Offset or not: guidance on accounting for sampling effort in generalized linear models

James A. Smith[1,2*]

[1] Port Stephens Fisheries Institute, New South Wales Department of Primary Industries, Locked Bag 1, Nelson Bay, NSW 2315, Australia

[2] School of Biological, Earth and Environmental Science, University of New South Wales, Sydney, NSW, 2052, Australia

[*] james.a.smith@dpi.nsw.gov.au

Abstract

1. Observed data are often dependent on a measure of sampling effort, such as counts measured per unit area. A common tool to account for differences in effort is the ‘offset term’ in a generalized linear model, which allows for a fixed proportional relationship between effort and the response variable. However, there is limited detailed guidance on the application of offsets and transformations or when an estimated effort covariate might be more appropriate.

2. This article explores the parametrisation and implementation of the offset term, plus additional methods to account for sampling effort in regression models. We evaluate the performance of offsets and covariates across various data characteristics through simulation.

3. When uncertainty regarding the effort–response relationship exists, modelling sampling effort as a log-transformed covariate, ideally as a constrained smoother, is ideal because it covers most scenarios: a proportional relationship, a non-linear (e.g. saturated) relationship, and flexibility in multi-species or hurdle models (e.g. allowing effort to influence detection probability in a binomial model). I show that parameter recovery in effort-as-covariate models is generally robust in simple models, so a log-transformed offset is only advantageous when: a proportional relationship is well-supported, model complexity or data availability hinders covariate estimation, or non-linearity at data limits is uncertain.

4. Although our simulation showed reasonable performance of all sampling effort parameterisations, how to model effort remains a key decision, and one that benefits from considered thought before modelling occurs. The nature of the effort–response relationship

(i.e. proportional, otherwise linear on the link or original scales, or non-linear), and how multiple effort variables could be included in the same model, will benefit from both statistical and practical contexts and experience.

Keywords: offset, generalized linear models, sampling effort, survey effort, catch standardisation

Code available at: https://github.com/smithja16/Effort_Offset_Simulation

1. Introduction

Survey programs and related data often have variation in sampling effort, such as differences in hours searched or areas surveyed. Accurately accounting for this variation is crucial when modelling response variables like abundance (e.g., counts or biomass) to ensure unbiased estimates of both the response and other predictor variables (e.g., environmental or spatiotemporal covariates). One common approach is to transform the response variable before modelling, such as calculating counts per unit of sampling effort. However, this approach can introduce statistical issues, such as violating the assumptions of count or biomass data distributions or misrepresenting the variance structure (Zuur et al 2009).

An alternative and often preferred method is to use a generalized linear model (GLM) that includes sampling effort as a transformed offset term or covariate (Maunder and Punt 2004, Zuur et al 2009). This approach models the response variable on its original scale while standardizing it to sampling effort. Effort variables, sometimes called 'detectability covariates' (Buckland et al. 2009; Thorson and Kristensen 2024) or 'catchability covariates' (Thorson 2019), influence only the observed magnitude of the response without altering the underlying variable. Many studies have included effort as an offset, for effort variables such as the number of trap nights (Kammerle et al 2018), distance walked (O'Kelly et al 2018), number of survey points (Ausprey et al 2023), and area trawled (Thorson et al 2020). However, there remain numerous decisions to make, and pitfalls to avoid, when including sampling effort in a GLM. These include understanding the assumed relationship between effort and the outcome variable, and the extent to which a fitted covariate can model potential relationships.

A variable included as an offset is used to adjust the expected value of the response without having an estimated parameter. This means the offset variable has a fixed coefficient value =

1. A sampling effort offset is typically only used when a GLM uses a log link function, which encompasses the common statistical distributions for abundance data: Poisson, negative binomial, and Tweedie, but also lognormal and gamma in delta (hurdle) models (Zuur et al 2009, Thorson 2018). Because the offset term is used to standardize the response (i.e. abundance per unit sampling effort), the response and effort variables need to be on the same scale, which means log-transforming effort to match the link function. In other words, the log link function allows the offset to scale the expected value of the response in a proportional way on the original scale, i.e. a 50% increase in effort means a 50% increase in abundance, all else being equal.

A general log link GLM can be written:

$$Y \sim F(\mu, \theta),\ \mathrm{E}(Y) = \mu$$

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 \log(T) \tag{1}$$

$$\mu = e^{(\beta_0 + \beta_1 X_1)} \times T^{\beta_2} \tag{2}$$

The response variable Y is a random variable with statistical distribution $Y \sim F(\mu, \theta)$, and the log link function determines the relationship between the expected value of abundance $\mathrm{E}(Y) = \mu$ and the predictor variables (the linear predictor). β_0 is the intercept, β_1 is the estimate (coefficient) for the first predictor variable X_1, and β_2 is the coefficient of the sampling effort term T. GLMs model the expected value of Y, not Y itself, and the GLM can be written on the link scale (equation 1), or the original 'response' scale (equation 2). Equation 2 shows that when log-transformed T is included in the GLM as an offset term, $\beta_2 = 1$ (the parameter is not estimated), so effort is proportional to expected abundance, i.e. we are essentially modelling $\frac{\mu}{T}$.

When modelling abundance (including outcome variables such as fishing catches) the options to 1) include sampling effort as an offset or 2) include it as a covariate, are often presented as equally able alternatives (Maunder and Punt 2004, Thorson 2019). The second option is often considered more flexible than an offset term – capable of fitting a proportional relationship as well as deviations from it – with deviations possible due to processes like the saturation of fishing gear (Kuriyama et al 2019). Non-linearity of the effort–abundance relationship may not be uncommon, especially for 'capture' sampling methods (Thorson 2019, Smith et al 2020, Smith & Johnson 2024). For this flexibility to be true, the effort variable included as a covariate must also be able to represent the proportional effort–abundance relationship implied by an offset (i.e. β_2 can be = 1 in equation 2). This depends on model structure and likely on the collinearity among covariates. One obvious pitfall is when effort T is not log-transformed before inclusion as a covariate:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 T$$

$$\mu = e^{(\beta_0 + \beta_1 X_1 + \beta_2 T)}$$

In this case, effort is no longer proportional to expected abundance and the relationship between effort and abundance is non-linear, even if $\beta_2 = 1$ (**Fig. 1a**). This model is only wrong if the user was assuming the model could act like an offset term and fit a proportional effort–abundance relationship if it existed.

