Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Statistics Lab

Final Report

Spring Term 2024

Supervised by:
Fabio Sigrist

Authors:
Tatiana Bielakova
Marvin Lob
Paul Stephan

Abstract

This report presents a methodology for modelling the growth patterns of soy plants' canopy cover based on environmental variables, utilising a three-parameter logistic mixed effects model. It expands upon prior research by fellow ETH statistics students and incorporates a refined definition of the soy plant's growth period and a novel characterisation of spatial grouping structures. Analysis of 151 soybean genotypes across two experimental fields over eight years identified genotypes number 10004, 10009, 10014, 10018, and 10020 as potential candidates resilient to varying weather conditions. While we find notable differences between measurements from the two platforms, further analysis is needed to discover location and platform effects. Our comparison of the bootstrap method and the basic intervals() function from the nlme package shows that, in most cases, the latter suffices. This finding is contributing to the discourse on correct inferences in non-linear mixed effects models. The report provides a reproducible pipeline for accurately modelling soybean growth in diverse environmental conditions, recommending resilient genotypes for further testing, and presenting groundwork for future tests assessing measurement platforms.

Contents

1 Introduction

This report aims to identify robust soybean genotypes resilient to varying weather conditions using the nonlinear mixed effect model. Our study focuses on 151 different genotypes, analyzed through 10,250 observations collected from 2015 to 2022 at two distinct locations: Eschikon (FIP drone) and Delley (UAV camera).

We model the canopy coverage of the various soybean genotypes with a logistic growth curve. This method is chosen for its ability to capture the sigmoidal growth pattern typical of plant canopies. The model includes random effects to account for the grouping of plots within each location, ensuring that location-specific variations are considered. This allows us to isolate and understand the inherent differences due to the environments at Eschikon and Delley.

In addition to random effects, the model incorporates fixed effects to analyze the influence of weather variables and genetic differences on canopy development. The weather variables considered are precipitation, radiation, and temperature, each averaged over the 14 days prior to each measurement. By including these fixed effects, we can assess the impact of environmental conditions on the growth of soybean canopies and determine how different genotypes respond to these conditions.

This approach aims to identify soybean genotypes that demonstrate stable performance across a range of weather conditions, highlighting those that are resilient to environmental fluctuations. By doing so, we identify 6 potential genotypes that are likely to perform well under variable climatic conditions, which is crucial for agricultural stability and productivity (and the main question of our client).

Additionally, we compare the two platforms using an NLME model, focusing solely on the intercepts for each platform. While our analysis reveals a significant difference between the results, we cannot determine whether this discrepancy is attributable to the location, the platform, or the combination, as the two variables are perfectly confounded. Finally, we compare the classical method of obtaining intervals with a bootstrap method on a smaller subsample of the selected candidates to determine the validity of our results.

This report builds upon the work of students from last year's Statistics Lab and is structured as follows. In Section 2, we introduce the definition of growth phases, followed by plot grouping in 3. Section 4 presents our proposed model. In Section 5, we continue with analyzing the model's diagnostics, and in Section 6, we discuss our approach to selecting potential robust candidates. Section 7 compares the two measurement platforms, and Section 8 addresses the issue of valid inference for the NLME model. Finally, Chapter 9 discusses the data repository, and we summarize our conclusions in Section 10.

2 New definition of the Growth Phase

The data on the canopy cover of the soybean plants encompass the entire life cycle of a soy plant. Thus, it includes the growth period, the period at which the soy plants are fully grown, and the period when the plants die. Thus, the graph over the entire year per plant has an inverted U-shape, as can be seen in Figure 3. Since the research questions of our client are targeted only at the growth period of the plants, we need to define the point at which the growth period ends and the senescence period starts.

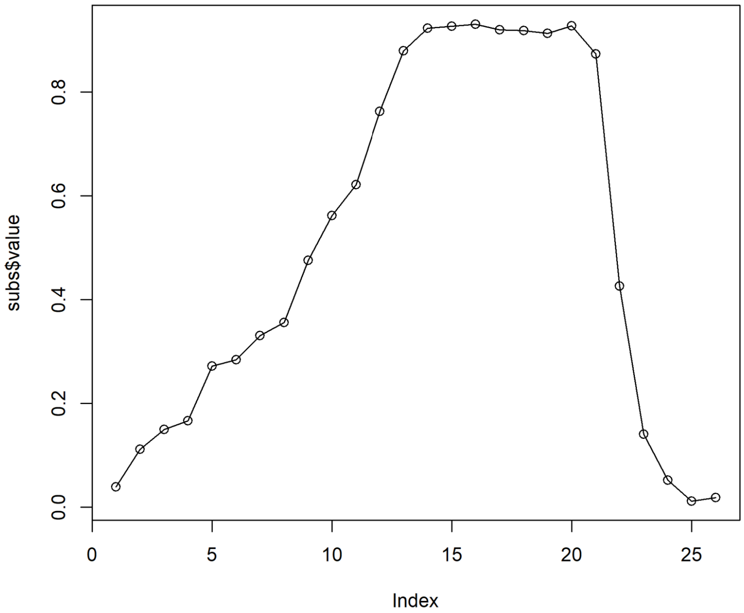
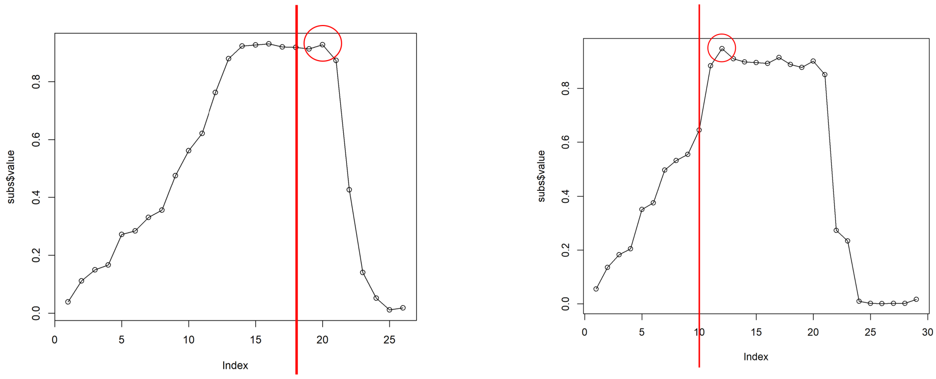


Figure 1: Lifecycle of Soy Plant

Defining the end of the growth period is not as trivial as picking the maximum canopy cover. This is because the canopy cover curve is characterized by a plateau phase with small fluctuations in the neighbourhood of the maximum. Because of these fluctuations, the maximum value might occur at any point in the plateau phase. In the data set that we were initially given, the growth and senescence periods have already been marked. However, upon closer inspection, we noticed that the border between the growth and senescence phase was defined as two data points before the maximum value. Since, as explained above, the maximum value may occur at any point of the plateau, this has led to either part of the true growth phase being cut off or considerable fluctuations being left in the annotated growth phase. This can be seen in Figure 2.



((a)) Fluctuations remain

((b)) Growth phase cut off

Figure 2: Examples for faulty definition of growth phase

To algorithmically determine the appropriate transition from growth to senescence phase, we implement the following procedure that is depicted in Figure 3. For every single plot, we note the timestamp of the maximum canopy cover. Since we found that the fluctuations occur at maximally 0.02 percentage points under the max value, we also note when the plant first reached the maximum canopy cover value minus 0.02 percentage points. Now, we inspect the period between the first time reaching 0.02 under the max and the maximum. If the plant monotonically grows until the maximum, we simply take the maximum as the end of the growth phase. If the canopy cover did not grow monotonically in that period, we take the last value before the first drop-off as the end of the growth phase. This way, the growth period is fully included, but the fluctuation phase is not.

