ORIGINAL ARTICLE

Strengthen forensic entomology in court—the need for data exploration and the validation of a generalised additive mixed model

Michèle Baqué · Jens Amendt

Abstract Developmental data of juvenile blow flies (Diptera: Calliphoridae) are typically used to calculate the age of immature stages found on or around a corpse and thus to estimate a minimum post-mortem interval (PMI$_{min}$). However, many of those data sets don't take into account that immature blow flies grow in a non-linear fashion. Linear models do not supply a sufficient reliability on age estimates and may even lead to an erroneous determination of the PMI$_{min}$. According to the Daubert standard and the need for improvements in forensic science, new statistic tools like smoothing methods and mixed models allow the modelling of non-linear relationships and expand the field of statistical analyses. The present study introduces into the background and application of these statistical techniques by analysing a model which describes the development of the forensically important blow fly *Calliphora vicina* at different temperatures. The comparison of three statistical methods (linear regression, generalised additive modelling and generalised additive mixed modelling) clearly demonstrates that only the latter provided regression parameters that reflect the data adequately. We focus explicitly on both the exploration of the data—to assure their quality and to show the importance of checking it carefully prior to conducting the statistical tests—and the validation of the resulting models. Hence, we present a common method for evaluating and testing forensic entomological data sets by using for the first time generalised additive mixed models.

Introduction

Forensic entomology is the analysis of insect evidence for forensic purposes [3]. Depending on the level of accessibility and environmental conditions, necrophagous insects will promptly colonise a fresh corpse. By calculating the age of developing insects on a body and by analysing the composition of its insect fauna, the expert may be able to deduce the time when insects first colonised the body which can infer a minimum time since death or post-mortem interval (PMI$_{min}$). Establishing this period of time is the most important task of a forensic entomologist.

Because blow flies are usually the first group to colonise a body, the focus is often on them when using entomological evidence to estimate the minimum PMI. Analysing the size of the larvae or identifying which stage of immature development they have achieved allows the approximation of age, and such methods are well supported by existing research and are widely described and accepted in the forensic community. However, due to the recent burst in development of the forensic sciences, new court criteria require the evaluation of scientific evidence prior to its submission to the court. The Daubert Standard states that scientific evidence should be testable, have a known error rate, be peer-reviewed and be accepted by the specific scientific community employing the technique [24]. Moreover, a 2009 US Research Council report indicated a need for major improvements in many disciplines of forensic science in order to increase accuracy and meet those standards. Clearly, one of these improvements may be the introduction of new

M. Baqué (✉) · J. Amendt
Institute of Forensic Medicine, Goethe-University Frankfurt,
Kennedyallee 104,
60596 Frankfurt am Main, Germany
e-mail: baque@em.uni-frankfurt.de

statistical analyses in forensic entomology for the validation of methods in court proceedings [16]. Estimations of insect age are calculated using curvilinear regression, isomegalen diagrams, isomorphen diagrams and thermal summation models [2, 12, 22]. However, estimation of juvenile age in forensic entomology suffers due to the fact that blow flies grow in a non-linear fashion and show, e.g., variable size distributions at different larval stages which unequally affect the estimation of their age [24]. Not surprisingly, scientists have reported different growth rates for the same species of fly, and experts have come to incongruent conclusions about a PMI based on the same entomological evidence depending on which growth data were utilised [24]. This is at least partly due to the wrongful application and misuse of statistical methods. Several authors realised that problem and have attempted to improve the methods used [e.g. 20, 21] or introduce new statistical tools and formulas in forensic entomology [12, 13, 24, 26]. However, it is our belief that a scientific paper used as a basic principle for a court report should provide detailed descriptions of the experimental settings, an appropriate presentation of the data exploration and finally a sound validation and justification of the applied statistics.

Hence, the present paper exemplary analyses some growth parameters of the blow fly *Calliphora vicina*, testing various statistical methods in order to reveal differences in the determined regression parameters and to show their impact on the examined characters of the specimens. Besides the validation of the chosen test, we focus especially on the evaluation of the data, which should reveal their quality and the importance of checking it carefully prior to conducting the statistical tests, a fact which is so far often neglected when publishing growth data in forensic entomology [see [29] for a general discussion of this problem in an ecological context].

Material and methods

Flies

Two laboratory cultures of *Calliphora vicina* were reared in wire cages at the Institute of Forensic Medicine in Frankfurt am Main, Germany. Flies were fed with sugar and water ad libitum. Blood was offered as a protein source every two days. Fresh beef liver was provided as an oviposition medium once a week. After oviposition, the eggs were transferred to a LinTek MKKL 600/2 incubator, set at 20±1°C under a lighting cycle of 12:12 h (light/dark). Constant temperature was validated using a DS1922L Temperatur Logger iButton (Maxim/Dallas), measuring every 1 h. The time needed for hatching was not recorded, but larvae were used soon after emergence.

Sampling

Resulting larvae were reared in population densities of 10, 50 and 100 individuals at three constant temperatures (5°C, 20°C and 30°C). To ensure comparability between the measurements at the three temperatures, sampling was not specified in "days after hatching", but converted to accumulated degree days (ADD). Based on previously published data regarding developmental landmarks in *Calliphora vicina* (larval stages two and three) [4, 5, 8, 10, 15], sampling times were defined as 40 ADD and 120 ADD after hatching, assuming a minimum developmental threshold of 0°C (Table 1). Each density/temperature/ADD-combination was repeated five times within a single generation (five replicates of each combination) (*n*=4800).

Hatched larvae were placed on ground beef in separate plastic cups (400 mL). To prevent intraspecific competition, the amount of meat was increased proportionally (10 individuals received 25g minced meat, 100 individuals received 250g ground beef, respectively).

Each cup was kept in a large plastic container (5000 mL) which was covered with a fine paper towel held in place by a rubber band. The container was filled with 3 cm of sawdust which served as medium for pupariation. The cups were moved daily within the incubators to count out incubator temperature effects and to prevent pseudoreplication. A datalogger (DS1922L) was placed in each incubator to record their temperature stability at 1-h intervals.

For each sample time, all larvae from the meat were removed and killed with boiling water. Impacts on larval development by handling effects could be ruled out because larvae were not disturbed until sampling time. They were then placed in 95% ethanol to avoid postmortem changes [1, 23]. Length measurements were performed using a geometrical micrometer to 0.1 mm [25], and their weight was estimated using an electronic balance (reading precision: *d*=0.001 g).

