

Modelling Maize Project: days.to.germination

Modelling with smooth effect

Author: Nisia Trisconi | Zurich Data Scientists
Reviewer: Dr. Luisa Barbanti | Zurich Data Scientists

October 30, 2023

Contents

1	Freezing Package versions	2
2	Load packages	2
3	Settings	2
4	Getting data	3
5	Design	5
6	Response variable: <i>days.to.germination</i>	5
6.1	Aim	5
6.2	Model fitting	5
6.3	Model selection	7
6.4	Model checking	11
7	Methods description	12
8	Session information	14

1 Freezing Package versions

The following code lines are commented out because the `{checkpoint}` package no longer works.

```
## (messages are omitted in this chunk)
##
# library(checkpoint)
# checkpoint(snapshot_date = "2022-11-15")
```

2 Load packages

```
## (messages are omitted from this chunk)
##
library(dplyr)
library(kableExtra)
library(ggplot2)
library(tibble) ## function rownames_to_column()
library(multcomp)
library(lme4)
library(mgcv)
# library(AICcmodavg)
```

3 Settings

Global settings:

```
Sys.setenv(lang = "en_US")
theme_set(theme_bw())

if (!dir.exists("Prepared_data_and_models")) {
  dir.create("Prepared_data_and_models")
}
```

4 Getting data

```
d.maize <- readRDS(file = paste0("Prepared_data_and_models/",  
                                "d.maize_PreparedData.RDS"))
```

Overview of the data:

```
dim(d.maize)
```

```
[1] 108 33
```

```
head(d.maize)[1:min(ncol(d.maize), 30)]
```

```
# A tibble: 6 x 30  
  pot    soil      well depth seed.weight fungus date.germinated observations  
  <chr> <chr>    <chr> <dbl>    <dbl> <chr>    <chr>          <chr>  
1 A1    Bio garden a      3      30 <NA>    2022-05-11    <NA>  
2 A1    Bio garden b      5      34 <NA>    2022-05-11    <NA>  
3 A1    Bio garden c      2      35 <NA>    2022-05-09    <NA>  
4 A1    Bio garden d      1      40 <NA>    2022-05-10    <NA>  
5 A1    Bio garden e      4      46 <NA>    2022-05-11    <NA>  
6 A1    Bio garden f      6      37 <NA>    2022-05-11    <NA>  
# i 22 more variables: height_2022_07_05 <chr>, cob_weight <chr>, ...12 <dbl>,  
# pot.fac <fct>, soil.fac <fct>, well.fac <fct>, seed.weight.grams <dbl>,  
# fungus.fac <fct>, date.germinated.asDate <date>, obs.time <fct>,  
# broken <lgl>, height_2022_07_05.num <dbl>, plant.found <lgl>,  
# cob_weight.num <dbl>, germinated.in.lab <lgl>, germinated.in.field <lgl>,  
# germinated.yes <lgl>, days.to.germination <dbl>,  
# days.to.germination.censored <dbl>, seed_coord_y <dbl>, ...
```

```
str(d.maize)
```

```
tibble [108 x 33] (S3: tbl_df/tbl/data.frame)  
$ pot           : chr [1:108] "A1" "A1" "A1" "A1" ...  
$ soil          : chr [1:108] "Bio garden" "Bio garden" "Bio garden" "Bio garden" ...  
$ well          : chr [1:108] "a" "b" "c" "d" ...  
$ depth         : num [1:108] 3 5 2 1 4 6 6 4 5 1 ...  
$ seed.weight   : num [1:108] 30 34 35 40 46 37 27 16 23 22 ...  
$ fungus        : chr [1:108] NA NA NA NA ...  
$ date.germinated : chr [1:108] "2022-05-11" "2022-05-11" "2022-05-09" "2022-05-10" ...  
$ observations   : chr [1:108] NA NA NA NA ...  
$ height_2022_07_05 : chr [1:108] "217" "131" "143" "194" ...  
$ cob_weight     : chr [1:108] "117" "26" "61" "109" ...  
$ ...12         : num [1:108] NA NA NA NA NA NA NA NA NA NA ...  
$ pot.fac       : Factor w/ 18 levels "A1","A2","A3",...: 1 1 1 1 1 1 2 2 2 2 ...  
$ soil.fac      : Factor w/ 4 levels "Bio garden","Composana",...: 1 1 1 1 1 1 3 3 3 3 ..  
$ well.fac      : Factor w/ 6 levels "a","b","c","d",...: 1 2 3 4 5 6 1 2 3 4 ...  
$ seed.weight.grams : num [1:108] 0.3 0.34 0.35 0.4 0.46 0.37 0.27 0.16 0.23 0.22 ...  
$ fungus.fac    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
$ date.germinated.asDate : Date[1:108], format: "2022-05-11" "2022-05-11" ...  
$ obs.time      : Factor w/ 2 levels "morning","night": 2 2 2 2 2 2 2 2 2 2 ...  
$ broken        : logi [1:108] FALSE FALSE FALSE FALSE FALSE FALSE ...  
$ height_2022_07_05.num : num [1:108] 217 131 143 194 206 233 158 282 241 232 ...  
$ plant.found    : logi [1:108] TRUE TRUE TRUE TRUE TRUE TRUE ...  
$ cob_weight.num : num [1:108] 117 26 61 109 106 156 57 286 51 120 ...  
$ germinated.in.lab : logi [1:108] TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```

$ germinated.in.field      : logi [1:108] FALSE FALSE FALSE FALSE FALSE FALSE ...
$ germinated.yes          : logi [1:108] TRUE TRUE TRUE TRUE TRUE TRUE ...
$ days.to.germination      : num [1:108] 11 11 9 10 11 11 11 11 NA 9 ...
$ days.to.germination.censored: num [1:108] 11 11 9 10 11 11 11 11 14 9 ...
$ seed_coord_y            : num [1:108] 1 1 2 2 3 3 1 1 2 2 ...
$ seed_coord_x            : num [1:108] 1 2 1 2 1 2 3 4 3 4 ...
$ position_field_x        : num [1:108] 1 1 1 1 1 1 1 1 1 1 ...
$ position_field_x_cm     : num [1:108] 50 50 50 50 50 50 50 50 50 50 ...
$ position_field_y        : int [1:108] 1 2 3 4 5 6 7 8 9 10 ...
$ position_field_y_cm     : num [1:108] 25 50 75 100 125 150 175 200 225 250 ...

