10.2478/msr-2014-0009

Linear mixed models: GUM and beyond

Barbora Arendacká[1], Angelika Täubner[2], Sascha Eichstädt[1], Thomas Bruns[2] and Clemens Elster[1]

Physikalisch-Technische Bundesanstalt
[1]Abbestr. 2-12, 10587 Berlin, Germany, barbora.arendacka@ptb.de
[2]Bundesallee 100, 38116 Braunschweig, Germany

In Annex H.5, the Guide to the Evaluation of Uncertainty in Measurement (GUM) [1] recognizes the necessity to analyze certain types of experiments by applying random effects ANOVA models. These belong to the more general family of linear mixed models that we focus on in the current paper. Extending the short introduction provided by the GUM, our aim is to show that the more general, linear mixed models cover a wider range of situations occurring in practice and can be beneficial when employed in data analysis of long-term repeated experiments. Namely, we point out their potential as an aid in establishing an uncertainty budget and as means for gaining more insight into the measurement process. We also comment on computational issues and to make the explanations less abstract, we illustrate all the concepts with the help of a measurement campaign conducted in order to challenge the uncertainty budget in calibration of accelerometers.

Keywords: Linear mixed models, uncertainty, GUM, ANOVA, random effects.

1. Introduction

The paper deals with methods suitable for an analysis of long-term repeated experiments, which are commonly conducted to assess reproducibility of a measurement. For instance, in case of accelerometer calibration, an accelerometer may be repeatedly mounted into the measurement setup and each time its frequency response function (FRF) may be measured over a range of different frequencies. Such an experiment reveals variability that may not be observed in a routine calibration, when an accelerometer is mounted into the setup only once. However, this variability is of interest when creating an uncertainty budget. Moreover, such an experiment may yield further insights into the measurement, either by revealing issues deserving further investigation or by confirming that things work as expected. The data collected in such a long-term experiment come usually in groups. For example, in case of the repeatedly mounted accelerometer, we have a set of measurements for every mounting and it seems natural to expect that measurements obtained within one mounting are more interrelated than measurements obtained for two different mountings. This already suggests using linear mixed models for the data analysis. We will elaborate on this, giving an explanation what linear mixed models are and what everything they can capture, in section 2. For now, we just note that linear mixed models are closely related to random effects ANOVA (ANalysis Of VAriance) or mixed effects ANOVA models, terms that the reader may be more familiar with, since the former appears in the GUM, Annex H.5 [1], or in the ISO/TS 21749:2005 [2]. Since ANOVA models are special cases of linear mixed models, we prefer to use the latter, more general term to refer to models that include several random (and possibly also some fixed) effects. Examples of practical use of ANOVA models with random effects can be found e.g. in [3, 4, 5]. These papers actually illustrate the span of ANOVA applications: uncertainty evaluation in the line with the GUM in [3], measurement process inspection in [4] and a mixture of questions about both fixed and random effects in [5].

While the GUM explains in detail only the most simple case of an ANOVA model (see [6] for its Bayesian treatment), we would like to show that further aspects observed in measurements in practice can be incorporated into the model and that besides components of uncertainty, such an analysis can answer further questions of interest. This is discussed in section 3. In section 4 we comment on computational issues. The closed-form formulas, as reported in Annex H.5 in [1], or in the ISO norm [2] are available only in relatively simple cases. In more complex situations one has to resort to other approaches, e.g. a (restricted) maximum likelihood estimation. However, this can be done using statistical software.

Since our presentation of the topic is closely linked to the already mentioned long-term experiment connected with the calibration of accelerometers, in section 5 we present some concrete results and discuss their interpretation. Section 6 offers concluding remarks. Before proceeding, let us now describe the nature of the accelerometer experiment underlying our presentation of linear mixed models.

1.1. A practical case

The data shown in this paper come from a measurement campaign conducted in connection with the EMRP project IND09 ’Traceable dynamic measurement of mechanical quantities’, where the focus is on torque, force and pressure. However, acceleration is a fundamental quantity for dynamic measurements and the campaign should serve as a model campaign for the other three quantities, exploiting the fact that a traceable primary calibration of accelerometers has already been realized in PTB [7, 8]. In the campaign, a single-ended (SE) accelerometer Brüel & Kjaer (B&K) type 8305-001 and a back-to-back (BB) accelerometer Brüel & Kjaer (B&K) type 8305