Table 1 Summary of the experimental treatments

Temperature Sampling time	5°C	20°C	30°C
ADD 40	5×10[a]	5×10[a]	5×10[a]
	5×50[b]	5×50[b]	5×50[b]
	5×100[c]	5×100[c]	5×100[c]
ADD 120	5×10[a]	5×10[a]	5×10[a]
	5×50[b]	5×50[b]	5×50[b]
	5×100[c]	5×100[c]	5×100[c]

[a] 10 individuals

[b] 50 individuals

[c] 100 individuals

Regression analysis

Various regression analyses served to introduce different statistical methods in order to identify faulty model validations and to find ways of reproducing the data adequately.

Statistical analyses like linear regression, GAM and GAMM were carried out using the R statistical package [19].

Length was used as response variable. *Weight*, *Density*, *Temperature* and *Time* were used as possible explanatory variables, where *Density* consisted of the characteristic values 10, 50 and 100 (coded in R as: 1, 2 and 3, respectively), *Temperature* consisted of the values 5°C, 20°C and 30°C (coded in R as: 1, 2 and 3, respectively) and *Time* was defined through the values 40 ADD and 120 ADD (coded in R as: 1 and 2, respectively).

To find the optimal model in terms of statistical criteria, the Akaike Information Criteria (AIC) was consulted, which is a measure of goodness of fit: The preferred model is the one with the lowest AIC value [18, 27].

Data exploration

For the detection of outliers, bivariate dotplots [7], boxplots and scatterplots were used. Collinearity and the relationships between variables were revealed using a pairplot. Preliminary exploration of the data indicated a strong collinearity between *Weight* and *Time*. Additionally, a non-linear relationship between *Weight* and *Length* was highlighted. Therefore, either *Weight* or *Time* should be used in the fixed part of the model but never both at the same time as this would lead to difficulties during the model selection process and give larger *p*-values.

Linear regression

The non-linear relationship between *Length* and *Weight* caused the elimination of *Weight* from the analysis. Therefore, the variables used in this model were *Length* as response and *Density*, *Temperature* and *Time* as nominal explanatory variables. Additionally, two- and three-way interaction terms were included.

Finding the optimal linear model involves verifying the main underlying assumptions: homogeneity, independence and normality [18, 27, 28]. To verify homogeneity, scatterplots were used in which the residuals were plotted against the fitted values as well as against each explanatory variable that was either used or not used in the model.

To detect dependent structures, the residuals were plotted against *Weight* using a scatterplot. Normality can be checked by a normal QQ-plot.

Homogeneity and independence are the most important assumptions, and if one assumption is not confirmed, the model should be rejected.

Generalised additive model

Because of non-linear relationships and heterogeneous and dependent structures, the next step was fitting a generalised additive model (GAM). Instead of a slope parameter, GAMs use a non-linear smoothing function to summarise the relationship between X and Y [24, 27].

Due to the collinearity between *Weight* and *Time*, the variables used in this model were *Length* as response and *Density*, *Temperature* and their interaction terms as nominal explanatory variables. *Time* was dropped out and replaced by a smoothing function of *Weight*.

Based on the results for the residual plots of this model, a normal (Gaussian) distribution with an identity link function was suggested. The optimal number of degrees of freedom for the smoother was estimated through cross-validation (a method in which observations are left out and instead are predicted while the smoother is used for the rest of the data.)

The model validation process is approximately the same as in linear regression, which means that the main underlying assumptions have to be verified. Here, the model still showed dependent structures due to the hierarchical, nested data set (five batches within one approach). One possible solution to improve the model without transforming variables, which means a loss of information, is a generalised additive mixed model.

Generalised additive mixed model (GAMM)

The generalised additive mixed model (GAMM) is an extension of generalised additive modelling [28]. Unknown smooth functions of different types of covariates as well as random effects can be added. Therefore, they provide a broad and flexible framework for regression analyses in complex situations.

For this data set, the consecutive number of the batches of all approaches (*BatchID*) was included as a random structure. To deal with heterogeneity, a variance structure for the residuals was included in which each temperature level was allowed to have a different variance. Because *Temperature* was a nominal variable, a VarIdent variance structure was used [17, 28].

The other variables used in this model were *Length* as response and *Weight* as a smoothing function. The model validation is still the same as in linear regression and in GAM.

Results and discussion

Data exploration

The exploration of the data was done using all variables. As three of the explanatory variables were categorical and as they therefore cannot have any different values than 1, 2 or

3, it was only checked for outliers in the response variable and in the continuous variable *Weight*. The spread in length was nearly the same in both time regimes, whereas the spread in weight was bigger for 120 ADD than for 40 ADD. The dotplots indicated that *Length* is the more accurate measureable variable, and therefore, it presented the better response variable.

Both boxplots and dotplots showed some points that are beyond the centre of the data, indicating that these might be points that could influence the analysis incorrectly. Transformation of data might help here, but this process would lead to a loss of information as it would compress the data, and because the points were not that extreme, it was decided not to transform the data.

Pairplots and conditional plots of *Length* versus each individual explanatory variable showed no linear relationship between *Length* and *Density*, a weak linear correlation between *Length* and *Temperature* and, as expected, a strong relationship between *Length* and *Time* (Fig. 1). Additionally a collinear relationship between *Weight* and *Time* could be highlighted which is obvious too as weight increases with time. A scatterplot of *Length* versus *Weight* showed linear patterns at the start of the development and non-linear patterns with a wider margin at length values from 1.5 to 2.2 cm, which means that maggots with a weight of 0.1 g, for example, may have length values between approximately 1.5 and 1.9 cm (Fig. 2).

Because of the weak correlation between *Length* and *Temperature* and *Density*, the growth rate patterns do not change by temperature and density (Fig. 3). Additionally, it became apparent that individuals who were exposed to the same amount of accumulated degree days, especially 40 ADD, were longer the higher the temperature was. This means that despite the supposed equality of development, the use of the standard ADD method is suspect and should be taken with care when calculating the age of the larvae as this may lead to an incorrect estimation of the PMI [4, 5] (Fig. 4).

After the evaluation of the data, it became clear that there were no extreme outliers and that *Weight* and *Time* were collinear. So, two options for further analysis opened up:

1. Drop *Weight* out, and use *Time* as the categorical variable for a linear regression.
2. Drop *Time* out, and use a smoothing function of *Weight* for a generalised additive model.