The goal of this article is to provide guidance on the appropriate model structures for accounting for sampling effort in GLMs, helping researchers avoid common pitfalls and understand the implications of different modelling choices. I achieve this by generating data with different effort-abundance relationships and levels of collinearity among covariates, and by evaluating which approach for including sampling effort in a model is most robust to these different relationships. Given that the effort–abundance relationship is not usually known, a specific aim is to identify whether there is one model structure that is generally robust for modelling an unknown effort–abundance relationship that could be either proportional or non-linear. I end by highlighting some issues that are easily overlooked when accounting for the wide variety of sampling effort variables in statistical models.

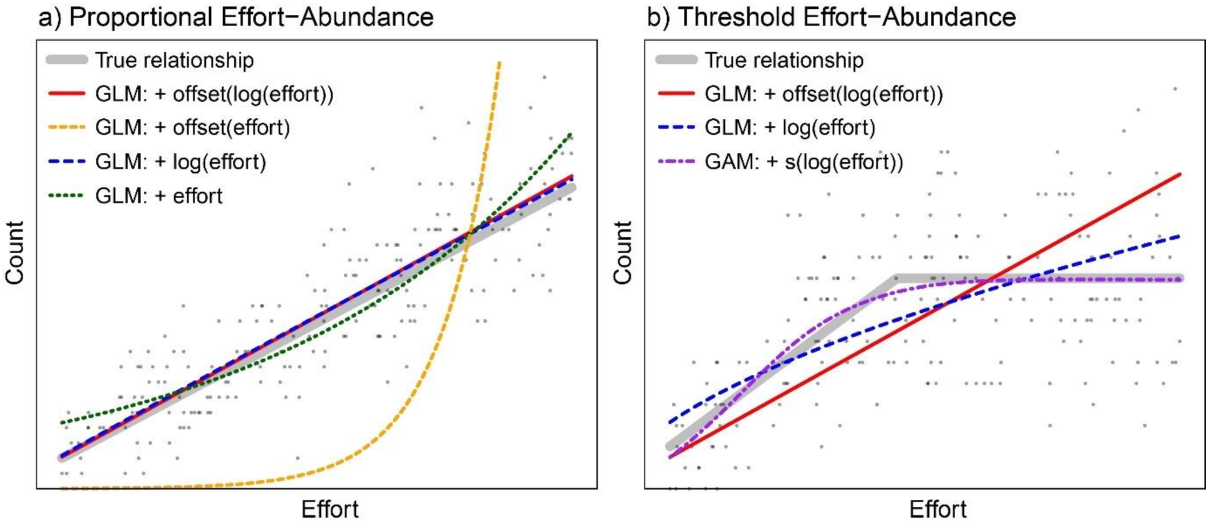

Figure 1. Example effort-abundance relationships and various effort parameterizations. This shows how different effort parameterisations can fit a proportional (a) or thresholding (b) effort–abundance relationship for simulated count data (a generalized additive model, GAM, is considered a special cased of GLM). In this simple case, the fit is poor when effort is not log transformed (a), and only a smoother can accurately estimate this thresholding relationship (b), although the +log(effort) option can be non-linear. See Table 2 for these model’s equations. Code for generating this figure is shared in the GitHub repository.

2. Testing different model structures

2.1 Data scenarios

Eight data scenarios and six model structures were tested, with R code (R Core Team 2024) available at https://github.com/smithja16/Effort_Offset_Simulation. Abundance data consisted of 1000 counts from a negative binomial distribution (theta = 3.0), generated from specified relationships with three predictors: sampling effort (arbitrary units), site (categorical, 3 levels), and temperature. These specified relationships allowed me to measure model success by the recovery of known patterns. Data scenarios varied across three factors: 1) the relationship between effort and abundance (none, proportional, threshold), 2) the collinearity among effort and the other predictor variables (collinear or not), and 3) the shape of the temperature effect (linear or domed on the scale of the linear predictor). The eight data scenarios are summarised in **Table 1** and cover sufficient combinations of the three factors to evaluate their influence.

The effort–abundance relationship determines whether an offset term is appropriate, i.e. when it is ‘proportional’ an offset term would be expected to perform well, but when it is not a more flexible approach may be better. The collinearity between predictors evaluates whether certain model structures are more accurate when effort is correlated to other variables (e.g more sampling occurs at certain sites or at specific temperatures). It is possible that a collinear effort covariate may lead to less accurate recovery of a proportional effort–abundance relationship than an offset term. The shape of the temperature effect (linear or domed) evaluates whether misspecifying one covariate (e.g. fitting a domed effect with a linear term) influences the accuracy of the effort term, especially if the two terms are collinear. Collinearity was induced using a multivariate normal distribution, and the domed temperature relationship was generated using a normal distribution.

2.2 Model structures

Seven model structures that differ primarily in how they model effort are shown in **Table 2**. Model 1 is the typical 'effort as offset' structure, and the others include effort as a covariate (Models 2-6), specifically as a smoother (Models 4-6) in a GAM to allow more flexibility. Model 7 is a log-linear model and was added for interest due to its occasional use (e.g. Shono 2008). For this analysis I tested models M1-M6, where I included two additional covariates, Site as a fixed factor and Temperature as a continuous variable. The log-linear M7 is largely identical to M1 so was not tested here. M6 was added to explore the misspecification data scenario described above, but it is identical to M4 in how it models effort. All models were fit using the 'mgcv' R package (Wood 2017).

