

# Predicting eel production

Andrew L Jackson & Andrew C Parnell

August 16, 2011

## 1 Introduction and notation

The aim of this modelling exercise is to develop a robust statistical model to estimate, with appropriate uncertainty, eel production at sites and locations where data are scarce. Foundational to this model are large datasets from a subset of regions around Europe where eels are present. From these data, and using existing knowledge of growth, mortality and recruitment, we develop a Bayesian model to allow production estimates to be derived using geographic and environmental data which are readily available for most locations in Europe where specific biological data might be absent.

Production ( $P$ ) in  $\text{Kg year}^{-1}\text{hectare}^{-1}$  can be inferred in a variety of ways, but from a population biology perspective, it can be described mechanistically by the equation

$$P(t) = \frac{N(0)S(t)M(t)}{A} \quad (1)$$

where,

- $P(t)$  is the yearly production in  $\text{kg/ha}$  at time  $t$ .
- $N(0)$  is the number of immature glass eels recruited to the population at time  $t = 0$ .
- $S(t)$  is the proportion surviving (survivorship) to time  $t$ .
- $M(t)$  is the mass in  $\text{Kg}$  of eels at time  $t$ .
- $A$  is the area in  $\text{ha}$  of the catchment under review.

We drop the explicit reference to  $t$  and take logs so that

$$\log P = \log S + \log N + \log M - \log A \quad (2)$$

This formulation is the underlying model assumed for all our subsequently described statistical models which aim to take existing data and models of  $N$ ,  $S$  and  $M$  as they relate to environmental variables in order to estimate production where such relationships are not known

directly. This formulation comes with some inherent assumptions. Firstly, it is a static model of a population process and assumes that an entire cohort of new recruits mature and contribute to production at the same time  $t$ . Similarly, recruitment from years around the defined  $t = 0$  are assumed to not contribute to production at time  $t$ . Given enough sequential years of data it would be possible to fit a dynamic population model to the data which would mechanistically describe the processes of recruitment survival and maturation. However, even a static model (as we are constrained to fit) will add considerable value to our understanding of this system.

In essence, the full model we construct involves modelling  $N$ ,  $S$  and  $M$  separately using a combination of mechanistic algebraic models and data-driven Bayesian models where the mathematical relationships are unknown. These models are described in detail subsequently, but here we provide a short summary of how these were estimated and included in the full model.

Recruitment is a major unknown variable in this model. Rather than having direct estimates of the number of recruited eels (hence  $N(0)$  is unknown), we have a dimensionless index of recruitment ( $R$ ), a relationship we approximate via  $N \approx aR^b$ . The estimates of the coefficients  $a$  and  $b$  represent the unknown and estimated transformations linking recruitment index and recruitment number. Similarly, since Mass and survivorship are observed with uncertainty, we include parameters for multiplicative and additive biases in estimating their relationship with production.

Survival  $S$  is obtained as a deterministic function  $h$  of the growth rate of Eels, as well as estimates of length at maturity, and various sex-related parameters which we denote as  $\Omega$ . We thus write  $S = h(G, L, \Omega)$  where  $G$  is the growth rate and  $L$  is the length at maturity. In our model the values of  $L$  and  $\Omega$  are fixed, being obtained from Bevacqua et al. (2011). The growth rate, however, is estimated (as discussed in Section 3) with predictions based on catchment covariates  $Z$ .

A directed acyclic graph (DAG) for the growth rate modelling (combined with that of the survival modelling outlined in Section 4) is shown in Figure 1. A DAG for the production modelling stage is shown in Figure 2.

The document is structured as follows. Section 2 gives a literature review on the subject of estimating production. Section 3 details the growth rate modelling component of the project. Section 4 covers the estimation of the survivorship  $S$ . Section 5 draws on the previous sections to derive an overall model for production. We conclude with a discussion in Section 6.

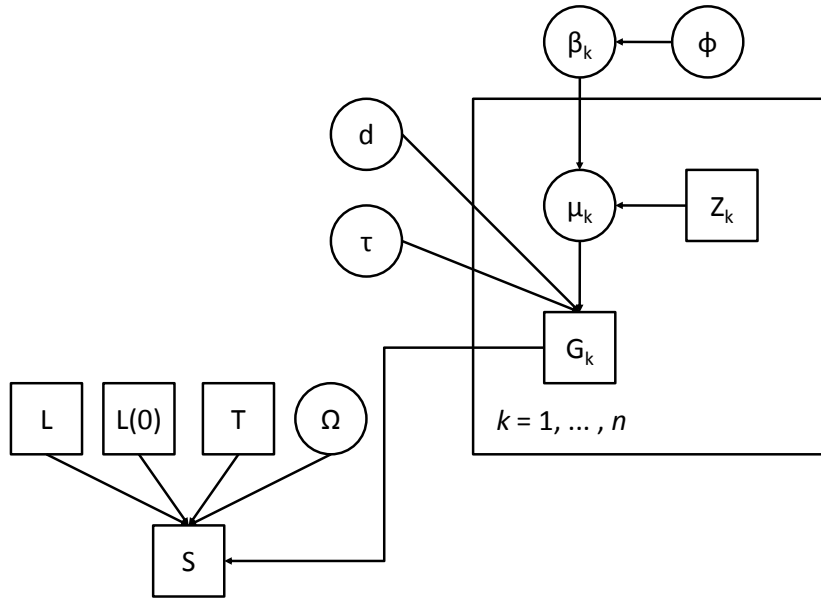


Figure 1: A DAG of the growth rate and mortality modelling process.

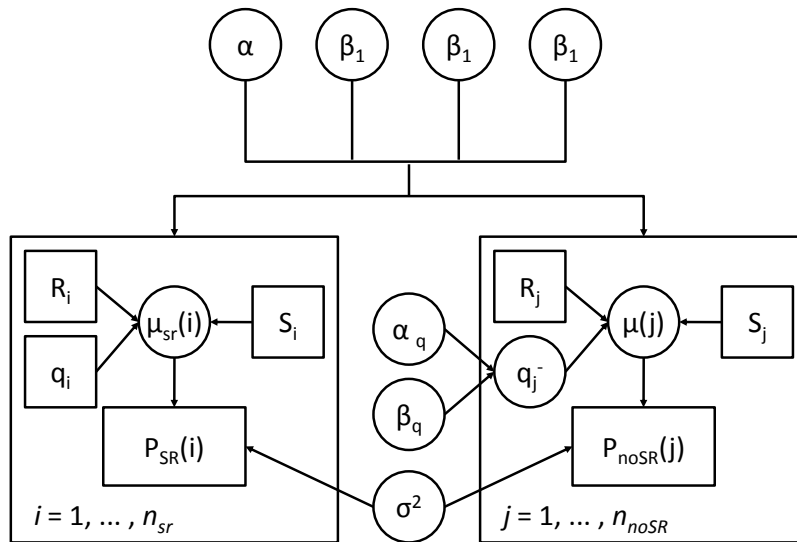


Figure 2: A DAG for the production modelling stage

## **2 Previous work on estimating Eel production**

### **2.1 Previous work on geographic variation in life history traits**

Large-scale models of eel production have hitherto not been attempted as far as we are aware. Where information on relevant life history parameters such as growth rates are available, such as the meta-analysis described in Jessop (2011), the models are a separate series of regression type analyses. While providing useful information, this approach does not consider subtle but important factors (with interactions) together which can mask and distort our understanding of the relationships. Nevertheless, these approaches provide considerable information about important key relationships such as strong effects of latitude, longitude and gender on growth rates.