Because of its simplicity, linear regression was the first analysis carried out.

Linear regression

The model in the linear regression analysis contained the explanatory variables *Temperature*, *Density*, *Time* and their interaction terms. The model has the form:

$$
\begin{aligned}
\text{Length}_{ij} = {} & \alpha + \beta 1 \times \text{Density}_i + \beta 2 \times \text{Temperature}_i \\
& + \beta 3 \times \text{Time}_i + \beta 4 \times \text{Density}_i \times \text{Temperature}_i \\
& + \beta 5 \times \text{Density}_i \times \text{Time}_i + \beta 6 \times \text{Temperature}_i \times \text{Time}_i \\
& + \beta 7 \times \text{Density}_i \times \text{Temperature}_i \times \text{Time}_i + \varepsilon_{ij}.
\end{aligned}
$$

$$(1)$$

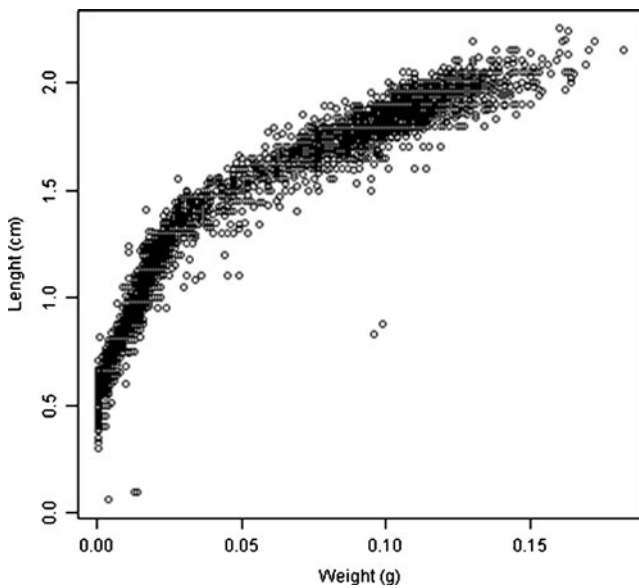**Fig. 1** Pairplot of the response variable *Length* against each explanatory variable
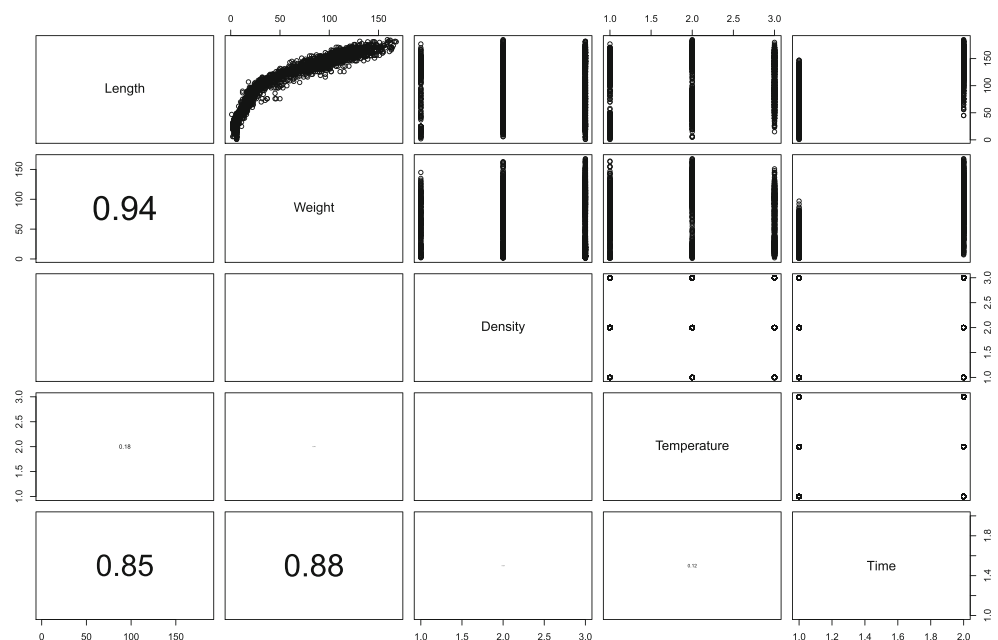
Fig. 2 Scatterplot of Length (cm) against Weight (g)

by this combination of predictors ($F_{7,3875}=3389$, $p<2.2e-16$).

All interaction terms and single variables were highly significant, which means that the null hypothesis which states that there is no interaction between Density, Temperature and Time can be rejected (Table 2).

After the model selection process, the model validation followed, in which the underlying assumptions of linear regression had to be verified. To test for normality, a QQ-plot can be used, in which the points should lay on a straight line. Only in the middle range are the data normally distributed. However, as this assumption is the least important one, some discrepancy can be accepted (Fig. 5). To verify homogeneity, the residuals were plotted against the fitted values (Fig. 6). Because the spread of the residuals is not the same along the gradient, a violation of homogeneity occurred. To detect the source of this heterogeneity, the residuals were plotted against each explanatory variable. The graphs of the residuals versus Density and Temperature showed differences in spread which was the cause of the variation in the residuals. To test for independence, the residuals were plotted against Weight (Fig. 7). Although Weight is not applied in the model, it can be used to check for a possible lack of independence. The graph shows a clear pattern in which all samples with very low and very high weight values had negative residuals, which means that the samples were overfitted since the residuals are calculated as observed minus fitted values. In contrast, samples with weight values lying in between had positive residuals, indicating that all of these samples were under-fitted.

The notation in Eq. 1 means that Length is modelled as a function of Density, Temperature, Time and their two- and three-way interaction terms. The variable $Length_{ij}$ is the length of individual i in batch j. The index j runs from 1 to 5, whereas the number of individuals per batch differs. The term ε_{ij} is the unexplained noise and is assumed to be normally distributed with a mean of 0 and variance of σ^2.

