**WEEK 8**

###Mining Time Series Data

set.seed(123)

data1<-rnorm(12)

#create time series data as quarterly data starting

#from 2nd quarter of 2022也就是按季度排，从22年第二季开始

X<-ts(data1, frequency = 4,start=c(2022,2))

#create monthly time series data starting

#from February 2021也就是按月份排，21年2月开始

X2<-ts(data1,frequency = 12, start=c(2021,2))

#检查变量类别用class

class(data1)

class(X)

X2

#r adopting ISO 8601 format

#系统的时间，也就是现在电脑显示的日期和时间

date<-Sys.Date()

time<-Sys.time()

date2<-as.Date("2023-5-12")

date2

class(date2)

#create date starting from 2016-01-01 on to 2018-12-31

daily\_index<-seq.Date(from=as.Date("2016-01-01"),

to=as.Date("2018-12-31"),

by="day")

#3-day interval三天一个数据

day3\_index<-seq.Date(from=as.Date("2016-01-01"),

to=as.Date("2018-12-31"),

by="3 day")

#data dates.formats2.csv

data\_df<-read.csv(file.choose(),stringsAsFactors = F,

sep=";")

head(data\_df)

class(data\_df)

str(data\_df)

#transform character to date format转化日期格式

attach(data\_df)

US\_format\_New<-as.Date(US\_format,format="%m/%d/%Y")

US\_format\_New

Japanese\_format\_new<-as.Date(Japanese\_format,format="%d/%m/%Y")

Japanese\_format\_new

#Excel\_Numeric\_Format

#in excel1, the origin of date is December 30, 1899

#in R, the origin January 1, 1970

#Excel数字日期格式转换为 R 中的日期格式,origin指定了Excel中日期的起点

Excel\_f<-as.Date(Excel\_Numeric\_Format,

origin=as.Date("1899-12-30"))

Excel\_f

Excel\_Numeric\_Format

#Class time series data in R

#POSIX class

time\_str<-"2017-01-20 23:59:34"

time\_str\_ctl<-as.list.POSIXct(time\_str)

#zoo

install.packages("TSstudio")

library(TSstudio)

library(zoo)

data(US\_indicators)

str(US\_indicators)

Vehicle\_Sales1<-zoo(x=US\_indicators$'Vehicle Sales',

frequency=12)

Vehicle\_Sales1

#ts class(下面那坨是TSstudio包的，可以敲这个ts\_plot(USgas)查看)

data(USgas)

class(USgas)

#xts class

library(xts)

version#检查版本用的，有些包可能要更新才能用

Vehicle\_Sales2<-xts(x=US\_indicators$'Vehicle Sales',

frequency=12,

order.by=US\_indicators$Date)

Vehicle\_Sales2

#####Time series decomposition#####

#seasonal component

data(USgas)

ts.plot(USgas, main="US Monthly Natural Gas consumption",

ylab="Billion Cubic Feet",xlab="year")#时间序列线图

ts\_heatmap(USgas)#时间序列热图

#cycle component

data(USUnRate)

ts.plot(USUnRate, main="US Monthly Unemployment Rate",

ylab="Unemployment Rate",xlab="year")

ts\_heatmap(USUnRate)

#Types of decomposition

#additive structure

ts.plot(USgas, main="US Monthly Natural Gas consumption",

ylab="Billion Cubic Feet",xlab="year")

usg\_decompose<-decompose(USgas)#时间序列进行分解，加法结构

plot(usg\_decompose)

#extract each component 列出各个组件的名称

names(usg\_decompose)

#seasonal component

usg\_decompose$seasonal

#trend component

usg\_decompose$trend

#random component

usg\_decompose$random

#multiplicative structure

data(AirPassengers)

ts.plot(AirPassengers, main="Monthly Airline Passenger",

ylab="Thousand of Passengers",xlab="year")

air\_decompose<-decompose(AirPassengers,

type="multiplicative")#分解，乘法结构

plot(air\_decompose)

#extract each component

names(air\_decompose)

#seasonal component

air\_decompose$seasonal

#trend component

air\_decompose$trend

#random component

air\_decompose$random

#convert multiplicative structure to additive structures.乘法结构变加法

library(forecast)

#multiplicative structure

data(AirPassengers)#名为 AirPassengers 的时间序列数据

#计算 AirPassengers 数据的 Box-Cox 变换的 lambda 值

AirP<-BoxCox.lambda(AirPassengers)

#根据之前计算的 lambda 值进行变换。这个过程旨在将数据转换为对数形式

AirPT<-BoxCox(AirPassengers,lambda = AirP)

par(mfrow=c(1,2))#一行两列，以便同时显示两张图

plot.ts(AirPassengers)

plot.ts(AirPT)

###Time series forecasting

#data Tovel.csv

data<-read.csv(file.choose(),header=T)

yt<-ts(data)

#图片边界par(mfrow=c(2,1))，这个可能大了，反正我换了个显示尺寸喵#

par(mar = c(1, 1, 1, 1))

plot.ts(yt, main="Paper Towel Daily Sales",ylab="yt")

#data is not stationary不平稳

#do a differentation差分操作用于使数据更加平稳

#diff() 函数进行差分操作，减去前一个时间点的值来消除趋势或周期性，使数据更平稳

ztl<-diff(yt,differences=1)

plot.ts(ztl, main="1st difference data",ylab="zt")

#ARIMA model identification

#自相关函数（ACF）和偏自相关函数（PACF），用于帮助识别 ARIMA 模型的参数。

#自相关函数显示了序列与其自身滞后版本之间的相关性，帮助识别时间序列中可能的自相关结构。

acf(ztl, main="Sample Autocorrelation Function")