2.3 Performance metrics

I evaluated success of our models across data scenarios by looking primarily at the 1) diagnostics of residuals, 2) model goodness of fit, and 3) parameter recovery and the effort marginal effect. Residuals (vs fitted values and vs the offset term) should be pattern free, indicating a well specified model. I used simulated residuals generated by the 'DHARMa' R package (Hartig 2022). For goodness of fit I calculated the similarity of simulated and fitted abundance counts using mean absolute error (MAE), and calculated MAE for: a) all data, b) for data at large effort values, c) for data at low effort values. MAE at large and small effort values test whether an effort–abundance relationship fits well throughout the data. Differences in collinearity and the temperature effect among data scenarios cause differences in mean abundance, so MAE should only be compared among models within each data scenario. Parameter recovery refers to the similarity in the specified and estimated values of the variable effects (i.e. coefficients or non-linear shapes) and dispersion, and a well specified model will have accurate parameter recovery. And the effort marginal effect refers to the shape of the estimated effort–abundance relationship (**Fig. 1**).

Table 1. Summary of the eight data scenarios (D1 to D8). Collinearity was induced with a covariance of 0.5 (effort–temperature) or -0.5 (effort–site). The threshold effort–abundance relationship was proportional up until max(effort)/2 and then constant at larger effort values.

No.	Effort–abundance relationship	Collinearity of predictors	Temperature effect
D1	Proportional	None	Linear
D2	Proportional	Effort–temperature	Linear
D3	Threshold	None	Linear
D4	Threshold	Effort–temperature	Linear
D5	Proportional	Effort–temperature	Domed
D6	Proportional	Effort–temperature and effort–site	Linear
D7	None	None	Linear
D8	Threshold	Effort–temperature and effort–site	Domed

Table 2. Syntax (using the R software language) and equations for common ways to specify effort in a GLM. All models have a log-link, except the last which has an identity link. For all models, abundance Y is a random variable with statistical distribution $Y \sim F(\mu, \theta)$, and the link function determines the relationship between the expected value of abundance $\mathrm{E}(Y) = \mu$ and the predictor variable(s). β_0 is the intercept, β_1 is the estimate (coefficient) for the first predictor variable X, and T is sampling effort. s indicates a smoother of function f. For the final model, c is a constant ($< \min(Y/T)$) to avoid log(0), and σ^2 is the variance of the

residuals on the log scale. The `offset` function in R ensures that no parameter is estimated for that variable (Zuur et al 2009).

No.	**R syntax and Equation**	**Implied effort–abundance relationship**
M1	`glm(Y ~ X + offset(log(effort)))` $\log(\mu) = \beta_0 + \beta_1 X + \log(T)$ $\mu = e^{(\beta_0 + \beta_1 X)} \times T$	Abundance is proportional to effort on the original scale, i.e. 10% higher effort means 10% higher mean catch; this is the typical model structure for an effort offset; transforming in the model formula `offset(log(effort))` or using a transformed effort variable `offset(log_effort)` are equivalent
M2	`glm(Y ~ X + log(effort))` $\log(\mu) = \beta_0 + \beta_1 X + \beta_2 \log(T)$ $\mu = e^{(\beta_0 + \beta_1 X)} \times T^{\beta_2}$	Abundance is proportional to effort when β_2 = 1 (as above), otherwise the relationship follows a power law, e.g. when $\beta_2 = 2$ if effort doubles mean abundance increases by a factor of 4
M3	`glm(Y ~ X + effort)` $\log(\mu) = \beta_0 + \beta_1 X + \beta_2 T$ $\mu = e^{(\beta_0 + \beta_1 X + \beta_2 T)}$	Abundance is non-linear with effort on the original scale; effort has a consistent multiplicative effect on the outcome, i.e. for every *unit* increase in effort abundance increases by a factor of $\exp(\beta_2)$
M4	`gam(Y ~ X + s(log(effort)))` $\log(\mu) = \beta_0 + \beta_1 X + f(\log(T))$ $\mu = e^{(\beta_0 + \beta_1 X + f(\log(T)))}$	The function *f* can create a variety of non-linear relationships, but when this is a straight line of slope = 1 it is equivalent to M1 with a proportional effort–abundance relationship
M5	`gam(Y ~ X + s(effort))` $\log(\mu) = \beta_0 + \beta_1 X + f(T)$ $\mu = e^{(\beta_0 + \beta_1 X + f(T))}$	This lies somewhere between M3 and M4 due to the flexibility of function *f*. But because most GAM smoothers *s* have the same degree of smoothness across a variable, the smoothness will change as μ changes, so this may not be capable of a fitting a perfectly proportional relationship (unlike M4)
M6	`gam(Y ~ s(X) + s(log(effort)))` $\log(\mu) = \beta_0 + f(X) + f(\log(T))$ $\mu = e^{(\beta_0 + f(X) + f(\log(T)))}$	Identical to M4 but with a smoother for temperature, so that the domed temperature effect of scenarios D5 and D8 could be fit properly; explores whether the effort covariate is less accurate due to misspecification of a collinear term (M4 vs M6 for D5)
M7	`glm(log(Y/effort + constant) ~ X, family=gaussian(link="identity"))` $\log\left(\frac{Y}{T} + c\right) = \beta_0 + \beta_1 X$ $Y = \left(e^{\left(\beta_0 + \beta_1 X + \frac{\sigma^2}{2}\right)} - c\right) \times T$	Equivalent to a linear model with log-transformed response, the relationship between abundance and effort is proportional by design; predictions of mean *Y* require bias-correcting the linear predictor with $+\frac{\sigma^2}{2}$, assuming that *Y* is log-normally distributed on the original scale (Duan 1983)

3. Model performance

3.1 Residuals and dispersion

All models showed sufficiently pattern-free residuals, except for D7 (no effort effect) for the one model which used an offset and thus assumed a proportional relationship (M1). Not accurately modelling a threshold effort relationship (e.g. M1 for D8) did not significantly affect the residuals or dispersion.