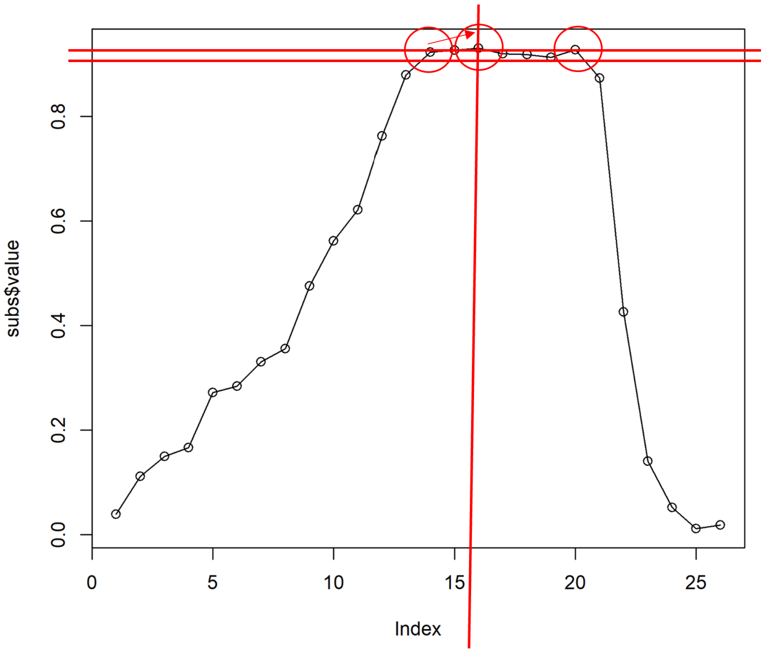


Figure 3: New definition of end of growth phase

3 Plot grouping

The random effects in our model are included as plots grouped in a 6x2 formation. By including these random effects, we capture the influence of unobserved variables that may (and very likely do) vary across different parts of the field, such as soil quality or wind strength. This inclusion helps to control for unmeasured confounders that could otherwise bias the fixed effects estimates.

We follow the original experimental design for each of the fields (Roth, Barendregt, C.-A. Bétrix, et al. 2022). One field represents a Latin rectangle with three replications and an overall dimension of 18 rows by 6 ranges (columns). Complete blocks in the range and row directions consist of 2 ranges and 6 rows, respectively. The covered size for the majority of grouped plots is then in line with other studies focused on soil spatial variation, where the usual plot is of a squared shape with length size varying from 7-13 meters (Stevens, Bogaert, and van Wesemael 2015; Yang et al. 2019).

4 Model Specification

We model the growth of soy plants using a non-linear mixed effects model. As a non-linear model, we utilise a three-parameter logistic curve since this model is frequently being used for modelling the growth of plants (Archontoulis and Miguez 2015), and the previous analyses using the soy plant data have also used this type of model. The equation below shows the equation of a general three-parameter logistic model.

$$f(x) = \frac{\text{Asym}}{1 + \exp\left(\frac{-(x - \text{xmid})}{\text{Scal}}\right)} + \epsilon$$

In the appendix, we provide a detailed explanation of what such a model entails. Here, we discuss how we apply it in our biological context. First, we explain what the parameters of the model mean. Then, we motivate the fixed effects by explaining the model parameters and, lastly, comment on the random effects in the model.

Understanding the model parameters

Each parameter of the model has a distinct role in determining the exact shape of the sigmoid curve. Figure 4 shows the effect of a change of every parameter.

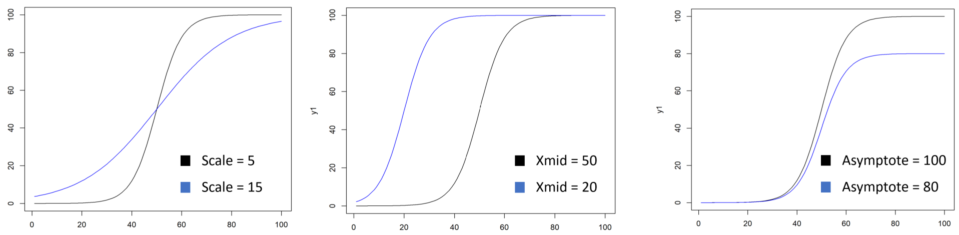


Figure 4: Fixed effects model specification

The scale parameter governs the steepness of the curve, with a lower coefficient indicating a steeper curve. The xmid parameter indicates the x-coordinate of the midpoint of the curve. Thus, different xmid coefficients represent a shift from left to right. Lastly, the asymptote gives the y-value that the curve approaches.

Explanation of the Fixed Effects

A central goal of our analysis is to find a well-fitting interpretable model that represents the environmental influences on the growth phase of soy plants. Further, we want our model to serve as a basis for testing research questions with a biological motivation. Therefore, our model specification is mainly guided by literature from the field of environmental sciences rather than just maximising the fit of the model. To be specific, we picked the variables that are used to predict the coefficients of our 3-parameter logistic curve according to biological theory. Since our client's main interest is the effect of environmental variables on different genotypes of soy plants, we picked the variables for our model specification accordingly. Further, we account for the spatial structure of the experimental fields by adding random effects.

First of all, we needed to choose how to reflect environmental conditions in the model. Guided by biological literature, we consider precipitation, radiation, and temperature (Beardsell, Mitchell, and Thomas 1973 and Roth, Barendregt, C. A. Bétrix, et al. 2022). This is because these variables serve as the basis for photosynthesis, which enables a plant to grow (Zhang et al. 2023). To balance the resolution of our data while still remaining able to capture variation in environmental conditions, we consider rolling time windows of 14 days length. Thus, we summed up the variable values over the past 14 days to create our measure for each environmental condition. The equation below gives an overview of the fixed and random effects structure of our model.

$$\phi_{ij} = \begin{bmatrix} Asym_{ij} \\ Scal_{ij} \\ xmid_j \end{bmatrix} = \underbrace{\begin{bmatrix} genotype_i \\ genotype_i * (Precip_j + Radiation_j) \\ Temp_j + Precip_j \end{bmatrix}}_{\text{fixed effects}} + \underbrace{\begin{bmatrix} plot\_grouped_{Asym,j} \\ 0 \\ plot\_grouped_{xmid,j} \end{bmatrix}}_{\text{random effects}}$$

For characterizing the growth phase of the soy plants, the most important parameter of the 3-parameter logistic curve is the scale parameter. This parameter describes how steep the slope of the exponential part of the logistic growth curve is. Hereby, a low scale value indicates fast growth and vice versa. Since this scale parameter is our main parameter of interest, we explain it by the interaction of the genotypes with the environmental conditions. For environmental conditions, we pick precipitation and radiation. We chose radiation over temperature or the combination of both as the model did not converge when picking both, and since radiation represents the sunlight that a leaf can use to generate energy for further growth.

The xmid parameter of the logistic curve signifies a temporal shift of the period of the quickest growth. Since this shift may be caused by lower temperatures in the very early germination phase, we explain this parameter by temperature. We do control for radiation as the leaves are not or not fully developed before the quick growth phase, and, therefore, temperature is more meaningful for the development of the plant. We also control for precipitation for the xmid parameter.

Since we are using canopy cover instead of plant height as our target variable, the asymptote should theoretically converge close to 100% for all plants. Therefore, this parameter is of smaller interest to us compared to especially the scale parameter. However, to account for potential growth issues of certain genotypes, we still explain the asymptote by the genotypes in our model.