#偏自相关函数显示了序列在移除其它滞后效应后与当前滞后的相关性，有助于确定时间序列中的部分自相关结构。

pacf(ztl,main="Sample Partial Autocorrelation Function")

#based on ACF & PACF plot

#the model should be ARIMA(0,1,1)

#fit ARIMA(0,1,1) to dataset

#arima() 函数拟合了 ARIMA(0,1,1) 模型到时间序列数据 yt 上

#order = c(0, 1, 1) 指定了模型的阶数，其中 0 表示非季节性自回归（AR）阶数

#"1" 表示非季节性移动平均（MA）阶数

model<-arima(yt,order=c(0,1,1))

summary(model)

#plot model to observed dataset

plot.ts(yt,main="Paper Towel Daily Sales",ylab="yt")

lines(fitted(model),col="red",lty=2)#拟合值

legend("bottomleft",

c("Observed Data","Model ARIMA(0,1,1)"),

col=c(1,2),lty=c(1,2))#图例

#Diagnostic check

#residual analysis

library(forecast)

f.value<-forecast(model,h=5)

residual<-f.value$residuals

#residual should not be autocorelated模型的残差（residual）的自相关函数图

acf(residual, lag.max = 20)

#based of ACF plot

#residual is not autocorelated

#residual should approximate normal distribution

#通过观察残差的直方图，可以初步检查残差是否近似于正态分布

hist(residual)

#the residual should having a constant variance over time series

#残差随时间变化的图形。这有助于观察残差是否在时间序列上保持恒定的方差

plot.ts(residual)

#since all the assumptions is fulfiled, the ARIMA(0,1,1)

#model can be used for prediction

#残差看起来在自相关性、正态性和方差稳定性方面都表现得很好，

#这意味着你的 ARIMA(0,1,1) 模型对于预测是合适的。

#这种检查有助于确认模型是否能够准确地捕捉数据的特征，

#并且验证了模型是否可以用于进一步的预测。

#Forecasting from ARIMA(0,1,1)model

plot.ts(yt, main="Paper Towel Daily Sales",ylab="yt")

lines(fitted(model),col="red",lty=2)

forecasting<-predict(model,n.ahead = 5)# 5 步的预测

#50% confidence interval常见的 50% 置信区间

U<-forecasting$pred+0.69\*forecasting$se

L<-forecasting$pred-0.69\*forecasting$se

#forecasting plot图形中加入了图例，标注了观测数据、预测值和置信区间

ts.plot(yt,forecasting$pred,U,L,col=c(1,2,4,4),lty=c(1,1,2,2))

legend("bottomleft",c("Observed Data","Forecast","Confidence Interval (50%)"),

col=c(1,2,4),lty=c(1,1,2))

#Time series clustering时间序列聚类

#打开week8中的sample2谢谢喵

#load("D:/RSTUDIO/R project/Week8/sample2.RData")储存位置可改，或File直接开

ts.plot(sample2)

par(mfrow=c(3,3))#绘图区域分割为 3 行 3 列

#plot.ts() 函数来逐一绘制 sample2 中的不同时间序列变量，每个变量占一个小图

plot.ts(sample2[,1]);plot.ts(sample2[,2]);plot.ts(sample2[,3])

plot.ts(sample2[,4]);plot.ts(sample2[,5]);plot.ts(sample2[,6])

plot.ts(sample2[,7]);plot.ts(sample2[,8]);plot.ts(sample2[,9])

#没有就下包喵~#determine the cluster动态时间规整

install.packages("dtw")

library(dtw)

#put a label on each time series data

#创建了标签，用于表示每个时间序列数据点所属的类别

D.Labels<-rep(1:60)

#compute distance matrix

#使用 dist() 函数计算了时间序列数据集 sample2 之间的距离矩阵。

#method = "DTW" 表示使用动态时间规整（DTW）方法计算距离

distMatrix<-dist(sample2,method="DTW")

#plot the cluster

#使用 hclust() 函数进行层次聚类，根据时间序列之间的距离构建聚类结构。

#method = "average" 表示使用平均链接法

TSCluster<-hclust(distMatrix,method = "average")

plot(TSCluster,labels=D.Labels,

main="Time Series Clustering")

#get 6 cluster将数据分为 6 类的矩形框

rect.hclust(TSCluster,k=6)

#justify your results

#cluster聚类

window(10,10)

par(mfrow=c(2,4))

plot.ts(sample2[,34]);plot.ts(sample2[,35]);plot.ts(sample2[,32])

plot.ts(sample2[,37]);plot.ts(sample2[,33]);plot.ts(sample2[,31])

plot.ts(sample2[,36])

window(10,10)

par(mfrow=c(1,3))

plot.ts(sample2[,18]);plot.ts(sample2[,16])

plot.ts(sample2[,20])

#Time series classification

install.packages("party")

library(party)

#创建了一个类别标签 classId，共有 6 个类别，每个类别包含 10 个样本

classId<-rep(as.character(1:6),each=10)

#打开Data.Ks喵谢谢喵#ctree() 函数进行分类建模，评估了分类模型的预测效果。

Class.Model<-ctree(classId~.,data=Data.Ks)

plot(Class.Model)#绘制了构建的条件推断树模型的可视化结果

#measure precison of prediction model

pClassId<-predict(Class.Model)#得到预测类别

#计算了实际类别和预测类别之间的交叉表，可以帮助评估模型的预测效果

table(classId,pClassId)