3.2 Goodness of fit

Models with effort included as a smoother (M4-M6) were better fitting, regardless of whether the effort–abundance relationship was proportional or thresholding (**Table 3**; raw values **Table S1**). As expected, this difference was larger for thresholding effort–abundance relationships. These models were also more accurate at high and low effort values, though this may represent some overfitting to noise (so predictive performance would also be a useful for model selection). The inaccuracy from using an untransformed effort variable (M3) was largest at high and low effort values (**Table 3**).

The difference in MAE among models was often small (**Table 3**), especially for proportional relationships, showing that most parameterizations fit the bulk of the data well (**Fig. 2**). This was somewhat a feature of our data generation, where there were fewer high and low effort to fit. The difference in MAE among models is also a function of the unexplained information or noise; e.g. when a Poisson distribution is used for data generation instead of a negative binomial the percentage differences among models more than doubles due to the higher signal-to-noise ratio.

3.3 Parameter recovery and Effort marginal effects

Parameter recovery was generally good (**Table S2, Fig. S1**), as seen in the true and estimated effort marginal effects (**Fig. 2**). All models accurately recovered the relevant temperature effects and site effects for all data scenarios. However, the intercept (representing reference site A) was poorly recovered by most models, especially when the model contained a smoother (M4-M6) or untransformed effort term (M3). The poor recovery would be typical of such data, due to the intercept representing a reference far from the simulated data (i.e. when effort or log(effort) and temperature = 0), and due to the compounded parameter uncertainty. Poor recovery was sometimes expected (e.g. a linear effort could not recover a thresholding relationship), but some cases were unexpected: M1 failed to recover the temperature effect when an offset was used to represent a thresholding effort–abundance relationship (D4), and M5 sometimes failed to recover the proportional effort relationship (D1, D5; **Fig. 2**).

Table 3. Percentage difference in mean absolute error (MAE) among models. This compares the models' relative ability to explain the data. The percent difference is relative to the model with the lowest MAE in each data scenario, e.g. M1 had an MAE on average 0.26% higher than the lowest MAE over the four proportional data scenarios. M6 is not included in this comparison because it models temperature differently, and it otherwise identical to M4.

	% MAE		% MAE high		% MAE low	
	Proportional	Threshold or constant	Proportional	Threshold or constant	Proportional	Threshold or constant
M1	0.26	3.41	1.79	23.35	1.77	2.28
M2	0.2	1	1.48	7.85	2.04	7.9
M3	0.81	1.58	4.68	9.01	10.38	17.53
M4	0.07	0.04	0.58	0.52	0.93	0.08
M5	0.01	0	0.23	0	1.2	0.31

