Metadata

Description of R code

Item	Description	Responsible
Name of the data / code	R code for modeling planting spot quality, seedling vitality, and seedling height	Author
Author & ORCID	Kemppainen, Kalle[1,2], ORCID 0009-0000-6184-8812 Miina, Jari[2], ORCID 0000-0002-8639-4383	Author
Authors' affiliation(s)	[1] School of Forest Sciences, University of Eastern Finland, Yliopistokatu 7, FI-80101 Joensuu, Finland (https://ror.org/00cyydd11) [2] Natural Resources Institute Finland, Yliopistokatu 6, FI-80100 Joensuu, Finland (https://ror.org/02hb7bm88)	Author
Owner of the material	Kemppainen, Kalle[1,2], ORCID 0009-0000-6184-8812 Miina, Jari[2], ORCID 0000-0002-8639-4383	Author
Publisher	Zenodo (after publication)	Author
Funder	University of Eastern Finland (https://ror.org/00cyydd11), Natural Resources Institute Finland (https://ror.org/02hb7bm88).	Author
Description	The aim of this analysis was to model ordinal responses (spot quality and seedling vitality) and a continuous response (seedling height). The dataset consisted of measurements carried out in 2022, 2023, and 2024. Spot quality was analyzed using data from 2022 (N = 4120). Seedling vitality and Norway spruce height were analyzed using follow-up data from 2023 (N = 1274) and 2024 (N = 989). The follow-up data was collected from repeated measurements within 60 permanent circular sample plots (N = 60). Note that the seedlings and spots of the sample plots were not tracked individually during the follow-up measurements (2023–2024).	Author
Methods	All the statistical analyses were executed with R (R version 4.5.1). Cumulative link models (CLM) and cumulative link mixed models (CLMM) were used to analyze the ordinal response variables: planting spot quality (CLM) and seedling vitality (CLMM). A linear mixed-effects model (LMEM) was used to analyze the height of Norway spruce (*Picea abies* (L.) Karst.) seedlings. The CLM and CLMM were fitted using the functions *clm* and *clmm2* from the *ordinal* package, respectively (R Core Team 2025; Christensen 2025). The LMEM was fitted using the function *lme* from *nlme* package (R Core Team 2025; Pinheiro and Bates 2026). Model fit was evaluated by the Akaike information criterion and by comparing nested models with likelihood ratio tests. Further details are provided in Kemppainen et al. (2026). The R code is available in Appendix 1.	Author
Variables	See the metadata file for the dataset.	Author
Author keywords	cumulative link model, cumulative link mixed model, linear mixed-effects model, R	Author
Vocabulary keywords (community standard)	-	Author
Discipline	Forest science	Archive/Repository/Publisher
Type of material	R code	Author
Language	eng	Author
Time range covered	-	Author
Geographic region	-	Author
Version	Version 1	Author
File format(s)	docx. & .txt	Author
Availability of the materials (open, embargo, registration, limited, registration required)	Open. The data and code are available upon request from Kalle Kemppainen or through the open research repository (Zenodo).	Author
Justification for access restrictions	-	Author
Licence	CC BY 4.0	Author
Connections with other research materials	-	Author

Access to the connected research materials	-	Author
Codes only: hardware/software requirements for running the code	R (R version 4.5.1)	Author
Connections to other products of research	Kemppainen et al. (2026)	Author
Personal data	No.	Author
Confidential or secret data	No.	Author
Publication date	26.6.2026 (Zenodo)	Archive/Repository/Publisher
Preservation policy	Permanent	Author
Permanent identifier (PID)	https://doi.org/10.5281/zenodo.20917313	Archive/Repository/Publisher

References

Christensen R (2025) Ordinal: Regression Models for Ordinal Data. R package version 2025. https://CRAN.R-project.org/package=ordinal. Accessed 17 June 2026.

Kemppainen K, Miina J, Tarvainen E, Kankaanhuhta V, Laitila J, Peltola H, Strandström M, Kärhä K (2026) Quality of drill-based site preparation and early performance of Norway spruce and Scots pine seedlings planted in different seasons. Silva Fenn 60, article id 26011. https://doi.org/10.14214/sf.26011.

Pinheiro J, Bates D (2026) Nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-169. https://cran.r-project.org/package=nlme. Accessed 17 June 2026.

R Core Team (2025) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/. Accessed 28 July 2025.

