数据分析与处理技术——时间序列分析

商学院 徐宁

参考资料

阅读：

1. 中文（第2版）：预测：方法与实践 (otexts.com)

2. 英文第三版：Forecasting: Principles and Practice (3rd ed) (otexts.com)

时间序列分析

时间序列变量

ts类变量操作

ts变量可视化

空缺值处理

时间序列数据

在R语言中，时间序列是向量的拓展结构

类型符号：ts变量

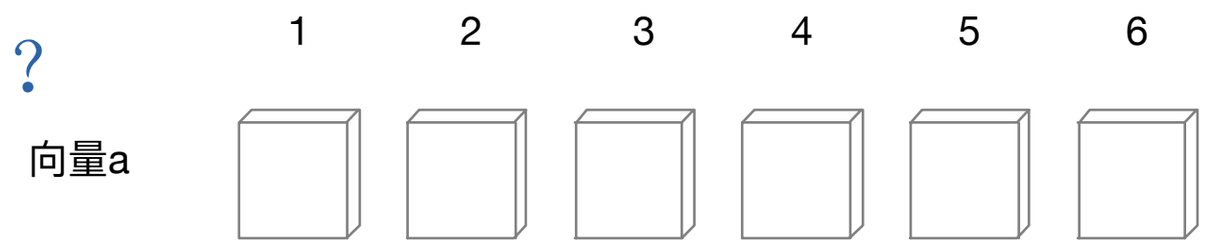
时间序列与普通序列有什么差别？

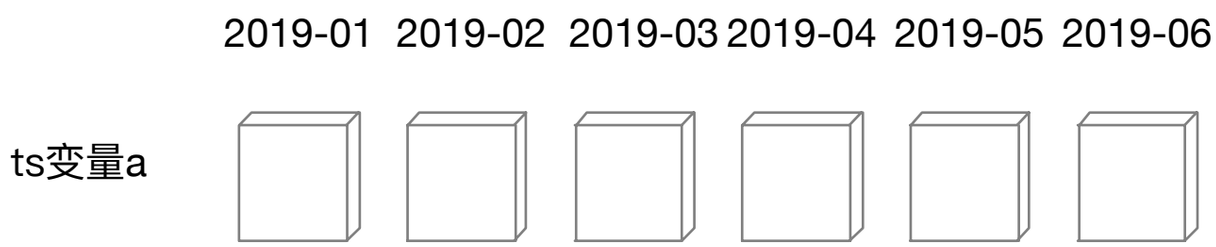
时间序列常见的类型

年度数据
季度数据
月份数据
周数据
日期数据

……

计算机如何组织时间序列数据





时间序列变量

ts()函数转化向量

frequency参数

365	日期数据
52	周数据
12	月数据
4	季度数据
1	年数据

```
> x=rnorm(20,mean=30,sd=5)
> t=ts(x,start = c(2016,1),frequency = 4)
> t
         Qtr1     Qtr2     Qtr3     Qtr4
2016 32.30360 28.86309 22.81964 24.36428
2017 24.79465 27.18220 22.63273 25.62494
2018 26.62748 40.13740 33.08386 28.66923
2019 32.62693 34.16754 32.38213 29.48067
2020 28.20993 31.62933 33.60743 35.59776
```

时间窗口

通过索引取子集

```
> x[1:3]
[1] 32.30360 28.86309 22.81964
```

window()取窗口数据

```
> window(t,start=c(2016,1),end=c(2017,4))
          Qtr1     Qtr2     Qtr3     Qtr4
2016 32.30360 28.86309 22.81964 24.36428
2017 24.79465 27.18220 22.63273 25.62494
```

多序列变量

创建两个相同长度的时间序列

```
a <- rnorm(100,mean = 10,sd=5)
b <- runif(100,min = 15,max = 30)
a <- ts(a,start = 2000,frequency = 4)
b <- ts(b,start = 2000,frequency = 4)
```

当并入变量y时，此时的y成为矩阵类型，同时也是时间序列类型

```{r}
y=cbind(a,b)
class(y)
```

```
[1] "mts"    "ts"     "matrix"
```

tsibble类变量是将ts、tibble类型结合的时间序列表格类型，参考：https://tsibble.tidyverts.org/

分析电力市场数据

发电厂非常关注电力市场的需求，通常要提前安排下一个月的燃料采购、人员配置等生产计划。某电厂拿出了从1991年7月至2008年6月的数据尝试进行分析以后的发展趋势。

时间序列变量a10（需加载fpp2）

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1991							3.53	3.18	3.25	3.61	3.57	4.31
1992	5.09	2.81	2.99	3.20	3.13	3.27	3.74	3.56	3.78	3.92	4.39	5.81
1993	6.19	3.45	3.77	3.73	3.91	4.05	4.32	4.56	4.61	4.67	5.09	7.18
1994	6.73	3.84	4.39	4.08	4.54	4.65	4.75	5.35	5.20	5.30	5.77	6.20
1995	6.75	4.22	4.95	4.82	5.19	5.17	5.26	5.86	5.49	6.12	6.09	7.42
1996	8.33	5.07	5.26	5.60	6.11	5.69	6.49	6.30	6.47	6.83	6.65	8.61
1997	8.52	5.28	5.71	6.21	6.41	6.67	7.05	6.70	7.25	7.82	7.40	10.10
1998	8.80	5.92	6.53	6.68	7.06	7.38	7.81	7.43	8.28	8.26	8.60	10.56
1999	10.39	6.42	8.06	7.30	7.94	8.17	8.72	9.07	9.18	9.25	9.93	11.53
2000	12.51	7.46	8.59	8.47	9.39	9.56	10.83	10.64	9.91	11.71	11.34	12.08
2001	14.50	8.05	10.31	9.75	10.85	9.96	11.44	11.66	10.65	12.65	13.67	12.97