The R^2 value for the model was 0.859, indicating that about 86% of the variation in length can be explained

Fig. 3 Coplot for Length against Weight, conditional on Density and Temperature (Density: 1=10 individuals, 2=50 individuals, 3=100 individuals; Temperature: 1=5°C, 2=20°C, 3=30°C)
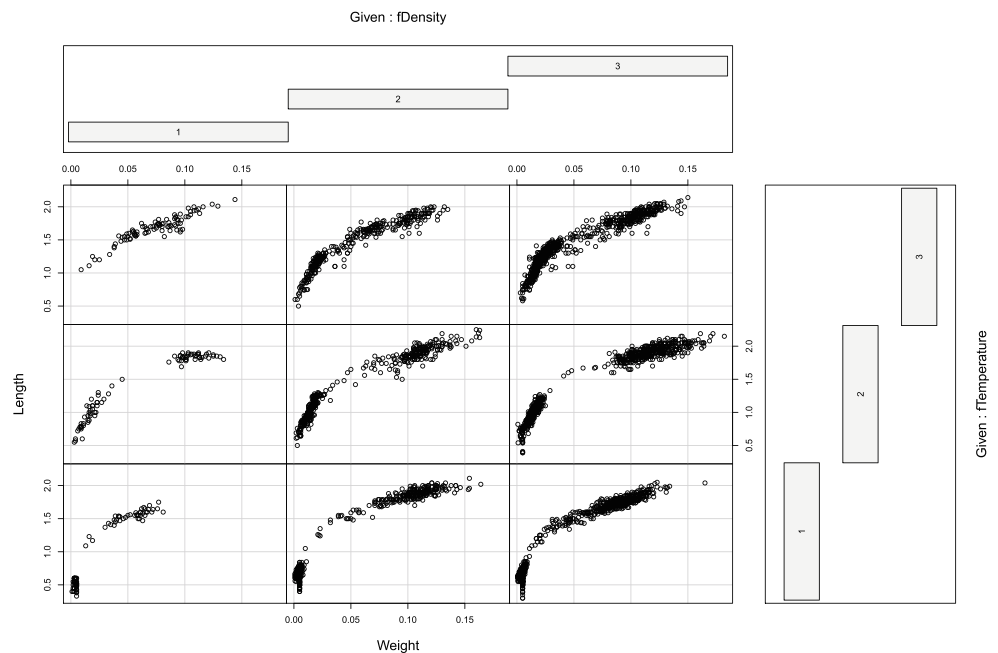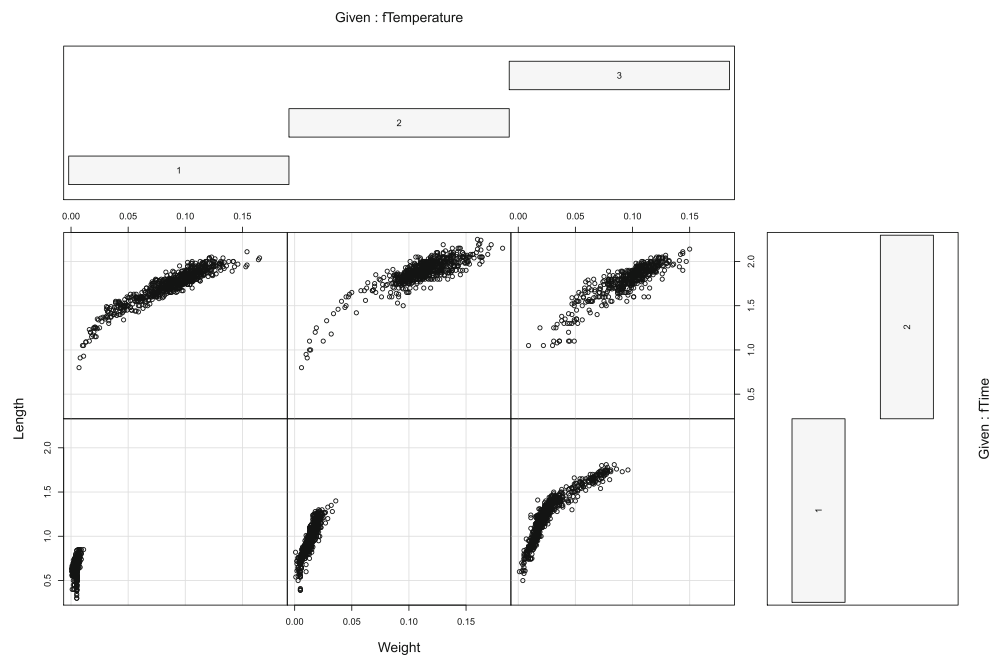
Fig. 4 Coplot for *Length* against *Weight*, conditional on *Temperature* and *Time* (*Temperature*: 1=5°C, 2=20°C, 3=30°C; *Time*: 1=40 ADD, 2=120 ADD)

The overall conclusion is that linear regression is not the right analysis for this set of data due to heterogeneity and a lack of independence. Although it provided significant values and the R^2 was about 86%, the main underlying assumptions were not verified. This caused the regression parameters to be unreliable, and they therefore should not be used in an expert opinion as the description of the developmental rate as a linear function of the ambient temperature may lead to an incorrect age determination of the specimen and therefore to an erroneous calculation of the minimum post-mortem interval.

The next step for improving the model was fitting a generalised additive model which allows for non-linear relationships.

Table 2 Numerical output from the linear regression model containing three explanatory variables and their two- and three-way interaction terms

Coefficients	Estimate β	Std. error	*t*-value	*p*-value
(Intercept)	−1.90811	0.10348	−18.440	< 2e−16
Density	0.34135	0.03960	8.621	< 2e−16
Temperature	1.13293	0.04762	23.791	< 2e−16
Time	1.86172	0.06550	28.425	< 2e−16
Density: temperature	−0.22379	0.01816	−12.323	< 2e−16
Density: time	−0.18735	0.02498	−7.500	7.84e−14
Temperature: time	−0.57889	0.03076	−18.817	< 2e−16
Density: temperature: time	0.12381	0.01169	10.595	< 2e−16