#prediction accuracy

#计算了模型的预测准确率，即预测正确的样本数量与总样本数量之比

sum(classId==pClassId)/nrow(Data.Ks)

#这些步骤帮助你构建了一个分类模型，并且评估了该模型对于给定数据的分类准确度。

#通过这些指标，你可以初步了解模型的预测效果和分类准确度。

**WEEK 10**

##mining sequence data

install.packages('TraMineR')

library(TraMineR)

#example of sequence data

data(mvad)

summary(mvad)#摘要

##define a sequence data

#define a labels and codes of codes for state

#例如:"employment"被定义为"EM"

mvad.labels<-c("employment","further education",

"higher education","joblessness",

"school","training")

mvad.scode<-c("EM","FE","HE","JL","SC","TR")

#create sequence object in R

#使用了从第17列到第86列的数据来创建序列对象，并使用之前定义的状态代码和标签。

mvad.seq<-seqdef(mvad, 17:86, states=mvad.scode,

labels=mvad.labels)

#statistical summary indicators

#mean time spent in each state

#overall dataset每种状态（就业、进修、失业等）的平均持续时间

seqmeant(mvad.seq)

#mean time spent in each state by group

#group:fathers employment status

#by()函数，这次是按照"XX"对数据进行分组，

#并计算了每个分组内每种状态的平均持续时间。

#这可以用来比较不同XX情况的年轻人在每种状态上的平均持续时间。

by(mvad.seq, mvad$funemp, seqmeant)

#group:gender

by(mvad.seq, mvad$male, seqmeant)

#number of transition状态转换的次数

nT<-seqtransn(mvad.seq)

hist(nT)#直方图

#transition rates转换率

#overall dataset

mvad.trate<-seqtrate(mvad.seq)

#if the transition rate different over time

#time varying transition rates

mvad.trate2<-seqtrate(mvad.seq, time.varying = T)

##Visualizing sequence data

#sequence index plot序列索引图

#每一条线代表一个序列，横轴表示序列的时间顺序，

#纵轴表示数据集中的序列编号。这有助于观察序列在数据集中的分布情况

seqiplot(mvad.seq, border=NA,

main="Sequence index plot")

#sequence frequency plot序列频率图

seqfplot(mvad.seq, border=NA,

main="Sequence frequency plot")

#sequence distribution plot序列分布图

#观察数据集中序列的多样性和分布情况

seqdplot(mvad.seq, border=NA,

main="Sequence distribution plot")

#model state plot模态状态图

#展示了每个时间点上最常见的状态

seqmsplot(mvad.seq, border=NA,

main="Modal state plot")

#transversal entropy横向熵图

#横向熵是一种度量序列模式多样性的指标，

#它表示在给定时间点上不同序列的多样性程度

seqHtplot(mvad.seq, main="Transversal Entropies")

#event sequence of transition

#define event:sequence of transition

mvad.seqe<-seqecreate(mvad.seq)

#look for frequent event subsequences

#seqefsub() 函数通过事件序列 mvad.seqe 找到频繁出现的事件子序列，

#参数 pmin.support = 0.05 表示最小支持度为 0.05，即出现频率至少为 5%

fsubseq<-seqefsub(mvad.seqe, pmin.support = 0.05)

#plot 15 most frequent event前15个最频繁的事件子序列

plot(fsubseq[1:15], col="blue")

##Categorizing pattern in sequence data聚类算法

library(cluster)

#compute optimal matching distance based on transition rate

#这行代码计算了基于转换率的最优匹配距离。

#seqsubm() 函数根据指定的方法（这里是"TRATE"，即基于转换率），

#计算了序列对象 mvad.seq 中每对序列之间的相似性矩阵。

#这个相似性矩阵描述了序列之间的相似程度

submat<-seqsubm(mvad.seq, method="TRATE")

#这行代码使用最优匹配距离（Optimal Matching，OM）来计算序列数据的距离矩阵。

#seqdist() 函数计算了基于最优匹配的序列之间的距离，

#并使用了之前计算的相似性矩阵 submat。这个距离矩阵描述了序列之间的差异程度。

dist.om<-seqdist(mvad.seq, method="OM", sm=submat)

#build hierarchical clustering based on

#optimal matching distance

#agnes()函数进行聚类分析。

#agnes()函数执行了AGNES（AGglomerative NESting）层次聚类算法，

#它接受一个距离矩阵 dist.om 作为输入，并使用ward方法来构建聚类。

#diss = TRUE 表示输入的是距离矩阵

clusterward<-agnes(dist.om, diss=T, method="ward")

plot(clusterward)#按<Return>键来看下一个图

#retrieve each sequence data for some cluster

#example:k=4 cluster

#cutree()将层次聚类的结果切割为指定数量的聚类， k = 4，即切割成4个聚类。

cl.4<-cutree(clusterward, k=4)

#每个聚类分配了一个标签

#用了paste()函数来生成标签，例如："type 1"、"type 2"、"type 3"、"type 4"

cl.4fac<-factor(cl.4, labels=paste("type",1:4))

seqiplot(mvad.seq,group=cl.4fac, border=NA,

main="Sequence index plot")

seqfplot(mvad.seq,group=cl.4fac, border=NA,

main="Sequence frequency plot")

seqdplot(mvad.seq,group=cl.4fac, border=NA,

main="Sequence distribution plot")

seqmsplot(mvad.seq, border=NA,group=cl.4fac,

main="Modal state plot")

seqHtplot(mvad.seq,group=cl.4fac, main="Transversal Entropies")