时间序列可视化

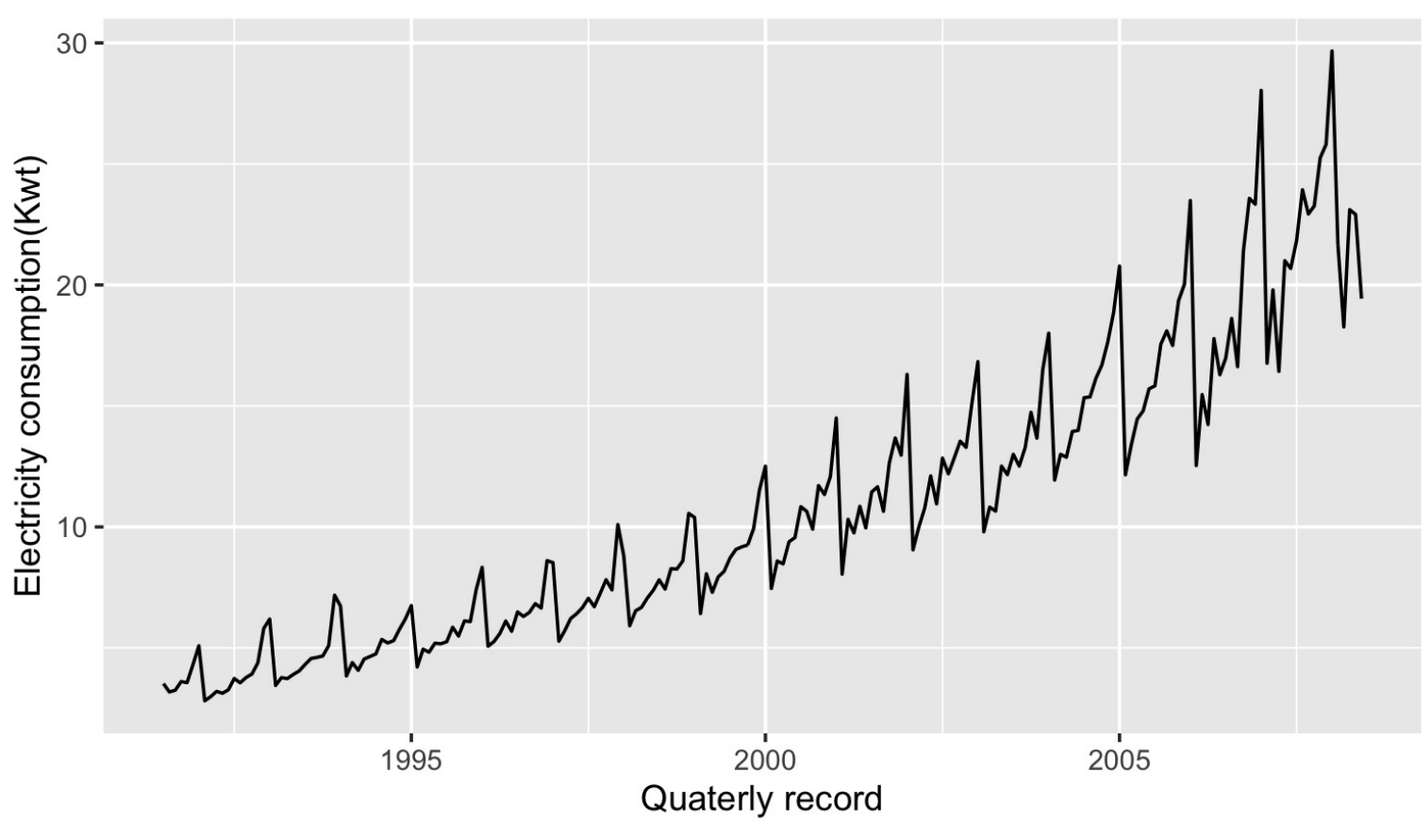
autoplot()/autolayer()函数取代了ggplot()函数功能，兼容ts类变量

案例：a10

需加载forecast工具包

```
> autoplot(co2)
```

```
autoplot(a10)+
  xlab("Quaterly record")
  ylab("Electricity consumption(Kwt)")
```