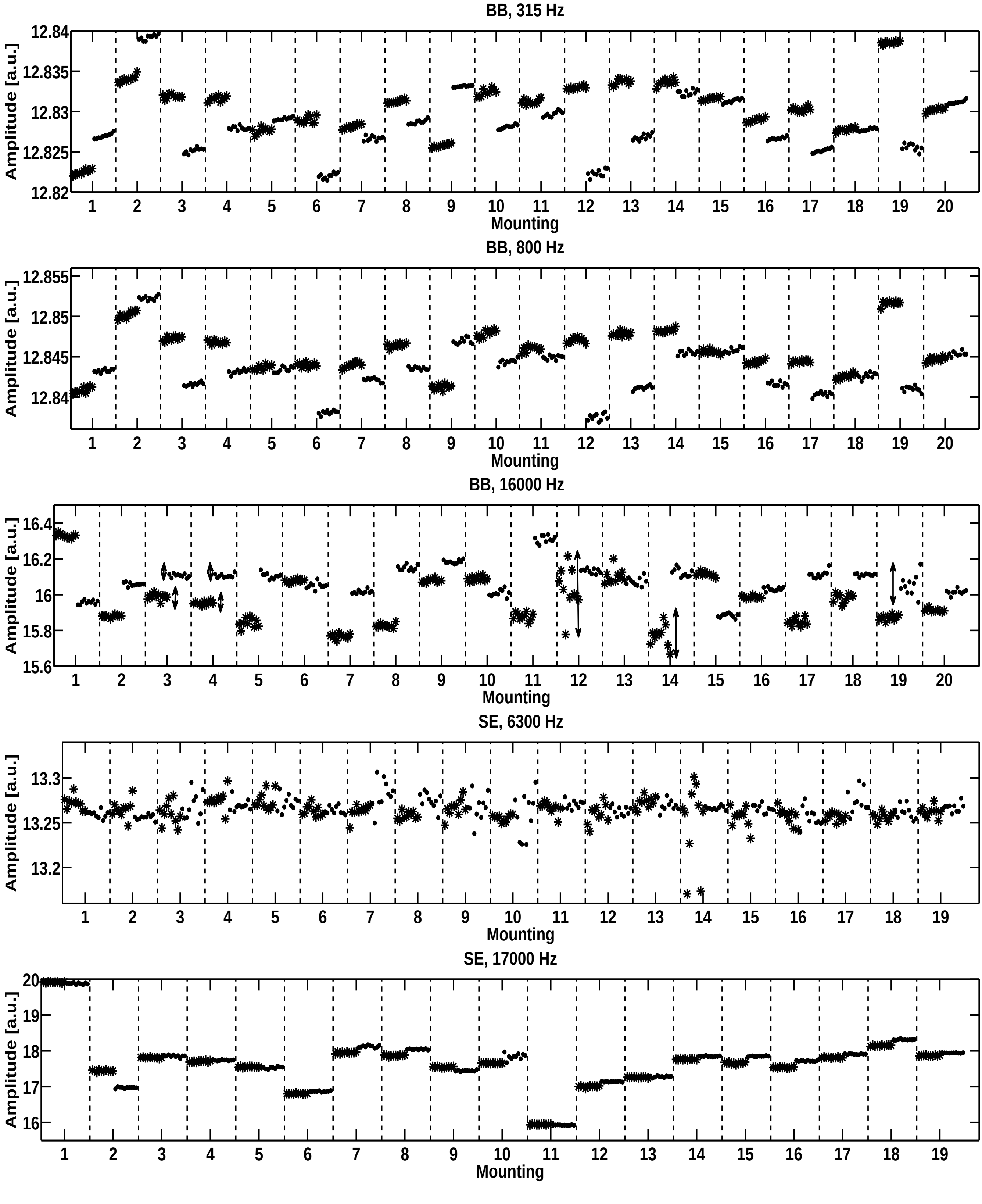

Fig. 1: Amplitudes (in a.u.) of the FRF obtained for selected frequencies with a repeatedly mounted accelerometer (SE = single-ended, BB = back-to-back) and input acceleration determined with the interferometer in position 90°(*) and 0°(·). The arrows indicate different variability observed due to random error (BB, 16000 Hz). Note that in order to show the finer patterns in fluctuations clearly, the y-axis scale varies between the plots.

were repeatedly (19 and 20 times, respectively) mounted into the measurement setup and values of the FRF were measured for a range of frequencies (10 Hz – 20 kHz) using a sinusoidal excitation of the system. The input acceleration was measured with a laser interferometer pointing its rays at two points on the top of the BB accelerometer and at two points next to the bottom of the SE accelerometer. In each case these two points lay on a line having two possible angles with a certain reference surface: 0° and 90°, which we will refer to as position 0° and position 90°. For each mounting, the FRF was measured 10 times in each position, amplitude and phase of the input and output signals being determined through a sine fit. The measurements were done within roughly 5 weeks with both accelerometers being mounted usually once on a given day. The order of the accelerometers on the given day was randomized, in order to avoid accidental bias, since comparison of relative standard uncertainties obtained for the two accelerometers was also of interest. Here, we will use only some parts of the measured data to show what features observed in reality may be captured by a linear mixed model and we will refer to the setup of the experiment when explaining the nature of the models. The repeated measurements for us will be always the repeatedly measured amplitudes of the FRF at a given frequency. Figure 1 shows these for several chosen frequencies as obtained in the experiment. Each plot depicts a dataset that will be analysed separately.

2. Linear Mixed Models

2.1. Basic ANOVA model

Looking at Fig. 1, it is clear that the simplest model of values varying randomly and independently around the measured quantity ($y_{mpr} = \mu + \varepsilon_{mpr}$) is not really applicable to the data obtained in our long-term experiment. The fluctuations the observed amplitudes exhibit have a finer structure, which may be described by adding random and/or fixed effects to the simple model. Which effects might be considered stems from the nature of the experiment and the way of modeling it. In our case, we vary mounting of the accelerometer and position of the laser beams of the interferometer, so we can think of effects of position, mounting and their interaction, mounting-position. The *effect of mounting* will be modelled as random. This means that we assume that measurements obtained within one mounting are influenced by a common value, which changes randomly and independently from mounting to mounting. However, these effect sizes or levels are centered at zero and have a finite variance. The mounting effect includes all the uncontrollable external influences that may vary between the different mountings as well as any influences caused by small imprecisions in the mounting procedure. In contrast to mounting, the *effect of position* will be modelled as fixed. That means that the same value is supposed to influence the measurements whenever they are taken at position 0°, and similarly, one and the same value is manifested when position 90° is measured. The common assumption is that these values, the levels of the effect (i.e. the values of the possible systematic biases for position 0° and for position 90°) sum to zero. This is necessary for the model to be identifiable and it ensures that as a whole, the measured values still vary around the measured quantity. The position effect is considered as fixed since the two positions are well defined and can be repeatedly realized. The small discrepancies that may arise from every (manual) adjustment of the interferometer are then covered by the random *mounting-position effect*. That covers also other influences that may arise from the interaction of a given mounting and a given adjustment of the position, i.e. effects observed due to changing both mounting and position. The discrepancy of an actual measured value (y_{mpr}) from the measured quantity (μ) is modelled as a sum of the different effects and the model

$$\begin{aligned} y_{mpr} = \mu + b_p + A_m + B_{mp} + \varepsilon_{mpr}, \\ m = 1,\ldots,20; p = 0^\circ, 90^\circ; r = 1,\ldots,10, \\ b_{0^\circ} + b_{90^\circ} = 0, \end{aligned} \tag{1}$$

is summarized (with distributional assumptions) in Table 1.

The model in Table 1 is a mixed effects ANOVA model. The unknown parameters to be estimated are μ, b_{0°, b_{90°, σ_M^2, σ_{MP}^2 and σ^2. Note that for the fixed effect the values of the effect levels are of interest (b_{0°, b_{90°), while in case of random effects the respective variances are in focus (σ_M^2, σ_{MP}^2, σ^2). One can estimate also the realized values of the random effects if desired (e.g. the values of $A_1,\ldots, A_{20}$). However, since these quantities are random, this is referred to as prediction in the literature (even though no future aspect is present).

Table 1 shows also the distributional properties for the measurements implied by the model. The normal distribution comes from the fact that a sum of normally distributed, mutually independent random variables is again normally distributed. Further, if we arrange all the y_{mpr} in a vector, this would have a multivariate normal distribution with a certain mean vector (entries $\mu + b_{0^\circ}$ for observations in position 0° and $\mu + b_{90^\circ}$ for observations in position 90°) and a covariance matrix Σ. Unlike in the simple model with independent, identically distributed random errors, Σ in the model in Table 1 is not diagonal. The observations are in a certain way correlated. This is not surprising, since, e.g. all amplitudes measured within one mounting are influenced by the same value of the mounting effect, which results in a positive correlation. Actually, the model in Table 1 does not allow negative correlations at all. Namely, the correlation between amplitudes obtained within one mounting but at different positions is $\frac{\sigma_M^2}{\sigma_M^2+\sigma_{MP}^2+\sigma^2}$, while amplitudes obtained within the same combination of mounting and position have correlation $\frac{\sigma_M^2+\sigma_{MP}^2}{\sigma_M^2+\sigma_{MP}^2+\sigma^2}$. Amplitudes obtained within different mountings are modelled as independent.