Explanation of the random effects

Next, we look at the random effects included in the model in the form of grouped plots with formation 6x2. Incorporating random effects helps control for unobserved heterogeneity. In other words, random effects account for variations in outcomes that observed variables alone cannot explain.

To present it as an example, in our case, the unobserved heterogeneity could have a form of different soil composition in different parts of the same field and between the fields. This could result in e.g. earlier growth of plants with better soil quality. As we do not have information on the soil quality at the time of observations but consider it to be important for determining the growth, we implicitly allow for variation depending on where and which year the different genotype is planted.

Considering the explanation from above, we use random effects on two of the three parameters, xmid and asymptote. Starting with the xmid, the parameter corresponds to the point where the growth rate transitions from accelerating to decelerating. Changing the value of xmid signifies a temporal shift of the period of the quickest growth (as explained in the section above). It can be, therefore, also interpreted as a parameter describing the start of the growth, for which the soil quality would be a crucial component. Thus, we take into account both fixed and random effects for this parameter.

Continuing with the asymptote parameter, theoretically, it should always converge close to 1. However, there are certain differences that we first capture by the fixed effect of genotype. To encounter further exogenous influences affecting the final coverage of the plant above the ground level, we include the random effects on the asymptote. Looking at some of the biological interpretations of these differences, the strong wind coming from certain directions could potentially affect the final height/ coverage of the plant as certain parts of the field would be then more hostile for planting.

We present the final specification of the NLME model in equation 1.

$$CanopyCover_{ij}(x) = \frac{Asym_{ij}}{1 + \exp\left[-\left(x - xmid_j\right)/Scal_{ij}\right]} + \epsilon_{ij},$$

$$\phi_{ij} = \begin{bmatrix} Asym_{ij} \\ Scal_{ij} \\ xmid_j \end{bmatrix} = \begin{bmatrix} genotype_i \\ genotype_i * (Precip_j + Radiation_j) \\ Temp_j + Precip_j \end{bmatrix} + \begin{bmatrix} plot\_grouped_{Asym,j} \\ 0 \\ plot\_grouped_{xmid,j} \end{bmatrix} \quad (1)$$

$$\begin{aligned} plot\_grouped_{Asym,j} &\sim \mathcal{N}(0, \psi_1), & & \\ plot\_grouped_{xmid,j} &\sim \mathcal{N}(0, \psi_3), & \qquad i &\in \{1, 2, \ldots, n_{genotypes}\}, \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) & j &\in \{1, 2, \ldots, n_{plots}\} \end{aligned}$$

5 Diagnostics

Model diagnostic plots

The performance of Model 1 can be visualised in diagnostic plots. First, we look at the Fitted values vs residuals plot in figure 5. Generally, the plot looks good. We see that the variance of the residuals does not change too much across the fitted values. Also, they mostly seemed centred at zero. However, at higher predicted values, we see consistently negative residuals, indicating our model predicts larger asymptotes than the data indicates. Fortunately, this is not too much of a concern for our analysis as we are not particularly interested in the value of the asymptote but more in the growth phase.

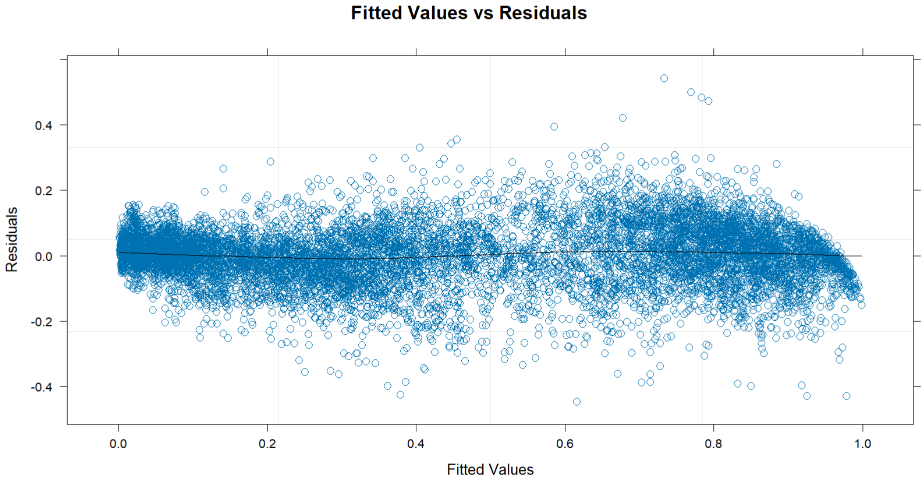


Figure 5: Fitted values vs Residuals

Furthermore, the normality assumption of the errors is investigated by checking the QQ plot in figure 6. We see mostly a well-behaved QQ plot with the exception of some outliers at the end and beginning. However, because we have many observations for most genotypes, our inference is not dependent on the normality assumption as the central limit theorem takes over. Last but not least, we can check the distributional assumption of the random effect. They should be normally distributed. Figure 7 shows that the mostly follows this assumption.

Model comparison

The students from last year identified the need for spatial variables. This, however, poses the risk that the random effects we include in our model capture most of the variability. A justification is needed that in our proposed model, the inclusion of genotypes and weather variables is actually necessary. For this, we compare our model with a model that only includes the random effects but now genotypes nor weather variables (called *no_geno*) with a model that does include only genotypes but no weather variables (called *only_geno*) and with our final model (called *final_model*).

We see in table 1, 2, and 3 that each bigger, more complex model was justified in a sense that the increase in variance it explained was significant compared to the increase in degrees of freedom.

