**Homework-09**

问题：

在doubs河流鱼群研究中，请筛选VOLPla站点有关CHE鱼类生物量和密度的记录，构建表3。其中，第1列为序号，第2列为时间戳（stamp），第3列为生物量（Biomass）。请根据下面问题作答。

|  |  |  |
| --- | --- | --- |
| no | Date | Biomass |
| 1854 | 1994-06-21 | 0.9500713 |
| 1855 | 1995-06-13 | 0.9519654 |
| 1856 | 1996-06-18 | 8.3632565 |
| ︙ | ︙ | ︙ |

1. library(tidyverse) #加载数据处理包 表3 doubs河1994-2020年CHE鱼类个体生物量
2. library(randomForest)#加载随机森林算法
3. data <- read.table("fishBiomassData.txt ")#读取生物量数据
4. mydata <- data |>
5. subset(STATION=="VOLPla" & SP == "CHE") |>

# 筛选站点为VOLPla、鱼种为CHE的记录

1. select(DATE, BIOMASS)

# 仅保留日期（DATE）和生物量（BIOMASS）列

1. ts <- ts(mydata$BIOMASS, start=1994, frequency=1

# 创建时间序列对象，起始年份1994，每年1个观测，实现将生物量数据转换为时间序列（ts）格式

1. ts\_org <- window(ts, end = 2018)

# 提取1994-2018作为训练集，用于模型训练

1. ts\_trf <- ts.org |> log() |> diff(1)

#训练集对数变换+一阶差分，模型更稳定

1. lag\_order <- 2 #前2年的数据作为特征
2. horizon <- 2 #预测未来2年

1. ts\_mbd <- embed(ts.trd, lag.order + 1)

#embed嵌套：形成时间序列的矩阵（其中每一行包含当前值及其过去两年的值）

1. Y\_train <- ts\_mbd[, 1]
2. X\_train <- ts\_mbd[, -1]

# Y\_train：矩阵的第一列，代表当前值（要预测的值）

# X\_train：矩阵的其余列，代表滞后值（用于预测的特征）

1. y\_test <- window(ts, start = 2019, end = 2020)

#获取2019和2020年的实际观测值用于验证预测效果

1. x\_test <- ts\_mbd[nrow(ts\_mbd), c(1:lag\_order)]

#提取最新的滞后变量，作为初始预测输入。

1. pred\_rf <- numeric(horizon)
2. for (i in 1:horizon){set.seed(1)
3. fit\_rf <- randomForest(X\_train, y\_train)
4. pred\_rf[i] <- predict(fit\_rf, X\_test)
5. y\_train <- y\_train[-1]
6. X\_train <- X\_train[-nrow(X\_train), ] }
7. pred\_rf

#使用随机森林模型滚动预测未来2年。每年训练一次模型，用上一次预测值作为下次预测的输入。

1. exp\_term <- exp(cumsum(pred\_rf))
2. last\_obs <- as.vector(tail(ts\_org, 1))
3. backtrans\_fc <- last\_obs \* exp\_term

#从差分对数空间还原为原始生物量

1. y\_pred <- ts(backtrans\_fc, start = 2019, frequency = 1)

#将预测结果重新转换为时间序列对象

1. library(fpp2)
2. forecast::accuracy(as.numeric(y\_pred), as.numeric(y\_test))

#accuracy() 评估预测值与真实值之间的误差

1. ts\_fc <- cbind(ts,pred = c(rep(NA, length(ts\_org)), y\_pred))
2. plot\_fc <- ts\_fc |> autoplot() + theme\_minimal()
3. plot\_fc

#将原始序列与预测值合并，并用plot() 画图

1) 这是构建机器学习时间序列预测模型，根据代码，请简述时间序列预测建模基本原理以及基本过程。

**一．基本原理：时间序列预测试剂上就是利用回归模型，基于历史数据来预测未来。**由于时间序列数据一般具有时间上的相关性，所以tn+1时刻的值可能很接菌tn时刻的值，即可拆分前后两个时间的观察值，并分别看作特征值和目标变量，由此构成一个训练样本。例如一个等时间间隔取样的时序t（13个时间点）要预测下一个时刻t14，将每组中第一个数据作为特征，第二个数据作为目标变量，构成12 个样本。利用1-11个样本模型进行训练，第12个样本用于预测，就可以预测t14。

**二．基本过程：**

|  |  |  |  |
| --- | --- | --- | --- |
| 步骤 | 目的 | 对应代码 | 说明 |
| 数据准备 | 读取和筛选感兴趣的站点和鱼种数据 | [1]-[6] | 仅保留VOLPla站点、CHE鱼种，选出日期和生物量 |
| 构建时间序列 | 将数据转换为时间序列格式 | [7] | 频率设为1，表示年度数据 |
| 构造训练集 | 获取1994–2018年训练数据 | [8] | 用于训练模型 |
| 数据变换 | 对数变换 + 一阶差分，稳定序列 | [9] | 稳定时间序列结构（使其平稳） |
| 特征工程 | 构建滞后特征矩阵 | [10]-[14] | 形成 X\_train（滞后值） 和 Y\_train（当前值） |
| 构造测试集 | 准备真实测试数据 | [15]-[16] | 真实值用于验证预测效果，x\_test作为初始预测输入 |
| 滚动预测 | 用随机森林迭代预测未来 | [17]-[23] | 每预测一年，将其作为下年输入 |
| 反变换还原 | 将预测结果还原为原始单位 | [24]-[26] | 从对数差分恢复为生物量数值 |
| 构建预测序列 | 构建时间序列格式的预测结果 | [27] | 将预测结果重新转换为时间序列对象 |
| 模型评估和可视化 | 评估精度，画出预测图 | [28]-[31] | 使用RMSE等指标，绘制预测图 |

2）除了基于原始值（生物量）构造滞后特征外，还可从时间戳提取并构造建模特征，请简述从时间戳构造建模特征的依据是什么？

从时间戳获取时间序列的预测特征是利用序列本身数据构建特征的方法。

归纳起来，有以下三种：

1. **日期时间特征：**观察值与具体日期或季节相关，如日低温和雨季等有重要关系，可以选择把日期或季节作为特征
2. **滞后特征：**如今年的11月份数据与往年份的11月份数据更相关，即需要关注历史上的今天
3. **窗口特征：**建模精度不仅与采用或选择的滑动窗口大小密切相关，而且与窗口内的均值，最大值等具体数据也有关。可以选取窗口中的和，最小值，最大值等作为特征，建立机器学习模型。

数据下载网址：

https://figshare.com/articles/dataset/Data\_for\_Contemporary\_loss\_of\_genetic\_diversity\_in\_wild\_fish\_populations\_reduces\_biomass\_stability\_over\_time\_/13095380?file=39686458

或者从GitHub下载[fishBiomassData.txt](https://github.com/flliu315/homework2025/blob/main/fishBiomassData.txt" \o "fishBiomassData.txt)：

https://github.com/flliu315/homework2025

要求：

一周内上传至各自GitHub托管的homework中。