To pool or not to pool: That is the question in microbial kinetics ☆

M.A.J.S. van Boekel [*]

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

ABSTRACT

Variation observed in heat inactivation of *Salmonella* strains (data from Combase) was characterized using multilevel modeling with two case studies. One study concerned repetitions at one temperature, the other concerned isothermal experiments at various temperatures. Multilevel models characterize variation at various levels and handle dependencies in the data. The Weibull model was applied using Bayesian regression. The research question was how parameters varied with experimental conditions and how data can best be analyzed: no pooling (each experiment analyzed separately), complete pooling (all data analyzed together) or partial pooling (connecting the experiments while allowing for variation between experiments).

In the first case study, level 1 consisted of the measurements, level 2 of the group of repetitions. While variation in the initial number parameter was low (set by the researchers), the Weibull shape factor varied for each repetition from 0.58–1.44, and the rate parameter from 0.006–0.074 h. With partial pooling variation was much less, with complete pooling variation was strongly underestimated.

In the second case study, level 1 consisted of the measurements, level 2 of the group of repetitions per temperature experiment, level 3 of the cluster of various temperature experiments. The research question was how temperature affected the Weibull parameters. Variation in initial numbers was low (set by the researchers), the rate parameter was obviously affected by temperature, the estimate of the shape parameter depended on how the data were analyzed. With partial pooling, and one-step global modeling with a Bigelow-type model for the rate parameter, shape parameter variation was minimal. Model comparison based on prediction capacity of the various models was explored.

The probability distribution of calculated decimal reduction times was much narrower using multilevel global modeling compared to the usual single level two-step approach. Multilevel modeling of microbial heat inactivation appears to be a suitable and powerful method to characterize and quantify variation at various levels. It handles possible dependencies in the data, and yields unbiased parameter estimates. The answer on the question "to pool or not to pool" depends on the goal of modeling, but if the goal is prediction, then partial pooling using multilevel modeling is the answer, provided that the experimental data allow that.

1. Introduction

Modeling heat inactivation of micro-organisms is important from both a practical and a scientific point of view. It is essential for optimization of heat processing, where balance is needed between desired inactivation and undesired damage (Van Boekel et al., 2020). Many publications on this topic have led to much knowledge, summarized in meta-analyses such as the one from Den Besten et al. (2018). Characteristic for microbial kinetics is the enormous variation observed between experiments, an inevitable phenomenon due to the biological nature of micro-organisms. Many factors affect how they behave,

including response to heat stress. Variation is not so easy to quantify. Repeating experiments provides information; the question is how to deal with such data. Variation can be seen as a nuisance, but it can also serve as a source of information. Averaging is sometimes applied but using all data without averaging is also possible. Then the question becomes whether all results should be pooled together or that every experiment should be analyzed on its own, a question that is addressed here.

Many types of inactivation models are proposed in literature. The first model ever applied is some 100 years old now, describing log-linear inactivation as in Eq. (1):

$$log_{10}N = log_{10}N_0 - \frac{t}{D} \qquad (1)$$

Parameter D represents decimal reduction time, $log_{10}N_0$ and $log_{10}N$ the logarithm of the number of micro-organisms at time zero and time t, respectively. Sometimes, the survival ratio $S = N/N_0$ is modeled as in Eq. (2):

$$log_{10}S = -\frac{t}{D} \qquad (2)$$

but it may be better to model $log_{10}N_0$ as a parameter since it is not a fixed number but subject to experimental variation. The parameters in Eq. (1) are usually derived via linear regression, though it was noted by Dolan and Mishra (2013) that parameter D should actually be considered as a nonlinear parameter as it is the *inverse* of the slope of the linear regression line, with consequences for calculating confidence intervals and predictions. In the course of time, various other models have been proposed because log-linear behaviour is rather the exception than the rule, e.g., Peleg and Cole (1998); Peleg (2006); Van Boekel (2002). Though the log-linear model is still popular, nonlinear models are more often applied the last two decades. Free web applications of various models are available (Garre et al., 2018; Garre et al., 2017; González et al., 2019; Peleg et al., 2017). The Weibull model in Eq. (3), is quite flexible because it is able to describe log-nonlinear as well as log-linear behaviour:

$$log_{10}N = log_{10}N_0 - \frac{1}{2.303}\cdot\left(\frac{t}{\beta}\right)^{\alpha} \qquad (3)$$

α represents the shape parameter (dimensionless), β is a rate-like parameter (dimension time). Parameter β is similar to but not the same as a D-value. It represents a characteristic time at which the survival function $log_{10}S(t) = \exp(-1) = 0.434$ regardless of α. A simpler form of the Weibull model is the equivalent Weibullian model (Peleg, 2006):

$$log_{10}N = log_{10}N_0 - k_r\cdot t^{n_t} \qquad (4)$$

Comparison of Eqs. (3) and (4) shows that $n_t = \alpha$ and $k_r = 1/(2.303 \cdot \beta^{\alpha})$. With $\alpha = n_t = 1$, Eq. (3) becomes Eq. (1) with $k_r = 1/D = 1/(2.303 \cdot \beta)$. However, the simple Weibull model is not a panacea for every survival curve; other models may do a better job, depending on the shape of the survival curve, see, for instance, Gil et al. (2017) for a recent overview of possible models. The purpose of this article is not to test various models, however. The question addressed is, for cases where the Weibull model applies, how parameters α and β vary as a function of experimental repetitions and temperature experiments. To characterize such variation, the possibilities of multilevel modeling are explored, for which the Bayesian approach is most suited (Gelman and Hill, 2007; McElreath, 2020) though a frequentist approach is also possible (Pinheiro et al., 2020).

Multilevel modeling is not new in food microbiology. Bayesian multilevel modeling using a Weibullian type model was recently described for microbial risk assessment (Garre et al., 2020). They characterized variability in thermal inactivation within strains and between strains of 20 strains of *Listeria monocytogenes*. Their dataset is quite unique because it allowed distinction between three levels: experimental variability, within-strain variability and between-strain variability, making it a comprehensive three-level model. Multilevel models have been used also in other cases, e.g., *Salmonella* survival in dry nuts (Santillana Farakos et al., 2016), *E. coli* survival in beef gravy (Juneja and Marks, 2005), *Salmonella* survival in chicken (Juneja et al., 2016). Multilevel models are also well-known in meta-analyses and risk management.

A Bayesian approach to heat inactivation using the same *Salmonella* data as in case study 2 in the current article was applied by Koyama et al. (2019), be it with the log-linear model (Eq. (2)) despite the nonlinear trend in the data. Moreover, $log_{10}N_0$ was not modeled nor did they apply

multilevel modeling. Next, a very brief introduction to the concept of multilevel modeling is given by explaining the differences between no pooling, complete pooling and partial pooling of data. Then, a very brief recapitulation about the Bayesian approach and its connection to multilevel modeling follows, after which the results are discussed.

1.1. Multilevel data resulting from experimental design

Experimental designs are obviously important for subsequent analysis and modeling. In that context, levels are to be understood as follows. Repetitions of similar experiments can be grouped; measurements within each repetition form the lowest level 1, while repetitions themselves are grouped into level 2, a subsample of the population of all possible repetitions. Similarly, experiments at various temperatures can be grouped per temperature: measurements at each temperature form level 1, while the various temperature experiments form groups at level 2. Another example is the multilevel approach by Garre et al. (2020), the lowest level being variability in repeated experiments with the same batch and strain (experimental variability), the second level variability measured for the same strain from different batches (biological variability within strains), and the third level variability between 20 different *Salmonella* strains. Typical for multilevel modeling is that experiments/observations are characterized by a grouping factor while model parameters are then allowed to vary per group around a central value. Response variables at level 1 are always measurements/observations. Response variables at higher levels are regression coefficients from the level below that.

The way data can be analyzed depends on how experiments are designed. Fig. 1 gives a schematic overview of a common design for heat inactivation experiments. Suppose that, as an example, six trials (i.e., j = 6 in Fig. 1) are done isothermally under the same conditions, to determine heat survival curves, i.e., number of surviving bacteria are measured as a function of heating time. Inexplicable variation between the trials will happen, in other words, experimental results will be similar but not completely identical. Several possibilities then arise for analysis. The *first* approach could be to pool the results from all six repetitions together and analyze them as if they are all generated in the same way without any variation between the trials. This is called **complete pooling**; it ignores grouping structures and may lead to underfitting (Gelman and Hill, 2007), i.e., not using all the information in the data. The *second* approach could be to average over the six measurements at each heating time, described as **pooling and averaging**. The data are then compressed, leading to even more underfitting. The *third* approach could be to analyze each repetition separately. This is the **no pooling** approach, which would lead (in the example) to six different modeling results at each temperature. The outcome of one repetition is then in no way connected to the outcome of another repetition. In the no pooling case, group means are estimated independently as if the variation between groups is infinitely large; this may lead to overfitting because it tends to make the groups more different than they actually are (Gelman and Hill, 2007). The *fourth* approach connects the groups (6 trials in the example) with each other, called **partial pooling**, which is achieved with multilevel modeling. Group means are then considered a random sample from an overarching common distribution. Partial pooling is in between no-pooling and complete pooling, and a compromise between under- and overfitting. Complete pooling and no pooling approaches are thus subsets of partial pooling. If variation at the group level of trials (Fig. 1) is characterized by a standard deviation σ_g, no pooling implies that $\sigma_g \to \infty$, while complete pooling implies $\sigma_g \to 0$. With partial pooling, $0 < \sigma_g < \infty$, and σ_g can be estimated by multilevel modeling. Common single level classical regression is actually a special case of multilevel regression with either $\sigma_g \to 0$ *or* $\sigma_g \to \infty$, depending on whether data are pooled or not. Another important aspect with heat inactivation experiments is that measurements within a trial are mostly not completely independent statistically as they may be correlated. Regression of such data without accounting for correlation
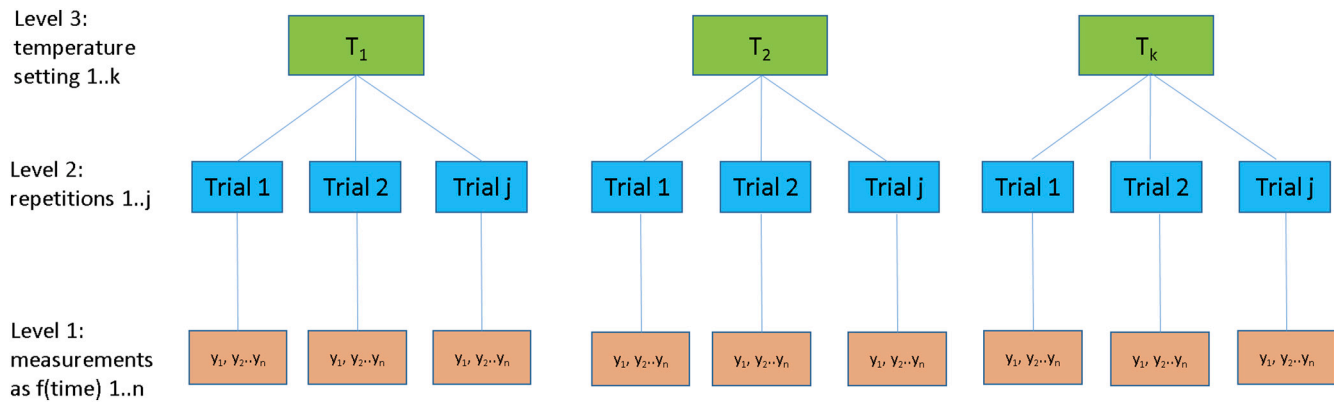
Fig. 1. Schematic overview of a common experimental design to determine microbial survival in which several isothermal experiments ($T_1…T_k$) are done with repetitions (trial 1…j) and 1…n measurements in each repetition. The three levels that can be distinguished form an example of a nested, hierarchical design.