### **2.2 Previous work on mortality**

The main paper on eel mortality is Bevacqua et al. (2011) and it provides a detailed algebraic model for calculating instantaneous mortality rates in eels as a function of gender, size, stocking density and water temperature. This model is a considerable improvement over the previous method which assumed a constant, discrete annual mortality rate of 14% per annum. However, the model as presented allows calculation of instantaneous mortality rates, whereas for our purposes, we need to know the probability of a given eel surviving to maturity. In this document we provide the equation to calculate this survivorship function algebraically allowing estimation of the proportion of eels surviving to any given age.

### **2.3 Previous work on production**

The Department of Communications, Energy and National Resources (2009) describe a methodology for estimating production of eels for catchments in the west of Ireland based on the underlying geology by linking this in turn to growth rates. This straight-forward approach forms the basis for our more complex and statistically robust method described here for the whole of Europe.

### 3 Growth rate modelling

One of the main determinants of eel production is assumed to be growth rate, which underpins the proportion of recruits surviving to maturity ( $S$ ) and their mass at maturation ( $M$ ). Growth rates vary considerably across Europe, with environmental factors such as local climate, geography and chemistry driving these patterns to a greater or lesser extent. The dataset provided by the Marine Institute comprised a total of 12,145 observations of either single or groups of eels for which growth rate is calculated from:

$$G = \frac{L - L(0)}{A} \quad (3)$$

where  $L$  and  $A$  are the lengths (in cm) and ages (in years, determined usually by otolith analysis) of the eels and  $L(0)$  is the length of glass eels entering the system and is taken to be 7cm in all analyses. In total, four models were fitted separating out on-land (OL) and at-sea (AS) data, and further separating the OL data into males-only (OLM), females-only (OLF) and all-sexes, OLALL, comprising males, females and undetermined sex. Growth rate was found to have a highly skewed distribution so  $\log(G)$  was chosen as the response variable. A range of explanatory environmental factors (denoted  $Z$ ) were available for each site from which we built a suite of models in order to identify the most suitable combination of variables for prediction of growth rates (Table 3).

Variable	On land	At sea
sample size	✓	✓
post-1980	✓	✓
latitude	✓	✓
longitude	✓	✓
mean temperature	✓	✓
tidal (yes or no)	✓	✓
river depth (deep or shallow)	✓	✓
geology (Siliceous/calcareous/clay)	✓	
mean elevation	✓	
mean slope	✓	
mean rain	✓	
trophic status	✓	
maximum Strahler order	✓	
area (km <sup>2</sup> )	✓	
lacustrine area (km <sup>2</sup> )	✓	
fluvial habitat area (km <sup>2</sup> )	✓	
saline (yes or no)	✓	

Table 1: Candidate explanatory variables ( $Z$ ) available for the on-land (OL) and at-sea (AS) data subsets.

Some observations were removed from the data set prior to model fitting. For the OL data it was agreed that Turkey would be removed from the analysis as it is situated some distance from

the other observations. Another 11 suspicious cases were removed where growth rate was found to be less than zero. No cases were removed from the AS data.

An explanatory variable ‘post-1980’ was included in all models so as to measure comparative ‘pristine’ growth rates. However, in all but one of the following models it was found to be unimportant.

### 3.1 Model details

We fit a selection of robust, regularised, weighted  $t$ -regressions to the growth rate data. We define  $G_i$  to be the growth rate for observation  $i$ ,  $Z_i$  to be a collection of explanatory variables including interactions as appropriate,  $\gamma_0$  to be an unknown intercept, and  $\gamma$  to be a vector of unknown parameters. The model can be written as:

$$\log(G_i) = \gamma_0 + \mathbf{Z}_i^T \boldsymbol{\gamma} + \epsilon_i \quad (4)$$

The extreme variability of some of the observations (especially those aggregated from multiple eels) suggests that a normal distribution is inappropriate for the residuals  $\epsilon_i$ . Instead we use a scaled  $t$  distribution, so that  $\epsilon_i \sim \tau t_d$ , with  $\tau$  a scale parameter and  $d$  a degrees of freedom parameter both to be estimated from the data. Large values of  $d$  are equivalent to letting  $\epsilon \sim N(0, \tau)$ , i.e. standard linear regression.

We note that some of the observations were obtained from averages of several eels, whereas others were derived from individuals. We thus adjust our scale parameter so that  $\epsilon_i \sim \frac{\tau}{n_i} t_d$  where  $n_i$  is the number of observations used to create observation  $i$  ( $n_i$  is given in the data sets as the variable ‘sample size’). Such a weighting technique allows for those samples made up of many eels to have more say in the final estimation of parameter values. It also enables easier identification of extreme values.

All the explanatory variables we use are standardised to have zero mean and unit standard deviation, so as to aid parameter identification.

### 3.2 Prior distributions

In a Bayesian model, prior distributions are required for the parameters. When little prior information is available, we use reference or Jeffreys priors (see e.g. Lee, 2004). In the case of the intercept parameter  $\gamma_0$ , we use a flat prior distribution. For the scale parameter  $\tau$  we use the transformation-invariant prior  $p(\tau) \propto \tau^{-1}$ . We give the degrees of freedom parameter  $d$  a  $U(3, 30)$  distribution. If  $d < 3$  the distribution no longer has finite variance making prediction impossible. When  $d > 30$  the residuals can be considered practically normally distributed so there is no further gain in increasing  $d$ .

A key part of this model is that we use a regularisation prior (Birkes and Dodge, 1993) on the regression coefficients  $\gamma_j$  so that:

$$\gamma_j \sim N(0, \phi) \quad (5)$$

for  $j = 1, \dots, m$  (with  $m$  the number of explanatory variables included in the model). Such a prior distribution has the effect of shrinking the estimated regression coefficients towards zero.

This behaviour is desirable for a number of reasons. First, when dealing with models with large numbers of covariates (or interaction terms in this case) the fitted model will give large parameter values only to those variables which are strongly related to growth rate. Second, as regression models tend to overstate the importance of certain explanatory variables (for examples see Birkes and Dodge, 1993) we are able to more honestly state their effects and hopefully provide improved predictions. The parameter  $\phi$ , which controls the degree of shrinkage, is estimated from the data and thus also requires a prior distribution. We again use the Jeffreys prior  $p(\phi) \propto \phi^{-1}$ .

