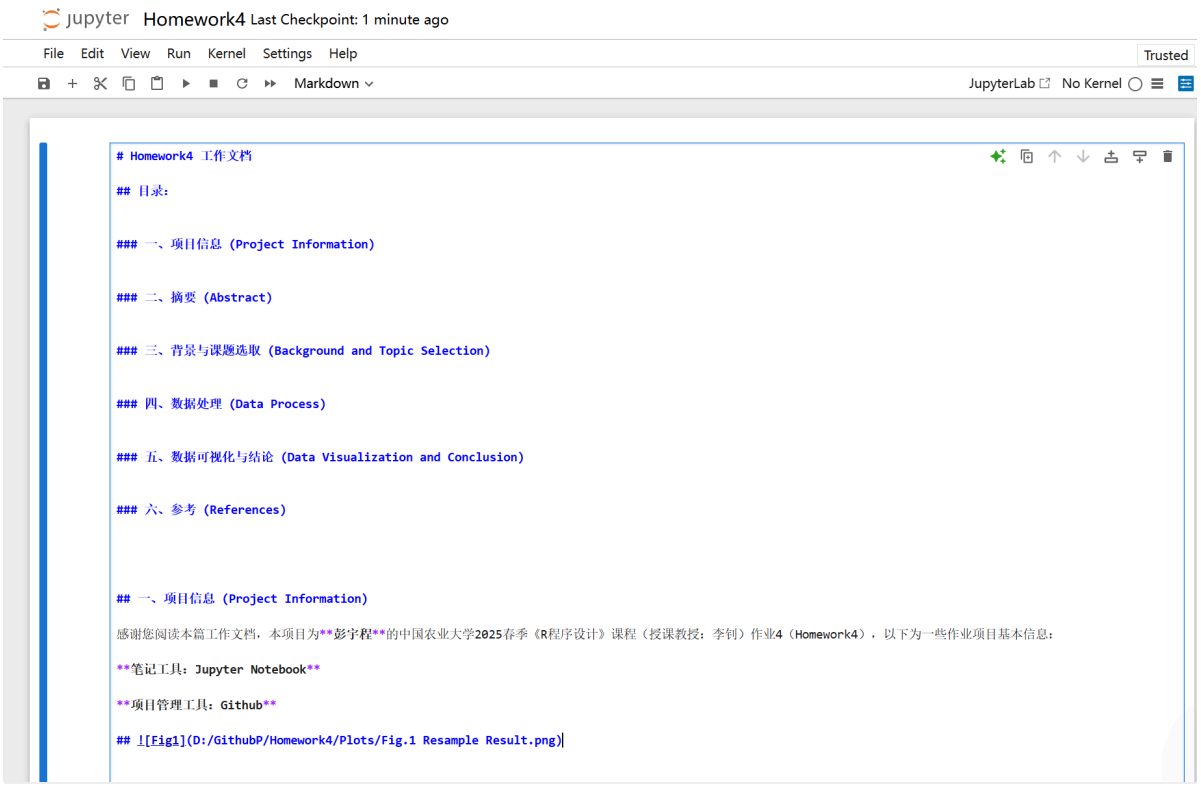
Homework4 工作文档

目录：

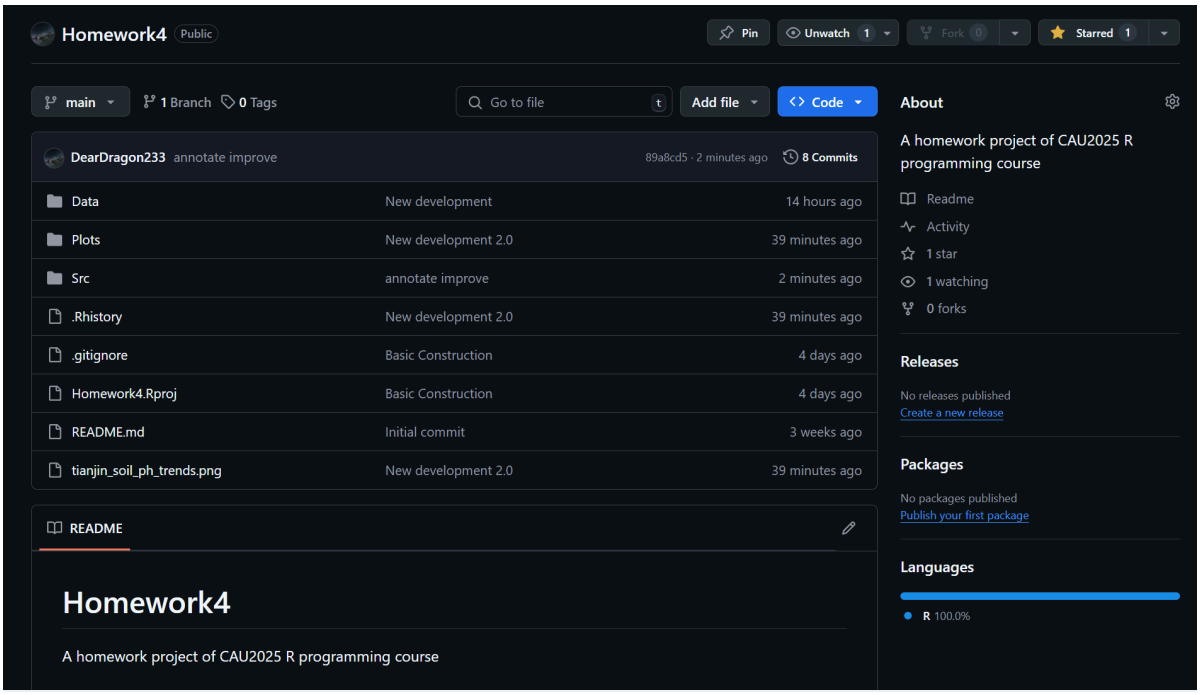
一、项目信息 (Project Information)

感谢您阅读本篇工作文档，本项目为**彭宇程**的中国农业大学2025春季《R程序设计》课程（授课教授：李钊）作业4（Homework4），以下为一些作业项目基本信息：

笔记工具：Jupyter Notebook



项目管理工具：Github URL：https://github.com/DearDragon233/Homework4



额外数据：National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://www.geodata.cn):

1. 天津市250米分辨率土壤有机碳密度数据集
2. 天津市250米分辨率土壤黏粒含量数据集
3. 天津市1：100万土壤类型图（2018年）
4. 天津市250米分辨率土壤酸碱度数据集
5. 天津市250米分辨率土壤全氮含量数据集

包含并行计算、程序嵌套，见数据处理第二部分

二、摘要 (Abstract)

本文聚焦天津市土壤状况，基于土壤黏粒含量、有机碳密度、全氮含量及酸碱度等因子展开研究。通过随机森林方法构建机器学习模型，剖析各因子与土壤类型关联。经数据处理、可视化及分析，明确土壤 pH 值在经纬度方向上的分布特征，获得了土壤因子重要性排序。

三、背景与课题选取 (Background and Topic Selection)

通过作业3(Homework3)，我们学习和实践了许多生态学和生态地理等领域的基础数据分析流程，我个人学习到了如Markdown语法、Github基础使用等关键的基础技能，在上次作业中，我们主要研究了Topic4下的部分课题，Topic4所提供的主要数据为世界植被覆盖类型（地形）分布的栅格数据，我们最后停留在世界范围内的耕地转化潜力(Potential)预测模型搭建的初始阶段。

在Homework4中，我开始将目光转向于更小地理范围下的课题研究。在浏览国家地球系统科学数据中心时，我偶然看到了天津市的土壤数据，这使我联想到前几年天津市生态治理中显露的治理问题，如2017年，**天津市宁河区内的七里海湿地曾因“在核心区和缓冲区违法建设湿地公园”而在第一轮中央环保督察中被“点名”、东丽区金钟街道欢坨村填埋有毒有害垃圾130多亩，6到8米深。**除此之外，作为天津市生人，我一直对家乡的生态情况保持着关注，希望能够对其进行深一步了解。

由此，我通过拓展土壤数据，基于**土壤粘粒含量、土壤有机碳密度、土壤全氮含量、土壤酸碱度**这四个土壤因子进行研究与拓展，我希望能够通过这四个土壤因子研究其与天津市地理范围内土壤类型的关联，探究各个土类亚类与这四个因子的关联程度，基于这个目标，我选择了随机森林方法以实现机器学习。

以下我引用了专著或论文中对这四个土壤因子的基本概念介绍：

土壤黏粒（Clay）是指粒径小于 2 μm 的土壤颗粒，是土壤物理化学性状的关键组成部分。它们具有较大的比表面积和高阳离子交换容量（CEC），在水分保持、养分供应、结构稳定性及土壤肥力中起着核心作用（Brady & Weil, 2008）。

