

# VAST model structure and user interface

James Thorson

## Purpose of document:

R package VAST includes many different forms of documentation including:

1. Doxygen documentation that can be accessed through the standard R-help interface when the library is loaded within R;
2. the VAST decision tree and user guide (currently in review at Fisheries Research)
3. two separate Rmarkdown “tutorials” that provide annotated code illustrating how to run VAST for single- or multi-species example using real-world data;
4. a searchable “issue tracker” available through GitHub; and
5. peer-reviewed articles describing development and applications for each feature (see list on GitHub).

This “VAST model structure and user interface” document is intended to complement these other resources by documenting and describing the model structure (all model equations and notation) while linking it to user-options that are available via the R interface to package VAST.

## Model description:

In the following, I use mathematical notation similar to the C++ code used to define the model in TMB: I use parentheses to indicate a parameter or variable that is indexed by the specified indices, and I use subscripts for naming (e.g., to indicate different parameters for different model components). Notation is close to common recommendations, e.g., Edwards and Auger-Méthé (2019), although I use parentheses to indicate indices of vectors,

matrices, and arrays, and reserve subscripts for naming. Feel free to change notation when describing the model to suit your purposes in reports or publications. For further details regarding terminology, motivation, and statistical properties, please read the papers listed on the GitHub main page.

## Overview

VAST predicts variation in density across multiple locations  $s$ , time intervals  $t$ , for multiple categories  $c$ . Categories could include either multiple species, and/or multiple size/age/sex classes for each individual species. VAST approximates the covariance between these multiple factors using a factor-model decomposition (Thorson et al. 2015a, 2016a), i.e., by summing across the contribution of multiple random effects (termed factors). If there is only a single category, the model reduces to a standard univariate spatio-temporal model.

After estimating variation in density across space, time, and among categories, VAST then predicts total abundance across a user-specified spatial domain. This is equivalent to an “area-weighting” approach to index standardization, and the resulting prediction of total abundance can be used as an index of abundance.

In addition to spatial and spatio-temporal covariance among multiple categories, VAST allows users to specify either density or catchability covariates. Both explain variation in observed catch-rate data, but VAST predicts density (for use in calculating the abundance index) using density covariates but not catchability covariates. Therefore, VAST “controls for” catchability covariates when calculating an index (i.e., removes their estimated effect) while “conditioning on” density covariates when calculating an index (i.e., uses them to improve interpolated/extrapolated predictions of density).

## Linear predictors

The model potentially includes two linear predictors (because it is designed to support delta-models, which include two components). The first linear predictor  $p_1(i)$  represents encounter probability in a delta-model, or zero-inflation in a count-data model:

$$\begin{aligned}
 p_1(i) = & \underbrace{\sum_{f=1}^{n_{\beta_1}} L_{\beta_1}(c_i, f) \beta_1(t_i, f)}_{\text{Temporal variation}} + \underbrace{\sum_{f=1}^{n_{\omega_1}} L_{\omega_1}(c_i, f) \omega_1(s_i, f)}_{\text{Spatial variation}} + \underbrace{\sum_{f=1}^{n_{\varepsilon_1}} L_{\varepsilon_1}(c_i, f) \varepsilon_1(s_i, f, t_i)}_{\text{Spatio-temporal variation}} \\
 & + \underbrace{\sum_{f=1}^{n_{\eta_1}} L_1(c_i, f) \eta_1(v_i, f)}_{\text{Vessel effects}} + \underbrace{\sum_{p=1}^{n_p} (\gamma_1(c_i, t_i, p) + \sigma_1(c, p) \xi_1(s, c, p)) X(x_i, t_i, p)}_{\text{Habitat covariates}} \\
 & + \underbrace{\sum_{k=1}^{n_k} \lambda_1(k) Q(i, k)}_{\text{Catchability covariate}}
 \end{aligned}$$

where  $p_1(i)$  is the predictor for observation  $i$ ,  $\beta_1(t_i, f)$  represents temporal variation for time  $t_i$  for factor  $f$  (of  $n_{\beta_1}$  factors representing temporal variation), and  $L_{\beta_1}(c_i, f)$  is the loadings matrix that generates temporal covariation among categories for this linear predictor. The number of factors  $n_{\beta_1}$  can range from zero to the number of categories  $n_c$ ,  $0 \leq n_{\beta_1} \leq n_c$ , where  $n_{\beta_1} = 0$  is equivalent to eliminating all temporal terms from the model. By default,  $n_{\beta_1} = n_c$ ,  $\beta_1(t, f)$  is treated as a fixed effect for each year  $t$  and factor  $f$ , and  $\mathbf{L}_{\beta_1}$  is an identity matrix; this formulation is equivalent to estimating a separate intercept  $\beta_1(t_i, c) = \beta_1(t_i, f)$  for each category and year. However, the intercepts can instead be treated as a random effect using the factor-model formulation, which allows for sharing information among years and categories. When treated as random,  $\beta_1(t_i, f)$  is assigned a normal distribution with unit variance, such that  $\mathbf{L}_{\beta_1}^T \mathbf{L}_{\beta_1}$  is the covariance among categories for a given process (Thorson et al. 2015a). When treating intercepts as random, and when there is only one category and using one factor ( $n_{\beta_1} = 1$ ), then  $\mathbf{L}_{\beta_1}$  is a 1x1 matrix (i.e. a scalar) such

