免出现由于非平稳时间序列不存在协整关系导致的伪回归现象[6]。可利用单位根检验的方法或时序图的方法进行预测数据平稳性检验。基于 Python 语言的平稳性检验代码如下:

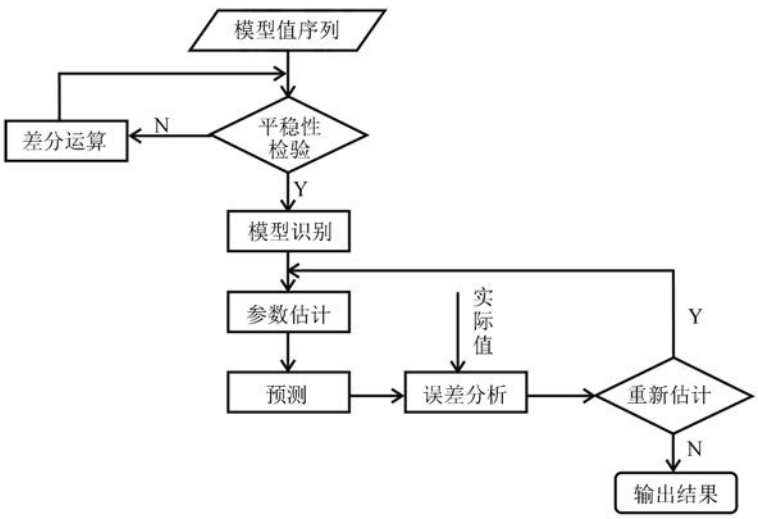


图 3 磁盘空间预测建模图

```
import pandas as pd
    discfile='../tmp/discsize_process.xls'
    data=pd.read_excel(discfile)
    data=data.iloc[:len(data)-5]        #不使用最后 5 个数据
    fromstatsmodels.tsa.stattools import adfuller as ADF        #利用 ADF 方法进行平稳性检测
    diff=0
    adf=ADF(data['CZTW_AWS:164:root\emc_data\waveform'])
    whileadf[1]>=0.05:    #p 值小于 0.05 认为平稳
    diff=diff+1
    adf=ADF(data['CZTW_AWS:164:root\emc_data\waveform'].diff(diff).dropna())
```

2.2　模型识别

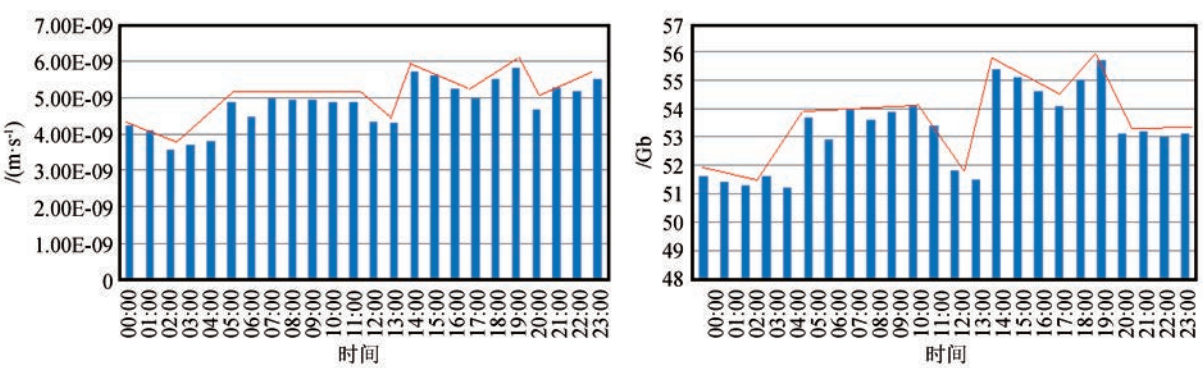
极大似然比方法是一种具有理论性的点估计法，也是一种反映真实性的指标。线性平稳时间序列模型的关键在于确定阶数，预测程序利用 BIC 信息准则对模型进行定阶，确定 ARIMA 时序模型的 p、q 参数。通过 BIC 信息准则确定与磁盘空间预测模型相关的准则函数，同时要考虑模型对原始模型数值的接近程度和待定参数个数。

3　模型应用

使用建立的 ARIMA 时序模型预测存储服务器磁盘空间时[7]，首先完成定时从 AWS 数据存储服务器上抽取数据，并对抽取数据进行预处理，将处理后的数据放到模型的初始数据中，作为该模型的输入数据；时序模型通过调用初始数据来预测存储服务器磁盘未来一周内的空间使用情况，从而判断未来磁盘使用率，当使用率达到预先设置的阈值时，就会以短信方式将报警信息发送给测震台网系统维护人员。同时可根据需要将预测得到的空间使用率对应不同的预警等级通知维护人员，以便其采取不同的处理措施[8]。