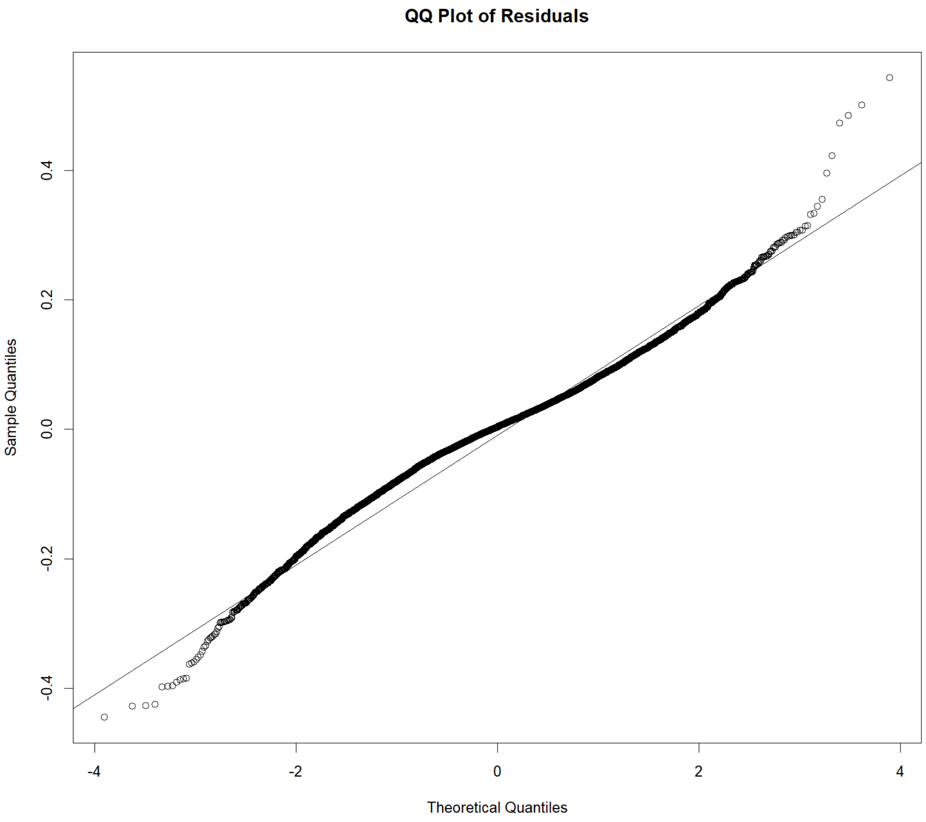


Figure 6: QQ-plot

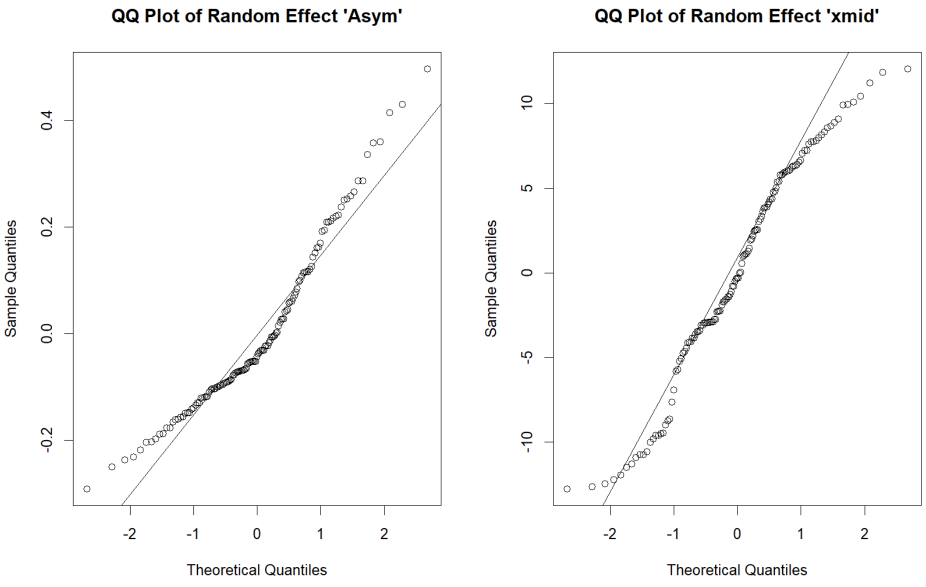


Figure 7: QQ-plot of random effects

Table 1: Model Comparison: no genotype vs only genotype

Model	df	AIC	BIC	logLik	L.Ratio	p-value
no_geno	8	-17744.08	-17686.17	8880.039		
only_geno	296	-18367.28	-16224.65	9479.641	1199.204	< .0001

Table 2: Model Comparison: no genotype vs final model

Model	df	AIC	BIC	logLik	L.Ratio	p-value
no_geno	8	-17744.08	-17686.17	8880.039		
final_model	588	-19079.66	-14823.34	10127.832	2495.586	< .0001

Table 3: Model Comparison: only genotype vs final model

Model	df	AIC	BIC	logLik	L.Ratio	p-value
only_geno	296	-18367.28	-16224.65	9479.641		
final_model	588	-19079.66	-14823.34	10127.832	1296.382	< .0001

6 Finding ideal genotypes

One of the core aims of our analysis of the growth of soy plants is to identify genotypes that grow fast and are robust to differing environmental conditions. Statistically, determining the robustness of a genotype in this observational setting is challenging. In our model, the main effect of the environmental conditions and the interaction between environmental variables and the genotypes represent how much a change in environmental conditions, like, for example, higher/lower temperatures, impacts the growth speed of the soy plant per genotype. A significant coefficient on the environmental variables thereby shows that the growth is impacted by varying environmental conditions. The absence of a significant coefficient, however, does not imply that environmental conditions have no effect on the genotype but just that the current study was not able to identify an effect. For determining ideal candidates, we, therefore, had to be more considerate than just picking genotypes without significant environmental effects.

Our approach to identifying promising candidates is based on the notion of equivalence tests Meyners 2012. Hereby, both the estimate and the confidence interval of the parameter are inspected. If the estimate is close to zero and the width of the confidence interval is small, it is an indication that the effect size is small and the effect was reliably estimated - which is equivalent to our definition of a genotype being robust with respect to an environmental condition. This logic is depicted in Figure 8. The green intervals are the ones that would pass the test, and the red ones are the ones that fail. For the genotype main effect, we slightly deviate from this idea since we aim to have a scale parameter that is as low as possible but reliably estimated. Thus, we impose a one-sided restriction on the estimate. Since this test does not guarantee that the genotypes are actually robust, we can only use it as a filter that enables us to sort out genotypes that likely are not good robust candidates. Therefore, we come up with certain thresholds for this test that will act as a filter that an ideal genotype must pass.

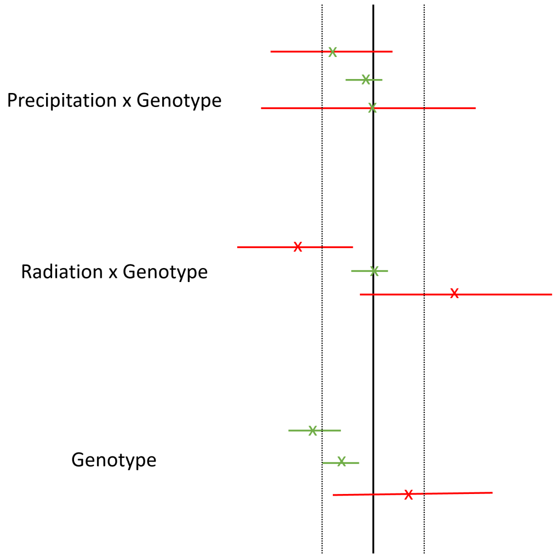


Figure 8: Idea of Equivalence test

An issue that arose when identifying the appropriate thresholds for filtering the genotypes was using the default values of the nlme-function yields coefficient estimates for factorial variables that are based on the treatment constraint. Thus, the main effect of the environmental variables is calculated for the reference genotype, and the interactions represent the deviations of each genotype-environmental interaction from the environmental effect of the reference genotype. Therefore, the confidence intervals yielded by the intervals function are regarding the deviation from the baseline and not the actual environmental effect per genotype. Since, for non-linear mixed effect models, there is no straightforward way of deriving the appropriate confidence intervals of the genotype-specific main effect from the interaction effect directly, we use the sum-to-zero constraint to estimate a baseline model. The main effect of this model represents the overall mean radiation and precipitation effect and the interaction effects represent deviations from the overall mean. The overall mean scale, asymptote, environmental effects and their corresponding confidence intervals can be found in Table 4.

Table 4: Overall mean coefficients

	Scale	Radiation	Precipitation	Asymptote
Estimate	32.89	-0.005	-0.082	0.9723
Interval width	8.15	0.002	0.004	0.004

As expected, the overall means of the environmental effects are negative. This means that, on average, more rainfall and more radiation leads to a steeper growth curve, so to quicker growth. Since our goal is to find genotypes that are robust to environmental conditions, we aim to find genotypes whose environmental effect coefficients are close to zero. For the genotype-specific main effect to be zero, the interaction effect must be of the same size as the negative of the overall main effect. For the precipitation effect, for example, the precipitation main effect is -0.08. So for the genotype-specific precipitation effect to be zero, the genotype-precipitation effect in the model must be +0.08, so that the effects cancel out. This serves as our logic for defining the threshold for the estimates of the environmental effects. The thresholds for the widths of the intervals were guided by the empirical distribution of the widths of the interaction effects. Hereby, the interval widths should be a filtering criterion of similar strength as the coefficient estimate filters.