### 3.3 Model fitting and selection

An important part of the modelling process is determining which of the explanatory variables contribute to estimating growth rate. Furthermore, some explanatory variables may interact, for example growth rate may be strongly increased when a catchment is both deep *and* saline. Including all possible explanatory variables and their interaction terms into a model is infeasible. We instead consider a large set of models including all the explanatory variables and all 2-way interaction terms. We prune the selection of models by first using a simple non-Bayesian stepwise regression (see Venables and Ripley, 2003, P176), and then sub-select models from these via the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002) and by looking at the values of the regularised coefficients. Values of the DIC for a model represent a balance between goodness-of-fit and complexity (as measured by the effective number of parameters). Models with small DIC are considered most preferable as they have good predictive power and interpretability. We give the DIC values obtained for the models to allow comparison with possible future models created on the same data.

Models were fitted via Markov chain Monte Carlo (MCMC) using the software R and jags (Plummer, 2011). We run the models for 20,000 iterations, removing an initial burn-in set of 4000, and thin the remaining values by 16. The posterior distribution for the degrees of freedom parameter for the  $t$ -distribution was found to be 3 or 4 in all cases indicating that the residual errors are long-tailed and do not match a simple normal distribution. Unfortunately this will also yield predictions that are less precise.

We subject all models to a variety of convergence tests. Specifically, we used autocorrelation functions, trace plots, and the Brooks-Gelman-Rubin diagnostic (Brooks and Gelman, 1998) to check convergence of the MCMC algorithm. For brevity, we do not show the detailed convergence results. Residuals were tested for their compatibility with the assumption of  $t$ -distributed residuals and extreme observations were identified for further analysis. Fitted values were tested for weak correlation against estimated residuals. Predicted values were checked against explanatory variables and compared with the observed data. Finally, both fitted and data values were tested against some of the excluded explanatory variables (i.e. not included in the final model) to determine an ‘out-of-sample’ prediction error.

## 3.4 Results

### 3.4.1 AS model results (DIC=135.2, effective number of parameters 16.2)

The final AS model was selected using just eight explanatory variables, including two interaction terms. The final explanatory variables used are shown in Figure 3. The residuals showed a good fit to the distribution as evidenced by the top two panels of Figure 4. It performed similarly well when comparing fitted values across explanatory variables included and excluded from the model.

The most extreme observations were ID numbers 50 (from Spain) and 13120 (Germany). These contain respectively the largest and smallest residual values. Observation 50 is based on 17 samples which have a very high growth rate (14.5) compared to other samples from Spain (mean of all the Spain observations 6.81). Obs 13120 is based on 137 samples which have a low growth rate (3.76) compared to other samples from Germany (mean all obs 5.86).

### 3.4.2 OLM model results (DIC = 358.9, effective number of parameters 37.2)

The OL models used a far larger selection of explanatory variables and interaction terms. The final OLM model is shown in Figure 5. The residuals again showed a reasonable fit to the  $t_4$  distribution, though there were some extreme outliers, particularly those made up from aggregating large numbers of eels. The predictive model performance on included and excluded explanatory variables was good. Residual diagnostics for the OLM model are shown in Figure 6.

The two most extreme observations are ID numbers 13500 and 2133, both from France. Observation 13500 is based on 216 samples which have a very high growth rate (13.4) compared to other samples from France (the mean of all French observations was 7.21). Observation 2133 is based on only 14 samples which have a low growth rate (4.8) compared to other samples from France.

### 3.4.3 OLF model results (DIC=976, effective number of parameters = 64)

The OLF model also used a large number of explanatory variables and interaction terms, with results consistent with that of males. Figure 7 shows the variables included. The OLF model identified several extreme observations, particularly those with very large growth rates. However, the predictive performance of the model was still good. The residual diagnostics for the OLF model are shown in Fig 7.

The two most extreme observations were again from France, ID numbers 13501 and 10806. Observation 13501 is based on 212 samples which have a very high growth rate (14.5 compared to other samples from France (the mean of all French observations is 6.15). Observation 10806 is based on 5 samples which have a low growth rate (0.13) compared to other samples from France.



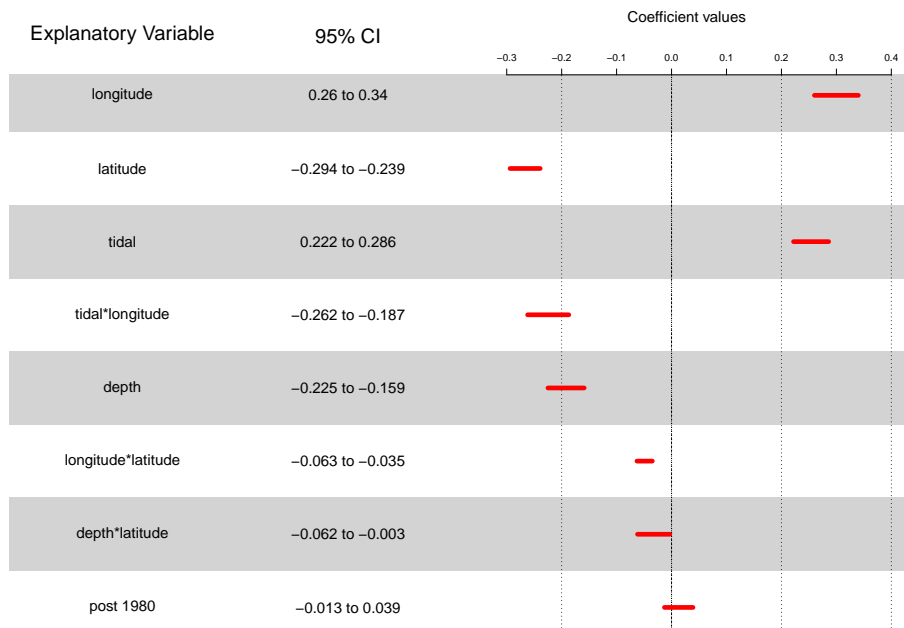


Figure 3: A graph of the final included explanatory variables for the AS model with 95% credible intervals given numerically and pictorially.

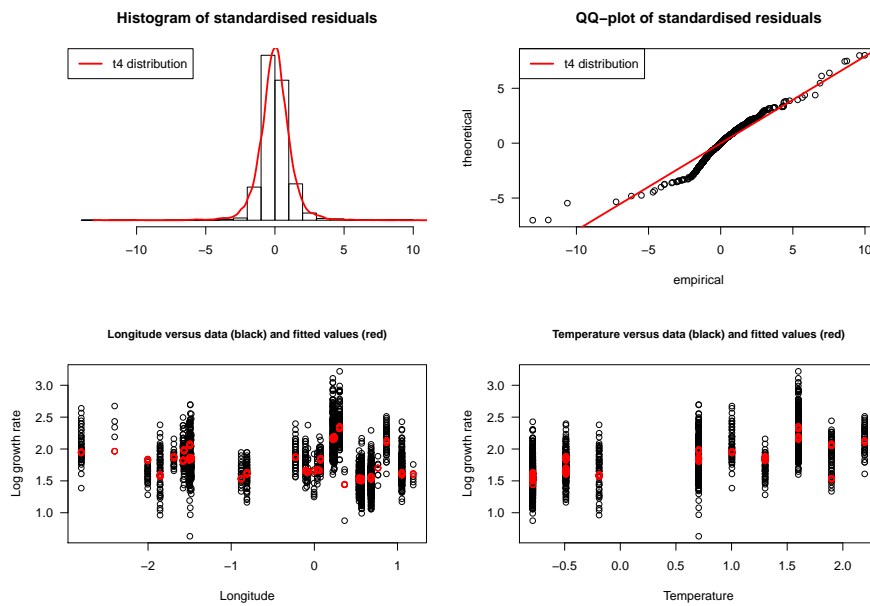


Figure 4: A four panel graph showing model performance for the AS model. The top left panel shows a histogram of the standardised residuals together with a theoretical  $t_4$  distribution. The top right panel shows the QQ-plot of the standardised residuals. The bottom left panel shows the fitted values (red) and data values (black) against an explanatory variable included in the model. The bottom right panel shows the fitted values (red) against the data values (black) against an explanatory variable not included in the model.