68  $\mathbf{L}_{\beta 1}^2$  is the variance and the absolute value,  $abs(\mathbf{L}_{\beta 1})$  is the standard deviation for temporal  
 69 variation.

70 Similarly,  $\omega_1(s_i, f)$  represents spatial variation at location  $s_i$  for factor  $f$  (of  $n_{\omega 1}$   
 71 factors representing spatial variation), and  $L_{\omega 1}(c_i, f)$  is the loadings matrix that generates  
 72 spatial covariation among categories for this linear predictor. Similarly,  $\varepsilon_1(s_i, f, t_i)$   
 73 represents spatio-temporal variation for each factor  $f$  (of  $n_{\varepsilon 1}$  factors representing spatio-  
 74 temporal variation), and  $L_{\varepsilon 1}(c_i, f)$  is the loadings matrix that generates spatio-temporal  
 75 covariation for this predictor.  $\eta_1(v_i, f)$  represents random variation in catchability among a  
 76 grouping variable (tows or vessels) for each factor  $f$  (of  $n_{\eta 1}$  factors representing  
 77 overdispersion), and  $L_1(c_i, f)$  is a loadings matrix that generates covariation in catchability  
 78 among categories for this predictor. All loadings matrices are specified similarly to  $\mathbf{L}_{\beta 1}$ , i.e.,  
 79 where factors have a variance of one such that  $\mathbf{L}^T \mathbf{L}$  represents the covariance among species.  
 80 The main difference is that spatial, spatio-temporal, and overdispersion factors can only be  
 81 specified as random effects, while the intercepts can be specified as either random or fixed  
 82 (where specifying as fixed “turns off” all factor-modelling for that intercept).

83 Finally,  $X(x_i, t_i, p)$  is an array of  $n_p$  measured density covariates that explain  
 84 variation in density for time  $t$  and knot  $x$ . VAST can include a separate, spatially-varying  
 85 effect of each habitat covariate  $p$  for each category  $c$ . The spatially varying slope is  
 86  $\gamma_1(c_i, t_i, p) + \sigma(c, p)\xi_n(s, c, p)$ , where  $\gamma_1(c_i, t_i, p)$  is the average effect of density covariate  
 87  $X(x_i, t_i, p)$  for category  $c$ ,  $\xi_n(s, c, p)$  represents spatial variation in that effect (which has a  
 88 mean of zero and standard deviation of one), and  $\sigma(c, p)$  represents the estimated standard  
 89 deviation of spatial variation of covariate  $p$  for category  $c$ .  $Q(i, k)$  is a matrix of  $n_k$   
 90 measured catchability covariates that explain variation in catchability, and  $\lambda_1(k)$  is the  
 91 estimated impact of catchability covariates for this linear predictor. By default, VAST

specifies that  $\gamma_1(c, t_1, p) = \gamma_1(c, t_2, p)$  for all years  $t_1$  and  $t_2$ , although users can relax this constraint by specifying a different structure for `Data_Fn(..., Map=NewMap)`.

Similarly, the second linear predictor  $p_2(i)$  represents positive catch rates in a delta-model, or the count-data intensity function in a count-data model:

$$\begin{aligned}
 p_2(i) = & \underbrace{\sum_{f=1}^{n_{\beta 2}} L_{\beta 2}(c_i, f) \beta_2(t_i, f)}_{\text{Temporal variation}} + \underbrace{\sum_{f=1}^{n_{\omega 2}} L_{\omega 2}(c_i, f) \omega_2(s_i, f)}_{\text{Spatial variation}} + \underbrace{\sum_{f=1}^{n_{\varepsilon 2}} L_{\varepsilon 2}(c_i, f) \varepsilon_2(s_i, f, t_i)}_{\text{Spatio-temporal variation}} \\
 & + \underbrace{\sum_{f=1}^{n_{\eta 2}} L_2(c_i, f) \eta_2(v_i, f)}_{\text{Vessel effects}} + \underbrace{\sum_{p=1}^{n_p} (\gamma_2(c_i, t_i, p) + \sigma_2(c, p) \xi_2(s, c, p)) X(x_i, t_i, p)}_{\text{Density covariates}} \\
 & + \underbrace{\sum_{k=1}^{n_k} \lambda_2(k) Q(i, k)}_{\text{Catchability covariates}}
 \end{aligned}$$

where all variables and parameters are defined similarly except using different subscripts (Thorson et al. In press, Thorson and Barnett 2017).

