

# Contrast Analysis and Multilevel Modeling in Food Science: A Case Study on Apples

M.A.J.S. van Boekel

Food Quality & Design Group, Wageningen University & Research, Wageningen, The Netherlands

## ABSTRACT

The power of Bayesian multilevel modeling and contrast analysis for food research is illustrated by meta-analysis of weight–diameter relations of apples (first level) at the factors cultivar, orchard site, year of harvest and tree (second level). Contrast analysis compares the effect of factors. Multilevel modeling connects levels, prevents over- and underfitting and accounts for dependencies in measurements/observations. A nonlinear power–law relation of weight–diameter is derived with parameters for spherical approximation and density. Literature data were analyzed in three ways: i) completely pooled (ignoring variation between groups), ii) per factor (no-pooling, ignoring similarities between groups) and iii) partially pooled (acknowledging variation and similarities, i.e. multilevel modelling). The power–law relation was obeyed at tree, orchard, harvest year and cultivar level. Very small differences were found per tree, orchard site and year of harvest, but differences between cultivars were more substantial. Parameter estimates from multilevel modeling differed between complete pooling and no-pooling. Spherical approximation of apples appeared reasonable, whereas apple densities differ considerably per cultivar. Bayesian analysis appears very suitable for model building, validation, predictive capacity and visualisation of parameter uncertainties. The approach can be applied in a much wider context and more complicated relations than considered here for apple characteristics.


## KEYWORDS

multilevel models; contrasts; Bayesian regression; apples; variability; predicting food quality

## Introduction

When products are harvested, all kinds of processes and handling actions may take place. Sorting and grading according to size are such handlings. Accounting for variation in properties is then an obvious aspect to consider. Natural products show inherent biological variability in composition, structure, size, colour and so on. This is due to variations in, among other things, genetic properties, climate conditions, agronomic conditions, management and storage conditions. Then, there is additional variation when making observations and/or performing experiments and doing analyses. In short, there are always uncontrollable and uncontrolled conditions as causes for variation. It is essential to be able to deal with such variability; statistics is then of course indispensable. What is not always obvious, perhaps, is that variation occurs at various levels, and that there are relations in the type of variation between objects. A statistical technique of multilevel modeling is designed to deal with this. It is a concept that has already been used for quite a while in the social sciences<sup>[1]</sup>, the medical and pharmaceutical sciences, e.g.<sup>[2–4]</sup> ecology (e.g.<sup>[5]</sup>). There are also recent textbooks dealing with multilevel modeling.<sup>[6–9]</sup> However, the concept seems to be relatively unknown in food science. Some

**CONTACT** M.A.J.S. van Boekel  Tiny.vanBoekel@wur.nl  P.O. Box 17, Wageningen 6700 AA, the Netherlands

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/87559129.2023.2228000>

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

exceptions exist, of course. It is used in food microbiology (e.g.,<sup>[10]</sup>), meta-analyses (e.g.,<sup>[11]</sup>), microbiological risk assessment<sup>[12]</sup> and food technology (<sup>[13–15]</sup>).

The present study attempts to show the power of contrast analysis and multilevel modeling in food research. It is illustrated with a meta-analysis of the relation between weight and diameter of apples from various cultivars. A meta-analysis can also be seen as a case of multilevel modeling where the lower level represents the individual studies while the higher level represents the central tendency of the trend across the various studies and, importantly, the uncertainty of that trend.<sup>[16]</sup> Furthermore, contrast analysis is used to study the effect of categorical variables where the goal is to compute differences between categories (analogous to ANOVA); the important aspect then is to compute the mean difference and not the differences between means.<sup>[17]</sup> The current paper starts with a very brief introduction of multilevel modeling, followed by an equally brief introduction of Bayesian analysis because it is the preferred method for contrast computations, multilevel and meta-analyses.<sup>[16,17]</sup>

### **Multilevel modeling**

Variability comes in many forms: between groups and between individuals/objects within a group or cluster. Imagine, for instance, that a researcher wants to learn about the polyphenol content in apples. Variation may be expected between apple cultivars, but also between apples of the same cultivar, while in addition there is variation in the region where apples are grown (due to a different climate, sunlight intensity, different soil type, management of the orchard, etc.). Sampling many apples to analyse polyphenol content without taking notice of such possible differences, will undoubtedly show variation that could be summarized by a mean and a standard deviation. Such an analysis is then at only one, single level. But in doing so, it is not seen whether or not there is perhaps a difference in cultivar or in regions, or between trees. Neglecting group effects and averaging obscures variation and leads to loss of information. Next to differences, similarities may also be expected between apples (for instance, the effect of ripening stage), so it would be nice if those common characteristics as well as the differences in individual characteristics are captured in models, thus accounting for variability at various levels. Multilevel models make it possible to preserve variation in the original data by modeling statistical phenomena at different levels (hence the name). Multilevel models are also known under the name of *hierarchical models* and *mixed effect models*, though the meaning may be slightly different. It all depends on how the data are obtained. The term ‘mixed effect models’ refers to a mix of fixed effects on the one hand, effects that are supposed to be explained as predictable at the population level (‘are fixed’) while there is also non-explained variance on the other hand that is called ‘random’ and these effects are thus variable. Statisticians nowadays seem to prefer the terms multilevel or hierarchical models. In any case, one should be aware that the terminology can be a bit confusing. According to,<sup>[9]</sup> it may be best to use the term ‘constant’ for effects that are assumed constant at the population level, and the term ‘varying’ for effects that are assumed to vary by groups or categories. A common characteristic of experimental design of many studies is that they result in dependencies between data. Common statistical analyses, however, assume statistical independence, meaning that the studied characteristics obtained should be completely independent. In practice this is not so easy to achieve. In shelf life studies, for instance, food samples are followed over time to study the change in quality attributes. If this concerns the same food sample, the measurements are not independent. If the same consumers judge a food over time, the measurements depend on the consumer and are not independent. If the growth of micro-organisms over time is studied, or their inactivation over time, dependencies are introduced in measurements. If samples are taken from the same reactor over time, they are not independent. If apples are taken from the same tree, there may be dependencies. There is an often overlooked phenomenon called ‘pseudoreplication’, e.g.<sup>[18–20]</sup> It means that the number of samples is overestimated by a certain experimental design. An example may help to appreciate this. Suppose one measures the polyphenol content of an

apple over time by some non-invasive measurement, say, 2 times per month over a period of three months, then there are 6 measurements per apple. However, these measurements are not independent because the same apple is measured all the time. Counting the number of measurements as 6 would lead to pseudoreplication; the actual number of measurements is 1 (if the researcher sticks to one apple). The result is that the statistical evaluation is inflated if it would be counted as 6 measurements. The example of measuring the same apple six times is known as longitudinal analysis, and can be handled as being multilevel, which takes into account the dependency in the data. Knowing how experiments are designed, or how observations are obtained, is thus very important to avoid pseudoreplication in the subsequent statistical analysis. One advantage of multilevel models is thus that they take these dependencies into account. This is the reason for the remark<sup>[17]</sup>:

Each proposed regression model should in principle be a multilevel one; use of a single level model should be explained

A major advantage is thus that multilevel models can handle dependencies between individuals from the same group, as well as the dependencies that arise when multiple observations are made on the same individual subject of investigation. By modeling the variation at group level, one does not need to model everything at the individual level, thereby reducing the number of parameters considerably. A second benefit is that if some groups contain not a lot of data, they can ‘borrow strength’ from the other groups, in other words, information is carried over between groups. When there are predictors at different levels of variation, use of multilevel models is recommended. So, multilevel modeling leads to:

- increase in the scope of inference
- better characterization of system uncertainty
- more efficient estimation.

In short, multilevel models are thus well suited for more complex data structures that result from longitudinal data, nested and crossed data, repeated measures design and block designs.

Regression analysis basically has two goals: i) understanding relationships between variables and the associated parameters, and ii) making predictions based on established relationships. The two goals can go together, of course, but in this paper, the second goal of prediction is the focus of attention. A possible danger in making predictions is over- and underfitting. Overfitting may happen when a model gets “too excited about the data”,<sup>[17]</sup> such that it depends too much on the data on which it is built and will not predict future observations well. Underfitting occurs when the model does not capture all the information from the data. Multilevel modeling helps in that sense to prevent over – as well as underfitting because it combines information from different data sources.

How does this work? Statistical models are built upon parameters for which inference is made. A parameter can also be seen as a placeholder for a missing model. If there is another model explaining how a particular parameter gets its value, this other model could be embedded in the original model, so that the models are built at different levels. In this way, models with multiple levels of uncertainty arise, each one feeding into the next one. Fixed effects pertain to variables that have an effect on the response variable of interest (similar to the explanatory variables in a standard regression model). Random effects pertain to categorical factors, groups, that can be controlled, not as an explanatory variable but to investigate how they may influence patterns. It is also a recognition of the fact that there are only finite sampling possibilities in research endeavors, resulting in uncertainty. So, if apples are collected to estimate the relation between weight and diameter, not all trees in the world can be sampled, obviously. The random effect then manifests itself in an estimation of the variation in the population from which it is sampled.

In this paper, the concept of multilevel modeling is explained by using a case study where the size (diameter) and weight of apples were measured while it was also recorded – in some cases – from which tree, orchard and cultivar the apples came from. The case study is trivial in the sense that it concerns a simple relationship between size (diameter) and weight of apples but it becomes non-trivial

when a tree effect or cultivar or orchard effect is taken into account. Because it is a simple case, it lends itself well to illustrate the multilevel method as well as a contrast analysis.

### **The Bayesian approach**

The Bayesian approach has gained much interest in the past decades. This is not because it is new; in fact, it is much older than the now traditional frequentist approach, which has become dominant in the 20<sup>th</sup> century. The reason for its increased popularity is due to the advancement of computer science and software. Some basics were discussed in relation to food science,<sup>[21]</sup> and the general approach can be found in textbooks.<sup>[6,7,9,17,22,23]</sup> This will not be repeated here, except for some very basic elements so that the uninitiated reader is able to follow the reasoning. The traditional statistical method is the so-called frequentist approach; it uses frequencies of events as a measure for probability. It considers parameters as unknown but fixed; what is considered variable is the data: obtained data from a particular sampling event, and other trials would yield different results. It leads to point estimates of parameters, p-values and null-hypotheses significance testing. The Bayesian approach, on the other hand, considers parameters as variables; they are characterized by probability distributions, not by a fixed value. Furthermore, the Bayesian approach requires expert knowledge to state something sensible about parameters before looking at the data; this is called the prior distribution. The prior when combined with the data (expressed in a likelihood function) yields a joint posterior distribution; this can be seen as an update of existing knowledge via information contained in the data. A posterior is a probability density distribution reflecting all that is known about parameters in view of prior knowledge, the proposed model and the obtained data. It is called a Bayesian approach because Bayes' theorem is used to reverse probabilities. A likelihood distribution tells something about how likely data are in view of a model or a hypothesis; this is where the frequentist method leaves it: it focuses on how probable the obtained data are. In the Bayesian approach, the focus is on how likely a model or hypothesis is in view of earlier knowledge and the data that are obtained. It does not result in point estimates and p-values (though Bayesian p-values do exist but with a different meaning). Bayes' rule is used to calculate the probability of a model/hypothesis/parameter from the likelihood of data and the prior expectation of a model/hypothesis/parameter. The reason that the Bayesian approach has become much more popular in recent years is that software has become available that numerically approximates posterior distributions based on the input of the prior and likelihood function. This numerical approximation routine is called Markov Chain Monte Carlo (MCMC). The state-of-the-art software in this respect is the statistical language called Stan.<sup>[24]</sup> To avoid confusion with the frequentist approach, the word significant is not used in the Bayesian context, a commonly used alternative is the word credible. Similarly, the term 'confidence interval' is not used, but instead the term 'credible interval' is used (the two concepts are conceptually different, though they may coincide numerically). McElreath<sup>[17]</sup> coined an alternative term 'compatibility interval' instead of 'credible interval' to indicate that the reported parameter values are compatible with the model and the data. A commonly used statement in the frequentist world that values are significantly different from zero is not used in the Bayesian context. Rather, statements can be made about probabilities of hypotheses and their differences and the judgement of whether or not that is relevant is left to the researcher rather than to a statistical test.

### **Contrast analysis**

Frequently, researchers want to investigate whether or not there are differences caused by treatments (in the case of experiments) or due to different conditions (in the case of observations). In the present context, a research question could be as follows: are apples harvested on one orchard different in size than harvested on another? A well-known statistical method to test such questions is ANOVA (Analysis of Variance), which analyzes differences in means between groups that received different treatments or were obtained under different

conditions. It does this by comparing variances within groups and between groups. If the ratio of the in-between variance is larger than that of the within variance a difference may be declared significant, or not, when this ratio is compared to a threshold from the Fisher probability distribution. This is commonly tested at the 5% significance level, and it is clearly a frequentist concept. However, if more than two groups are tested, ANOVA does not tell which differences are significant, if any, only that there are differences. To learn about which group means are different and how much, so-called post-hoc analysis is needed. A contrast analysis is such a post-hoc analysis: it compares differences pairwise in the case of more than two treatments. This can be done in the frequentist as well as in the Bayesian approach. In the frequentist approach, it is about differences in marginal means, whereas in the Bayesian approach, it is about differences in whole posterior distributions (a marginal mean refers to a mean resulting from a statistical model, not the mean of the raw data). Moreover, in the Bayesian approach, there is no significance testing, it is left to the researcher whether or not differences are relevant or not.

## Methods and data

### Methods

As already mentioned, the Bayesian approach lends itself better for multilevel modeling than the frequentist approach according to many authors,<sup>[17]</sup> though it is also done in the frequentist way (e.g.,<sup>[25–27]</sup>). The reason multilevel modeling follows naturally from the Bayesian approach is because of the very fact that parameters/models are considered variable, i.e., can be characterized by a probability distribution, as discussed above. In the current paper, multilevel modeling is applied in the Bayesian framework. The applied software was R version 4.2.2,<sup>[28]</sup> used in RStudio version 2022.12.0.<sup>[29]</sup> Bayesian regression was done using the package brms,<sup>[30,31]</sup> version 2.17.0. brms is an interface between R and the probabilistic language Stan, which is very suitable for Bayesian statistics<sup>[24,32]</sup>; it uses state-of-the-art Markov Chain Monte Carlo (MCMC) calculations, see also.<sup>[21]</sup> It is always necessary to check the MCMC performance to see whether or not the algorithm converged and the model results make sense; several diagnostics are available for that.<sup>[21]</sup> In this paper, such checks are not shown, but they were done and some of them are reported in the Supplement, section 4. In all cases reported below, these checks were found to be OK. The R codes and data sets can be found on the author's Github page <https://github.com/TinyvanBoekel/apples>. All the R packages used are listed in the Supplement (section 13).