2.2. Extension - linear trend

The model in Table 1 belongs to the ANOVA family. One speaks about an ANOVA model when the observations are

Model: $y_{mpr} = \mu + b_p + A_m + B_{mp} + \varepsilon_{mpr}$, $m = 1,\dots,20; p = 0^\circ, 90^\circ; r = 1,\dots,10$

Term	Random	Levels	Distribution	Implications
common mean	no	μ	–	
position	no	b_{0°, b_{90°, $(b_{0^\circ} + b_{90^\circ} = 0)$	–	$y_{mpr} \sim N(\mu + b_p, \sigma_M^2 + \sigma_{MP}^2 + \sigma^2)$ (for correlations see text in Sec. 2.1)
mounting	yes	$A_1,\dots, A_{20}$	$A_m \sim N(0, \sigma_M^2)$, iid	
mounting-position	yes	$B_{1,0^\circ}$, $B_{1,90^\circ},\dots,$ $B_{20,0^\circ}$, $B_{20,90^\circ}$	$B_{mp} \sim N(0, \sigma_{MP}^2)$, iid	$\bar{y}_{mp\cdot} \sim N(\mu + b_p, \sigma_M^2 + \sigma_{MP}^2 + \sigma^2/10)$
random error	yes	$\varepsilon_{1,0^\circ,1},\dots, \varepsilon_{1,90^\circ,10},\dots,$ $\varepsilon_{20,90^\circ,1},\dots, \varepsilon_{20,90^\circ,10}$	$\varepsilon_{mpr} \sim N(0, \sigma^2)$, iid	$\bar{y}_{m\cdot\cdot} \sim N(\mu, \sigma_M^2 + \sigma_{MP}^2/2 + \sigma^2/20)$

Table 1: y_{mpr} is the rth measured amplitude within the mth mounting at position p, μ is the value of the amplitude of the FRF (i.e. the measured quantity), b_p is the fixed effect of the position, A_m denotes the random effect of mounting, B_{mp} the random effect of mounting-position, ε_{mpr} the random error and iid stands for independent, identically distributed. All the random effects and random errors are assumed to be mutually independent. Shown are also the distributional implications of the model for the measured amplitude y_{mpr}, the average amplitude over one mounting-position combination $\bar{y}_{mp\cdot}$ and the average amplitude taken over measurements obtained within one mounting $\bar{y}_{m\cdot\cdot}$. The latter average would be reported as the amplitude for the given frequency in a routine calibration.

simple sums of the different effects. Whenever the dependence is more complicated (but still linear in the effects), one calls the model a linear mixed model. For example, if we include a trend for observations obtained within each mounting-position, the model becomes

$$\begin{aligned} y_{mpr} = \mu + b_p + cr + A_m + B_{mp} + C_{mp}r + \varepsilon_{mpr}, \\ m = 1,\dots,20; p = 0^\circ, 90^\circ; r = 0,\dots,9, \\ b_{0^\circ} + b_{90^\circ} = 0, \end{aligned} \tag{2}$$

where the parameter c represents a common trend in the repeated measurements of amplitude for the given frequency and the random effect C_{mp} provides its modification for each mounting-position combination. We assume $C_{mp} \sim N(0, \sigma^2_{MPtrend})$. The index r denoting the order of the measurement starts now from 0, which reflects the fact that we assume to start from the value of the measured quantity and observe a slight increase for the repetitions. That such an extended model may be useful when describing real measurements is supported by the amplitudes we observed at frequency 315 Hz, see Fig. 1. There, for most mounting-position combinations the repeatedly measured amplitudes exhibit a slight linear increase. Note that for the average over a mounting it now holds:

$$\bar{y}_{m\cdot\cdot} \sim N\left(\mu + c\bar{r}, \sigma_M^2 + \frac{\sigma_{MP}^2}{2} + \frac{\sigma^2_{MPtrend}}{2}\bar{r}^2 + \frac{\sigma^2}{20}\right), \tag{3}$$

where $\bar{r} = \sum_{i=0}^{9} r/10$ and there is a systematic bias ($c\bar{r}$) when the simple average $\bar{y}_{m\cdot\cdot}$ is used as an estimate of the measured quantity.

2.3. Extension - heteroscedasticity

Other modification of the basic model supported by our data is related to the random error variability. The model in Table 1 assumes that the fluctuations due to the random error have the same variance σ^2. This may be a good approximation for measurements at 800 Hz, see Fig. 1. However, looking at the amplitudes obtained at 16000 Hz, the variability of measurements within the mounting-position combinations is rather different. Compare e.g. the fluctuations in mounting 12 at 90°, mounting 14 at 90° or mounting 19 at 0° with the fluctuations of the repeated measurements obtained within mounting 2 or 4 at either position. This can be incorporated into our model by allowing the variance of the random error to vary between the mounting-position combinations, i.e.

$$\begin{aligned} &\varepsilon_{1,0^\circ,r} \sim N(0, \sigma^2_{1,0^\circ}),\ r = 1,\dots,10; \\ &\varepsilon_{1,90^\circ,r} \sim N(0, \sigma^2_{1,90^\circ}),\ r = 1,\dots,10; \\ &\dots \\ &\varepsilon_{20,90^\circ,r} \sim N(0, \sigma^2_{20,90^\circ}),\ r = 1,\dots,10. \end{aligned}$$

The number of parameters in the model is then increased by 39 in our case, 1 common σ^2 is replaced by 40 (the number of mounting-position combinations) different variances: $\sigma^2_{1,0^\circ}$, $\sigma^2_{1,90^\circ}$, $\sigma^2_{2,0^\circ}$,$\sigma^2_{1,90^\circ}$,..., $\sigma^2_{20,0^\circ}$, $\sigma^2_{20,90^\circ}$. In this case, for the average over a mounting we have:

$$\bar{y}_{m\cdot\cdot} \sim N\left(\mu, \sigma_M^2 + \frac{\sigma_{MP}^2}{2} + \frac{1}{4}\left(\frac{\sigma^2_{m0^\circ}}{10} + \frac{\sigma^2_{m90^\circ}}{10}\right)\right). \tag{4}$$

2.4. Extension - correlations

So far all the random effects and their levels were considered independent and the correlations between the measurements were induced only by the influence of the same effect level on several measured values. However, the model may be generalized by allowing correlations also between the random effects. For example, we may consider a model with only mounting-position effects, which would be however correlated within one mounting. This would allow, e.g. for the possibility of negative correlation between measurements at different positions within the same mounting.