#Determine the most discriminating transitions

#between cluster and plot the frequencies by cluster

#seqecmpgroup()函数来确定不同聚类之间具有区分性的转换。

#它分析了之前找到的频繁事件子序列 fsubseq 在每个聚类中的出现频率差异，

#从而确定哪些转换最能区分不同的聚类。group = cl.4fac 指定了要分析的聚类变量。

discr<-seqecmpgroup(fsubseq, group=cl.4fac)

plot(discr[1:6])#频率最高的前6个

**WEEK 11**

#Mining Text Data#

install.packages("tm")

install.packages("SnowballC")

install.packages("wordcloud")

install.packages("RColorBrewer")

install.packages("syuzhet")

install.packages("ggplot2")

library(tm)#文本挖掘

library(SnowballC)

library(wordcloud)#词云

library(RColorBrewer)

library(syuzhet)

library(ggplot2)

#import the data选择text文件

text<-readLines(file.choose())

class(text)

#define a corpus text创建了一个文本语料库（corpus）

docs<-Corpus(VectorSource(text))

class(docs)

inspect(docs)#查看创建的语料库的内容

#Data cleaning of text data

#(i)remove special character from text

#define blank space

#替换为空格

toSpace<-content\_transformer(function(x,pattern)

gsub(pattern,"",x))

#convert all special character into blank space

docs2<-tm\_map(docs,toSpace,"/")

docs2<-tm\_map(docs2,toSpace,"@")

docs2<-tm\_map(docs2,toSpace,"!")

docs2<-tm\_map(docs2,toSpace,",")

docs2<-tm\_map(docs2,toSpace,":")

#(ii)Convert the text to lower case小写

docs2<-tm\_map(docs2,content\_transformer(tolower))

#(iii)remove numbers移除文本中的数字

docs2<-tm\_map(docs2,removeNumbers)

#(iv)remove the stopwords移除英文停用词

docs2<-tm\_map(docs2,removeWords,

stopwords("english"))

#(v)remove punctuation移除标点符号

docs2<-tm\_map(docs2,removePunctuation)

#(vi)eliminate extra unnecessary spaces in the text

#消除文本中额外的空格

docs2<-tm\_map(docs2,stripWhitespace)

###Word tokenization

library(tidyverse)

library(tokenizers)

#tokenize\_words函数对文本进行了词语分词化，

#将文本拆分成单词列表，保存在words.Tk变量中

words.Tk<-tokenize\_words(text)

#Text stemming提取词干（stem），即将单词转换为其词根形式，存储在docs3变量

docs3<-tm\_map(docs2,stemDocument)

inspect(docs3)#展示了处理后的文本数据

#Document-term matrix

#TermDocumentMatrix函数将经过处理的文本数据集docs3转换为文档-词项矩阵（DTM），

#将单词的出现频率映射到一个矩阵中，保存在dtm变量中

dtm<-TermDocumentMatrix(docs3)

m<-as.matrix(dtm)

m

#the frequency table of words词频表

#rowSums(m)用于计算每个单词在整个文本数据集中的总出现次数

#sort(..., decreasing = TRUE)对单词按照频率进行排序

v<-sort(rowSums(m),decreasing = T)

#word列存储单词，freq列存储对应的频率

d<-data.frame(word=names(v),freq=v)

#plot word frequency

#first 20 most frequent words最高的前20个单词的频率

#las=2参数将标签文本旋转90度

#names.arg=d[1:20,]$word设置了条形图中每个条形的标签，前20个高频单词内容

barplot(d[1:20,]$freq,las=2, names.arg=d[1:20,]$word,

main="Most frequent words")#main="Most frequent words"图表标题

###################################################

#word cloud

set.seed(12)

#words=d$word是用于词云的单词列表，freq=d$freq包含了每个单词的频率信息

#min.freq=2设置词云中单词最小出现频率，max.words=100定义了词云中最大单词量

#random.color=F禁用了随机颜色

wordcloud(words=d$word, freq=d$freq, min.freq=2,

max.words = 100,random.color=F,

color=brewer.pal(8,"Dark2"))

#Word association（findAssocs函数用于分析词语之间的关联性）

#analyze the association between frequent words

#Example:which words are associated "freedom"?

#freedom”相关联的单词，其相关性阈值为0.3

findAssocs(dtm,terms="freedom",corlimit=0.3)$freedom

#Example:association between some particular words

#展示三个单词之间的相关性

findAssocs(dtm,terms=c("freedom","dream","will"),corlimit=0.25)

#Example:find association for words that occur

#at least 10 times至少出现10次的词

findAssocs(dtm,terms=findFreqTerms(dtm,lowfreq = 10),corlimit = 0.25)

##Sentiment Analysis

library(sentimentr)

x<-'Sentiment analysis is super fun'

sentiment(x)#正面情感

y<-'Sentiment analysis is super boring. I do love work'

sentiment(y)#包含正面和负面情感的复合情感

#get\_sentiment函数则用于对整个文本进行情感分析

#method="syuzhet"来采用了"syuzhet"方法进行情感分析

sentiment\_sz<-get\_sentiment(text,method="syuzhet")

hist(sentiment\_sz)

summary(sentiment\_sz)

##Emotion classification用了NRC情感词典（NRC Sentiment Lexicon）

#get\_nrc\_sentiment函数用于基于NRC情感词典对文本进行情感分类

#列代表不同的情感类别，行代表文本中的每个词或短语

Ec<-get\_nrc\_sentiment(text)

td<-data.frame(t(Ec))#计数

td\_new<-data.frame(rowSums(td))#新数据框，每行表示一种情感类别的总计数

td\_new

##transformation & cleaning

names(td\_new)[1]<-"count"#数据框td\_new的第一列更改列名为"count"