### Data

A case study is done with data obtained from literature to illustrate the concept; it is about the relation between diameter and weight of apples. Some data were extracted from papers by digitization using WebPlotDigitizer <https://automeris.io/WebPlotDigitizer/>. Other data were kindly supplied by authors (see Acknowledgments) for Elstar and Pinova cultivars,<sup>[33]</sup> for red Elstar and Holyday<sup>[34]</sup> and for Fuji, Gala, Honeycrisp cultivars.<sup>[35]</sup> Generally, researchers are looking for a relation between weight and diameter, for instance, as a quick method to estimate the size of apples: since weighing apples is simpler and quicker than measuring their size, a statistical relation between the two with weight as the predictor for size would make this possible. Weights and diameters of apples of the same and different apple cultivars were measured and, in addition, in some cases it was recorded from which tree the apples came from and from which orchard and the year of harvest. This extra information was not used in the original papers but kindly supplied to the present author. It opens up the possibility of investigating whether or not there is perhaps a specific cultivar/orchard/tree/year of harvest effect on the relation between diameter and weight.

(To avoid an excessive number of Figures in this article, informative but not essential Figures are shown in the Supplement; reference is made to these Supplement Figures in the article.)

## Results

### *Characterization of the data: exploratory data analysis*

It is always a good idea to check out the available data to make sensible decisions on how to analyze them.<sup>[36]</sup> The available data can be subdivided at the level of cultivars, sites, year of harvest and trees.

### *Overview of all cultivars*

As a first impression, all cultivars are shown separately as well as in one scatterplot: see Figure S1 in the Supplement. It makes clear that the data cover a wide range of weight and diameter and that there could be a difference between cultivars, making it worthwhile to investigate this quantitatively (in statistical jargon: the variable ‘cultivar’ can be seen as a *categorical* factor, so not a numerical continuous variable). Figure S1 also shows that, overall, the relationship between weight and diameter is of a curvilinear nature. Some authors assumed a linear relationship,<sup>[34]</sup> others a second-order polynomial<sup>[35]</sup> or a power-law model<sup>[33]</sup> to describe the relation between weight and diameter. It is to some extent arbitrary whether diameter should be a function of weight or the other way around, and both relations can be found in literature. Here, the reasoning of Clarke<sup>[37]</sup> and Pflanz et al.<sup>[33]</sup> is adopted that, from an experimental point of view, it is easier to predict diameter from weight than the other way around, in other words, weight will be considered the predictor variable for diameter in this meta-analysis.

### *Orchard site and year effect*

For some cultivars, an orchard effect can be investigated as apples were picked at different sites, while the same cultivars were sampled from different orchards in two subsequent years (2016 and 2017), so orchard site and year are now the categorical factors. Figure S2 in the Supplement shows the results (the codes rs, bd and bv refer to orchard sites Rock Springs, Bedford and Biglerville, respectively, in PA, USA; full geographical details are available).<sup>[35]</sup> There could be some differences, also depending on the year, but it is hard to say from the plot how big or small differences are. Moreover, it is dangerous to draw conclusions just based upon eye-balling, a proper statistical analysis is needed, as attempted below.

### *Tree effects*

Despite being from the same cultivar, apples may differ by tree. Moreover, apples picked from the same tree are not independent in the statistical sense, and their properties may be correlated because they have the tree in common. If that is the case, this has its bearing on statistical analysis. Data showing a possible tree effect for all available cultivars can be found in the Supplement. By way of example, Figure S3 in the Supplement shows an extended exploratory data analysis for the cultivar Pinova with box plots and scatter plots. In any case, there seem to be differences for apples coming from different trees but again it is hard to quantify without a proper statistical analysis.

This finishes the section on exploratory data analysis, which shows that there are possible effects that are interesting to further explore. Also, the curvilinear relationship between weight and diameter is obvious and the nature of this relationship is discussed in the next section.

### *A theoretical model for the relation diameter-weight of apples*

The figures in the Supplement show variability in the data, depending on the factors considered. The question now is how such data are best analyzed. It depends, of course, on the

research question at hand. In this case, the question was how diameter (as a dependent variable) relates to weight (as the independent or predictor variable). This question can be investigated by applying regression, which assumes a relation between variables, a relation that might be causal or not causal. Here, a causal relationship is derived, as explained next. Most publications about the relation between diameter and weight of apples use an empirical relation such as a linear model,<sup>[34]</sup> a higher-order polynomial<sup>[35]</sup> or a power-law relation.<sup>[33]</sup> Clarke,<sup>[37]</sup> on the other hand, derived a theoretical relation between weight and diameter on the assumption that apples are perfect spheres. That same analysis is followed here (Clarke<sup>[37]</sup> studied the cultivars Cox orange Pippin, Golden Delicious and Bramley but, unfortunately, the raw data he used are not publicly available). If an apple is considered as a sphere, the relation between volume  $V$  (m<sup>3</sup>) and diameter  $d$  (m) is:

$$V = \frac{1}{6}\pi d^3 (\text{m}^3) \quad (1)$$

The relation between weight  $w$  (kg) and volume  $V$  is dictated by density  $\rho$  (kg/m<sup>3</sup>) as depicted in the following equation:

$$\rho = \frac{w}{V} (\text{kg}/\text{m}^3) \quad (2)$$

If Equations (1) and (2) are combined, the following relation between diameter and weight is found:

$$\frac{w}{\rho} = \frac{1}{6}\pi d^3 \quad (3)$$

Algebraic rearrangement results in Equation (4):

$$\begin{aligned} d^3 &= \frac{6w}{\pi\rho} \\ d &= \left(\frac{6w}{\pi\rho}\right)^{\frac{1}{3}} = c \cdot w^{\frac{1}{3}} \\ c &= \left(\frac{6}{\pi\rho}\right)^{\frac{1}{3}} \end{aligned} \quad (4)$$

In Equation (4)  $c$  is a constant if  $\rho$  is a constant. The density  $\rho$  is, of course, not a fixed constant and possibly different per cultivar; the expected range of values is between 0.0005 and 0.001 g/mm<sup>3</sup>. The important conclusion from this small algebraic exercise is that, from a theoretical point of view, the relation between diameter and weight should be a power-law relation with two parameters, constant  $c$  and exponent  $n$ , where  $n$  is expected to be 1/3 and  $c$  to be between 12 and 16, if diameter is expressed in mm and weight in g. However, this theoretical relation is only an approximation since apples are not perfect spheres and also the density is not fixed. Therefore, it is better to apply the derived relation as a power-law model:

$$d = c \cdot w^n \quad (5)$$

and to estimate its parameters from the data rather than fix them at a certain value. Equation (5) is obviously a nonlinear model, not only because the relation between  $d$  and  $w$  is not linear but also in the statistical sense that the parameters are nonlinear. Therefore, nonlinear regression is needed and that will be done in a Bayesian framework in the next sections. Clarke<sup>[37]</sup> linearized Equation (5) by taking its logarithm:

$$\log d = \log c + n \cdot \log w \quad (6)$$

He then applied linear regression to find parameters  $\log c$  and  $n$  (in this respect, it may be interesting to read about the “logarithms of apples”<sup>[38]</sup>). Here, a different route is taken by using Bayesian nonlinear regression with Equation (5) as the basic model.

## A statistical model

A statistical model implies in the Bayesian context that priors and likelihood need to be proposed for the parameters and the data, respectively. It does not seem unreasonable to propose that the data follow a normal distribution (more specifically, the resulting residuals) or, to put it differently, to assume that the response variable diameter  $d_i$  is generated according to a normal distribution with mean  $\mu_i$  and standard deviation  $\sigma_e$  where the standard deviation is supposed to be constant for all data. This assumption is shown in Equation (7):

$$d_i \sim N(\mu_i, \sigma_e) \quad (7)$$

and represents the likelihood of the data. (An alternative could be a log-normal distribution, which allows only positive diameters. A normal distribution runs in principle from  $-\infty$  to  $+\infty$  and diameters can only be positive. However, with the distributions used for apple diameters, the probability for diameter values  $<0$  is virtually zero so that a normal distribution will also suffice.) The regression routine will yield an expected response value  $\mu_i$  that is supposed to change non-linearly with weight  $w_i$  according to the derived power-law model displayed in Equation (5):

$$\mu_i = c \cdot w_i^n \quad (8)$$

The task is now to find estimates of the parameters  $c$ ,  $n$  and  $\sigma_e$  from the available data ( $\mu_i$  does not need to be estimated as it follows directly from Equation (8) when parameters  $c$  and  $n$  have been estimated). These three parameters need prior distributions in the Bayesian framework. Priors should be based as much as possible on expert knowledge: what can be said about these parameters before the data are known? From the theoretical analysis above, it follows that the numerical value of parameter  $c$  is expected to be between 12 and 15 and the exponent  $n$  to be around 0.33. To indicate that there is uncertainty about these values, a relatively large standard deviation can be given to the prior distribution, which will make it a weakly regularizing prior, as indicated in Equation (9):

$$\begin{aligned} c &\sim N(12, 2) \\ n &\sim N(0.3, 0.03) \\ \sigma_e &\sim \text{exponential}(1) \end{aligned} \quad (9)$$

A weakly regularizing prior means that the MCMC software is gently pushed into a certain direction in the sense that impossible values are avoided while allowing all values that are possible in principle. The two parameters  $c$  and  $n$  are given a normal distribution and the standard deviation  $\sigma_e$  an exponential distribution; this latter distribution prevents negative values (a standard deviation cannot be negative, other possible priors for a standard deviation that are used in literature are a half-Cauchy distribution, a log-normal or a gamma-distribution, which all prevent the parameter from becoming negative). It is instructive to perform simulations with assumed priors (i.e., without considering data yet), to check whether or not the priors yield reasonable results.<sup>[17]</sup> The ultimate objective of this action is to prevent the model from coming up with impossible values (like negative diameters) but also to come up with extreme but not impossible values. This procedure is called a prior predictive check.

## Prior predictive check

Above, priors and likelihood have been presented. As mentioned, it pays off to check whether or not the proposed model yields reasonable values. This is possible because Bayesian models are generative, that is to say that data can be simulated from the assumed distributions. This is done with the priors shown in Equation (9) and yields simulated regression lines as shown in the Supplement in Figure S4, showing that a wide range of diameters is covered with these priors, some even absurdly wide (compared to the experimental data in Figure S1). Extreme values are thus seen to be possible; these priors can be considered as weakly informative as they allow a wide range of outcomes, also with some



extreme but not completely impossible outcomes. Having established reasonable priors, the next action is to perform actual regression by combining the priors with the likelihood for the data. This is the task of a statistical model and is the problem in reverse compared to a generative model: estimating the values of model parameters from the observed data.

### Bayesian regression of all data

Having established a scientifically based power-law model as well as a statistical model, Bayesian regression analysis can now be applied to the data and is, as a first exercise, applied to all the available pooled cultivar data. Since it concerns a nonlinear model, nonlinear regression is needed but when done in the Bayesian way there is not really a difference between linear and nonlinear regression. The only possible difficulty with nonlinear models can be that the MCMC procedure runs into numerical problems with very complex models, but this is not expected with a simple power-law model as used here. Figure S6A in the Supplement shows the posterior parameter densities, pair plots and correlation coefficients, resulting from complete pooling of all the cultivar data together. The very strong parameter correlation between  $c$  and  $n$  stands out, which is due to the mathematical formulation of the power-law model. This is not necessarily a problem because this information about correlation is stored in the posterior and is automatically taken into account when doing calculations with the posterior. This is different with frequentist results where further calculations need to work with covariances that are not always supplied by software programs (Microsoft's Excel is one of them, but, fortunately, that can be remedied by using macros).<sup>[39]</sup> However, it is better to avoid parameter correlation when possible. There is a way to avoid this strong parameter correlation by reparameterization of the predictor variable. For linear models, a well-known procedure to avoid parameter correlation is to center and standardize the independent variable (weight in this case).<sup>[17]</sup> A procedure specifically for the non-linear power-law model is described.<sup>[40]</sup> The derivation is as follows. Based upon a mathematical analysis of how correlation coefficients are built up, it can be derived that Equation (5) can be rewritten as:

$$\begin{aligned} d_i &= c_{rp} \cdot \left(\frac{w_i}{w_{rp}}\right)^n \\ c_{rp} &= c \cdot (w_{rp})^n \end{aligned} \quad (10)$$

The value for the reparameterized independent variable  $w_{rp}$  is calculated as in Equation (11) for  $N$  measurements:

$$w_{rp} = \exp \left[ \frac{\sum_{i=1}^N d_i^2 \cdot \ln(w_i)}{\sum_{i=1}^N d_i^2} \right] \quad (11)$$

As shown, parameter  $n$  remains as is, but parameter  $c$  is transformed into another parameter, called  $c_{rp}$ , where the subscript 'rp' stands for 'reparameterized'. Furthermore, the independent variable  $w$  is transformed into a new independent variable  $w/w_{rp}$  having the same dimension as  $w$  (g in this case). Note that  $w_{rp}$  is just a number. Also, note that the dependent variable ( $d$  in this case) is not affected by the reparameterization.

The likelihood for the data and priors for parameters  $n$  and  $\sigma_e$  in the above-described reparameterized model remain the same but the numerical value for  $c_{rp}$  needs to be adapted. Knowing that parameter  $c$  is somewhere in the range between 10 and 15 and from the relation between  $c$  and  $c_{rp}$ , the prior  $c_{rp} \sim \mathcal{N}(64, 5)$  was proposed as a weakly regularizing prior. Some MCMC checks can be found in the Supplement (section 2, Figure S5). The resulting pairs and parameter plot for the completely pooled cultivar data are in Figure S6B in the Supplement, showing that parameter correlation has completely disappeared when compared to Figure S6A. Obviously, the 'new' parameter  $c_{rp}$  has

a different numerical value than  $c$  but it can be easily recalculated back, including its uncertainty, into a value for the original parameter  $c$  from the information available in the posterior using Equations (10). It was checked that parameter values for  $n$  and  $c$  obtained without parameterization are identical to those obtained with parameterization by recalculation of  $c_{rp}$  into  $c$ . There was hardly any difference (results not shown), indicating that despite the high correlation in the non-parameterized case, estimation still went well, probably due to the use of weakly regularizing priors and the fact that it is a relatively simple model so that the MCMC software can handle the strong correlation. However, to avoid possible complications due to parameter correlation altogether, the reparameterized model will be used in the remainder of this article. In the resulting fit plots, the original weight value  $w$  will be plotted for ease of interpretation, even though the reparameterized weight value  $w/w_{rp}$  was used for estimation. That it makes no difference in the fit can be seen in Figures S6(C,D). (To be sure, this reparameterization is not a specific Bayesian thing, it can – and should – also be used in the frequentist paradigm.)

The fact that information stored in the posterior can be used in further calculations where correlations are automatically taken into account is yet another advantage of working with complete posterior distributions. In the frequentist framework, taking correlation into account in calculations is doable but less straightforward. It requires usage of the parameter variance-covariance matrix, which is not always supplied by software programs.<sup>[39]</sup>

To show that the priors were not dominating the outcome, Figure S7 in the Supplement compares the priors with the posteriors, showing clearly that the posterior parameter distributions are much narrower than the priors, which shows that the data are dominating in determining the posteriors.

Figure S6D in the Supplement shows the resulting fit of the model to the data with the 95% credible interval as well as 95% prediction interval for all data together. The model seems to do a good job; the 95% credible interval is almost coinciding with the regression line, which means that the expected value  $\mu_i$  (the mean) is extremely precisely predicted, due to the large amount of data. The 95% prediction interval for new, not yet measured values, however, does not cover all data and seems therefore to underestimate variation. Table 1 shows a numerical summary of the posterior obtained from the reparameterized model. The estimate of parameter  $n$  is lower than the theoretically derived value of 0.33, based on the assumption that apples are perfect spheres. The estimate of parameter  $c_{rp}$  is on the higher side of the expected value between 60 and 70. However, this is not where the analysis stops, below it will be shown that these first results are not completely trustworthy because of the complete pooling that was applied. But first some other possible effects on the regression outcome will be considered.