土壤有机碳（Soil Organic Carbon, SOC 或 OCD）是土壤有机质的主要组成部分，对维持土壤肥力、改善结构、促进微生物活动及碳循环具有至关重要的作用。它不仅影响土壤的物理和化学特性，还在全球碳平衡中扮演着关键角色（Lal, 2004）。

氮是植物必需的主要营养元素之一，在土壤中的存在形式包括无机氮（硝态氮、铵态氮）和有机氮。土壤氮循环是维持农业生产力的核心过程，直接影响作物的生长和土壤生态系统的功能（Fageria & Baligar, 2005）。

土壤 pH 是衡量土壤酸碱状态的重要指标，影响土壤中养分的有效性、重金属的生物可利用性及微生物群落的组成。不同作物对土壤 pH 的适应性不同，因此 pH 调节是土壤管理和农作物生产中的核心措施（Horneck et al., 2011）。不同作物对土壤 pH 的适应性不同，例如，茶树适宜生长在酸性土壤（pH 4.5 - 5.5）中，而甜菜则更适合在中性至微碱性土壤（pH 6.5 - 7.5）中生长。土壤 pH 调节是土壤管理和农作物生产中的核心措施。

基于以上对土壤因子的认识，我希望通过研究土壤黏粒含量、土壤有机碳密度、土壤全氮含量、土壤酸碱度这四个土壤因子与天津市地理范围内土壤类型的关联，探究各个土类亚类与这四个因子的关联程度。为了实现这一目标，我选择了**随机森林方法**进行机器学习。随机森林是一种集成学习算法，它通过构建多个决策树并综合其结果来提高模型的准确性和稳定性。其基本原理是从原始数据集中有放回地抽取多个样本，构建多个决策树，每个决策树在节点分裂时随机选择一部分特征进行最优分裂。最终，通过对多个决策树的预测结果进行投票或平均，得到随机森林的预测结果。这种方法能够有效减少模型的过拟合现象，提高模型的泛化能力。

此外，我还尝试对 Topic4 中的核心地形栅格数据进行截取，希望获得天津市地理范围内的地形值分布，将其与这四个土壤因子进行关联性分析。然而，由于源数据的分辨率无法支撑小范围内地形像元的划分，这一思路未能实现。

```
## 四、数据处理 (Data Process)
```

```
# 加载包
library(terra)
library(caret)
library(randomForest)
library(ggplot2)
library(ggExtra)
library(ggpmisc)
library(ggpubr)

# 读取 shapefile
shp_path <- "Data/Resource/TJ-Landkind(2018)-1m/tianjin100.shp"
tianjin_shp <- vect(shp_path)

# 读取土壤特征tif 文件
clay_raster <- rast("Data/Resource/TJ-土壤黏粒含量-250/clay_0_5cm_mean.tif")
nitrogen_raster <- rast("Data/Resource/TJ-土壤全氮含
量-250/nitrogen_0_5cm_mean.tif")
ph_raster <- rast("Data/Resource/TJ-土壤酸碱度-250/phh2o_0_5cm_mean.tif")
ocd_raster <- rast("Data/Resource/TJ-土壤有机碳密度-250/ocd_0_5cm_mean.tif")

# 提取每个地块的土壤特征
```

```
clay_values <- extract(clay_raster, tianjin_shp, fun = mean, na.rm = TRUE)[,2]
nitrogen_values <- extract(nitrogen_raster, tianjin_shp, fun = mean, na.rm =
TRUE)[,2]
ph_values <- extract(ph_raster, tianjin_shp, fun = mean, na.rm = TRUE)[,2]
ocd_values <- extract(ocd_raster, tianjin_shp, fun = mean, na.rm = TRUE)[,2]

# 合成数据框
soil_data <- data.frame(
  clay.ID = clay_values,
  nitrogen.ID = nitrogen_values,
  ph.ID = ph_values,
  ocd.ID = ocd_values,
  亚类 = tianjin_shp$亚类
)

# 检查缺失值并去除
soil_data <- na.omit(soil_data)

# 将目标变量转换为因子
soil_data$亚类 <- as.factor(soil_data$亚类)

# 设置过采样的训练控制参数
train_control <- trainControl(method = "cv", number = 10, sampling = "up",
savePredictions = "final")

# 训练模型并进行过采样
model <- train(亚类 ~ clay.ID + nitrogen.ID + ph.ID + ocd.ID,
               data = soil_data,
               method = "rf",
               trControl = train_control)

# 给出模型摘要，查看交叉验证性能
print(model)
```

terra包用于处理地理空间数据，包括读取和操作栅格数据；caret包提供了统一的接口来实现各种机器学习算法和模型评估方法；randomForest包专门用于实现随机森林算法；而ggplot2、ggExtra、ggpmisc和ggpubr则用于数据可视化和图形美化。接下来，我使用extract()函数从栅格数据中提取每个地块（shapefile 中的多边形）的土壤特征平均值。通过设置fun = mean和na.rm = TRUE，确保计算平均值时忽略缺失值。然后，我将提取的四种土壤特征值与 shapefile 中的土壤亚类属性合并，创建了一个数据框soil_data。
为了保证模型训练的质量，我使用na.omit()函数删除了包含缺失值的记录。最后，将目标变量 "亚类" 转换为因子类型，因为我们要进行的是分类预测任务。在模型构建阶段，我使用了caret包中的train()函数来训练随机森林模型。首先，我设置了训练控制参数train_control，选择了 10 折交叉验证方法（method = "cv"，number = 10），并启用了过采样（sampling = "up"）来处理类别不平衡问题。然后，我指定了模型公式，以 "亚类" 为目标变量，以四种土壤特征为预测变量。通过设置method = "rf"，我选择了随机森林算法进行建模。训练过程中，模型会自动进行参数调优，以找到最优的模型配置。

```
# 加载包
library(dplyr)
library(ggplot2)
library(foreach)
library(doParallel)
library(broom)
```

```
# 提前需运行获得model.R中的summary_long
source("Src/Process/Model.R")
#v分组
group1 <- c("和平区", "河东区", "河西区", "河北区", "南开区", "红桥区")
group2 <- c("滨海新区", "东丽区", "西青区", "北辰区", "武清区", "静海区")
group3 <- c("津南区", "宁河区", "宝坻区", "蓟州区")

# 为每个区分配组
summary_long <- summary_long %>%
  mutate(District = case_when(
    name %in% group1 ~ "Group1",
    name %in% group2 ~ "Group2",
    name %in% group3 ~ "Group3",
    TRUE ~ "Other"
  ),
  Factor = recode(factor,
                  "Organic Carbon Density" = "OCD",
                  "Clay Content" = "Clay",
                  "Nitrogen Content" = "Nitrogen",
                  "pH Value" = "pH"),
  Value = value)

# 设置并行运算
numCores <- parallel::detectCores() - 1
cl <- makeCluster(numCores)
registerDoParallel(cl)

# 嵌套并行运算
results <- foreach(d = unique(summary_long$District), .combine = rbind, .packages
= c("dplyr", "broom")) %:%
  foreach(f = unique(summary_long$Factor), .combine = rbind) %dopar% {
    sub_data <- summary_long %>% filter(District == d, Factor == f)
    model <- lm(Value ~ 1, data = sub_data)
    tidy_out <- broom::tidy(model)
    data.frame(District = d,
               Factor = f,
               Estimate = tidy_out$estimate[1],
               p.value = tidy_out$p.value[1])
  }

# 停止并行
stopCluster(cl)

print(results)
```

dplyr用于数据处理和转换，ggplot2用于绘图，foreach和doParallel用于实现并行计算，broom则用于将统计模型的结果转换为整洁的数据框格式。然后，我通过source()函数导入了之前运行的 Model.R 脚本中生成的summary_long数据集，该数据集包含了土壤因子的相关统计信息。

由此，我对天津市的各个行政区进行分组。根据地理位置和土壤特性的相似性，我将 16 个区分为三组：Group1 包含中心城区的 6 个区，Group2 包含环城区的 6 个区，Group3 包含远郊区的 4 个区。然后，我使用case_when()函数为summary_long数据集中的每个观测分配对应的组标签。同时，我将土壤因子的名称进行了缩写重编码，使后续分析和可视化更加简洁。

在处理大规模土壤数据时，为了提升分析效率，我采用了并行计算与程序嵌套相结合的优化策略。具体来说，我先通过系统函数检测计算机的可用 CPU 核心数，并保留一个核心以确保电脑运行流畅。然后，

我创建了一个计算集群，将这些核心组织起来协同工作，就像组建了一个高效的并行计算的群。
进一步，我尝试设计了一个嵌套循环结构。外层循环按照之前定义的区域分组（比如中心城区、环城区、远郊区）对数据进行划分，内层循环则针对不同的土壤因子（如有机碳密度、黏粒含量等）进行处理。这样的程序嵌套结构，让我能够系统地对每个 "区域 - 因子" 组合进行单独分析。
接下来，将这些组合分配给计算集群中的不同核心同时处理。这就像是把一项大任务拆分成多个小任务，让多人进行，提高了处理速度。对于每个组合的数据，我使用统计模型计算平均值并检验其显著性，然后将结果整理成规范的表格。所有计算完成后，关闭行计算群，释放系统资源。