#将数据框td\_new的行名（情感类别）添加为一列，并命名为"sentiment"

td\_new<-cbind("sentiment"=rownames(td\_new),td\_new)

rownames(td\_new)<-NULL#td\_new的行名设为NULL，将其重置

td\_new2<-td\_new[1:8,]#只显示前8个

#条形图的高度代表了每种情感类别的计数，通过不同的颜色进行区分

quickplot(sentiment, data=td\_new2,weight=count,

geom="bar",fill=sentiment,ylab="count")+

ggtitle("Survey Sentiments")

#################################################

#Mining Spatial Data

library(sp)

library(raster)

load("D:\\RSTUDIO\\R project\\week11\\wst.RData")

attach(wst)#变量直接放入当前环境

lonlat<-cbind(longitude,latitude)#数据集的经度和纬度组成一个矩阵lonlat

pts<-SpatialPoints(lonlat)#经纬度坐标点

class(pts)

plot(pts)

#define coordinate system

crdref<-CRS('+proj=longlat +datum=WSG84')#定义了坐标系

#坐标系信息应用到 SpatialPoints 对象

pts<-SpatialPoints(lonlat,proj4string = crdref)

#embed the atribut to SpatialPoints vector

df<-data.frame(ID=name,precip)#新数据框包含了名称和降水量

ptsdf<-SpatialPointsDataFrame(pts,data=df)

plot(ptsdf)

showDefault(ptsdf)

#spatial lines data这里spLines()函数用于创建空间线数据

lonlat<-cbind(longitude,latitude)

linesD<-spLines(lonlat,crs=crdref)

plot(linesD)

#spatial polygom data创建空间多边形数据

lonlat<-cbind(longitude,latitude)

polyD<-spPolygons(lonlat,crs=crdref)

plot(polyD)

#raster form data创建栅格形式的空间数据

r<-raster(ncol=10,nrow=10,xmx=-80,xmn=-150,

ymn=20,ymx=60)

area<-rnorm(ncell(r))#随机数，范围r

values(r)<-area

plot(r)

#raster stack & raster brick

#stack() 函数将这四个栅格对象组合成一个栅格堆叠（stack）。

#栅格堆叠是将多个栅格对象堆叠到一个对象中，每个栅格对象对应于堆叠中的一个层

r2<-r\*r

r3<-r^3

r4<-2\*r2+r3

s<-stack(r,r2,r3,r4)

plot(s)

#data manipulation in spatial analysis

#下面找到了 raster 包中的 lux.shp 文件的路径，矢量空间数据文件 lux.shp

f<-system.file("external/lux.shp",package = "raster")

p<-shapefile(f)

#presenting data in data frame format

d<-data.frame(p)

#extract spesific atributes提取空间数据对象 p 中的特定属性

p$NAME\_1

p$AREA

#add new variable

set.seed(12)

xp<-sample(letters,length(p))

p$new<-xp#生成的随机字母序列 xp 添加为 p 对象的一个新变量 new

data.frame(p)#创建了一个数据框，显示了经过变量添加和删除后的空间数据对象 p

#delete any variable删除了 p 对象中的 new 变量

p$new<-NULL

#data integration

dfr<-data.frame(District=p$NAME\_1,Canton=p$NAME\_2,

Value=runif(length(p)))

cD<-merge(p,dfr,by.x=c('NAME\_1','NAME\_2'),

by.y=c('District','Canton'))

#data aggregation绘制地图

pa<-aggregate(p,by='NAME\_2')#aggregate() 函数按NAME\_2 属性进行聚合

plot(p)

plot(pa,add=T,col=rainbow(12))

#map manipulation

**WEEK 12**

##Minig Spatial Data##

#空间数据的简单处理

library(sp)

library(raster)

#raster data manipulation

#其中参数ncol和nrow指定了栅格的列数和行数，

#而xmx、xmn、ymn和ymx则定义了栅格的空间范围，分别表示最大和最小的经度和纬度

r<-raster(ncol=10,nrow=10,xmx=-80,xmn=-150,

ymn=20,ymx=60)

income<-rnorm(ncell(r))#生成一个与栅格单元数相同的随机数向量

values(r)<-income#将之前生成的收入数据向量income赋值给raster对象r的值

plot(r)

#algebra in raster data

r2<-r+10

r3<-r^3

r4<-2\*r2+r3

s<-stack(r,r2,r3,r4)#四个图层的栅格数据的堆叠

plot(s)

#extract raster layer from raster stack/brick

rL3<-raster(s,layer=3)#提取第三个图层（layer=3），并将其存储在 rL3 中

plot(rL3)

#add new value in the raster data

#生成一个随机数向量 debt，其中包含了长度与 r 栅格对象的栅格单元数相同的数值

#这些数值是通过 rgamma 函数生成的伽马分布的随机数

#并且每个数值的相反数被计算（(-1) \* rgamma(...)）

debt<-(-1)\*rgamma(length(r),20,30)

r$debt<-debt#将新图层添加

plot(r)#添加前以及添加后的两图

#crop raster data裁剪栅格数据，即限制数据集的范围以仅包含感兴趣的部分

#横坐标（经度）范围在 -150 到 -120，纵坐标（纬度）范围在 20 到 40 之间的部分

r1<-crop(r,extent(-150,-120,20,40))

plot(r1)

r2<-crop(r,extent(-120,-110,40,60))#这个和上面同理

plot(r2)