### Orchard site effects

Some data are available on apples from the same cultivar but harvested at different orchard sites. This gives the option to investigate whether or not there is an effect of site on the relationship between weight and diameter. Causal effects for such differences could be due to differences in soil, climate/weather conditions, crop load and management. Figure S2 in the Supplement gave a first impression but the question to be investigated now is whether or not there is a noticeable effect. Since there are data from only three sites available, this seems too low for a proper multilevel analysis (discussed below) for which some 5–6 different groups are usually thought to

**Table 1.** Numerical summary of the posterior parameter distributions resulting from Bayesian regression of the completely pooled cultivar data (reparameterized power-law model). SE = standard error. The lower and upper bounds represent 95% credible intervals of the parameters.

|                       | Mean  | SE    | Lower bound | Upper bound |
|-----------------------|-------|-------|-------------|-------------|
| $c_{rp}(-)$           | 69.77 | 0.027 | 69.71       | 69.82       |
| $n(-)$                | 0.31  | 0.001 | 0.31        | 0.31        |
| $\sigma_e(\text{mm})$ | 1.92  | 0.020 | 1.88        | 1.96        |

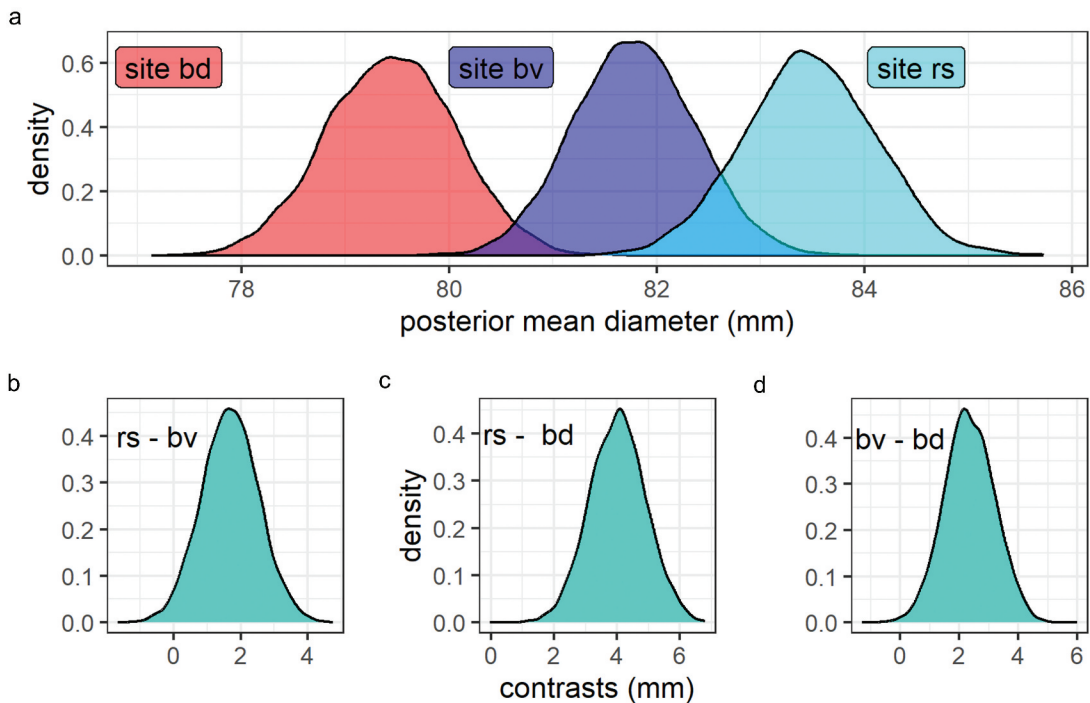
be required to give reliable results. However, what remains possible is to perform a *Bayesian contrast analysis*. It is imaginable that the orchard site has an effect on diameter in two ways. First, an orchard site may have an effect on weight, which, in turn, influences diameter. Second, there could also be a direct separate effect of site on diameter without an interfering effect via weight. The task is to unravel these two possibilities. The following analysis is for cultivar Fuji, harvested in year 2016. The other available data are analyzed in a similar way in the Supplement.

### Overall effect of site on diameter

The first question to address is whether or not there is an overall effect of site on diameter. **Figure 1a** shows the effect of site on the distribution of the posterior *average* value of diameter (where the width of the distribution indicates the uncertainty in estimating this expected value). The plot suggests differences but also overlaps due to site. To conclude whether these differences are statistically relevant, one needs to calculate the so-called contrast: this is the posterior distribution of the *difference* between the posterior distributions per site; this difference is simply obtained by subtracting one posterior distribution from the other. One should never *compare* means as such; instead, one should *calculate* mean differences, i.e., contrasts.<sup>[17]</sup> The result of doing this for the effect of orchard site is shown in **Figures 1b, c, d**.

The contrast plots show indeed differences between sites in average diameters, for instance, the average diameter on site rs is about  $4 \pm 1.5$  mm larger than on site bd.

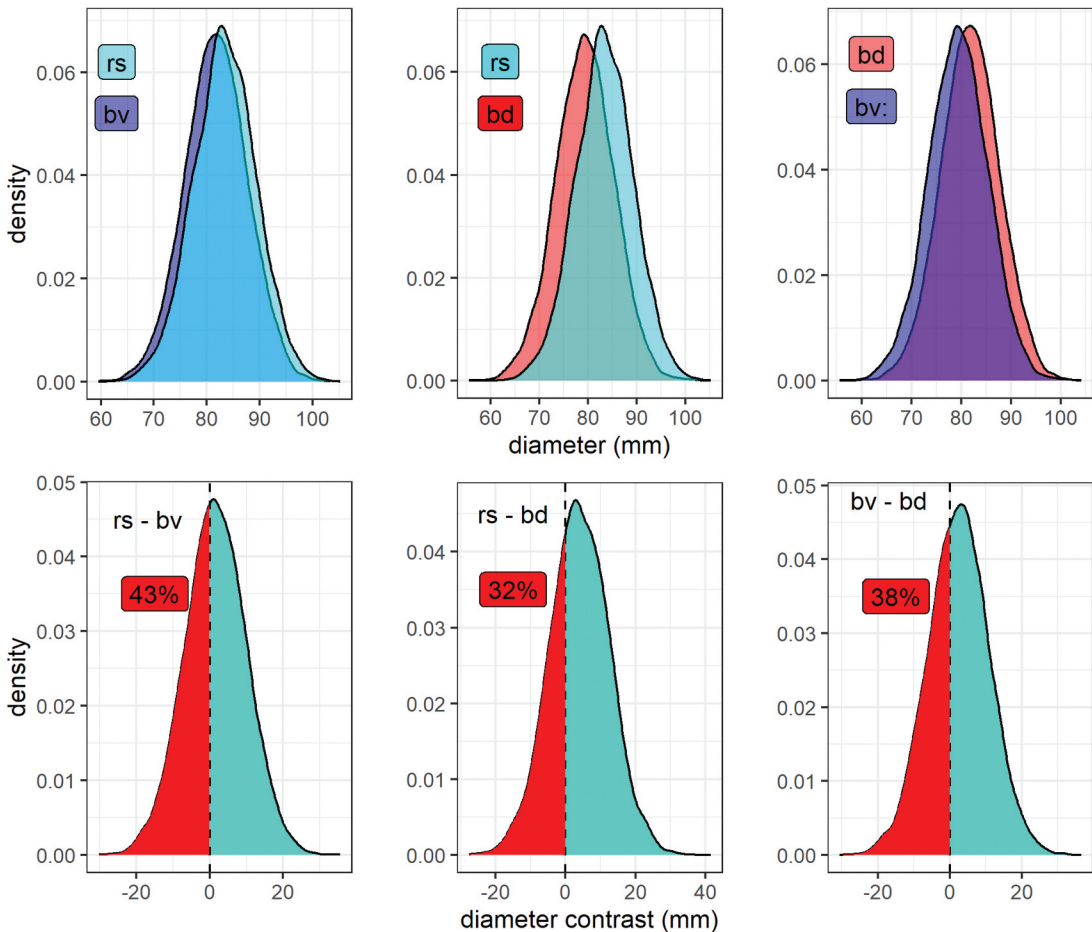
Contrasts can be calculated not just for the mean but also for the whole distribution of apple diameters. This is so because the posterior mean and standard deviation are now available as estimates so that a whole normal distribution can be simulated. This is called a posterior predicted distribution, in this case of diameters, and that can, of course, be done for each site. Two steps are thus needed to calculate such a contrast in the full distribution: i) the first step is to perform many



**Figure 1.** Probability distributions of the posterior average diameter for cultivar Fuji harvested in year 2016 on three sites (rs, bd, bv). b, c, d: contrasts in average diameter (mm) of Fuji apples for sites rs-bv, rs-bd and bv-bd..

simulations (e.g., 10,000 or so) using the mean and standard deviation obtained from the regression analysis; this results in posterior predicted distributions, and ii) the second step is to calculate the contrast of these two distributions by subtracting them from each other. [Figure 2](#) shows the results as whole distributions as well as the calculated contrasts of the pairwise differences in predicted diameter between two sites. Considerable overlap is seen between diameters measured for each of the two compared sites, even though, *on average*, differences remain. In other words, even though the average diameter on a particular site may be smaller than on another site, a considerable amount of individual apples on that particular site may be actually be larger.

From these simulations, the percentage of differences in diameters between sites can be calculated. For instance, for sites rs and bv, the percentage of apples larger on site rs in comparison to site bv is 57%, meaning that even though on average apple diameters are 1.7 mm larger on site rs than on site bv, still  $100 - 57 = 43\%$  of the fuji apples harvested on site bv are larger in diameter than the ones harvested on site rs (this latter percentage is marked in red in [Fig. 2](#)). This quantifies the overlap in the whole distribution. Similarly, the percentage of apples larger in diameter on site bd than on site bv is 38% and the percentage of apples larger on site bd than on site rs is 32%.



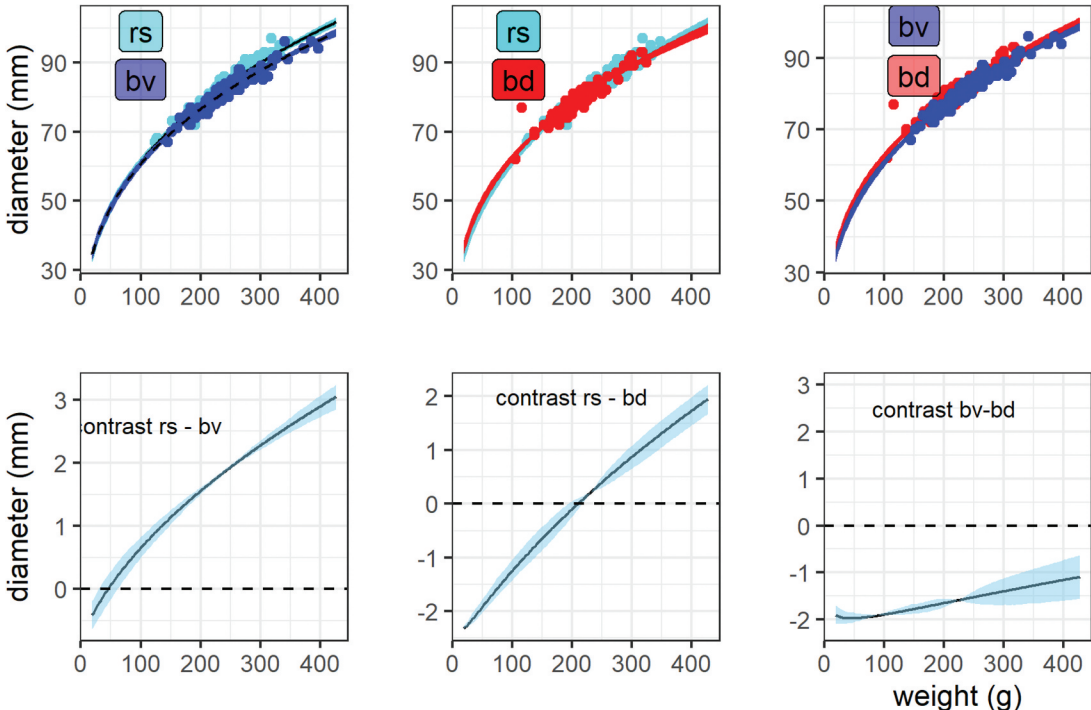
**Figure 2.** Upper row: posterior predicted whole distributions (10,000 simulations) of Fuji apples harvested on sites bv, bd and rs in 2016. Lower row: corresponding contrast plots of diameters; the red areas indicate the proportion of samples with contrast < 0.

Therefore, in conclusion, there seems to be a slight overall effect (i.e., indirect via weight and direct on diameter) of site on diameter in the order of 2–4 mm. Note that this analysis did not consider the variable weight at all, it was only investigated in how far measured diameter values differ per site. To account also for the variable weight, regression analysis needs to be done per site using the model shown in Equation (5).

### **Effect of site on the relation between diameter and weight**

To investigate the effect of site on diameter via the variable weight, nonlinear regression of diameter on weight is performed using the reparameterized power-law model, as was done before for all cultivars, but now only for cultivar fuji, year 2016 and three different sites. To see a possible effect of sites, regression is done per site (this is called ‘stratifying by site’ in statistical jargon). The top row in Fig. 3 shows the fits pairwise for the sites while the lower row in Fig. 3 shows the contrast between the fits (obtained by subtracting the respective posteriors). It reveals a small effect of site on the relation between weight and diameter.

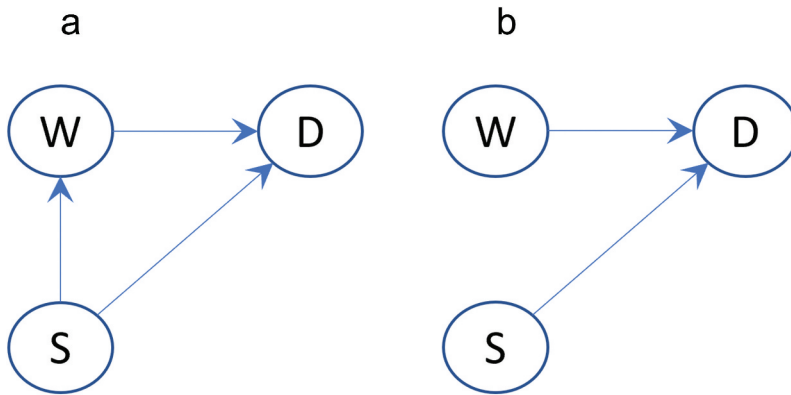
The question now is how to interpret the results in Fig. 3. Imagine that the contrast for the effect of site on the relation diameter – weight was zero, that would mean that site has no direct effect on this relation and the regression lines should coincide (within experimental error). However, an effect is noticeable, albeit small. Comparing sites rs – bv, apples tend to be a bit larger in diameter with increasing weight when harvested on site rs as compared to site bv. This is also the case in the comparison of sites rs-bd but there the lower weight apples harvested on site bd tend to be larger in diameter than on site rs while the opposite is true for the higher weights. Comparing sites bv and bd shows yet another picture: apples harvested on these two sites show a parallel relationship that does not change with increasing weight but there is a consistent difference in diameter: apples harvested on site bv are about 1.5 mm smaller in diameter over the whole range of weight values.