3. Benefits

As already mentioned, the GUM [1] shows in Annex H.5 an instance where a random effects ANOVA model should be employed for the analysis of measured data. Namely, the example deals with a Zener voltage standard calibrated against a stable voltage reference over a two-week period. The measurements are done repeatedly (5 times) on 10 different days. The statistical model underlying the analysis in [1] is

$$y_{dk} = \mu + A_d + \varepsilon_{dk}, \quad d = 1,\dots,J(=10), k = 1,\dots,K(=5), \tag{5}$$

where y_{dk} is the kth measurement of the voltage on day d, $A_d \sim N(0,\sigma_D^2)$ stands for a random effect of day and $\varepsilon_{dk} \sim N(0,\sigma^2)$ are the random errors. From the measurements, one wants to obtain the common mean, μ, which is in this case the measured voltage, and the associated type A uncertainty. These values are, as stated in the GUM, not the final results of the calibration. Of course, the estimated value of μ has to be compared with the stable voltage reference and the uncertainty has to be amended to include type B contributions. The voltage is estimated as the average over all measurements $\bar{y}_{..}$, and since from (5) we have $\bar{y}_{..} \sim N(\mu, \sigma_D^2/J + \sigma^2/(JK))$, it is straightforward to obtain the associated type A uncertainty, once the model is fitted to the data, i.e. the parameters are estimated: $u_{\bar{y}_{..}} = \sqrt{\widehat{\sigma}_D^2/J + \widehat{\sigma}^2/(JK)}$ (the hats denote estimates). In the GUM notation, $\widehat{\sigma}^2$ is s_W^2, $\widehat{\sigma}_D^2$ is s_B^2 and the formula for the uncertainty may be compared with formulas (H.32) and (H.29) in [1]. If the calibration of accelerometers was always conducted with the accelerometer mounted repeatedly into the setup, we could proceed similarly as in the GUM, estimate the model parameters by fitting one of the models from the previous section to the measured amplitudes and report the estimated $\hat{\mu}$ (the amplitude of the FRF at the given frequency) and its type A uncertainty (depending on the method of estimation). However, in our case μ is not in the center of interest, since the experiment was not done with the aim to calibrate the given accelerometer. The aim was to assess the reproducibility of the calibration procedure. Thus we are primarily interested in the variance parameters σ_M^2, σ_{MP}^2, σ^2, also called variance components, since they decompose the overall variability into several separate sources. Considering this decomposition may be of interest in itself, since it provides further knowledge about the measurement process. For example, in the case reported in the GUM, Annex H.5 [1], it is recommended that apparent day-to-day variability (i.e. larger positive value of σ_D^2) should be a reason for investigation of its cause.

Coming back to our experiment with accelerometers, the variability due to mounting and mounting-position is not observed when the accelerometer is mounted only once[1], however, one would like to cover it in the overall uncertainty. If the variability due to these sources is considered inherent to the measurement setup, one can use the same values of $\widehat{\sigma}_M^2$, $\widehat{\sigma}_{MP}^2$, coming from the long-term repeated experiment, for results from future calibrations. Thus these variance components might be included as entries in an uncertainty budget. For a future calibration experiment, they would be type B components. Similarly, it is often the case that an uncertainty budget includes also a fixed term for the repeatability (σ^2 in model in Table 1), since this is considered a property of the measurement setup, see e.g. [9], section 5.6. Examining results from a long-term repeated experiment using mixed linear models enables us to check whether the repeatability variance remains constant over time. This means examining whether a model with common σ^2 fits the measurements well enough as compared to a model with, for example, different repeatability variances for each mounting-position combination. This assessment may be done by examining a plot of residuals obtained after the respective model is fitted, or by carrying out a likelihood ratio test comparing both fits (see e.g. [10], p. 83). As a by-product, the revealed changes in the repeatability variance may trigger further investigations of the measurement setup.

Last but not least, fitting a mixed linear model to a long-term repeated experiment provides quantitative answers to questions about the fixed effects as well. For example, in our case we can determine the presence and size of the systematic bias between amplitudes measured at the two positions, or quantify the effect of neglecting a trend present in the measurements when calculating just a simple average of the amplitudes and reporting it as the final calibration result.

4. How to squeeze the numbers out

In the previous two sections we have seen that data generated in a long-term repeated experiment exhibit a structured behaviour that can be modelled by linear mixed models. We have also pointed out what benefits such an analysis may bring regarding both the overall understanding of the measurement process and the establishment of an uncertainty budget. In this section we would like to discuss practical aspects of carrying out such an analysis, mentioning ways how the models can be fitted to particular data (i.e. how the estimates of the parameters may be obtained).

4.1. ANOVA table

Consider first the basic model in Table 1. In what follows, we will describe some apparently reasonable estimators for the unknown parameters: μ, b_{0°, b_{90°, σ_M^2, σ_{MP}^2, σ^2, which will establish a connection to the formulas stated in [1, 2] and which in fact turn out to have also optimal statistical properties in this case (they have minimum variance among all unbiased estimators in their class (linear, quadratic respectively)), for details see [11], pp. 129, 161.

- $\hat{\mu}$, $\hat{b}_{0^\circ}$, $\hat{b}_{90^\circ}$: It seems indeed natural to estimate the common mean μ (the measured quantity) as the average over all measurements: $\bar{y}_{...}$, and the position effect b_{0° (b_{90°) as the difference between the average of all measurements at position 0° (90°) and the overall mean:

[1] Strictly speaking, the component σ_{MP}^2 could be estimated from measurements done within one mounting, but it would be like estimating a variance from two observations (within one mounting we have only two mounting-position combinations).

$\bar{y}_{\cdot 0^\circ \cdot} - \bar{y}_{\dots}$ ($\bar{y}_{\cdot 90^\circ \cdot} - \bar{y}_{\dots}$). These are the (weighted) least squares estimators in the model, see [11], p. 160.

- $\widehat{\sigma}_M^2$: Since measurements obtained for different mountings are modelled as independent and each mounting mean $\bar{y}_{m\cdot\cdot}$ has the same distribution, see Table 1, their sampling variance $S_M^2 = \frac{1}{20-1}\sum_{m=1}^{20}(\bar{y}_{m\cdot\cdot} - \bar{y}_{\dots})^2$ estimates the variance from Table 1, i.e. $\sigma_M^2 + \sigma_{MP}^2/2 + \sigma^2/20$. Thus, if we had estimates $\widehat{\sigma}_{MP}^2$, $\widehat{\sigma}^2$ for σ_{MP}^2 and σ^2, we could estimate the variance of the mounting effect, σ_M^2, as $\widehat{\sigma}_M^2 = S_M^2 - \widehat{\sigma}_{MP}^2/2 - \widehat{\sigma}^2/20$.