```

5 Design

108 maize seeds are planted in 18 different pots, each with 6 wells.

Inside one pot, the same soil is used. The soils that were used are: Bio garden (4 pots), Composana (4 pots), herbs (6 pots), mixture (4 pots).

In each well, one maize seed is planted at a pre-defined depth (in cm), which is allocated randomly to the well. The maximum value for depth is 6cm and this corresponds to planting the seed directly in the coconut fiber that makes up the pot.

Wells in the same pots are allocated as follows:

```
[,1] [,2]
[1,] "e" "f"
[2,] "c" "d"
[3,] "a" "b"
```

The pots are arranged as follows on a table in the lab:

```
[,1] [,2] [,3] [,4] [,5] [,6]
[1,] "C1" "C2" "C3" "C4" "C5" "C6"
[2,] "B1" "B2" "B3" "B4" "B5" "B6"
[3,] "A1" "A2" "A3" "A4" "A5" "A6"
```

Seeds are watered for the first time on 04.30.2022 and are transferred to the field on 05.15.2022 according to the same scheme.

Some seeds are broken when planted, one seed develops a fungus.

Some seeds germinate in the lab, others in the field, while some seeds never germinate.

On 07.05.2022, the height of all maize plants is measured in cm. The plants that were not measured for time reasons receive a height value of “not measured”, while the plants that were not found and could hence not be measured present missing values.

On 09.16.2022, the weight of the cob is measured for all plants that have a cob. The variety of maize that was planted typically yields 1 cob per plant.

6 Response variable: *days.to.germination*

6.1 Aim

We are interested in testing whether *days.to.germination* is influenced by:

- Position in the lab (i.e. *seed_coord_x* and *seed_coord_y*)
- Soil (variable *soil.fac*)
- Depth in soil (variable *depth*)
- Seed weight (variable *seed.weight*)

6.2 Model fitting

The variable *days.to.germination* is a continuous variable (in particular it is an amount), whose density is already well-centered, so there is no need to log transform it.