## Number of spatial and spatio-temporal factors

The user controls the number of spatial and spatio-temporal factors used for each component via input:

```
FieldConfig = c("Omega1"=1, "Epsilon1"=1, "Omega2"=1, "Epsilon2"=1)
```

where `FieldConfig[1]` controls  $n_{\omega 1}$ , `FieldConfig[2]` controls  $n_{\varepsilon 1}$ , `FieldConfig[3]` controls  $n_{\omega 2}$ , and `FieldConfig[4]` controls  $n_{\varepsilon 2}$ , and a value of zero “turns off” that component of spatial or spatio-temporal covariation.

## Number of overdispersion factors

112 The user controls the number of catchability factors used for each component via input:

```
113 OverdispersionConfig = c("Eta1"=0, "Eta2"=0)
```

114

115 where `OverdispersionConfig[1]` controls  $n_{\delta_1}$ , and `OverdispersionConfig[2]` controls  $n_{\delta_2}$ ,

116 and a value of zero again “turns off” that component of random covariation in catchability.

117 For example, if the user inputs:

```
118 OverdispersionConfig = c("Eta1"=1, "Eta2"=1)
```

119

120 then there will be one random effect estimated for each unique level of `Data_Geostat$Vessel`

121 for both the first and second linear predictors.

122

### 123 **Link functions and observation error distributions**

124 There are user-controlled options that control the observation error distribution and the link-  
125 functions used to calculate expected encounter probabilities and positive catch rates based on  
126 the two linear predictors.

127 The `ObsModel` vector has two components, controlling the observation error distribution and  
128 link function respectively.

```
129 ObsModel = c("PosDist"=2, "Link"=0)
```

130 There are currently four options for the link function. For the latest set of options see the R  
131 help documentation by typing into the R terminal `?VAST::Data_Fn``.

132 1. `ObsModel[2]=0` applies a logit-link for the first linear predictor:

$$133 \quad r_1(i) = \text{logit}^{-1}(p_1(i))$$

134 where  $r_1(i)$  is the predictor encounter probability in a delta-model, or zero-inflation in a  
135 count-data model, and  $\text{logit}^{-1}(p_1(i))$  is the inverse-logit (a.k.a. logistic) function of  
136  $p_1(i)$ , and:

$$137 \quad r_2(i) = a_i \times \log^{-1}(p_2(i))$$

where  $r_2(i)$  is the predicted biomass density for positive catch rates in a delta-model or mean-intensity function for a count-data model,  $\log^{-1}(p_2(i))$  is the exponential function of  $p_2(i)$ , and  $a_i$  is the area-swept for observation  $i$ , which enters as a linear offset for expected biomass given an encounter.

2. `ObsModel[2]=1` corresponds to a “Poisson-link” delta-model that approximates a Tweedie distribution:

$$r_1(i) = 1 - \exp(-a_i \times \exp(p_1(i)))$$

where  $r_1(i)$  is the predictor encounter probability and  $1 - \exp(-a_i \times \exp(p_1(i)))$  is a complementary log-log link of  $p_1(i) + \log(a_i)$ , and:

$$r_2(i) = \frac{a_i \times \exp(p_1(i))}{r_1(i)} \times \exp(p_2(i))$$

where  $r_2(i)$  is the predicted biomass given that the species is encountered. In this “Poisson-process” link function,  $\exp(p_1(i))$  is interpreted as the density in number of individuals per area such that  $a_i \times \exp(p_1(i))$  is the predicted number of individuals encountered, and  $\exp(p_2(i))$  is interpreted as the average weight per individual. Area-swept  $a_i$  therefore enters as a linear offset for the expected number of individuals encountered (Thorson 2018). This Poisson-link function should only be used for delta-models, and not for count-data models, but can also be used to combine encounter, count, and biomass-sampling data (see section below for details).

### Observation models:

There are different user-controlled options for observation models for available sampling data, which are controlled by `ObsModel_ez[1]`.

```
# Control observation error
ObsModel_ez = c("PosDist"=2, "Link"=0)
```

I distinguish between observation models for continuous-valued data (e.g., biomass, or numbers standardized to a fixed area), and observation models for count data (e.g., numbers treating area-swept as an offset). However, both are parameterized such that the expectation for sampling data  $\mathbb{E}(B_i) = r_1(i) \times r_2(i)$ .