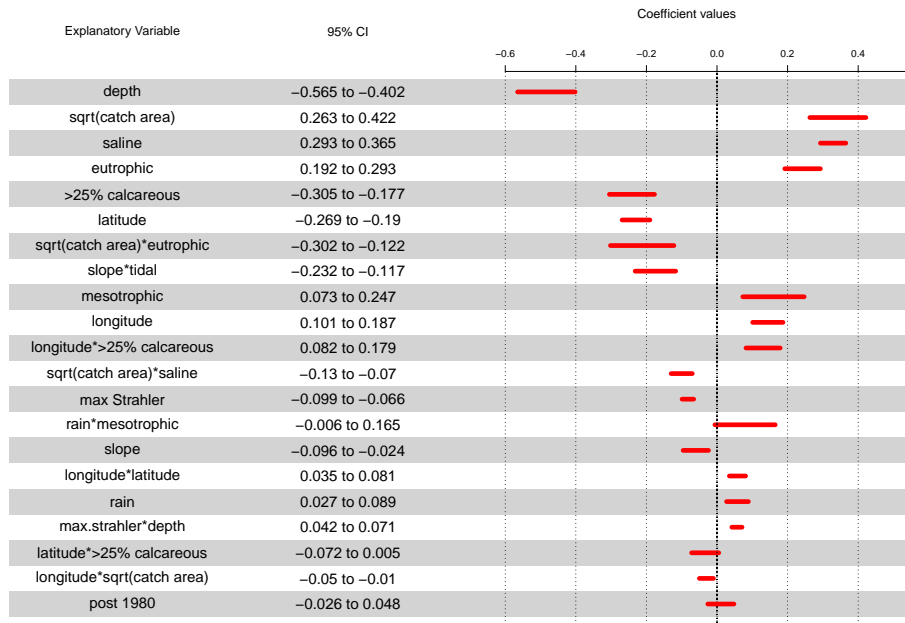


Figure 5: A graph of the final included explanatory variables for the OLM model with 95% credible intervals given numerically and pictorially.

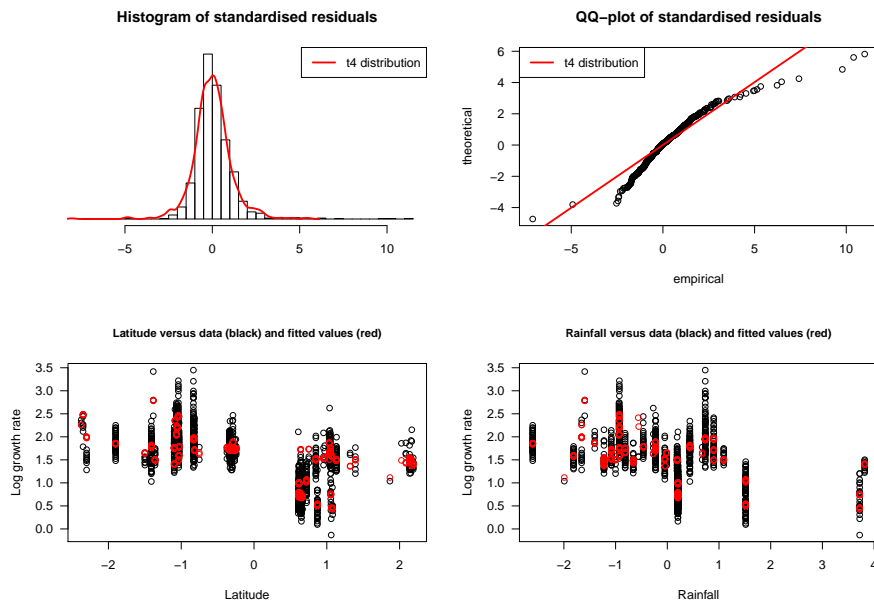


Figure 6: A four panel graph showing model performance for the OLM model. The top left panel shows a histogram of the standardised residuals together with a theoretical  $t_4$  distribution. The top right panel shows the QQ-plot of the standardised residuals. The bottom left panel shows the fitted values (red) and data values (black) against an explanatory variable included in the model. The bottom right panel shows the fitted values (red) against the data values (black) against an explanatory variable not included in the model.

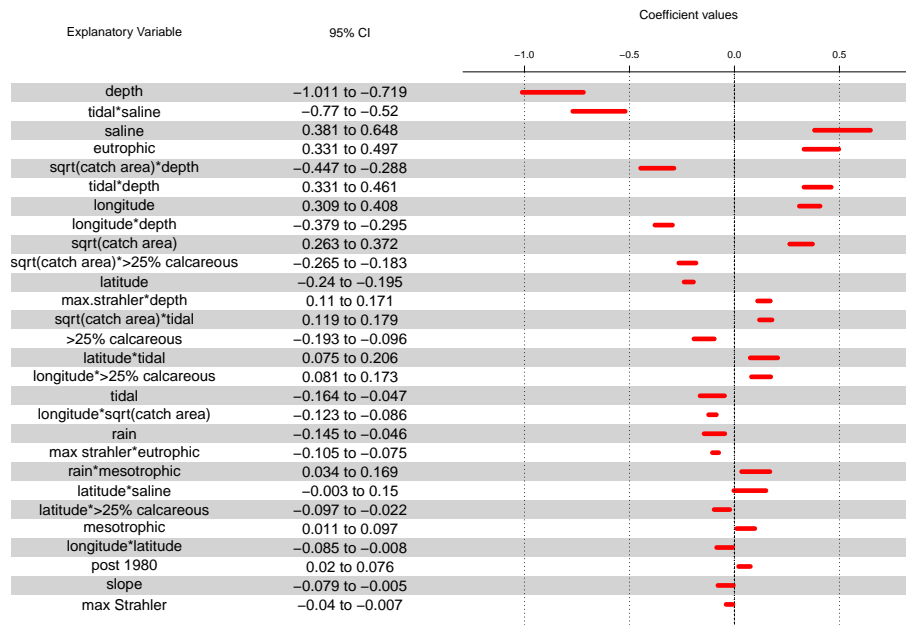


Figure 7: A graph of the final included explanatory variables for the OLF model with 95% credible intervals given numerically and pictorially.