Based on the graphical analysis, it is not necessary to transform any of the explanatory variables. Therefore, they can be used as they are in the model.

depth and *seed.weight* are numeric variables which are introduced in the model as continuous variables.

seed_coord_x and *seed_coord_y* are numeric variables which are introduced in the model as smooth variables.

For this reason, a generalised additive model is fitted.

```
gam.days.to.germination <- gam(days.to.germination ~
                                s(seed_coord_x,
                                  seed_coord_y) +
                                soil.fac +
                                depth +
                                seed.weight,
                                data = d.maize)

##
summary(gam.days.to.germination)
```

Family: gaussian

Link function: identity

Formula:

```
days.to.germination ~ s(seed_coord_x, seed_coord_y) + soil.fac +
                        depth + seed.weight
```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.1644790	0.7335842	12.4927	< 2.2e-16 ***
soil.facComposana	-0.4544784	0.4915974	-0.9245	0.3586
soil.facherbs	-0.1548137	0.4318388	-0.3585	0.7211
soil.facmixture	-0.6941419	0.5245769	-1.3232	0.1903
depth	0.6698200	0.0690838	9.6958	2.547e-14 ***
seed.weight	-0.0081194	0.0203995	-0.3980	0.6919

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

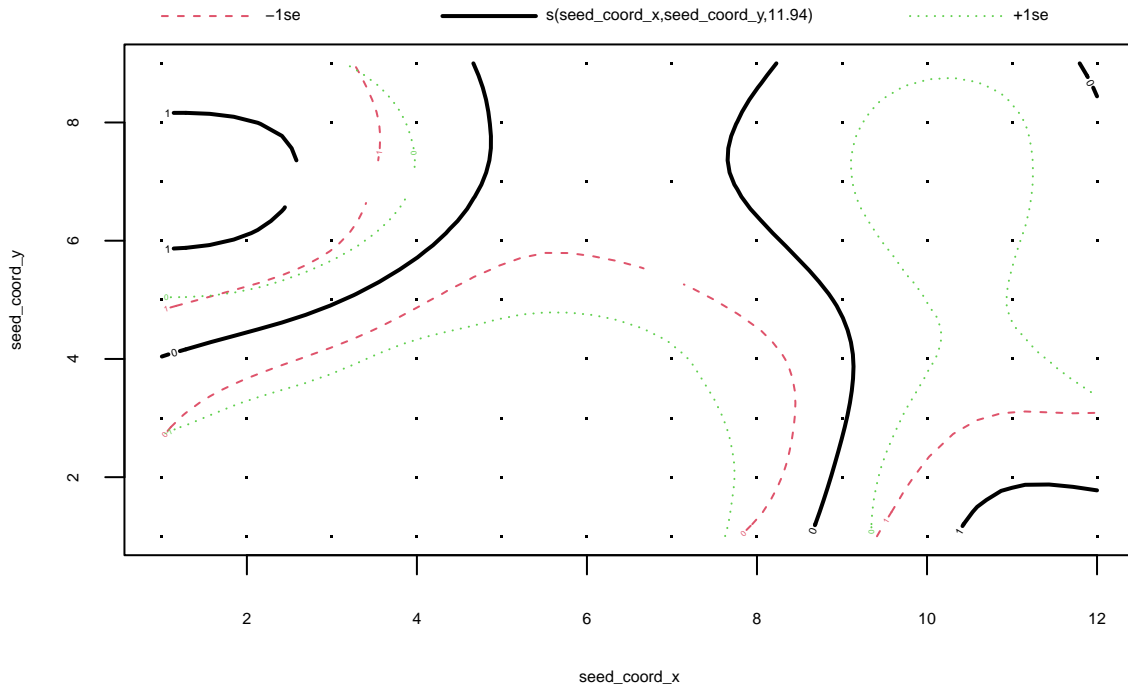
	edf	Ref.df	F	p-value
s(seed_coord_x,seed_coord_y)	11.941	16.274	1.7202	0.06461 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.59 Deviance explained = 67.4%

GCV = 1.2732 Scale est. = 1.0012 n = 84

```
plot(gam.days.to.germination)
```



Only the coefficient of *depth* seems to be significantly different from 0.

This makes sense since, if the seed is deep in the soil, it takes more time for the seedling to emerge from it.

The effective degrees of freedom (referred to as edf) are high, suggesting that the position in the lab has a non-linear relationship with the response variable. The plot displays the gradient of the estimated smooth terms, confirming a non-linear trend.