underestimates variation, resulting in biased parameter estimates and their uncertainties. The advantage of multilevel modeling is that it takes correlation into account so that an unbiased estimate of variability is obtained, thereby counteracting the danger of "pseudoreplication" (Lazic et al., 2018; Lazic et al., 2020). Parameters describing the population level are sometimes named 'fixed' while parameters describing variation on the cluster/group level are called 'random.' The term 'mixed effect models,' sometimes used instead of multilevel modeling, refers to a mixture of random and fixed effects. The terminology is somewhat confusing but the approach is not. The strength of the multilevel method is that the various levels inform each other; individual variation between experiments is allowed but similarities between experiments are also taken into account. The term 'borrowing strength' is also used sometimes: information from one experiment is used in the analysis of another. Yet another term is hierarchical modeling, pointing at hierarchy in the levels, shown also in Fig. 1. This short introduction attempted to clarify that models taking various levels into account are to be preferred, provided that group levels can be distinguished in the data. To loosely quote McElreath (2020): "if researchers use single level classical regression, they should make clear why they did NOT use multilevel modeling", in other words, multilevel modeling should be the standard. More background material is provided in Garre et al. (2020), Gelman and Hill (2007), Gelman et al. (2013), Kruschke (2015), Lambert (2018), McElreath (2020).

1.2. Bayesian regression

A detailed description of Bayesian regression for kinetics is given by Van Boekel (2020). The Bayesian method considers every unknown as a random variable. Measurements and observations, once available, are considered fixed and are called data, while non-observable variables such as rate constants, $D-$ and $Z-$values are considered random variables and called parameters. A random variable is described by a probability distribution. In Bayesian analysis, unknown parameters are given a probability distribution before data analysis, called the prior, reflecting the expert knowledge that is available before data analysis. The data themselves are described by a data-generating model and a probability distribution called the likelihood function, and by combining the prior and the likelihood according to Bayes' theorem an updated probability distribution is obtained called the posterior. This posterior density gives the probability for certain parameter values, reflecting the state of knowledge about the parameters, conditional on the data and the proposed model. In the frequentist framework only point estimates are obtained after regression because parameters cannot have distributions: they are considered fixed. Confidence intervals and *p*-values are not probability statements about parameters but reflect the theoretical range of values obtained if the experiments are repeated

infinitely. While application of Bayes' theorem faced computational difficulties in the past, it is nowadays well possible, also for very complex models, because of the Markov Chain Monte Carlo (MCMC) method with which the posterior parameter distribution is approximated, based on the input of the likelihood function and the proposed prior distributions. The details are not explained here but can be found in many textbooks, e.g. McElreath (2020). Software packages are available to do that routinely, and the one used here is the Stan language (Gelman et al., 2015), referred to in the Material and methods section. Multilevel modeling goes naturally in the Bayesian approach where parameters are considered variable, while parameters become response variables at the higher levels. Assigning probability distributions to parameters is not possible in the frequentist approach. Nevertheless, multilevel modeling is also done in the frequentist way, be it that maximum likelihood methods are then used in parameter estimation, as was, for instance, done by Juneja et al. (2016).

1.3. Outline of the paper

The goal of this paper is to explore the potential of multilevel regression using two case studies of microbial heat inactivation. The present work differs from the multilevel approach by Garre et al. (2020) in the micro-organism studied (*Salmonella* rather than *Listeria monocytogenes*); the present *Salmonella* dataset only allows within-strain variability analysis, however. Another difference is the use of the basic Weibull model, rather than a reparameterized Weibull model. The dimension of rate parameter β does not depend on shape parameter α in the basic Weibull model which makes interpretation and comparison easier. In models such as the Weibullian model (Eq. (4)), the dimension of k_r is t^{-n_t}, thus depending on shape parameter n_t.

Two case studies about microbial inactivation kinetics are reported here. The first one comprises trials (repetitions) for one strain at one isothermal condition, allowing to investigate quantitatively by multilevel modeling how parameter estimates vary between similar experiments. This has not been reported before in literature, to the best of the author's knowledge. The second case investigates if and how isothermal experiments at various temperatures can be analyzed in a multilevel way. This is also not described before in literature. The difference with Koyama et al. (2019) is the use of the Weibull model instead of the log-linear model and the use of multilevel modeling. At the end of the article, the question raised in the title: "to pool or not to pool" will be discussed.

2. Material and methods

Both case studies investigate thermal inactivation of strains of *Salmonella* serovar Typhimurium DT104, described in Mattick et al. (2001),

for experiments at several temperatures, a_w and pH values, all in broth. They are archived in the database Combase. The first case study used an isothermal dataset at $T = 65$°C, $a_w = 0.8$, pH $= 6.5$ containing 18 repetitions The second case study used data about the same *Salmonella* strain studied at 9 different temperatures at pH $= 7$, $a_w = 0.75$. The detailed IDs for the data extracted from Combase are provided in the Supplement.

R version 4.0.3 (R Core Team, 2020) was used from within RStudio version 1.4.1103 (RStudio Team, 2020). A list of all R packages used is in the Supplement.

Bayesian regression was done with the R package brms (Bürkner, 2017, 2018), version 2.14.4, using tidyverse R extensions for data handling as described by Kurz (2020). brms is an interface between R and the language Stan, which is state-of-the-art software for Markov Chain Monte Carlo (MCMC) calculations (Gelman et al., 2015). MCMC performance needs to be checked for convergence with diagnostics described in, for instance, McElreath (2020). An example of such checks is shown in the Supplement; in all cases reported here, these checks were found to be OK. All R codes and data used can be found at the author's Github repository https::/github.com/TinyvanBoekel/IJFM.

3. Results and discussion

3.1. Multilevel analysis of 18 isothermal repetitions

Fig. 2 shows the dataset of 18 similar experiments with the same strain at 65 °C, $a_w = 0.8$, pH $= 6.5$. With reference to Fig. 1, one temperature is considered (k $= 1$), with number of trials j $= 18$, and number of measurements in each trial $n = 10$. The experimental conditions were similar for each experiment but between-strain and within-strain variability could not be distinguished because information on which days experiments were done, and whether or not the same stock culture was used is not available. The variability depicted in Fig. 2 is therefore assumed to be within-strain. Despite similar experimental conditions, considerable variation between experiments exists, commonly observed with microbial counts. Many factors may cause such variation, some can be controlled (like temperature, pH and a_w, kept constant in this case) but not all factors affecting inactivation can be controlled completely. For instance, bacteria were perhaps grown in the same media but at different points in time, causing small variations in behaviour. The question is also whether the data can be considered statistically independent, a necessary assumption when doing regression. Probably, the data in one heating experiment were not independent, if samples were taken from the same batch during one experiment. Such data lend themselves well to multilevel modeling, provided that there are enough

repetitions available (a minimum of 5–6 experiments is usually indicated, though there is no clear-cut limit; in this case 18 repetitions were available, more than enough for multilevel analysis). The case study is analyzed step by step, from pooled and pooled-averaged, to completely pooled, and finally to partially pooled analysis with multilevel modeling.

3.1.1. Regression settings

Bayesian regression needs a likelihood function representing the data generating process, for which the Weibull model was chosen, and an assumption of how the data are distributed. When microbiological counts are expressed logarithmically, $\log_{10}N$ values are assumed to be distributed normally (counts themselves are distributed log-normally). This implies that each measured $\log_{10}N$ is supposed to be centered around a mean μ with a dispersion characterized by a constant standard deviation σ_e. The Weibull model in Eq. (3) contains three parameters: $\log_{10}N_0$ (dimensionless), parameter β (dimension time) and α (dimensionless). Since N_0 was measured, it is subject to variation and therefore estimated (as $\log_{10}N_0$). Including σ_e, four prior distributions are thus needed (note that priors describe expert knowledge before data are analyzed, they do not describe actual variation in the real world). Normal distributions were assumed for parameters α, β as this is the most natural expression of uncertainty (McElreath, 2020). For σ_e a half-cauchy distribution is often used, where 'half' means: bounded at zero, thus forcing the standard deviation to be positive, with rather 'fat' tails characteristic for a Cauchy distribution, allowing for unlikely but not impossible large values. See Van Boekel (2020) for more details on priors in relation to kinetics. An exponential distribution could be an alternative prior for σ_e, which is also strictly positive but has less fat tails. This leads to the statistical model in Eq. (5):

$$
\begin{aligned}
\log_{10}N &\sim \mathcal{N}(\mu, \sigma_e) \\
\mu &= \log_{10}N_0 - 1/2.303 \cdot (t/\beta)^\alpha \\
\log_{10}N_0 &\sim \mathcal{N}(7, 1) \\
\alpha &\sim \mathcal{N}(1, 1) \ (\text{lb} = 0) \\
\beta &\sim \mathcal{N}(0.03, 0.03) \ (\text{lb} = 0) \\
\sigma_e &\sim \text{half} - \text{Cauchy}(0, 10)
\end{aligned}
\tag{5}
$$

Expert knowledge tells that $\log_{10}N_0$ varies around 7 (set by the researchers). For parameter β, a very crude rule of thumb is that $(1/\beta) \cdot t_{end} \approx 1 - 10$ leading to a value of 0.03 but it is given a large standard deviation to indicate uncertainty about the actual value. The lower bound at zero (lb $= 0$) makes this a truncated normal distribution because this parameter cannot be negative. Parameter α has usually values between 0.1 and 2 and so its average is expected to be around 1 but again with a large standard deviation and a lower bound of zero. These probability distributions are available in the software; they are coded via the package brms and then transferred to the software Stan. Prior predictive checks (McElreath, 2020) comprise simulations with the model and priors without involving the data yet, to see whether the settings make sense. The result with the settings in Eq. (5) is shown in the Supplement (Fig. S1); a wide range of model outcomes appears to be possible without impossible outcomes such as increasing numbers of micro-organisms with time.