**Figure 3.** Fits resulting from regression of diameter on weight for cultivar Fuji harvested on three sites rs, bd and bv in 2016 (upper row) and contrast plots showing the pairwise difference in regression lines for the three sites rs, bv and bd. The ribbons indicate the 95% credible intervals.

In passing, it is noted that such a contrast analysis as performed here can also be described in the framework of DAGs (Directed Acyclic Graphs) and causal models.<sup>[17,41]</sup> A DAG for the effect of orchard site on the relationship between diameter and weight can be described as in Fig. 4a. The arrows indicate the relationships between the variables, which in this case proposes that site has an effect both on diameter and weight. In fact, variable S (site) shows a so-called backdoor path: it influences diameter indirectly via weight as well as directly. Variables that affect both the independent predictor variable and the dependent response variables cause this backdoor path. In order to find the true relationship between weight and diameter, this backdoor path needs to be closed, as shown in Fig. 4b. This is achieved by stratifying on site (i.e., study the relationship between weight and diameter separately per site). This is, in fact, what was done in this section on the effect of site. Without stratifying for site, the relationship is obscured by the variable site and regression analysis will happily report outcomes without any notification of confounds.

The same analyses were done for the Fuji harvest in 2017 and the Gala harvest in 2016, which are reported in detail in the Supplement. Numerical summaries of the contrast analyses for Fuji apples harvested both in 2016 and 2017 and the Gala apples harvested in year 2016 can be found in Table 2, using the function ‘hypothesis’ in the R package brms.<sup>[42]</sup> This function returns not only the numerical values for the differences in means of diameters and their uncertainties, it also reports the value for the contrast posterior probability. It indicates the proportion of contrast samples that is above zero. In cases where it is almost 1 or zero, it is very credible that the large majority of samples really differ



**Figure 4.** Directed Acyclic Graphs (DAG) showing a causal relationship between site (S), and diameter (D). a: indirect effect of S on D (backdoor of S via W). b: direct effect of S on D obtained by stratifying for site for the relation between W and D (backdoor path of S on W closed).

**Table 2.** Numerical summary of the contrast in mean diameter for sites rs, bv, bd in years 2016 (cultivars Fuji and Gala) and 2017 (cultivar Fuji). SE = standard error, lower and upper bounds represent 95% credible intervals, and the posterior probability indicates the proportion of samples from the posterior of the contrast that is above zero.

|                         | Mean  | SE    | Lower bound | Upper bound | Posterior probability contrast > 0 |
|-------------------------|-------|-------|-------------|-------------|------------------------------------|
| rs – bv, Fuji 2016 (mm) | 1.68  | 0.857 | 0.27        | 3.10        | 0.9760                             |
| rs – bd, Fuji 2016 (mm) | 4.03  | 0.891 | 2.56        | 5.50        | 0.9999                             |
| bv – bd, Fuji 2016 (mm) | 2.34  | 0.854 | 0.95        | 3.75        | 0.9970                             |
| rs – bv, Fuji 2017 (mm) | -1.44 | 1.255 | -3.52       | 0.63        | 0.1250                             |
| rs – bd, Fuji 2017 (mm) | 3.80  | 1.215 | 1.79        | 5.76        | 0.9982                             |
| bv – bd, Fuji 2017 (mm) | 5.23  | 0.999 | 3.61        | 6.89        | 1.0000                             |
| rs – bv, Gala 2016 (mm) | -2.60 | 1.011 | -4.28       | -0.93       | 0.0062                             |
| rs – bd, Gala 2016 (mm) | -0.06 | 0.982 | -1.69       | 1.54        | 0.4792                             |
| bv – bd, Gala 2016 (mm) | 2.54  | 0.701 | 1.37        | 3.70        | 1.0000                             |

between sites. In two cases ( $rs - bv$  for Fuji 2017 and  $rs - bd$  for Gala 2016), the differences between these sites are not so clear.

Table 2 shows that there is no consistent effect of site when compared over 2 years; for instance, Fuji apples are, on average, somewhat larger on site  $rs$  in 2016 but smaller in 2017 in comparison to site  $bv$ . So, all in all, the effect of site on diameter is noticeable but only on the order of a few mm. A similar picture emerges from the Gala data.

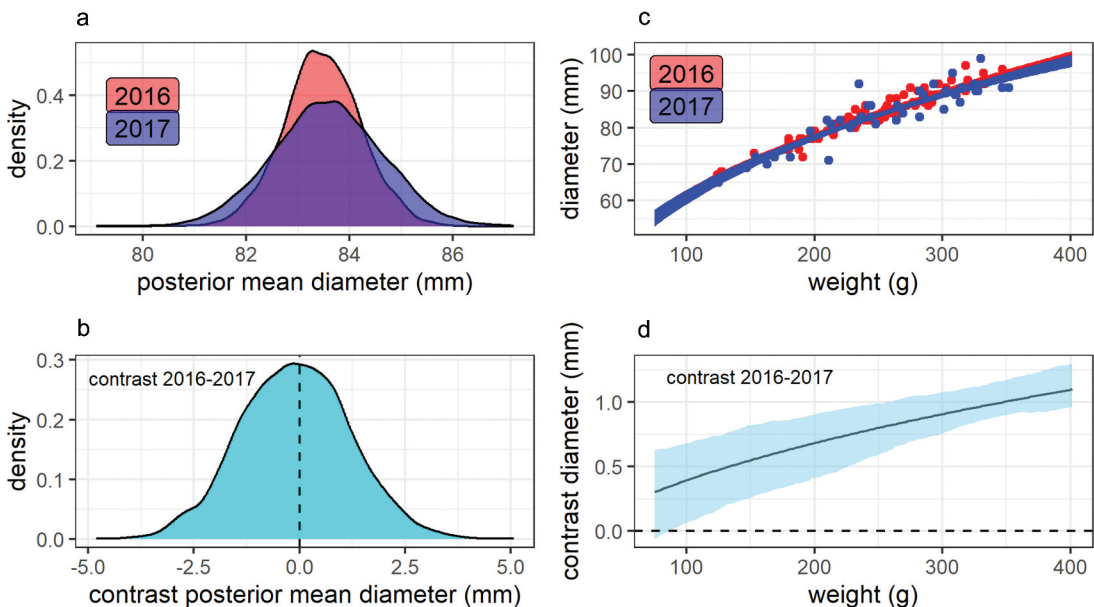
In conclusion about the effect of orchard site, credible differences can be shown, but overall, the effects are rather small, on the order of a few mm, which is not much considering that diameter values range from 50 to 100 mm. Nevertheless, this contrast analysis shows how to characterize such data and their mean differences to analyze data as shown in Figure S2 in the Supplement in a quantitative way. It is remarkable that even small differences are picked up by such a contrast analysis.

### Year effect for Fuji apples per orchard

As was already apparent from the previous section for cultivar Fuji, data are available for 2 years (2016 and 2017) and three orchard sites, which enables an analysis on whether or not there is an effect of year on the relationship between diameter and weight for each site. A possible causal effect for year could be different climate/weather conditions and/or different management conditions. The first analysis is on the total effect of year on diameter, the second on the effect of year on the relation between diameter and weight.

### Total effect of year for Fuji apples on site $rs$

The total effect of year on diameter is found by comparing this per year, in this case for site  $rs$ . The distributions of the posterior mean are shown in Figure 5a. The two distributions are seen to largely overlap (a somewhat larger spread in diameter in 2017), but the ultimate test is to calculate the contrast



**Figure 5.** a: Posterior density distribution of average posterior diameter for Fuji apples harvested on site  $rs$  in year 2016 and 2017. b: Contrast in posterior mean 2016–2017. c: Fit resulting from Bayesian regression of diameter on weight for year 2016 and 2017. d: Contrast for regression lines for 2016–2017; the light blue ribbon indicates the 95% credible interval.

between the 2 years as shown in [Figure 5b](#). This clearly shows the absence of any systematic effect of year on average diameter for site rs.

### ***Effect of year on the relation between weight and diameter for Fuji apples harvested on site rs***

By stratifying for year, it can be investigated in how far the relation between weight and diameter is influenced by the factor year. [Figures 5\(c,d\)](#) show the regression lines and contrast in predicted diameter for years 2016 and 2017. The figure shows a consistent trend: diameter increases with weight more in year 2016, but the effect is very small, amounting to 0.5–1 mm only.

So, all in all there is no effect of year on the average diameter, while there is a distinct, but very small year effect noticeable on the relation between diameter and weight, only 1 mm at most, so that can safely be neglected if one is interested in the relation between weight and diameter, as is the purpose of this article. A similar analysis and conclusion for the two other sites is reported in the Supplement: a distinct but very small effect of year and so pooling the 2 years does not lead to big mistakes. Even though the effect is seen to be very small, if one is interested in the possible effect of sites, this analysis opens up the possibility to explore that in a quantitative way, but then it would answer a different research question. Clarke<sup>[37]</sup> noticed some effect of year and site as well and he also concluded that these effects were limited in predicting diameter from weight over a whole range of weights.

Another way to show that year of harvest has hardly an effect is to make a prediction for the year 2017 using the regression result of year 2016 and then compare that to the actual observed data in 2017. This is shown in [Figure S16](#) in the Supplement with the 95% prediction interval, an interval that shows where new to be observed data are expected with a 95% probability. Nearly all the observed data are seen to fall within this prediction ribbon. In other words, when it comes to predicting diameter based on weight, there is no real difference between the years. Note that this is an independent check on the validity of the regression model since the data from 2017 were not used in the model for 2016. Hence, it is concluded that pooling the data from different years does not lead to big mistakes.

### ***Conclusion about year effect***

In conclusion, the effect of year is noticeable on some sites, but it is not a large effect. Of course, there were only 2 years to compare so it is impossible to generalize from these results, but the main purpose was to show how such effects could be analyzed if relevant data are present. For the moment, it is concluded that possible year effects can be neglected when it comes to predicting diameter from weight, at least for the data sets that are currently available. Nevertheless, it is remarkable that even such small differences can be detected by the applied procedure. For purposes other than predicting diameter from weight, this might be useful.

### ***Tree effects within a cultivar***

As a next step, it is investigated in how far a tree effect exists for the same cultivar. If it does, then the subsequent analysis for a possible cultivar effect depends on that, and if it does not, then apples from different trees but from the same cultivar can be pooled. Since for some data, but not all, it was registered which apples came from which tree, it can be analyzed in how far apples from the same cultivar differ per tree, in other words, tree then becomes the categorical factor within a particular cultivar. Here, this is investigated for cultivar Pinova for which data from 11 trees are available. A similar exercise is done for cultivars Elstar, Fuji and Gala in the Supplement. There are several ways in which such data can be analyzed. The first possibility is to pool all the data to obtain one overall relation (complete pooling), the second option is to do separate regressions per tree (no-pooling) and the third option is to do multilevel modeling (partial pooling). [Figure S3](#) in the Supplement already gave an overview of the Pinova data and displays some variation in the relation between diameter and



weight per tree, and the task is now to investigate if this difference is statistically relevant. The priors and likelihood were the same as shown in Equation (9).

### ***Complete pooling of the Pinova tree data***

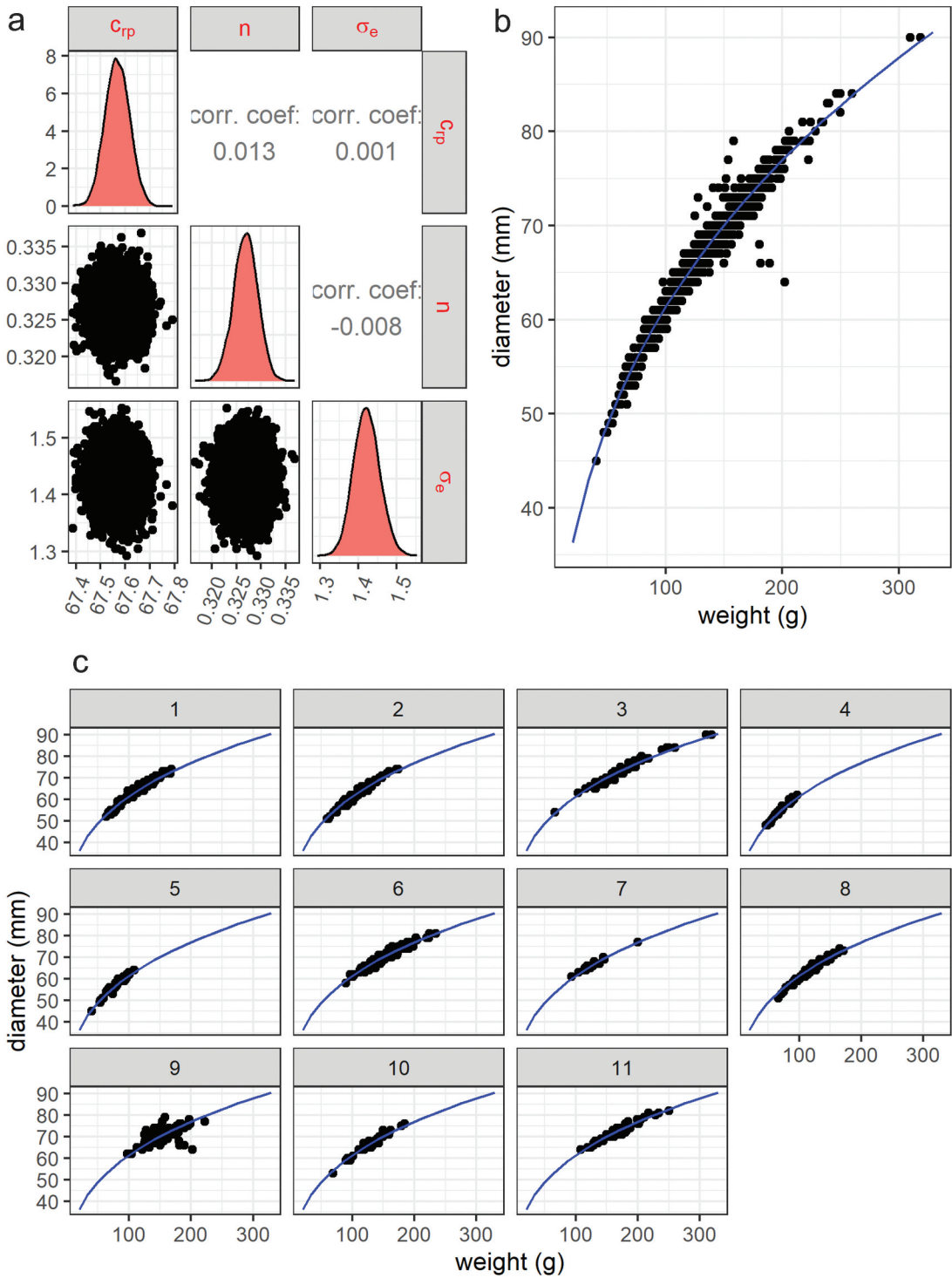
Figure 6a shows the result of Bayesian regression of the completely pooled Pinova tree data in the form of posterior distributions, pair plots and correlation coefficients. The posteriors look quite alright and parameter correlation between the parameters  $c_{rp}$  and  $n$  is virtually absent due to the reparameterization applied. The resulting fits from regression of the completely pooled Pinova data are shown in Fig. 6b. The fits resulting from regression of completely pooled data but displayed for the individual trees are shown in Fig. 6c. The fits look quite alright, both at the overall level as at the individual tree level, which is a first indication that there is not much of a tree effect. Table 3 shows a numerical summary of the posterior distributions.