The explained deviance is relatively high, indicating that the model is a reasonably good fit to the true distribution of the response variable.

6.3 Model selection

We now want to check whether it is best to introduce the smooth terms interacting with each others (i.e. in a bivariate way, as above) or in an additive way.

To test this hypothesis, we fit a model with the smooth terms introduced additively.

We need to adjust the number of basis in the `s()` function (the function estimating the smooth term) because otherwise the model cannot be fitted; this is done by modifying `k`, which sets the upper limit on the degrees of freedom associated with the `s()` smooth.

```
k.x <- d.maize %>%
  pull(seed_coord_x) %>%
  n_distinct()
k.y <- d.maize %>%
  pull(seed_coord_y) %>%
  n_distinct()
gam.days.to.germination.add <- gam(days.to.germination ~ s(seed_coord_x, k = k.x) +
  s(seed_coord_y, k = k.y) +
```

```

soil.fac +
depth +
seed.weight,
data = d.maize)
summary(gam.days.to.germination.add)

```

Family: gaussian

Link function: identity

Formula:

```

days.to.germination ~ s(seed_coord_x, k = k.x) + s(seed_coord_y,
  k = k.y) + soil.fac + depth + seed.weight

```

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.600802	0.725415	13.2349	< 2.2e-16 ***
soil.facComposana	-0.473481	0.417483	-1.1341	0.2606
soil.facherbs	-0.132244	0.381958	-0.3462	0.7302
soil.facmixture	-0.462299	0.452715	-1.0212	0.3106
depth	0.631473	0.071986	8.7721	6.231e-13 ***
seed.weight	-0.020220	0.020719	-0.9759	0.3324

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value
s(seed_coord_x)	4.5831	5.6770	1.9709	0.07624 .
s(seed_coord_y)	2.4695	3.0879	1.1144	0.37616

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

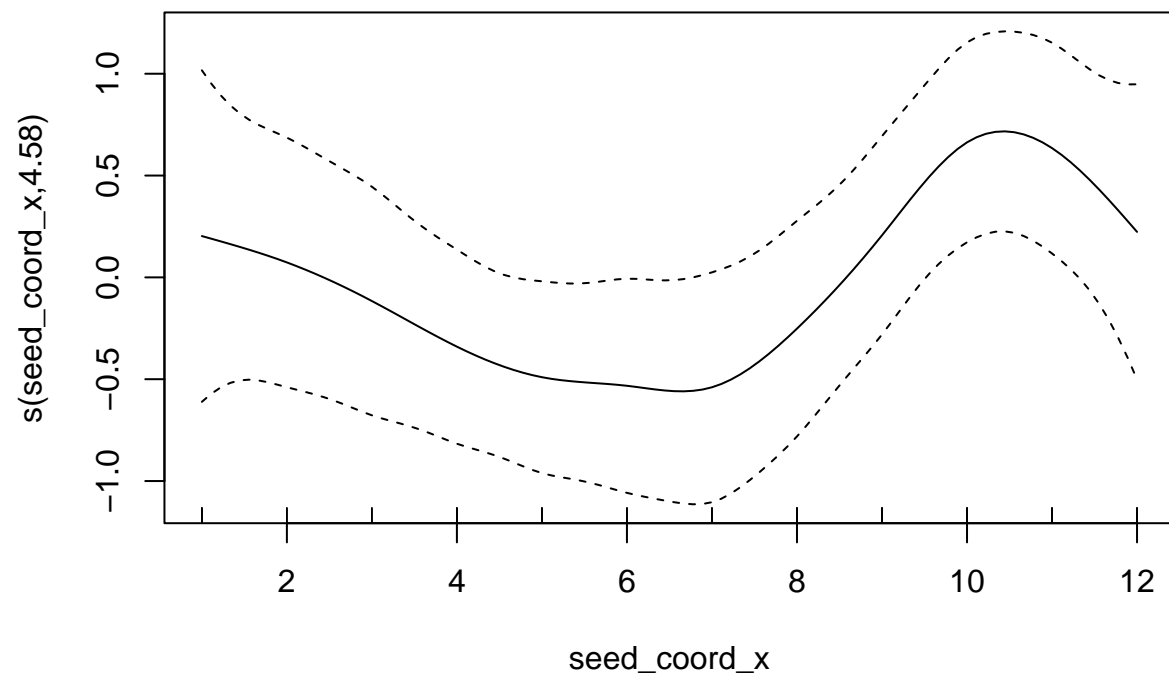
R-sq.(adj) = 0.534 Deviance explained = 60.2%