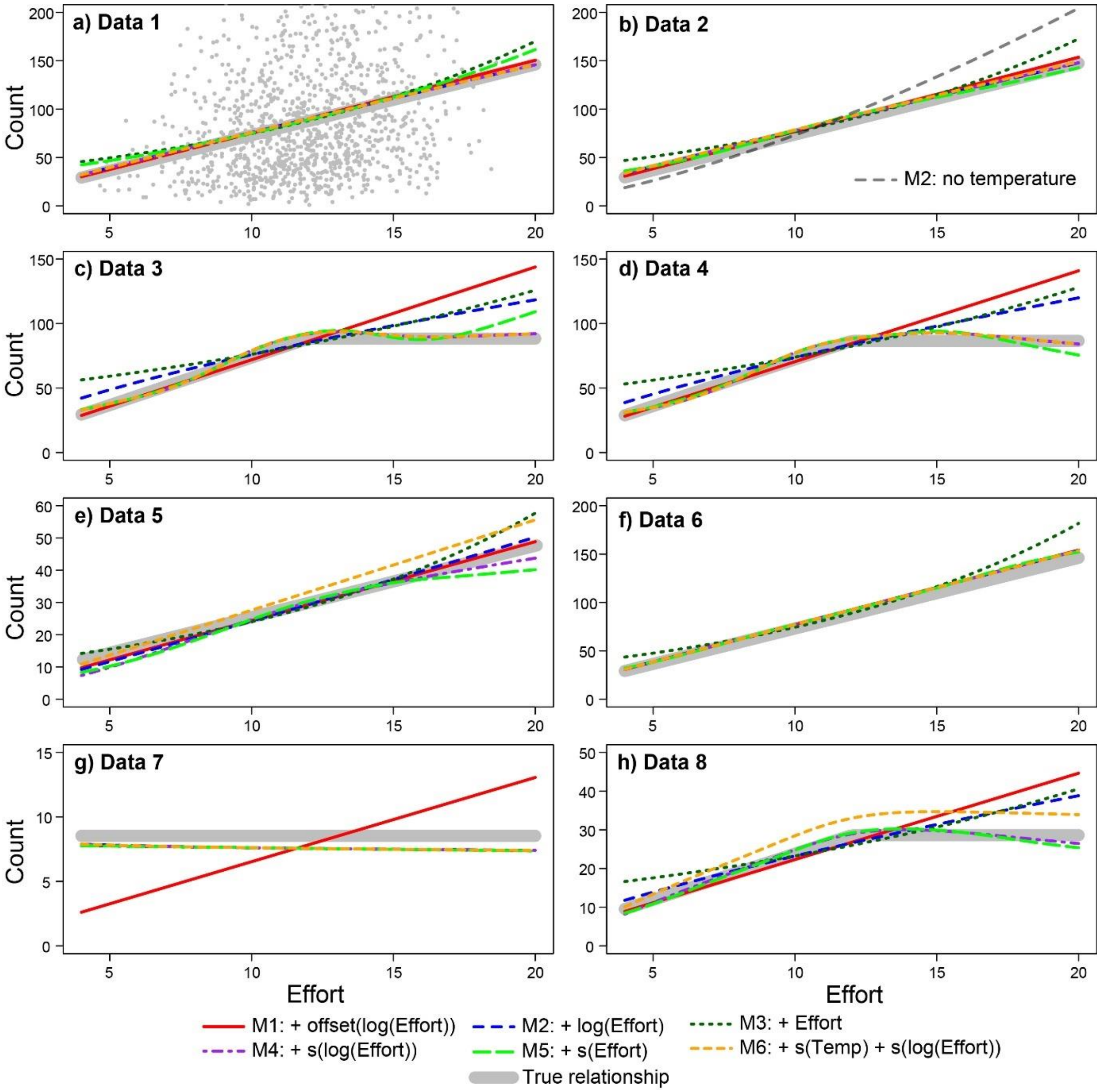

Figure 2. The true and estimated marginal effort–count relationships for the eight data scenarios (a-h) and six model structures. The true relationships used to generate the data are shown as thick grey lines (this marginal is approximate in any scenario with collinear variables). The generated data are shown in a), highlighting the noise inherent in the negative binomial distribution as well as variation in the temperature and site effects. In b) is shown the effort marginal estimated by a version of M2 without the temperature covariate (grey dotted line); this illustrates that inducing collinearity between temperature and effort does influence the estimated effort effect.

4. What this simulation says about offsets vs covariates

The data scenarios and model structures allowed me to evaluate the following questions: 1) *Which models failed to fit a proportional effort–abundance relationship?* Only the models with untransformed effort covariates (M3 and M5). 2) *Was a proportional relationship better estimated by an offset than a covariate when there was collinearity among covariates?* No, there was no evidence of this given the moderate collinearity in this simulation – provided that all collinear variables were included in the model (**Fig. 2b**). The issues of collinearity and parameter estimation are well discussed elsewhere and apply equally here (Zuur et al 2010, Dormann et al 2013). 3) *Was an effort smoother affected by the misspecification of the non-linear but correlated temperature covariate?* Yes, a small amount. This is seen comparing M4 and M6 for D4 and D8. When temperature was correctly specified (M6) the model seemed to better recover the shape of the effort–abundance relationship (Fig. 2), but the change was small. 4) *Were there any limitations to using an effort smoother?* There was some overfitting at the edges due to fewer data and less certainty estimating the expected value (Fig. 2), but otherwise no. For scenarios with fewer observations, or high model complexity, having additional parameters and increased estimation uncertainty could be considered a limitation. 5) *Are there limitations to an offset term?* Only when there is non-proportionality in the effort–abundance relationship, which when unmodelled can also affect the accuracy of other estimates (e.g. poor recovery of temperature for M1 in D4).

5. Recommendations

Based on this analysis and other studies, I suggest the following recommendations and guidance for the modelling of sampling effort and the use of offset terms:

- Including effort as either an offset `+offset(log(effort))` or covariate `+log(effort)` appear equally suitable for modelling a proportional effort–abundance relationship, provided that the effort covariate is log-transformed (for a log-link GLM). Only the offset guarantees proportionality, so is preferred when proportionality is highly likely; this avoids deviations form proportionality due to measurement error or model misspecification. Effort variables likely to be proportional would be the area or duration surveyed by observational methods.
- If some non-proportionality is possible (e.g. through gear saturation) include effort as a covariate, as this can model a proportional relationship plus deviations from it; and using a smoother `+s(log(effort))` term when possible seems most useful due its flexibility and robust shrinkage (the log-transformation of effort is important even for smoothers).
- Consider constraining the wiggliness of the effort smoother to avoid unlikely relationships, e.g. using parameter 'k': `+s(log(effort),k=4)`, stronger penalization using the '`select`' and '`gamma`' arguments (Wood 2017), and additional shape constraints if necessary (Pya and Wood 2015).
- Using an effort covariate is an obvious choice for including effort in both parts of a hurdle model, as offsets have a different use in binomial models and are typically only used in the positive component (Thorson 2019; but see Shelton et al 2014), although there are related alternatives to delta models (Thorson 2018). However, the effort–probability relationship has less theoretical basis and cannot be proportional. Adding an untransformed effort variable (to a logit-link model) assumes that *absolute* changes in effort have a consistent effect on the log-odds of sampling, which is probably more realistic than a log-transformed effort variable (**Fig. S2**).
- For most machine learning methods, effort must be included as a covariate or the response changed to *Y*/*effort* (Leathwick et al 2006, Li et al 2015, Smith et al 2020).

Given that the lack of link function, log-transforming the covariate is no longer essential but is probably still useful, especially for methods like Bayesian additive regression trees which have more constraints on their fitted responses (Chipman et al 2010) compared to methods like random forests.

- Understanding the data is essential: consider whether your measure of effort is likely to be proportional based on controlled experiments or expert advice, e.g. a doubling of trawl area can double abundance, but a doubling of underwater viewing distance of a camera probably does not (Smith et al 2017) and should be converted to viewing volume.
- Understanding the data is essential: evaluate aspects such as interactions between effort and other variables, e.g. imagine catches of fish are highest at dusk and dawn but nets can be left overnight for practical purposes – the effort–abundance relationship would vary across a 'duration' variable; in this case a smoother covariate would be ideal, and a categorical effort covariate (< 6 h, > 6 h) worth exploring. Evaluate cautiously a plot of effort vs abundance to evaluate the relationship as it is likely conditional on other effects.
- Multiple offsets can be used if there is more than one detectability covariate (e.g. Kortello et al 2024) but this is the same as using their product: $Y/(effort_1 \times effort_2)$. Calculating this product can evaluate whether this model structure is logical. Otherwise use either one offset term and additional variables as covariates, or all terms as covariates (Maunder and Punt 2004), although care should be taken to note which covariates are assumed multiplicative and which proportional (Grüss et al 2019).
- Like with any covariates, very strong collinearity will likely bias estimation of parameters and their errors, so remove such covariates when inference of the effort effect is important. Including effort as an offset vs a covariate does somewhat delineate an interest in standardisation (an offset) or inference (covariate).