For determining the ideal candidate, we chose a threshold for the confidence interval of precipitation as the width of 0.05 and for radiation the width of 0.005. Since an ideal genotype should be quickly growing in general, we also filtered for the genotype-specific scale parameter. The ideal candidates should have a scale parameter that is smaller than 20. Also, this choice was made after a descriptive inspection of the sample. We also pose a restriction on the Asymptote, as the ideal genotype should reach a sufficiently high maximum canopy cover value. The exact choices for all thresholds, as well as the number of genotypes that fulfill each condition, are summarized below in Table 5.

Table 5: Filters for finding ideal genotypes

	Scal[1] est[5]	Scal int[6]	Rad[2] est	Rad int	Prec[3] est	Prec int	Asym[4] est	Asym int
Thresh	<0	<20	0.002-0.1	<0.005	0.07-1	<0.05	> -0.03	<0.07
$\#n$	95	56	54	45	25	20	90	24

[1]Scale, [2]Radiation, [3]Precipitation, [4]Asymptote, [5]Estimate,[6]Interval width

[7]Number of genotypes fulfilling the threshold

To sum up, ideal candidates should have scale parameters that are not affected by precipitation and radiation, whose environmental effects have been estimated with sufficient precision, and that are generally growing at a fast pace. These criteria were fulfilled by six genotypes (Genotypes 10004, 10009, 10014, 10015, 10018, and 10020), our final candidates. We refit the model with each one of the candidates at a time as a reference level to obtain the correct confidence intervals for each environmental variable. The genotype-specific environmental effects on the scale parameter are depicted in Figure 9.

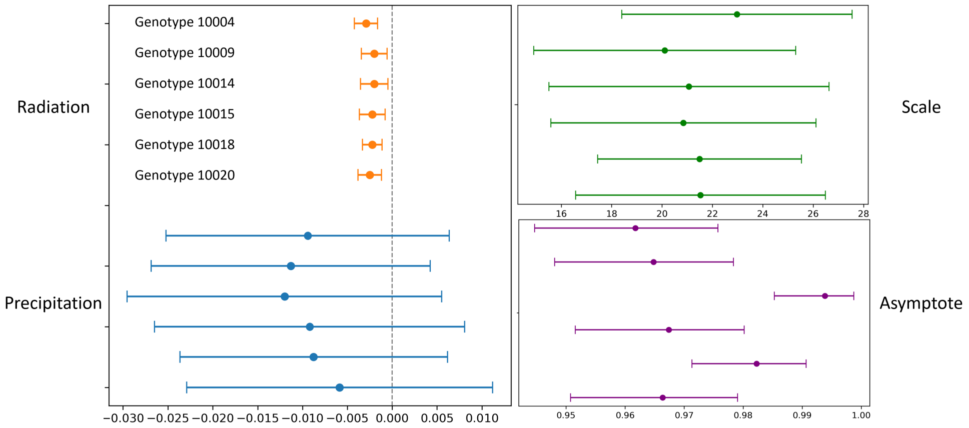


Figure 9: Details on the ideal candidate genotypes

7 Platform comparison

As a next step, we did some initial comparisons of the two measuring platforms - UAV and FIP. We face several problems, the main one being the fact that the locations and platforms are perfectly confounded (see Table 6). Thus, we cannot reliably distinguish to which extent the difference in estimates can be attributed to location, platforms, or the combination of the two. This is a fundamental issue that cannot be addressed in the current experimental design. Therefore, a reader should bear in mind this issue, and the validity of our results should be interpreted with caution.

Platform	Year	Location	# Observations
FIP	2015	Eschikon	506
FIP	2016	Eschikon	589
FIP	2017	Eschikon	798
FIP	2018	Eschikon	817
FIP	2019	Eschikon	503
FIP	2020	Eschikon	1296
FIP	2021	Eschikon	914
FIP	2022	Eschikon	841
UAV	2019	Delley	1054
UAV	2020	Delley	2969

Table 6: Data summary of platform, year, location, and observations

To account for the variability in genotypes, potential location specifics, and differences in yearly measurements, we again utilize the grouped plots and work with the mixed effect model. We work with the whole dataset, despite certain differences in yearly observations and genotypes planted, prioritising the size of the datasets and relying on the fact that the plot grouped structure should capture some of the effects.

The idea of the model is similar to the main model specified in Equation 1. We are working with logistic growth curves with parameters - Asym, xmid, and scal, consisting of the fixed and random effect parts. For the fixed effects, we use a simple approach using only dummy variables for intercept, testing whether there is a significant difference between the two platforms.

As the model with random effects on all three parameters failed to converge, we needed to make a choice on which two of the three parameters we should put the random effects. We decided to follow the approach from the main model, putting random effects on xmid and Asym. The xmid, being the symmetry point for the logistic growth curve, also determines the starting point of the growth. As the soil quality is detrimental to the initial growth, including random effects for the xmid makes sense.

The final model is presented in Equation 2. According to our analysis, there is a difference between the two platforms/ locations for all three parameters at a 5% significance level. However, as stated before, we cannot determine to which extent it can be attributed to platforms or location specifics. To do so, further analysis would require using the two measurement platforms in the same location and comparing observations in this setting.

$$
CanopyCover_{jl}(x) = \frac{Asym_{jl}}{1+\exp\left[-\left(x - xmid_{jl}\right)/Scal_{l}\right]} + \epsilon_j,
$$

$$
\phi_{jl} = \begin{bmatrix} Asym_{jl} \\ Scal_{l} \\ xmid_{jl} \end{bmatrix} = \begin{bmatrix} platform_{l} \\ platform_{l} \\ platform_{l} \end{bmatrix} + \begin{bmatrix} plot\_grouped_{Asym,j} \\ 0 \\ plot\_grouped_{xmid,j} \end{bmatrix} \tag{2}
$$

$$
\begin{aligned}
plot\_grouped_{Asym,j} &\sim \mathcal{N}(0,\psi_1), & & \\
plot\_grouped_{xmid,j} &\sim \mathcal{N}(0,\psi_3), & & j \in \{1,2,\dots,n^{l}_{plots}\}, \\
\epsilon_j &\sim \mathcal{N}(0,\sigma^2) & & l \in \{1,2\}
\end{aligned}
$$

Model Diagnostics

For the chosen model, we look at several diagnostics plots to visually check whether we strongly violate some of the assumptions. Plotting fitted values vs. residuals as depicted in Figure 10 suggests a relatively stable variance of the residuals with a mean around zero. Similarly, as in the case of the main model and Figure 6, the asymptote in the case of higher values is a bit off. Continuing with the QQ-plot in Figure 11, there are no strong outliers, and residuals seem to follow a normal distribution with constant variance. Lastly, we look at the QQ plots for the random effects in Figure 12. Although far from perfect, they do not violate the normality assumptions too strongly.

