University of Zurich[UZH]

Measurement error and uncertainty in data: a fascinating statistical challenge

Stefanie Muff

Statistics Seminar, NTNU Trondheim

23. October 2017

Research interests

- Measurement error modeling (methods & applications) [1]
- Bayesian statistics (ideal for taming measurement error!)
- Population biology / quantitative genetics [2]
- Movement ecology [3]
- The proper handling of statistical methods (p-values, model selection,..)

[1] *e.g.* Muff et al. (2015); Muff and Keller (2015); Muff et al. (2017a,b)
[2] Ponzi et al. (in prep)
[3] e.g. Weinberger et al. (2016), Gehr et al. (2017) or Muff et al. (in prep).

Research interests

- **Measurement error modeling** (methods & applications) [1]
- **Bayesian statistics** (ideal for taming measurement error!)
- Population biology / quantitative genetics [2]
- Movement ecology [3]
- The proper handling of statistical methods (p-values, model selection,..)

[1] *e.g.* Muff et al. (2015); Muff and Keller (2015); Muff et al. (2017a,b)
[2] Ponzi et al. (in prep)
[3] e.g. Weinberger et al. (2016), Gehr et al. (2017) or Muff et al. (in prep).

Why is “measurement error” an exciting research topic?

- Ubiquituous
- Cross-disciplinary
- Often neglected/ignored (also in introductory textbooks)
- Consequences of error are often unknown
- Many open questions
- Challenging methodology

$\rightarrow$ Ideal “playground” for a statistician....

https://quotefancy.com/

Sources of measurement measurement error (ME)

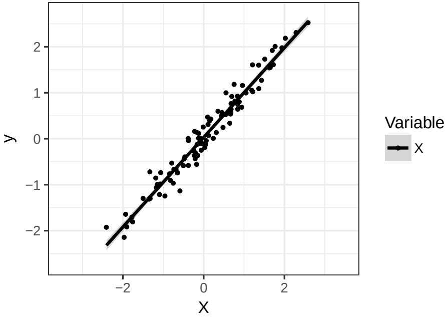
- Measurement **imprecision**.
- **Incomplete** or **biased observations**.
- **Preferential sampling**.
- **Misalignment** error in spatial models.
- **Misclassification** error.
- ...

In addition, missing data is a special and extreme case of ME.

Short preamble on measurement error in regression models

Find regression parameters β_0 and β_x for the model with covariate $\mathbf{x}$:

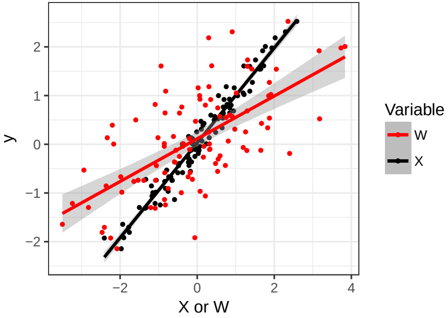
$$y_i = \beta_0 + \beta_x \cdot x_i + \epsilon_i , \quad \epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2) \ .$$



Short preamble on measurement error in regression models II

However, assume that only an erroneous proxy **w** is observed with

$$w_i = x_i + u_i \qquad u_i \sim \mathrm{N}(0, \sigma_u^2) \quad \text{with} \quad \sigma_u^2 = \sigma_x^2 \ .$$



Short preamble on measurement error in regression models III

Now assume that the erroneous proxy **w** is given as

$$x_i = w_i + u_i \qquad u_i \sim \mathrm{N}(0, \sigma_u^2) \quad \text{with} \quad \sigma_u^2 = \sigma_x^2 \ .$$

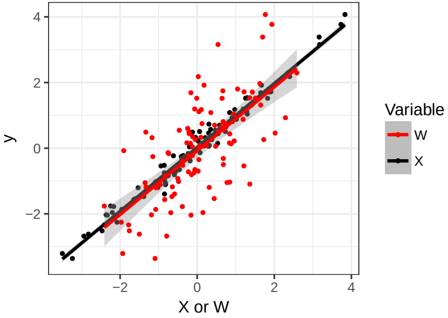
(I have only flipped x_i and w_i!)