3.1.2. Regression of pooled and averaged data of the 18 repetitions

Results obtained at each heating time from each trial were completely pooled and will be compared with multilevel modeling results later on. The posterior distributions of the parameters are shown in Fig. 3, along with a pairs plot and parameter correlation coefficients. The posterior distributions look well behaved, but correlation between parameters α and β is substantial, a known problem with the Weibull model. This parameter correlation did not hamper estimation as such because the model converged without problems. Strong parameter correlation needs to be taken into account when parameters are used in further calculations such as decimal reduction times based on the Weibull parameters α and β, an example of which is given below. In the
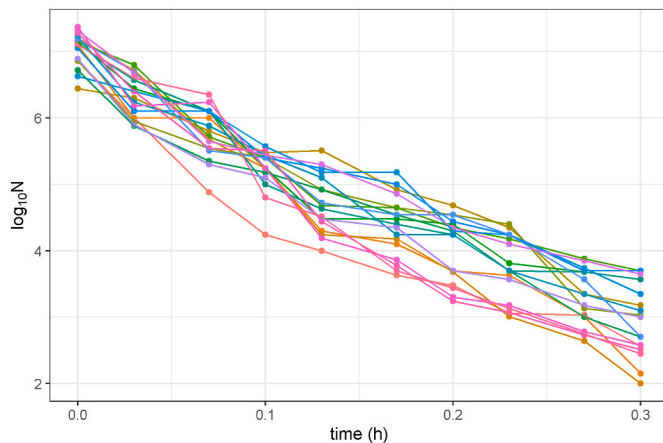

Fig. 2. Overview of the 18 experimental repetitions (trials) of heat inactivation of *Salmonella* at 65 °C, $a_w = 0.8$, pH $= 6.5$.
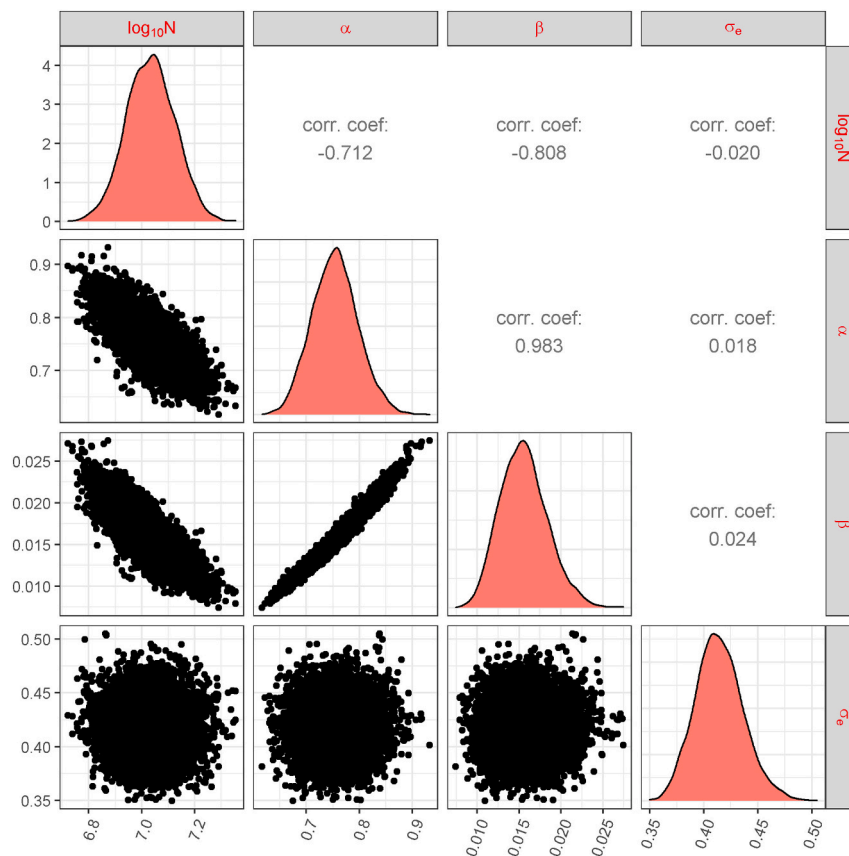Source: Mattick et al. (2001), data extracted from Combase.

Fig. 3. Pairs plot, parameter posterior distributions and Pearson correlation coefficients resulting from the nonlinear regression of the completely pooled *Salmonella* data.

frequentist framework, covariances are then needed with complicated formulas but if that is done, correct uncertainty estimates can be obtained even in the presence of strong collinearity (De Levie, 2012). In the Bayesian framework, however, further calculations with parameters take this correlation automatically into account leading to correct uncertainty estimates of the calculated result, a big advantage of having a posterior distribution available.

The numerical summary of the posterior in terms of the mean, standard error and credible intervals is given in Table 1. Bayesian 95% credible and prediction intervals indicate with 95% probability mean and future values, respectively, conditional on the model and the data. In the frequentist interpretation, 95% confidence and prediction intervals mean: if the experiment is repeated 100 times, values are in the interval 95 times (and 5 times not). The result $0.67 < \alpha < 0.84$ indicates that log-linear behaviour is not happening with this dataset as $\alpha = 1$ is not in the credible interval. Least-squares estimates reported by Mattick et al. (2001) using Eq. (4) were $\alpha = 0.73$ and (recalculated from k_r) $\beta = 0.019$ h, in the same range as found here. The fit based on the completely-pooled data is shown in Fig. 4A; such a fit is called 'retrodiction' by McElreath (2020) because it shows how the model matches

with the data in retrospect, it is not a prediction. Note that the regression line is not the result of minimizing least squares, but of the combination of the most likely values for the parameters in the posterior distribution, conditional on the model and the data. The 95% credible interval for the mean in Fig. 4A is seen to be quite narrow, suggesting that it can be estimated precisely (**as a preliminary note, this credible interval is strongly underestimating variation in the mean as shown when discussing the multilevel modeling results**). Prediction intervals refer to the ability of the model to predict future, not yet observed values; this calculation adds, besides the uncertainty of the parameters, also the uncertainty due to sampling, for which σ_e is used. Obviously, the prediction interval is much larger than the credible interval for the mean because of the large scatter in the data.

Sometimes, data are averaged; doing that per heating time over all the trials leads to pooled-and-averaged data. Bayesian regression of these data led to a similarly looking pairs and correlation and posterior parameter distributions plot as in Fig. 3 (shown in Supplement Fig. S3), while the fit is shown in Fig. 4B. The numerical results of the averaged data are also presented in the Supplement (Table S1) and are virtually the same compared to Table 1, except for $\sigma_e = 0.09$ in the pooled-averaged case whereas it was $\sigma_e = 0.41$ for the non-averaged pooled data. The result of $\sigma_e = 0.09$ is strongly underestimating the true variability. Comparison between Fig. 4A and B shows the difference to be consequently in the 95% prediction interval, which is much wider in the case of the completely pooled data. The error bars (representing experimental standard deviations in the data, hence variation in the real world) are much wider than what the model predicts with 95% probability for future values, so this pooled-and-averaged data based model is definitely not useful to predict variability in future data. This is clearly a case of underfitting because the information in the data is not fully used. Averaging is, in general, not good practice because it removes

Table 1
Numerical summary of the posterior resulting from Bayesian regression of the Weibull model to the completely pooled data of *Salmonella* case study 1with 18 repetitions. SE = standard error, lower and upper bounds represent 95% credible intervals.

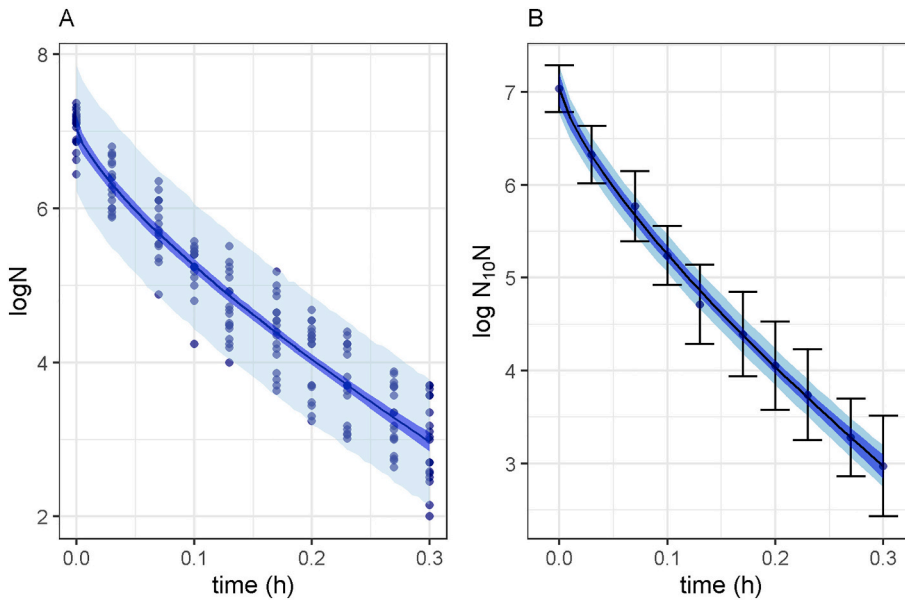|	Mean	SE	Lower bound	Upper bound
$\log_{10}N_0$	7.03	0.09	6.86	7.21
α	0.75	0.04	0.67	0.84
β (h)	0.016	0.003	0.011	0.022
σ_e	0.41	0.02	0.37	0.46

Fig. 4. Fit resulting from Bayesian nonlinear regression of the Weibull model to the completely pooled *Salmonella* data (A) and to the completely pooled-and-averaged data (B). The dark blue ribbon represents the 95% credible interval for the mean and the light blue ribbon the 95% prediction interval for future data. The error bars in panel B represent the standard deviations of the data at each heating time ($n = 18$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

information. A possible remedy in this case could, perhaps, be weighted regression. However, a different route is taken in the next sections.

3.1.3. Regression of the 18 individual datasets: the case of no-pooling

In the previous section, all data were pooled (and after that averaged) and analyzed in one go. Alternatively, each experiment could be analyzed separately, resulting in 18 regression results. The fits to the individual datasets do not look bad at all per experiment (shown in Supplement Fig. S4) but they are very different from each other. The between-experiment variation appears to be much larger than the within-experiment variation. The posterior distributions of the three parameters resulting from the individual regressions are summarized in Fig. 5 in so-called forest plots. The large variation in the estimates as well as in the posterior distributions is clear; these are the most likely values for the most optimal fit to the data per experiment. The disadvantage of these individual regressions is that nothing is shared between them. When the next regression is done, it is "ignorant" about the

outcome of the previous one. Moreover, if the data are correlated within one run, this may inflate the regression procedure. The question is how to interpret these regression results overall. Should they be averaged to know what is happening on the population level? The subsequent question would then be how to weigh the results, some parameters are much better estimated in one run (e.g., trial 1) than in another (e.g., trial 4). This is the point where multilevel modeling comes into play by acknowledging, characterizing and exploiting the observed variation, which can be accomplished by sharing information between regressions. This sharing of information is based on partially pooled data, as discussed in the next section.