6. Additional considerations

Additional points to consider when modelling effort include:

- In multi-species models, should all species be considered to have the same effort–abundance relationship? Perhaps, although an effort covariate could be fitted for each species in case of species differences (Smith et al 2024). An offset may be more robust for rare species for which a covariate may have considerable uncertainty.
- A similar alternative to M7 (**Table 2**) is modelling abundance per unit effort (*Y*/*effort*) but avoiding pre-model transformation by using a delta model combining binomial and (commonly) gamma components (e.g. Panzeri et al 2024).
- What do our GLMs assume about zeros in the data? Is a zero from (say) 2 hours effort equivalent to a zero from 24 hours effort? Probably not, and such an occurrence would have no influence on a model with an effort offset term. This would, however, influence an estimated effort covariate. These patterns could also influence a binomial model by including effort as a model 'weight' if a covariate is deemed inappropriate.
- Treating an effort as a factor is also possible (Grüss et al 2019), e.g. soak time with four levels (Groeneveld et al 2003), which could be useful when there are clusters of effort values or insufficient data, and such variables could estimate flexible relationships, with each level estimated separately.

References:

Ausprey, I. J., Newell, F. L., & Robinson, S. K. (2023). Sensitivity of tropical montane birds to anthropogenic disturbance and management strategies for their conservation in agricultural landscapes. *Conservation Biology*, *37*(6), e14136.

Buckland, S. T., Russell, R. E., Dickson, B. G., Saab, V. A., Gorman, D. N., & Block, W. M. (2009). Analyzing designed experiments in distance sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, *14*, 432-442.

Chipman H. A., George, E. I., McCulloch, R. E. (2010) BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266-298.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, *78*(383), 605-610.

Groeneveld, J. C., Butterworth, D. S., Glazer, J. P., Branch, G. M., & Cockcroft, A. C. (2003). An experimental assessment of the impact of gear saturation on an abundance index for an exploited rock lobster resource. *Fisheries Research*, *65*(1-3), 453-465.

Grüss, A., Walter III, J. F., Babcock, E. A., Forrestal, F. C., Thorson, J. T., Lauretta, M. V., & Schirripa, M. J. (2019). Evaluation of the impacts of different treatments of spatio-temporal variation in catch-per-unit-effort standardization models. *Fisheries Research*, *213*, 75-93.

Hartig F (2022). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models. R package version 0.4.6, https://CRAN.R-project.org/package=DHARMa.

Kämmerle, J. L., Corlatti, L., Harms, L., & Storch, I. (2018). Methods for assessing small-scale variation in the abundance of a generalist mesopredator. *PloS One*, *13*(11), e0207545.

Kortello, A., Hausleitner, D., Mowat, G., Barrueto, M., Heim, N., Larson, L., ... & Clevenger, A. (2024). How to find a wolverine: factors affecting detection at wolverine (*Gulo gulo*) bait stations in western Canada. *Ecosphere*, *15*(7), e4929.

Kuriyama, P. T., Branch, T. A., Hicks, A. C., Harms, J. H., & Hamel, O. S. (2019). Investigating three sources of bias in hook-and-line surveys: survey design, gear saturation, and multispecies interactions. *Canadian Journal of Fisheries and Aquatic Sciences*, *76*(2), 192-207.

Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., & Taylor, P. J. M. E. P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, *321*, 267-281.

Li, Z., Ye, Z., Wan, R., & Zhang, C. (2015). Model selection between traditional and popular methods for standardizing catch rates of target species: a case study of Japanese Spanish mackerel in the gillnet fishery. *Fisheries Research*, *161*, 312-319.

Maunder, M. N., & Punt, A. E. (2004). Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, *70*(2-3), 141-159.

O'Kelly, H. J., Rowcliffe, J. M., Durant, S. M., & Milner-Gulland, E. J. (2018). Robust estimation of snare prevalence within a tropical forest context using N-mixture models. *Biological Conservation*, *217*, 75-82.

Panzeri, D., Russo, T., Arneri, E., Carlucci, R., Cossarini, G., Isajlović, I., ... & Libralato, S. (2024). Identifying priority areas for spatial management of mixed fisheries using ensemble of multi-species distribution models. *Fish and Fisheries*, *25*(2), 187-204.

Pya and S. N. Wood. Shape constrained additive models. *Statistics and Computing*, 25:543–559, 2015.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Shelton, A. O., Thorson, J. T., Ward, E. J., & Feist, B. E. (2014). Spatial semiparametric models improve estimates of species abundance and distribution. *Canadian Journal of Fisheries and Aquatic Sciences*, *71*(11), 1655-1666.

Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, *93*(1-2), 154-162.

Smith, J. A., Cornwell, W. K., Lowry, M. B., & Suthers, I. M. (2017). Modelling the distribution of fish around an artificial reef. *Marine and Freshwater Research*, *68*(10), 1955-1964.

Smith, J. A., Tommasi, D., Sweeney, J., Brodie, S., Welch, H., Hazen, E. L., ... & Jacox, M. G. (2020). Lost opportunity: Quantifying the dynamic economic impact of time-area fishery closures. *Journal of Applied Ecology*, *57*(3), 502-513.

