ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.13, No.5, February 2017
E-mail: wltx@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于R语言时间序列的轿车销量分析及预测

赵玉新

(北京信息职业技术学院, 北京 100018)

摘要:该文数据来自数据堂网站,主要使用R语言为工具,进行数据分析,本次分析主要侧重于某型号轿车的时间序列分析,首先提取该轿车每月的销量情况,作为历史数据,然后进行分析预测,得出结论。

关键词:R语言;数据分析;轿车销量;时间序列分析

中图分类号:TP311 文献标识码:A 文章编号:1009-3044(2017)05-0016-03

DOI:10.14004/j.cnki.ckt.2017.0568

时间序列是以固定时间间隔为单位的一列数据,最常见的比如某只股票的每日股价走势图,每天的气象数据等。时间序列分析是统计分析的一个重要内容,由于基于历史数据可以进行预测,因此几乎每种统计分析软件都有时间序列的分析及预测功能。

时间序列常见的分析方法有:简单平均法、加权平均法和移动平均法等。还有 time series 里面两个强大的算法: Holt-Winters 和 ARIMA。

R语言具有功能强大的程序包,在数据计算,统计分析以及数据挖掘等方面都所向披靡,本文介绍轿车销量时间序列数据在R中统计分析及预测的实现。

1 数据情况

我们要对车型中大众朗逸的车型进行分析预测,所以在EXCEL中首先筛选出大众朗逸的销售数据。

月份	厂商	品牌	车型	本月销量
2013,10	上海大众	大众	大众朗逸	30272
2013,9	上海大众	大众	大众朗逸	29952
2013,8	上海大众	大众	大众朗逸	27096
2013,7	上海大众	大众	大众朗逸	24241
2013,6	上海大众	大众	大众朗逸	25610
2013,5	上海大众	大众	大众朗逸	30534
2013,4	上海大众	大众	大众朗逸	33933
2013,3	上海大众	大众	大众朗逸	38334
2013,2	上海大众	大众	大众朗逸	34889
2013,1	大众	上海大众	Lavida朗逸	48267
2012,12	大众	上海大众	Lavida朗逸	20036
2012,11	大众	上海大众	Lavida朗逸	31697
2012,10	大众	上海大众	Lavida朗逸	30637
2012,9	大众	上海大众	Lavida朗逸	24600
2012,8	大众	上海大众	Lavida朗逸	20239
2012,7	大众	上海大众	Lavida朗逸	11471

图1 大众朗逸销量月份数据(部分)

数据从2011年4月到2013年10月,大约是两年半的月销售数据。

为了操作方便,我们按月份升序排序,结果如下:

月份	厂商	品牌	车型	本月销量
Apr-11	大众	上海大众	Lavida朗逸	23439
May-11	大众	上海大众	Lavida朗逸	22602
Jun-11	大众	上海大众	Lavida朗逸	22980
Jul-11	大众	上海大众	Lavida朗逸	3591
Aug-11	大众	上海大众	Lavida朗逸	17293
Sep-11	大众	上海大众	Lavida朗逸	23852
Oct-11	大众	上海大众	Lavida朗逸	23121
Dec-11	大众	上海大众	Lavida朗逸	24071
Jan-12	大众	上海大众	Lavida朗逸	15718
Feb-12	大众	上海大众	Lavida朗逸	21996
Mar-12	大众	上海大众	Lavida朗逸	20263
Apr-12	大众	上海大众	Lavida朗逸	20000
May-12	大众	上海大众	Lavida朗逸	19705
Jun-12	大众	上海大众	Lavida朗逸	10325
Jul-12	大众	上海大众	Lavida朗逸	11471
Aug-12	大众	上海大众	Lavida朗逸	20239
Sep-12	大众	上海大众	Lavida朗逸	24600
Oct-12	大众	上海大众	Lavida朗逸	30637
Nov-12	大众	上海大众	Lavida朗逸	31697

图2 大众朗逸销量按月份升序排序的数据(部分)

2数据处理

首先我们将excel数据导入到R中,将excel文件以剪贴板的格式进行保存。然后使用read.table函数进行导入:

```
a<-read.table("clipboard",he=T)
```