绘制多序列

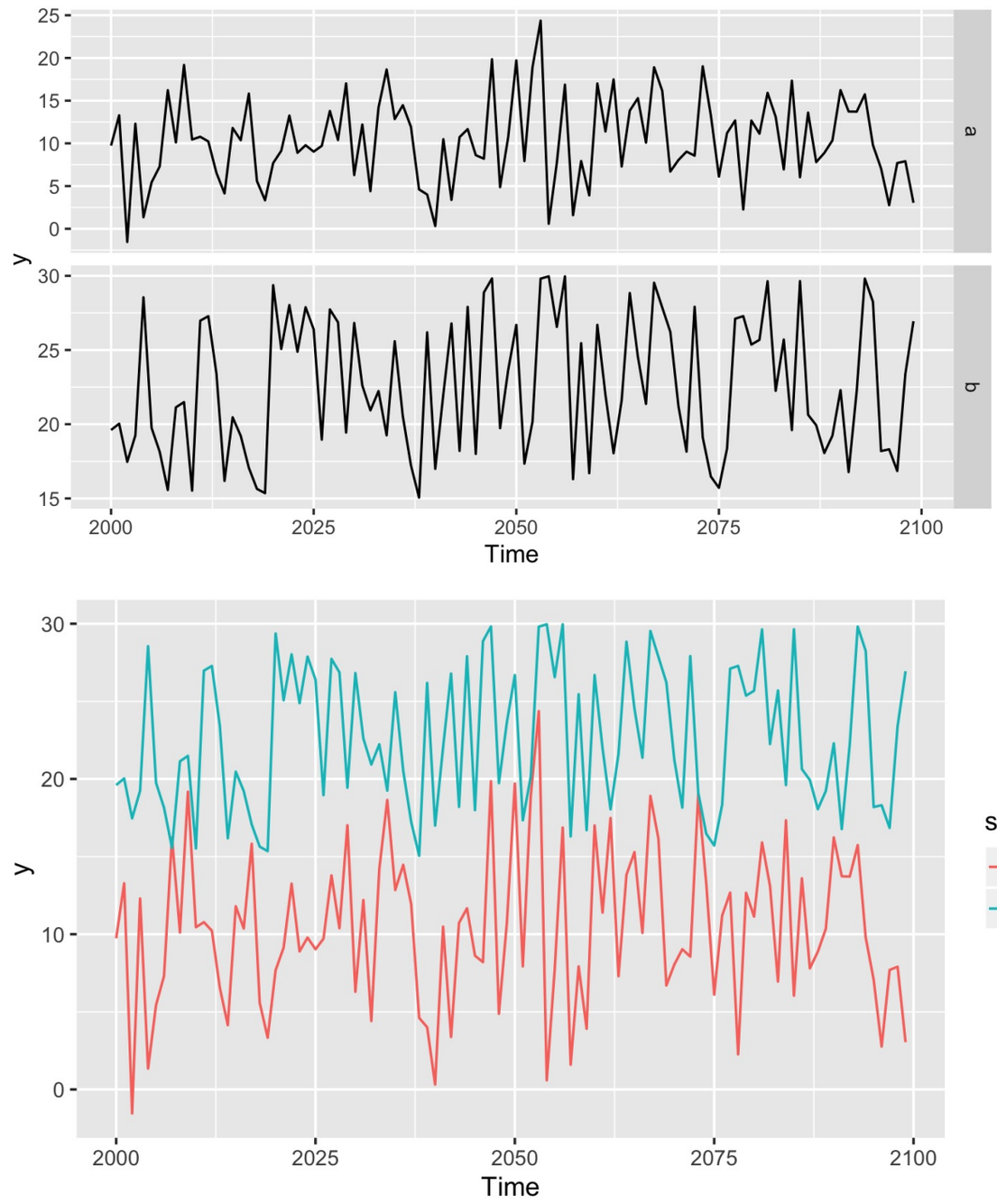
当绘制多序列时，一种方式通过facets参数进行分面

```
autoplot(y,facets = T)
```

另外一种则利用autolayer逐层加序列

```
autoplot(a,series = "Series a")+
  autolayer(b,series = "Series b")
```