3.1.4. Bayesian regression of the partially pooled data from the 18 repetitions: multilevel modeling

With multilevel modeling, the 18 repetitions are allowed to be different, while similarities between experiments are also acknowledged so that information can be shared. As a bonus, statistical dependencies
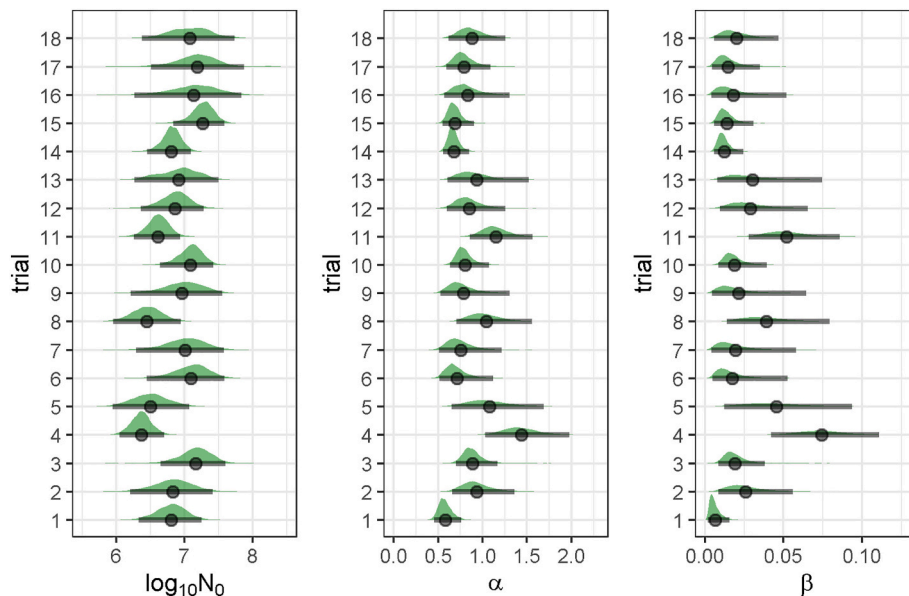

Fig. 5. Forest plots showing posterior parameter densities for the Weibull parameters $\log_{10}N_0$, α, β, resulting from the individual regression per trial dataset. The dot represents the mean of the distribution.

are taken into account that are more often present than realized, possibly leading to the earlier mentioned phenomenon of 'pseudo-eplication', meaning that data are assumed to be independent when they are not. Statistical dependencies in the data may lead to serious bias in the estimates of parameters. A way to achieve that regressions inform each other is by postulating that the parameters are connected to each other via a multivariate normal distribution. The parameters are assumed to vary around a common value with variation characterized by a variance/standard deviation and a covariance matrix that makes the parameters dependent on each other. Mathematically, parameters in each cluster/group level are allowed to deviate from the population mean by a certain positive or negative amount u_i, v_i, w_i as expressed in Eq. (6) as an expansion of the Weibull Eq. (3):

$$log_{10}N = \left((log_{10}N_0 + u_i) - \frac{1}{2.303} \cdot \left(\frac{t}{\beta + v_i} \right)^{\alpha + w_i} \right) \quad (6)$$

On average, u_i, v_i, w_i are assumed to be zero, i.e., their mean corresponds to the population mean of the corresponding parameter, with variation around that mean characterized by standard deviations $\sigma_u, \sigma_v, \sigma_w$, respectively. This is the same reasoning as for residuals ϵ_i: their mean is also assumed to be zero with variation characterized by standard deviation σ_e. The distribution of the parameters is assumed to be characterized by a multivariate normal distribution (MVN):

$$\begin{bmatrix} u_i \\ v_i \\ w_i \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma \right) \quad (7)$$

Σ is a covariance matrix and can be rewritten as in Eq. (8):

$$\Sigma = \begin{pmatrix} \sigma_u & 0 & 0 \\ 0 & \sigma_v & 0 \\ 0 & 0 & \sigma_w \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_u & 0 & 0 \\ 0 & \sigma_v & 0 \\ 0 & 0 & \sigma_w \end{pmatrix} \quad (8)$$

$\mathbf{R}$ holds the symmetric correlation matrix with the correlation coefficients ρ:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_{uv} & \rho_{uw} \\ \rho_{vu} & 1 & \rho_{vw} \\ \rho_{wu} & \rho_{wv} & 1 \end{pmatrix} \quad (9)$$

Upon modeling these equations, the population parameters $log_{10}N_0$, α, β, the standard deviations $\sigma_u, \sigma_v, \sigma_w, \sigma_e$ and the correlation coefficients $\rho_{uv}, \rho_{uw}, \rho_{vw}$ form the output. Note that u_i, v_i and w_i are not directly given as model output, but they can be calculated as deviations from the grand mean, as shown below. Covariances are directly related to the correlation coefficient and can be calculated from it; for instance, for correlation coefficient ρ_{uv} and covariance σ_{uv}:

$$\rho_{uv} = \frac{\sigma_{uv}}{\sqrt{\sigma_u^2 \sigma_v^2}} \quad (10)$$

$$\sigma_{uv} = \rho_{uv} \cdot \sigma_u \cdot \sigma_v$$

As always, Bayesian regression needs priors for every estimated parameter. A commonly used one for the correlation matrix is the LKJ prior (McElreath, 2020); with a value = 1, it gives equal probability for the correlation coefficient to be between −1 and +1, which seems reasonable because there is no way of knowing correlation beforehand (remember that priors are not describing real word variation but prior knowledge). The resulting statistical model with likelihood function and priors is:

$$log_{10}N \sim \mathcal{N}(\mu_i, \sigma_e)$$
$$\mu_i = log_{10}N_0 - \frac{1}{2.303} \cdot \left(\frac{t}{\beta} \right)^{\alpha}$$
$$log_{10}N_0 \sim \mathcal{N}(7, \sigma_u)$$
$$\alpha \sim \mathcal{N}(1, \sigma_w) \text{ lb} = 0$$
$$\beta \sim \mathcal{N}(0.03, \sigma_v) \text{ lb} = 0 \quad (11)$$
$$\sigma_u \sim \text{half} - \text{cauchy}(0, 10)$$
$$\sigma_v \sim \text{half} - \text{cauchy}(0, 10)$$
$$\sigma_w \sim \text{half} - \text{cauchy}(0, 10)$$
$$\sigma_e \sim \text{half} - \text{cauchy}(0, 10)$$
$$\mathbf{R} \sim \text{LKJ}(1)$$

Note that the priors for the parameters α, β and $log_{10}N_0$ contain now so-called hyperpriors for their standard deviation instead of a number, while these hyperpriors have priors on their own. The numerical summaries of the posteriors are in Table 2 (the pairs and correlation plots are in Supplement Fig. S5). In comparison to Table 1, the extra parameters characterize variation between trials on group-level and population-level effects (completely pooled data only provide information on population effects because groups are considered equal). The population level effects are the values for parameters $logN_0, \alpha$ and β. Their means shifted a little bit compared to the completely pooled data, while their standard errors decreased a bit. This is a typical consequence of multi-level modeling: some of the variance has shifted from population level to group-level effects. Note, however, that the experimental standard deviation σ_e has decreased considerably in the multilevel approach ($\sigma_e = 0.26$) as compared to the model based on the pooled data ($\sigma_e = 0.41$) because of this shifting of variance. The multilevel approach thus gives a more detailed picture of the factors that cause variation. The group- or cluster-level effects $\sigma_u, \sigma_v, \sigma_w, \rho_{uw}, \rho_{uv}, \rho_{vw}, \sigma_e$ characterize the variation in parameters between trials and how they are correlated. This is the within-strain variability described by Garre et al. (2020) for their *Listeria monocytogenes* case. Note that the point estimates of the correlation coefficients are not very high but also that their estimates are quite uncertain, reflecting the information that is present in the data, information that apparently does not allow a more precise estimation. To clarify possibly confusing terminology in literature where random and fixed effects are mentioned, in the Bayesian framework parameters are never considered fixed but random. Better terms would be 'varying effects' instead of 'random effects' and 'population effects' instead of 'fixed effects' (Gelman et al., 2013). In the Bayesian framework, coefficients can be common across groups (the population level) but deviations in the coefficients are also allowed so that they vary across groups (the varying effects). Furthermore, note the difference between standard deviations and standard errors (SE) in numerical outputs. Standard deviations characterize *variation* in the real world. With

Table 2
Numerical summary of multilevel modeling of the 18 repetitions of *Salmonella* inactivation in case study 1. SE = standard error, lower and upper bounds indicate 95% credible intervals. Standard deviations and correlation coefficients are reported for parameters *u, v, w* defined in Eq. (6).

	Mean	SE	Lower bound	Upper bound
$log_{10}N_0$	7.01	0.07	6.87	7.15
α	0.77	0.03	0.71	0.84
β (h)	0.017	0.002	0.01	0.02
σ_u	0.13	0.08	0.01	0.29
σ_w	0.06	0.03	0.003	0.13
σ_v	0.004	0.002	0.001	0.009
ρ_{uw}	0.09	0.47	−0.81	0.91
ρ_{uv}	−0.23	0.42	−0.88	0.70
ρ_{vw}	0.50	0.43	−0.64	0.94
σ_e	0.26	0.02	0.23	0.29

completely pooled data, σ_e represents the variance/standard deviation not explained by the predictor variables; with partially pooled data real world variation is also attributed to grouping factors, thereby decreasing unexplained variance. SE refers to *uncertainty* in the estimates present in the posterior after combining priors with data. Although the same metric is used, standard deviations and standard errors refer to different entities, namely variation as a property of the real world and uncertainty about their values as a property of the researcher. The variation in parameters u, v, w at the group level (see Eq. (7)) as deviations from their respective population level is shown in Fig. 6. This variation is much smaller than the one shown for the no-pooled data in Fig. 5. This is because in the multilevel procedure, the regressions have 'informed' each other giving a much more consistent picture. This sharing of information leads to a phenomenon called 'shrinkage', which guards against under- and overfitting, also discussed by Garre et al. (2020). The term shrinkage indicates that the results from the individual regressions are drawn ("shrunk") towards the grand mean. Shrinkage can be illustrated in more ways, for instance, by comparing fits obtained from partial pooling at the population level, partial pooling at the trial level and no-pooling (Supplement Figs. S6 and S7 show this in detail). Multilevel modeling takes away extremes, narrows down the variation and prevents thereby under- or overfitting. This is obviously useful when making predictions as in risk analysis. In addition, a realistic quantitative impression of the variability in parameter estimates is obtained that did account for dependencies in the data. Of course, results are conditional on the data and the model used. The analysis as given in Fig. 6 shows, by the way, that there are some individual experiments with a rather wide and skewed posterior distribution (e.g., trials 1 and 4, see also Supplement Fig. S7). The individual fit with trials 1 and 4 was quite good but appears to deviate considerably from the "grand mean."