	月份	厂商	品牌	车型	本月销量
1	Apr-11	大众	上海大众	Lavida朗逸	23439
2	May-11	大众	上海大众	Lavida朗逸	22602
3	Jun-11	大众	上海大众	Lavida朗逸	22980
4	Jul-11	大众	上海大众	Lavida朗逸	3591
5	Aug-11	大众	上海大众	Lavida朗逸	17293
6	Sep-11	大众	上海大众	Lavida朗逸	23852
7	Oct-11	大众	上海大众	Lavida朗逸	23121
8	Dec-11	大众	上海大众	Lavida朗逸	24071
9	Jan-12	大众	上海大众	Lavida朗逸	15718
10	Feb-12	大众	上海大众	Lavida朗逸	21996
11	Mar-12	大众	上海大众	Lavida朗逸	20263
12	Apr-12	大众	上海大众	Lavida朗逸	20000
13	May-12	大众	上海大众	Lavida朗逸	19705
14	Jun-12	大众	上海大众	Lavida朗逸	10325
15	Jul-12	大众	上海大众	Lavida朗逸	11471
16	Aug-12	大众	上海大众	Lavida朗逸	20239
17	Sep-12	大众	上海大众	Lavida朗逸	24600
18	Oct-12	大众	上海大众	Lavida朗逸	30637

图3 R中的大众朗逸车型销量数据

根据历史数据,首先绘制时间序列图,如下:

```
plot.ts(a$本月销量,xlab="月份")
```

收稿日期:2017-02-15

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

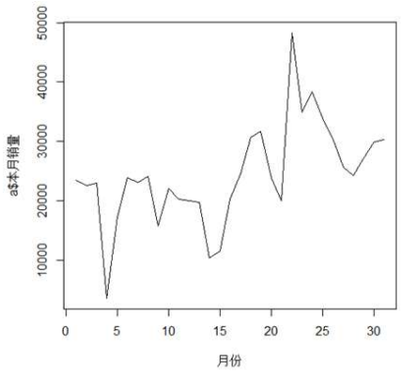

图4　朗逸月度销售数据时序图

从图中可以看出，是大众朗逸31个月的销售数据，没有明显的周期和季节趋势，2013年1月，创下销售记录，48267台，应该是春节前，是车辆销售旺季。2011年7月出现了销售销售的低谷，销量只有3000多台。

3 时间序列检验分析

3.1 自相关检验

对于非平稳数列的数据，ACF自相关图不会趋向于0，或者趋向0的速度很慢。自相关图中的两条虚线表示置信界限，是自相关系数的上下界。

下面绘制原始数列的自相关图：

```
acf(a$本月销量,lag.max=30)
```

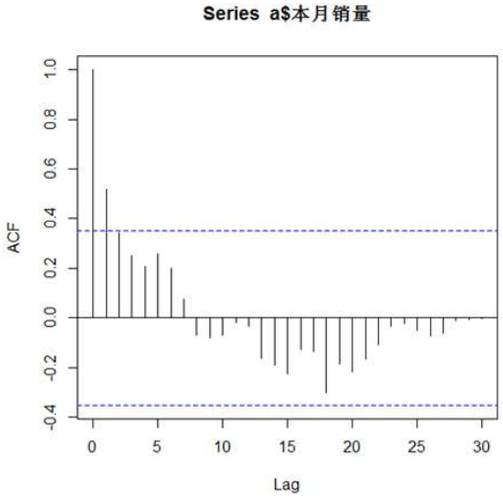

图5　原始数列的自相关图

3.2 单位根检验

```
unitrootTest(a$本月销量)
```

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    DF: -0.3973
  P VALUE:
    t: 0.5321
    n: 0.5823

Description:
 Mon Feb 13 20:08:34 2017 by user: bitc024
```

图6　单位根检验结果图

从以上几幅图进行分析，图4中的时序图，可以看出有连年递增趋势，为非平稳序列。从自相关检验结果看，自相关系数长期大于零，进一步表明为非平稳序列；单位根检验结果p值显著大于0.05，也判断其为非平稳序列。

4 ARIMA建模分析

4.1 非平稳序列差分

差分，即Integrated。一阶差分是把原数列每一项减去前一项的值。二阶差分是一阶差分基础上再来一次差分。差分一直得到平稳序列为止。R中使用diff()函数对时间序列进行差分运算。

```
> diffsale<-diff(a$本月销量)
> diffsale
 [1] -837  378 -19389  13702  6559 -731  950 -8353
6278 -1733
[11] -263 -295 -9380  1146  8768  4361  6037  1060
-7776 -3885
[21] 28231 -13378  3445 -4401 -3399 -4924 -1369
2855  2856  320
```

差分后再进行检验：