Short preamble on measurement error in regression models III

Now assume that the erroneous proxy **w** is given as

$$x_i = w_i + u_i \qquad u_i \sim \mathrm{N}(0, \sigma_u^2) \quad \text{with} \quad \sigma_u^2 = \sigma_x^2 \ .$$

(I have only flipped x_i and w_i!)

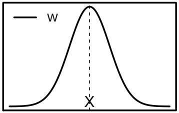


Two fundamentally different error types

- **Classical measurement error:** the "mismeasurement" type:

  Example: uncertainty in measuring tarsus length

$$\mathbf{w} = \mathbf{x} + \mathbf{u}$$
$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \sigma_u^2 \mathbf{D})$$

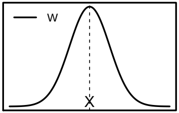


Two fundamentally different error types

- **Classical measurement error:** the "mismeasurement" type:

  Example: uncertainty in measuring tarsus length

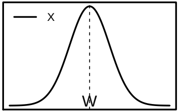
$$\mathbf{w} = \mathbf{x} + \mathbf{u}$$
$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \sigma_u^2 \mathbf{D})$$



- **Berkson measurement error**: The "rounding" type.

  Examples: Experiments; limited resolution of a measurement device

$$\mathbf{x} = \mathbf{w} + \mathbf{u}$$
$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \sigma_u^2 \mathbf{D})$$



Possible bias induced by ME

- **Attenuation** (bias towards the null)

  $\rightarrow$ Underestimated regression coefficients

  $\rightarrow$ Conservative estimates

- **No bias**

  $\rightarrow$ But more uncertainty...

- **Reverse attenuation** (bias away from null)

  $\rightarrow$ **Over**estimated regression coefficients

  $\rightarrow$ **Anti**conservative estimates

Simulations and apps

Illustration with shiny apps for two error types in linear, logistic and Poisson regression:

- Classical error
- Berkson error

Correcting for the error: Error modeling

The **two most popular approaches**:

- **Bayesian methods**: Prior information about the error enters a model.

  Then use

  $$\text{posterior} = \text{likelihood} \times \text{prior}$$

  to calculate the parameter distribution after error correction (with MCMC or INLA).

Correcting for the error: Error modeling

The **two most popular approaches**:

- **Bayesian methods**: Prior information about the error enters a model.

  Then use

  $$\text{posterior} = \text{likelihood} \times \text{prior}$$

  to calculate the parameter distribution after error correction (with MCMC or INLA).

- **SIMEX**: SIMulation EXtrapolation, a heuristic and intuitive idea.

Correcting for the error: Error modeling

The **two most popular approaches**:

- **Bayesian methods**: Prior information about the error enters a model.

  Then use

  $$\text{posterior} = \text{likelihood} \times \text{prior}$$

  to calculate the parameter distribution after error correction (with MCMC or INLA).

- **SIMEX**: SIMulation EXtrapolation, a heuristic and intuitive idea.

Prerequisite for error modeling:

Assessing the bias and modeling the error is **only possible if the error structure** (model) **and the respective model parameters** (e.g., error variances) **are known!**

(It is sometimes better **not** to model the error..)

Why Bayesian ME modeling?

1. **Simple and general:**
   The formulation of Bayesian error models is usually straightforward (hierarchical modeling).

2. **Identifiability issues:**
   Most models with error components are nonidentifiable, e.g.:

$$w_i = x_i + u_i \quad \text{with} \quad \sigma_w^2 = \sigma_x^2 + \sigma_u^2 \ .$$

   The error variance σ_u^2 and the sampling variance σ_x^2 are *confounded*.