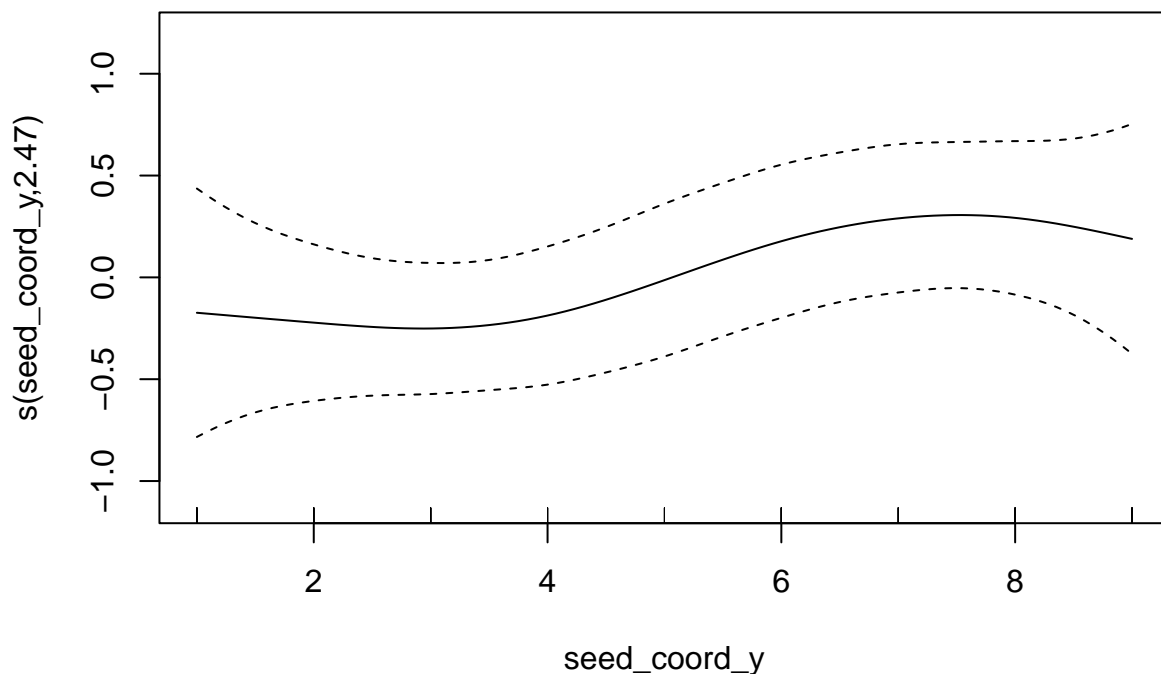
GCV = 1.3473 Scale est. = 1.138 n = 84

```

plot(gam.days.to.germination.add)

```





We compare the models using the AIC and the BIC criteria.

We use this method instead of comparing the models using the `anova()` method because the models are not nested. Indeed, when fitting a GAM with smooth terms, if we modify the these smooth terms, the smooths change, and as a result, the models are not nested.

```
AIC(gam.days.to.germination.add,
    gam.days.to.germination)
```

	df	AIC
gam.days.to.germination.add	14.052565	263.15752
gam.days.to.germination	18.940970	256.18459

```
##
BIC(gam.days.to.germination.add,
    gam.days.to.germination)
```

	df	BIC
gam.days.to.germination.add	14.052565	297.31674
gam.days.to.germination	18.940970	302.22662

Both criteria agree that the best model is the simpler one, where the smooth terms are introduced in the model in an additive manner.

In addition, we check whether the variable *soil.fac* has an influence on the response variable as a whole. Indeed, the above summary only shows the relative influence of each level compared to the reference level of *soil.fac*.

To achieve this result, we fit another model without this variable, and then we compare the two models using the `anova()` function.