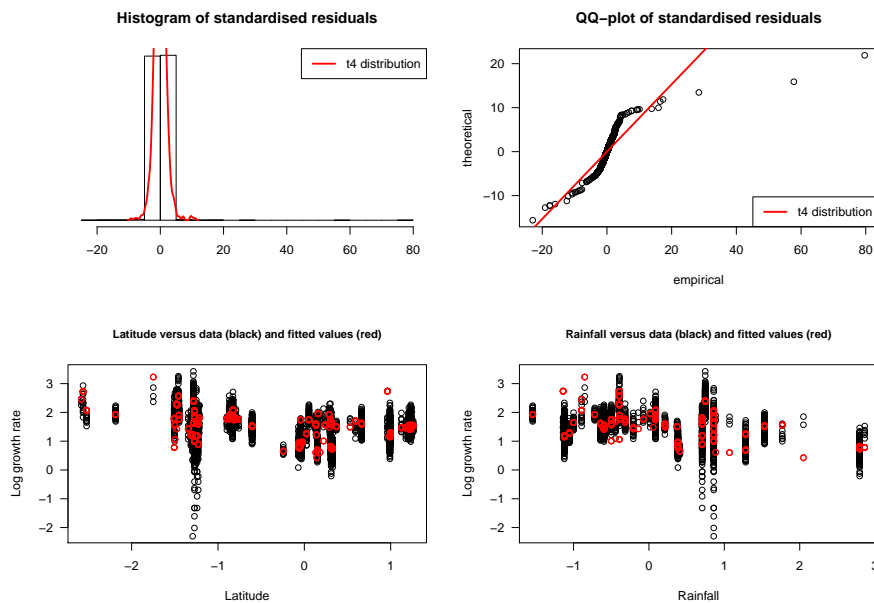


Figure 8: A four panel graph showing model performance for the OLF model. The top left panel shows a histogram of the standardised residuals together with a theoretical  $t_4$  distribution. The top right panel shows the QQ-plot of the standardised residuals. The bottom left panel shows the fitted values (red) and data values (black) against an explanatory variable included in the model. The bottom right panel shows the fitted values (red) against the data values (black) against an explanatory variable not included in the model.

#### 3.4.4 OLALL model results (DIC=6108, effective number of parameters =90)

The OLALL model used data from the OLF and OLM models, as well as those of undetermined sex. The final explanatory variables and interaction terms used are shown in Figure 9. Residuals were satisfactory, showing a few extreme observation and generally good predictive performance (Figure 10).

The two most extreme observations were from the UK and Spain (ID numbers 255 and 86 respectively). Observation 255 is based on 77 samples which have a very high growth rate (9.2) compared to other samples from UK (the mean of all UK observations is 2.055). Observation 86 is based on a very large number of samples (243) for which the model predicts the growth rate poorly.

### 3.5 Discussion

The growth rate models perform reasonably well both in terms of their predictive accuracy and theoretical properties. A number of key variables appear in all of the models in a consistent way (for example, the coefficient for depth is always large and negative). The extreme observations found by the models likely arise either due to a missing explanatory variable (classification as a lagoon has been identified as a possibility) or as erroneously recorded data.

As a measure of overall model performance, we can calculate the ratio of standard error to mean for each of the models (often written as the coefficient of variation or CV value). A large CV value indicates poor overall performance (the standard error is a large proportion of the mean) whereas a small value indicates relatively precise estimates. For the growth rate models, we have 17.1% (AS model), 25.4% (OLALL model), 18.2% (OLM model) and 15.6% (OLF model). The values indicate that the undetermined sex data included in the OLALL model causes a substantial deterioration in model performance.

Once incorporated into production estimates, it should be noted that the uncertainty in the estimation of growth rates will cascade into uncertainty on estimates of production. A key component in reducing production estimation uncertainty will be to improve the growth rate models. We cover this further in the discussion (Section 6).

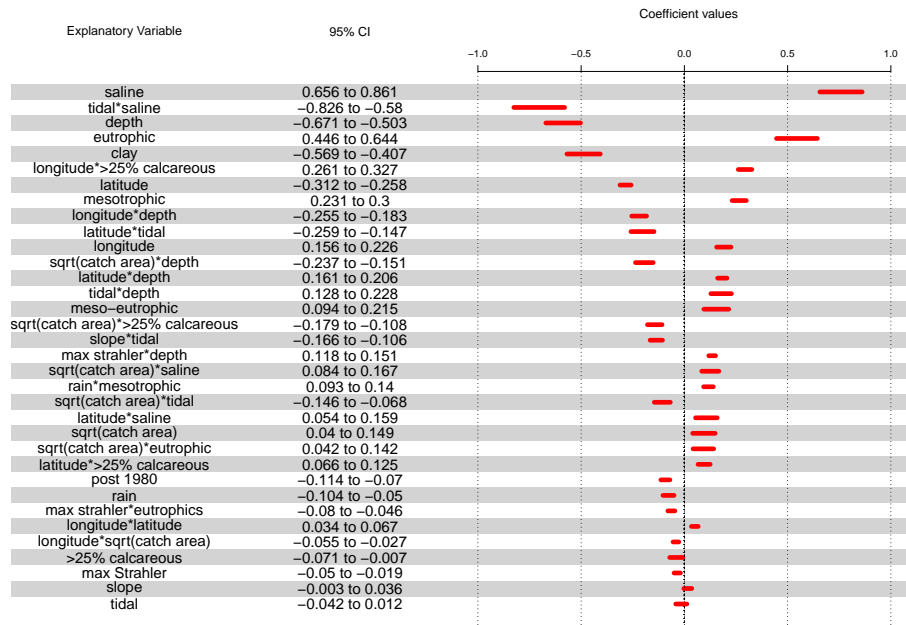


Figure 9: A graph of the final included explanatory variables for the OLALL model with 95% credible intervals given numerically and pictorially.

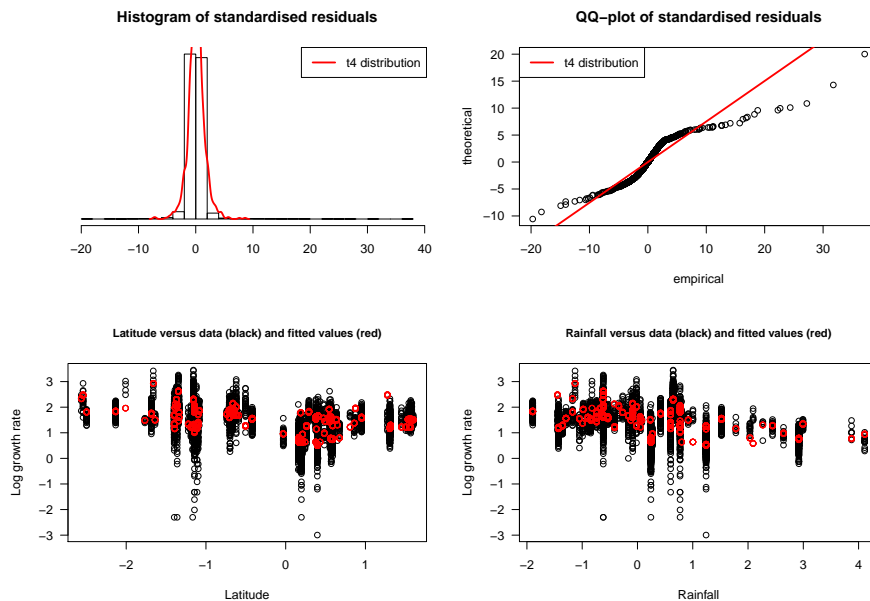


Figure 10: A four panel graph showing model performance for the OLALL model. The top left panel shows a histogram of the standardised residuals together with a theoretical  $t_4$  distribution. The top right panel shows the QQ-plot of the standardised residuals. The bottom left panel shows the fitted values (red) and data values (black) against an explanatory variable included in the model. The bottom right panel shows the fitted values (red) against the data values (black) against an explanatory variable not included in the model.