   $\rightarrow$ The "Bayesian crank" can be turned even if a model is nonidentifiable.

   $\rightarrow$ All you need is a legitimate prior distribution.

   $\rightarrow$ "Partially identified models" (Gustafson, 2005).

Hierarchical Bayesian error models

Hierarchical Bayesian modeling is truly universal. Regression model with response $\mathbf{y}$ and covariates $\mathbf{x}$ and $\mathbf{z}$ and inverse link function $h()$.
Classical error in the covariate $\mathbf{x}$ can be modeled as

$$
\begin{aligned}
\mathrm{E}(\mathbf{y} \mid \mathbf{x}) &= h(\beta_0 + \beta_x \mathbf{x} + \mathbf{z}\boldsymbol{\beta}_z) \ , & & \\
\mathbf{w} &= \mathbf{x} + \mathbf{u} \ , & \mathbf{u} &\sim \mathrm{N}(0, \tau_u \mathbf{D}_u) \ , \\
\mathbf{x} &= \alpha_0 + \mathbf{z}\boldsymbol{\alpha}_z + \boldsymbol{\varepsilon}_x \ , & \boldsymbol{\varepsilon}_x &\sim \mathrm{N}(0, \tau_x \mathbf{D}_x) \ .
\end{aligned}
$$

Hierarchical Bayesian error models

Hierarchical Bayesian modeling is truly universal. Regression model with response $\mathbf{y}$ and covariates $\mathbf{x}$ and $\mathbf{z}$ and inverse link function $h()$.
Classical error in the covariate $\mathbf{x}$ can be modeled as

$$
\begin{aligned}
\mathrm{E}(\mathbf{y} \mid \mathbf{x}) &= h(\beta_0 + \beta_x \mathbf{x} + \mathbf{z}\boldsymbol{\beta}_z) \ , && \\
\mathbf{w} &= \mathbf{x} + \mathbf{u} \ , && \mathbf{u} \sim \mathrm{N}(0, \tau_u \mathbf{D}_u) \ , \\
\mathbf{x} &= \alpha_0 + \mathbf{z}\boldsymbol{\alpha}_z + \boldsymbol{\varepsilon}_x \ , && \boldsymbol{\varepsilon}_x \sim \mathrm{N}(0, \tau_x \mathbf{D}_x) \ .
\end{aligned}
$$

Berkson error is modeled as

$$
\begin{aligned}
\mathrm{E}(\mathbf{y} \mid \mathbf{x}) &= h(\beta_0 + \beta_x \mathbf{x} + \mathbf{z}\boldsymbol{\beta}_z) \ , && \\
\mathbf{x} &= \mathbf{w} + \mathbf{u} \ , && \mathbf{u} \sim \mathrm{N}(0, \tau_u \mathbf{D}_u) \ .
\end{aligned}
$$

$\rightarrow$ Can be fitted as a joint model with MCMC (since the 1990's) or INLA (since 2015, see Muff et al., 2015, jointly with A. Riebler, L. Held, H. Rue).

Hierarchical Bayesian models with INLA

INLA is able to deal with latent Gaussian hierarchical models.

Three sub-models (here for classical ME):

- Observation model
  - **Regression model**: $\mathrm{p}(\mathbf{y}|\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}_1)$

    $$\mathrm{E}(\mathbf{y}) = h(\beta_0 + \beta_x \mathbf{x} + \mathbf{z}_{[i,]}\boldsymbol{\beta}_z)$$

  - **Error model**: $\mathrm{p}(\mathbf{w}|\mathbf{x}, \boldsymbol{\theta}_2)$

    $$\mathbf{w} = \mathbf{x} + \mathbf{u} \ , \qquad \mathbf{u} \sim \mathrm{N}(\mathbf{0}, \tau_u \mathbf{D}_u)$$