Smith, J. A., & Johnson, D. D. (2024). Evaluating drivers and predictability of catch composition in a highly mixed trawl fishery using stacked and joint species distribution models. *Fisheries Research*, *279*, 107151.

Thorson, J. T. (2018). Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(9), 1369-1382.

Thorson, J. T. (2019). Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. *Fisheries Research*, *210*, 143-161.

Thorson, J. T., Adams, C. F., Brooks, E. N., Eisner, L. B., Kimmel, D. G., Legault, C. M., ... & Yasumiishi, E. M. (2020). Seasonal and interannual variation in spatio-temporal models for index standardization and phenology studies. *ICES Journal of Marine Science*, *77*(5), 1879-1892.

Thorson, J., & Kristensen, K. (2024). *Spatio-temporal Models for Ecologists*. CRC Press.

Wood, S. N. (2017). *Generalized Additive Models: an Introduction with R*. 2nd Ed. Chapman and Hall/CRC.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3-14.

Supplementary Material

Table S1. Raw mean absolute error (MAE) values. Raw mean absolute error (MAE) values for the six models and eight data scenarios. Values should only be compared within data scenarios.

	M1	M2	M3	M4	M5	M6
D1	44.24	44.18	44.22	44.18	44.20	44.18
D2	45.54	45.44	45.72	45.44	45.36	45.31
D3	43.87	42.96	43.12	42.52	42.47	42.52
D4	43.27	42.74	43.03	42.10	42.09	42.10
D5	15.34	15.36	15.48	15.28	15.27	14.80
D6	44.02	44.01	44.42	44.01	43.98	44.01
D7	4.60	4.37	4.37	4.37	4.37	4.37
D8	13.32	13.18	13.34	13.01	13.01	12.54

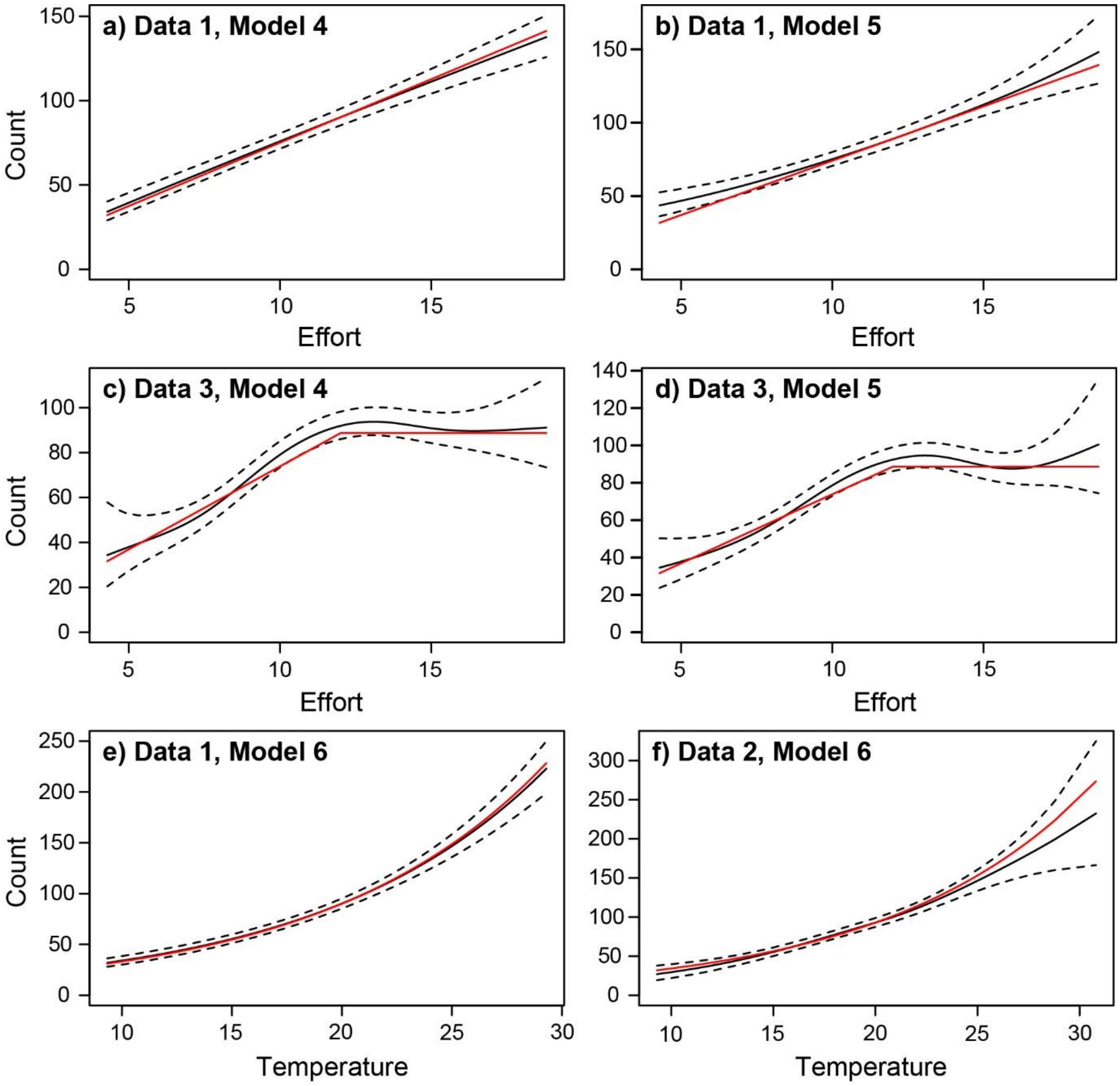

Fig. S1. Examples of how the success of effort and temperature smoothers were evaluated. In a-d) if the true effort–abundance relationship (red line) was within the 95% confidence interval for the estimated smoother (black lines), then the model was deemed successful at recovering the true effort–abundance relationship. In e-f) the same process was used but for the ‘linear’ temperature–abundance relationship.