3.1.4.1. Retrodiction, model comparison and prediction. Fig. 7 displays the overall regression plot resulting from multilevel modeling (partial pooling), with 95% credible and prediction bands, and is compared to results from completely pooled data. This is retrodiction, checking in how far the model matches with the data on which it is based in retrospect. Fig. 7A shows the 95% credible interval for the multilevel model to be wider than for the one based on pooled data, illustrating the remark made earlier with Fig. 4 that the uncertainty in predicting the

regression mean is underestimated with completely pooled data. In contrast, Fig. 7B shows the opposite where the multilevel model gives a narrower prediction band than the model based on the completely pooled data. This is because the multilevel model "borrows strength" from all the experiments, thereby increasing confidence in model predictions. This is, once again, illustrating that multilevel modeling counteracts over- and underfitting that may occur with single-level modeling (McElreath, 2020). Another way of retrodiction is to do so-called posterior predictive simulations. It indicates how well predictions made from the posterior distribution match the densities of the actual measured data. A graphical impression of that is given in the Supplement (Fig. S8), showing a slightly better match for the multilevel model, especially in the range $3 < \log_{10}N > 6$. However, fitting is one thing, prediction another, as shown next. Several models have been tested above and comparing them should be part of a kinetic workflow (Van Boekel, 2021). The goal is not to select the best performing one and to throw away the rest, but rather to compare them in their performance. A well-known selection criterion in the frequentist world is the Akaike Information Criterion (AIC) but that is not suitable for Bayesian models (McElreath, 2020). For Bayesian models the Widely Applicable Information Criterion (WAIC) can be used and/or equivalently the 'leave-one-out-cross-validation' (loo-cv) method (Vehtari et al., 2017). It leaves one observation out, refits the model and checks how well the model is able to predict the observation that was left out; this is then repeated for other observations (there are procedures in the software to speed up this process). The software package brms used here has that option incorporated from the R package loo. The loo-cv method is really about predictive performance of a model. Note that loo-cv values have no meaning in absolute sense, they become valuable in comparing. The model with the highest loo-cv value performs the best and is put to zero, less performing values then get a negative value. An additional feature of loo-cv (and also of WAIC) is that also an uncertainty measure about the value can be provided from the posterior, something that is not possible with frequentist model comparison methods such as AIC. The loo-cv results for this case study are in Table 3, showing that the multilevel model based on partially pooled data performs considerably better. This is not model discrimination but rather model comparison: it does not mean that the single level model based on completely pooled data is not performing well, it only indicates that the multilevel model is performing better in predicting future results. It demonstrates the value
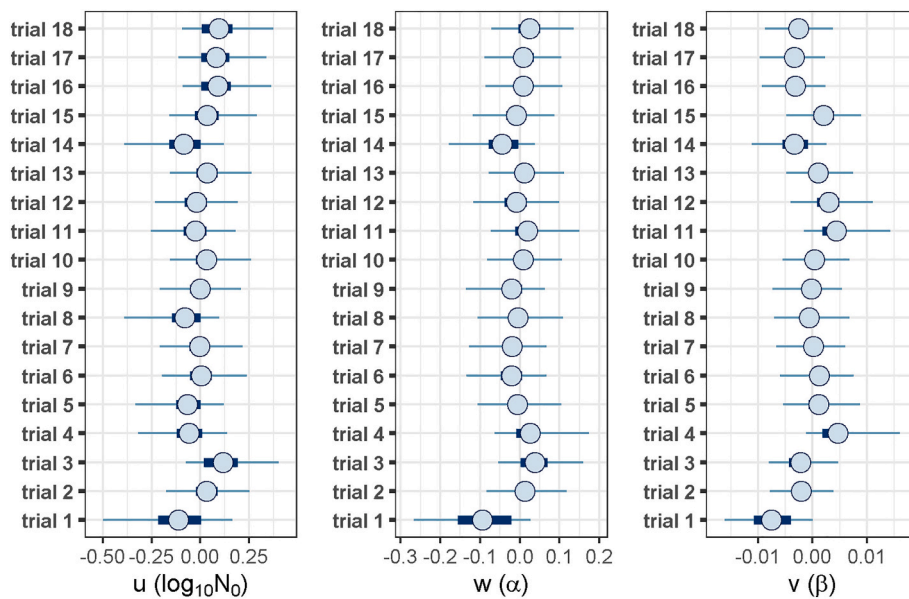

Fig. 6. Variation per trial of parameters u (deviation from the population value of parameter $\log_{10}N_0$), v (deviation from the population value of parameter β) and w (deviation from the population value of parameter α); the population mean is given the value = 0. The circle indicates the mean, the thick line the 50% probability range, the thin line the 95% probability range.
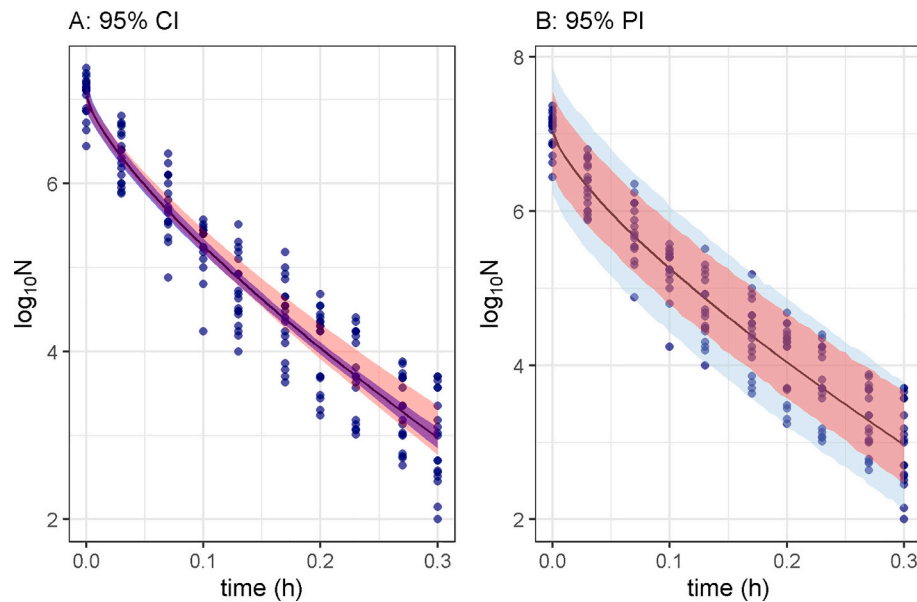
Fig. 7. Regression line + 95% credible interval (CI) for the multilevel model with partially pooled data (red ribbon) as compared to the model based on completely pooled data (blue ribbon) (A). 95% prediction interval (PI) for the multilevel model (red ribbon) as compared to the model based upon completely pooled data (blue ribbon) (B). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Results of the leave-one-out-cross-validation (loo-cv) method applied to the model based upon partially pooled data (multilevel) and the completely pooled data (single level) for the *Salmonella* case study 1 with 18 repetitions. The best performing model is put to 0.

	loo-cv value	SE
Multilevel partially pooled	0.0	0.0
Single level completely pooled	−68.3	9.5

of not only judging fits ("retrodiction") but also to study predictive performance. While posterior predictive checks of the model with complete pooling differ only slightly from the one with partial pooling (Fig. S8 in the Supplement), the predictive capacity of the multilevel model is considerably better (Table 3).

This concludes the first case study. All analyses indicated that the Weibull shape parameter consistently indicated nonlinear behaviour with upward curvature. The point estimates obtained from averaged and pooled data were not that different from those from multilevel modeling, but their estimated variation was. If the objective is only to obtain insight which model is suitable to describe observed or measured behaviour, multilevel modeling is less relevant. However, when the objective is to also obtain insight in sources of variation, multilevel modeling is very useful. The biggest advantage is that a more realistic impression of variation is obtained. It counteracts thereby under- and overfitting, as was found to be the case when data were averaged or completely pooled, respectively. When the goal is prediction, partial pooling is shown to be very useful. Moreover, multilevel modeling avoids possible bias in parameter estimation by taking into account dependencies in the data. The disadvantage is of course that more experimental data are needed that can be grouped. There is some debate among statisticians whether or not a minimum of 5 or 6 levels is needed to be able to apply multilevel modeling. According to Gelman and Hill (2007), however, such advice is misguided; they acknowledge that estimation of variance parameters is a concern with small sample sizes, but they maintain that it should work as well as classical regression. In food microbiology, two or three repetitions are common. The dataset with 18 repetitions as used in this paper is in that sense not representative but was useful for illustration purposes. The reward of doing more

repetitions will be a much better characterization of the variation involved and unbiased parameter estimates, which is of particularly high value when the goal is to make predictions, especially in risk analysis. Having explored the possibilities of multilevel modeling for analyzing repetitions done in food microbiology with two levels, the next case study investigates its use in exploring temperature effects in which also repetitions were done, which calls for a three level analysis.

3.2. Multilevel analysis of temperature effects

The data reflecting the various isothermal treatments are shown in Fig. 8: $\log_{10}N$ decreases as a function of time, the more so at higher temperature. With reference to Fig. 1, the number of temperature experiments is k = 9, the number of trials per temperature varied from j = 3–4, and the number of measurements at each trial varied from n = 2–10. The data are quite scattered, as is common in such experiments. The decrease does not seem to be log-linear in time overall, but that will be investigated. The same dataset was used by Koyama et al. (2019) who imposed a linear relation without testing that assumption.

Bayesian regression was again applied with this dataset. The procedure is the same as with the first case study: proposing a likelihood function (the same Weibull model as above) and priors for the parameters. Since this workflow was already explained in detail in the first case study, it is not reported in detail again for the second case study but of course it was adhered to. Where relevant, additional information is provided in the Supplement. The difference with the first case study is the number and the nature of levels.