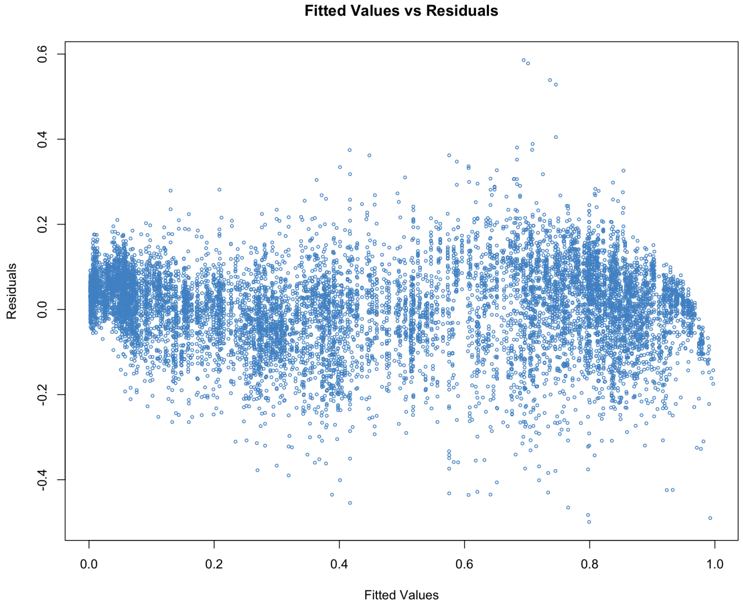


Figure 10: UAV-FIP Model: Fitted values vs Residuals

Lastly, as a final step in the analysis of the two platforms, we look at the comparison with a simple version of having a common intercept for the two platforms in all three parameters. As can be seen in Table 7, the more complex version is preferred based on all criteria.

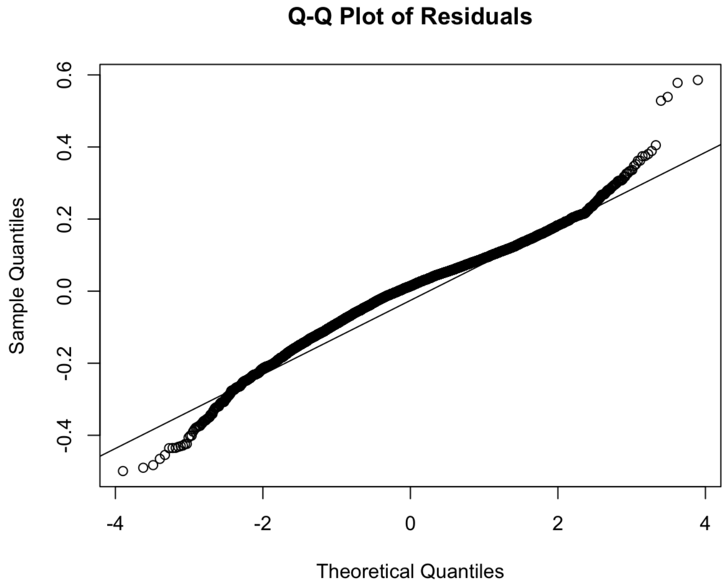


Figure 11: UAV-FIP Model:QQ-plot

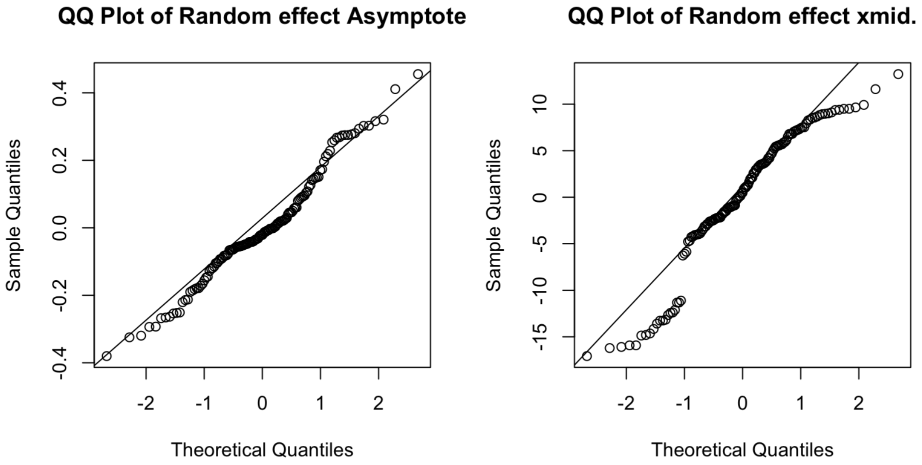


Figure 12: UAV-FIP Model: QQ-plot of random effects

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
only intercept	7	-16836.36	-16785.69	8425.18			
platform	10	-17111.77	-17039.38	8565.88	1 vs 2	281.49	< .0001

Table 7: Model Comparison

8 Comparison to bootstrapped intervals

During the process of modelling the growth using the `nlme` function, we were facing the issue of creating a valid inference for the nlme model. Based on the documentation of `nlme` package, the function interval is supported for `gls`, `lme`, and `lmList` classes. Although `nlme` is not directly mentioned, the received intervals appear without any errors or warnings. After contacting the `R-community`, we were suggested to compare the approach with the bootstrap method `boot_nlme`.

Due to the computational burden of the bootstrap method, we decided to do a proof-of-concept and run the method on a smaller subsample of our data. We decided to work with the selected genotype candidates specified in Section 6. The subsample thus consisted of 1268 observations of six genotypes - 10004, 10009, 10014, 10015, 10018, and 10020.

As a first step, we run the NLME model on the random subsample and derived intervals for the estimates using the classical `intervals()` function. Next, we run the bootstrap method `boot_nlme` on the classical setting with the default 999 sample setting and obtained the bootstrap intervals using `confint()` function.

We present the confidence intervals obtained with the methods from above in Figure 13. The red colour depicts the confidence intervals using the bootstrap method, and the blue colour depicts the classical intervals. The numbers next to the confidence intervals indicate the combined coverage, e.g., the ratio of the interval length shared by both methods to the total length. The coefficients are ordered decreasingly using this metric. As the length varies across different coefficients, we provide Figure 14 with the narrow cases and Figure 15 depicting the wide coefficients.

With a sufficiently large number of observations for the sample, the combined coverage ranges from 64-76 %. In the majority of cases, the bootstrapped confidence intervals are wider on both sides, which is in line with the intuition behind the method. As the `intervals()` is much more computationally effective without being too much off, we suggest sticking to this function in the future as well. However, with a small number of observations (e.g., less than 20, which is the case of some genotypes), the resulting confidence intervals vary by a lot.