```
plot.ts(diffsale)
acf(diffsale,lag.max=30)
unitrootTest(diffsale)
```

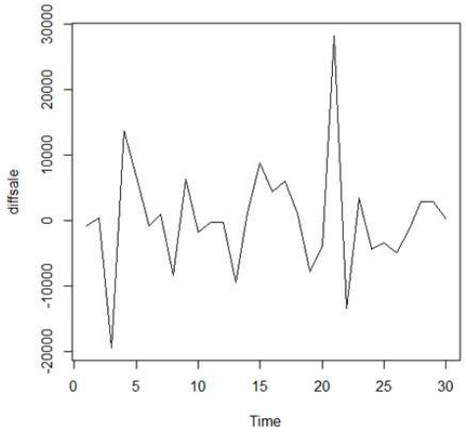

图7　一阶差分后的时序图

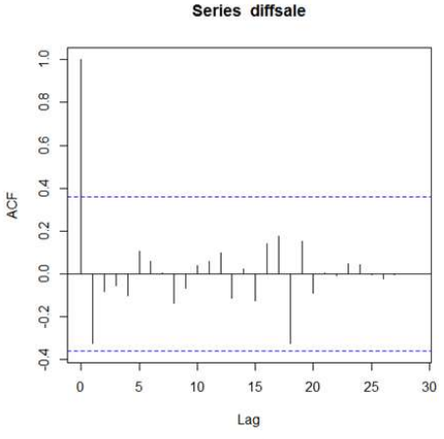

图8　自相关检验图

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    DF: -5.0996
  P VALUE:
    t: 9.78e-06
    n: 0.1046

Description:
 Mon Feb 13 20:55:34 2017 by user: bitc024
```

图9　单位根检验图

一阶差分后，时序图在均值附近平稳波动，自相关有很强的短期相关性，单位根检验P值远小于0.05。所以一阶差分后序列表现为平稳。

4.2 时间序列模型识别定阶

从一阶差分后的自相关图可以看出，ACF值迅速跌入置信区间，没有收敛趋势，显示出拖尾性。所以考虑选用AR模型一阶差分后的序列，即对原始序列使用ARIMA(1,1,0)模型。

```
> arima<-arima(a$本月销量,order=c(1,1,0))
> arima
Call:
arima(x = a$本月销量, order = c(1, 1, 0))
Coefficients:
          ar1
      -0.3116
s.e.   0.1696
sigma^2 estimated as 62185411:  log likelihood = -311.8,
aic = 627.61
```

4.3 白噪声检验

检验残差序列是否为白噪声序列，使用Box.test函数

```
> Box.test(arima$residuals,lag=5,type="Ljung-Box")
        Box-Ljung test
data:  arima$residuals
X-squared = 3.5647, df = 5, p-value = 0.6136
```

可以看出，p值=0.6136，大于0.05，通过白噪声检验。

5 ARIMA模型预测

R中可以通过forecast包对未来的序列值进行预测，预测未来5个月朗逸的月销量以及置信度上下界，语句如下：

```
> forecast.Arima(arima,h=5,level=c(80,95))
   Point Forecast  Lo 80    Hi 80    Lo 95    Hi 95
32       30172.28 20066.26 40278.31 14716.452 45628.11
33       30203.36 17934.32 42472.39 11439.491 48967.22
34       30193.67 15580.54 44806.81 7844.816  52542.53
35       30196.69 13710.44 46682.95 4983.139  55410.24
36       30195.75 11988.24 48403.26 2349.771  58041.73
```

可以清晰地看到预测值。还可以绘制原始及预测值图形，使用plot完成。

```
> forecast<-forecast.Arima(arima,h=5,level=c(80,95))
> plot.forecast(forecast)
```

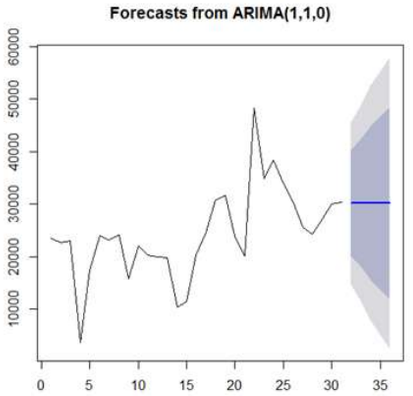

图10　时间序列预测图

6 结束语

以上是笔者对朗逸轿车月度销售数据分析研究，主要使用R语言的时间序列分析方法，绘制序列图，检验其是否为平稳序列，非平稳序列进行差分处理，直到平稳为止。然后使用ARIMA方法进行分析建模，再进一步完成预测。

参考文献：

[1] 张良均,等 .R语言与数据挖掘[M].
[2] 数据堂网站. http://www.datatang.com/.

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net