Appendix 1

```
####################################################################################
#
# Modeling ordinal and continuous responses
# Data: Drill-based site preparation (2022–2024)
# Original code by Jari Miina
# Modifications by Kalle Kemppainen
#
####################################################################################
# Fitted models (1–3):
#
# 1. Cumulative Link Model (CLM)
#     Quality of planting spots (ordinal response)
#     Data from 2022
#     All site preparation spots (N = 4120)
#
# 2. Cumulative Link Mixed Model (CLMM)
#     Seedling vitality (ordinal response)
#     Data from 2023 and 2024
#     Repeated follow-up measurements within 60 permanent sample plots (N = 2256)
#
# 3. Linear Mixed-Effects Model (LME)
#     Height of Norway spruce seedlings (continuous response)
#     Data from 2023 and 2024
#     Repeated follow-up measurements within 36 permanent sample plots (N = 1343)
####################################################################################

# R packages used

# for data processing
library(dplyr)
library(data.table)

# for visualization
library(ggplot2)
library(vioplot)
library(lattice)

# for modeling
library(nlme)
library(ordinal)

# for diagnostics
library(car)
library(lmfor)

####################################################################################
```

```
################################################################################
#
# Drill-based site preparation - Quality of planting spots
#
################################################################################

# Loading data (dataset_drill_2022_spot_quality)

d0 <- fread("x:/xxx/data.csv", sep = ";")

# Getting familiar with the data

#View(d0)
#str(d0)

# Convert all categorical variables and notes to factors

d0$site_id <- as.factor(d0$site_id)
levels(d0$site_id)

d0$spot_quality <- recode_factor(d0$spot_quality, "1" = "Poor", "2" = "Satisfactory", "3" = "Good")
levels(d0$spot_quality)
table(d0$spot_quality, d0$site_id)

d0$tree_species <- recode_factor(d0$tree_species, "1" = "NS", "2" = "SP")
levels(d0$tree_species)

d0$site_type <- recode_factor(d0$site_type, "2" = "MT", "3" = "OMT", "1" = "VT")
levels(d0$site_type)

d0$season <- recode_factor(d0$season, "1" = "Early", "2" = "Mid", "3" = "Late")
levels(d0$season)
table(d0$season, d0$site_id)

# Create dummy variable "stump or root":
d0$stump_root <- ifelse(d0$note_dummy_1_stump == 1 | d0$note_dummy_2_root ==1, 1, 0)
d0$note_dummy_1_stump_2_root <- as.factor(d0$stump_root)
table(d0$note_dummy_1_stump_2_root)

d0$note_dummy_3_stone <- as.factor(d0$note_dummy_3_stone)
table(d0$note_dummy_3_stone)

d0$note_dummy_4_logging_residues <- as.factor(d0$note_dummy_4_logging_residues)
table(d0$note_dummy_4_logging_residues)

d0$note_dummy_5_water <- as.factor(d0$note_dummy_5_water)
table(d0$note_dummy_5_water)

d0$note_dummy_6_humus <- as.factor(d0$note_dummy_6_humus)
```

```
table(d0$note_dummy_6_humus)

d0$note_dummy_7_logging_trail <- as.factor(d0$note_dummy_7_logging_trail)
table(d0$note_dummy_7_logging_trail)

# Boxplots and violin plots (for the depth and diameter of planting spots):
boxplot(d0$spot_depth~d0$spot_quality)
boxplot(d0$spot_diameter~d0$spot_quality)
vioplot(d0$spot_depth~d0$spot_quality)
vioplot(d0$spot_diameter~d0$spot_quality)

# Histograms (for the depth and diameter of planting spots):
hist(d0$spot_depth)
hist(d0$spot_diameter)

# Mean spot diameters and depths by seasons:
d0 %>% group_by(season) %>% summarise(mean_diameter = mean(spot_diameter, na.rm=TRUE),
                                      sd_diameter = sd(spot_diameter, na.rm=TRUE),
                                      mean_depth = mean(spot_depth, na.rm=TRUE),
                                      sd_depth = sd(spot_depth, na.rm=TRUE))

# Mean spot diameters and depths by site types:
d0 %>% group_by(site_type) %>% summarise(mean_diameter = mean(spot_diameter, na.rm=TRUE),
                                      sd_diameter = sd(spot_diameter, na.rm=TRUE),
                                      mean_depth = mean(spot_depth, na.rm=TRUE),
                                      sd_depth = sd(spot_depth, na.rm=TRUE))