### ***No pooling analysis of the Pinova data***

As a next step, we investigate what the regression results are when the data are not pooled, i.e., every tree dataset is analyzed on its own, resulting in 11 separate results. The resulting posterior parameter densities per tree are displayed in Fig. 7, indicating some variation in parameter estimates per tree, though they all overlap to some extent. (In fact, contrasts should be calculated to investigate how far the parameters really differ, but a different route will be taken here by applying multilevel modeling.) The fits to the individual datasets were quite acceptable (results not shown). However, these regression results are obtained without any relation between them, the outcome of one regression is completely independent of another, while they describe the same relationship of apples from the same cultivar. It neglects possible similarities between each regression. This brings up the topic of multilevel regression where regression is done on the same data but now the outcome of one regression “informs” the outcome of the other: they share information.

### ***Partial pooling of the pinova data per tree: multilevel modeling***

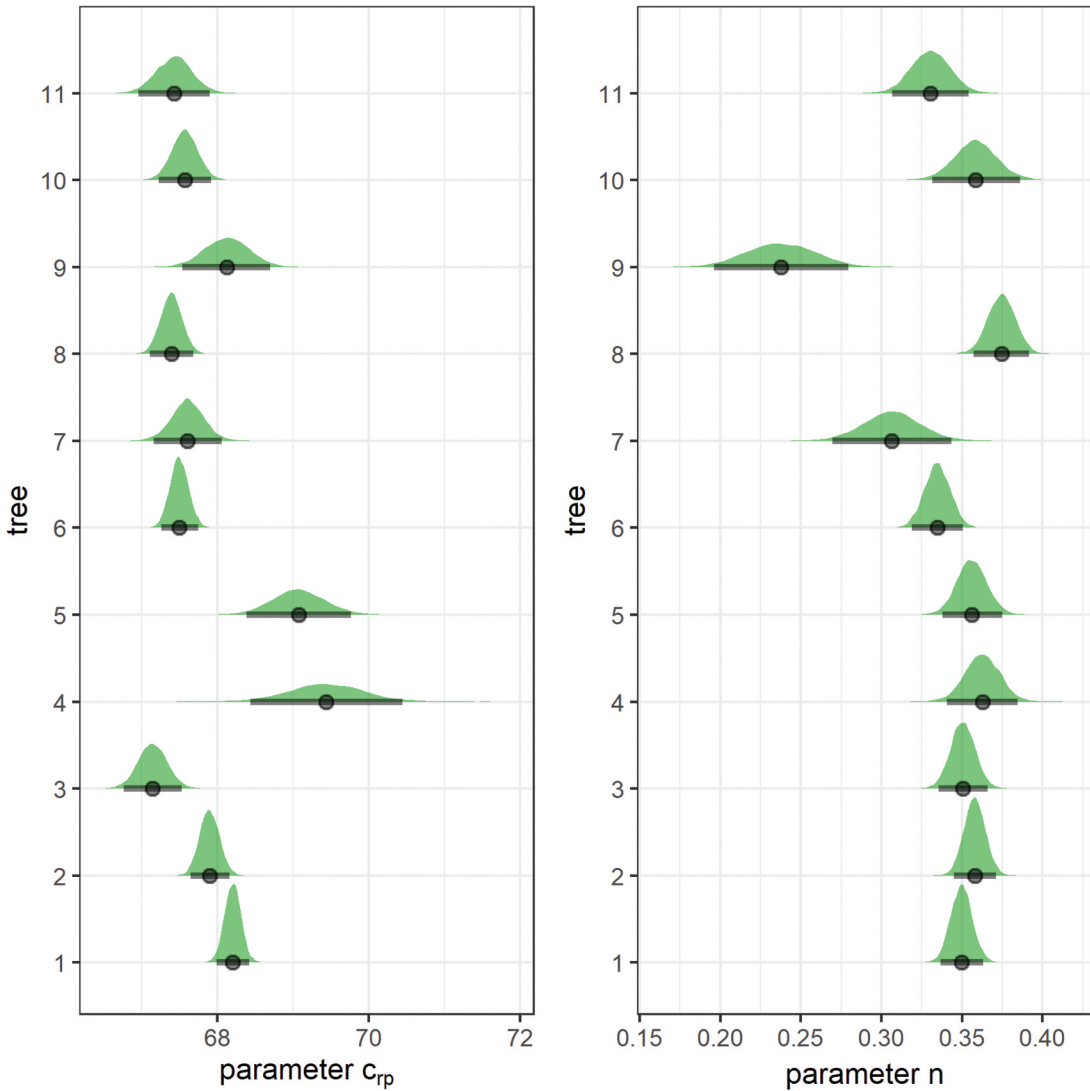
The third possibility is thus to apply multilevel modeling; this is also called partial-pooling. This deals with the situation where the model is able to make use of similarities while also allowing for differences. The obvious question is how this can be achieved. The answer is as follows. It is assumed that there is an overarching multivariate distribution for the two parameters in the model where the mean of this overarching distribution is considered as the population value, while its standard deviation characterizes the dispersion from the mean for the individual cases (note that this standard deviation does not express uncertainty, as is expressed in standard errors, but variation). The model used so far (Equation (5)) remains the same but the statistical notation changes as follows. The term multilevel modeling comes from the fact that variation is considered at different levels. In the current case study, there are two levels (in other cases, there might be more, see for instance.<sup>[12,13,21]</sup>). Level 1 is at the level of individual measurements within each experiment (apples measured from the same tree), level 2 is at the level of the cultivar where measurements on apples from the same tree are clustered together. Each result can be seen as a subsample of the population of all possible trials. The parameters derived from each measurement per tree can be seen as a representation of the population of all possible parameter values. It is thus attempted to characterize the random variation within measurements on apples from each tree as well as between all the measurements for all trees. That is why the term random effect (in contrast to a fixed effect) is sometimes used, while the term mixed effect refers to the fact that there is a mixture of random and fixed effects. The terminology may be somewhat confusing but, fortunately, the approach is not. Its strength is that the levels inform each other; room is given for individual variation per tree but it is also acknowledged that there are similarities for all trees (they share the same model, but the model parameters are allowed to vary). Multilevel modeling can be



**Figure 6.** Regression results for cultivar Pinova. a: Posterior parameter distribution, pairs plot and correlation coefficients of the completely pooled data set. b: fit resulting from the completely pooled pinova data analysis displayed for all data together. c: fit resulting from the completely pooled Pinova data analysis displayed per tree.

**Table 3.** Numerical summary of posterior parameter distributions resulting from Bayesian regression of the completely pooled Pinova data. SE = standard error, and lower and upper bounds indicate a 95% credible interval.

|                       | Mean  | SE    | Lower bound | Upper bound |
|-----------------------|-------|-------|-------------|-------------|
| $c_{rp}(-)$           | 67.57 | 0.049 | 67.47       | 67.67       |
| $n(-)$                | 0.33  | 0.002 | 0.32        | 0.33        |
| $\sigma_e(\text{mm})$ | 1.42  | 0.034 | 1.36        | 1.49        |



**Figure 7.** Ridgeplot showing posterior parameter densities resulting from individual regression of the pinova data per tree (no pooling). The thick black lines indicate the 95% credible interval and the black dots the mean of the posterior distribution.

done both in the Bayesian and frequentist framework but the Bayesian approach is more natural where parameters are considered to be random anyway, whereas they are considered fixed in the frequentist way, so there is then a bit of a tension to consider them random. As always in Bayesian modeling, the likelihood function for the data and prior distributions for

the parameters need to be specified, but it requires some extra work for multilevel modeling. Random effects are introduced to accommodate the between-tree variation. This is done by acknowledging that the parameters are related because they are estimating the same effect. That is to say, the two parameters  $c_{rp}$ ,  $n$  are allowed to vary from tree to tree, while characterizing that variation by introducing two new parameters  $u$  and  $v$ . Variation in  $c_{rp}$  is then  $c_{rp} \pm u$ , variation in  $n$  is  $n \pm v$ . The trick is that they can be connected by assuming that these two random effects are represented by a multivariate normal distribution (MVN), as displayed in Equation (12):

$$u, v \sim \text{MVN}(0, \Sigma) \quad (12)$$

The variance-covariance matrix  $\Sigma$  is a (in this case)  $2 \times 2$  random effects matrix holding the variances of  $u$ ,  $v$  as well as their covariance from which correlation coefficients can be derived. This is done as follows.  $\Sigma$  can be rewritten as in Equation (13):

$$\Sigma = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} R \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} \quad (13)$$

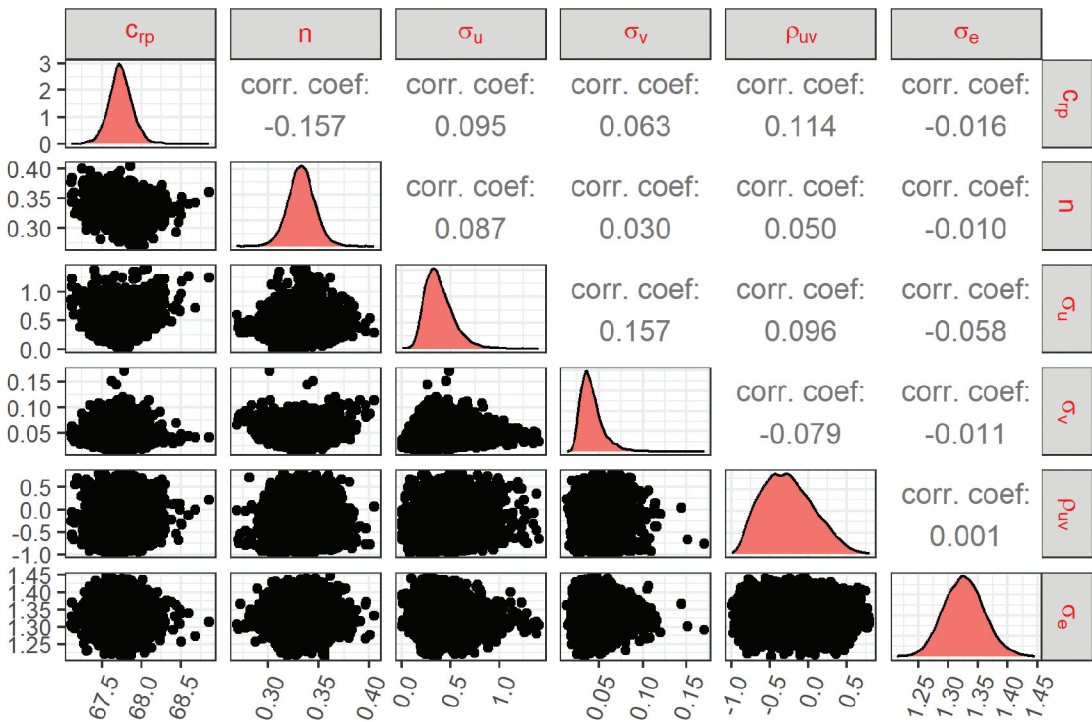
$R$  holds the symmetric correlation matrix with the parameter correlation coefficient  $\rho_{uv} = \rho_{vu}$ :

$$R = \begin{pmatrix} 1 & \rho_{uv} \\ \rho_{vu} & 1 \end{pmatrix} \quad (14)$$

Note that it is assumed that **on average** the random effects are zero, hence the 0 in Equations (12). In other words, these random effects are considered as deviations from the population value, or the “grand mean”, obtained for all data. Note that not  $u$ ,  $v$  will be estimated but rather  $\sigma_u$ ,  $\sigma_v$ , though  $u$ ,  $v$  can be calculated afterward. Since these are extra parameters, additional priors are needed for them. A normal distribution is again proposed for the likelihood function, as well as for the parameters  $c_{rp}$  and  $n$  with a lower bound at zero, an exponential distribution is used for both the random effects and the experimental standard deviation, and the so-called LKJ prior is used for the correlation coefficients between the random parameters (LKJ stands for Lewandowski-Kurowicka-Joe, named after the people who proposed this distribution); with a parameter value = 1, it is a weakly informative prior.<sup>[17]</sup> This translates to the statistical notation in Equation (15).

$$\begin{aligned} d_i &\sim N(\mu_i, \sigma_e) \\ \mu_i &= c_{rp} \cdot (w/w_{rp,i})^n \\ c_{rp} &\sim N(65, 5) \\ n &\sim N(0.3, 0.1) (\text{lb} = 0) \\ \sigma_u &\sim \text{exponential}(1) \\ \sigma_v &\sim \text{exponential}(1) \\ \sigma_e &\sim \text{exponential}(1) \\ \rho_{uv} &\sim \text{LKJ}(1) \end{aligned} \quad (15)$$

The results are as follows. [Figure 8](#) shows the posterior distributions, along with the pairs plots and correlation coefficients. [Table 4](#) gives a numerical summary of the parameter posterior distributions. It appears that the standard deviations for the category tree are rather low:  $\sigma_u$  quantifies the deviation of parameter  $c_{rp}$  and  $\sigma_v$  that for parameter  $n$ . [Figures 9a, b](#) show this in another way: except for tree 8 and 9, all parameter values are close to the grand mean. Another indication that the group effect is low is by comparing the experimental standard deviation  $\sigma_e$  from completely pooled data ( $\sigma_e = 1.42$ ) with that from partially pooled data ( $\sigma_e = 1.33$ ). The latter one is a little bit lower because part of the variance has shifted towards the group effect but not much. Finally, [Fig. 9c](#) shows the fits resulting from completely pooled data, from partially pooled data as the grand mean as well as at the group level. The plots almost coincide, indicating that the tree effect is only minor, except perhaps for tree 9.



**Figure 8.** Posterior parameter distributions, pair plots and correlation coefficients resulting from Bayesian multilevel regression of the Pinova data for the category tree.