Generalised additive model

Due to the high collinearity between *Weight* and *Time*, the categorical variable *Time* was replaced by the continuous variable *Weight*. A smoothing function of *Weight* has been used to model its effects. Therefore, the model for the generalised additive model contained the variables *Density* and *Temperature* (and their interaction term) which were
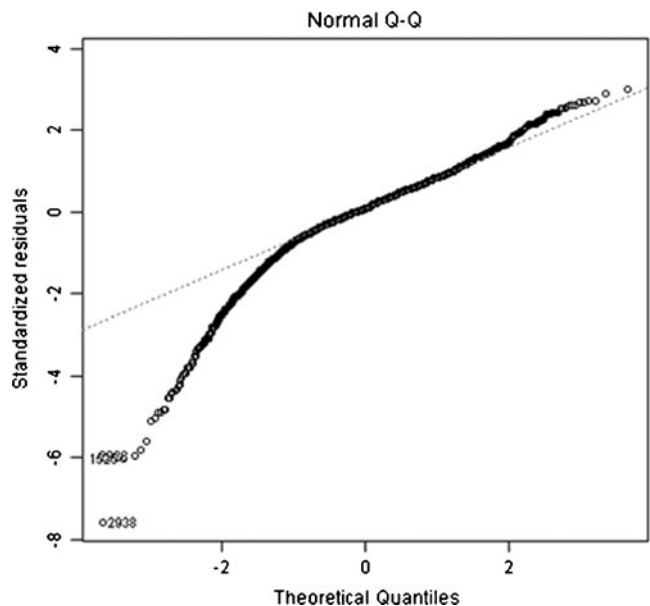

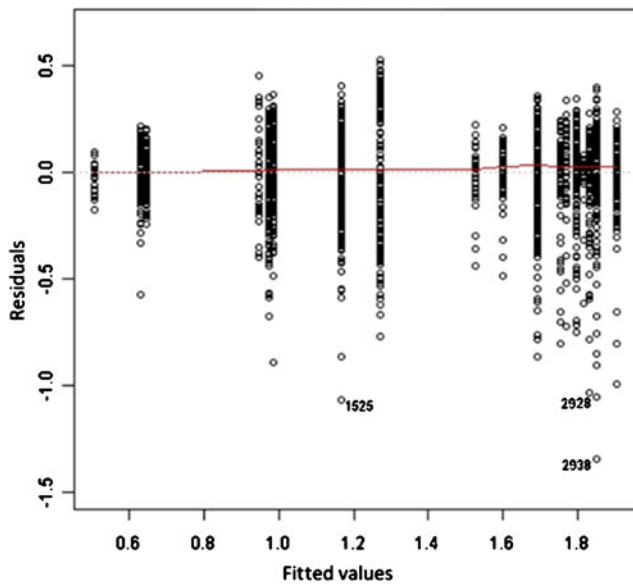**Fig. 5** Normal QQ-plot checking for normality

Fig. 6 Residuals against fitted values. Violation of homogeneity occurred because the spread is increasing with increasing fitted values

used as factors and a smoothing function of *Weight*. It had the form:

$$\text{Length}_{ij} = \alpha + \beta1 \times \text{Density}_i + \beta2 \times \text{Temperature}_i$$
$$+ \beta3 \times \text{Density}_i \times \text{Temperature}_i + f\left(\text{Weight}_{ij}\right) + \varepsilon_{ij}. \tag{2}$$

The aforementioned notation is almost like the linear regression model with the exeption of the smoothing function of *Weight*. The term ε_{ij} is assumed to be normally distributed with a mean of 0 and variance of σ^2.
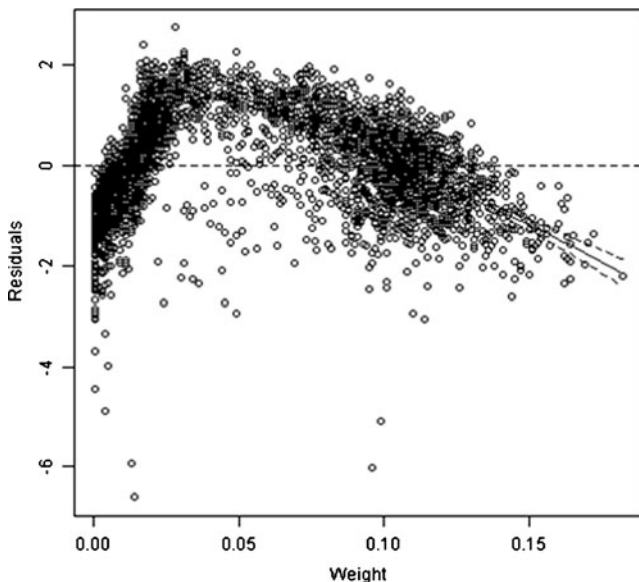

Fig. 7 Residuals against *Weight*. Note the lack of independence due to the pattern of positive and negative residuals

Table 3 Numerical output from the final generalised additive model containing the numerical variable Temperature

Coefficients	Estimate β	Std. error	*t*-value	*p*-value
(Intercept)	1.354169	0.003720	364.024	< 2e−16
Temperature	0.009809	0.001781	5.508	3.87e−08

During the model selection process, the *Density/Temperature* interaction term was the least significant and could be dropped out. After refitting the model without the interaction, *Density* was not significant and dropped out too. The final model included the nominal variable *Temperature* and a smoothing function of *Weight* (Tables 3 and 4).

The equivalent statistic to the linear R^2 value in GAMs is the explanation for deviance. For the final model, this was 97.6%, indicating a very good model fit (GCV=0.0059136). The optimal number of degrees of freedom for the smoother was 8.51, affirming the non-linear relationship between *Length* and *Weight* ($p < 2e-16$).

For the final model, the null hypothesis which states that there is no effect for Temperature and Weight can be rejected.

The model validation process is approximately the same as in a linear regression analysis. To check the normality, the distribution of the residuals is illustrated in a QQ-plot (Fig. 8). The plots showed a little discrepancy, but since normality is not the most important assumption to be verified, the suggested normal (Gaussian) distribution could be accepted. The plot of the residuals versus the fitted values showed no bias in residuals based on *Length*, which means that the assumption of homogeneity could be verified. To test for independence, the residuals were plotted against all explanatory variables. Their distribution for the variable *Weight* showed no dependent structures (Fig. 9).

However, a model misspecification could be highlighted in the graphs for the residuals versus *Temperature*. They showed heterogeneity and a clear violation of independence. The samples showed more negative residuals than positive ones. Additionally, the imbalance increased with increasing *Temperature* level, indicating a dependent structure (Fig. 10). The dependent structures in the model were due to multiple observations within a batch, which means that the five length values within one approach were more related to each other than to values from another approach. As a consequence, the standard errors and variances of the estimated parameters were inadequate, and therefore, they should not be used in an expert opinion either.