# Some observations showed extreme values for spot depth and diameter.
# However, these values were verified as correct and were therefore retained in the analysis.

# Interaction term (season * site type):
int_season_sitetype <- as.factor(interaction(d0$season, d0$site_type, drop=TRUE))
summary(int_season_sitetype)

#str(d0)

# Let's check how we should interpret the coefficients (clm):

m0 <- clm(spot_quality ~ 1, data = d0, Hess = TRUE)
summary(m0)

m1 <- clm(spot_quality ~ site_type, data = d0, Hess = TRUE)
summary(m1)

m2 <- clm(spot_quality ~ season, data = d0, Hess = TRUE)
summary(m2)

m3 <- clm(spot_quality ~ site_type + season, data = d0, Hess = TRUE)
summary(m3)
```

```
m4 <- clm(spot_quality ~ int_season_sitetype, data = d0, Hess = TRUE)
summary(m4)

# All variables without interaction term (site type * season):
m5 <- clm(spot_quality ~
        + site_type
        + season
        + spot_diameter
        + spot_depth
        + I(spot_depth^2)
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_logging_residues
        + note_dummy_5_water
        + note_dummy_6_humus
        + note_dummy_7_logging_trail
        , data = d0, Hess = TRUE)
summary(m5)

# All variables with interaction term (site type * season):
m6 <- clm(spot_quality ~
        + int_season_sitetype
        + spot_diameter
        + spot_depth
        + I(spot_depth^2)
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_logging_residues
        + note_dummy_5_water
        + note_dummy_6_humus
        + note_dummy_7_logging_trail
        , data = d0, Hess = TRUE)
summary(m6)

# Without spot depth and diameter:
m66 <- clm(spot_quality ~
        + int_season_sitetype
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_logging_residues
        + note_dummy_5_water
        + note_dummy_6_humus
        + note_dummy_7_logging_trail
        , data = d0, Hess = TRUE)
summary(m66)

# A comparison of the nested models
anova(m66,m6)
```

```
# Without spot diameter:
m666 <- clm(spot_quality ~
        + int_season_sitetype
        + spot_depth
        + I(spot_depth^2)
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_logging_residues
        + note_dummy_5_water
        + note_dummy_6_humus
        + note_dummy_7_logging_trail
        , data = d0, Hess = TRUE)
summary(m666)

# A comparison of the nested models
anova(m666, m6)

# Without spot depth:
m6666 <- clm(spot_quality ~
        + int_season_sitetype
        + spot_diameter
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_logging_residues
        + note_dummy_5_water
        + note_dummy_6_humus
        + note_dummy_7_logging_trail
        , data = d0, Hess = TRUE)
summary(m6666)

# A comparison of the nested models
anova(m6666, m6)

# The diameter and depth of site preparation spot should be included in the model

# Without logging trail (only 12 observations):
m7 <- clm(spot_quality ~
        + int_season_sitetype
        + spot_diameter
        + spot_depth
        + I(spot_depth^2)
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_logging_residues
        + note_dummy_5_water
        + note_dummy_6_humus
        #+ note_dummy_7_logging_trail
        , data = d0, Hess = TRUE)
```

```
summary(m7)

anova(m7, m6)

Anova(m7, type = "III")

# Several observations had extreme values for planting spot dimensions (depth and diameter).
# Next, we examine whether these outliers affected the model estimates.

# Excluding observations with depth > 30 cm or diameter > 75 cm or < 20 cm:
d00 <- subset(d0, spot_depth < 30 & spot_diameter < 75 & spot_diameter > 20)

# N = 4092

str(d00)

# Boxplots and violin plots (for the depth and diameter of planting spots):
boxplot(d00$spot_depth~d00$spot_quality)
boxplot(d00$spot_diameter~d00$spot_quality)
vioplot(d00$spot_depth~d00$spot_quality)
vioplot(d00$spot_diameter~d00$spot_quality)

# Interaction term (season * site type)
int_season_sitetype <- as.factor(interaction(d00$season, d00$site_type, drop=TRUE ))
summary(int_season_sitetype)

# Model m7 fitted without extreme values
m77 <- clm(spot_quality ~
        + int_season_sitetype
        + spot_diameter
        + spot_depth
        + I(spot_depth^2)
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_logging_residues
        + note_dummy_5_water
        + note_dummy_6_humus
        #+ note_dummy_7_logging_trail
        , data = d00, Hess = TRUE)
summary(m77)

# The parameter estimates changed slightly.
# However, all key results remained unchanged.
# Retaining the outliers in the dataset is therefore justified.


# CONCLUSIONS

# The cumulative link model (m7) was the final model for planting spot quality.
```

```
# The poorest spot quality was predicted for early summer site preparation.
# The spots with the highest probability of being good quality were estimated
# for midsummer and late summer site preparation at the MT and OMT sites.
# The model also estimated that when the depth or diameter of a spot increased,
# the probability of it belonging to a lower spot quality class decreased.
# As spot depth increased by 10–20 cm below ground level,
# the positive effect began to decline.
# This decline was due to a second-order polynomial for the effect of depth,
# which in turn reduced the probability of spots belonging to higher-quality classes.
# All dummy variables included in the model,
# especially notes related to water, logging residues, and humus layer residues,
# reduced the odds of belonging to a good spot quality class.

# Note.
# The variable "logging trail" was a significant explanatory variable in model m6.
# However, there were only 12 observations for logging trail (0.3%).
# Moreover, logging trail is not a particularly informative variable in the context of this analysis.
# Consequently, it was excluded from the final model (m7).


#################################################################################
#
# Drill-based site preparation - Seedling vitality
#
#################################################################################

# Loading data (dataset_drill_2023_2024_seedling_vitality_height)

d <- fread("x:/xxx/data.csv", sep = ";")

# Excluding seedlings which were loose from the ground (N = 7):
d <- subset(d, note_dummy_8_loose_seedling != 1)

# Getting familiar with the data

#View(d)
#str(d)

# Convert measurement year as well as site and plot identification number to factors:
d$year <- as.factor(d$year)
levels(d$year)

d$site_id <- as.factor(d$site_id)
levels(d$site_id)

d$int_year_site_id <- as.factor(interaction(d$year, d$site_id, drop=TRUE ))
summary(d$int_year_site_id)

d$plot_id <- as.factor(d$plot_id)
```

```
levels(d$plot_id)
table(d$plot_id, d$int_year_site_id)