## 4 Mortality and survivorship

The mortality rate of eels affects the proportion of a recruitment cohort that survives to maturity and hence regulates production at a given site. In the simplest case, a fixed mortality rate, that remains constant for all eels over their lifetime can be defined. Traditionally, a discrete rate of 14% mortality per annum was used, which ignores any sex or mass dependent effects known to affect mortality (Bevacqua et al., 2011). In order to move to more complex models of mortality, it is easier to define instantaneous mortality rates ( $\mu$ ), and use calculus to explicitly model the dynamic nature of a system where the mass and hence mortality rate of individual eels changes over time. Specifically, constant percentage mortality rates per annum ( $R$ ) are related to instantaneous mortality rates ( $\mu$ ) by:

$$\mu = \log 1 - R/100 \quad (6)$$

By way of example, a mortality rate of 14% per annum equates to an instantaneous mortality rate of approximately  $\mu = 0.15$ . In order to calculate the proportion of individuals surviving to time  $S(t)$ , we specify the rate of change of  $S$  with respect to  $t$  as:

$$dS/dt = -\mu S \quad (7)$$

We then integrate with respect to time to yield:

$$S = e^{-\mu t} \quad (8)$$

Knowing then how many years an eel takes to reach maturity  $t$ , it is then straight forward to evaluate  $S(t)$ .

However, knowing that mortality rates vary with mass (Bevacqua et al., 2011), we must allow  $\mu$  to vary with respect to time and hence be influence by the estimated growth rate of eels. Eel mortality has been show to be mass dependent according to,

$$\mu = aM^b \quad (9)$$

where  $b$  is a scaling exponent (and takes the value -0.46, Bevacqua et al., 2011), and  $a$  is dependent on both gender and stock density according to

$$a = e^{-c/(273.15+T)+q} \quad (10)$$

where  $c$  takes the value 14176 for males, and 14395 for females;  $q$  is dependent on both gender and stocking density of eels and takes values described in [Table 4]; and  $T$  is the temperature in degrees Celsius of the water, with the constant 273.15 converting to degrees Kelvin.

The mass ( $M$ , measured in g) of an eel is dependent on time as the eel grows, and is related to its length ( $L$ , measured in cm) according to

	low	med	high
female	49.9	50.4	50.8
male	48.5	49.3	49.7

Table 2: Values of  $q$  taken from Bevacqua et al. (2011) by stocking density (for low, medium and high) and gender.

$$M = \alpha L^\beta \quad (11)$$

where  $\alpha = 8.34 \times 10^{-4}$ , and  $\beta = 3.17$ , Bevacqua et al. (2011). As is usual in the literature, yellow eel growth is assumed to be linear with respect to time and is given by

$$L = Gt + L(0) \quad (12)$$

where  $L(0)$  is the initial length of glass eels entering the system. Then, combining Equations 7 to 12 and integrating over time we obtain:

$$S = \exp\left(\frac{L(0)a(\alpha L(0)^\beta)^b}{G(b\beta + 1)}\right) \exp\left(\frac{-(Gt + L(0))a(\alpha(Gt + L(0))^\beta)^b}{G(b\beta + 1)}\right) \quad (13)$$

where the first exponential term on the right hand side is a scaling term (independent of  $t$ ) to ensure that  $(S(t = 0)) = 1$ , and the second exponential term determines the shape of the curve. Owing to the declining instantaneous mortality rate with mass, and hence time, the survival curve for this more complex model is inherently steeper early in the yellow eel's life than a model assuming a fixed mortality rate, but slows considerably for older eels (Figure 11).

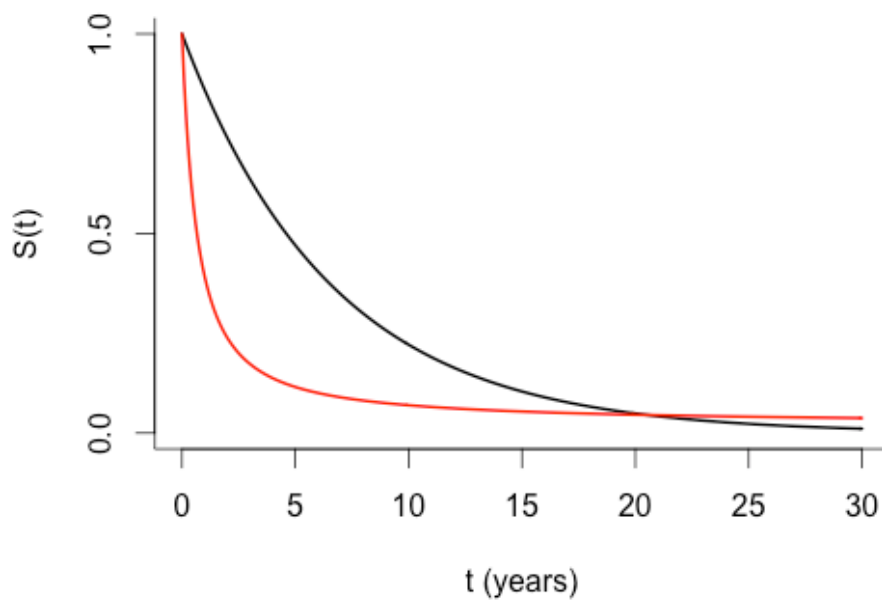


Figure 11: The proportion of individuals surviving as a function of time in years ( $S$ ). The black line assumes a fixed mortality rate of  $\mu = 0.1508$ . The red line depicts survival for a dynamic mortality rate for a male eel, at intermediate stocking density, an initial length of 7cm, a growth rate of 4.9 cm per annum (the mean for our entire dataset), in water at 14 degrees Celsius.



## 5 Production modelling

Returning to Equation 2 we use a calibration dataset comprising observations of eel production to estimate the relationship production  $P$  has with recruitment index  $R$ , survivorship  $S$  and mass at maturation  $M$ . The values of  $S$  and  $M$  are underpinned by the estimated  $G$  from the growth rate model described in section.

The mechanistic formula for deriving production given in Equation 2 was:

$$\log P = \log S + \log N + \log M - \log A$$

In this section we describe a model which estimates production from growth rate (and hence survivorship  $S$ ), recruitment index ( $R$ , a transformation of  $N$ ), mass ( $M$ ) and area ( $A$ ). We again use Bayesian modelling techniques, though we advise caution in the use of this part of the model as there are so few data. We have just 29 observations with sex ratios and 51 without.