3.2.1. Single level modeling at each temperature (no pooling)

Each temperature experiment consisted of several trials. In the first analysis, data were not pooled at the group level of temperature but they were pooled within each temperature experiment at the trial level (without applying averaging). The proposed likelihood function and priors are in Supplement Eq. (2). While the resulting fits look good (Supplement Fig. S9), posterior parameter distributions are displayed in Fig. 9 in a forest plot. Shape parameter α shows quite some variation per temperature, while variation in $\log_{10}\beta$ is seemingly erratic as a function of temperature. The shape parameter values reported by Mattick et al. (2001) for the same data, but obtained by single-level least-squares regression, ranged from α = 0.24–1.06, while here a range from
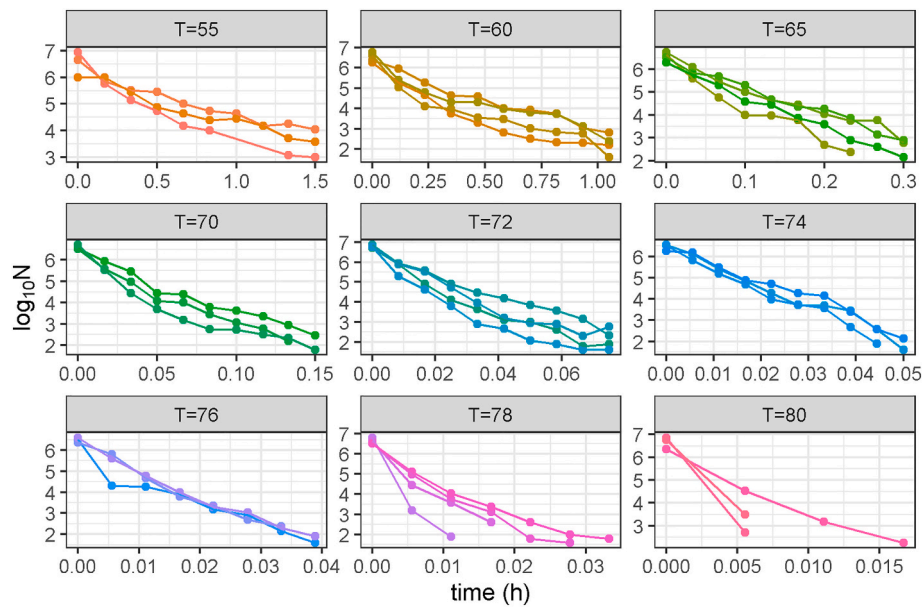
Fig. 8. Overview of the data describing isothermal inactivation of *Salmonella* for temperatures 55–80 °C.
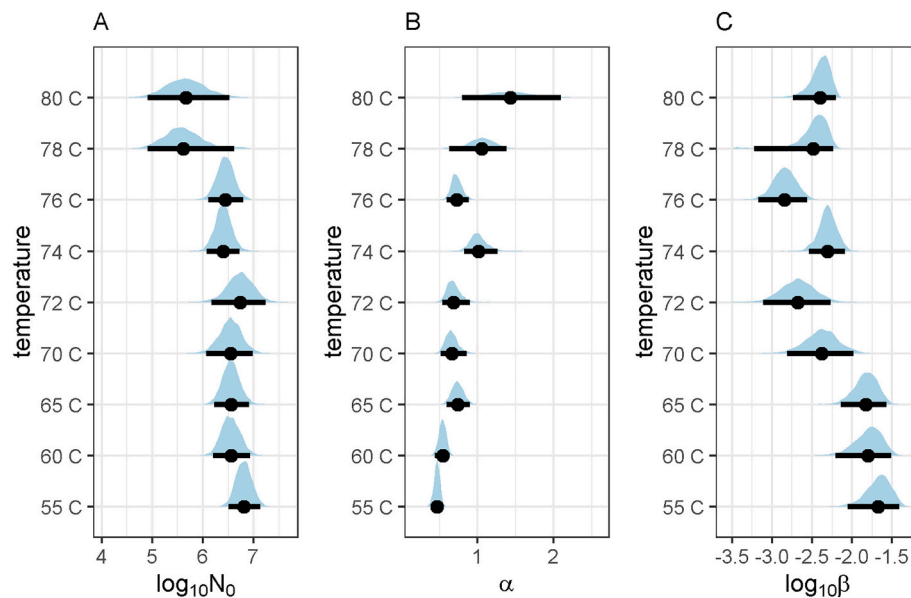Source: Mattick et al. (2001), data extracted from Combase.

Fig. 9. Forest plots showing posterior parameter distributions for the parameter $\log_{10}N_0$ (A), α (B) and $\log_{10}\beta$ (C) resulting from individual regression with the Weibull model per temperature studied.

0.47–1.43 was found. There is, as usual with Weibull parameters, a rather strong parameter correlation between α and β, as shown before in Fig. 3 (an example is shown in Supplement Fig. S10, for the 55 °C experiment). The effects shown in Fig. 9 result from two correlated temperature effects: variation of shape factor α and of rate parameter β, the latter one obviously depending on temperature. When the temperature dependence of rate parameter β needs to be calculated, parameter α should be constant, otherwise there are two sources of variation. The question is which constant value for α is then suitable. Couvert et al. (2005) made observations about the shape parameter variation per temperature. André et al. (2019) also studied this via individual regression at various temperatures, be it at higher temperatures and for spores. They also found variation in the shape parameter per temperature experiment but concluded that it could be fixed at a sort of average

value. They then went on to estimate the dependence of the individually estimated rate parameters as a function of temperature in a subsequent regression procedure, hence a two-step analysis. Here, a different route is taken as explained in the next section. A comparison of this different route with a two-step approach will be done later on, in the section on prediction of decimal reduction times.

3.2.2. Single-level global modeling of completely pooled temperature-dependent data

The different route taken here consists of letting the data decide about the most likely value of parameter α by regressing all data in one step, while integrating the temperature dependence of rate parameter β in the primary model. This is global modeling when using completely pooled data to estimate an overall value for α. The global aspect is that a

model connects the various temperature experiments with temperature as predictor. Pooling all the data implies statistically that data are considered to come from one identical probability distribution, regardless of the temperature at which they were collected; it also implies that the data are considered completely independent and uncorrelated. This last assumption may not be true if data are collected from one batch at a certain temperature treatment (depending on how the measurements were done). Rate parameter β needs to be replaced by a model in which its temperature dependence is made explicit to make it a global model, so that parameter α can subsequently be estimated as a global parameter. Several models are proposed in literature that describe temperature dependence of a rate parameter like β. A recent overview and evaluation can be found in Gil et al. (2017) and Milkievicz et al. (2021). The well-known Bigelow model, normally used to capture the temperature dependence of D-values, is also proposed for rate parameter β:

$$\beta = \beta_{ref} \cdot 10^{\frac{T_{ref}-T}{Z}} \tag{12}$$

A reference temperature T_{ref} is introduced with the corresponding parameter β_{ref}, the value of β at the reference temperature. According to Schwaab and Pinto (2007), the choice for T_{ref} is not arbitrary; the uncertainties in Z and β_{ref} depend on it when applying regression. If calculated according their guidelines, $T_{ref} = 67.25\,°C$ for the present case study. The temperature coefficient Z expresses the increase in temperature needed to reduce β by a factor 10, as with the traditional $D - Z$ model (but note that parameter β is not the same as a D-value!). An alternative model could be the log-logistic model suggested by Peleg (2006), which may be suited for the Weibullian model Eq. (4). To enhance MCMC convergence, parameter values should preferably be in the same range numerically; for that reason, β_{ref}, which has a value in the order of 10^{-3}, was expressed as $\beta_{ref} = 10^{bref}$; the exponent $bref$ is then estimated and has a value in the order of -2 to -3. In the results reported below β_{ref} was calculated back from that exponent, which is easily done from the posterior distribution, a clear advantage of the Bayesian approach; in the frequentist approach this would require more complicated propagation of error calculations.

The proposed likelihood function and priors are shown in the Supplement, Eqs. (3) and (4). Since showing individual fits was not the first goal of this analysis, the resulting fits with their 95% prediction intervals are also shown in Supplement Fig. S11, with the pair plots, posterior parameter densities and correlation coefficients in Fig. S12. It all looks well-behaved; the resulting fits with the global Weibull model show overall, sort of average fits, looking much better than the global log-linear model fits reported in Fig. 2 in Koyama et al. (2019). The numerical parameter summaries of the global Weibull model are reported in Table 4, with $\alpha = 0.62$ showing convincingly again log-nonlinear behaviour and a Z-value of 10.8 °C (note that this was derived using the β parameter, not a D-value).

Single-level completely pooled modeling does not allow individual fits to deviate from the average, it considers each experiment to be the same with the temperature dependence of parameter β accounted for by the Bigelow-type model in this case. These results will be compared to the multilevel outcome, which is the next analysis step.

3.2.3. Multilevel global modeling of partially pooled temperature-dependent data

The previous section showed the regression result for an overall, population value of parameter α. Multilevel modeling allows to investigate if and how parameter α varies per group level (temperature in this case) while still capturing the temperature effect on the rate parameter β with the Bigelow-type model. This will yield, again, an overall value of α at the global, population level, a "fixed" level, but it also allows to capture variation of α at the group level temperature, a "random" effect. The question here, however, is also: to pool or not to pool? The number of trials (repetitions) available at each temperature is limited (3–4 per temperature): it is questionable whether or not there is enough information in the data to extract a varying α at the trial level next to the temperature level. When it is done nonetheless, the results are as follows.

The likelihood function and priors for this case of Bayesian regression are shown in the Supplement (Eq. (5)), as is the pairs plot, posterior parameter distributions and correlation coefficients (Fig. S13). The research question was specifically about the random variation of the parameter α per temperature, so that is shown here in Fig. 10. This variation appears to be rather small at the level of temperature experiments but larger at the level of trial within each temperature group. The large variation of α at the temperature level found with the no-pooled analysis (Fig. 9), is now much smaller due to sharing information between the groups (i.e., temperature experiments). The individual fits ("retrodictions") obtained when the data are partially pooled per temperature, and per trial within the temperature groups, are shown in Supplement Fig. S14, and the fits look remarkably well, also per trial nested in each temperature experiment. The overall fit resulting from the population parameters only is in Fig. 11, which does not differ much from the fit obtained with single-level global modeling with completely pooled data shown in Supplement Fig. S11. But, as already discussed in the first case study, fits should not be the point where the analysis stops. There is a notable difference in partitioning of variance, which has its consequences for predictions. The 95% prediction intervals are narrower than with the completely pooled results (Table 5). Note that there is a small difference in the population parameters ($\alpha = 0.67$ and $\beta = 0.007$ resulting from partial pooling, Table 5) as compared to complete pooling ($\alpha = 0.62$ and $\beta = 0.006$, Table 4). The parameters $\sigma_{\alpha-T}$ and $\sigma_{\alpha-trial}$ confirm what also can be seen in Fig. 10: most of the variation in parameter α comes from the variation due to the group effect of trial, much less from a temperature effect.

3.2.3.1. Model comparison. As with the first case study, model comparison was also done for the second case study by comparing the model based on completely pooled data with the one based on partially pooled data using loo-cv. Results are in Table 6, showing convincingly that the partially-pooled model, with trials nested in temperature clusters, performs much better than the model using completely pooled data, even though the differences in fits were hardly noticeable. Keep in mind, though, that this is a comparison, it does not state that the other model is not performing well; as can be seen in the fits obtained they do perform well in retrodiction but much less so in prediction.

3.2.4. Prediction of decimal reduction values

An important goal of modeling is to test the capacity of a model to predict new, not yet observed values, indicated by prediction bands around the observed values. An important parameter to predict is the heating time needed to cause a certain number of decimal reductions of micro-organisms. With the classical decimal reduction time D for log-linear reduction (if that applies), this is an easy job. With the log-nonlinear dataset analyzed here with the Weibull model, it is also possible to calculate the time needed to reach a predefined number of decimal reductions d provided parameters α and β are known. The general formula is (Van Boekel, 2008):

Table 4

Numerical summary of the parameter estimates from the global completely pooled *Salmonella* data in case study 2, using the Bigelow model to capture the temperature dependence of rate parameter β.