- Latent model for $\mathbf{v} = (\beta_0, \boldsymbol{\beta}_z^\top, \alpha_0, \boldsymbol{\alpha}_z^\top, \mathbf{x}^\top)^\top$
  - **Exposure model** for $\mathbf{x}$: $\mathrm{p}(\mathbf{x}|\boldsymbol{\theta}_2)$

    $$\mathbf{x} = \alpha_0 + \mathbf{z}\boldsymbol{\alpha}_z + \boldsymbol{\varepsilon}_x, \quad \boldsymbol{\varepsilon}_x \sim \mathrm{N}(\mathbf{0}, \tau_x \mathbf{D}_x)$$

  - Independent Gaussian priors for $(\beta_0, \boldsymbol{\beta}_z^\top, \alpha_0, \boldsymbol{\alpha}_z^\top)$

- Hyperpriors $\mathrm{p}(\boldsymbol{\theta}_1)$, $\mathrm{p}(\boldsymbol{\theta}_2)$ with $\boldsymbol{\theta}_2 = (\beta_x, \tau_u, \tau_x)^\top$

Example 1: Two error mechanisms in a single variable

(Swiss National Cohort study on cardiovascular disease mortality [4])

Goal: To find factors that influence the risk of cardiovascular disease mortality.

Model: Weibull survival model

$$\begin{aligned} \eta_i &= \beta_0 + x_i \beta_x + \mathbf{z}_i^\top \boldsymbol{\beta}_z \ , \\ h_i(t) &= \exp(\eta_i)\, \gamma\, t^{\gamma - 1} \ , \end{aligned}$$

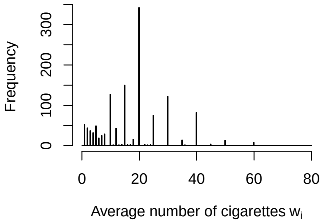
with hazard function $h_i(t)$.

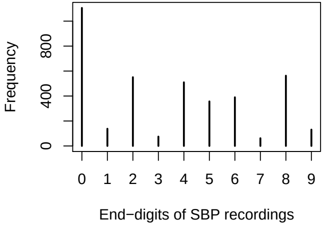
Problem: Measurement error in

- self-reported mean number of cigarettes smoked per day
- systolic blood pressure (SBP)

[4] Von Gunten et al. (2013)

- Distributions of self-reported cigarette numbers and end-digits of SBP measurements [5]:





$\rightarrow$ Rounding behaviour $\rightarrow$ Berkson error

- In addition, there is a component of misremembering (cigarettes) and mismeasurement (SBP) $\rightarrow$ classical error.

[5] Muff et al. (2017b)

Formulation of a Classical/Berkson error model

The problem:

- The correct variable $\mathbf{x}$ is first mismeasured.
- The mismeasured variable is then rounded.
- Observation $\mathbf{w}$.

Formulation of a Classical/Berkson error model

The problem:

- The correct variable $\mathbf{x}$ is first mismeasured.
- The mismeasured variable is then rounded.
- Observation $\mathbf{w}$.

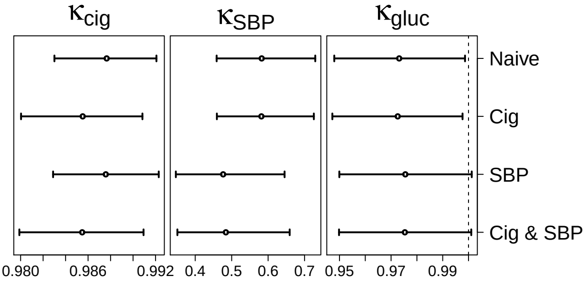
$\rightarrow$ Trick: Introduce an additional latent variable $\mathbf{r}$, such that

$$\begin{aligned} \mathbf{r} &= \mathbf{x} + \mathbf{u}_c \ , && \mathbf{u}_c \sim \mathrm{N}(\mathbf{0}, \tau_{u_c}\mathbf{D}_c) \quad \text{and} \\ \mathbf{r} &= \mathbf{w} + \mathbf{u}_b \ , && \mathbf{u}_b \sim \mathrm{N}(\mathbf{0}, \tau_{u_b}\mathbf{D}_b) \ , \end{aligned}$$

with classical error $\mathbf{u}_c$ and Berkson error $\mathbf{u}_b$.

