**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) 表3 doubs河1994-2020年CHE鱼类个体生物量
2. library(randomForest)
3. data <- read.table("fishBiomassData.txt ")
4. mydata <- data |>
5. subset(STATION=="VOLPla" & SP == "CHE") |>
6. select(DATE, BIOMASS)
7. ts <- ts(mydata$BIOMASS, start=1994, frequency=1)
8. ts\_org <- window(ts, end = 2018)
9. ts\_trf <- ts.org |> log() |> diff(1)
10. lag\_order <- 2
11. horizon <- 2
12. ts\_mbd <- embed(ts.trd, lag.order + 1)
13. Y\_train <- ts\_mbd[, 1]
14. X\_train <- ts\_mbd[, -1]
15. y\_test <- window(ts, start = 2019, end = 2020)
16. x\_test <- ts\_mbd[nrow(ts\_mbd), c(1:lag\_order)]
17. pred\_rf <- numeric(horizon)
18. for (i in 1:horizon){set.seed(1)
19. fit\_rf <- randomForest(X\_train, y\_train)
20. pred\_rf[i] <- predict(fit\_rf, X\_test)
21. y\_train <- y\_train[-1]
22. X\_train <- X\_train[-nrow(X\_train), ] }
23. pred\_rf
24. exp\_term <- exp(cumsum(pred\_rf))
25. last\_obs <- as.vector(tail(ts\_org, 1))
26. backtrans\_fc <- last\_obs \* exp\_term
27. y\_pred <- ts(backtrans\_fc, start = 2019, frequency = 1)
28. library(fpp2)
29. forecast::accuracy(as.numeric(y\_pred), as.numeric(y\_test))
30. ts\_fc <- cbind(ts,pred = c(rep(NA, length(ts\_org)), y\_pred))
31. plot\_fc <- ts\_fc |> autoplot() + theme\_minimal()
32. plot\_fc

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。

基本原理：时间预测实际上就是利用回归模型，基于历史数据来预测未来。时间序列预测旨在利用历史数据中蕴含的时间依赖性（如趋势、季节模式、周期性）和 自相关性，通过传统统计模型（指数平滑法、自回归移动平均模型）或机器学习（随机森林、随机森林等）模型预测未来值。

基本过程（基于代码流程）

1.数据预处理：

* 筛选数据：提取特定站点（STATION=="VOLPla"）和物种（SP=="CHE"）的记录（代码第4-6行）。
* 时间序列转换：将生物量数据转换为时间序列对象 ts，起始年为1994，频率为年数据（代码第7行）。

[4] mydata <- data |>

[5] subset(STATION=="VOLPla" & SP == "CHE") |> # 筛选站点为VOLPla、物种为CHE的记录

[6] select(DATE, BIOMASS) # 仅保留日期和生物量列

[7] ts <- ts(mydata$BIOMASS, start=1994, frequency=1) # 转换为时间序列（年度数据）

2.数据划分与变换：

* 训练集与测试集：以2018年为界，划分训练集 ts\_org 和测试集 y\_test（代码第8、15行）。
* 平稳化处理：对训练集进行对数转换和差分（log() |> diff(1)），消除非平稳性（代码第9行）。

[8] ts\_org <- window(ts, end = 2018) # 训练集（1994-2018）

[9] ts\_trf <- ts\_org |> log() |> diff(1) # 对训练集取对数后一阶差分

3.特征工程：

* 滞后特征构造：通过 embed() 生成滞后矩阵（代码第12行），例如 lag\_order=2 表示用前两年的生物量作为预测特征（代码第10-12行）。

[10] lag\_order <- 2 # 滞后阶数（用前2年预测下一年）

[11] horizon <- 2 # 预测未来2年（2019-2020）

[12] ts\_mbd <- embed(ts\_trd, lag\_order + 1) # 生成滞后矩阵

4.模型训练与预测：

* 随机森林建模：以滞后特征为输入，生物量为输出，训练模型（代码第19行）。
* 滚动预测：在循环中逐步预测未来两年（2019-2020），并动态更新训练集（代码第18-22行）。

[13] Y\_train <- ts\_mbd[, 1] # 目标变量（当前值）

[14] X\_train <- ts\_mbd[, -1] # 特征矩阵（滞后1和滞后2）

[15] y\_test <- window(ts, start = 2019, end = 2020) # 测试集真实值（2019-2020）

[16] x\_test <- ts\_mbd[nrow(ts\_mbd), c(1:lag\_order)] # 测试集特征（最后一行滞后值）

[17] pred\_rf <- numeric(horizon) # 初始化预测结果存储

[18] for (i in 1:horizon){

[19] set.seed(1) # 固定随机种子

[20] fit\_rf <- randomForest(X\_train, y\_train) # 训练模型

[21] pred\_rf[i] <- predict(fit\_rf, x\_test) # 预测未来值

[22] y\_train <- y\_train[-1] # 删除最早的历史值

[23] X\_train <- X\_train[-nrow(X\_train), ] # 删除对应的特征行

5.结果反变换与评估：

* 指数反变换：将对数差分预测值还原为原始尺度（代码第24-26行）。
* 准确性评估：计算预测值与真实值的误差指标（如RMSE、MAE）（代码第29行）。
* 可视化：绘制原始序列与预测值的对比图（代码第30-32行）。

[24] exp\_term <- exp(cumsum(pred\_rf)) # 将对数差分预测值还原为原始尺度

[25] last\_obs <- as.vector(tail(ts\_org, 1)) # 获取2018年的原始观测值

[26] backtrans\_fc <- last\_obs \* exp\_term # 计算最终预测值

[27] y\_pred <- ts(backtrans\_fc, start=2019, frequency=1) # 转换为时间序列

[28] library(fpp2) # 加载时间序列可视化包

[29] forecast::accuracy(as.numeric(y\_pred), as.numeric(y\_test)) # 计算预测误差

[30] ts\_fc <- cbind(ts, pred = c(rep(NA, length(ts\_org)), y\_pred)) # 合并原始序列与预测值

[31] plot\_fc <- ts\_fc |> autoplot() + theme\_minimal() # 绘图

[32] plot\_fc

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？

（1）日期时间特征（date time features）。即观察值与具体日期或季节有关，如日低温与雨季等有重要关系，可以选择把日期或季节作为特征。

（2）滞后特征（lag features）。如今年的11月份数据与以往年份的11月份数据更相关，即要关注历史上的今天。

（3）窗口特征（window features）。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。

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)：

https://github.com/flliu315/homework2025

要求：

一周内上传至各自GitHub托管的homework中。