Nearby farms are a source of lice for wild salmonids: a reply to Jansen et al. (2016)

R. M. Serra-Llinares[*], P. A. Bjørn, B. Finstad, R. Nilsen, L. Asplin

*Corresponding author: rosa.maria.serra.llinares@imr.no

This Supplement provides further details regarding the re-analysis of the data from Serra-Llinares et al. (2014) using zero-altered negative-binomial (ZANB) mixed models. The aim of this re-analysis was to model the observed number of attached stages (copepodites and chalimi, hereafter referred to as "attached lice") of salmon lice (*Lepeoptheirus salmonis* Kröyer) on wild salmonids (sea trout *Salmo trutta* L. and Arctic charr *Salvelinus alpinus* L.) using fish length, temperature, infestation pressure from nearby fish farms and year (categorical variable with the levels 2010 and 2011) as explanatory variables.

Methods

Prior to the analysis, the infestation pressure data used in Serra-Llinares et al (2014) was quality-controlled against analogous data currently available. The reason for this is that, at the time these data were gathered, we only had access to biomass and lice reports from fish farms separately; due to the reporting system at the time, some of the information required for merging both data sets (biomass and lice reports) was occasionally lacking, and thus some of the data was at times lost. Currently, we have access to biomass and lice data (included historical data) from fish farms through the Norwegian Marine Data Center at the Institute of Marine Research, where biomass and lice reports from all Norwegian fish farms are collected and merged together, the resulting data being of presumably higher quality than what was previously available. A comparison with analogous data currently available through the Norwegian Marine Data Center showed that our original data set tended to underestimate the amount of in-farm lice in some areas (data not shown), and thus we decided to use the newest data for this re-analysis. No changes were made on the data regarding lice counts on wild salmonids from the original data set.

As described in Serra-Llinares et al. (2014), water temperatures reported by the fish farms on a monthly basis were used in the study. However, the re-analysis of the data set following the approach by Helland et al. (2015) required gathering water temperature data also for reference locations, i.e. those locations without any active farms within 30 km. For those locations, water temperatures at 3 m depth were extracted from the results of the NorKyst800 numerical current model (Albretsen et al. 2011). The NorKyst800 has a grid resolution of 800 m, and comparison of model results with observations exhibits a typical error less than 1°C. Mean monthly temperatures were obtained from the model grid (Albretsen et al. 2011) closest to the sampling sites, for the period May-August in the years 2010 and 2011. For each wild fish sampling occasion, mean water temperature corresponding to the actual sampling month was used. It was not possible to verify the robustness of model predictions for two locations (Handelsbukt in Porsangerfjord and Løksa in Salangen), and thus these locations were excluded from the analyses (137 fish of 2959 fish were removed).

Prior to the analyses, data exploration was applied following the protocol in Zuur et al. (2010). In brief, Cleveland dotplots were used for outlier detection; pair plots, Pearson's correlations and variance inflation factors (VIF) were used to detect collinearity.

As a first step, the number of attached salmon lice observed on wild salmonids was analyzed using a generalized linear mixed model (GLMM) with a Poisson distribution. Fish length (FL), temperature (Temp), infestation pressure from nearby fish farm (Pres) and year (categorical variable with the levels 2010 and 2011) were used as covariates, together with the interactions between each covariate and fish farm pressure. Infestation pressure was calculated as the number of female lice in nearby farms (<30 km), linearly down-weighted according to distance as described in Serra-Llinares et al. (2014). To account for repeated measurements at the same location, location was used as a random effect. The fixed covariates in the Poisson GLMM explained 25% of the variation but the model was over-dispersed. Model validation techniques (comparison of the percentage of zeros from simulated vs. observed data) indicated that the main reason for over-dispersion was zero inflation.

Zero-inflated and zero-altered models have been developed to cope with data sets presenting an excess of zeros (as is often the case with parasitological data). One limitation of standard count models is that the zeros and the non-zeros (positives) are assumed to come from the same data-generating process. With zero-altered models (also called "hurdle models"), these two processes are not constrained to be the same. The basic idea is that a Bernoulli probability governs the binary outcome of whether a count variable has a zero or positive realization. If the realization is positive, the hurdle is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero count data model. With zero-inflated models, the response variable is modeled as a mixture of a Bernoulli distribution and a Poisson distribution (or any other count distribution supported on non-negative integers). Thus, the main difference between the zero-inflated negative binomial (ZINB) GLMM used by Helland et al. (2015) and the ZANB GLMM used for the analysis presented here is the way they interpret and analyze zero counts. For more detail and formulae, see, for example, Gurmu & Trivedi (1996) or Dalrymple et al. (2003).

The advantage of using a zero-altered GLMM is that it can be used to simultaneously investigate the following two questions: What is driving the absence and presence of lice? And when lice are present, what is driving their numbers? Thus, a zero-altered negative binomial (ZANB) GLMM was used for this analysis, with a full model specification as follows:

$$Ch_{ij} \sim ZANB(\mu_{ij}, \pi_{ij}, k)$$

$$E\left(Ch_{ij}\right) = \frac{1-\pi_{ij}}{1-P_0} \times \mu_{ij} \qquad \text{where } P_0 = \left(\frac{k}{\mu_{ij}-k}\right)^k \tag{1}$$

$$\log\left(\mu_{ij}\right) = FL + Temp + Pres + Year + FL \times Pres + Temp \times Pres + Year \times Pres + a_i$$

$$\text{logit}\left(\pi_{ij}\right) = FL + Temp + Pres + Year + FL \times Pres + Temp \times Pres + Year \times Pres + b_i$$

where Ch_{ij} is the number of attached lice for the jth observation at location i (n = 30). The random effects a_i and b_i capture the within-location dependency, and are assumed to be normally distributed. Prior to analysis, all continuous covariates were standardized by subtracting the mean and dividing by the standard deviation, which is common practice when fitting GLMMs.