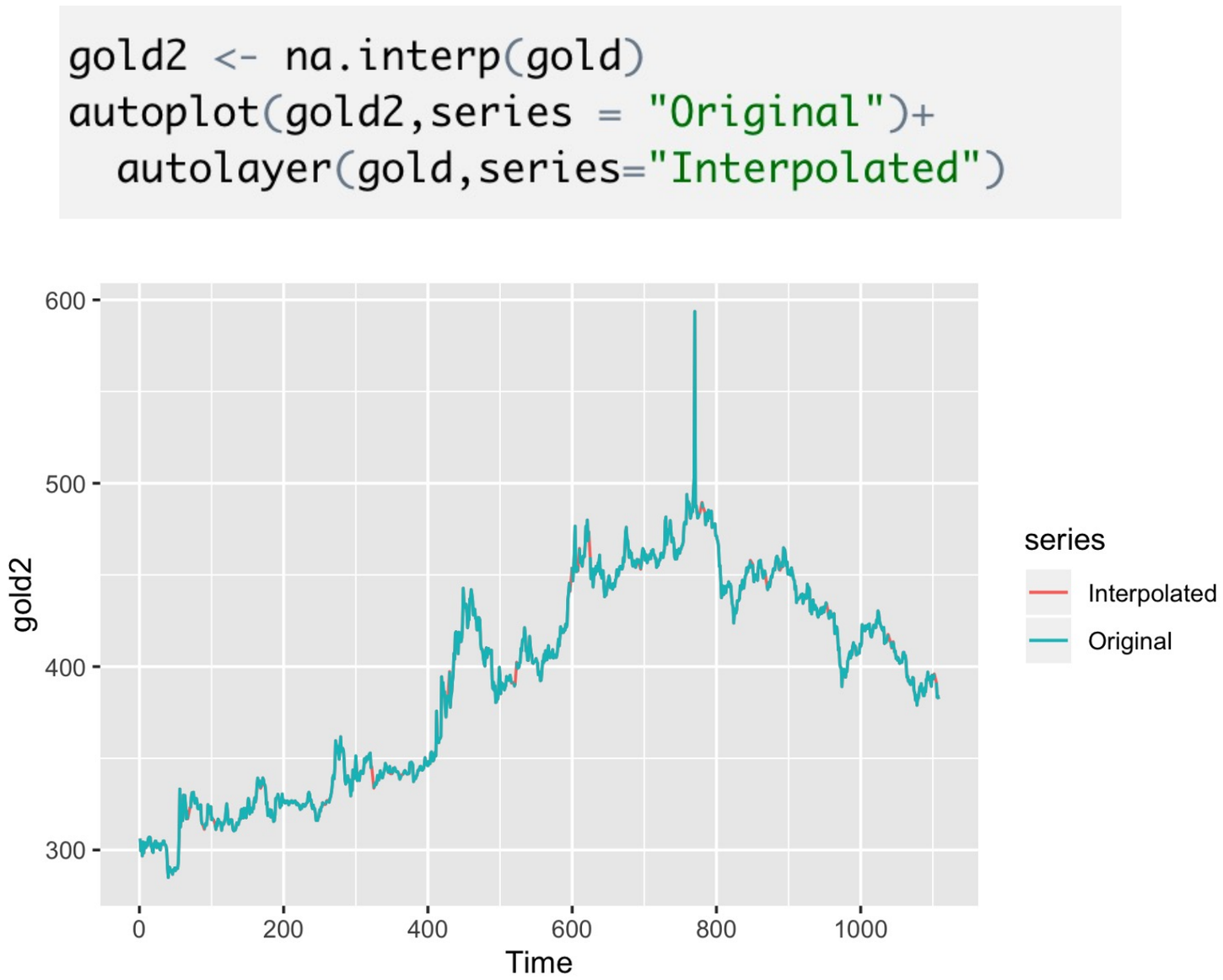
此时的序列将处于同一幅页面中



ts变量空值处理

时间序列中空缺值不宜删除处理，通常利用前后数据的连贯性进行补缺估算

```
gold2 <- na.interp(gold)
autoplot(gold2,series = "Original")+
  autolayer(gold,series="Interpolated")
```



特征分解

时间序列特征分解

序列噪声分析

特征分解原理

时间序列的特征可以大致分成如下几类

- Trend：长期趋势，记做T
- Seasonal：季节变动，记做S
- Cyclic：周期趋势，记做C
- 剩余的特征被作为剩余量记做Remainder，记做R

由于C特征通常长于两年，与T特征可以合并为T-C特征，也简化记为T特征。时间序列通常可以分为长期趋势、季节变动和周期趋势，分解方法分为加法型和乘法型，即：

$$y_t = S_t + T_t + R_t,$$

$$y_t = S_t \times T_t \times R_t.$$

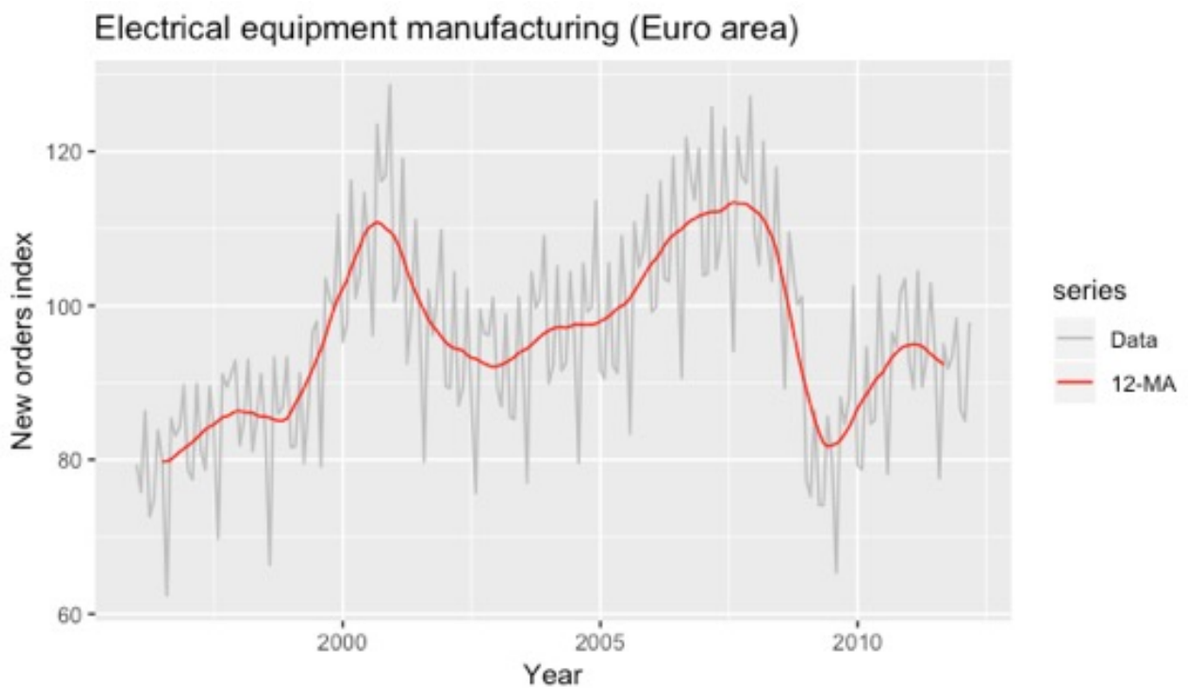
移动平均提取趋势线

移动平均方法能够抹平由于周期带来的数据波动，这中特性为提取趋势带来了方便

$$\hat{y}_{T+h|T} = \frac{1}{T}\sum_{t=1}^{T} y_t,$$

```
autoplot(elecequip,series = "Data")+
  autolayer(ma(elecequip,12),series = "12-MA")
```