- $\widehat{\sigma}^2$: Since within each mounting-position combination we have 10 repeated measurements, their sampling variability estimates the random error variance. We can thus obtain 40 estimates of the same quantity, namely, for each mp combination $S_{mp}^2 = \frac{1}{10-1}\sum_{r=1}^{10}(y_{mpr} - \bar{y}_{mp\cdot})^2$, where $\bar{y}_{mp\cdot} = \frac{1}{10}\sum_{r=1}^{10} y_{mpr}$ is the average of measurements within one mounting-position combination. The random error (or the repeatability) variance may be then estimated as the average of the individual estimates, namely $\widehat{\sigma}^2 = S^2 = \frac{1}{20*2}\sum_{m=1}^{20}\sum_{p\in\{0^\circ,90^\circ\}} S_{mp}^2$.

- $\widehat{\sigma}_{MP}^2$: Before motivating the estimate for σ_{MP}^2, observe that the estimate for σ^2 is based on fluctuations of the individual observations with respect to the mounting-position mean, and the estimate for σ_M^2 on fluctuations of the mounting means with respect to the overall mean. The estimate of σ_{MP}^2 should be then based on fluctuations of the mounting-position means. However, looking at such fluctuations with respect to the overal mean would not leave the mounting part of the variation, and thus σ_M^2, out. Thus we need to consider these fluctuations with respect to the overal mean adjusted for the position effect and the mounting effect: $\mu_{mp}^{adj} = \bar{y}_{\dots} + (\bar{y}_{\cdot p \cdot} - \bar{y}_{\dots}) + (\bar{y}_{m\cdot\cdot} - \bar{y}_{\dots})$. So we calculate $S_{MP}^2 = \frac{1}{(20-1)(2-1)}\sum_{m=1}^{20}\sum_{p\in\{0^\circ,90^\circ\}}(\bar{y}_{mp\cdot} - \mu_{mp}^{adj})^2$, which can be rewritten in the more familiar form $S_{MP}^2 = \frac{1}{(20-1)(2-1)}\sum_{m=1}^{20}\sum_{p\in\{0^\circ,90^\circ\}}(\bar{y}_{mp\cdot} - \bar{y}_{\cdot p\cdot} - \bar{y}_{m\cdot\cdot} + \bar{y}_{\dots})^2$. S_{MP}^2 is an unbiased estimate for $\sigma_{MP}^2 + \sigma^2/10$, thus $\widehat{\sigma}_{MP}^2 = S_{MP}^2 - \widehat{\sigma}^2/10$.

Before proceeding, note that the estimates $\widehat{\sigma}_M^2$ and $\widehat{\sigma}_{MP}^2$ may be negative. The common practice is to take this as an indication that the estimated variance is zero, i.e. the affected random effect is not present in the model. For a full discussion see [11], p. 130.

4.2. *Need for a general approach*

The closed form formulas as stated above are the so called ANOVA estimators and the sums of squares S^2, rS_{MP}^2, rpS_M^2 are the mean sums of squares from the so called ANOVA table. These tables for a variety of ANOVA models can be found in [12], the most common models are usually covered in standard statistical textbooks as well and these tables underlie the development in Annex H.5 of the GUM [1] and the ISO norm [2]. Apart from the fact that one has to look up the form of the table for each type of the ANOVA models, these procedures are based on two other important assumptions and that is the balancedness of the model and common repeatability variance.

A model is balanced if we have the same number of repetitions on each level, i.e. in our case we have always two positions measured within each mounting and we have always 10 repeated measurements within each mounting-position combination. The importance of balancedness becomes clear when we consider that the estimates for the parameters described above are based on averages at different levels. If these averages are obtained from the same number of observations, combining them together is straightforward.

Even though one can usually plan a reproducibility experiment in this balanced way, there might be outliers observed and one may want to omit them from the analysis, ending up with a slightly unbalanced situation. For example, look at Fig. 1 and the amplitudes measured at 6300 Hz. There seems to be some outliers in mounting 14, or one may consider all the measurements at position 90° in this mounting as strange. If such measurements were always discarded in the calibration, it would be reasonable to omit them from the reproducibility analysis as well. That would, however, lead to an unbalanced design - for mounting 14 we would have only measurements at one position. Use of formulas for a balanced case in a slightly unbalanced situation may not lead to completely unreasonable results, however, one should give it a serious consideration and the analysis becomes less straightforward. Especially, realization of statistical tests in such situations is more complicated (see e.g. [12]).

Apart from the balancedness, the closed form formulas above are derived for models with a common repeatability variance in all groups. If this is not the case, e.g. we want to fit our data with the model from Table 1 but with different random error variances for each mounting-position combination, the availability of closed form formulas for estimation is very limited. Similarly, if the model to be fitted is not a simple ANOVA model, but we want to consider trends (like e.g. in model (2)) or some correlations between the effects, the closed form formulas are not available at all. Thus one has to consider more general approaches to fitting linear mixed models that are applicable in any situation. Such a universal method is the Maximum Likelihood (ML) or the REstricted(or REsidual) Maximum Likelihood (REML).

4.3. *(Restricted) Maximum Likelihood*

As already mentioned, any of the models in section 2 implies a likelihood for our measured amplitudes, namely a multivariate normal distribution with a certain mean vector and a certain covariance matrix. By the ML method one then maximizes the likelihood in the parameters (e.g. μ, b_{0°, b_{90°, σ_M^2, σ_{MP}^2, σ^2) that appear in the mean vector and the covariance matrix. The maximization is constrained, since $b_{0^\circ} + b_{90^\circ} = 0$ and $\sigma_M^2 \geq 0$, $\sigma_{MP}^2 \geq 0$, $\sigma^2 > 0$. An alterna-

		BB, 315 Hz	BB, 16000 Hz	SE, 17000 Hz
μ:	est.	12.829	16.023	17.670
b:	est.(95% CI)	-0.0024 (-0.0047,0.000032)	0.1131 (0.0383,0.1880)	0.0510 (-0.0226,0.1245)
c:	est.(95% CI)	6.2e-5 (5.4e-5,6.9e-5)	-	-
σ_M:	est.(95% CI)	0.00076 (4e-6,0.1467)	0(-)	0.7459 (0.5362,1.038)
σ_{MP}:	est.(95% CI)	0.0036 (0.0026,0.0050)	0.1167(0.0931,0.1462)	0.1077 (0.0776,0.1496)
$\sigma_{MPtrend}$:	est.(95% CI)	1.5e-5 (1e-5,2.4e-5)	-	-
σ_{mp}:	range	6e-5–5.1e-4	0.0066–0.1198	0.0055–0.0855