Table S2. Summary table of parameter recovery for relevant models and data types; T = temperature, Int = intercept, SB = Site B, SC = Site C. Green cells indicate when the 95% confidence interval of an estimate encompassed the true value used to generate the data. Three results which reflect unexpected poor recovery of parameters are highlighted purple.

	Model 1				Model 2					Model 3					Model 4					Model 5					Model 6				
	T	Int	SB	SC	T	Int	SB	SC	E[1]	T	Int	SB	SC	E	T	Int	SB	SC	E[3]	T	Int	SB	SC	E[5]	T[4]	Int	SB	SC	E[5]
D1														[2]															
D2														[2]															
D3									[7]					[7]															
D4									[7]					[7]															
D5	[6]				[6]					[6]				[2]	[6]					[6]					[5]				
D6														[2]															
D7																													
D8	[6]				[6]				[7]	[6]				[7]	[6]					[6]					[5]				

[1] Parameter recovery for the 'log(effort)' variable in M2 was considered successful when the 95% interval for the slope included 1, or 0 (D7)
[2] It is not possible for the 'effort' variable in M3 to recover the correct parameter which is on the log scale, except for a slope = 0 for D7
[3] Parameter recovery for the 's(log(effort))' variable in M4-M6 was considered successful for the proportional data scenarios (D1, D2, D5, D6) when a slope=1 (or slope=0 for D7) linear-predictor effort term was within the confidence interval for the smoother for the extent of effort values; likewise for the threshold data scenarios (D3, D4, D8; Fig. S1)
[4] Parameter recovery for the 's(Temperature)' variable in M6 was considered successful for when a slope=0.1 linear-predictor temperature term was within the confidence interval for the smoother for the extent of temperature values (Fig. S1); when the temperature effect was non-linear (D5, D8) success was evaluated by eye (see [5])
[5] The temperature smoother for D5 was accurately symmetrical and correctly centred on T_{mean} = 20; the smoother for D8 was less symmetrical and centered on 18
[6] It was expected that linear temperature terms could not recover a domed effect (D5 and D8)
[7] It was expected that the linear effort terms could not recover the threshold effort relationship (D3, D4, D8)

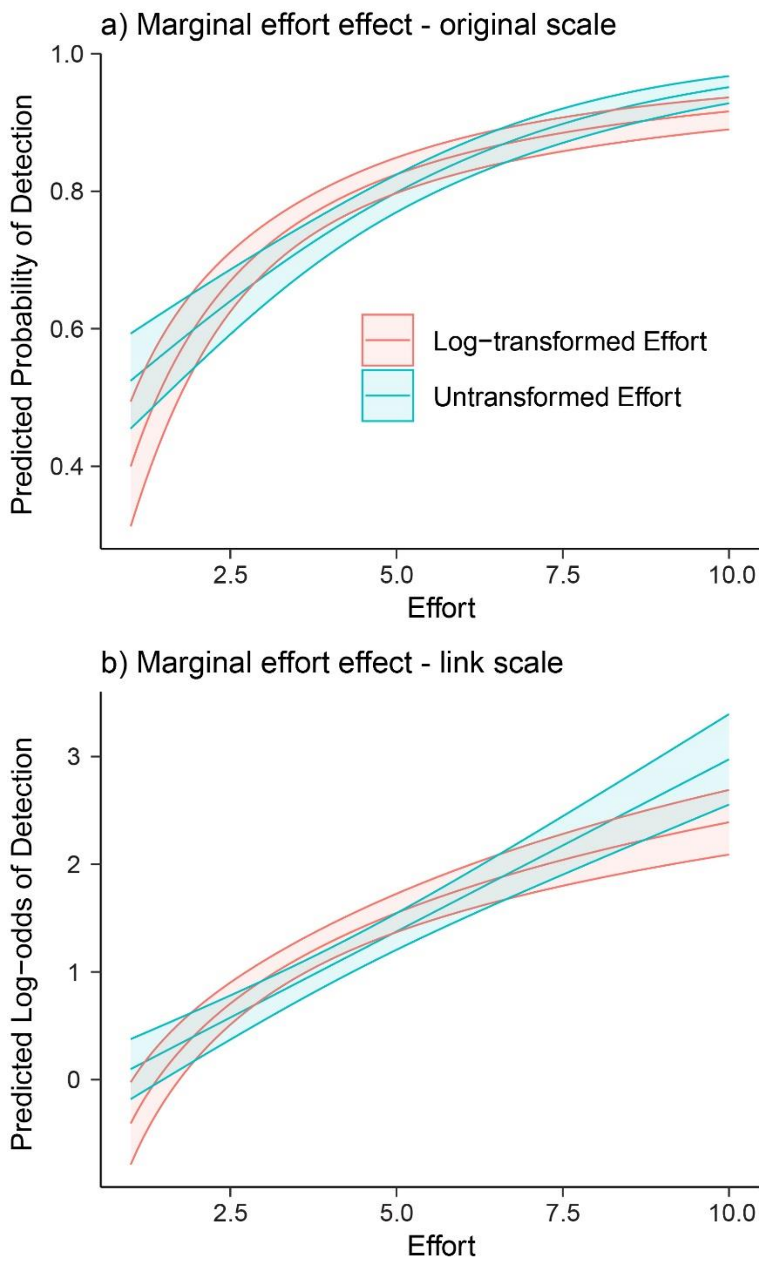

Fig. S2. The difference in the assumed effort-probability relationship when effort is included as a log-transformed (red) or untransformed (blue) covariate in a binomial GLM. The untransformed covariate is linear with log-odds (b, link scale), whereas the log-transformed covariate is non-linear on both scales. These are simulated data, and code for generating this figure is shared in the GitHub repository.