#merge raster data

#将两个裁剪后的栅格数据对象 r1 和 r2 合并成一个新的栅格数据对象 m

m<-merge(r1,r2)

plot(m)

#Descriptive statistics

#cellStats 函数用于计算栅格数据的统计量

#第一个参数是栅格对象

#第二个参数是要计算的统计量的函数

#mean均值

cellStats(r,mean)

#standard deviation标准差

cellStats(r,sd)

#median中位数

cellStats(r,median)

#Spatial Autocorrelation空间自相关

#shapefile 函数读取一个 Luxemburg 行政区划的形状文件（shapefile）

p<-shapefile(system.file("external/lux.shp",

package = "raster"))

plot(p)

data.frame(p)#对象 p 转化为数据框，以便查看其中的数据

#创建一个新的对象 p1，该对象包含原始数据中行政区划名称为 "Diekirch" 的部分

p1<-p[p$NAME\_1=="Diekirch",]

plot(p1)

p1$value<-c(10,6,4,11,6)#向选定的 "Diekirch" 区域添加一个名为 "value" 的新变量

#没有就install.packages("spdep")

library(spdep)

#poly2nb 函数创建一个空间邻接矩阵（spatial weights matrix）

#这个邻接矩阵描述了地理空间上的相邻关系

w<-poly2nb(p1,row.names = p1$ID)

#nb2listw 函数将邻接矩阵转换为 "listw" 对象

#这是进行空间统计分析时常用的一种表示方式

ww<-nb2listw(w)

#对选定区域 "Diekirch" 中的 "value" 变量执行莫兰指数（Moran's I）检验

#ww 是之前创建的空间权重矩阵，randomisation = FALSE 表示不进行随机化检验

moran.test(p1$value,ww,randomisation = F)

#莫兰指数是一种用于测量地理空间中数据的空间相关性的统计指标

#检验的假设是数据是否在空间上呈现显著的自相关性

#如果 p 值小于显著性水平，就可以拒绝空间自相关性的零假设，表示存在空间自相关性。

#由于 p 值小于显著性水平（0.008524 < 0.05），我们可以拒绝零假设

#得出结论：在 "Diekirch" 区域中的 "value" 变量具有正的空间自相关性

#表明在该地区，相邻地理单元之间的 "value" 值更可能相互关联，而不是随机分布

#Spatial Interpolation空间插值

#install.packages("devtools")

library(devtools)

#install\_github("rspatial/rspat")

library(rspat)

#spat\_data 函数加载一个名为 'precipitation' 的空间数据集

#这个数据集可能包含有关降水量的信息，可以在空间上进行插值

d<-spat\_data('precipitation')

#plot a data

#指定经度（LONG）和纬度（LAT）作为空间坐标

#设置坐标参考系（CRS）。在这里，CRS 被设置为 WGS84

dsp<-vect(d,c("LONG","LAT"),

crs="+proj=longlat+datum=WGS84")

#加载名为 "counties" 的示例空间数据集

#plot(CA) 用于绘制美国的县级边界

CA<-spat\_data("counties")

plot(CA)

#determine the cumulative precipitation

#这段代码用于计算累积降水量，并将结果添加到空间数据对象 d 中

mnts<-toupper(month.abb)#创建一个包含英文缩写月份的字符向量，将其转换为大写形式

d$prec<-rowSums(d[,mnts])#创建了一个新的列，包含了每个点的累积降水量（月）

#老样子设置坐标参考系（CRS）

dsp<-vect(d,c("LONG","LAT"),

crs="+proj=longlaat+datum=WGS84")

#define some threshold

#数值将被用作累积降水量的分割点，将其分成不同的类别

#一个路过的吸血鬼说：“你刚刚用了cuts了，对吧？！”

cuts<-c(0,200,300,500,1000,3000)

#plot the data based on threshold

#颜色渐变将用于表示不同累积降水量类别

blues<-colorRampPalette(c('yellow','orange','blue','dark blue'))

#函数绘制美国县级边界，border='dark gray' 设置边界的颜色为深灰色

plot(CA,border='dark gray')

#参数 "prec" 表示使用 dsp 对象中的 "prec" 列

#breaks 参数指定了阈值，将累积降水量分成不同的类别：“用了cuts，对吧？！”

plot(dsp,"prec",breaks=cuts, col=blues(10))

#Spatial interpolation

#proximity polygons

windows(10,10)#10x10 的绘图窗口

#生成 dsp 中点的 Voronoi 多边形

#Voronoi 多边形是由空间中离散点到最近邻的距离确定的多边形。

v<-voronoi(dsp)

plot(v)

#points 函数将原始的 dsp 数据点绘制到 Voronoi 图形上

points(dsp)

#crop 函数将 Voronoi 多边形限制在美国的县级边界范围内

vca<-crop(v,CA)

#绘制裁剪后的 Voronoi 多边形，其中颜色表示预测的累积降水量

#我对吸血鬼说：“喂，你刚才又用了prec是吧？”

plot(vca,"prec")

**WEEK 13**

##local spatial regression

houses<-read.csv(file.choose(),header = T,sep=";")

head(houses)

#convert to spatial data

library(rspat)

#vect 函数将数据框 houses 转换为 SpatialPointsDataFrame 对象 hvect

hvect<-vect(houses,c("longitude","latitude"))#指定名字为经纬的列

hvect

plot(hvect,cex=0.5,pch=1)#cex = 0.5 控制绘图中点的大小，pch = 1 设置点的形状

###ordinary regression

hd<-houses

#glm 函数拟合一个普通的线性回归模型

#定义了回归方程，其中 houseValue 是因变量（看那个“~”标），其他是自变量

m<-glm(houseValue~income+houseAge+roomhead+bedroomhead+population,data=hd)

summary(m)