# Note that not all seedlings were found in the 2024 follow-up measurement (285 missing seedlings).

# Convert all categorical variables to factors

d$seedling_vitality <- recode_factor(d$seedling_vitality, "1" = "Dead", "2" = "Weakened", "3" = "Healthy")
levels(d$seedling_vitality)
table(d$seedling_vitality, d$site_id, d$year)

d$tree_species <- recode_factor(d$tree_species, "1" = "NS", "2" = "SP")
levels(d$tree_species)
table(d$seedling_vitality, d$tree_species, d$year)

d$site_type <- recode_factor(d$site_type, "2" = "MT", "3" = "OMT", "1" = "VT")
levels(d$site_type)
table(d$seedling_vitality, d$site_type, d$year)
table(d$tree_species,d$site_type,d$year)

d$season <- recode_factor(d$season, "1" = "Early", "2" = "Mid", "3" = "Late")
levels(d$season)

d$seedling_position_1 <- recode_factor(d$seedling_position, "1" = "Middle", "2" = "Side", "3" = "Outside")
table(d$seedling_vitality, d$seedling_position_1, d$year)

d$planting_depth_1 <- recode_factor(d$planting_depth, "1" = "Proper", "2" = "Shallow", "3" = "Deep", "0" = "Unstable seedling")
table(d$seedling_vitality, d$planting_depth_1, d$year)

d$seedling_compaction_1 <- recode_factor(d$seedling_compaction, "1" = "Compacted", "2" = "Uncompacted", "0" = "Unstable seedling")
table(d$seedling_vitality, d$seedling_compaction_1, d$year)

d$seedling_angle_1 <- recode_factor(d$seedling_angle, "1" = "Upright", "2" = "Askew", "3" = "Prone", "0" = "Unstable seedling")
table(d$seedling_vitality, d$seedling_angle_1, d$year)

d$spot_type_1 <- recode_factor(d$spot_type, "1" = "Hole", "2" = "Elevated", "3" = "Flat")
table(d$spot_type_1, d$year)

# Convert notes to factors:

# Create dummy variable "stump or root":
d$note_dummy_1_stump_2_root <- ifelse(d$note_dummy_1_stump == 1 | d$note_dummy_2_root == 1, 1, 0)
d$note_dummy_1_stump_2_root <- as.factor(d$note_dummy_1_stump_2_root)
table(d$note_dummy_1_stump_2_root, d$year)

d$note_dummy_3_stone <- as.factor(d$note_dummy_3_stone)
table(d$note_dummy_3_stone, d$year)

d$note_dummy_4_water <- as.factor(d$note_dummy_4_water)
```

```
table(d$note_dummy_4_water, d$year)

d$note_dummy_5_humus <- as.factor(d$note_dummy_5_humus)
table(d$note_dummy_5_humus, d$year)

d$note_dummy_6_peat <- as.factor(d$note_dummy_6_peat)
table(d$note_dummy_6_peat, d$year)

d$note_dummy_7_tree_top_browsing <- as.factor(d$note_dummy_7_tree_top_browsing)
table(d$note_dummy_7_tree_top_browsing, d$year)

d$note_dummy_9_logging_residues <- as.factor(d$note_dummy_9_logging_residues)
table(d$note_dummy_9_logging_residues, d$year)

d$note_dummy_10_tree_top_breakage <- as.factor(d$note_dummy_10_tree_top_breakage)
table(d$note_dummy_10_tree_top_breakage, d$year)

d$note_dummy_11_brush <- as.factor(d$note_dummy_11_brush)
table(d$note_dummy_11_brush, d$year)

d$note_dummy_12_logging_trail <- as.factor(d$note_dummy_12_logging_trail)
table(d$note_dummy_12_logging_trail, d$year)

# Interaction term (year * planting season * tree species * site type):
int_year_season_treespecies_sitetype <- as.factor(interaction(d$year, d$season, d$tree_species, d$site_type, drop=TRUE ))
summary(int_year_season_treespecies_sitetype)

#str(d)

# Models with a random effect for sample plot (N = 60) using clmm2

# Let's check how we should interpret the coefficients:

m0 <- clmm2(seedling_vitality ~ 1, random = plot_id, data = d, Hess = TRUE)
summary(m0)

m1 <- clmm2(seedling_vitality ~ year, random = plot_id, data = d, Hess = TRUE)
summary(m1)

m2 <- clmm2(seedling_vitality ~ season, random = plot_id, data = d, Hess = TRUE)
summary(m2)

m3 <- clmm2(seedling_vitality ~ tree_species, random = plot_id, data = d, Hess = TRUE)
summary(m3)

m4 <- clmm2(seedling_vitality ~ site_type, random = plot_id, data = d, Hess = TRUE)
summary(m4)

# Interaction term (year * planting season * tree species * site type):
```

```
m5 <- clmm2(seedling_vitality ~ int_year_season_treespecies_sitetype, random = plot_id, data = d, Hess = TRUE)
summary(m5)

# Interaction term (year * planting season * tree species * site type) and other variables:
m6 <- clmm2(seedling_vitality ~
        + int_year_season_treespecies_sitetype
        + seedling_position_1
        + planting_depth_1