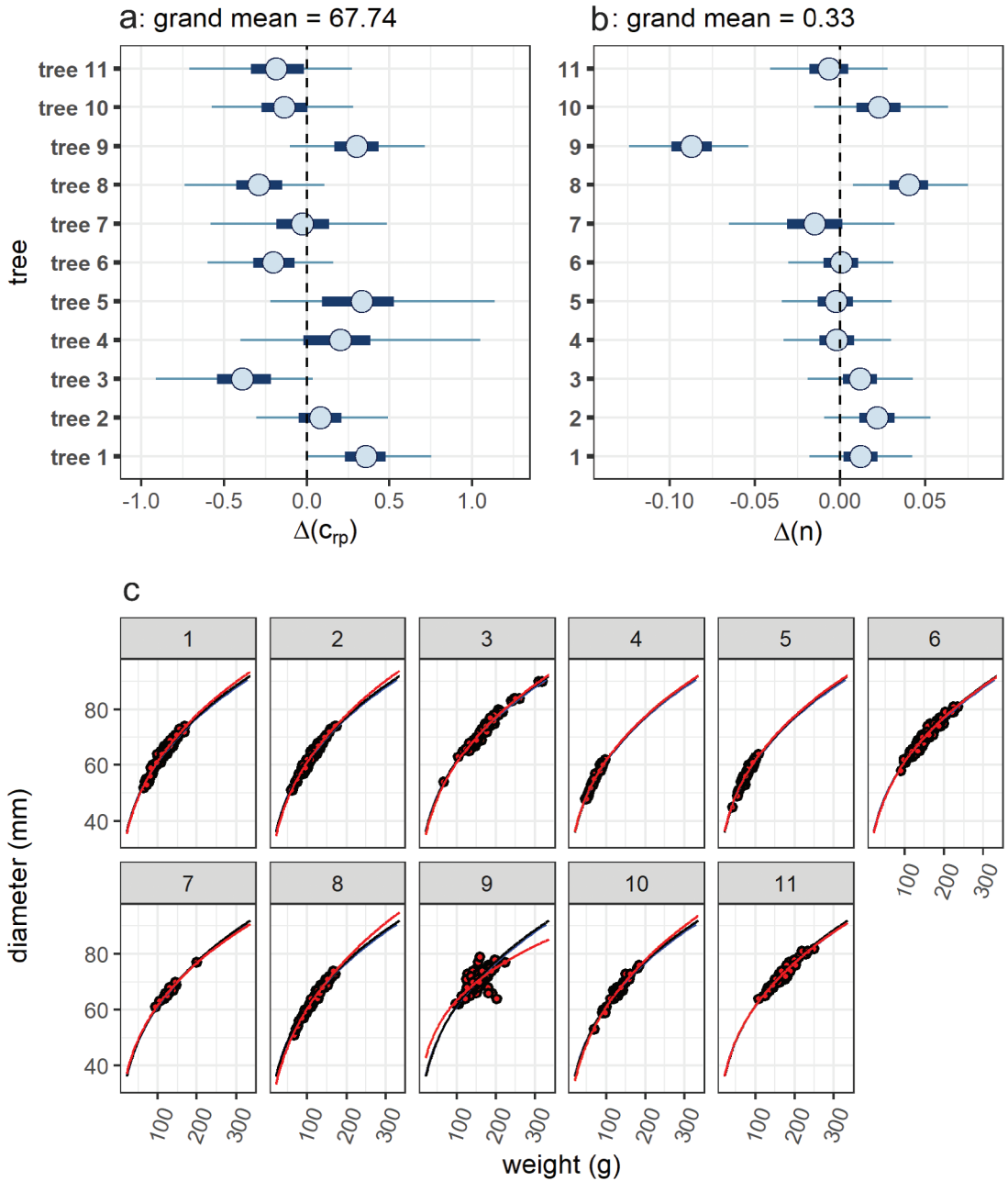
**Table 4.** Numerical summary of Bayesian regression of the partially pooled Pinova data with tree as category.  $\sigma_u$  indicates the standard deviation for parameter  $c_{tp}$ ,  $\sigma_v$  that of parameter  $n$ . SE = standard error, and the lower and upper bounds represent 95% credible intervals of the parameters.

| Parameter             | Mean  | SE    | Lower bound | Upper bound |
|-----------------------|-------|-------|-------------|-------------|
| $c_{tp}(-)$           | 67.74 | 0.152 | 67.45       | 68.05       |
| $n(-)$                | 0.33  | 0.014 | 0.3         | 0.36        |
| $\sigma_u(-)$         | 0.39  | 0.159 | 0.165       | 0.775       |
| $\sigma_v(-)$         | 0.04  | 0.013 | 0.024       | 0.073       |
| $\rho_{u,v}(-)$       | -0.27 | 0.33  | -0.81       | 0.424       |
| $\sigma_e(\text{mm})$ | 1.33  | 0.034 | 1.26        | 1.40        |

This same analysis was done for all those cultivars where data with information on the trees were available, as described in the Supplement. The conclusion was in all cases that there is only a minor tree effect. This means that apples from the same cultivar but harvested from different trees can be pooled without obscuring the relation between weight and diameter.

### Variance decomposition

To quantify the effect of multilevel modeling versus single-level modeling, a measure called variance decomposition can be calculated. It is the same idea as behind the ‘intra-class correlation coefficient’  $ICC$  that measures the proportion of variance explained by grouping structure in a population. <sup>[43,44]</sup>  $ICC = 0$  means: no information in grouping, while  $ICC = 1$  indicates a strong group effect: all observations within a group are identical. Another way of saying is that if  $ICC = 0$  the observations do not depend on a group membership (residuals are independent of the group). If  $ICC = 1$ , observations only vary between clusters (complete interdependence of residuals). If



**Figure 9.** Multilevel modeling results of the Pinova data per tree. a: Overview of the deviation  $\Delta$  from the grand mean of parameter  $C_{rp}$  per tree. b: Overview of the deviation  $\Delta$  from the grand mean of parameter  $n$  per tree. The thick and thin lines represent the 50% and 95% credible interval, respectively, and the blue circle represents the mean. c: Resulting fits to the data; blue line: fit resulting from completely pooled data, black line: fit resulting from grand mean result from partially pooled data, red line: fit at tree level resulting from partially pooled data.

$ICC \approx 0$ , multilevel modeling makes no sense and single-level regression analysis will suffice. Variance decompositions for Bayesian multilevel models using brms can be calculated from the posterior predictive distribution using the R package performance (see <https://easystats.github.io/performance/reference/icc.html>). Variance for each of the draws from the predictive distribution

not conditioned on group-level terms is compared to variances of each of the draws from the posterior predictive distribution conditioned on all varying effects. The ratio between these two variances displays the same information as the ICC. For the Pinova data, the variance conditioned on fixed effects is 47.14 (95% credible interval is from 45.41 to 48.84), while the variance conditioned on random effects is 48.77 (95% credible interval from 41.60 to 56.56) and therefore the variance ratio is virtually zero taking the uncertainty into account; in other words, this variance decomposition suggests that it makes no sense to apply multilevel modeling with tree as categorical factor. So, the overall conclusion of this section on possible tree effects is that there is not really a tree effect. Additional data analysis reported in the Supplement for other cultivars comes to the same conclusion. Therefore, the next analysis is to investigate whether or not there is a cultivar effect on the relation.

### **Cultivar effect**

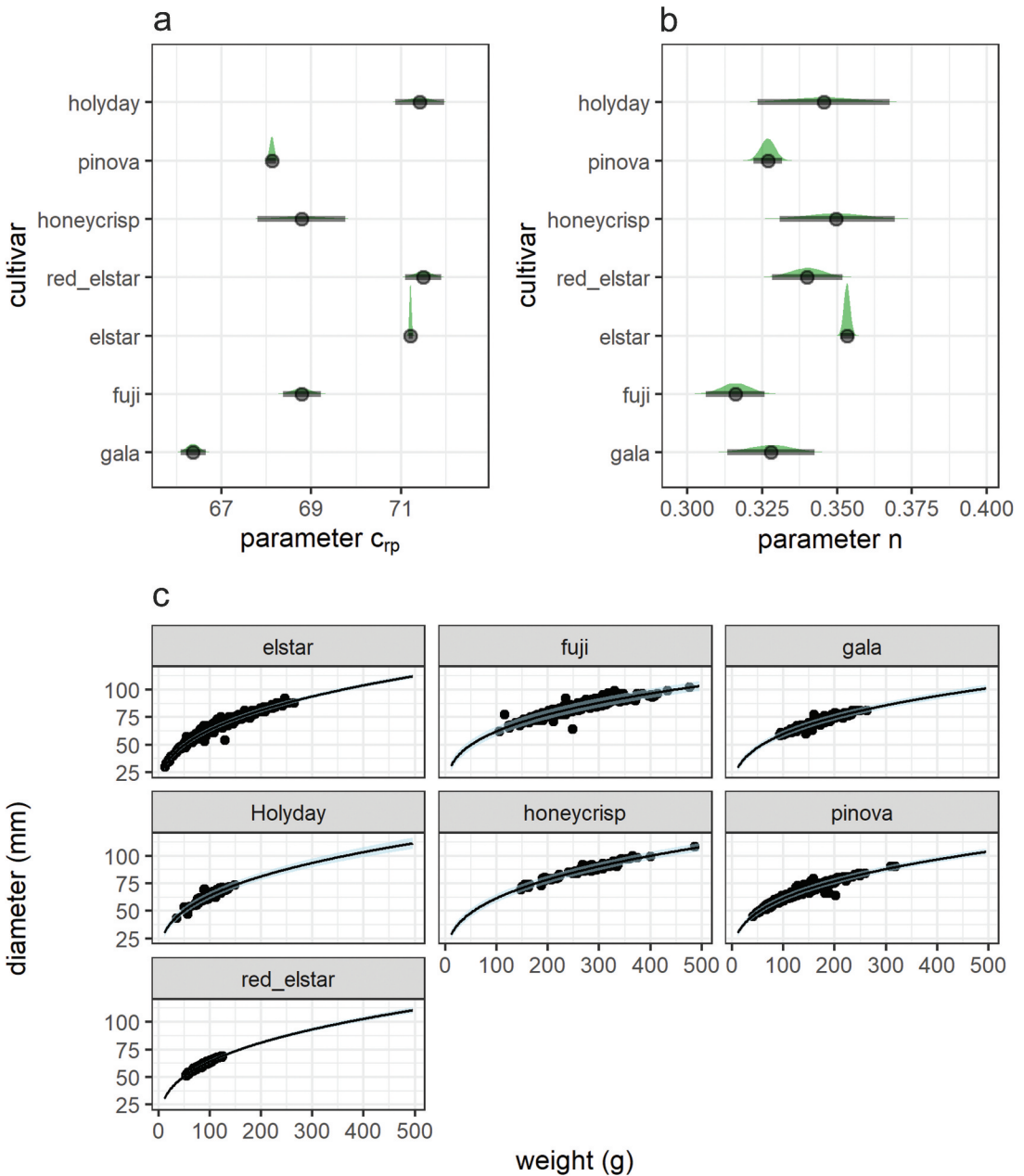
As concluded in the previous sections, apples from all trees of the same cultivar can be pooled because there is only a very minor tree, orchard and year effect. It was subsequently investigated whether or not there is a cultivar effect on the relation between weight and diameter of apples, in other words, cultivar becomes now the categorical variable. The same procedure was applied as in the section on the tree effect: first, regression with complete pooling, then regression with non-pooled data, followed by regression with partial pooling.

#### **Regression with complete pooling of the cultivar data**

The result of regression of all the pooled cultivar data is already shown in Figure S6 in the Supplement and Table 1 in this paper. The same result per cultivar is shown in Figure S28 in the Supplement, which shows at a glance whether or not there could be differences per cultivar. Although the overall trend is followed quite well, it is also clear that the fit is less for some cultivars, e.g., for the cultivars Elstar (diameter underestimated by the model), Gala (diameter overestimated by the model) and Honeycrisp (data underestimated by the model). It leads to a false impression of precision. The 95% prediction interval (for to-be-expected future data) is rather narrow, but it does not cover all observed data. These results make it worthwhile to investigate possible cultivar effects in more detail, continuing with a non-pooling approach, finally followed by a partial pooling multilevel modeling approach.

#### **Regression of the individual cultivar data (no pooling)**

After pooling all data for all cultivars, the second possibility is to analyze the data per cultivar separately. In doing so, individual regression results are obtained per cultivar, and then the results can be compared to each other per regression. Figure 10a, b compares the resulting posterior densities for parameters  $c_{rp}$  and  $n$ , respectively, showing the variation in estimates as well as in the uncertainties that go with them. The variation is seen not to be negligible between cultivars with very different magnitudes of uncertainty, which is partly due to the difference in the number of available data per cultivar: samples containing more apples lead, obviously, to less uncertainty in estimates. The resulting fits were all OK as shown in Fig. 10c. However, the analysis per cultivar emphasizes the differences between them and tends to lead to overfitting since each dataset is used on its own. Not all information is used: the next regression does not 'know' the outcome of the previous one, while there are similarities that could help to improve estimation. A perhaps even more serious effect is that such regressions assume that samples are independent. However, the apples used in one regression come from the same cultivar and some of them from the same tree, which makes them not independent. As a result, because of violation of the assumption of independence, parameter estimates may be biased. These two disadvantages can be mitigated by applying multilevel modelling, or partial pooling, as described in the next section (the machinery behind multilevel modeling is the same as already explained for multilevel modeling of the Pinova data with tree as category, so that is not repeated here).



**Figure 10.** Ridgeplot for parameters  $c_{rp}$  (a) and  $n$  (b) resulting from individual regression per cultivar. Fits per cultivar resulting from individual regressions per cultivar (c).

**Regression of the partially pooled cultivar data (multilevel modeling)**

Figure S29 in the Supplement shows the resulting posterior distributions, pairs plot and parameter correlations. The parameter correlations are low, obviously due to the reparameterized model. Clearly, some parameter distributions are rather skewed. Table 5 shows a numerical summary of the posterior distribution. It is instructive to check the differences between the completely pooled regression and the partially pooled one. The estimates for parameters  $c_{rp}$ ,  $n$  and  $\sigma_e$  slightly shifted from 67.74, 0.33 and 1.33 (completely pooled) to 69.39,



**Table 5.** Numerical summary resulting from Bayesian multilevel modeling of the partially pooled cultivar data with cultivar as the categorical factor. SE = standard error, and the lower and upper bounds reflect the 95% credible interval.  $\sigma_u$  represents the standard deviation of parameter  $c_{rp}$ ,  $\sigma_v$  that of parameter  $n$ ,  $\sigma_e$  the residual standard deviation.

| Parameter        | Mean  | SE    | Lower bound | Upper bound |
|------------------|-------|-------|-------------|-------------|
| $c_{rp}$ (-)     | 69.39 | 0.792 | 67.79       | 70.93       |
| $n$ (-)          | 0.34  | 0.007 | 0.32        | 0.35        |
| $\sigma_u$       | 2.02  | 0.549 | 1.23        | 3.38        |
| $\sigma_v$       | 0.02  | 0.007 | 0.01        | 0.04        |
| $\rho_{u,v}$ (-) | 0.47  | 0.310 | -0.26       | 0.91        |
| $\sigma_e$ (mm)  | 1.31  | 0.014 | 1.28        | 1.34        |