### 5.1 Model details

We expand equation 2 to write:

$$\log P_i = \alpha + \beta \log R_i + q_i^{(f)} (\log S_i^{(f)} + \log M_i^{(f)}) + q_i^{(m)} (\log S_i^{(m)} + \log M_i^{(m)}) - \log A_i + \epsilon_i \quad (14)$$

for each data point  $i = 1, \dots, n$ , with  $\epsilon_i \sim N(0, \sigma^2)$ . Since most of the underlying key components of the model are known *a priori* to be gender dependent, we superscript all parameters with  $(m)$  and  $(f)$  to indicate the sex to which they are applied. (Where values do not have a superscript we are referring to the collection of male and female values together.) The proportion of females in each population is denoted by  $0 < q_i^{(f)} < 1$  with the corresponding proportion of males given by  $q_i^{(m)} = 1 - q_i^{(f)}$ . Mathematically Equation 14 performs a weighted sum of the fixed offsets to predict production from the recruitment index.

In fact, we find the model in Equation 14 to be inflexible in explaining variability in production and prefer a simpler explanatory model:

$$\log P_i = \alpha + \beta_1 \log R_i + \beta_2 q_i^{(f)} + \beta_3 \log S_i^{(f)} + \epsilon_i \quad (15)$$

Such a model seems more appropriate when taking into account uncertainties in Survival and Mass themselves (both being estimates from external parameters). Including all such terms in a linear model here will fail due to the high collinearity between Mass, Area and Survival, and between the sexes. We prefer to work with just Survivorship because it is the best understood of the covariates, and only with females for consistency (Equation 15 could equally be re-written for males and would produce near identical parameter estimates).

The sex ratios  $q_i^{(f)}$  are not always available, and so are estimated in such cases as part of the modelling framework. We write  $q_j^{(-)}$  for an observation with a missing sex ratio, and fit a second model to the missing sex ratio data:

$$\log P_j = \alpha + \beta_1 \log R_j + \beta_2 q_j^- + \beta_3 \log S_j^{(f)} + \epsilon_j \quad (16)$$

for  $j = 1, \dots, n_{miss}$ . The second model will now estimate the sex ratios, and yet still provide precision in the estimation of the other parameters.

For ease of notation, we define  $\Theta = (\alpha, \beta_1, \beta_2, \beta_3, \sigma^2, q^{(-)})$  to be the full set of unknown parameters to be estimated for the production model. Recall that  $Z$  represents the explanatory variables used to estimate growth rate. We write the posterior as:

$$p(\Theta|R, P, S^{(f)}, q^{(f)}, Z) \propto p(P|\Theta, R, P, S^{(f)}, q^{(f)}, Z) p(\Theta) \quad (17)$$

The likelihood relies on female survivorship ( $S^{(f)}$ ) which, as stated in Section 1 is a function of growth and length at maturity. We write  $\hat{G}$  as the predicted growth and  $\hat{L}$  for the predicted length at maturity for the observations in the production data set. The likelihood can thus be re-written as:

$$\begin{aligned} p(P|\Theta, R, P, S^{(f)}, q^{(f)}, Z) &= p(P|\Theta, R, P, \hat{L}^{(f)}, \hat{G}^{(f)}, q^{(f)}, Z) \\ &= \int_{\hat{G}^{(f)}, \hat{L}^{(f)}} p(P, \hat{G}^{(f)}, \hat{L}^{(f)}|\Theta, R, \Omega, Z) d\hat{G}^{(f)} d\hat{L}^{(f)} \\ &= \int_{\hat{G}^{(f)}, \hat{L}^{(f)}} p(P|S^{(f)}, \Theta, R) \times p(\hat{L}^{(f)}|\hat{G}^{(f)}, \Omega) \\ &\quad \times p(\hat{G}^{(f)}|Z) d\hat{G}^{(f)} d\hat{L}^{(f)} \end{aligned} \quad (18)$$

The likelihood for the model thus requires us to sample a set of female growth rates for the production data (a collection of  $\hat{G}^{(f)}$ ), and hence sample female lengths at maturity ( $\hat{L}^{(f)}$ ) and female survivorships ( $\hat{S}^{(f)}$ ). From these we can calculate a likelihood for production and hence fit a model.

We find the expression for the likelihood in Equation 18 to be overly complex given the small amount of data available. We thus avoid the numerical integration by taking an empirical Bayes approach (Carlin and Louis, 2000) and estimating the parameters in  $\Theta$  based on fitted values for  $\hat{G}^{(f)}$  and  $\hat{L}^{(f)}$ . After the model is fitted, however, the predictions for production include uncertainties in growth rate and length at maturity.

## 5.2 Prior distributions

As with the previous Bayesian models, prior distributions are required for the parameters in  $\Theta$ . In the case of regression intercept, slope and variance parameters we use Jeffreys prior distributions (Lee, 2004). We do not place a regularisation prior here as there is no discussion with regard to the inclusion of explanatory variables. The  $Beta(0.5, 0.5)$  reference prior distribution (being constrained between 0 and 1) is used as a prior distribution on the  $q_i^{(f-)}$  proportional parameters. More informative priors could be used here, especially with reference to the variations among parameters for females and/or males. It is likely that these are linked, and the correlation may be of interest. At this stage, however, we do not pursue this approach.

## 5.3 Model fitting

The final model was fitted via Markov chain Monte Carlo using the JAGS software (Plummer, 2011). The models were run for 100,000 iterations with a burn-in period of 20,000 and subsequent thinning by a factor of 80. We run two chains with over-dispersed (more extreme than we would expect) starting values to ensure convergence and check convergence through trace plots, autocorrelation plots and the Brooks-Gelman-Rubin diagnostic. We again omit the convergence diagnostics for reasons of brevity.

## 5.4 Results

The values of the coefficients for the production model are shown in Figure 12. None of the explanatory variables appear to exhibit a particularly strong relationship with production. However, recruitment index seems to be broadly positive, and female survivorship appears broadly negative. The last finding may seem counter-intuitive (an increase in female survivorship decreases production), however recall that female survivorship is strongly correlated with mass at maturity and area of catchment. The value here is being driven by all three of these variables. The sex ratio appears highly uncertain in its relationship with production. It has been left in the present model because it was found to increase the predictive variability in line with the data.

Residual and fitted values plots are shown in Figure 13. With such a small data set, it is hard to comment on general performance of the model but it seems adequate. The fitted values plots show that the model has matched the mean relationships present in the data.