$\rightarrow$ Combining this model with the survival model led to corrected parameter estimates. We used INLA to fit the model.

Results given in terms of event time ratios. These quantify the proportional change in survival times expected from a change by one unit.



In words:

- The daily consumption of 20 cigarettes shrinks the expected lifetime by a factor of 0.75 and not just 0.78 (without error modeling)
- An increase in blood pressure from 120 to 160 mm Hg shrinks the expected lifetime by a factor of 0.71 and not just 0.78 (without error modeling).

Example 2: Miscounting error in the response of a ZINB regression

COPD: Chronic obstructive pulmonary disease

Exacerbation: A sudden worsening of symptoms that requires treatment with antibiotics, corticosteroids or hospitalization.

Goal: Investigate the effect of a pharmacotherapy vs placebo ($x_i \in \{0, 1\}$) on the number of exacerbations (y_i) of COPD patients [6].

[6] Calverley et al. (2007)

Example 2: Miscounting error in the response of a ZINB regression

COPD: Chronic obstructive pulmonary disease

Exacerbation: A sudden worsening of symptoms that requires treatment with antibiotics, corticosteroids or hospitalization.

Goal: Investigate the effect of a pharmacotherapy vs placebo ($x_i \in \{0, 1\}$) on the number of exacerbations (y_i) of COPD patients [6].

Model: Negative binomial regression

$$y_i \sim \text{NBin}\left(\exp(\log(t_i) + \beta_0 + x_i\beta_x + \mathbf{z}_i\boldsymbol{\beta}_z), \theta\right) \ .$$

Additional covariates $\mathbf{z}_i$, t_i=actual time under treatment (offset).

[6] Calverley et al. (2007)

Example 2: Miscounting error in the response of a ZINB regression

COPD: Chronic obstructive pulmonary disease

Exacerbation: A sudden worsening of symptoms that requires treatment with antibiotics, corticosteroids or hospitalization.

Goal: Investigate the effect of a pharmacotherapy vs placebo ($x_i \in \{0,1\}$) on the number of exacerbations (y_i) of COPD patients [6].

Model: Negative binomial regression

$$y_i \sim \text{NBin}\left(\exp(\log(t_i) + \beta_0 + x_i\beta_x + \mathbf{z}_i\boldsymbol{\beta}_z), \theta\right) \ .$$

Additional covariates $\mathbf{z}_i$, t_i=actual time under treatment (offset).

Problem: Exacerbation numbers y_i are self-reported by the patients, and thus miscounted.

[6] Calverley et al. (2007)

Miscounting error model

- External study [7] investigated the error in the number of self-reported exacerbations .

- Comparison between patient self-reports s_i and consensus classifications by a central adjudication committee, consisting of several experienced physicians ("gold standard", y_i).

- The external validation data were used to estimate the parameters of a zero-inflated negative binomial error model:

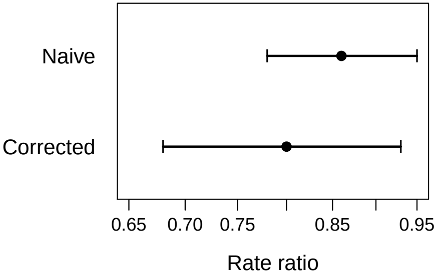
$$s_i \mid y_i \sim \text{ZINB}\left(\gamma_0 + \gamma_1 y_i, p_i, \theta_E\right) \; .$$

with $\text{logit}(p_i) = \delta_0 + \delta_1 \text{I}(y_i > 0)$, where y_i is unobserved.