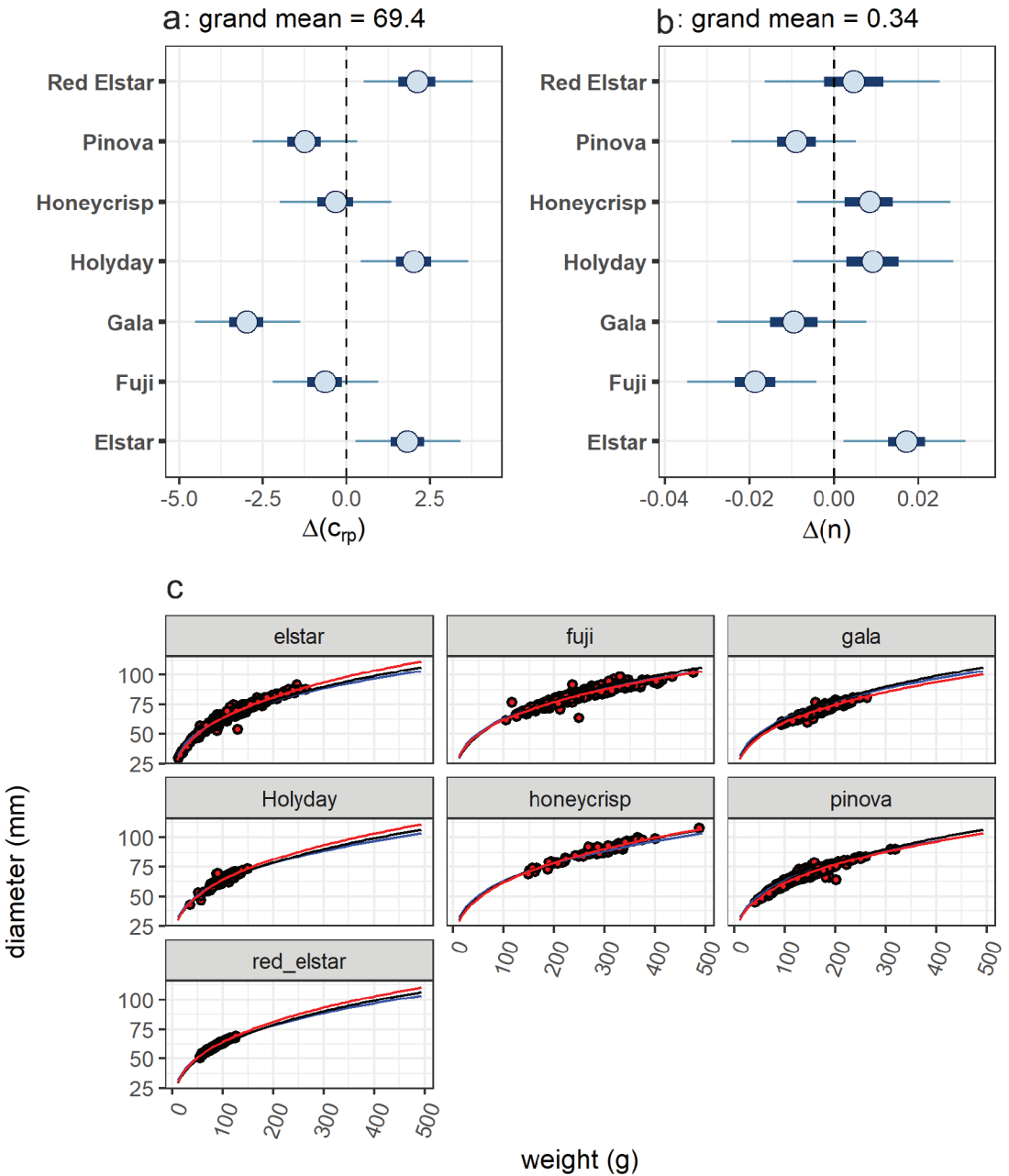
0.34 and 1.31 (partially pooled), respectively. It is important to realize that multilevel modeling is a compromise between over- and underfitting. The (slight) reduction in parameter  $\sigma_e$  is the result of variance partitioning; part of the variance is now explained in terms of variance of the parameters at the group level. In this way, the variation between parameters describing the characteristics of cultivars is quantified. Note that the model is exactly the same for each cultivar and that the variation is in the model parameters (i.e., the second level). Clarke<sup>[37]</sup> found values of  $0.333 < n < 0.362$  and  $11.96 < c < 13.84$  for Cox Orange Pippin and Golden Delicious, and  $0.357 < n < 0.37$  and  $11.48 < c < 12.3$  for Bramley, values that are in the same order of magnitude but definitely different from what is found here. Clarke<sup>[37]</sup> derived the values from a log-linear relation between  $\log(\text{diameter})$  and  $\log(\text{weight})$  with subsequent individual regressions without pooling. Clarke<sup>[37]</sup> also concluded that there is a cultivar effect on the relation between size and weight, as well as for different treatments and rootstocks. Fig. 11a, b shows a comparison of the parameter estimates at the group level from the population (“grand mean”) level. These estimates vary around zero, as implied by Equations (13). In most cases, but not all, the 95% credible intervals do cover the population mean, indicating that the deviation is noticeable but not too strong. Cultivars Fuji and Elstar deviate the most from the population (grand mean) values. Figure 11c compares the global fit resulting from complete pooling and from partial pooling. The difference in prediction intervals is almost not visible, the prediction band for the multilevel model is just a little bit wider. As derived above, the theoretical value of parameter  $n$  should be 0.33 if the assumption that apples are perfect spheres is correct. As shown in Table 5 the 95% credible interval is  $0.326 < n < 0.353$ , suggesting that a spherical approximation is not too bad.

### Variance decomposition

As was done in the section on the effect of Pinova trees, also here a variance decomposition analysis was applied to get an impression on how far variance has shifted due to partial pooling. The result is that the ratio is 0.12 with 95% credible interval ranging from 0 to 0.21; in other words, 12% of the variance is due to clustering of the data per cultivar. The variance conditioned on fixed effects is 72.73 (95% credible interval from 71.85 to 73.63) and conditioned on random effects is 82.54 (95% credible interval from 73.86 to 91.96). There are no strict rules to tell whether or not this is substantial, but it is, at least, noticeable.

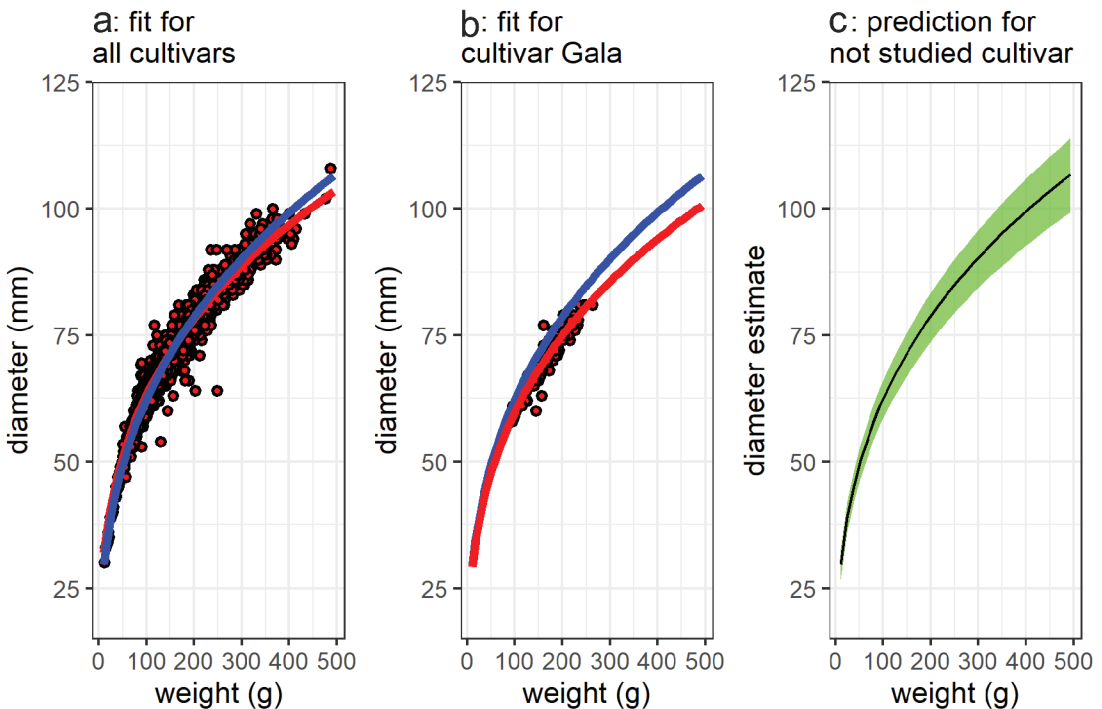
### Discussion

Now that various methods of analysis have been shown, the question is how to use the results. The results clearly show that variations due to tree origin, year and orchard site are noticeable but very small. The variation due to cultivar is not negligible, though also not very strong. What is gained by applying multilevel modeling on the cultivar level? Using results from multilevel models requires some consideration as there are several possibilities, depending on the research question one has.



**Figure 11.** Multilevel modeling of the data per cultivar. a: Overview of the deviation  $\Delta$  of parameter  $c_{rp}$  from the grand mean. b: Overview of the deviation  $\Delta$  of parameter  $n$  from the grand mean. The thick and thin lines in (a) and (b) represent the 50% and 95% credible parameter interval, respectively, and the blue circle represents the mean. c: Fit of the global model to all cultivars as resulting from complete pooling (blue line), from partial pooling at the population level (black line) and from partial pooling at the group level (red line).

- (1) If the interest is only in the population level without attention to group-level effects (cultivar in this case), then it suffices to use the population level (the global grand mean) resulting from multilevel modeling. The prediction will then result in an expected value while ignoring any cultivar-specific effect. However, since multilevel modeling is used, this expected value will be a compromise between under- and overfitting and it will have the best predictive capacity,



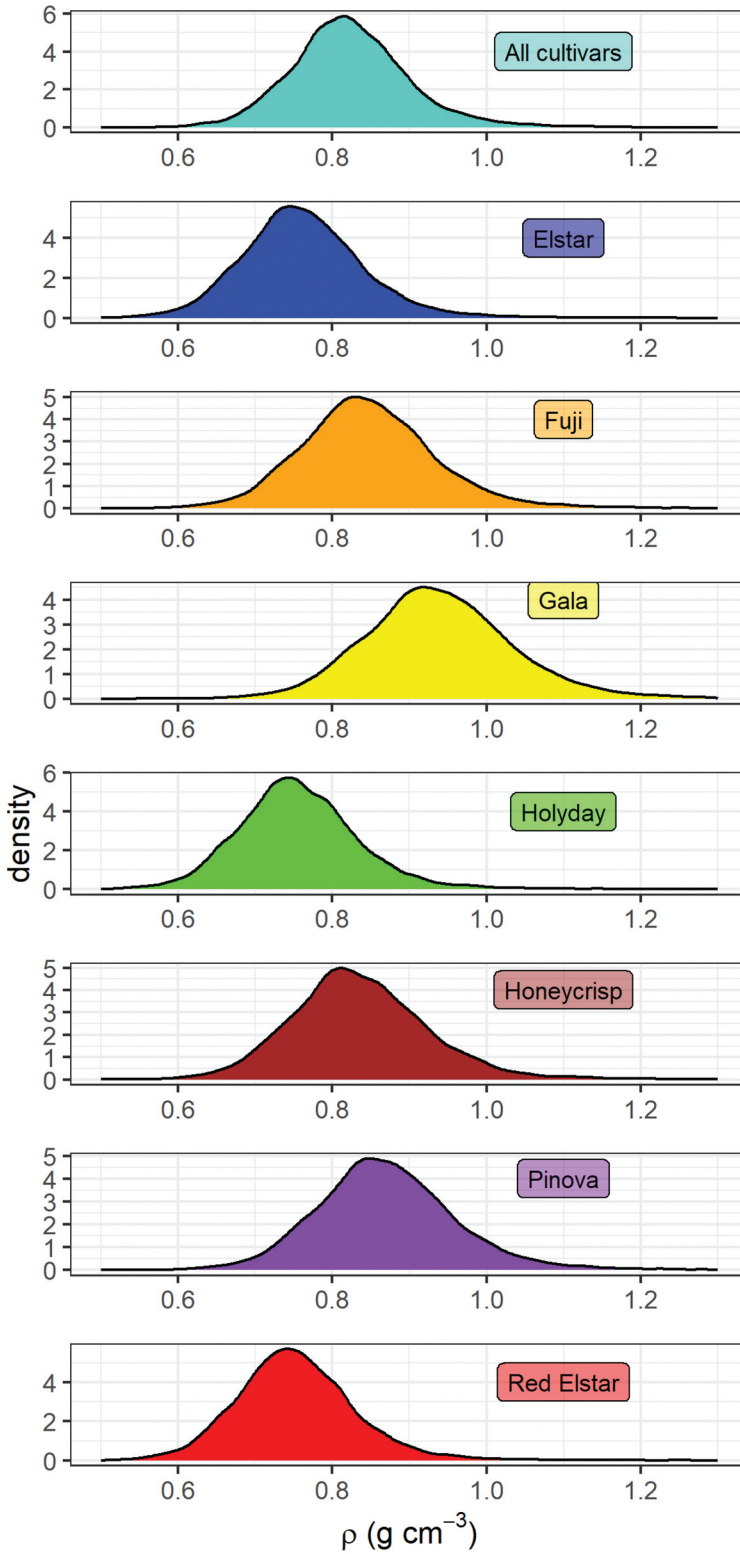
**Figure 12.** a: Fit of the regression lines resulting from complete pooling (red line) and from partial pooling to all cultivar data (blue line). b: Regression line for the specific Gala group level resulting from multilevel modeling at the population level (blue line, the blue lines in (a) and (b) are identical) and at the group level (red line). c: Prediction of diameter for a cultivar not studied yet with 95% prediction interval.

better than without multilevel modelling.<sup>[17]</sup> The effect can be seen in Figure 12a where the difference in regression line resulting from complete pooling and partial pooling is compared. It appears, in this case, that using results from complete pooling underfits diameter at higher weight values.

- (2) If the interest is in predicting the relationship for a group (a specific cultivar in this case), then one can use the information about the group-level effects based on the existing data (i.e., conditional on existing groups). So, if the interest is in predicting diameter from weight for a cultivar that was part of the underlying data, say, for instance, cultivar Gala, then a result is obtained as in Figure 12b (this can, of course, also be done for the other cultivars).
- (3) If the interest is in predicting what would happen for new groups (i.e., not based on the data that were used to build the model), then that is also possible, for instance, to predict diameter from weight for a cultivar that was not part of the data set used to build the model. To do that, it needs to be specified how to handle variation for such a new group. That can be done in two ways: i) based on the observed variation in the existing groups, ii) simulate variation based on the model's existing parameters. Both methods give more or less the same result. An example of such a prediction is shown in Figure 12c.

### Calculation of densities of apple cultivars

In the derivation of the relation between weight and diameter, the assumption was made that apples can be considered as globular and the estimates for parameter  $n$  suggest that this was not a bad



**Figure 13.** Calculated probability density for the density  $\rho$  of apples, using Equation (5) and parameter estimates resulting from partial pooling. The result for all cultivars pooled together is compared to that calculated for each cultivar separately.

approximation, the value 0.33 is contained within the 95% credible interval of parameter  $n$  (Table 5). There are a few literature references that measured sphericity of apples and those available values suggest that apples are indeed almost spherical.<sup>[45–49]</sup> If that assumption holds, it implies that parameter  $c_{rp}$  (and therefore also  $c$ ) is an estimator for density  $\rho$  (see Equation (5)). Since the multilevel modeling approach has led to both a population estimate for all cultivars considered, an overall density can be calculated from this, but also for each cultivar separately. Since all the information needed to do that calculation is available in the posterior, complete probability distributions can be calculated easily from the posterior. This is shown in Fig. 13. It shows differences between cultivars, the most striking one between the cultivars Gala and Elstar. However, it also clearly shows the uncertainty involved in these calculations, as indicated by the width of the distributions. Once again, this result should be treated with caution since it is not known whether or not there is a difference between cultivars in how far a spherical approximation holds; Fig. 11b does show some variation in parameter  $n$  for the various cultivars and that may be a reflection of a difference in sphericity. However, the main objective of this exercise is to show how posterior distributions can be used to do further calculations, thus showing how an impression of *derived* parameters can be obtained from the information present in posterior distributions; it takes into account correlations automatically, for instance, something that would require substantially more complex calculations with covariance matrices in the frequentist world.