Table 4 Numerical output for the smoothing function from the final generalised additive model

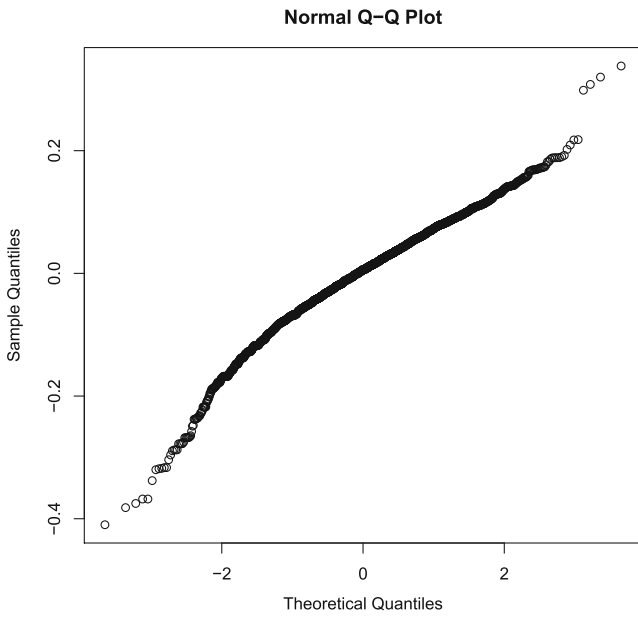| Smooth term	edf	Ref. df	*F*-value	*p*-value
s (Weight)	8.512	8.907	16,740	<2e−16

Normal Q–Q Plot

Fig. 8 Checking for normality with a QQ-plot

Because all linear and additive models assume independence within the observations, a generalised additive model is an improvement in terms of the smoother for *Weight*, although it is still not the right model. To allow for a hierarchical data set, the next step was applying a generalised additive mixed model.

Generalised additive mixed model

The residual spread in the generalised additive model differed with *Temperature*, which means that the assumption of normally distributed residuals with a mean of 0 and variance of σ^2
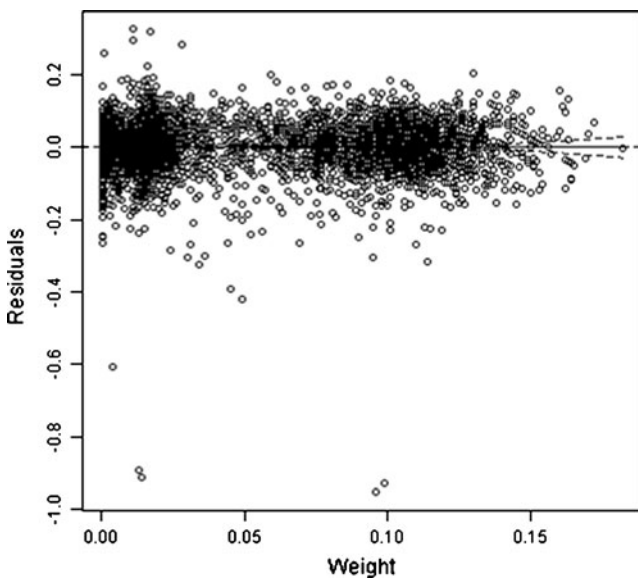

Fig. 9 Residuals versus *Weight* to test for independence
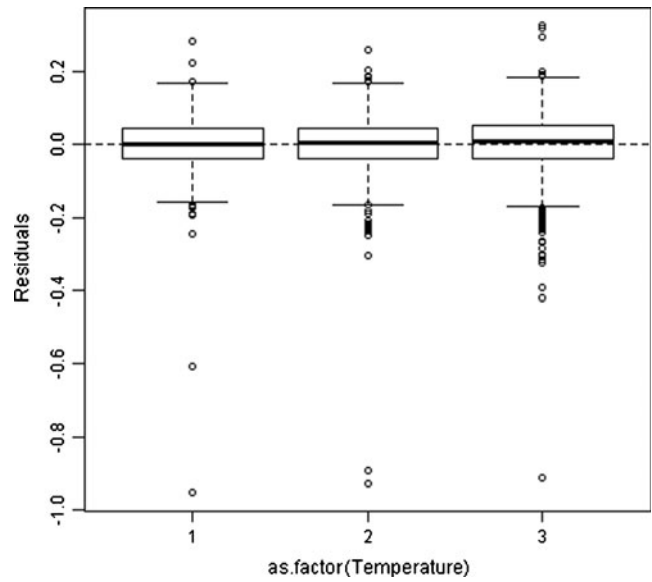

Fig. 10 Residuals versus *Temperature*. Note that the spread of the residuals is not the same along the gradient

is wrong. One solution for dealing with heterogeneity is including a variance structure in the residuals. Since *Temperature* is included as a factor in the fixed part of the model, it was possible to use the VarIdent variance structure for *Temperature* as a variance covariate. The first model had the form:

$$\text{Length}_{ij} = \alpha + \beta 1 \times \text{Temperature}_i + f\left(\text{Weight}_{ij}\right) + \varepsilon_{ij}$$
$$\text{where } \varepsilon_{ij} \sim N(0, \sigma^2{}_a).$$

(3)

To take the dependent structures into account, a second model with the *BatchID* as the random part was fitted. Both models were compared using the AIC. The second model had the form:

$$\text{Length}_{ij} = \alpha + \beta 1 \times \text{Temperature}_i + f\left(\text{Weight}_{ij}\right) + a_i + \varepsilon_{ij}$$
$$\text{where } a_i \sim N(0, \sigma^2{}_a) \text{ and } \varepsilon_{ij} \sim N\left(0, \sigma_j{}^2\right).$$

(4)

The aforementioned notations are almost the same as the generalised additive model with the exemption of the random intercept a_j which allows for variation between batches.

The model, when using a random intercept (4), has the lowest AIC and is therefore selected as the optimal model (model (4): −9518.457 compared to model (3): −8837.617).

Table 5 Estimated parameters from the final mixed effect model

	Estimate	Std. error	*t*-value	*p*-value
(Intercept)	1.316149	0.016696	78.830	<2e−16
Temperature	0.023695	0.007774	3.048	0.00232

Table 6 Numerical output for the smoothing function from the final generalised additive mixed model

Smooth term	edf	Ref. df	*F*-value	*p*-value
s (Weight)	8.751	8.751	2,281	<2e−16

The R^2 value for the model was 0.975, indicating that about 98% of the variation in *Length* can be explained by this combination of predictors ($F_{3,3791}=15995.52$, $p< 0.0001$). The optimal number of degrees of freedom for the smoother was 8.75, affirming the non-linear relationship between *Length* and *Weight* ($p <2e−16$).