[7] Frei et al. (2016)

Error-corrected results

The actual treatment effect estimate increased:



Naive rate ratio $\exp(\hat{\beta}_x) = 0.86$ (95% CI from 0.78 to 0.95)
Corrected rate ratio $\exp(\hat{\beta}_x) = 0.80$ (95% CI from 0.68 to 0.93)
(smaller=stronger)

Example 3: Pedigree error in song sparrows

(With Erica Ponzi and Lukas Keller)

Goal: Estimate heritability and inbreeding depression for a wild bird population using pedigree data.

Model: The animal model is a mixed model, in a simple form given as

$$y_i = \mu + \beta_f f_i + a_i + e_i \ ,$$

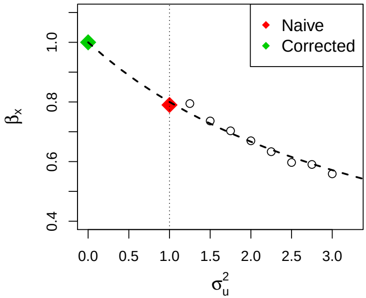
with $(a_1, \ldots, a_n)^\top \sim \mathsf{N}(0, \sigma_a^2 \mathbf{A})$, $e_i \sim \mathsf{N}(0, \sigma_e^2)$, inbreeding depression β_f and $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$.

Problem: The pedigree is known to contain misassigned paternities. This may lead to bias in estimates of heritability, inbreeding depression etc.


www.vogelwarte.ch

SIMEX: A very intuitive idea (Cook and Stefanski, 1994)

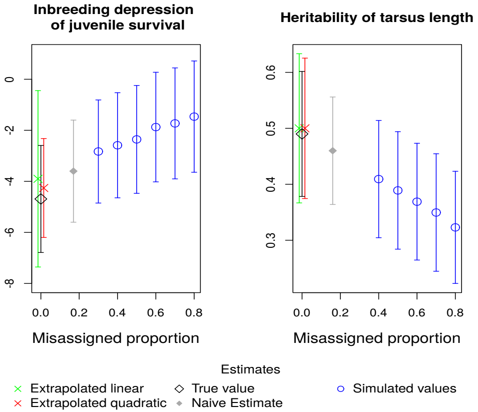
- **Simulation phase:** The error in the data is progressively aggravated.
- **Extrapolation phase:** The observed trend is then extrapolated back to a hypothetical error-free value.



Example:
A regression slope β_x, but x was estimated with error
$w = x + u, u \sim \mathrm{N}(0, \sigma_u^2)$.

Pedigree-SIMEX

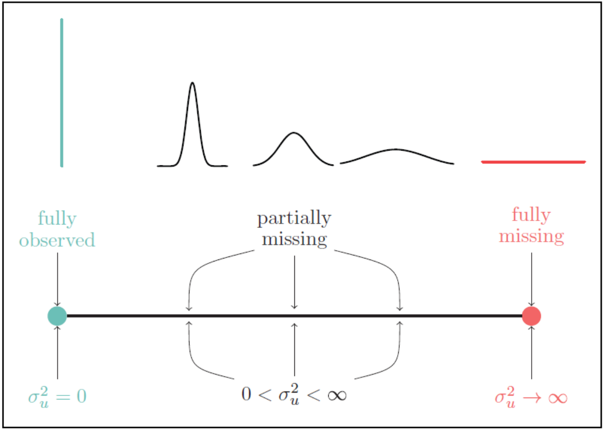
The idea can be transferred to pedigree error by a successive increase of the error proportion (up to 100%) and extrapolation to zero error.



→ PSIMEX package on CRAN (written by Erica Ponzi).

Do you care about missing data?

… then you might want to care about error too: It is a **special case of classical measuremet error**.