        + seedling_compaction_1
        + seedling_angle_1
        + spot_type_1
        + spot_diameter
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_water
        + note_dummy_5_humus
        + note_dummy_6_peat
        + note_dummy_7_tree_top_browsing
        # note_dummy_8_loose_seedling      # excluded
        + note_dummy_9_logging_residues
        + note_dummy_10_tree_top_breakage
        + note_dummy_11_brush
        + note_dummy_12_logging_trail
        , random = plot_id, data = d, Hess = TRUE)
summary(m6)

# Variables with p-values close to the significance threshold (p < 0.1):

m66 <- clmm2(seedling_vitality ~
        + int_year_season_treespecies_sitetype
        #+ seedling_position_1
        #+ planting_depth_1
        + seedling_compaction_1
        + seedling_angle_1
        + spot_type_1
        #+ spot_diameter
        #+ note_dummy_1_stump_2_root
        #+ note_dummy_3_stone
        + note_dummy_4_water
        #+ note_dummy_5_humus
        #+ note_dummy_6_peat
        + note_dummy_7_tree_top_browsing
        # note_dummy_8_loose_seedling      # excluded
        #+ note_dummy_9_logging_residues
        #+ note_dummy_10_tree_top_breakage
        #+ note_dummy_11_brush
        #+ note_dummy_12_logging_trail
        , random = plot_id, data = d, Hess = TRUE)
summary(m66)
```

```
# A comparison of the nested models
anova(m66, m6)

# Significant variables (p < 0.05):

m666 <- clmm2(seedling_vitality ~
        + int_year_season_treespecies_sitetype
        #+ seedling_position_1
        #+ planting_depth_1
        + seedling_compaction_1
        + seedling_angle_1
        #+ spot_type_1
        #+ spot_diameter
        #+ note_dummy_1_stump_2_root
        #+ note_dummy_3_stone
        + note_dummy_4_water
        #+ note_dummy_5_humus
        #+ note_dummy_6_peat
        + note_dummy_7_tree_top_browsing
        # note_dummy_8_loose_seedling    # excluded
        #+ note_dummy_9_logging_residues
        #+ note_dummy_10_tree_top_breakage
        #+ note_dummy_11_brush
        #+ note_dummy_12_logging_trail
        , random = plot_id, data = d, Hess = TRUE)
summary(m666)

# A comparison of the nested models
anova(m666, m6)
anova(m666, m66)

# Without interaction term (year * planting season * tree species * site type):

m6666 <- clmm2(seedling_vitality ~
        #+ int_year_season_treespecies_sitetype
        #+ seedling_position_1
        #+ planting_depth_1
        + seedling_compaction_1
        + seedling_angle_1
        #+ spot_type_1
        #+ spot_diameter
        #+ note_dummy_1_stump_2_root
        #+ note_dummy_3_stone
        + note_dummy_4_water
        #+ note_dummy_5_humus
        #+ note_dummy_6_peat
        + note_dummy_7_tree_top_browsing
        # note_dummy_8_loose_seedling    # excluded
```

```
    #+ note_dummy_9_logging_residues
    #+ note_dummy_10_tree_top_breakage
    #+ note_dummy_11_brush
    #+ note_dummy_12_logging_trail
    , random = plot_id, data = d, Hess = TRUE)
summary(m6666)

# A comparison of the nested models
anova(m6666, m666)

# Visualizing the random effects of sample plot level (N = 60)
# See: https://petrilankoski.com/2013/01/05/r-ordinal-scrips-part-2/

levels(d$plot_id)
raneff <- m666$ranef
rnd <- m666$condVar
ci <- m666$ranef + qnorm(0.975) * sqrt(rnd) %o% c(-1,1)
alat <- cbind(levels(d$plot_id),raneff,ci)

tata <- data.frame(cbind(ci,raneff))

ord.re <- tata[order(raneff),]

alat.re <- alat[order(raneff),]

n <- length(raneff)

plot(1:n, ord.re$raneff, axes=FALSE, ylim = c(-3,3), ylab = "Random effect", xlab = "Plot")
axis(1, at=1:n, labels=alat.re[,1], las=1)
axis(2)
for(i in 1:n) {
  segments(i, ord.re[i,1], i, ord.re[i,2])
  abline(h=0, lty=2)
  }

# CONCLUSIONS

# The cumulative link mixed model (m666) was the final model for seedling vitality.
# The model predicted that the more askew a seedling was,
# the more likely it was to fall into a lower vitality class.
# Moreover, poor soil compaction around the planted seedling,
# notes indicating seedlings with eaten tops,
# and the presence of water in site preparation spots
# reduced the likelihood of belonging to a healthy vitality class.