ma(y,n)函数对序列y进行n期平滑生成



季节性特征分析

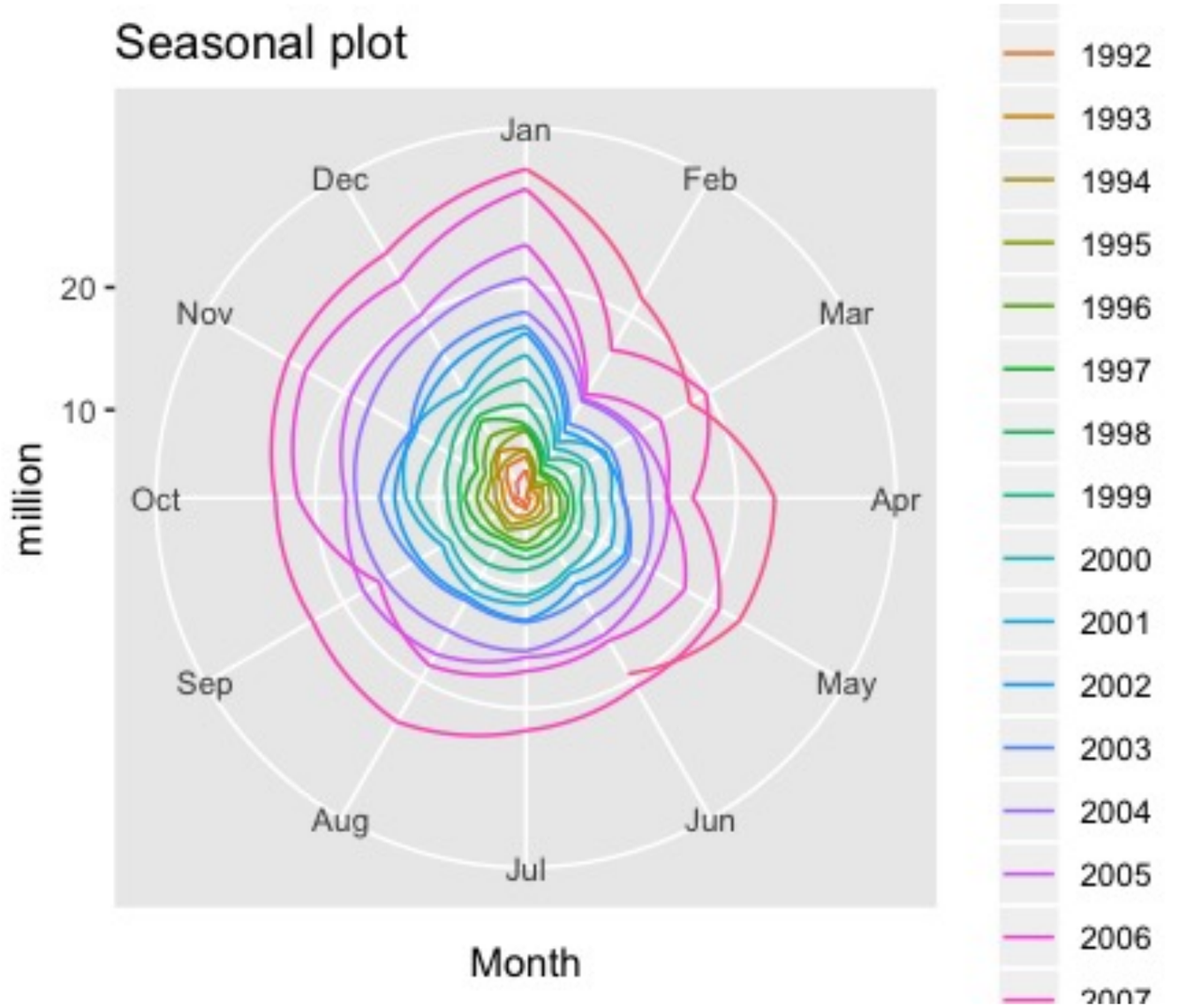
分析思路：将每年数据拆分成单独数据段，进行比对。

问题1:你从图中看出了每年度季节特征如何变化吗？

问题2:随着年度的发展季节特征在如何变化？

问题3:这两种图之间是什么关系？

```
> ggseasonplot(a10)
> ggseasonplot(a10,polar=TRUE)
```



特征子序列分解

如何根据特征将原来的序列拆分成多个子序列的组合？

时间序列分解为三个特征子序列：

```
> deseries=decompose(a10)