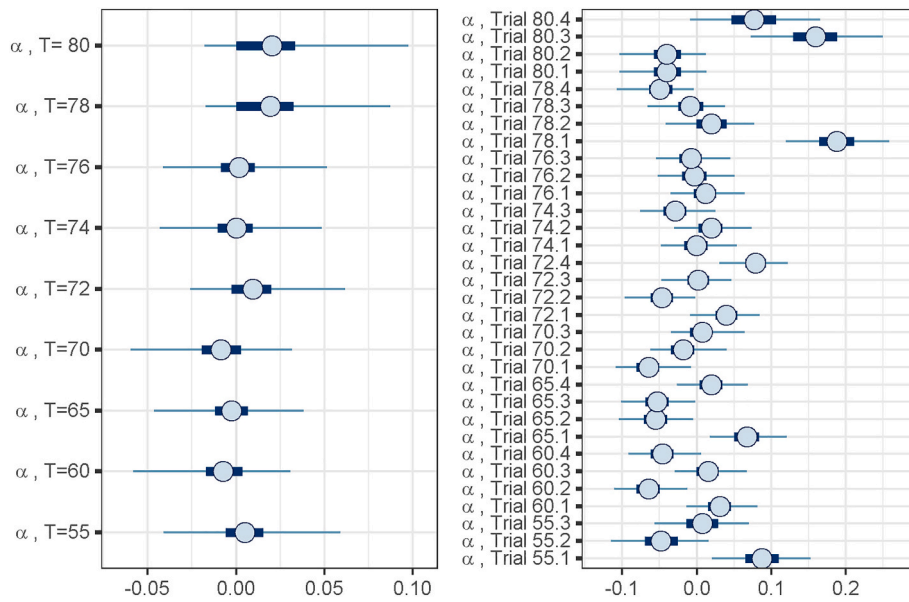|	Mean	SE	Lower bound	Upper bound
$\log_{10}N_0$	6.50	0.01	6.48	6.52
Z (°C)	10.77	0.13	10.51	11.03
β_{ref} (h)	0.006	0.0007	0.005	0.007
α	0.62	0.02	0.58	0.67
σ_e	0.48	0.02	0.44	0.53

Fig. 10. Plot of the variation of parameter α per temperature (left panel) and per trial (right panel), expressed as deviation from the population mean that is given the value 0. The circle indicates the mean, the thin line the 95% credible interval and the thick line the 50% interval.
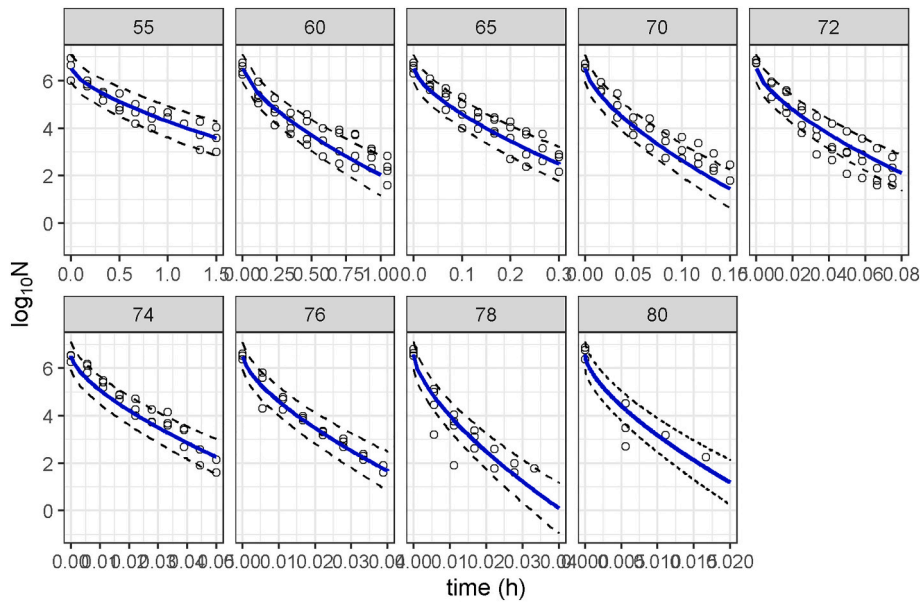

Fig. 11. Fits on the population level with 95% prediction intervals resulting from global multilevel (partial pooling) modeling of the *Salmonella* data using the Bigelow-type model to capture the temperature dependence of β.

Table 5

Numerical summaries of the parameters resulting from the global multilevel partial pooling regression of the *Salmonella* data in case study 2, using the Bigelow model to capture the temperature dependence of parameter β.

	Mean	SE	Lower bound	Upper bound
$\log_{10}N_0$	6.50	0.001	6.50	6.50
Z (°C)	11.04	0.33	10.47	11.78
β_{ref} (h)	0.007	0.0004	0.006	0.008
α	0.67	0.001	0.67	0.67
$\sigma_{\alpha-T}$	0.02	0.02	0.001	0.07
$\sigma_{\alpha-trial}$	0.07	0.012	0.05	0.10
σ_e	0.30	0.01	0.27	0.33

Table 6

Results of 'leave-one-out-cross-validation' for the multilevel partially-pooled-temperature-trial model and for the completely pooled data single level model for the heat inactivation of *Salmonella* in case study 2. The best performing model is put to 0.

	loo-cv-value	SE
Multilevel partially pooled	0.00	0.00
Single level completely pooled	−106.12	14.00

$$t_d = \beta \cdot \left(-ln\left(10^{-d}\right)^{\frac{1}{a}} \right) \tag{13}$$

For a 6D reduction, for instance, this would be:

$$t_d = \beta \cdot \left(-ln\left(10^{-6}\right)^{\frac{1}{a}} \right) = \beta \cdot 6 \cdot ln\left(10\right)^{\frac{1}{a}} = \beta_{ref} 10^{\frac{T_{ref}-T}{Z}} \cdot 6 \cdot ln\left(10\right)^{\frac{1}{a}} \tag{14}$$

In the previous section, values for α, β_{ref} and Z were obtained, so decimal reduction times can be calculated at any desired temperature with these parameters, on the assumption that the models hold at the conditions specified, and, of course, the uncertainty that goes with it. This can be done directly from the posterior parameter distribution. The grand mean population estimates can be used for this, but it could also be done for a specific temperature cluster with its own α parameter, if so desired. Such calculations were done for 6D reduction of the *Salmonella* data analyzed here, by way of example. It might be interesting to compare this result with the two-step method, which consists of first deriving parameters by applying a primary model, followed by subsequent regression of the rate parameter in a secondary model; this was, for instance, done recently by André et al. (2019). To obtain such a two-step result, the data were also analyzed per temperature by fixing the shape parameter at $\alpha = 0.67$ (the grand mean for α) to derive rate parameter β at this fixed value at each temperature, still using the Bayesian approach. The logarithm of this rate parameter was subsequently regressed versus temperature according to the logarithmic version of Eq. (12) to find parameters $log\beta_{ref}$ and Z (a TDT curve, see Supplement Eq. (6) and Fig. S15 and Table S2 for detailed results). The point estimates resulting from this two-step method were $\beta_{ref} = 0.007$ (h) and $Z = 12.5$ °C (the values from the global multilevel regression were $\beta_{ref} = 0.007$ (h) and $Z = 11.0$ °C, Table 5). Decimal reduction times and their uncertainties were subsequently calculated according to Eq. (14), one with the global multilevel model parameters and another with the two-step single level model parameters. By using the posterior parameter distributions, correlations and propagation of uncertainties are automatically taken into account. The time needed for 6 decimal reductions calculated in these two ways are shown in Fig. 12, by way of example at 75 °C. A small difference appeared in the estimate for a 6D reduction time, but more striking is that the global multilevel method shows much less uncertainty as its density distribution is considerably narrower. Also, the density resulting from the two-step single-level modeling approach is skewed to the right, making confidence/credible intervals non-symmetric. Conditional on the model and dataset used here, global multilevel modeling with partial pooling is therefore preferable from the point of view of reducing uncertainty in predicting inactivation. It would be interesting to investigate this finding further with other datasets and models.

4. Conclusion

The possibilities of multilevel modeling in kinetic analysis of heat-induced microbial inactivation were investigated. The findings are expected to be applicable to other datasets, as well as to other models than the Weibull model, but that needs further work. Whether multilevel modeling can be used for non-isothermal processes needs also further investigation. Provided that the data structure allows it, the conclusion is that multilevel modeling is a very powerful and promising method. The following arguments apply in favour of multilevel modeling:

- it connects individual regressions by partial pooling of the data so that an overall impression of the variation of inactivation parameters can be obtained rather than the variation of parameters per individual regression
- the shrinkage phenomenon leads to realistic parameter estimation at the population level suitable for prediction
- it compensates for the possible correlations in data and thereby gives a more realistic impression of the variability involved, which may be underestimated with single level regression
- the prediction capabilities of a multilevel model appear to be (much) better than from single level modeling
- lower parameter uncertainty is not a goal to strive for in selecting models if it means that a model does not describe variation in a system properly; multilevel models are the preferred ones as they guard against over- and underfitting
- Predictive capacities of models should get more attention than their retrodictive capacities and multilevel models appear to be better in that

In addition, it is shown that global modeling of the rate parameter in the Weibull model via a Bigelow-type relation is well possible in a global regression procedure, making traditional separation of primary and secondary models in food microbiology not really necessary for
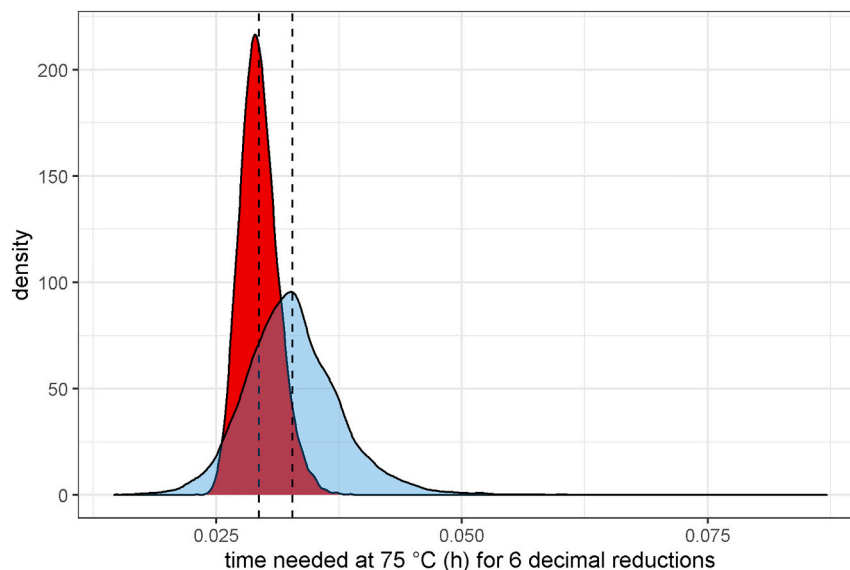

Fig. 12. Probability densities calculated for the time needed for 6 decimal reductions at 75 °C using the parameters from the one-step multilevel model (red fill) and the two-step single level model (turquoise fill). The vertical dashed lines indicate the mean of the distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

parameter estimation. Global modeling reduces uncertainty in predictions such as decimal reduction times as compared to the common two-step approach. Other secondary models such as the log-logistic model may also be suitable to be incorporated in primary models. However, testing primary and secondary models as a first step may still be valuable if the goal is to check what type of model is most applicable to the data at hand. As mentioned in the introduction, the Weibull model is not a panacea for every inactivation experiment. The current analysis is definitely not limited to the Weibull model, it should be applicable to other models as well. This paper used Bayesian regression because it is well suited for multilevel modeling with varying effects for parameters, but it also allows interesting visualizations of how parameters behave. The posterior parameter distribution is a rich source of information for further calculations, as shown here for calculations of decimal reduction times and their uncertainties. Multilevel modeling is also possible in the frequentist framework but subsequent calculations are less straightforward. It is perhaps an investment in time to learn Bayesian regression but it will be worth the effort.