五、数据可视化与结论(Data Visualization and Conclusion)

```
# ggplot2启动!
ggplot(resampling_results_long, aes(x = mtry, y = Value, color = Metric, group =
Metric)) +
  geom_line(linewidth = 1) +
  geom_point(size = 3) +
  labs(title = "Resample Result", x = "mtry", y = "值") +
  scale_x_continuous(breaks = resampling_results$mtry) +
  theme_minimal(base_size = 14, base_family = "serif") +
  theme(
    plot.title = element_text(hjust = 0.5),
    panel.grid.major = element_line(color = "gray90"),
    panel.grid.minor = element_blank()
  )
```

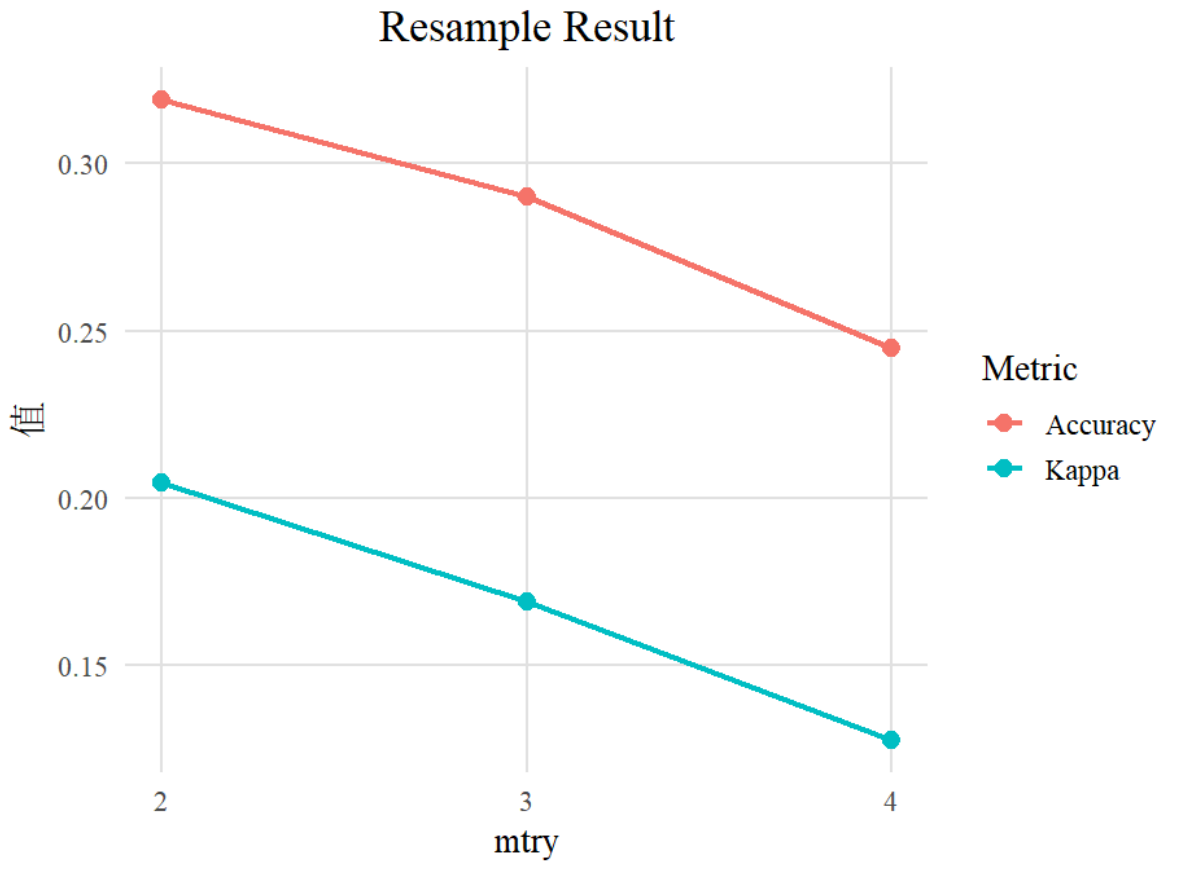


Fig.1 Resample Result

该图聚焦于重采样结果，横坐标 “mtry” 为随机森林模型中的重要参数，它代表在每次分裂时考虑的变量数。纵坐标展示了准确率（**Accuracy**）和 **Kappa** 系数两个关键指标。从图中曲线走势可知，随着 “mtry” 值在 2 - 3 区间变动，准确率和 **Kappa** 系数均出现波动。

准确率反映模型正确分类样本的比例，**Kappa** 系数则在考虑了随机分类情况后，更精准地衡量模型分类效果与随机分类的差异程度 。当 “mtry” 取值不同时，模型对样本的分类能力有所不同。在该区间内，若某一 “mtry” 值对应的准确率和 **Kappa** 系数较高，表明在此参数设定下，模型能更有效地识别样本特征，对数据分类更为准确。例如，当 “mtry” 为某一特定值时，模型在训练集和测试集上都能保持较高的准确率，意味着模型在该参数设置下具有较好的泛化能力，能更准确地对新数据进行分类预测。不过，要确定 “mtry” 的最优值，还需结合实际数据的分布特征、研究目的以及多次实验结果综合判断。

```
# ggplot启动
ggplot(importance_scores, aes(x = reorder(Feature, -Overall), y = Overall, fill =
factor(Feature))) +
  geom_col(width = 0.4) +
  scale_fill_manual(values = color_palette) +
  labs(title = "Feature Importance Ranking", x = "Feature", y = "Importance") +
  theme_minimal(base_size = 14, base_family = "serif") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 12, color =
"black"),
    axis.text.y = element_text(size = 12, color = "black"),
    axis.title.x = element_text(size = 14, color = "black"),
    axis.title.y = element_text(size = 14, color = "black"),
    plot.title = element_text(size = 16, color = "black", hjust = 0.5),
    panel.grid.major = element_line(color = "gray90"),
    panel.grid.minor = element_blank(),
    axis.line = element_line(color = "black"),
    legend.position = "none"
  )
```

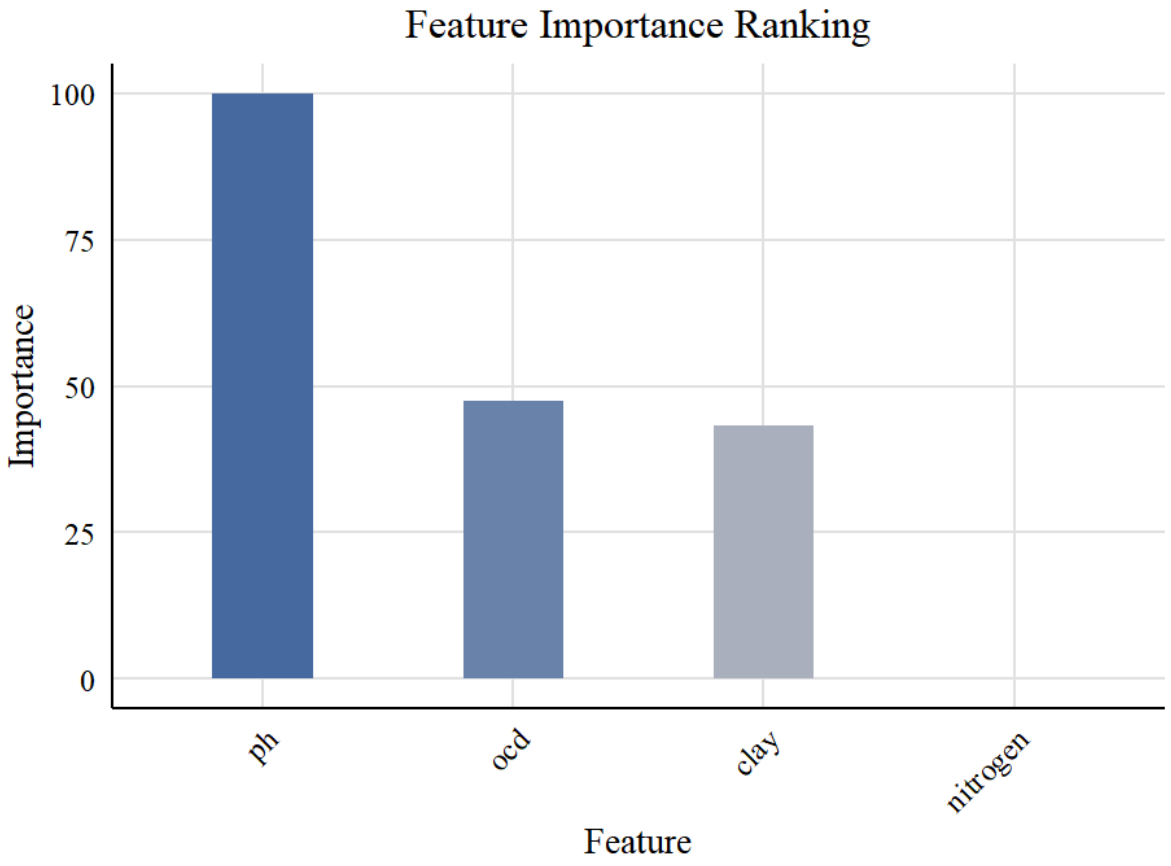


Fig.2 Feature Importance Ranking

从 Fig.2 “Feature Importance Ranking” 图呈现的 “ph>organic carbon density>clay>nitrogen” 且 “nitrogen 近乎为 0” 这一结果来看，可从数据和模型构建两个关键角度去剖析。