(Blackwell et al., 2015)

Modeling missing data and classical ME in a unified framework

Again the hierarchical error model:

$$\eta_i = \beta_0 + \beta_x x_i + \mathbf{z}_i \boldsymbol{\beta}_z \qquad \text{Regression model} \; ,$$

$$w_i = x_i + u_i \qquad \text{Error model} \; ,$$

$$x_i = \alpha_0 + \mathbf{z}_i \alpha_z + (\text{other terms}) + \varepsilon_i \; , \qquad \text{Exposure model} \; .$$

Idea:

- For error variances $\sigma^2_{u_i} \to \infty$ (missing case), the error model is uninformative.
- Information about x_i is retrieved only by the exposure model, similar to *e.g.* Goldstein (2011).

Modeling missing data and classical ME in a unified framework

Again the hierarchical error model:

$$
\begin{aligned}
\eta_i &= \beta_0 + \beta_x x_i + \mathbf{z}_i \boldsymbol{\beta}_z && \text{Regression model ,} \\
w_i &= x_i + u_i && \text{Error model ,} \\
x_i &= \alpha_0 + \mathbf{z}_i \alpha_z + (\text{other terms}) + \varepsilon_i \ , && \text{Exposure model .}
\end{aligned}
$$

Idea:

- For error variances $\sigma^2_{u_i} \to \infty$ (missing case), the error model is uninformative.
- Information about x_i is retrieved only by the exposure model, similar to *e.g.* Goldstein (2011).

Many open points:

- Missings in the outcome? Missing NAR?
- Non-Gaussian data (misclassification)?
- Berkson ME?

THANK YOU!

References:

Blackwell, M., J. Honaker, and G. King (2015). A unified approach to measurement error and missing data: Details and extensions. *Sociological Methods & Research 46*(3), 342–369.

Calverley, P. M., J. A. Anderson, B. Celli, G. T. Ferguson, C. Jenkins, P. W. Jones, J. C. Yates, and J. Vestbo (2007). Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *New England Journal of Medicine 356*, 775–789.

Cook, J. and L. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association 89*, 1314–1328.

Frei, A., L. Siebeling, C. Wolters, L. Held, P. Muggensturm, A. Strassmann, M. Zoller, G. ter Riet, and M. Puhan (2016). The inaccuracy of patient recall for COPD exacerbation rate estimation and its implications: Results from central adjudication. *CHEST 150*(4), 860–868.

Gehr, B., E. Hofer, S. Muff, A. Ryser, E. Vimercati, K. Vogt, and L. F. Keller (2017). Spatial scale and behavioral state interact in shaping temporal dynamics of habitat selection in Eurasian lynx. *Oikos*. In press.

Goldstein, H. (2011). *Multilevel Statistical Models* (4 ed.). Chichester: Wiley.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science 20*, 111–140.

Muff, S. and L. F. Keller (2015). Reverse attenuation in interaction terms due to covariate error. *Biometrical Journal 57*, 1068–1083.

Muff, S., M. Ott, J. Braun, and L. Held (2017b). Bayesian two-component measurement error modelling for survival analysis using INLA – A case study on cardiovascular disease mortality in Switzerland. *Computational Statistics & Data Analysis*. In press.

Muff, S., M. A. Puhan, and L. Held (2017a). Bias away from the null due to miscounted outcomes? *Statistical Methods in Medical Research*. in press, DOI: 10.1177/0962280217694403.

Muff, S., A. Riebler, L. Held, H. Rue, and P. Saner (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 64*, 231–252.

Reid, J. M., L. F. Keller, A. B. Marr, P. Nietlisbach, R. J. Sardell, and P. Arcese (2014). Pedigree error due to extra-pair reproduction substantially biases estimates of inbreeding depression. *Evolution 3*, 802–815.

Weinberger, I. C., S. Muff, A. Kranz, and F. Bontadina (2016). Flexible habitat selection paves the way for a recovery of otter populations in the European Alps. *Biological Conservation 199*, 88–95.