```

列表变量deseries
- seasonal
- trend
- random

```
> autoplot(deseries)
```

a10

||

季节子序列

+

趋势子序列

+

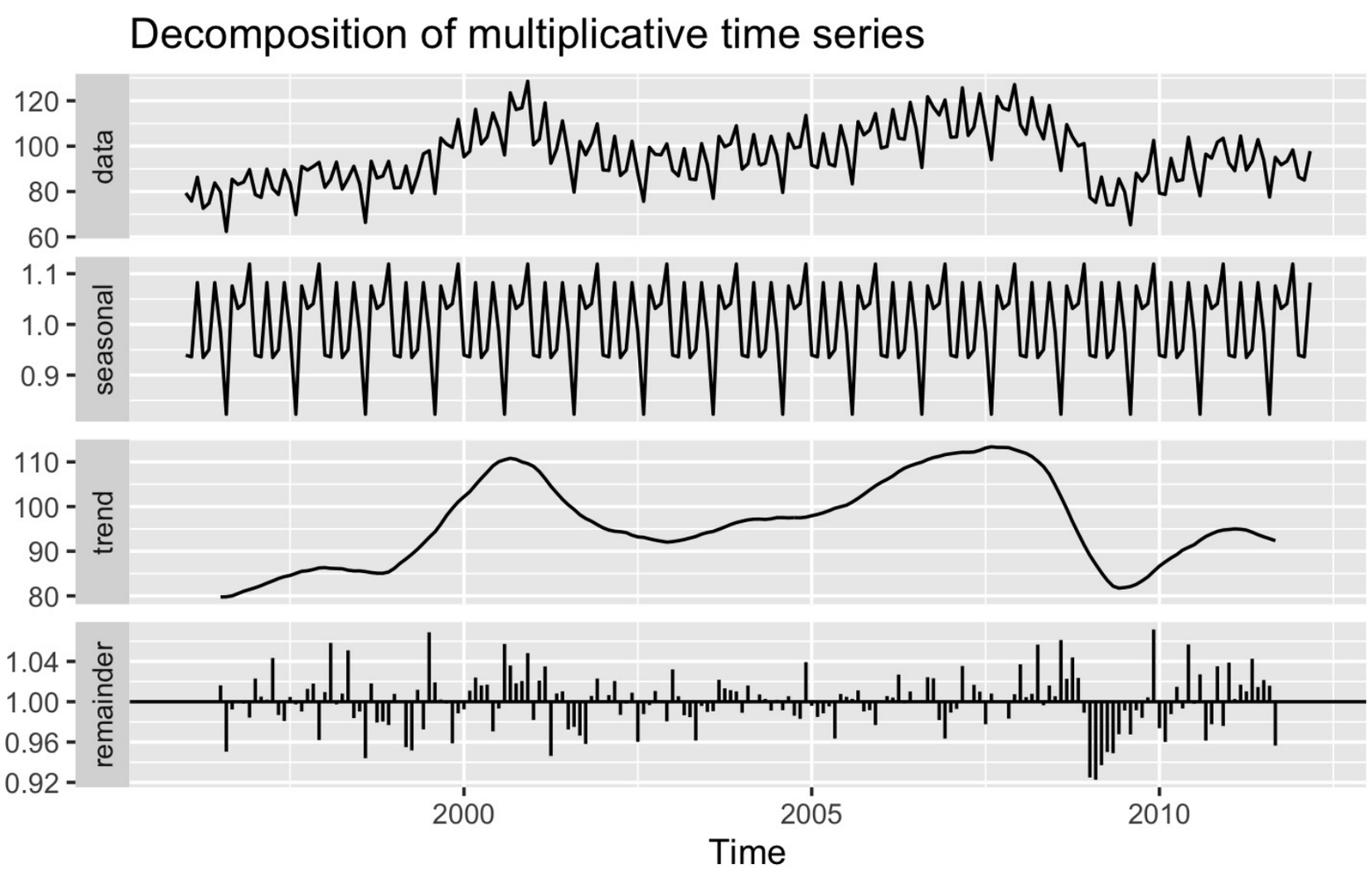
白噪声序列

特征子序列分解

decompose()函数将序列拆分成季节特征、趋势特征和剩余量三个子序列。type参数用于指定加法分解或乘法分解，默认为加法

```
deseries=decompose(elecequip,type = "multiplicative")
autoplot(deseries)
```