All variables were statistically significant, which means that none of these terms can be left out (Tables 5 and 6).

One part of the output from a generalised additive mixed model shows information about the different multiplication factors of σ, which means that each *Temperature* treatment has different variances. The estimated value for σ is 0.068. The treatment at 30°C has the largest variance, namely, $(0.068 \times 1.022)^2$, whereas individuals reared at 20°C have the slightest residual spread $(0.068 \times 0.96)^2$. The residual spread at 5°C lies in between (0.068×1^2).

The model validation process is the same as in a linear regression analysis. The plot of the residuals versus the fitted values showed no extreme bias in residuals based on *Length*, which means that the homogeneity assumption could be verified (Fig. 11). To test for independence, the residuals were plotted against all explanatory variables. For the variable *Weight*, the residuals showed no dependent structures (Fig. 12). In contrast, the graphs for the residuals versus *Temperature* still showed patterns with more negative than positive residuals, indicating that a violation of independence
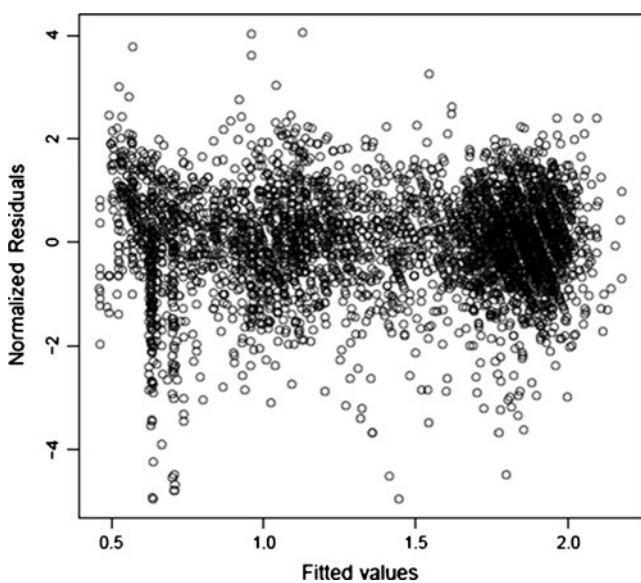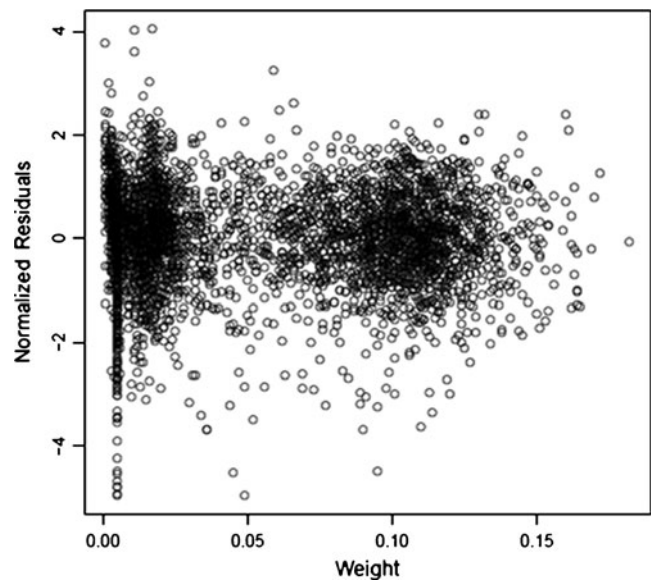

Fig. 12 Normalised residuals versus *Weight* to test for independence

occurred (Fig. 13). As the residuals in the model validation are not ordinary but standardised residuals, they should not show any pattern. However, because their distribution is not that rigorous and all terms in the model are highly significant, the results are quite robust. According to this, a generalised additive mixed model is by now the best method for that kind of data set because it allows for non-linear relationships and correlation between the batches.

Conclusion and outlook

The common methods for the determination of age in immature blow flies include measurements of the length and
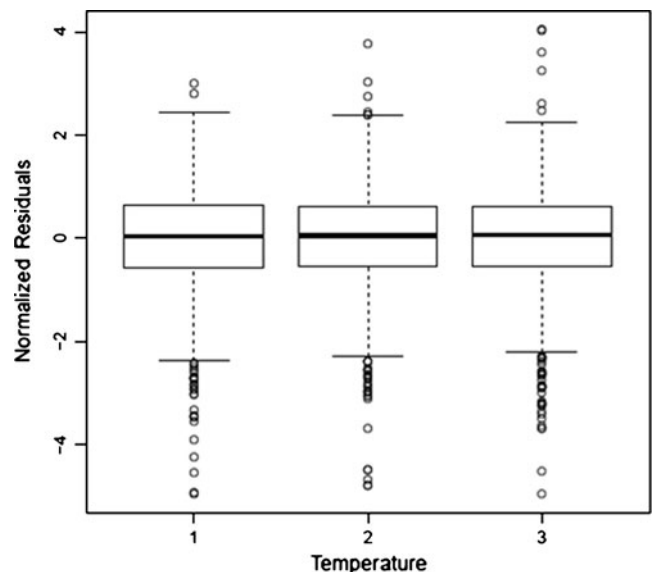

Fig. 11 Normalised residuals versus fitted values to test for homogeneity

Fig. 13 Normalised Residuals of the final model, including a variance structure. Residuals are grouped per *Temperature*

the principle of the accumulated degree hours or days (ADH, ADD). The basis of these methods for the determination of individual age is a continuous assumed linear relationship between ambient temperature and development. Since the growth rate of blow flies is only linear during a small period of development and within a certain temperature range [3, 14, 24], the applied linear models do not supply a sufficient reliability. In linear models, the variability of a series of individual observations (y_i) is explained by a series of independent variables ($x1, x2,…,xn$) and a source of error (ε). One of the greatest limitations of linear models is the underlying assumption of independent observations and homoscedasticity [29]. This means that observations within an experimental approach or vessel have to be independent from each other. If the observations show dependent structures, the required replicates are only pseudoreplicates. Therefore, the use of linear models with a single region of variability (ε) becomes inappropriate.