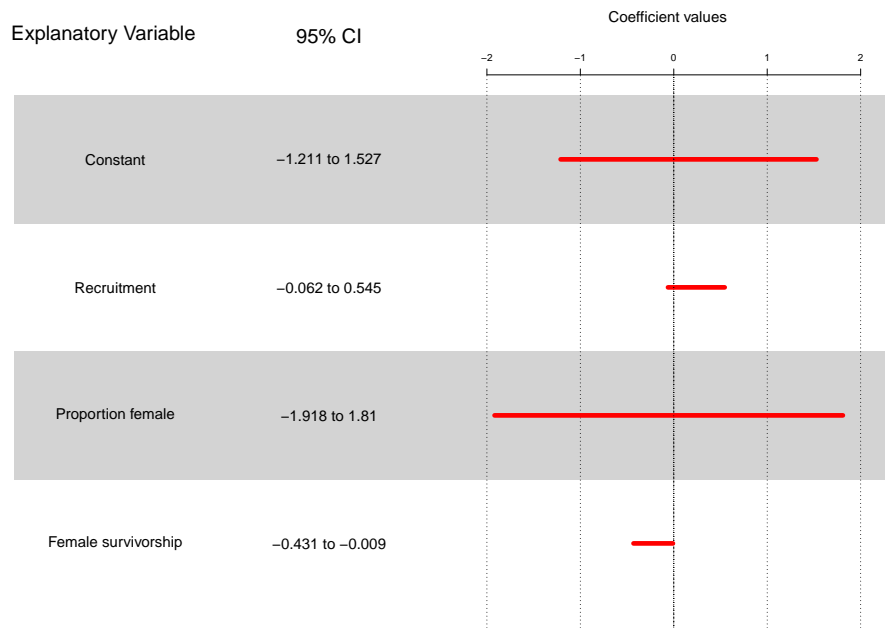


Figure 12: A graph of the final included explanatory variables for the production model with 95% credible intervals given numerically and pictorially.

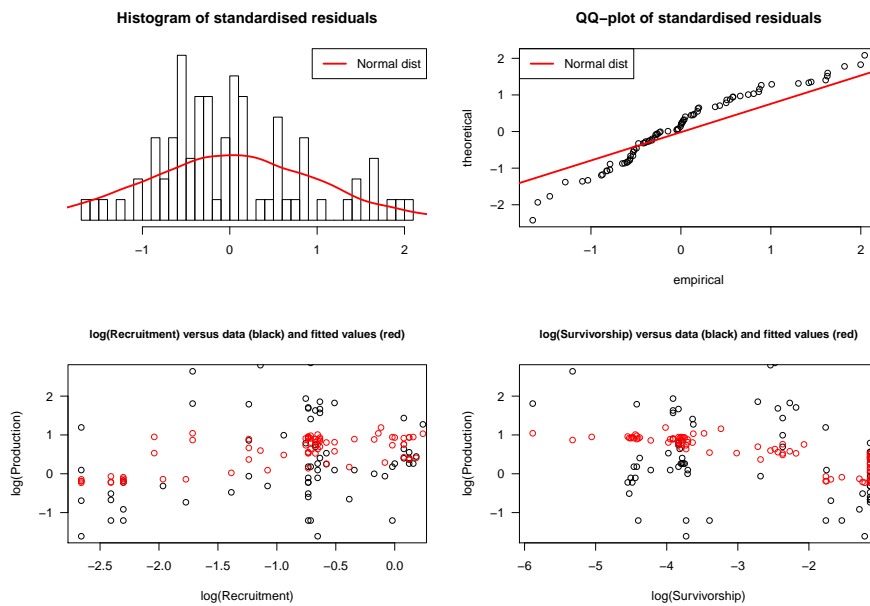


Figure 13: A four panel graph showing model performance for the production model. The top left panel shows a histogram of the standardised residuals together with a theoretical normal distribution. The top right panel shows the QQ-plot of the standardised residuals. The bottom left panel shows the fitted values (red) and data values (black) against log recruitment. The bottom right panel shows the fitted values (red) against the data values (black) against log survivorship.

## 6 Discussion

The work above outlines a method for estimating eel production based on explanatory variables via the estimation of growth rate and survivorship. There is a large set of data available for estimating growth rate; this seems the most robust part of our modelling framework. Conversely both the survivorship model and the production suffer from a shortage of adequate data. The models we suggest for survivorship are strongly based on Bevacqua et al. (2011), but this work lacks uncertainty estimates on many of the key parameters. Similarly the production models we have developed are based on very small amounts of data, some of which have key variables missing (e.g. sex ratio). The prediction intervals we thus provide for production are highly uncertain. Worse, because the uncertainty in the survivorship parameters is not included, the uncertainty is possibly an under-estimate. They are, however, our most honest guess at the production for the site given by users. In particular, 95% credible intervals for the mean production can be obtained with reasonable precision.

A major concern with the production model is that the mechanistic approach for estimating production (Equation 14) does not produce any better (in terms of accuracy and precision) predictions than the phenomenological model we arrived at in Section 5.1. Given that the mechanistic model is mathematically consistent with the units of production as given and therefore is unlikely to be incorrect in form, the mis-match must arise elsewhere. This means that at least one (and most likely more than one) of the variables used in the production model are incorrectly specified. The most obvious source is the estimated recruitment index; a more complicated bias adjustment than that used here may be necessary (though again we are hampered by the lack of calibration data at present). Other problems might include biases in the estimation of production for the calibration data set, or that of derived growth and hence survivorship. On a positive note, however, that there is some evidence of relationships between recruitment index ( $I$ ) and production ( $P$ ), and this suggests there is at least some explanatory power in the model.

Lastly, we outline possible recommendations for future modelling of eel production. The most obvious of these is an increase in the amount of production calibration data. Without more data here it remains extremely difficult to estimate the relationship between growth, survivorship and production. With further data, a number of advanced analysis options become available which may allow for far superior precision in the estimation of production. If data were available on an increased spatial scale, kriging methods (eg Banerjee et al., 2004) may become more appropriate (they were attempted for the growth rate modelling stage here, but were found insufficiently superior to the  $t$ -regression modelling). With increased data on a temporal scale, state-space approaches (eg Pole et al., 1994) will allow for dynamic tracking of production over time, removing noise from individual catchments. As a final note, we recommend perhaps focussing on catchment level relationships at fewer sites with a greater focus on understanding the relationships within a catchment. This would add levels of hierarchy to the Bayesian model and allow a more complete understanding of the important dynamic relationships.

## References

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2004). *Hierarchical modelling and analysis for spatial data*. Chapman and Hall.

- Bevacqua, D., P. Melia, G. D. Leo, and M. Gatto (2011). Intra-specific scaling of natural mortality in fish: the paradigmatic case of the european eel. *Oecologia* 165, 333–339.
- Birkes, D. and Y. Dodge (1993). *Alternative Methods of Regression*. Wiley.
- Brooks, S. P. and A. Gelman (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of computational and graphical statistics* 7, 434–455.
- Carlin, B. P. and T. A. Louis (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall.
- Department of Communications, Energy and National Resources (2009). *National Report for Ireland on Eel Stock Recovery Plan*. Department of Communications, Energy and National Resources.
- Jessop, B. (2011). Geographic effects on american eel (*anguilla rostrata*) life history data. *Canadian Journal of Fisheries and Aquatic Sciences* 67(2), 326–346.
- Lee, P. M. (2004). *Bayesian Statistics: An Introduction*. Oxford, England: Oxford.
- Plummer, M. (2011). *rjags: Bayesian graphical models using MCMC*. R package version 2.2.0-4.
- Pole, A., M. West, and J. Harrison (1994). *Applied Bayesian Forecasting and Time Series Analysis*. Chapman and Hall.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society (Series B)* 64, 583–639.
- Venables, W. and B. Ripley (2003). *Modern Applied Statistics with S*. Springer.