除基础分解方法之外，其他优化的时间序列特征分解方法还有x11 seats stl等



序列噪声分析

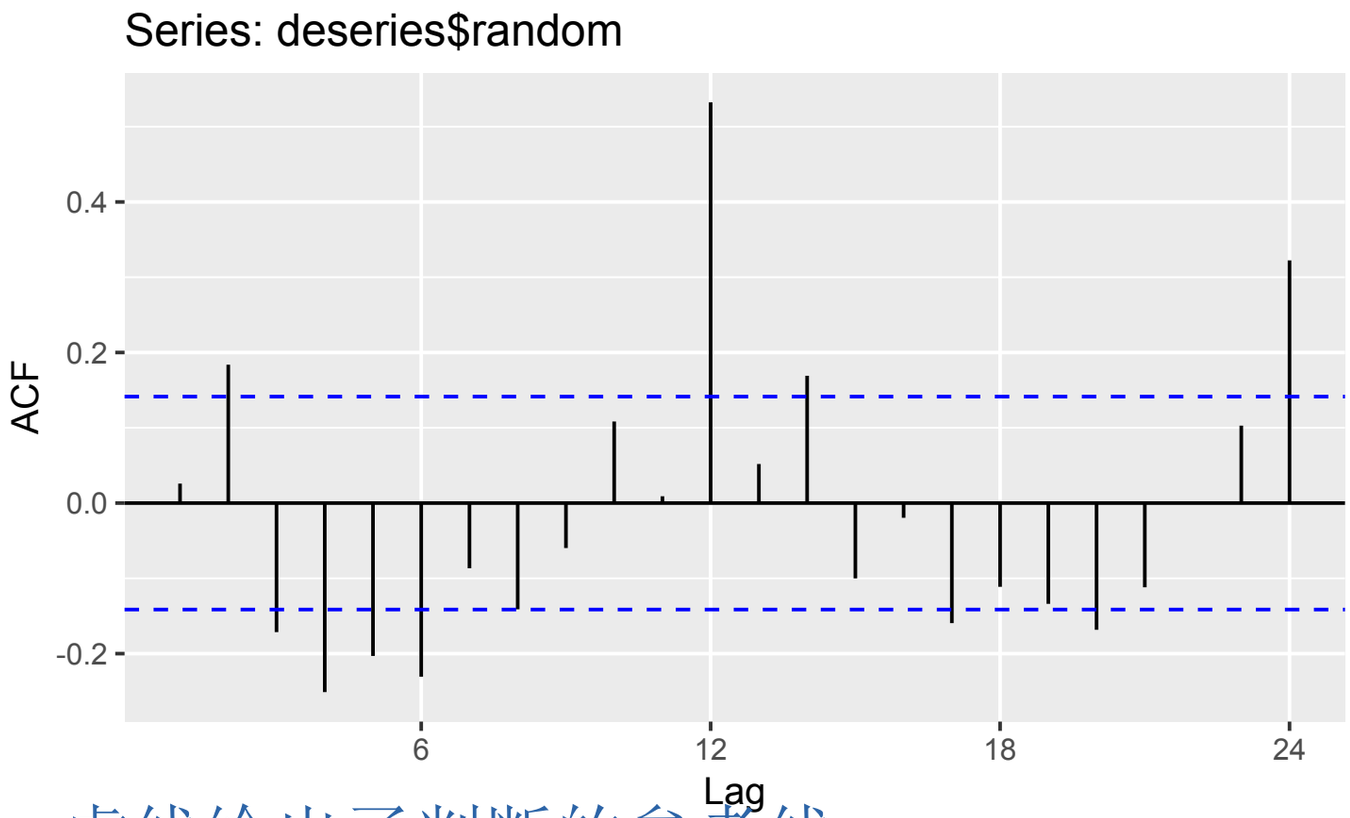
判断白噪声序列的思路：剩余子序列是否还包含规律性特征

原始数据与提前1期数据序列计算相关系数

原始数据与提前2期数据序列计算相关系数

计算代码

```
> ggAcf(deseries$random)
```



序列自相关降低到0显然不现实，虚线给出了判断的参考线。

数据的差分与滞后

Arima预测模型

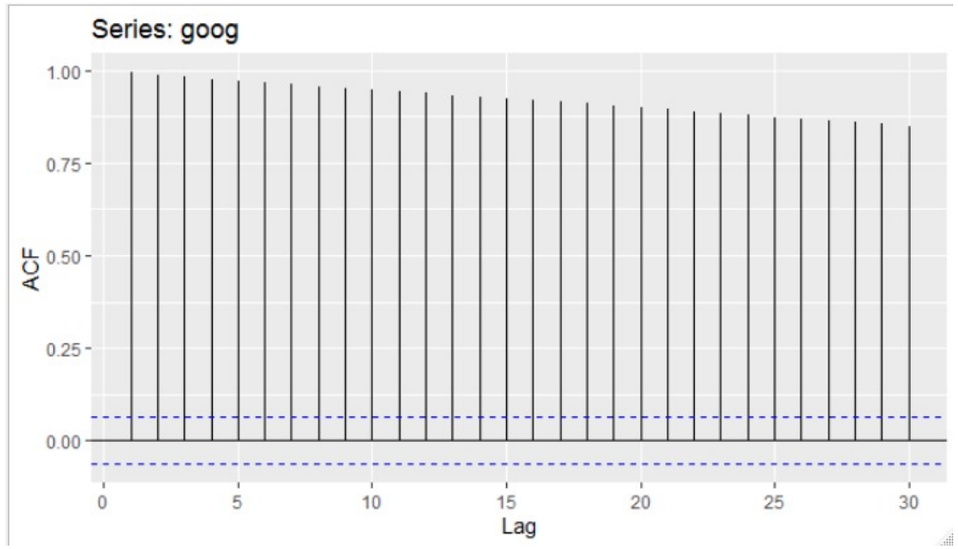
时间序列预测模型

差分序列

平稳的时间序列不随观测时间的变化而变化，例如像白噪声序列那样的几乎完全随机，与观察的时间没多大关系。

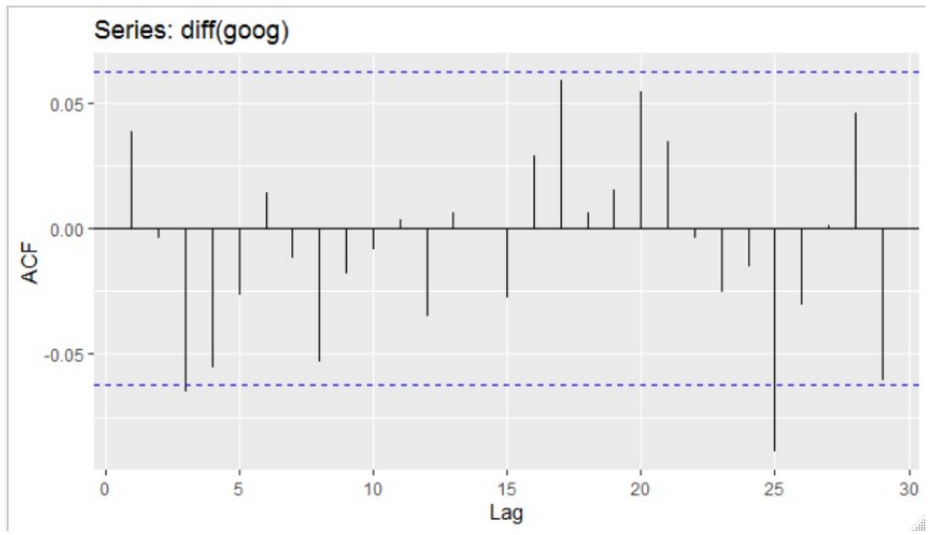
以goog数据(股市价格)为例，直接计算其ACF图如图

ggAcf(goog) →



差分之后的数据ACF图已经趋于平稳，这类数据称为平稳时间序列

做差分序列：diff(goog)
差分序列ACF图：ggAcf(diff(goog)) →



向量自回归

向量自回归

Autoregression model(简称AR)利用序列自身的滞后期作为自变量做回归，它的阶数p指模型中的自回归变量个数，记做AR(p)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

移动平均模型

Moving average model(简称MA)不同于AR用滞后变量做回归，MA用白噪声作为自变量做回归，阶数q指模型中的滞后变量个数。

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

arima模型计算

差分移动平均自回归模型

ARIMA(p,d,q)模型则是综合了AR和MA模型，其中p为自回归项数、q是移动平均项数，d则是差分阶数

```
fc=auto.arima(elecequip)
```