The error values or the variance of the residuals are negatively influenced and finally lead to incorrect inferences [6, 9]. In addition, completely independent approaches are hardly ensured in biological work because individuals that are reared in the same incubator but in separate vessels cannot be classified as independent [11].

The objective of this study was to present various statistical methods applied to a forensic entomological data set. The differences of these methods concerning the validation of the final model were demonstrated, and a method was found that reflects the data set most adequately.

After the exploration of the data, a step-by-step analysis from a linear regression to a generalised additive model was conducted, finalised by a generalised additive mixed model.

Both the linear regression model and the generalised additive model provided incorrect regression parameters due to existing dependence and correlation structures between the observations. Only with a mixed effects model regression parameters can be calculated, which reflect the data adequately.

Since laboratory experiments are used for the establishment of reference data for forensic entomological casework, they are usually carried out with multiple repetitions. These data sets are automatically nested and imply the use of a mixed model. This is the only way to provide valid regression parameters that lead to statistical valid confidence intervals for the determination of age in insects and therefore to an accurate determination of the minimum time since death.

We used here a simple setting for reasons of clarity, but our model is certainly useful for producing relevant development data for forensically relevant fly species. Amendt et al. [2] compiled development data published so far for some forensically important blow fly species. Besides a certain amount of consensus, it was shown that there is discrepancy for one and the same species between different authors. This might be related to possible geographic variability and population specific features, but could be as well the result of wrong methods and erroneous statistics. However, due to a lack of transparency, this is not always verifiable and could lead to difficult situations in court when using such publications as a reference. Hence, we believe that there is not just the need for a detailed description of the experimental settings when publishing development data, but also for an appropriate exploration of the data and the validation of the statistics applied in such a study. In a next step toward an accepted forensic entomological tool, some could allow for making the published data sets available (and therefore checkable) through an online repository. While this reads desirable at a first glance and might contribute to a traceable use of entomological data sets, there is the question how to organise and control such platforms and repositories. This could be the task of organisations like the European Association for Forensic Entomology (EAFE) or the North American Forensic Entomology Association (NAFEA).

References

1. Adams ZJO, Hall MJR (2003) Methods used for the killing and preservation of blow fly larvae, and their effect on post-mortem larval length. Forensic Sci Int 138:50–61

2. Amendt J, Campobasso CP, Gaudry E, Reiter C, LeBlanc HN, Hall MJR (2007) Best practice in forensic entomology—standards and guidelines. Int J Leg Med 121:90–104

3. Amendt J, Goff ML, Grassberger M, Campobasso CP (2010) Current concepts in forensic entomology. Novel arthropods, environments and geographical regions. Springer, New York

4. Ames C, Turner B (2003) Low temperature episodes in development of blow flies: implications for postmortem interval estimation. Med Vet Entomol 17:178–186

5. Anderson GS (2000) Minimum and maximum development rates of some forensically important Calliphoridae (Diptera). J Forensic Sci 45:824–832

6. Chaves LF (2010) An entomologist guide to demistify pseudoreplication: data analysis of field studies with design constraints. J Med Entomol 47:291–298

7. Cleveland WS (1993) Visualizing Data. Hobart Press, Summit, New Jersey, U.S.A.

8. Donovan SE, Hall MJR, Turner BD, Moncrieff CB (2006) Larval growth rates of the blow fly, *Calliphora vicina*, over a range of temperatures. Med Vet Entomol 20:106–114

9. Faraway JJ (2006) Extending the linear model with R. Generalized linear, mixed effects and nonparametric regression models. Chapman & Hall/CRC, Boca Raton, Florida

10. Greenberg B (1991) Flies as forensic indicators. J Med Entomol 28:565–577

11. Heffner RA, Butler MJ IV, Reilly CK (1996) Pseudoreplication revisited. Ecology 77:2558–2562
12. Higley LG, Haskell NH (2001) Insect development and forensic entomology. In: Byrd JJ, Castner JL (eds) The utility of arthropods in legal investigations. CRC Press, Boca Raton, Florida, pp 389–406
13. Ieno EN, Amendt J, Fremdt H, Saveliev AA, Zuur AF (2010) Analysing forensic entomological data using additive mixed effects modelling. In: Amendt J, Goff ML, Grassberger M, Campobasso CP (eds) Current concepts in forensic entomology. Novel arthropods, environments and geographical regions. Springer, New York, pp 139–162
14. Liu SS, Zhang GM, Zhu J (1995) Influence of temperature variations on rate of development in insects: analysis of case studies from entomological literature. Ann Entomol Soc Am 88:107–119
15. Marchenko MI (2001) Medicolegal relevance of cadaver entomofauna for the determination of the time of death. Forensic Sci Int 120:89–109
16. Michaud JP, Moreau G (2011) A statistical approach based on accumulated degree-days to predict decomposition-related processes in forensic studies. J Forensic Sci 56:229–232
17. Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS. Springer, New York
18. Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, UK
19. R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
20. Richards CS, PatersonID VMH (2008) Estimating the age of immature *Chrysomya albiceps* (Diptera: Calliphoridae), correcting for temperature and geographical latitude. Int J Legal Med 122:271–279
21. Richards CS, Villet MH (2008) Factors affecting accuracy and precision of thermal summation models of insect development used to estimate post-mortem intervals. Int J Legal Med 122:401–408
22. Richards CS, Crous KL, Villet MH (2009) Models of development for blow fly sister species *Chrysomya chloropyga* and *Chrysomya putoria*. Med Vet Entomol 23:56–61
23. Tantawi TI, Greenberg B (1993) The effect of killing and preservative solutions on estimates of maggot age in forensic cases. J Forensic Sci 38:702–707
24. Tarone AM, Foran DR (2008) Generalized additive models and *Lucilia sericata* growth: assessing confidence intervals and error rates in forensic entomology. J Forensic Sci 53:942–948
25. Villet MH (2007) An inexpensive geometrical micrometer for measuring small, live insects quickly without harming them. Entomol Exp Appl 122:279–280
26. Wells JD, LaMotte LR (1995) Estimating maggot age from weight using inverse prediction. J Forensic Sci 40:585–590
27. Zuur AF, Ieno EN, Smith GM (2007) Analysing ecological data. Springer, New York, London
28. Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer, New York
29. Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14