Table 2: REML estimates for the different parameters in mixed linear models fitted to our data. CI stands for confidence interval as calculated by the software, 'est.' stands for estimate, '-' means the respective parameter was not considered in the model (e.g. for 16000 Hz no trend was considered). All the fitted models were heteroscedastic, b estimates the systematic difference between position 0° and 90°. 0 in the column for 16000 Hz is not a direct estimate from R, but the inspection of the likelihood showed that the maximum is attained at the border of the parametric space (for more details, see section 5.2.1). In such a case, however, there is no automatic statement for the confidence interval.

tive is the REML method, which first fits the fixed effect part with a least squares fit and then the residuals are used for estimating the variance parameters. At the end, the fixed effects are reestimated by weighted least squares with weights derived from the estimated covariance matrix. The maximum likelihood part in the REML procedure is in the estimation of the variance components, since these are determined to maximize the likelihood of the residuals. The maximization is constrained in this case as well, the variance components are non-negative. It is advisable to use a statistical software both for the ML and the REML method. A useful reference in this context is [13] explaining the model fitting with SAS, SPSS, Stata, R/S-plus, and HLM. While the optimization routines in general are not trivial, for certain groups of linear mixed models - without correlations between the random effects and with a common repeatability variance - one can obtain the ML/REML estimates also using an iterative algorithm based on the so-called Henderson's equations. See [11], p. 278-286, or [14] for the theory and procedure *mixed.m* on Matlab File Exchange [15] for a MATLAB implementation. We note that an adjustment of the iterative algorithm to a case when the random error variances differ in the different groups of data is possible and we plan to report on it elsewhere.

Both the ML and the REML methods are accompanied by an asymptotic theory regarding the distribution of the estimators, which then enables carrying out statistical tests about or constructing confidence intervals for the parameters. One may worry about the validity of these tests for finite number of measurements (the theory operates with infinite datasets) and there exist lots of adjustments trying to improve the test performance in practice, see e.g. [14]. Moreover, in each concrete example, one may use a simulation to look closer at the quality of the tests. This is described in a simple way in [16], Chapter 8. Another source of worry may be the assumption of normality, however, this seems to be acceptable in many practical situations, see e.g. the examples in [10]. The REML estimation is usually preferred, because in case of a balanced ANOVA model the results coincide with the results obtained from the closed form formulas from the ANOVA table (as long as these formulas yield positive estimates - i.e. estimates from the parametric space), see [11], p. 253. As already mentioned, in balanced models, these closed form formulas are not only intuitively appealing, but they are in a sense optimal, see section 4.1. Thus in these special models, REML yields better estimates of the variance components than ML. This feature is commonly described as accounting for the error due to estimation of fixed effects, and it is assumed to carry over to more general models.

Remark 1. As to the estimation of fixed effects, there are usually some constraints of the type $b_{0^\circ} + b_{90^\circ} = 0$. In our case, a possible way how to take these into account is to replace b_{0° by $b/2$ and b_{90° by $-b/2$, for some b. This b then represents the systematic difference between position 0° and 90°, which we may be directly interested in. (Interpretation of μ remains unchanged.) And we have to estimate only μ and b and the variance components in the fitting procedure, instead of formally: μ, b_{0°, b_{90° and the variance components.

5. Results and their interpretation

Table 2 shows results of fitting mixed linear models to amplitudes of the FRF at 315 Hz, 16000 Hz and 17000 Hz. The estimates were obtained by REML method using function *lme()* in the library *nlme* in R.

5.1. Diagnostics

Since application of any fitting procedure to any data yields some numerical results, before interpreting the estimated values or applying statistical tests, one should check whether the fitted model seems plausible for the data. By this we mean to check whether the assumptions we made do not appear to be in sharp conflict with the data. There exist lots of different graphical diagnostic tools, exploring different aspects of the mixed linear model, and we will show here only a few examples. For a more extensive treatment see [13, 16, 10].

The first to check are the residuals. A software usually offers a possibility to calculate normalized residuals, which correspond to normalized random errors. These should follow the standardized normal distribution and be independent. The residuals can be plotted against the fitted values or against the

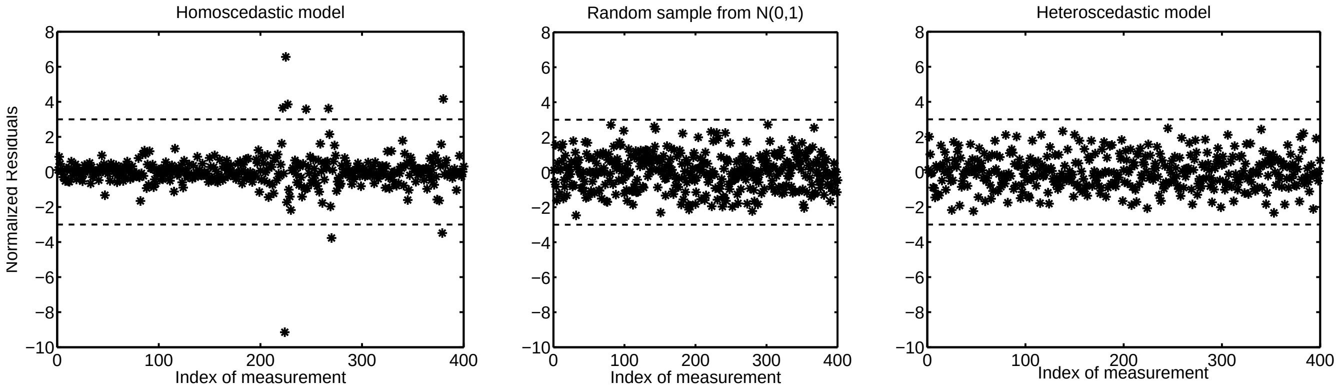

Fig. 2: Normalized residuals from fitting a mixed model with the same variance for all random errors (left) and a mixed model assuming different variances within mounting-position groups (right) to amplitudes at 16000 Hz for the back-to-back accelerometer. The middle plot shows a random sample from a standardized normal distribution for a better comparison.

index of measurement, the plots can be further subdivided according to the groups that appear in the model. What should not be visible in these plots are systematic trends or unequal variability in the values. The first would suggest that there are some systematic influences left out of our model (and thus we might think of extending the fixed effects part), the latter would suggest that our assumption about the variances for random errors has to be adjusted. For example, Fig. 2 shows normalized residuals from fitting a homoscedastic model (the same variance for all random errors) to the data from the back-to-back accelerometer at 16000 Hz. It is immediately clear, that under our assumptions (i.e. in our model), the normalized residuals do not have the same variability, contrary to what is expected. A model allowing for different variances in individual groups captures the data better. Besides visual comparison of residuals, the two models can be compared using a likelihood ratio test as well. In this test the maximum of the (restricted) likelihood in both models is compared and lack of difference would suggest that adopting the more complicated model is not necessary. In R the test can be done using function *anova()* (library *nlme*). For 16000 Hz, the likelihoods are 757.33 and 974.06 for the homoscedastic and the heteroscedastic model, respectively, and the accompanying p-value is less than 0.0001, supporting the conclusion that the heteroscedastic model is better suited.