数据方面，可能存在的情况是，本次研究获取的土壤数据中，**氮元素的含量变化极小**。土壤中氮元素虽然是植物生长必需的营养元素，但在天津市的这些土壤样本里，也许由于采样区域相对集中，或者当地整体的地质条件、农业活动等因素影响，使得不同土壤样本间氮含量差异不大。这就好比在一个区域内，所有土壤都来自相似的母质，并且当地施肥习惯比较统一，导致氮含量基本处于同一水平。在这种情况下，模型在学习过程中，就很难从氮含量的差异上区分不同的土壤类型，所以它在决定土壤类型时所起的作用就微乎其微，**反映在图中就是重要性近乎为 0**。

再从模型构建角度分析，随机森林模型是基于大量决策树的组合来进行预测和分类的。模型在构建过程中，会根据各个变量对目标变量（这里是土壤类型）的影响程度来分配权重。有可能在划分决策树节点时，模型发现土壤酸碱度（ph）、有机碳密度（organic carbon density）和土壤黏粒（clay）这几个变量能更有效地将不同土壤类型区分开。比如说，在某个节点上，依据土壤酸碱度的不同取值，可以很明显地将某类土壤划分到不同分支，从而更精准地识别土壤类型。而氮元素由于变化不明显，在决策树的构建过程中，很难为划分土壤类型提供有价值的信息，所以在模型中得到的重要性就很低。这提醒我在后续研究中，如果想要提高氮元素在模型中的重要性，或许可以尝试扩大采样范围，增加土壤样本的多样性，让氮含量的差异更明显，这样模型就能更好地学习到氮元素与土壤类型之间的关系，进而更准确地反映它们之间的真实联系。

```
# 绘制 OOB error-Number of Trees
oob_plot <- ggplot(oob_error, aes(x = Trees, y = OOB)) +
  geom_line(color = "#e41a1c", size = 1.2) +
  geom_point(color = "#377eb8", size = 1) +
  labs(
    title = "OOB Error-Number of Trees",
    x = "Number of Trees",
    y = "OOB Error Rate"
  ) +
  theme_minimal(base_size = 14, base_family = "serif")
```

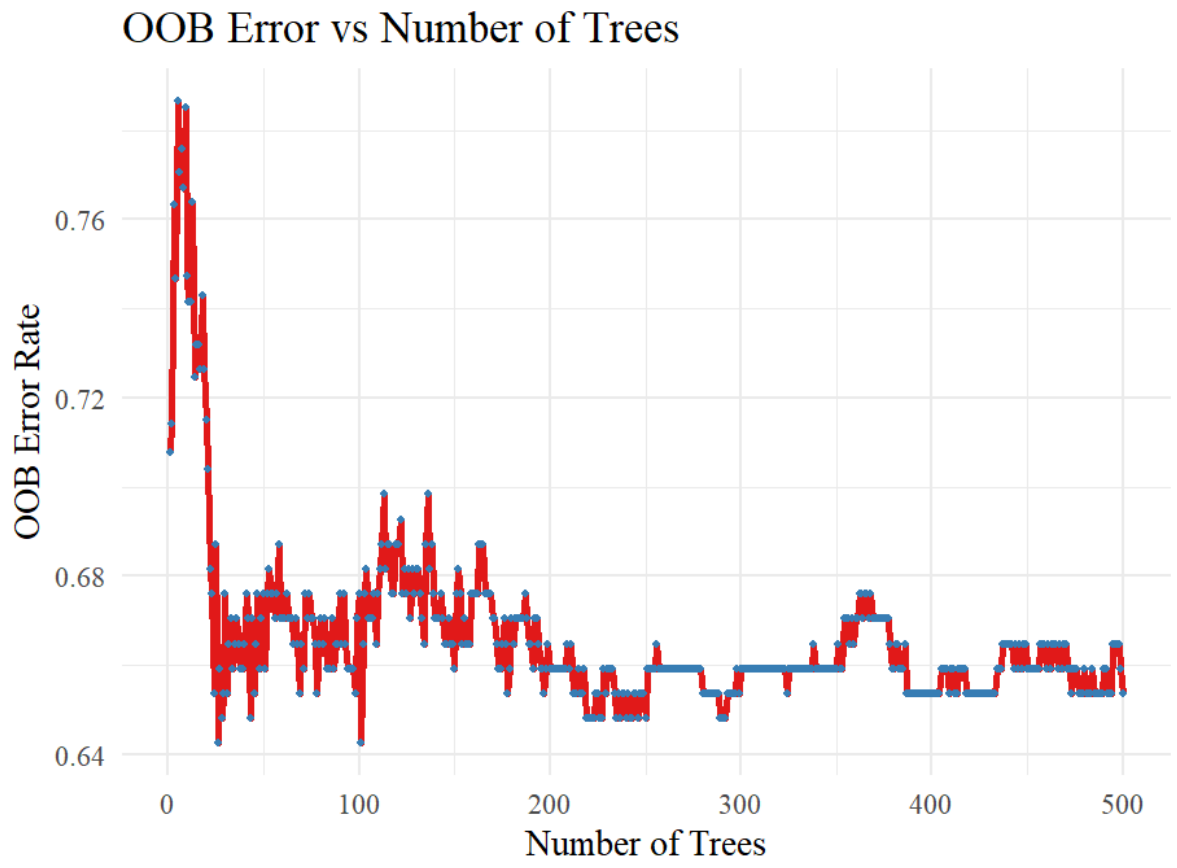


Fig.3 OBB Error - Number of Trees

该图呈现了袋外误差（**Out - of - Bag Error**，OOB Error）与树的数量（Number of Trees）之间的关系。袋外误差是随机森林模型特有的评估指标，它利用未参与每棵树构建的样本（袋外样本）来评估模型性能，避免了传统交叉验证的繁琐计算。

随着树的数量从 0 逐渐增加到 500，袋外误差呈现出先下降后趋于稳定的趋势。在初始阶段，增加树的数量使得模型能够学习到更多的数据特征和规律，不同树之间的组合可以更全面地覆盖数据的变化情况，从而降低模型的误差，提高模型的稳定性和准确性。例如，当树的数量较少时，模型可能无法充分捕捉到数据中的复杂关系，导致误差较大；而随着树的数量增多，模型能够从更多角度对数据进行拟合，误差随之减小。但当树的数量达到一定程度后，模型逐渐趋于饱和，新增加的树对模型性能提升的贡献越来越小，误差变化不再明显。此时继续增加树的数量，不仅会增加模型的计算成本，还可能引发过拟合问题。因此，在实际应用中，需要依据该曲线的变化趋势，选取误差相对较低且稳定时对应的树的数量作为模型参数，以平衡模型的准确性和计算效率。

```
# ggplot2启动
g <- ggplot(resamples_clean, aes(x = obs_num, y = pred_num)) +
  geom_jitter(alpha = 0.6, width = 0.2, height = 0.2, color = "#1f78b4") +
  geom_smooth(method = "lm", se = FALSE, color = "#33a02c") +
```

```
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "grey50") +
  annotate("text",
           x = min(resamples_clean$obs_num, na.rm = TRUE),
           y = max(resamples_clean$pred_num, na.rm = TRUE),
           label = paste0("Accuracy: ", acc_cv, "\nKappa: ", kappa_cv),
           hjust = 0, vjust = 1, size = 6, family = "serif") +
  labs(x = "Observed", y = "Predicted",
       title = "Cross Validation-New Model") +
  theme_minimal(base_size = 14, base_family = "serif")
# 添加边缘的直方图
g1 <- ggMarginal(g, type = "histogram", fill = "transparent")
```