# Note.
# More than half of the individual interaction coefficients
# between tree species, site types and measurement years were not statistically significant,
# although the interaction term was significant overall.
```

```
###############################################################################
#
# Drill-based site preparation - Height of Norway spruce seedlings
#
###############################################################################

# Loading data (dataset_drill_2023_2024_seedling_vitality_height)

dd <- fread("x:/xxx/data.csv", sep = ";")

# Excluding (Norway spruce) seedlings which were dead, whose top was broken or eaten, and which were loose from the ground:
d <- subset(dd, tree_species != 2 & seedling_vitality != 1 & note_dummy_7_tree_top_browsing == 0 & note_dummy_8_loose_seedling == 0 & note_dummy_10_tree_top_breakage == 0)

# Getting familiar with the data

#View(d)
#str(d)

# Convert measurement year as well as site and plot identification number to factors:
d$year <- as.factor(d$year)
levels(d$year)

d$site_id <- as.factor(d$site_id)
levels(d$site_id)

d$plot_id <- as.factor(d$plot_id)
levels(d$plot_id)

# Convert all categorical variables to factors:

d$seedling_vitality <- recode_factor(d$seedling_vitality, "2" = "Weakened", "3" = "Alive")
table(d$year, d$seedling_vitality)

d$seedling_position_1 <- recode_factor(d$seedling_position, "1" = "Middle", "2" = "Side", "3" = "Outside")
table(d$year, d$seedling_position_1)

d$planting_depth_1 <- recode_factor(d$planting_depth, "1" = "Proper", "2" = "Shallow", "3" = "Deep")
table(d$year, d$planting_depth_1)

d$seedling_compaction_1 <- recode_factor(d$seedling_compaction, "1" = "Compacted", "2" = "Uncompacted")
table(d$year, d$seedling_compaction_1)

d$seedling_angle_1 <- recode_factor(d$seedling_angle, "1" = "Upright", "2" = "Askew", "3" = "Prone")
table(d$year, d$seedling_angle_1)

d$spot_type_1 <- recode_factor(d$spot_type, "1" = "Hole", "2" = "Elevated", "3" = "Flat")
table(d$year, d$spot_type_1)

# Convert all notes to factors:
```

```
d$note_dummy_1_stump_2_root <- ifelse(d$note_dummy_1_stump == 1 | d$note_dummy_2_root == 1, 1, 0)
d$note_dummy_1_stump_2_root <- as.factor(d$note_dummy_1_stump_2_root)
table(d$note_dummy_1_stump_2_root, d$year)

d$note_dummy_3_stone <- as.factor(d$note_dummy_3_stone)
table(d$note_dummy_3_stone, d$year)

d$note_dummy_4_water <- as.factor(d$note_dummy_4_water)
table(d$note_dummy_4_water, d$year)

d$note_dummy_5_humus <- as.factor(d$note_dummy_5_humus)
table(d$note_dummy_5_humus, d$year)

# Note. There were no observations for peat in this subset
d$note_dummy_6_peat <- as.factor(d$note_dummy_6_peat)
table(d$note_dummy_6_peat, d$year)

d$note_dummy_7_tree_top_browsing <- as.factor(d$note_dummy_7_tree_top_browsing)
table(d$note_dummy_7_tree_top_browsing, d$year)

d$note_dummy_9_logging_residues <- as.factor(d$note_dummy_9_logging_residues)
table(d$note_dummy_9_logging_residues, d$year)

d$note_dummy_10_tree_top_breakage <- as.factor(d$note_dummy_10_tree_top_breakage)
table(d$note_dummy_10_tree_top_breakage, d$year)

d$note_dummy_11_brush <- as.factor(d$note_dummy_11_brush)
table(d$note_dummy_11_brush, d$year)

d$note_dummy_12_logging_trail <- as.factor(d$note_dummy_12_logging_trail)
table(d$note_dummy_12_logging_trail, d$year)

# Convert tree species, site types and seasons to factors
d$tree_species <- as.factor(d$tree_species)
d$tree_species <- recode_factor(d$tree_species, "1" = "NS")
levels(d$tree_species)

d$site_type <- as.factor(d$site_type)
d$site_type <- recode_factor(d$site_type, "2" = "MT", "3" = "OMT")
levels(d$site_type)

d$season <- as.factor(d$season)
d$season <- recode_factor(d$season, "1" = "Early", "2" = "Mid", "3" = "Late")
levels(d$season)