We use the `anova()` function instead of the `drop1()` function, because the latter is not implemented for generalised additive models.

```
gam.days.to.germination.add.reduced <- gam(days.to.germination ~ s(seed_coord_x, k = k.x) +
                                           s(seed_coord_y, k = k.y) +
                                           depth +
                                           seed.weight,
                                           data = d.maize)

##
anova(gam.days.to.germination.add, gam.days.to.germination.add.reduced, test = "Chisq")
```

Analysis of Deviance Table

Model 1: days.to.germination ~ s(seed_coord_x, k = k.x) + s(seed_coord_y,
k = k.y) + soil.fac + depth + seed.weight

Model 2: days.to.germination ~ s(seed_coord_x, k = k.x) + s(seed_coord_y,
k = k.y) + depth + seed.weight

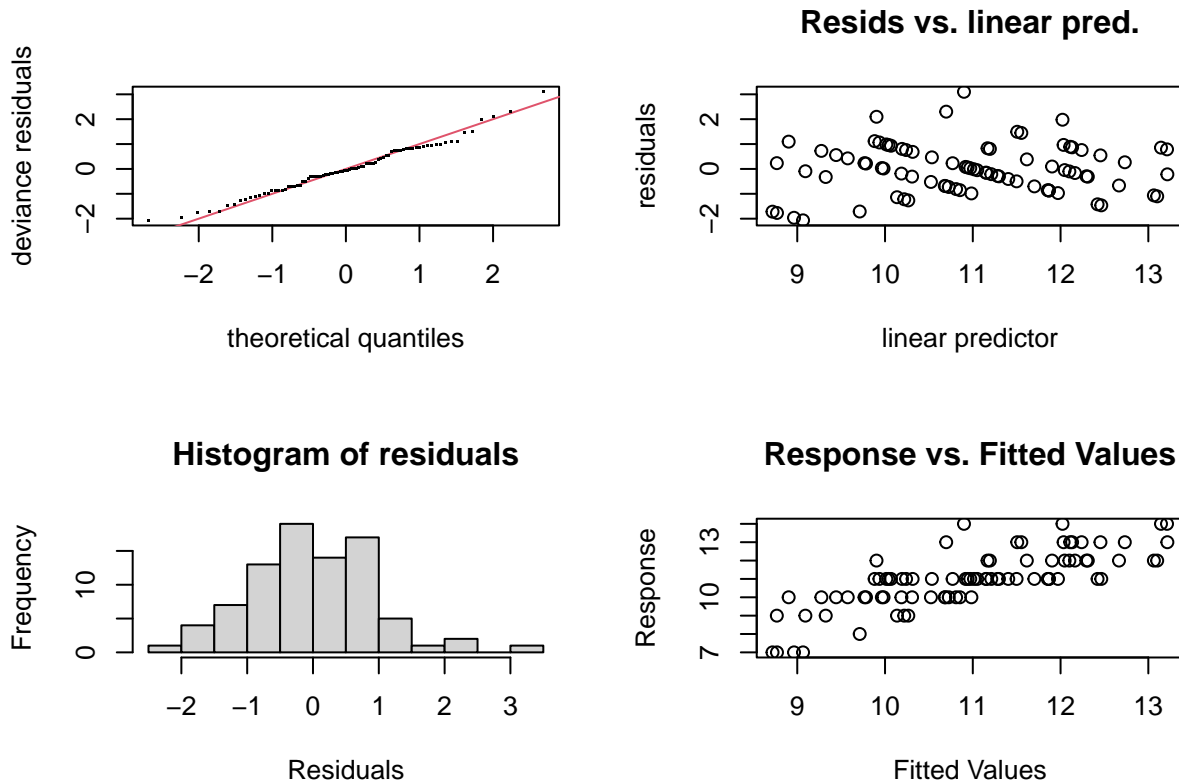
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	69.235	80.7358			
2	71.563	81.7780	-2.32795	-1.04222	0.7062

The test shows that we have no evidence that including the variable *soil.fac* improves the model.

However, because the variable is in the design, we do not drop it from the model.

6.4 Model checking

```
set.seed(12)
gam.check(gam.days.to.germination.add)
```



Method: GCV Optimizer: magic
Smoothing parameter selection converged after 10 iterations.
The RMS GCV score gradient at convergence was 2.3306397e-05 .
The Hessian was positive definite.
Model rank = 25 / 25

Basis dimension (k) checking results. Low p-value (k-index<1) may indicate that k is too low, especially if edf is close to k'.

	k'	edf	k-index	p-value
s(seed_coord_x)	11.00	4.58	1.09	0.76
s(seed_coord_y)	8.00	2.47	1.16	0.94

The model does not seem to violate any of the assumptions.