#显著性 (Significance):\*\*\* 表示非常显著，< 2e-16 表示极其显著

#离散参数 (Dispersion parameter):指定为高斯分布的离散参数

#迭代次数 (Number of Fisher Scoring iterations)

#这个模型中，所有的系数估计都是显著的(看summary的结果)

#local regression绘制县级边界(和上周的差不多)

counties<-spat\_data("counties")

crs(hvect)<-crs(counties)

plot(counties)

#从数据框 hd 中选择那些属于特定县（通过 NAME 列匹配）的观测数据

#线性回归模型，其中 houseValue 是因变量

#coefficients(m): 返回线性回归模型的系数估计值

regfun<-function(x){

dat<-hd[hd$NAME==x,]

m<-glm(houseValue~income+houseAge+roomhead+

bedroomhead+population,data=dat)

coefficients(m)}

#all region names

#新数据框包含所有不是缺失值的

hd2<-hd[!is.na(hd$NAME),]

#提取数据中所有唯一的县名

countrynames<-unique(hd2$NAME)

#每一列对应一个县，每一行对应一个模型系数

res<-sapply(countrynames,regfun)

#convert the data into data frame

resdf<-data.frame(NAME=colnames(res),t(res))

#visualize the results

#aggregate 函数，对县级边界数据的 "NAME" 列进行聚合，以得到唯一的县名

dcounties<-aggregate(counties[,"NAME"],"NAME")

#merge 函数，将县级边界数据和模型系数数据按照 "NAME" 列进行合并

cnres<-merge(dcounties,resdf,by="NAME")

plot(cnres,"income")

cnres2<-cnres

#对 cnres2 中除 "NAME" 列之外的所有列进行标准化（z-score 标准化）

values(cnres2)<-

as.data.frame(scale(as.data.frame(cnres)[,-1]))

#1:6 表示绘制除 "NAME" 列外的所有列(在前面可以head(resdf)看看)

plot(cnres2,1:6,plg=list(x="topright"),mar=c(1,1,1,1))

###Mining Graph Data

#install.packages("igraph")

library(igraph)

#undirected graph无向图

#例如，1-2 表示连接节点 1 和节点 2 的一条边

g<-graph\_from\_literal(1-2,1-3,1-7,3-4,2-3,2-4,3-5,4-5,4-6,4-7,5-6,5-8,6-7,7-8)

#input of nodes

#为图的节点添加标签（这里是8个人对应8个节点）

V(g)$name<-c("Adam","Judy","Bobby","Sam","Frank",

"Jay","Tom","Jerry")

plot(g)

#directed graph有向图，表示三个机场之间的航班方向关系

#three airport:NY,BJ,PR

#有向边通过 -> 和 + 符号来定义，例如，NY-+BJ 表示从纽约（NY）到北京（BJ）的有向边

#那个++就是双向，-+单向，可以搞三角恋了

dg<-graph\_from\_literal(NY-+BJ,BJ-+PR,PR++NY)

plot(dg)

#weighted graph带权重的无向图

#创建一个数据框，其中包含了一个权重矩阵

m<-read.table(row.names = 1,header = T,text =

"A B C D E

A 0 2 0 5 1.3

B 1 8 1 0 -4

C 0 0 0 0 3

D 5 2 1 0 0

E 1 1 1 0 3")

m<-as.matrix(m)

#graph.adjacency 函数创建一个带权重的无向图 ig

#参数 mode = "undirected" 表示创建的图是无向图

#weighted = TRUE 表示图中的边带有权重

ig<-graph.adjacency(m,mode = "undirected",weighted = T)

plot(ig,edge.label=E(ig)$weight)

#tree graph

#参数 40 指定图中的节点数量，children = 4 指定每个节点最多有 4 个孩子节点

#mode = "undirected" 表示图是无向的

tr<-make\_tree(40,children=4,mode="undirected")

plot(tr,vertex.size=10)

#Bipartite graph二分图

set.seed(23)#设置随机数种子

#10: 第一部分的节点数量，5: 第二部分的节点数量，p = 0.4: 边存在的概率

gb<-sample\_bipartite(10,5,p=0.4)

plot(gb)

#define some color & shape为二分图的节点定义了颜色和形状

col<-c("blue","red")

shape<-c("circle","square")

#vertex.color 和 vertex.shape 参数指定节点的颜色和形状

#V(gb)$type 表示每个节点所属的部分

#as.numeric(V(gb)$type) + 1 将部分标识转换为颜色和形状的索引

plot(gb,vertex.color=col[as.numeric(V(gb)$type)+1],

vertex.shape=shape[as.numeric(V(gb)$type)+1])

#hypergraph超图

#install.packages("HyperG")

library(HyperG)

#函数的参数是一个包含多个集合的列表，每个集合表示超图中的一个超边

h<-hypergraph\_from\_edgelist(list(1:2,2:5,3:7,

c(1,3,5,7,9)))

plot(h)

#representation of graph

#get.adjlist 函数获取了不同类型图的邻接列表

#Adjacency lists

Adj.list<-get.adjlist(g)#无向图

Adj.list2<-get.adjlist(dg)#有向图

Adj.list3<-get.adjlist(ig)#带权重的无向图

Adj.list4<-get.adjlist(tr)#树状图

Adj.list5<-get.adjlist(gb)#二分图

#Edge list

#get.edgelist 函数获取了不同类型图的边列表

Ed.list<-as.data.frame(get.edgelist(g))