由于该系统基于历史波形数据进行建模，随着时间变化，每天定时将新增数据添加到初始建模数据中，通过数据初始分析发现磁盘空间日变化量较小，因此每半月对模型进行调整，以减少预测结果误差[9]。

天津市测震台网采用 ARIMA 时序模型对 AWS 存储服务器磁盘进行监控预警的同时，利用该模型对磁盘负载情况进行分析，统计 AWS 存储目录下徐庄子地震台全天每小时数据量变化，并与该台站全天各时段的台基噪声 *RMS* 统计情况进行对比（图 4）。通过对比发现台站波形数据量与台基噪声 *RMS* 值变化趋势相对一致[10]，因此可将台站观测数据变化情况作为台基噪声评估的一项指标。



(a) 徐庄子台基噪声*RMS*值全天统计　　(b) 徐庄子台24小时数量变化统计

图 4　基于 Python 时序分析法台站数据量变化与台站台基噪声对比

4　结语

通过基于 Python 语言的时间序列算法建模，天津测震台网实现了对地震波形数据存储空间使用情况的分级预测，对于系统维护人员准确判断系统运行情况、采取相应措施保障测震系统正常运行起到了极其重要的作用。同时利用 ARIMA 模型对测震台站数据实时变化量情况进行监控，达到评估台站环境噪声的目的。在建模应用过程中，要定时对模型进行调整，保证模型适用随时改变的初始数据，以达到预测结果准确的目的。

（下转第 196 页）

自发自收载波噪声比、接收中心站参考载波噪声比和第一信道发射衰减值在该时间段内数值是否正常，免去人工查阅每天值班日志的工作。

4 结束语

不同于一般的预警系统，该系统能够实时监测卫星站的工作状态和运行状态，一旦发现异常会立即发出报警，同时自动生成检测报告供工作人员分析，节省了大量的人力物力，从而更高效智能地保证了卫星站稳定运行。该系统实现了值班日志自动生成，这不仅给值班人员带来了方便，也有效解决了人工监测不准确、不及时、成本高等问题。这对于提升地震应急救援综合能力具有重要作用。

参考文献

[1] 李庆德.船用卫星通信地球站控制系统的研究与应用[D].南京：南京邮电大学，2011：34-36.

[2] 姚晔.低轨卫星通信系统的路由策略及星群管理关键技术研究[D].成都：电子科技大学，2005：23-26.

[3] 朱月秀.现代通信技术[M].北京：电子工业出版社，2003：46-48.

[4] 魏晓光.应急通信中的极限传输研究[D].北京：北京邮电大学，2009：41-43.

[5] 许一丁.卫星通信便携站应急选址模型研究[D].沈阳：东北大学，2012：22-25.

（上接第 173 页）

参考文献

[1] 孙志林，卢雅倩，黄赛花.港口吞吐量的马氏链-时序分析预测[J].浙江大学学报（工学版），2012，46(7)：1289-1294.

[2] 黄大荣，李劲.基于粗糙集理论的数据清洗模型[J].自动化技术与应用，2004，23(3)：11-13.

[3] 李红梅，唐岚.基于 Python 的交通流数据清洗[J].电子技术与软件工程，2016(9)：208-208.

[4] 刘芳，何飞.基于聚类分析技术的数据清洗研究[J].计算机工程与科学，2005，27(6)：70-71.

[5] 王巍.基于 ARIMA 模型的安徽省城镇化水平预测研究[J].赤峰学院学报（自然科学版），2016，32(14)：15-17.

[6] 彭海波，向洪普.基于 Python 的空间数据批量处理方法[J].测绘与空间地理信息，2011，34(4)：81-82.

[7] 叶丰艳.ARIMA 模型在门诊季节性时序预测中的应用[J].中国医院统计，2009，16(3)：239-241.

[8] 朱宏，王晓磊，刘磊，等.天津市地震局数据监控与展示系统的设计与实现[J].地震工程学报，2016，38(增刊 2)：231-235.

[9] 孙路强，刘磊，栗连弟，等.基于 PHP 语言的天津市测震台网综合管理系统的研制[J].地震工程学报，2015，37(增刊 1)：108-113.

[10] 王建国，刘春国，王伟，等.地震前兆数据库综合管理系统[J].大地测量与地球动力学，2013，33(增刊 1)：114-116.