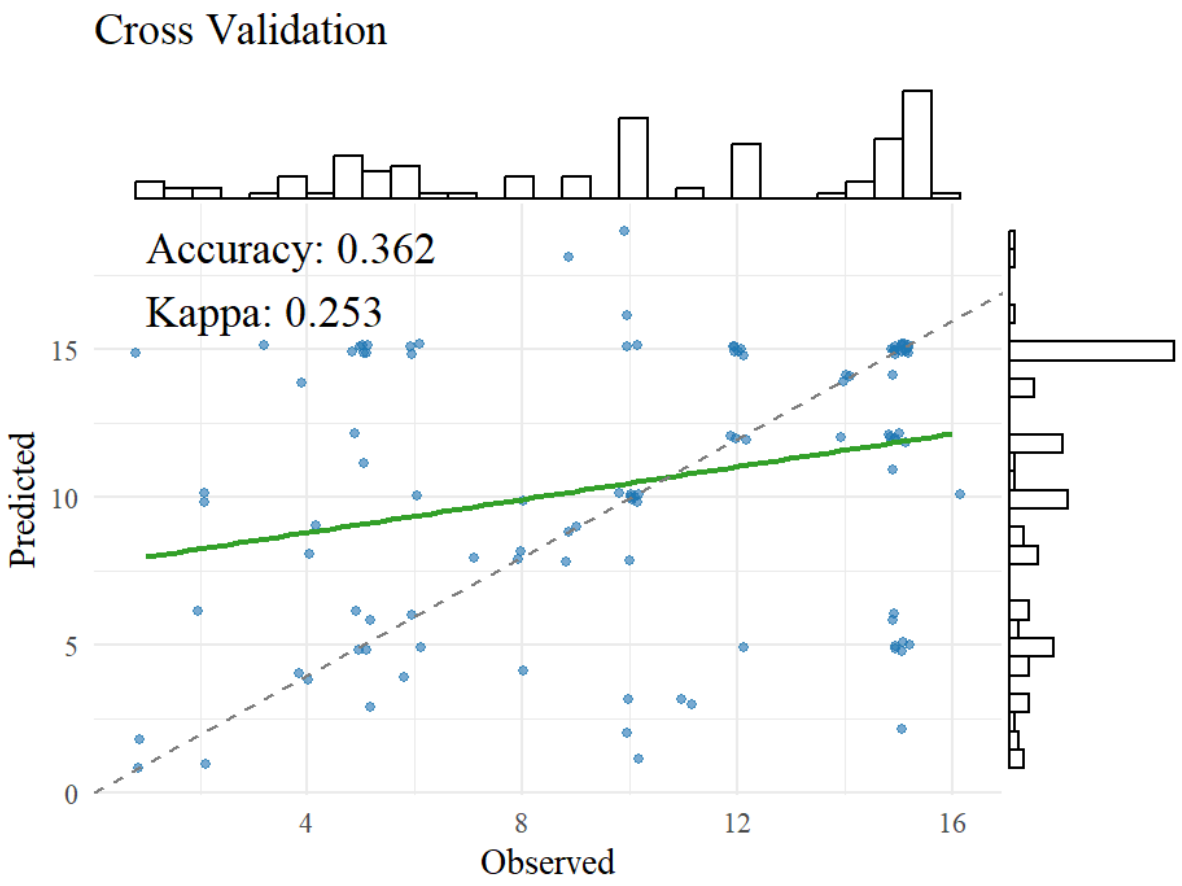


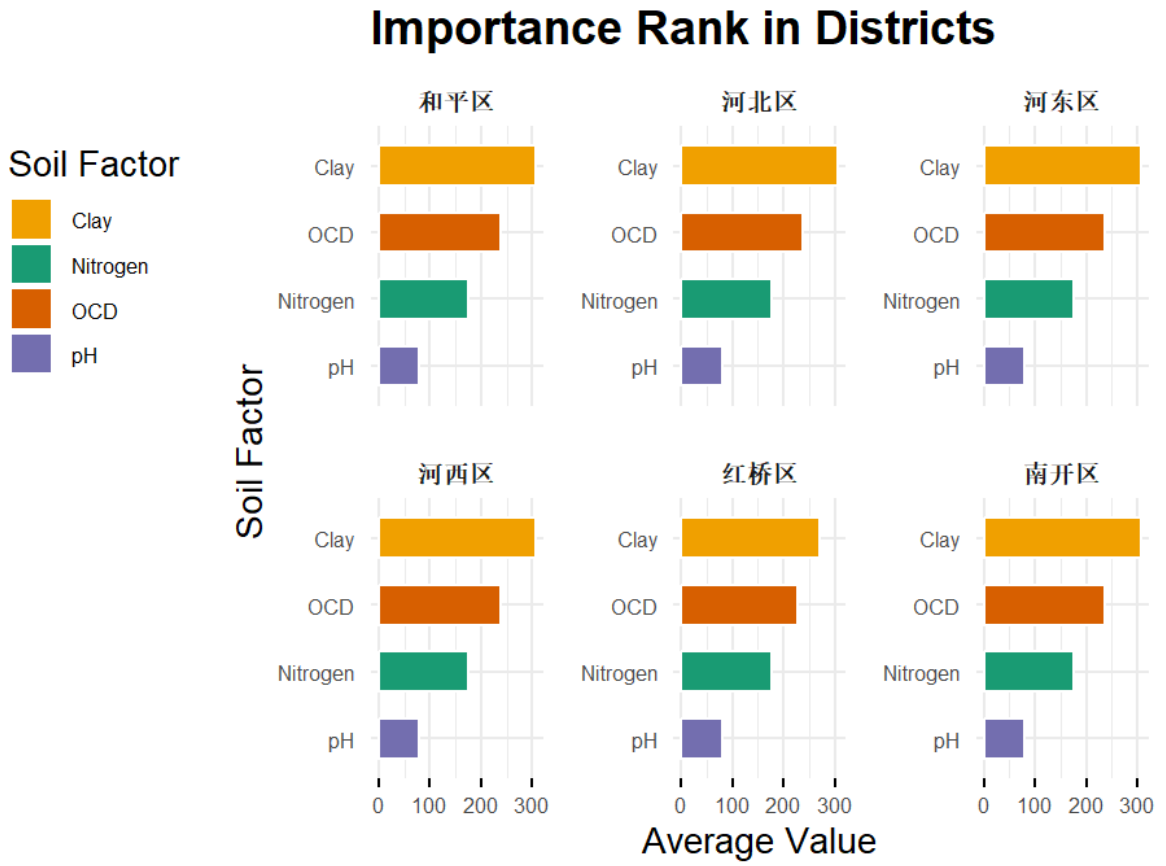
Fig.4 Cross Validation - New Model

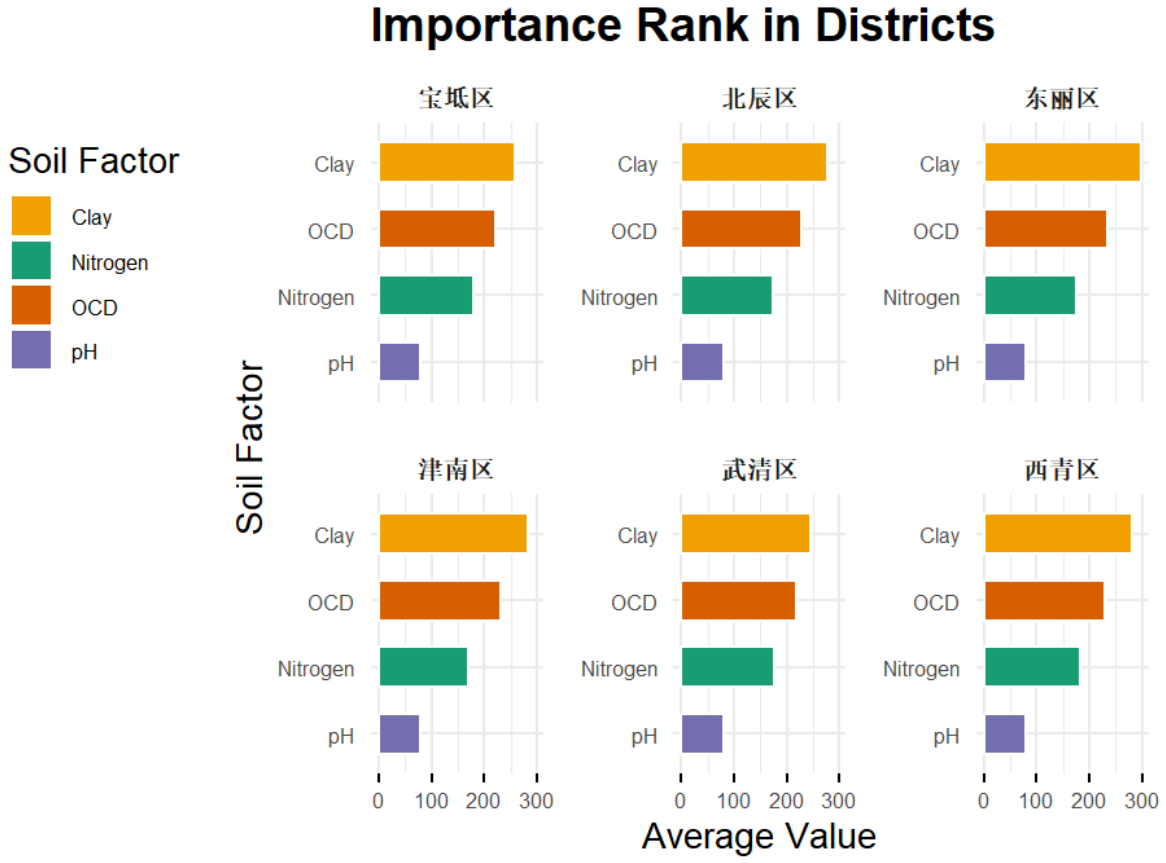
这张图展示了新模型的交叉验证结果，其中准确率（**Accuracy**）为 0.362，**Kappa** 系数为 0.253 。交叉验证是一种常用的评估模型泛化能力的方法，它将数据集划分为多个子集，通过在不同子集上进行训练和测试，得到模型在不同数据分布下的性能表现，从而更全面、客观地评估模型的优劣。