### Concluding remarks

This paper has combined data from various sources for a meta-analysis of the relation between weight and diameter of apples. Contrast analysis and multilevel modeling were used in the analysis and these appeared to be powerful tools for characterizing the relation as well as the uncertainties involved. A theoretically derived curvilinear relation between diameter and weight was experimentally confirmed at all levels (cultivar, year, sites, trees). It appeared that there is hardly an effect of trees, of orchards and of year of harvest within a cultivar on this relation, but there is a more noticeable effect of cultivar on this relation. Multilevel modeling made it possible to quantify this variation at the cultivar level. It is also evident that parameter estimates do differ when partial pooling/multilevel modeling is applied. Since multilevel modeling is a compromise between over- and underfitting, estimates resulting from multilevel modeling are to be preferred when using parameters to predict new results. It has also been shown how Bayesian posteriors can be used for further analysis, if so desired.

The investigated relation in this paper was chosen to be relatively simple to show the principles of multilevel modeling and contrast analysis. However, for more complicated relations, the same principles apply. Food researchers inevitably have to deal with variation at various levels. However, in order to be able to apply these methods, experimental design and reporting of how data are collected becomes more important than ever. One needs to be able to differentiate characteristics into clusters or groups. The methodology discussed here can be used for all kinds of food-related questions: yields of crops, quality characteristics, sensory analysis, analysis of repeated experiments, shelf life studies, eating rates, digestion studies. Characterizing variability rather than hiding it in averages should become the norm in food science. It is hoped that this paper has contributed to awareness on this.

### Acknowledgments

The author would like to acknowledge Dr. Pflanz (Leibniz Institut für Agrartechnik, Potsdam-Bornim (ATB), Germany) for supplying the raw data on Elstar and Pinova cultivars, and Dr. Martini (Department of Plant Science, The Pennsylvania State University, PA, USA) for supplying the data on Fuji, Gala and Honeycrisp cultivars. Furthermore, Drs. A. Garre (Departamento de Ingenieria Agronomica, Universidad Politecnica de Cartagena, Spain) and E. Woltering (Wageningen University & Research, the Netherlands) are acknowledged for critical review of an earlier version of the manuscript.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## References

- [1] Hair, J. F.; Favero, L. P. Multilevel Modeling for Longitudinal Data: Concepts and Applications. *RAUSP Manage. J.* 2019, 54(4), 459489. DOI: [10.1108/RAUSP-04-2019-0059](https://doi.org/10.1108/RAUSP-04-2019-0059).
- [2] Austin, P. C.; Goel, V.; van Walraven, C. An Introduction to Multilevel Regression Models. *Rev. Canadienne de Sante Publique.* 2001, 92(2), 150154. DOI: [10.1007/BF03404950](https://doi.org/10.1007/BF03404950).
- [3] Guthrie, B.; Donnan, P. T.; Murphy, D. J.; Makubate, B.; Dreischulte, T. Bad Apples or Spoiled Barrels?. Multilevel Modelling Analysis Variation in High-Risk Prescribing in Scotland Between General Practitioners and Between the Practices They Work in. *BMJ Open.* 2015, 5(11), 8270. DOI: [10.1136/bmjopen-2015](https://doi.org/10.1136/bmjopen-2015).
- [4] Vis, D. J.; Bombardelli, L.; Lightfoot, H.; Iorio, F.; Garnett, M. J.; Wessels, L. F. Multilevel Models Improve Precision and Speed of IC50 Estimates. *Pharmacogenomics.* 2016, 17(7), 691–700. DOI: [10.2217/pgs.16.15](https://doi.org/10.2217/pgs.16.15).
- [5] Oddi, F. J.; Miguexz, F. E.; Ghermandi, L.; Bianchi, L. O.; Garibaldi, L. A. A Nonlinear Mixed-Effects Modeling Approach for Ecological Data: Using Temporal Dynamics of Vegetation Moisture as an Example. *Ecol. Evol.* 2019, 9(18), 10225–10240. DOI: [10.1002/ece3.5543](https://doi.org/10.1002/ece3.5543).
- [6] Johnson, A. A.; Ott, M. Q.; Dogucu, M. *Bayes Rules! An Introduction to Applied Bayesian Modeling*; Chapman and Hall/CRC: Boca Raton, FL, 2022.
- [7] Kruschke, J. K. *Doing Bayesian Data Analysis*, 2nd ed.; London, UK: Academic Press, 2015.
- [8] Gelman, A.; Carlin, J. B.; Stern, D. B.; Rubin, A.; Vehtari, D. B.; Rubin, D. B. *Bayesian Data Analysis, 3rd*; Chapman and Hall/CRC, 2013. [10.1201/b16018](https://doi.org/10.1201/b16018)
- [9] Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge, UK: Cambridge University Press, 2007. [10.1017/CBO9780511790942](https://doi.org/10.1017/CBO9780511790942)
- [10] Cao, R.; Francisco-Fernandez, M.; Quinto, E. J. A Random Effect Multiplicative Heteroscedastic Model for Bacterial Growth. *BMC Bioinf.* 2010, 11(1), 77–89. DOI: [10.1186/1471-2105-11-77](https://doi.org/10.1186/1471-2105-11-77).
- [11] Jaloustre, S.; Guillier, I.; Morelli, L.; Noel, V.; Delignette-Muller, M. L. Modeling of Clostridium Perfringens Vegetative Cell Inactivation in Beef-In-Sauce Products: A Meta Analysis Using Mixed Linear Models. *J. Food Microb.* 2012, 154(1–2), 44–51. DOI: [10.1016/j.jfoodmicro.2011.12.013](https://doi.org/10.1016/j.jfoodmicro.2011.12.013).
- [12] Garre, A.; Zwietering, M. H.; Den Besten, H. M. W. Multilevel Modelling as a Tool to Include Variability and Uncertainty in Quantitative Microbiology and Risk Assessment. Thermal Inactivation of Listeria Monocytogenes as Proof of Concept. *Food Res. Int.* 2020, 137, 109374. DOI: [10.1016/j.foodres.2020.109374](https://doi.org/10.1016/j.foodres.2020.109374).
- [13] Van Boekel, M. A. J. S. Kinetics of Heat-Induced Changes in Foods: A Workflow Proposal. *J. Food Eng.* 2021, 306 (April), 110634. DOI: [10.1016/j.jfoodeng.2021.110634](https://doi.org/10.1016/j.jfoodeng.2021.110634).
- [14] Van Boekel, M. A. J. S. Kinetics of Heat-Induced Changes in Dairy Products: Developments in Data Analysis and Modelling Techniques. *Int. Dairy. J.* 2022, 9, 105187. DOI: [10.1016/j.idairyj.2021.105187](https://doi.org/10.1016/j.idairyj.2021.105187).
- [15] Van Boekel, M. A. J. S. To Pool or Not to Pool: That is the Question in Microbial Kinetics. *Int J Food Microbiol* 2021, No. June, 109283. [10.1016/j.jfoodmicro.2021.109283](https://doi.org/10.1016/j.jfoodmicro.2021.109283).
- [16] Kruschke, J. K.; Liddell, T. M. The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective. *Psychon. Bull. Rev.* 2018, 25(1), 178–206. DOI: [10.3758/s13423-016-1221-4](https://doi.org/10.3758/s13423-016-1221-4).
- [17] McElreath, R. *Statistical Rethinking. A Bayesian Course with Examples in R and Stan, Second Edition*; Boca Raton, FL, USA: CRC Press, 2020. [10.1201/9780429029608](https://doi.org/10.1201/9780429029608)
- [18] Vaux, D. L.; Fidler, F.; Cumming, G. Replicates and Repeats—What is the Difference and is It Significant? *EMBO Rep.* 2012, 13(4), 291–296. DOI: [10.1038/embor.2012.36](https://doi.org/10.1038/embor.2012.36).
- [19] Lazic, S. E.; Clarke-Williams, C. J.; Munafò, M. R. What Exactly is ‘N’ in Cell Culture and Animal Experiments? *PLoS Biol.* 2018, 16(4). DOI: [10.1371/journal.pbio.2005282](https://doi.org/10.1371/journal.pbio.2005282).
- [20] Lazic, S. E.; Mellor, J. R.; Ashby, M. C.; Munafò, M. R. A Bayesian Predictive Approach for Dealing with Pseudoreplication. *Sci. Rep.* 2020, 10(1), 1–10. DOI: [10.1038/s41598-020-59384-7](https://doi.org/10.1038/s41598-020-59384-7).
- [21] Van Boekel, M. A. J. S. On the Pros and Cons of Bayesian Kinetic Modeling in Food Science. *Trends Food Sci. Technol.* 2020, 99, 181–193. DOI: [10.1016/j.tifs.2020.02.027](https://doi.org/10.1016/j.tifs.2020.02.027).
- [22] Lambert, B. *A Student's Guide to Bayesian Statistics*; SAGE publications Ltd: London, 2018.
- [23] Gelman, A.; Hill, J.; Vehtari, A. *Regression and Other Stories*, Cambridge, UK: Cambridge University Press, 2021. [10.1017/9781139161879](https://doi.org/10.1017/9781139161879)
- [24] Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan: A Probabilistic Programming Language. *J. Stat. Soft.* 2017, 76(1), 1–32. DOI: [10.18637/jss.v076.i01](https://doi.org/10.18637/jss.v076.i01).
- [25] Pinheiro, J. C.; Bates, D. M. *Mixed Effect Models in S and S-Plus*. Springer, 2000. [10.1007/978-1-4419-0318-1](https://doi.org/10.1007/978-1-4419-0318-1)
- [26] Bates, D.; Mächler, M.; Bolker, B. M.; Walker, S. C. Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Soft.* 2015, 67(1), 1. DOI: <https://doi.org/10.18637/jss.v067.i01>.

- [27] Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar, S.; Team, R. C. *Nlme: Linear and Nonlinear Mixed Effects Models, Version 3.1-151*. 2020. <https://cran.r-project.org/package=nlme>.
- [28] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020. <https://www.r-project.org/>.
- [29] RStudio Team. *RStudio: Integrated Development for R. Desktop V. 1.4.869 "Wax Begonia"*. Boston, MA, USA: PBC, 2020. <http://www.rstudio.com/>.
- [30] Bürkner, P. C. Brms: An R Package for Bayesian Multilevel Models Using Stan. *J. Stat. Soft.* 2017, 80(1). DOI: 10.18637/jss.v080.i01.
- [31] Bürkner, P. C. Advanced Bayesian Multilevel Modeling with the R Package Brms. *The R J.* 2018, 10(1), 395–411. DOI: 10.32614/RJ-2018-017.
- [32] Gelman, A.; Lee, D.; Guo, J. Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *J. Edu. Behav. Stat.* 2015, 40(5), 530–543. DOI: 10.3102/1076998615606113.
- [33] Pflanz, M.; Gebbers, R.; Zude, M. Influence of Tree-Adapted Flower Thinning on Apple Yield and Fruit Quality Considering Cultivars with Different Predisposition in Fructification. *Acta Hort.* 2016, 1130(1130), 605–612. DOI: 10.17660/ActaHortic.2016.1130.90.
- [34] Dobrzański, B.; Rybczyński, R.; Dobrzańska, A.; Wójcik, W. Some Physical and Nutritional Quality Parameters of Storage Apple. *Int. Agrophys.* 2001, 15, 13–18.
- [35] Marini, R. P.; Schupp, J. R.; Baugher, T. A.; Crassweller, R. Relationships Between Fruit Weight and Diameter at 60 Days After Bloom and at Harvest for Three Apple Cultivars. *HortScience.* 2019, 54(1), 86–91. DOI: 10.21273/HORTSCI13591-18.
- [36] Kruschke, J. K. Bayesian Analysis Reporting Guidelines. *Nat. Hum. Behav.* 2021, 5(10), 1282–1291. DOI: <https://doi.org/10.1038/s41562-021-01177-7>.
- [37] Clarke, G. M. The Relation Between Weight and Diameter in Apples. *J. Hortic. Sci.* 1990, 64(4), 385–393. DOI: 10.1080/00221589.1990.11516070.
- [38] Boggs, J. E. The Logarithm of Ten Apples. *J. Chem. Educ.* 1958, 35(1), 30–31. DOI: 10.1021/ed035p30.
- [39] De Levie, R. Collinearity in Least-Squares Analysis. *J. Chem. Educ.* 2012, 89(1), 68–78. DOI: 10.1021/ed100947d.
- [40] Schwaab, M.; Pinto, J. C. Optimum Reparameterization of Power Function Models. *Chem. Eng. Sci.* 2008, 63(18), 4631–4635. DOI: 10.1016/j.ces.2008.07.005.
- [41] Cinelli, C.; Forney, A.; Pearl, J. A Crash Course in Good and Bad Controls. *SSRN Elect. J.* 2020. March, 1–30. DOI: 10.2139/ssrn.3689437.
- [42] Vuorre, M. *How to Calculate Contrasts from a Fitted Brms Model*. <https://sometimesir.com/posts/2020-02-06-how-to-calculate-contrasts-from-a-fitted-brms-model> (accessed 2020-08-08).
- [43] Lüdecke, D.; Ben-Shachar, M.; Patil, I.; Waggoner, P.; Makowski, D. Performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *J. Open Source Softw.* 2021, 6(60), 3139. DOI: 10.21105/joss.03139.
- [44] Nakagawa, S.; Johnson, P. C. D.; Schielzeth, H. The Coefficient of Determination R<sup>2</sup> and Intra-Class Correlation Coefficient from Generalized Linear Mixed-Effects Models Revisited and Expanded. *J. R. Soc. Interface.* 2017, 14(134). DOI: 10.1098/rsif.2017.0213.
- [45] Altuntas, E.; Ozturk, B.; Özkan, Y.; Yildiz, K. Physico-Mechanical Properties and Colour Characteristics of Apple as Affected by Methyl Jasmonate Treatments. *Int. J. Food Eng.* 2012, 8(1). DOI: 10.1515/1556-3758.2388.
- [46] Gorji Chakespari, A.; Rajabipour, A.; Mobli, H. Post Harvest Physical and Nutritional Properties of Two Apple Varieties. *J. Agric. Sci.* 2010, 2(3), 61–68. DOI: 10.5539/jas.v2n3p61.
- [47] Kumar, S.; Neeraj, S.; S, V. A Study on Colour and Dimensional Assessment of Different Apple Cultivars Present in Domestic Fruit Market of NCR Region. *Int. J. Trop. Agric.* 2017, 35, 849–856.
- [48] Lak, M. B. Geometric Properties of Kohanz Apple Fruits. *Agric. Eng. Int. Cigr J.* 2011, 13(4), 1–8.
- [49] Meisami-Asl, E.; Rafiee, S.; Keyhani, A.; Tabatabaeefar, A. Some Physical Properties of Apple Cv. 'Golab. *Agric. Eng. Int. Cigr J.* 2009, XI, 1–7.