The normality assumption may be checked by looking at a quantile-quantile plot. It plays a role when interpreting outcomes of statistical tests, since in case of serious violation of normality, the reported p-values may be misleading. Maximizing normal likelihood in order to estimate the parameters may then become questionable, too.

Second, we should look at the predicted random effects. We assume that random effects are normally distributed and (depending on the model) independent. The normality assumption may be checked by the quantile-quantile plots, plotting the effects against each other might help to discover correlations that are actually not present in the model. Take, for example the 315 Hz measurements. Figure 3 shows a quantile-quantile plot for the predicted mounting-position effects, B_{mp}, which does not indicate a violation of normality. The other plots are trying to discover a violation of the independence assumption. Plotting predicted C_{mp} values against the predicted B_{mp} might have revealed some correlation between the mounting-position effects and the random trend effects, which are, however, assumed to be independent. The scatter plot of pairs $(B_{m0^\circ}, B_{m90^\circ})$ might have revealed a dependence within the mounting-position effects, which would again violate our assumptions.

For both 315 Hz and 16000 Hz, the described diagnostic tools do not indicate a serious problem with the assumed model. Checking the fit for the amplitudes at 17000 Hz and single-ended accelerometer, mounting 1 and 11 would be suggested as possibly outlying. In addition, there seems to be some correlation between the B_{mp} effects within one mounting. This may suggest considering an alternative model without the mounting effect and with correlated $B_{m0^\circ}, B_{m90^\circ}$ for a common m. Then we would get $\widehat{\sigma}_{MP} = 0.7537$ and $corr(B_{m0^\circ}, B_{m90^\circ}) = 0.98$.

5.2. *Interpretation and use*

Let us now have a look at the results in Table 2.

5.2.1. **Effect of mounting**

Looking at the estimated variability for the mounting effect for 315 Hz (BB accelerometer), we see that the value is low compared to the mounting-position variability and in addition, it is accompanied with a confidence interval stretching over several orders of magnitude. This, according to [10], p.27, usually indicates problems with the model definition, in our case it suggests that the mounting effect may be redundant. This is further supported by comparison of the maximum of likelihoods for model with (2690.314) and without the mounting effect (2690.296), which are practically identical. (Formally testing that a certain random effect is not present in the model requires some caution, since the p-values reported by *anova()* in R tend to be larger than the true p-values, see [10], p. 87.)

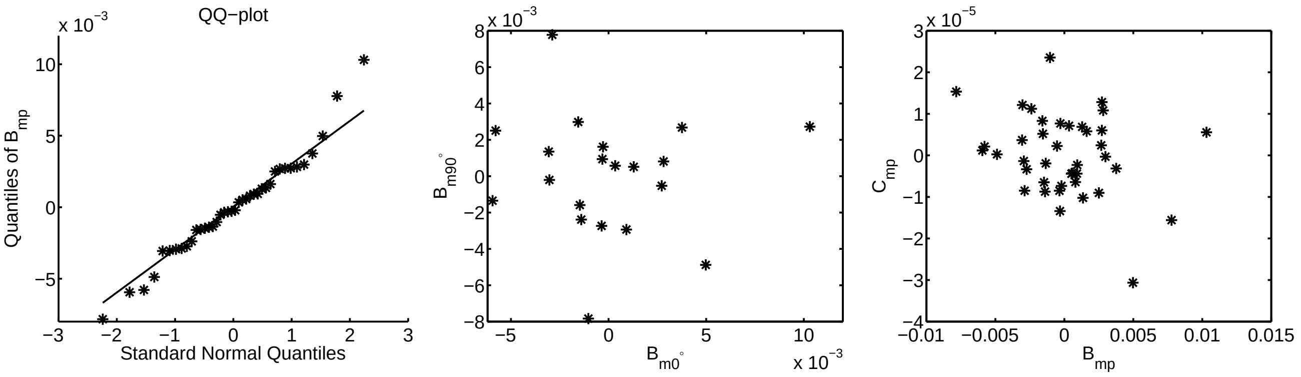

Fig. 3: Left: Quantile-quantile plot of the predicted mounting-position effects (B_{mp}). Not to violate the assumption of normality, the plot should be close to linear. Middle: Scatter plot of mounting-position effects at position 90° against these effects at position 0° not showing any obvious correlation. Right: Scatter plot of mounting-position effects B_{mp} against the predicted sizes of random trends C_{mp}, again without any obvious correlation.

The situation is similar for 16000 Hz (BB accelerometer). When trying to fit a model with mounting effect, the *lme()* function had problems achieving convergence when searching for the maximum of the restricted likelihood. An inspection of the restricted likelihood showed that the maximum is likely to be attained at the border of the parametric space, i.e. for $\sigma_M = 0$, which again suggests that the mounting effect is negligible.

The only case showing a dominant common effect for measurements obtained within one mounting was the case of amplitudes measured at 17000 Hz with the SE accelerometer. The common effect within each mounting is actually observable also from Figure 1, where all measurements within one mounting are strongly shifted in the same direction. This frequency is, however, known to be the transverse resonance frequency of the accelerometer, which is reflected also in the huge variability of the repeated measurements. As such, it is not reported in a calibration certificate and we mention it only to demonstrate a span of possible results from fitting a mixed linear model to repeated measurements.

5.2.2. Effect of mounting-position

For both 315 Hz and 16000 Hz the mounting-position effect (σ_{MP}) dominates the variability. This may not be surprising, considering that the position of the interferometer is each time manually adjusted. However, the analysis does not exclude that this effect has also contributions coming from a combination of mounting and position. Unlike the mounting effects, these are then independent between the positions.