从本次交叉验证的结果来看，模型的准确率和 **Kappa** 系数相对较低，表明模型的性能有待提升。准确率仅为 0.362，意味着模型正确分类的样本比例不高，可能在识别样本特征时存在一定困难。**Kappa** 系数为 0.253，说明模型的分类效果仅略优于随机分类。不过，这并不意味着该模型完全不可用。一方面，模型性能不佳可能是由于数据量不足，导致模型无法充分学习到数据中的规律；另一方面，模型参数设置可能不合理，未能充分发挥模型的潜力。后续可以考虑增加数据量，使模型有更多的数据进行学习和训练；同时，对模型参数进行进一步调优，如调整决策树的深度、节点分裂的最小样本数等，以提升模型的准确率和稳定性。

```
# 循环ggplot2
plots <- lapply(split_districts, function(d_group) {
  ggplot(summary_long %>% filter(name %in% d_group),
         aes(x = reorder(factor, value), y = value, fill = factor)) +
```

```
    geom_col(width = 0.6, color = "white") +
    facet_wrap(~ name, scales = "free_y", ncol = 3) +
    coord_flip() +
    scale_fill_viridis(discrete = TRUE, option = "D") +
    labs(
      title = "Importance Rank in Districts",
      x = "Soil Factor", y = "Average Value", fill = "Soil Factor"
    ) +
    theme_minimal(base_size = 14) +
    theme(
      plot.title = element_text(size = 18, face = "bold"),
      strip.text = element_text(size = 10, face = "bold"),
      axis.text.y = element_text(size = 8),
      axis.text.x = element_text(size = 8),
      legend.position = "left",
      legend.justification = "top",
      legend.box = "horizontal",
      legend.text = element_text(size = 8),
      panel.spacing = unit(1, "lines"),
      axis.ticks.x = element_line(size = 0.5)
    ) +
    scale_x_discrete() +
    scale_fill_manual(values = c("Clay" = "#F4A300", "OCD" = "#D95F02",
"Nitrogen" = "#1B9E77", "pH" = "#7570B3")) +
    scale_y_continuous(labels = scales::comma)
})
for (i in 1:length(plots)) {
  print(plots[[i]])
}
```


Importance Rank in Districts
Soil Factor
Clay
Nitrogen
OCD
pH
和平区
河北区
河东区
河西区
红桥区
南开区
Clay
OCD
Nitrogen
pH
0
100
200
300
Soil Factor
Average Value


Importance Rank in Districts
Soil Factor
Clay
Nitrogen
OCD
pH
宝坻区
北辰区
东丽区
津南区
武清区
西青区
Clay
OCD
Nitrogen
pH
0
100
200
300
Soil Factor
Average Value

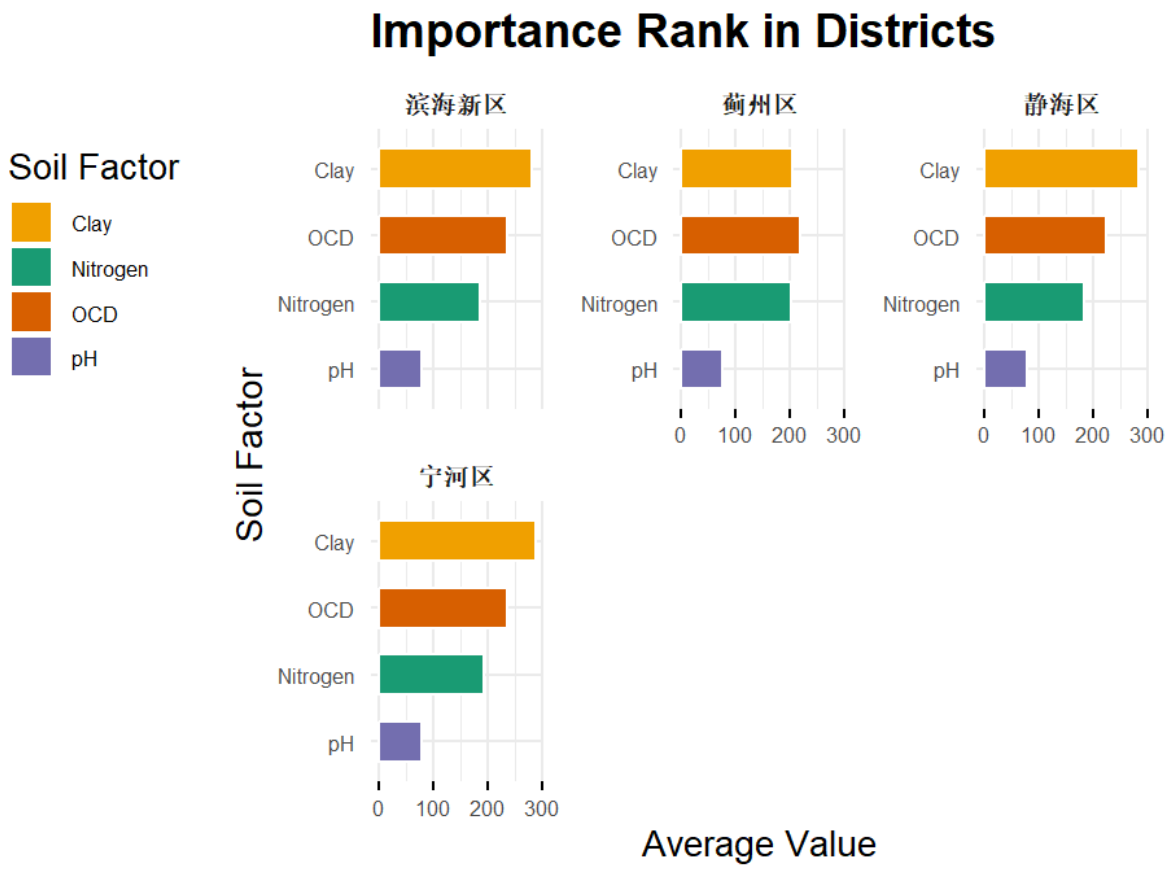


Fig.5 Importance Rank in Districts - 1、Fig.5 Importance Rank in Districts - 2、Fig.5 Importance Rank in Districts - 3

这三张图展示了天津市不同区域土壤因子的重要性排名情况。在和平区、河东区等各个区域中，土壤黏粒（Clay）、土壤有机碳（OCD）、氮（Nitrogen）和土壤酸碱度（pH）的重要性排序基本保持一致，土壤黏粒始终排在前列，随后依次是土壤有机碳、氮，土壤酸碱度排在最后。

这一结果表明，在天津市不同区域，土壤黏粒对土壤类型的影响最为显著。土壤黏粒的物理化学性质决定了它在土壤结构形成、养分储存和交换等方面的关键作用，进而在土壤类型的划分和特性塑造中占据主导地位。不同区域之间，虽然土壤因子的重要性排序相同，但这些因子的平均数值可能存在差异。例如，滨海新区和宝坻区的土壤黏粒平均含量可能不同，这种差异反映了不同区域的土壤特性存在一定区别。这些区域特性差异与当地的地质条件、气候因素、人类活动等密切相关。在进一步研究土壤差异和生态治理时，需要充分考虑这些区域特性，以便制定更具针对性的措施。

```
# ggplot2启动!
ggplot(results, aes(x = Factor, y = Estimate, fill = District)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.75), width =
0.7) +
  geom_text(aes(label = p_significance, color = p_significance_factor),
            vjust = -0.5, size = 5, position = position_dodge(width = 0.75)) +
  scale_fill_manual(values = c("Group1" = "lightblue", "Group2" = "lightgreen",
"Group3" = "lightcoral", "Other" = "lightgray")) +
  scale_color_manual(values = c("4*" = "red", "3*" = "lightcoral", "2*" =
"orange", "*" = "green", "ns" = "blue")) +  # 为星号设定颜色
  labs(title = "Soil Factor Estimates by District with P-value", x = "Soil
Factor", y = "Estimate Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  guides(fill = guide_legend(title = "District"),
         color = guide_legend(title = "Significance", order = 2))
```