The question posed in the title of this paper: "to pool or not to pool" can be answered as follows. Averaging data should never be done because that removes useful information. Also, analyzing data per individual group may lead to huge variation in parameters and between-group comparison may give a wrong impression about interpretation of what is happening at the population level; no-pooling does therefore not seem to be a good idea either. As for complete pooling or partial pooling, it depends on the goal of the research. The point estimates of parameters appeared to not differ that much between completely-pooled and partially pooled data. However, complete pooling tends to underestimate variation and partition of variance is not possible. If the goal is to get a realistic impression of the variability involved, and especially to make predictions, multilevel modeling using partially pooled data is the way to go. Such a conclusion, however, has its bearings on experimental design because it does require an experimental setup that allows to analyze the data with multilevel modeling, and not unimportantly, it will require more experimental work in the form of more repetitions. The benefits, however, seem very worthwhile in order to be able to deal with the inevitable variability in food microbiology.

Declaration of competing interest

None

Appendix A. Supplementary data

The supplemental material contains additional analyses as well as references to the R packages used. The raw data and the R scripts can be found at the author's Github repository https::/github.com/Tiny-vanBoekel/IJFM. Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijfoodmicro.2021.109283.

References

André, S., Leguerinel, I., Palop, A., Desriac, N., Planchon, S., Mafart, P., 2019. Convergence of Bigelow and Arrhenius models over a wide range of heating temperatures. Int. J. Food Microbiol. 291 (November 2018), 173–180. https://doi.org/10.1016/j.ijfoodmicro.2018.11.019.

Bürkner, P.-C., 2017. brms: an R package for Bayesian multilevel models using Stan. J. Stat. Softw. 80 (1) https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C., 2018. Advanced Bayesian multilevel modeling with the R package brms. R J. 10 (1), 395–411. Retrieved from. http://arxiv.org/abs/1705.11123.

Couvert, O., Gaillard, S., Savy, N., Mafart, P., Leguerinel, I., 2005. Survival curves of heated bacterial spores: effect of environmental factors on Weibull parameters. Int. J. Food Microbiol. 101, 73–81.

De Levie, R., 2012. Collinearity in least-squares analysis. J. Chem. Educ. 89 (1), 68–78. https://doi.org/10.1021/ed100947d.

Den Besten, H.M.W., Wells-Bennik, M.H.J., Zwietering, M.H., 2018. Natural diversity in heat resistance of bacteria and bacterial spores: impact on food safety and quality. Annu. Rev. Food Sci. Technol. 9 (1), 383–410. https://doi.org/10.1146/annurev-food-030117-012808.

Dolan, K.D., Mishra, D.K., 2013. Parameter estimation in food science. Annu. Rev. Food Sci. Technol. 4, 401–422.

Garre, A., Fernández, P.S., Lindqvist, R., Egea, J.A., 2017. Bioinactivation: software for modelling dynamic microbial inactivation. Food Res. Int. 93, 66–74. https://doi.org/10.1016/j.foodres.2017.01.012.

Garre, A., Clemente-Carazo, M., Fernández, P.S., Lindqvist, R., Egea, J.A., 2018. Bioinactivation FE: a free web application for modelling isothermal and dynamic microbial inactivation. Food Res. Int. 1121, 353–360. https://doi.org/10.1016/j.foodres.2018.06.057.

Garre, A., Zwietering, M.H., Den Besten, H.M.W., 2020. 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. 137, 109374 https://doi.org/10.1016/j.foodres.2020.109374.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambirdge University Press, Cambridge, p. 651.

Gelman, A., Carlin, J.B., Stern, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, 3rd edition. Chapman & Hall/CRC, p. 675.

Gelman, A., Lee, D., Guo, Ji, 2015. Stan: a probabilistic programming language for Bayesian inference and optimization. J. Educ. Behav. Stat. 40 (5), 530–543. https://doi.org/10.3102/1076998615606113.

Gil, M.M., Miller, F.A., Brandão, T.R.S., Silva, C.L.M., 2017. Mathematical models for prediction of temperature effects on kinetic parameters of microorganisms' inactivation: tools for model comparison and adequacy in data fitting. Food Bioprocess Technol. 10 (12), 2208–2225. https://doi.org/10.1007/s11947-017-1989-x.

González, S.C., Possas, A., Carrasco, E., Valero, A., Bolívar, A., Posada-Izquierdo, G.D., Pérez-Rodríguez, F., 2019. 'MicroHibro': a software tool for predictive microbiology and microbial risk assessment in foods. Int. J. Food Microbiol. 290 (July 2018), 226–236. https://doi.org/10.1016/j.ijfoodmicro.2018.10.007.

Juneja, V.K., Marks, H.M., 2005. Heat resistance kinetics variation among various isolates of *Escherichia coli*. Innov. Food Sci. Emerg. Technol. 6 (2), 155–161. https://doi.org/10.1016/j.ifset.2004.03.008.

Juneja, V.K., Cadavez, V., Gonzales-Barron, U., Mukhopadhyay, S., Friedman, M., 2016. Effect of pomegranate powder on the heat inactivation of *Escherichia coli* O104: H4 in ground chicken. Food Control 70, 26–34. https://doi.org/10.1016/j.foodcont.2016.05.027.

Koyama, K., Aspridou, Z., Koseki, S., Koutsoumanis, K., 2019. Describing uncertainty in salmonella thermal inactivation using Bayesian statistical modeling. Front. Microbiol. 10 (SEP), 1–7. https://doi.org/10.3389/fmicb.2019.02239.

Kruschke, J.K., 2015. Doing Bayesian Data Analysis, 2nd ed. Academic Press, p. 759.

Kurz, A.S., 2020. Statistical Rethinking With brms, ggplot2, and the tidyverse: Second Edition, second edi. Retrieved from. https://bookdown.org/content/4857/.

Lambert, B., 2018. A Student's Guide to Bayesian Statistics. SAGE publications Ltd, London, p. 498.

Lazic, S.E., Clarke-Williams, C.J., Munafò, M.R., 2018. What exactly is 'N' in cell culture and animal experiments? PLoS Biol. 16 (4) https://doi.org/10.1371/journal.pbio.2005282.

Lazic, S.E., Mellor, J.R., Ashby, M.C., Munafo, M.R., 2020. A Bayesian predictive approach for dealing with pseudoreplication. Sci. Rep. 10 (1), 1–10. https://doi.org/10.1038/s41598-020-59384-7.

Mattick, K.L., Jørgensen, F., Wang, P., Pound, J., Vandeven, M.H., Ward, L.R., Humphrey, T.J., 2001. Effect of challenge temperature and solute type on heat tolerance of Salmonella serovars at low water activity. Appl. Environ. Microbiol. 67 (9), 4128–4136. https://doi.org/10.1128/AEM.67.9.4128-4136.2001.

McElreath, R., 2020. Statistical Rethinking. A Bayesian Course With Examples in R and Stan, Second edition. CRC Press, Boca Raton, p. 612.

Milkievicz, T., Badia, V., Souza, V.B., Longhi, D.A., Galvão, A.C., Robazza, W. da S., 2021. Modeling Salmonella spp. inactivation in chicken meat subjected to isothermal and non-isothermal temperature profiles. Int. J. Food Microbiol. 344 (December 2020) https://doi.org/10.1016/j.ijfoodmicro.2021.109110.

Peleg, M., 2006. Advanced Quantitative Microbiology for Foods and Biosystems. Taylor & Francis/CRC Press, p. 417.

Peleg, M., Cole, M.B., 1998. Reinterpretation of microbial survival curves. Crit. Rev. Food Sci. Nutr. 38, 353–380.

Peleg, M., Normand, M.D., Corradini, M.G., 2017. A new look at kinetics in relation to food storage. Annu. Rev. Food Sci. Technol. 8 (1), 135–153.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, S., R Team, 2020. nlme: linear and nonlinear mixed effects models, version 3.1-151. Retrieved from. https://cran.r-porject.org/package=nlme.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from. https://www.r-project.org/.

RStudio Team, 2020. RStudio: Integrated Development for R. Desktop V. 1.4.869 "Wax Begonia". PBC, Boston MA. Retrieved from. http://www.rstudio.com/.

Santillana Farakos, S.M., Pouillot, R., Anderson, N., Johnson, R., Son, I., Van Doren, J., 2016. Modeling the survival kinetics of Salmonella in tree nuts for use in risk assessment. Int. J. Food Microbiol. 227, 41–50. https://doi.org/10.1016/j.ijfoodmicro.2016.03.014.

Schwaab, M., Pinto, J.C., 2007. Optimum reference temperature for reparameterization of the Arrhenius equation. Part 1: problems involving one kinetic constant. Chem. Eng. Sci. 62, 2570–2764.

Van Boekel, M.A.J.S., 2002. On the use of the Weibull model to describe thermal inactivation of microbial vegetative cells. Int. J. Food Microbiol. 74, 139–159.

Van Boekel, M.A.J.S., 2008. Kinetic Modeling of Reactions in Foods. CRC/Taylor & Francis, Boca Raton.

Van Boekel, M.A.J.S., 2020. On the pros and cons of Bayesian kinetic modeling in food science. Trends Food Sci. Technol. 99, 181–193. https://doi.org/10.1016/j.tifs.2020.02.027.

Van Boekel, M.A.J.S., 2021. Kinetics of heat-induced changes in foods: a workflow proposal. J. Food Eng. 306, 110634 https://doi.org/10.1016/j.jfoodeng.2021.10634.

Van Boekel, M.A.J.S., Ter Steeg, P.F., Dahoe, A.E., 2020. Co-optimization of safety, quality and legislation: opening Pandora's box? Curr. Opin. Food Sci. 35, 65–71. https://doi.org/10.1016/j.cofs.2020.02.001.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27 (5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.