5.2.3. Uncertainty budget

As mentioned in section 3, by fitting a mixed linear model to measurements obtained in a reproducibility experiment we can assess contributions to the uncertainty budget. In our case, the uncertainty budget would be created for a result of a calibration, in which an accelerometer is mounted into the setup and its frequency response function is measured at several frequencies. Taking amplitude at a certain frequency as an example, one would repeatedly (let's say 10 times) determine the amplitude with interferometer at position 0° and at position 90°. Then the average amplitude, $\bar{y}_{m\cdot\cdot}$, will be reported as the final result. Consider as an example the measurements at 315 Hz. The formula for the uncertainty observed directly in the long-term experiment, combining (3) and (4), is

$$u_{\bar{y}_{m\cdot\cdot}} = \sqrt{\sigma_M^2 + \frac{\sigma_{MP}^2}{2} + \frac{\sigma_{MPtrend}^2}{2}\bar{r}^2 + \frac{1}{4}\left(\frac{\sigma^2_{m0^\circ}}{10} + \frac{\sigma^2_{m90^\circ}}{10}\right)}. \tag{6}$$

As already discussed in section 3, considering the variability due to mounting and mounting-position (including the random trend) inherent to the experiment, we may use the estimated values in the formula (6) also for future measurements. The random error variability (or repeatability), $\sigma^2_{m0^\circ}$, $\sigma^2_{m90^\circ}$, may be obtained in each calibration for each position as the sampling variance of the repeated measurements. The other option is to observe that even though the variability of the random error varies, it does not cause much difference in the final expression for the uncertainty (6) and thus in practice we may consider some fixed value for it. For the measurements at 315 Hz, (6) equals 0.00266 when calculated with $\sigma_{m0^\circ} = \sigma_{m90^\circ} = 6.5e-5$, as well as when calculated with $\sigma_{m0^\circ} = \sigma_{m90^\circ} = 5.1e-4$, which are the minimal and maximal random error variabilities observed in the experiment. In future calibrations we just have to check that the repeated measurements do not exhibit variability that is substantially different from what we observed in the reproducibility experiment.

For a comparison, uncertainty of 0.00266 corresponds to a relative uncertainty of 0.021 % (0.00266/12.829, see Table 2), which constitutes 21 % of the overall relative uncertainty that would be calculated according to the current uncertainty budget. This was considered satisfactory.

5.2.4. Fixed effects

Last but not least, let us look at the estimates of the fixed effects. The parameter b estimates the systematic difference between the amplitudes measured with interferometer at positions 0° and 90°. This is accelerometer specific and not of direct interest, since the averaging over two positions is done with the intention to cancel out the biases. More interesting is the estimate of the trend (c). It appears in measurements at certain frequencies and as mentioned in section 2.2, ignoring it by the final averaging causes a systematic bias $c\bar{r}$ in the estimate of the measured quantity. Our analysis enables us to quantify the size of this bias relative to the mean: for 315 Hz it is 2.79e-4/12.829, which is, at the level of 0.002 %, negligible and well covered by the uncertainty.

6. Discussion and conclusions

Mixed linear models are well established in statistics. Even though there might still be a space for improvement, the state-of-the-art and its implementation in software packages enable a convenient application of these models to real data. In this paper we tried to point out benefits of application of mixed linear models for the analysis of long-term repeated experiments from a metrological perspective. We showed what all features, commonly observed in data (heteroscedasticity, trends), can be easily incorporated into a mixed linear model, discussed fitting procedures and interpretation of the results. Long-term repeated experiments are not the only situations when mixed linear models appear in metrology. A recent application appeared e.g. in calibration of flow meters [17] and a large amount of literature dealing with common mean estimation and key comparisons deals with a special case of a mixed linear model, too.

Acknowledgement

The research leading to these results has received funding within an EMRP project IND09. The EMRP is jointly funded by the EMRP participating countries within EURAMET and the European Union.

References

[1] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML. (2008). *Guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections)*. JCGM 100:2008. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

[2] ISO. (2005). *Measurement uncertainty for metrological applications - repeated measurements and nested experiments*. ISO/TS 21749:2005.

[3] Sakurai, H., Ehara, K. (2011). Evaluation of uncertainties in femtoampere current measurement for the number concentration standard of aerosol nanoparticles. *Measurement Sience and Technology*, 22, 024009.

[4] Lee, J., Yang, J., Yang, S., Kwak, J. (2007). Uncertainty analysis and ANOVA for the measurement reliability estimation of altitude engine test. *Journal of Mechanical Science and Technology*, 21 (4), 664–671.

[5] Wang, D.Y., Lin, K.-H., Lo Huang, M.-N. (2002). Variability studies on EMI data for electronic, telecommunication and information technology equipment. *IEEE Transactions on Electromagnetic Compatibility*, 44 (2), 385–393.

[6] Toman, B. (2006). Linear statistical models in the presence of systematic effects requiring a Type B evaluation of uncertainty. *Metrologia*, 43 (1), 27–33.

[7] von Martens, H.-J., Link, A., Schlaak, H.-J., Täubner, A., Wabinski, W., Göbel, U. (2004). Recent advances in vibration and shock measurements and calibrations using laser interferometry. In *Sixth International Conference on Vibration Measurements by Laser Techniques: Advances and Applications*. SPIE, Vol. 5503, 1–19.

[8] ISO. (1999). *Methods for the calibration of vibration and shock transducers — Part 11: Primary vibration calibration by laser interferometry*. ISO 16063-11:1999.

[9] Jackett, R.J., Barham, R.G. (2013). Phase sensitivity uncertainty in microphone pressure reciprocity calibration. *Metrologia*, 50 (2), 170–179.

[10] Pinheiro, J.C., Bates, D.M. (2000). *Mixed-effects Models in S and S-PLUS*. Springer.

[11] Searle, S.R., Casella, G., McCulloch, C.E. (1992). *Variance Components*. John Wiley & Sons.

[12] Burdick, R.K., Graybill, F.A. (1992). *Confidence Intervals on Variance Components*. Marcel Dekker.

[13] West, B.T., Welch, K.B., Gałecki, A.T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC.

[14] Witkovský, V. (2012). Estimation, testing, and prediction regions of the fixed and random effects by solving the Henderson's mixed model equations. *Measurement Science Review*, 12 (6), 234–248.

[15] Witkovský, V. (2000). *mixed.m — Matlab algorithm for solving Henderson's mixed model equations*. http://www.mathworks.com/matlabcentral/fileexchange/200-mixed.

[16] Gelman, A., Hill, J. (2009). *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge University Press.

[17] Wübbeler, G., Mickan, B., Elster, C. (2013). Bayesian analysis of sonic nozzle calibration data. In: *FLOMEKO 2013: 16th International Flow Measurement Conference*, 24-26 September 2013. CEESI.

Received December 16, 2013.
Accepted March 28, 2014.