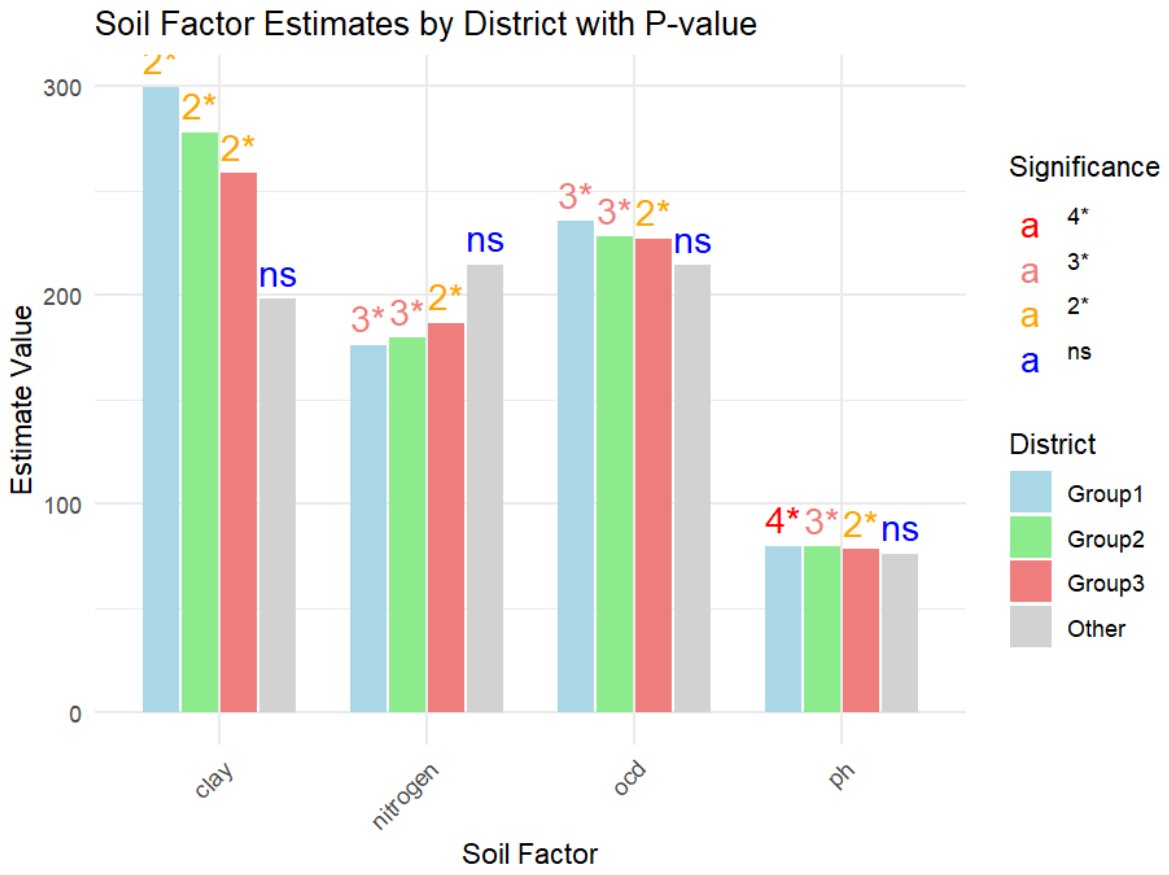


Fig.6 Soil Factor Estimates by District with P - value

该图呈现了不同区域（District）土壤因子（nitrogen、ocd、clay、ph）的估计值以及对应的 **P 值**。P 值用于判断样本结果是否具有统计学意义，通常以 0.05 为临界值。

从图中可以观察到，不同区域的土壤因子估计值存在差异，并且部分区域和因子组合带有显著性标记（星号）。带有显著性标记的区域和因子，表明它们之间的关系在统计学上具有显著意义。例如，在某些区域，土壤黏粒的估计值与其他区域相比有明显差异，且通过 P 值判断这种差异并非由随机因素导致，而是具有实际的生物学或地理学意义。这意味着在这些区域，土壤黏粒与土壤类型之间的关联更为特殊，可能存在独特的土壤形成过程或影响因素。对这些具有显著差异的区域和因子进行深入研究，有助于揭示不同区域土壤特性和土壤类型形成的内在机制，为土壤分类、土地利用规划以及生态环境保护提供更科学的依据。

```
# ggplot2启动!
map_plot <- ggplot() +
  geom_sf(data = admin_sf, fill = NA, color = "black", size = 0.5) +
  geom_sf(data = tianjin_shp, aes(fill = as.factor(亚类)), color = NA) +
  scale_fill_viridis_d(name = "Landkind Subclass") +
  coord_sf(xlim = c(lon_min, lon_max), ylim = c(lat_min, lat_max), expand =
FALSE) +
  theme_minimal() +
  labs(title = "Tianjin Landkind Subclass Map") +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text.x = element_text(size = 4),
    axis.ticks.x = element_line(linewidth = 1),
    legend.position = "bottom",
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 8)
  )
```

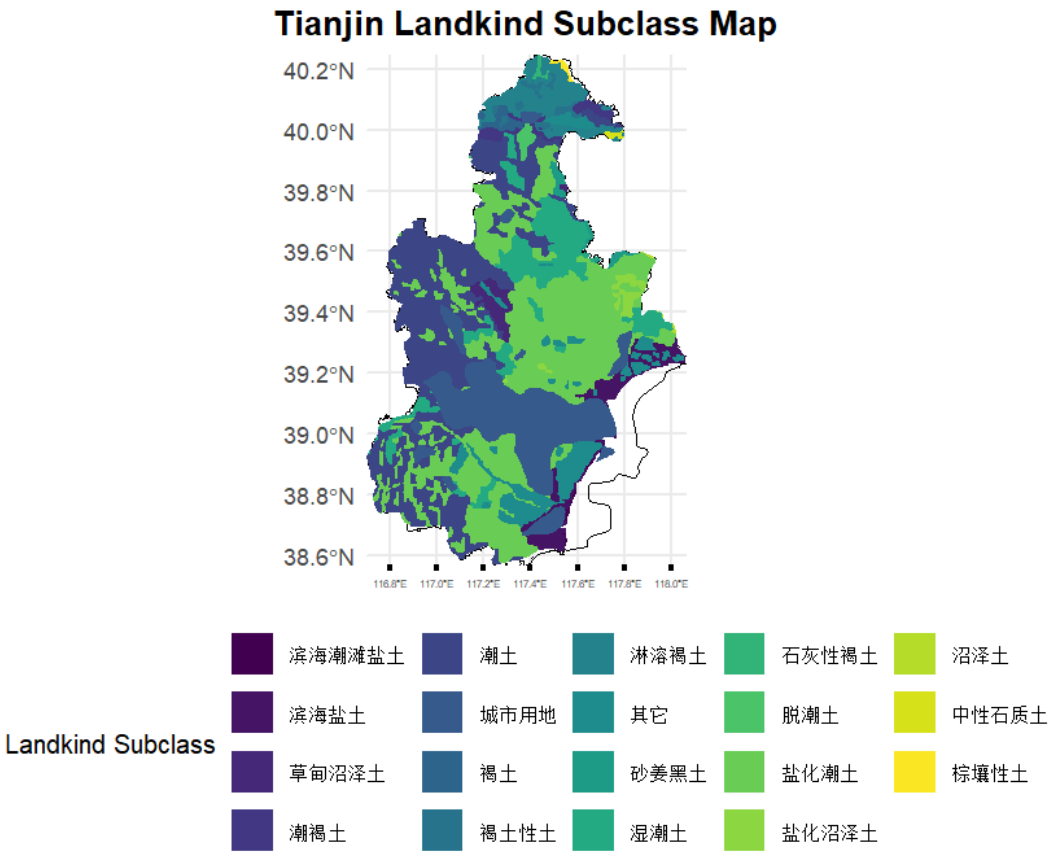


Fig.7 Tianjin Landkind Subclass Map

天津市的地理范围，纵坐标纬度范围大致在 38.6°N - 40.2°N，横坐标经度范围在 116.8°E - 118.0°E，明确了地图所涵盖的区域位置。

土地类型亚类分布

- **滨海潮滩盐土**：以深紫色标识，主要分布在天津市沿海区域。这是由于沿海地区受海水影响，盐分积聚，形成了潮滩盐土，其独特的盐分条件影响着该区域的植被生长和生态系统，一般植被多为耐盐植物 。