The ZANB GLMM was fitted in two separate steps (Zuur et al. 2012). First a Bernoulli GLMM was fitted to the absence and presence data. In the second step, the presence-only data (i.e. all the non-zero Ch_{ij} data) were analyzed using a zero-truncated negative binomial GLMM (NB GLMM). Both models were fitted using the glmmADMB package (Skaug et al. 2015) in R (R Development Core Team 2015). Once the 2 separate models were fitted, we combined the 2 components to calculate the fitted values and Pearson residuals for the combined model (ZANB GLMM).

Model validation was performed by plotting the Pearson residuals from the ZANB GLMM vs. fitted values, each covariate in the model, and each covariate not in the model. The percentage of zeros obtained by simulating data using the ZANB GLMM was compared with the percentage of zeros in the original data set. Comparison was also made between the sum of squared Pearson residuals for simulated and original data, and between maximum values of simulated and original data.

RESULTS

Based on data exploration results, 32 fish with questionable length values (based on condition factor) were removed, and fish farm pressure was square-root transformed to deal with its high dispersion. Scatterplots, Pearson's correlation and VIF values did not indicate any collinearity problems, not even between temperature and infestation pressure (VIF <2). Further, modeling fish farm pressure as a function of temperature using a linear mixed effects model (using location as random effect) identified a temperature effect, but the effect size was small, explaining only 8% of the variation.

The ZANB GLMM was fitted and successfully validated. According to the results from the first step (i.e. the Bernoulli GLMM) (Table 1), the fitted model for the years 2010 and 2011 can be written as follows:

$$\text{logit}(\pi_{ij}) = -0.59 + 0.09 \times FL + 0.81 \times Temp + 0.26 \times Pres - 0.18 \times FL \times Pres + 0.30 \times Temp \times Pres \quad (2010)$$
$$\text{logit}(\pi_{ij}) = 0.02 + 0.09 \times FL + 0.81 \times Temp + 0.85 \times Pres - 0.18 \times FL \times Pres + 0.30 \times Temp \times Pres \quad (2011)$$

$$(2)$$

For a graphical representation of these results, the numerical output of the Bernoulli GLMM was used to predict the probability of presence of attached lice on wild salmonids for increasing fish farm pressure values at 3 different temperatures (Fig. 2 in the main article). These results suggest that increasing values of length, temperature and fish farm pressure lead to an increase in the probability of presence of lice larvae on wild salmonids, the fish farm pressure effect being considerably stronger in 2011.

Results from zero-truncated NB GLMM model (Table 2 & Fig. 3 in the main article) suggest a significant effect of both fish farm pressure and temperature also on the actual lice counts on infested wild salmonids, but there is a large variation around the fitted values. In this case, the underlying equations can be specified as follows:

$$\text{logit}(\mu_{ij}) = 2.11 - 0.05 \times FL + 0.52 \times Temp + 1.18 \times Pres + 0.00 \times FL \times Pres + 0.08 \times Temp \times Pres \quad (2010)$$
$$\text{logit}(\mu_{ij}) = 2.23 - 0.05 \times FL + 0.52 \times Temp + 0.51 \times Pres + 0.00 \times FL \times Pres + 0.08 \times Temp \times Pres \quad (2011)$$

$$(3)$$

Note that, in this case, the effect of fish farm pressure is weaker in 2011 compared to 2010.

Model validation tools indicated that the ZANB GLMM produced simulated data sets with similar percentages of zeros as the original data, and the sum of squared Pearson residuals for the simulated data sets are comparable to the sum of squared Pearson residuals for the observed data, indicating that there are no major issues with the model.

LITERATURE CITED

Albretsen J, Sperrevik AK, Staalstrøm A, Sandvik AD, Vikebø F, Asplin L (2011) NorKyst-800 report no. 1: user manual and technical descriptions. Fisken og Havet no. 2/2011

Dalrymple ML, Hudson IL, Ford RPK (2003) Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. Comput Stat Data Anal 41:491–504

Gurmu S, Trivedi PK (1996) Excess zeros in count models for recreational trips. J Bus Econ Stat 14:469–477

Helland IP, Uglem I, Jansen PA, Diserud OH, Bjørn PA, Finstad B (2015) Statistical and ecological challenges of monitoring parasitic sea lice infestations in wild salmonid fish stocks. Aquacult Environ Interact 7:267–280

R Development Core Team (2015) R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna. www.r-project.org

Serra-Llinares RM, Bjørn PA, Finstad B, Nilsen R, Harbitz A, Berg M, Asplin L (2014) Salmon lice infection on wild salmonids in marine protected areas: an evaluation of the Norwegian ´National Salmon Fjords'. Aquacult Environ Interact 5:1–16

Skaug H, Fournier D, Bolker B, Magnusson A, Nielsen A (2015) Generalized linear mixed models using 'AD Model Builder'. R package version 0.8.2. http://r-forge.r-project.org/projects/glmmadmb/

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14

Zuur AF, Saveliev A A, Ieno EN (2012) Zero inflated models and generalized linear mixed models with R. Highland Statistics Limited, Newburgh