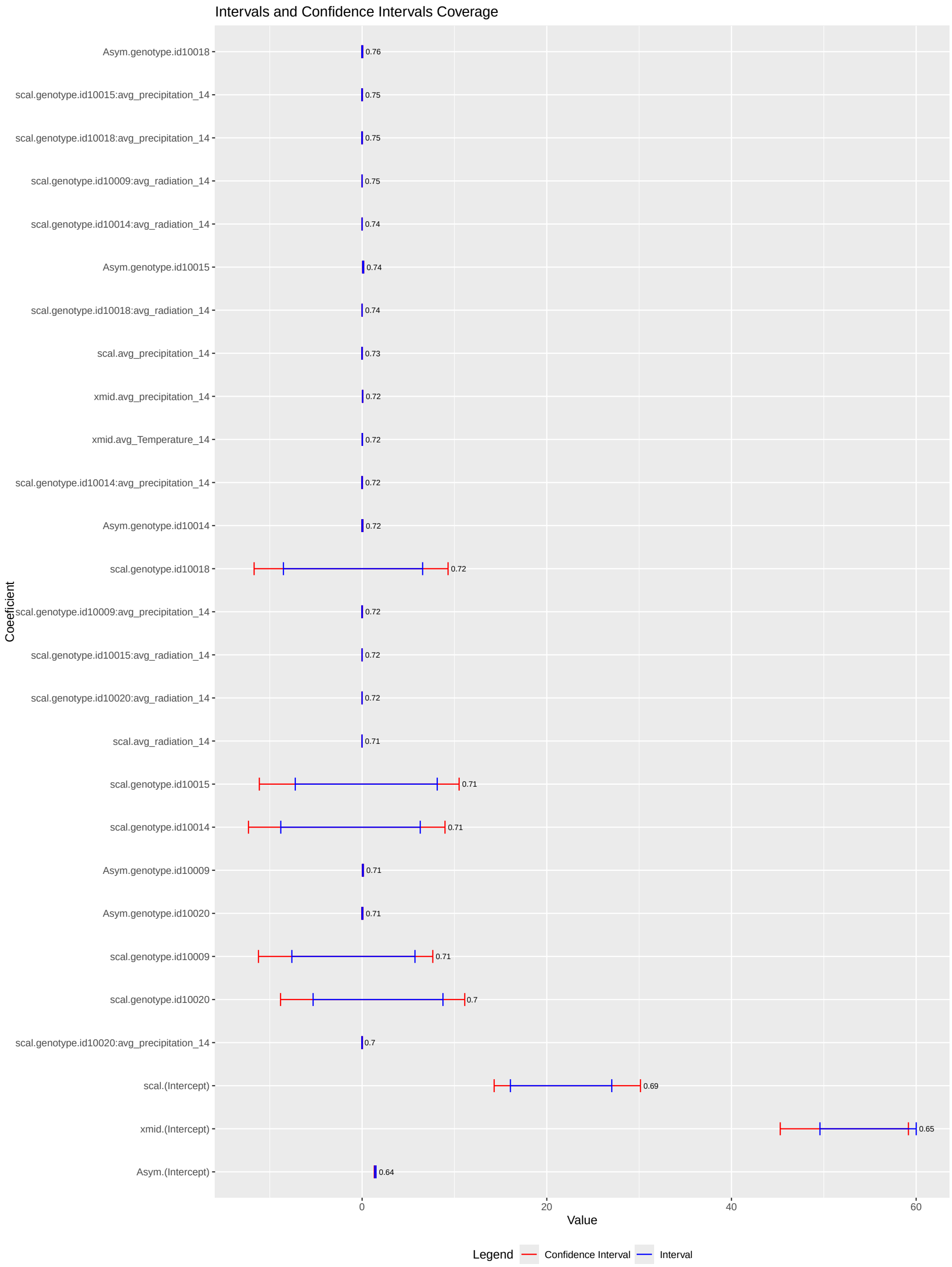


Figure 13: Comparison of the confidence intervals

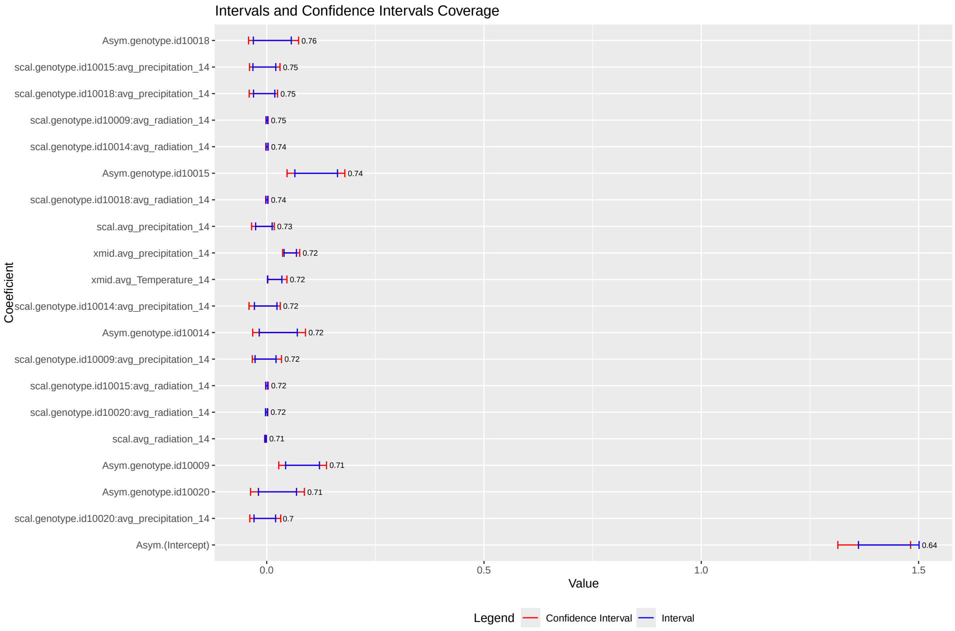


Figure 14: Comparison of the confidence intervals: Narrow

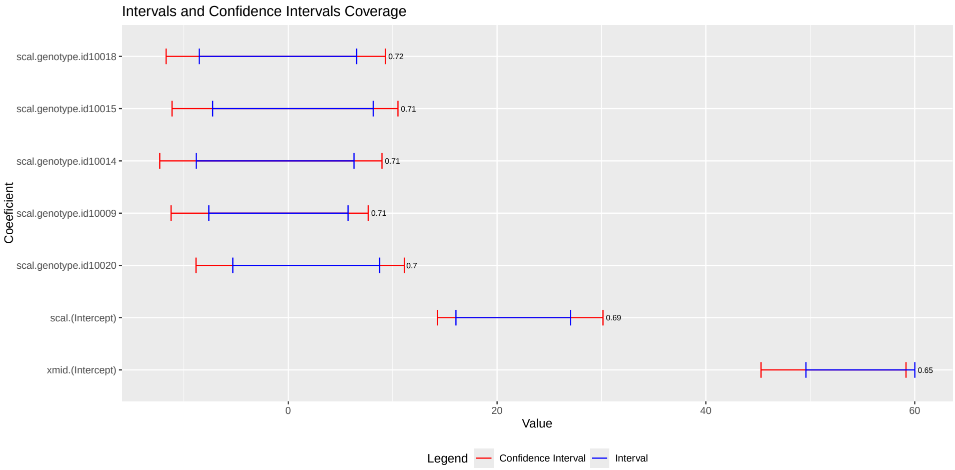


Figure 15: Comparison of the confidence intervals: Wide

9 Repository

In this section, we will investigate how to navigate the repository we build. You can find the repository under `https://gitlab.ethz.ch/marlob/canopy-cover-stats-lab`. It is logically divided into files for certain tasks:

- **data_pipeline_soybeans.R** takes in the *soybean_pixels_data.csv* file and constructs the growth definition from section 2, imputes missing dates. It outputs *soybean_data_for_modelling.csv.*
- **data_pipeline_weather.R** takes the *Weather_imputed.csv* file and performs the feature engineering. Importantly, it creates aggregated weather variables for different lags. It outputs *weather_data_for_modelling.csv.*
- **data_pipeline_join_df.R** joins the two previously created data frames while also renaming and dropping columns. Lastly, we perform the grouping defined in section 3. The output is the final model file names *model_data.csv.*
- **modelling.R** performs the main modelling of section **??**. This also includes submodels that are used for outlier detection and getting good starting values for the final model.
- **diagnostics.R** uses the final model to make the diagnostic plots of section 5.
- **anova_weather.R** uses the final model and submodels to the ANOVA tables from section 5.
- **plots_for_genotypes.R** is a file that helps to make prediction plots for certain genotypes.
- **get_ideal_candidates.R** finds the ideal candidates (see section 6) and gets the coefficients and confidence intervals for their environmental variable effects.
- **bootstrap-modelling.R** calculates the bootstrapped confidence intervals for a given subsample (section 8).
- **bootstrap-visualization.R** makes the visualization of how the bootstrap method works differently from the intervals() function (figure 13, 14, and 15).
- **uav_fip_modelling.R** computes and compares the different platforms with a simple model as described in section 7.
- **uav_fip-visualizations.R** depicts the diagnostics of the model from the previous file.