# Details of the heights of Norway spruce seedlings (mean, median, etc.)
summary(d$seedling_height[d$tree_species== "NS"])
```

```
# The height distribution of Norway spruce seedlings
hist(d$seedling_height[d$tree_species== "NS"])

# The mean heights of Norway spruce seedlings (ggplots, vioplots, and boxplots) across categorical variables (year, season, site type, and vitality)

vioplot(d$seedling_height~as.factor(interaction(d$year, d$site_type, drop=TRUE)))
boxplot(d$seedling_height~as.factor(interaction(d$year, d$site_type, drop=TRUE)))

vioplot(d$seedling_height~as.factor(interaction(d$year, d$season, drop=TRUE)))
boxplot(d$seedling_height~as.factor(interaction(d$year, d$season, drop=TRUE)))

vioplot(d$seedling_height~as.factor(interaction(d$year, d$seedling_vitality, drop=TRUE)))
boxplot(d$seedling_height~as.factor(interaction(d$year, d$seedling_vitality, drop=TRUE)))

vioplot(d$seedling_height~as.factor(interaction(d$year, d$site_type, d$seedling_vitality, drop=TRUE)))
boxplot(d$seedling_height~as.factor(interaction(d$year, d$site_type, d$seedling_vitality, drop=TRUE)))

vioplot(d$seedling_height~as.factor(interaction(d$year, d$season, d$seedling_vitality, drop=TRUE)))
boxplot(d$seedling_height~as.factor(interaction(d$year, d$season, d$seedling_vitality, drop=TRUE)))

str(d)

# Next, we fit a linear mixed-effects model (LME) for the height of Norway spruce seedlings
# The LME is fitted using the function lme from the nlme package

# Let's check how we should interpret the coefficients

# Intercept-only model
m0 <- lme(seedling_height ~ 1
        ,random=~1|plot_id, data=d)
summary(m0)

# LME with seedling-level variables
m1 <- lme(seedling_height ~
        #+ year*season
        #+ year*site_type
        #+ season*site_type
        + seedling_position_1
        + planting_depth_1
        + seedling_compaction_1
        + seedling_angle_1
        + seedling_vitality
        + spot_type_1
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_water
        + note_dummy_5_humus
        #+ note_dummy_6_peat                 # no observations
        #+ note_dummy_7_tree_top_browsing    # excluded
```

```
        # note_dummy_8_loose_seedling             # excluded
        + note_dummy_9_logging_residues
        #+ note_dummy_10_tree_top_breakage    # excluded
        + note_dummy_11_brush
        + note_dummy_12_logging_trail
        , random=~1|plot_id, data=d)
summary(m1)

anova(m1, type = "marginal")  # Type-III-like

# LME with plot-level and seedling-level variables (all variables)
m2 <- lme(seedling_height ~
        + year*season
        + year*site_type
        + season*site_type
        + seedling_position_1
        + planting_depth_1
        + seedling_compaction_1
        + seedling_angle_1
        + seedling_vitality
        + spot_type_1
        + note_dummy_1_stump_2_root
        + note_dummy_3_stone
        + note_dummy_4_water
        + note_dummy_5_humus
        #+ note_dummy_6_peat                          # no observations
        #+ note_dummy_7_tree_top_browsing      # excluded
        # note_dummy_8_loose_seedling             # excluded
        + note_dummy_9_logging_residues
        #+ note_dummy_10_tree_top_breakage    # excluded
        + note_dummy_11_brush
        + note_dummy_12_logging_trail
         , random=~1|plot_id, data=d)
summary(m2)

anova(m2, type = "marginal")  # Type-III-like

# Updating estimation using the maximum likelihood method
m1_ml <- update(m1, method="ML")
m2_ml <- update(m2, method="ML")

# A comparison of the nested models
anova(m1_ml, m2_ml)

# Significant variables (p < 0.05)

m3 <- lme(seedling_height ~
        + year*season
        + year*site_type
```

```
        + season*site_type
        + seedling_position_1
        #+ planting_depth_1
        + seedling_compaction_1
        #+ seedling_angle_1
        + seedling_vitality
        #+ spot_type_1
        #+ note_dummy_1_stump_2_root
        #+ note_dummy_3_stone
        #+ note_dummy_4_water
        #+ note_dummy_5_humus
        #+ note_dummy_6_peat                    # no observations
        #+ note_dummy_7_tree_top_browsing       # excluded
        # note_dummy_8_loose_seedling           # excluded
        #+ note_dummy_9_logging_residues
        #+ note_dummy_10_tree_top_breakage      # excluded
        #+ note_dummy_11_brush
        + note_dummy_12_logging_trail
        , random=~1|plot_id, data=d)
summary(m3)

anova(m3, type = "marginal")  # Type-III-like

# Updating estimation using the maximum likelihood method
m3_ml <- update(m3, method="ML")

# A comparison of the nested models
anova(m3_ml, m2_ml)

# Significant and relevant variables

m4 <- lme(seedling_height ~
        + year*season
        + year*site_type
        + season*site_type
        #+ seedling_position_1
        #+ planting_depth_1
        #+ seedling_compaction_1
        #+ seedling_angle_1
        + seedling_vitality
        #+ spot_type_1
        #+ note_dummy_1_stump_2_root
        #+ note_dummy_3_stone
        #+ note_dummy_4_water