训练模型

```
pre=forecast(fc,15)
```

对模型进行预测，设置h=15

```
autoplot(pre)
```

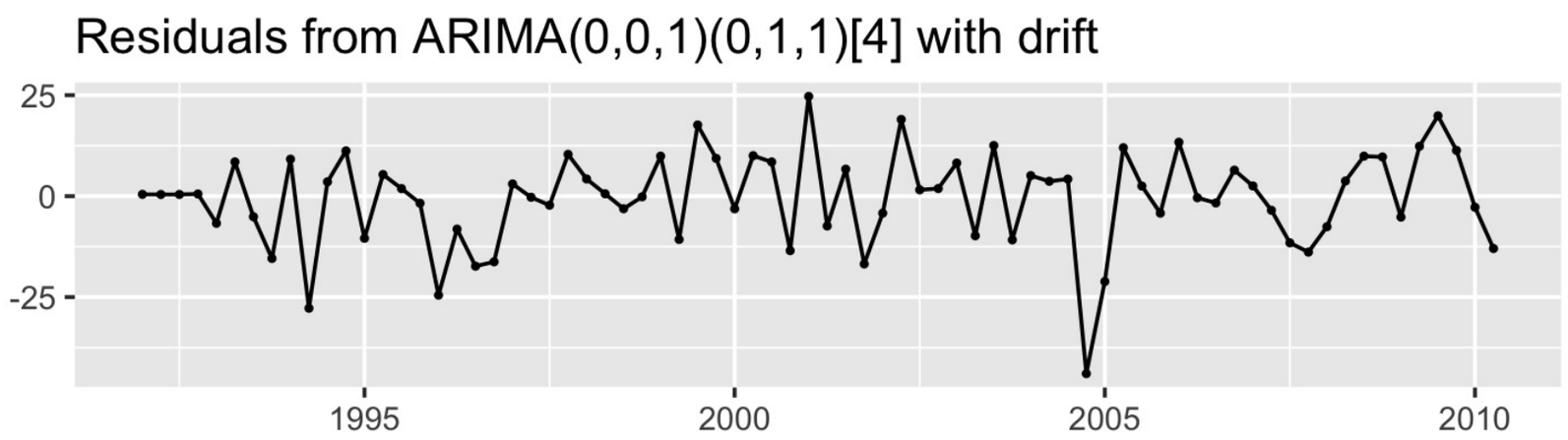
autoplot可以直接对时间序列的训练模型进行绘图

精度检验

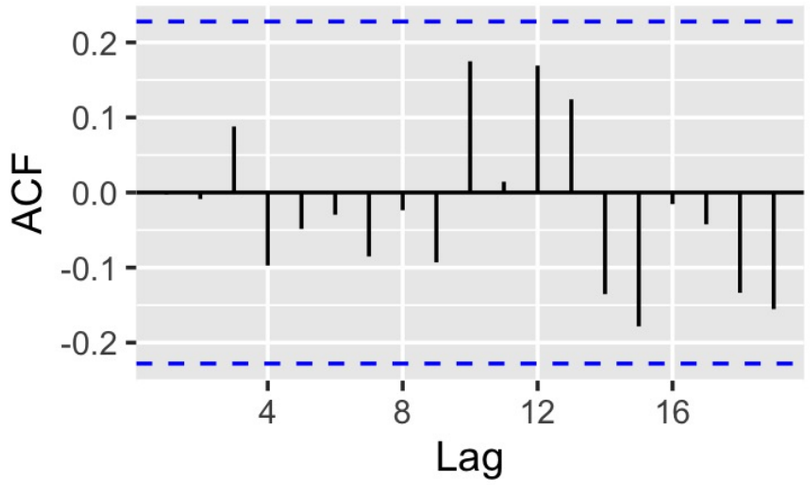
利用chechresiduals函数检验模型的拟合效果

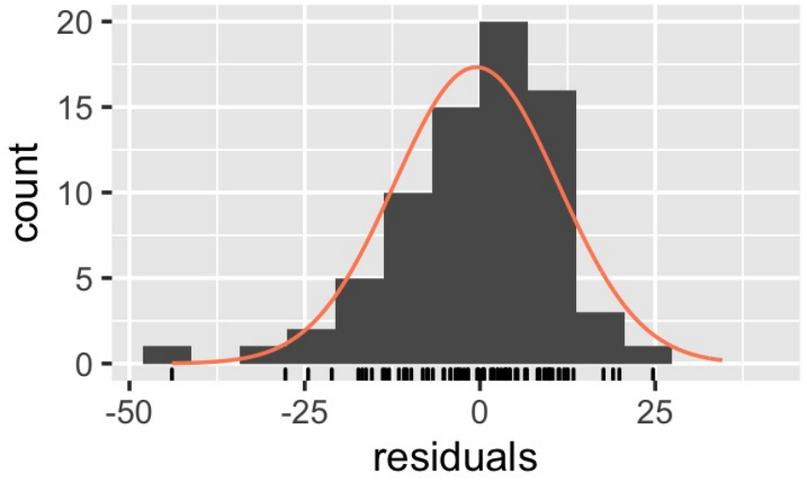
```
checkresiduals(fcarima)
```

残差值 →



残差自相关分析及分布状况 →





练习

1.尝试新的分解方法对elecequip序列进行分解

- x11分解法——seasonal工具包seas()函数
- stl分解法——基础包stl函数

2.创建一个正态分布随机数，观察其自相关图的特征，分析a10数据集的acf图

练习数据

练习数据

- ausbeer 啤酒厂销量
- melsyd 墨尔本-悉尼经济舱乘客客流量
- a10
- goog 谷歌公司在纳斯达克股票收盘价（2013年）

拓展阅读：《Forecasting:Principles and Practice》
https://otexts.com/fppcn