Ed.list2<-as.data.frame(get.edgelist(dg))

Ed.list3<-as.data.frame(get.edgelist(ig))

Ed.list4<-as.data.frame(get.edgelist(tr))

#Adjacency mmatrix获取邻接矩阵

Adj.M<-get.adjacency(g)

Adj.M2<-get.adjacency(dg)

Adj.M3<-get.adjacency(ig)

#Graph manipulation

plot(g)

#从图 g 中移除节点 "Adam" 和 "Judy"，创建了新的图

h1<-g-vertices(c("Adam","Judy"))

#windows(10,10)可加可不加

plot(h1)

#generate subgraph by manual input

#通过 graph\_from\_literal 函数手动输入了子图的边，创建了一个子图 h2

h2<-graph\_from\_literal("Adam"-"Farah","Adam"-"Lee",

"Adam"-"Bobby","Adam"-"Sam")

plot(h2)

#join graph

#union 函数将两个图 h1 和 h2 进行了合并

h3<-union(h1,h2)

plot(h3)

#modify nodes data

#对图 h3 的节点进行了属性修改

V(h3)#返回图 h3 中的节点列表

#用向量指定了每个节点的性别

V(h3)$gender<-c("male","male","male","female",

"male","male","male","female",

"male")

vertex\_attr(h3)#显示图 h3 中节点的属性信息

#modify edges data

#对图 h3 的边进行了属性修改

E(h3)$type<-c("email","phone","facebook","twitter",

"facebook","email","class","phone",

"twitter","email","facebook","twitter",

"facebook")#用向量指定了每条边的类型

E(h3)$weight<-c(10,1,3,4,12,4,4,7,5,10,11,3,2)#设置新的权重属性

edge\_attr(h3)#显示图 h3 中边的属性信息

plot(h3,edge.label=E(h3)$weight)

#node prominence analysis

#install.packages("statnet")

#install\_github("DougLuke/UserNetR")

library(statnet)

library(devtools)

library(UserNetR)

data(Bali)

plot(Bali)

#information about nodes

#获取节点属性中名为 "vertex.names" 的信息

Bali%v%"vertex.names"

#获取节点属性中名为 "role" 的信息

Bali%v%"role"

#information about edges

Bali%e%"IC"

#degree

node<-Bali%v%"vertex.names"

#计算 Bali 数据集中每个节点的度

deg<-degree(Bali)

#数据框 npa，其中包含节点的名称和对应的度

npa<-data.frame(node,deg)

#closeness measure

#closeness 函数计算 Bali 数据集中每个节点的紧密度

cls<-closeness(Bali)

#将之前创建的数据框 npa 与紧密度 cls 合并

npa2<-data.frame(npa,cls)

npa2

#betweeness centrality

#计算 Bali 数据集中每个节点的介数中心性

btw<-betweenness(Bali)

npa3<-data.frame(npa2,btw)

#eigen vector centrality

#计算了 Bali 数据集中节点的特征向量中心性（eigenvector centrality）

evc<-evcent(Bali)

#information centrality score

#计算了 Bali 数据集中节点的信息中心性分数

inf<-infocent(Bali)

npa4<-data.frame(npa3,evc,inf)

#cutpoints

#计算了 Bali 数据集中的割点

net<-Bali

#mode = "graph" 表示计算整个图的割点

#return.indicator = TRUE 表示返回一个逻辑向量，指示每个节点是否为割点

#可以通过打印 cpnet 或直接查看其内容来查看每个节点是否为割点

cpnet<-cutpoints(net,mode="graph",return.indicator = T)

#subgroup analysis

#clique

#该图是一个包含多个团（clique）的例子

library(igraph)

#团是一个图中的完全子图，即其中的每一对节点都相互连接

clqexmp<-graph.formula(A:B:C:D--A:B:C:D,D-E,E-F-G-E)

plot(clqexmp)

#cliques 函数从图 clqexmp 中找到包含至少3个节点的所有团

cliques(clqexmp,min=3)

#large number of clique极大团是在图中不能再添加任何一个节点而保持团的集合

maximal.cliques(clqexmp,min=3)#min = 3 表示只返回包含至少3个节点的极大团

#Bali dataset

#install.packages("intergraph")

library(intergraph)

#将 Bali 数据集转换为 igraph 对象 Bali2

Bali2<-asIgraph(Bali)

#找到包含至少8个节点的所有团

cliques(Bali2,min = 8)

#large number of clique

#maximal.cliques 函数从图 Bali2 中找到所有的极大团，每个团至少包含8个节点

maximal.cliques(Bali2,min=8)

#k-core

#转换 DHHS 数据集为 igraph 对象

data(DHHS)

iDHHS<-asIgraph(DHHS)

plot(iDHHS)

#计算每个节点的核心度

core<-graph.coreness(iDHHS)

table(core)

#modularity

data(Facebook)

plot(Facebook)

#通过 V(Facebook)$group 获取了网络中节点的分组信息

#使用 levels(factor(V(Facebook)$group)) 查看分组的水平

levels(factor(V(Facebook)$group))

#转换分组信息为数值型

grp\_num<-as.numeric(factor(V(Facebook)$group))

#计算网络的模块性

modularity(Facebook,grp\_num)

#community detection

data(Moreno)

plot(Moreno)

#asIgraph 函数将 Moreno 数据集转换为 igraph 对象

iMoreno<-asIgraph(Moreno)

#cluster\_walktrap 函数进行社团检测，并绘制检测到的社团结构

cw<-cluster\_walktrap(iMoreno)

plot(cw,iMoreno)