        #+ note_dummy_5_humus
        #+ note_dummy_6_peat                    # no observations
        #+ note_dummy_7_tree_top_browsing       # excluded
        # note_dummy_8_loose_seedling           # excluded
        #+ note_dummy_9_logging_residues
```

```
        #+ note_dummy_10_tree_top_breakage    # excluded
        #+ note_dummy_11_brush
        #+ note_dummy_12_logging_trail
        , random=~1|plot_id, data=d)
summary(m4)

anova(m4, type = "marginal")  # Type-III-like

# Updating estimation using the maximum likelihood method
m4_ml <- update(m4, method="ML")

# A comparison of the nested models
anova(m4_ml, m3_ml)

# Checking residuals (residuals vs. fitted) for evaluating heteroscedasticity:

# Pearson residuals vs. Fitted
plot(predict(m4), residuals(m4, type="pearson"),
    main = "Residuals vs. Fitted",
    xlab = "Fitted values",
    ylab = "Pearson residual")
abline(h = 0, col = "red")
mywhiskers(predict(m4), residuals(m4, type="pearson"), add=TRUE, lwd=2,se=FALSE)

# Raw residuals vs. Fitted
plot(predict(m4), residuals(m4),
    main = "Residuals vs. Fitted",
    xlab = "Fitted values",
    ylab = "Raw residuals")
abline(h = 0, col = "red")
mywhiskers(predict(m4), residuals(m4), add=TRUE, lwd=2,se=FALSE)

# Variance seems to increase slightly as fitted values increase
# Let's try to reduce heteroskedasticity using a power variance function

# LME using variance function varPower()
m5 <- lme(seedling_height ~
        + year*season
        + year*site_type
        + season*site_type
        #+ seedling_position_1
        #+ planting_depth_1
        #+ seedling_compaction_1
        #+ seedling_angle_1
        + seedling_vitality
        #+ spot_type_1
        #+ note_dummy_1_stump_2_root
        #+ note_dummy_3_stone
        #+ note_dummy_4_water
```

```
        #+ note_dummy_5_humus
        #+ note_dummy_6_peat                    # no observations
        #+ note_dummy_7_tree_top_browsing       # excluded
        # note_dummy_8_loose_seedling           # excluded
        #+ note_dummy_9_logging_residues
        #+ note_dummy_10_tree_top_breakage      # excluded
        #+ note_dummy_11_brush
        #+ note_dummy_12_logging_trail
        , random=~1|plot_id, data=d, weights=varPower())
summary(m5)

anova(m5, type = "marginal")  # Type-III-like

# Model m5 has a lower AIC than m4 (8116 < 8166)

# Let's check residuals again (m5):

# Pearson residuals vs. Fitted
plot(predict(m5), residuals(m5, type="pearson"),
    main = "Residuals vs. Fitted",
    xlab = "Fitted values",
    ylab = "Pearson residual")
abline(h = 0, col = "red")
mywhiskers(predict(m5), residuals(m5, type="pearson"), add=TRUE, lwd=2,se=FALSE)

# Residuals by categorical variables:

plot(d$season,resid(m5, type="pearson"),
    main = "Residuals vs. Season",
    xlab="Season", ylab="Pearson residual")
abline(h = 0, col = "red")

plot(d$site_type,resid(m5, type="pearson"),
    main = "Residuals vs. Site type",
    xlab="Site type", ylab="Pearson residual")
abline(h = 0, col = "red")

plot(d$year,resid(m5, type="pearson"),
    main = "Residuals vs. Year",
    xlab="Year", ylab="Pearson residual")
abline(h = 0, col = "red")

plot(interaction(d$year, d$site_type), resid(m5, type="pearson"),
    main = "Residuals vs. Year * Site type",
    xlab="Year * Site type", ylab="Pearson residual")
abline(h = 0, col = "red")

plot(interaction(d$year, d$season), resid(m5, type="pearson"),
    main = "Residuals vs. Year * Season",
```

```
    xlab="Year * Season", ylab="Pearson residual")
abline(h = 0, col = "red")

plot(interaction(d$site_type, d$season), resid(m5, type="pearson"),
    main = "Residuals vs. Site type * Season",
    xlab="Site type * Season", ylab="Pearson residual")
abline(h = 0, col = "red")

plot(d$seedling_vitality, resid(m5, type="pearson"),
    main = "Residuals vs. Seedling vitality",
    xlab="Seedling vitality", ylab="Pearson residual")
abline(h = 0, col = "red")

# Residuals by plot ID
boxplot(residuals(m5, type="pearson") ~ d$plot_id,
        main = "Residuals vs. Sample plot",
        xlab="Sample plot", ylab="Pearson residual",
        las = 2)
abline(h = 0, col = "red")

# Normal Q-Q Plot
qqnorm(residuals(m5, type = "pearson"))
qqline(residuals(m5, type = "pearson"))

# Now the residuals seem to be OK

# CONCLUSIONS

# The linear mixed model including a power variance function
# was the final model (m5) for the height of Norway spruce seedlings.
# The two-level interactions Year × Season and Year × Site type,
# as well as seedling vitality and planting position,
# were significant and relevant predictors.
# The residuals of the model (m5) appeared satisfactory
# in terms of both homoscedasticity and normality.

# Note.
# The variables "seedling compaction", "seedling position", and "logging trail"
# were also significant predictors, but they were excluded from the final model.
# For example, some categories contained only a small number of observations,
# which may result in unreliable estimates and misleading interpretations.
```