- **潮土**：用深蓝色表示，分布较为广泛，在天津市中部和部分靠近河流的区域都有分布。潮土是在长期受河水泛滥沉积，经过人类耕作熟化而形成的土壤，肥力状况受灌溉、施肥等人为因素影响较大。
- **淋溶褐土**：蓝绿色区域，多见于山区或地势较高、排水较好的地方。淋溶作用较强，使得土壤中的一些物质被淋洗，具有特定的土壤结构和养分特征。
- **石灰性褐土**：绿色区域，分布有一定范围。这类土壤富含碳酸钙，在一定程度上影响土壤酸碱度和养分有效性，对当地农作物种植和生态环境有相应作用。
- **沼泽土**：浅绿色标识，主要在地势低洼、排水不畅，长期或季节性积水的区域形成，其富含有机质，但通气性较差。

其他土地类型

- **滨海盐土**、**草甸沼泽土**等：也在图中有各自的分布区域，不同的颜色对应不同的形成条件和土壤特性。比如滨海盐土主要在沿海受海水影响的区域，草甸沼泽土则多在具有草甸植被和沼泽化过程的地段 。
- **城市用地**：用深蓝色系表示，集中在天津市的城市区域。

```
# ggplot2启动!
plot_longitude <- ggplot(ph_data, aes(x = longitude, y = ph_value, color =
ph_value)) +
  geom_point(alpha = 0.7) +
  scale_color_viridis_c(name = "pH Value") +
  labs(title = "Soil pH Variation by Longitude", x = "Longitude (°E)", y = "pH
Value") +
  theme_minimal()

plot_latitude <- ggplot(ph_data, aes(x = latitude, y = ph_value, color =
ph_value)) +
  geom_point(alpha = 0.7) +
  scale_color_viridis_c(name = "pH Value") +
  labs(title = "Soil pH Variation by Latitude", x = "Latitude (°N)", y = "pH
Value") +
  theme_minimal() +
  coord_flip()

# 使用 cowplot 拼接
title_plot <- ggdraw() +
  draw_label("Soil pH Spatial Distribution in Tianjin", fontface = "bold", size =
14, x = 0.5, y = 0.85)

combined_plot <- plot_grid(
  plot_longitude,
  plot_latitude,
  ncol = 1,
  rel_heights = c(0.7, 0.7)
)

final_plot <- plot_grid(title_plot, combined_plot, ncol = 1, rel_heights = c(0.5,
9))
```

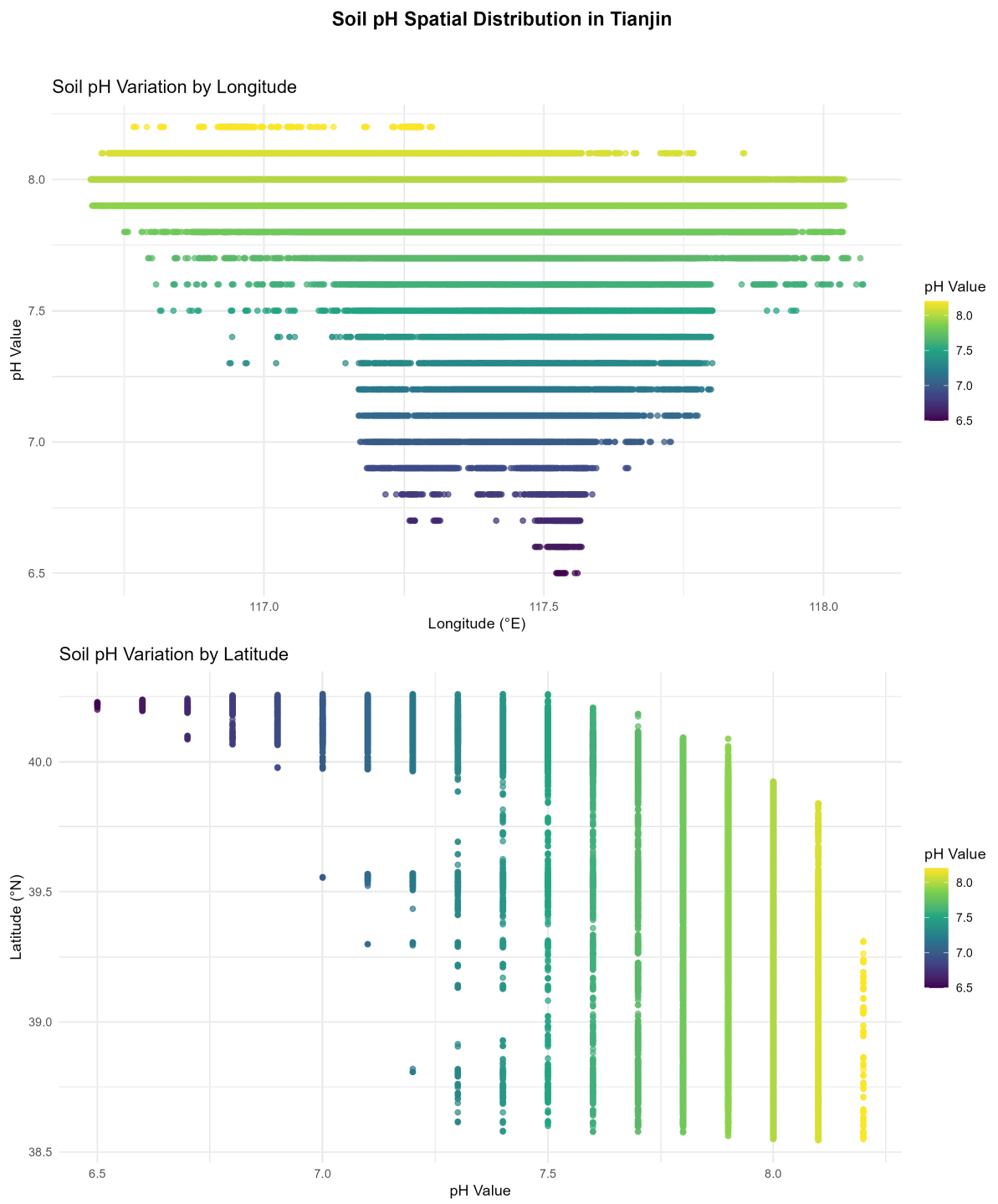


Fig.8 Soil pH Spatial Distribution in Tianjin

按经度变化（Soil pH Variation by Longitude）

横坐标展示的是经度，范围从 117.0°E 到 118.0°E ，纵坐标则是土壤 pH 值，跨度在 6.5 - 8.0 之间。不同颜色的点与线条代表着各异的 pH 值，对应右侧图例。

在研究过程中，我发现随着经度变化，土壤 pH 值波动明显。像在 117.0°E - 117.2°E 附近，pH 值分布极为分散，从酸性到碱性范围都有涉及。这可能是因为该区域内成土母质复杂多样，不同母质所含化学成分不同，经过长期风化和土壤发育过程，导致土壤酸碱度差异大。而且，此区域可能存在不同程度的人类活动干扰，比如工业排放、农业施肥等，这些活动会向土壤中引入不同化学物质，进而影响 pH 值。

而在 117.6°E - 117.8°E 附近，pH 值相对集中在 7.0 - 7.5 之间。也许该区域地形相对均一，排水条件相似，使得土壤中酸碱物质淋溶程度相近，从而 pH 值较为稳定。或者在土地利用方式上较为统一，例如都是以某种特定的农业种植为主，且采用相似的施肥和管理措施，所以对土壤酸碱度的影响较为一致 。

按纬度变化（Soil pH Variation by Latitude）

横坐标为土壤 pH 值，范围 6.5 - 8.0 ，纵坐标是纬度，从 38.5°N 到 40.0°N 。不同颜色表示不同 pH 值。

从图中能看出，随着纬度升高，土壤 pH 值变化趋势明显。在较低纬度，像 38.5°N - 39.0°N 区域，pH 值分布范围广。这可能是因为低纬度地区降水相对较多，淋溶作用较强，使得土壤中碱性物质被淋洗，导致 pH 值波动较大。同时，该区域植被类型可能丰富多样，不同植被根系分泌物质以及残体分解对土壤酸碱度的影响不同。

而在较高纬度，接近 40.0°N 时，pH 值有向较高值集中的趋势。这或许是由于高纬度地区气温较低，微生物活动相对较弱，土壤中有机物分解缓慢，有机酸产生较少，而且蒸发较弱，盐分积累相对较多，使得土壤呈现相对偏碱性 。

六、参考(References)

1. Brady, N. C., & Weil, R. R. (2008). The nature and properties of soils (14th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
2. Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. Science, 304(5677), 1623–1627. https://doi.org/10.1126/science.1097396
3. Fageria, N. K., & Baligar, V. C. (2005). Enhancing nitrogen use efficiency in crop plants. Advances in Agronomy, 88, 97–185. https://doi.org/10.1016/S0065-2113(05)88004-6
4. Horneck, D. A., Sullivan, D. M., Owen, J. S., & Hart, J. M. (2011). Soil acidity and liming: Basic information for farmers and gardeners (EC 1478-E). Oregon State University Extension Service.