7 Methods description

To understand which factors influence the germination time of seeds, a generalised additive model was employed.

We used smooth terms because the explanatory variables *seed_coord_x* and *seed_coord_y* do not follow any clear distribution, thus the assumptions made on linear or generalised linear models did not hold.

The statistical analysis was performed using the R programming language, specifically version 4.3.1 (see citation below). The generalised linear model was fitted with the `gam()` function in the `{mgcv}` add-on

package (see citation below).

Citations

```
citation()
```

To cite R in publications use:

```
R Core Team (2023). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
```

A BibTeX entry for LaTeX users is

```
@Manual{,
  title = {R: A Language and Environment for Statistical Computing},
  author = {{R Core Team}},
  organization = {R Foundation for Statistical Computing},
  address = {Vienna, Austria},
  year = {2023},
  url = {https://www.R-project.org/},
}
```

We have invested a lot of time and effort in creating R, please cite it when using it for data analysis. See also 'citation("pkgname")' for citing R packages.

```
citation("mgcv")
```

2011 for generalized additive model method; 2016 for beyond exponential family; 2004 for strictly additive GCV based model method and basics of gamm; 2017 for overview; 2003 for thin plate regression splines.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1):3-36

Wood S.N., N. Pya and B. Saefken (2016) Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association* 111:1548-1575.

Wood, S.N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*. 99:673-686.

Wood, S.N. (2017) *Generalized Additive Models: An Introduction with R* (2nd edition). Chapman and Hall/CRC.

Wood, S.N. (2003) Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* 65(1):95-114.

To see these entries in BibTeX format, use 'print(<citation>, bibtex=TRUE)', 'toBibtex(.)', or set 'options(citation.bibtex.max=999)'.

8 Session information

```
sessionInfo()
```

```
R version 4.3.1 (2023-06-16)
```

```
Platform: aarch64-apple-darwin20 (64-bit)
```

```
Running under: macOS Sonoma 14.0
```

```
Matrix products: default
```

```
BLAS: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
```

```
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib; LAPACK vers
```

```
locale:
```

```
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
time zone: Europe/Zurich
```

```
tzcode source: internal
```

```
attached base packages:
```

```
[1] stats      graphics  grDevices  utils      datasets  methods    base
```

```
other attached packages:
```

```
[1] mgcv_1.8-42      nlme_3.1-162      lme4_1.1-34      Matrix_1.6-1.1  
[5] multcomp_1.4-25 TH.data_1.1-2     MASS_7.3-60      survival_3.5-5  
[9] mvtnorm_1.2-3    tibble_3.2.1      ggplot2_3.4.4    kableExtra_1.3.4  
[13] dplyr_1.1.3      knitr_1.44
```

```
loaded via a namespace (and not attached):
```

```
[1] sandwich_3.0-2    utf8_1.2.4        generics_0.1.3    xml2_1.3.5  
[5] stringi_1.7.12    lattice_0.21-8    digest_0.6.33     magrittr_2.0.3  
[9] evaluate_0.22     grid_4.3.1        fastmap_1.1.1     httr_1.4.7  
[13] rvest_1.0.3       fansi_1.0.5       viridisLite_0.4.2 scales_1.2.1  
[17] codetools_0.2-19 cli_3.6.1          rlang_1.1.1       munsell_0.5.0  
[21] splines_4.3.1     withr_2.5.1       yaml_2.3.7        tools_4.3.1  
[25] nloptr_2.0.3      minqa_1.2.6       colorspace_2.1-0  webshot_0.5.5  
[29] boot_1.3-28.1     vctrs_0.6.4       R6_2.5.1          zoo_1.8-12  
[33] lifecycle_1.0.3   stringr_1.5.0     pkgconfig_2.0.3   pillar_1.9.0  
[37] gtable_0.3.4      Rcpp_1.0.11       glue_1.6.2        systemfonts_1.0.5  
[41] xfun_0.40         tidyselect_1.2.0  rstudioapi_0.15.0 htmltools_0.5.6.1  
[45] rmarkdown_2.25    svglite_2.1.2     compiler_4.3.1
```