10 Conclusion

In this project, we modelled the growth of soy plants that had been planted on two experimental fields of land over the course of eight years. The analysis builds on the code base of a prior project. To conduct our analyses, we created a new data pipeline that makes it easy to process new data. We newly defined the end of the growth period so it entails the entire growth of the soy plant without the fluctuations close to the maximum canopy cover, and we newly defined the spatial grouping structure of the plots of land.

Following extant literature, we used a three-parameter logistic mixed effects model for modelling the growth of soy plants. As the target variable, we used the canopy cover instead of the plant height. According to biological theory, we explained the model's parameters by the environmental variables precipitation, temperature, and radiation. To account for spatial effects, we divide the field into grouped plots for which we estimate random effects. According to performance metrics and diagnostics plots, the model seems to fit very well. Model comparisons show that adding environmental effects and spatial random effects significantly improves the fit of the model.

A key research question of our client was to find genotypes whose growth curve does not change with different levels of environmental conditions. While we cannot directly test for robust genotypes, we posed criteria that ideal genotypes must fulfill. Following these criteria, we were able to identify six candidates with ideal characteristics: Genotypes number 10004, 10009, 10014, 10018, and 10020. We propose that further (experimental) tests are carried out on these candidates.

Further, our client was interested in investigating the fit of the model depending on the measurement platform that was used for measuring the canopy cover of the soy plants. We receive significant differences in intercepts for all the parameters specifying the logistic growth curve. However, as they are perfectly confounded, we cannot say whether it is due to the location or measurement platform. Further analysis needs to be done to compare the platforms directly.

Since the built-in intervals function works for non-linear mixed effect models without giving warnings despite not being particularly mentioned in the documentation, we decided to compare the results of that function to the intervals yielded by bootstrapping. The bootstrap method, in the majority of cases, produces wider confidence intervals, and the combined coverage ranges from 64-77%. As the method is computationally more expensive, we suggest sticking to the classic *intervals()* in case of many observations.

In conclusion, this project managed to yield a reproducible pipeline for accurately modelling the growth of different soy plant genotypes in varying environmental conditions. We propose candidates for robust genotypes, give an intuition and ideas for further tests regarding different measurement platforms, and contribute to the scientific discussion regarding correct inference in non-linear mixed effects models.

Appendix

1 Nonlinear Mixed Effects (NLME) Model

A Nonlinear Mixed-Effects (NLME) model is a statistical model used to analyze data that includes both fixed and random effects. This model is particularly useful when data is collected from multiple subjects or experimental units, as it accounts for variability both within and between these units. NLME models are commonly employed in fields where data exhibit hierarchical or clustered structures, such as clinical research, developmental studies, or growth modelling.

The model can be generally described in the following form:

$$y_{ij} = f(\boldsymbol{\theta}_i, \boldsymbol{x}_{ij}) + \epsilon_{ij}$$

where:

- y_{ij} is the response variable for the i-th subject at the j-th observation.
- $f(\boldsymbol{\theta}_i, \boldsymbol{x}_{ij})$ is a nonlinear function of the parameters $\boldsymbol{\theta}_i$ and explanatory variable $\boldsymbol{x}_{ij}$ of length r.
- ϵ_{ij} is the normally distributed noise term

The parameter vector $\boldsymbol{\theta}_i$ can be decomposed into the fixed effects and $\boldsymbol{\mu}$ and random effects $\boldsymbol{\alpha}_i$ together with the design matrices $\boldsymbol{U_i}$ and $\boldsymbol{A_i}$:

$$\boldsymbol{\theta}_i = \boldsymbol{U_i}\boldsymbol{\mu} + \boldsymbol{A_i}\boldsymbol{\alpha}_i$$

where:

- $\boldsymbol{\mu}$ represents the fixed effects, which are constant across all subjects and are the focus of the analysis.
- $\boldsymbol{\alpha}_i$ represents the random effects for the i-th subject, assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$.
- $\boldsymbol{U_i}$ and $\boldsymbol{A_i}$ are design matrices of size $r \times p$ and $r \times q$ for fixed and random effects, respectively

This represents a general form of the nonlinear mixed effects model, as any nonlinear function of fixed and random effects can be expressed as $f(\boldsymbol{U_i}\boldsymbol{\mu} + \boldsymbol{A_i}\boldsymbol{\alpha_i}, \boldsymbol{x}_{ij})$.

The parameters of the NLME model, including the fixed effects $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, are typically estimated using maximum likelihood or restricted maximum likelihood methods. For further information on the NLME model, we refer a curious reader to Lindstrom and Bates (1990).

Bibliography

Archontoulis, Sotirios V. and Fernando E. Miguez (Mar. 2015). “Nonlinear Regression Models and Applications in Agricultural Research”. In: *Agronomy Journal* 107 (2), pp. 786–798. ISSN: 1435-0645. DOI: 10.2134/AGRONJ2012.0506. URL: https://onlinelibrary.wiley.com/doi/full/10.2134/agronj2012.0506%20https://onlinelibrary.wiley.com/doi/abs/10.2134/agronj2012.0506%20https://acsess.onlinelibrary.wiley.com/doi/10.2134/agronj2012.0506.

Beardsell, M. F., K. J. Mitchell, and R. G. Thomas (June 1973). “Transpiration and photosynthesis in soybean: Effects of temperature and vapour pressure deficit”. In: *Journal of Experimental Botany* 24 (3), pp. 587–595. ISSN: 00220957. DOI: 10.1093/JXB/24.3.587.

Lindstrom, Mary J. and Douglas M. Bates (1990). “Nonlinear Mixed Effects Models for Repeated Measures Data”. In: *Biometrics* 46.3, pp. 673–687. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2532087 (visited on 05/27/2024).

Meyners, Michael (Dec. 2012). “Equivalence tests – A review”. In: *Food Quality and Preference* 26 (2), pp. 231–245. ISSN: 0950-3293. DOI: 10.1016/J.FOODQUAL.2012.05.003.

Roth, Lukas, Christoph Barendregt, Claude Alain Bétrix, et al. (Feb. 2022). “High-throughput field phenotyping of soybean: Spotting an ideotype”. In: *Remote Sensing of Environment* 269, p. 112797. ISSN: 0034-4257. DOI: 10.1016/J.RSE.2021.112797.

Roth, Lukas, Christoph Barendregt, Claude-Alain Bétrix, et al. (2022). “High-throughput field phenotyping of soybean: Spotting an ideotype”. In: *Remote Sensing of Environment* 269, p. 112797. ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2021.112797. URL: https://www.sciencedirect.com/science/article/pii/S0034425721005174.

Stevens, François, Patrick Bogaert, and Bas van Wesemael (2015). “Detecting and quantifying field-related spatial variation of soil organic carbon using mixed-effect models and airborne imagery”. In: *Geoderma* 259-260, pp. 93–103. ISSN: 0016-7061. DOI: https://doi.org/10.1016/j.geoderma.2015.05.008. URL: https://www.sciencedirect.com/science/article/pii/S0016706115001627.

Yang, Yi et al. (Feb. 2019). “Soil carbon sequestration accelerated by restoration of grassland biodiversity”. In: *Nature Communications* 10.1, p. 718. ISSN: 2041-1723. DOI: 10.1038/s41467-019-08636-w. URL: https://doi.org/10.1038/s41467-019-08636-w.

Zhang, Lihao et al. (May 2023). “Spatial variations in the response of spring onset of photosynthesis of evergreen vegetation to climate factors across the Tibetan Plateau: The roles of interactions between temperature, precipitation, and solar radiation”. In: *Agricultural and Forest Meteorology* 335, p. 109440. ISSN: 0168-1923. DOI: 10.1016/J.AGRFORMET.2023.109440.