#### *Continuous-valued data (e.g., biomass)*

If using an observation model with continuous support (e.g., a normal, lognormal, gamma, or Tweedie models), then data  $b_i$  can be any non-negative real number,  $b_i \in \mathcal{R}$  and  $b_i \geq 0$ . VAST calculates the probability of these data as:

$$\Pr(b_i = B) = \begin{cases} 1 - r_1(i) & \text{if } B = 0 \\ r_1(i) \times g\{B|r_2(i), \sigma_m^2(c)\} & \text{if } B > 0 \end{cases}$$

where `ObsModel[1]` controls the probability density function  $g\{B|r_2(i), \sigma_m^2(c)\}$  used for positive catch rates (see `?Data_Fn` for a list of options), where each options is defined to have with expectation  $r_2(i)$  and dispersion  $\sigma_m^2(c)$ , where dispersion parameter  $\sigma_m^2(c)$  varies among categories by default.

#### *Discrete-valued data (e.g., abundance)*

If using an observation model with discrete support (e.g., a Poisson, negative-binomial, Conway-Maxwell Poisson, or lognormal-Poisson models), then data  $b_i$  can be any whole number,  $b_i \in \{0, 1, 2, \dots\}$ . VAST calculates the probability of these data as:

$$\Pr(B = b_i) = \begin{cases} (1 - r_1(i)) + g\{B = 0|r_2(i), \dots\} & \text{if } B = 0 \\ r_1(i) \times g\{B = b_i|r_2(i), \dots\} & \text{if } B > 0 \end{cases}$$

where `ObsModel[1]` controls the probability mass function  $g\{B|r_2(i), \dots\}$  used (again, see `?Data_Fn` for a list of options), where I use  $\dots$  to signify that these probability mass functions generally can have one or more parameter governing dispersion, and the precise number and interpretation varies among observation models (i.e., the value of `ObsModel[1]`). For these count-data models,  $(1 - r_1(i))$  is the “zero-inflation probability” (i.e., the proportion of habitat in the immediate vicinity of location  $s_i$  and time  $t_i$  that is never occupied), while  $r_2(i)$



is the expected value for probability mass function  $g\{B = b_i | r_2(i), \dots\}$  (i.e., the number of individuals that are in the vicinity of sampling in habitat that is occupied), and  $g\{B = 0 | r_2(i), \dots\}$  is the probability of not encountering category  $c$  given that sampling occurs in occupied habitat (Martin et al. 2005).

## Settings regarding spatial domain

VAST approximates spatial and spatio-temporal variation as being piecewise-constant. To do so, the user specifies a number of knots  $n_x$ :

```
# Number of knots
n_x = 1000
```

VAST then uses a k-means algorithm to identify the location of  $n_x$  knots to minimize the total distance between the location of available data and the location of the nearest knot. This distributes knots as a function of the spatial intensity of sampling data.

VAST then uses a stochastic partial differential equation (SPDE) approximation to the probability density function for spatial and spatio-temporal variation (Lindgren et al. 2011). This SPDE approximation involves generating a triangulated mesh that has a vertex of a triangle at each knot, and VAST generates this triangulated mesh using package *R-INLA* (Lindgren 2012). Outputs from this triangulated mesh can then be used to calculate the precision (inverse-covariance) matrix for a multivariate normal probability density function for the value of a spatial variable at each mesh vertex. Specifically, the correlation  $\mathbf{R}_1(s, s + h)$  between location  $s$  and location  $s + h$  for spatial and spatio-temporal terms included in the first linear predictor is approximated as following a Matern function:

$$\mathbf{R}_1(s, s + h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \times (\kappa_1 |h\mathbf{H}|)^{\nu} \times K_{\nu}(\kappa_1 |h\mathbf{H}|)$$

where  $\mathbf{H}$  is a two-dimensional linear transformation representing geometric anisotropy (with a determinant of 1.0),  $\nu$  is the Matern smoothness (fixed at 1.0), and  $\kappa_1$  governs the

213 decorrelation distance for that first linear predictor ( $\kappa_2$  is also separately estimated for the  
 214 second linear predictor). By default, the two degrees of freedom in  $\mathbf{H}$  are estimated as fixed  
 215 effects, but the user can specify isotropy (i.e.,  $\mathbf{H} = \mathbf{I}$ ) by specifying:

```
216 # Turn of geometric anisotropy
217 Data = Data_Fn( ..., Aniso=FALSE )
218
```

219 VAST then specifies that the spatial and spatio-temporal Gaussian random fields each  
 220 have a variance of 1.0. By default VAST specifies these as follows:

$$221 \quad \boldsymbol{\omega}_1(\cdot, f) \sim MVN(\mathbf{0}, \mathbf{R}_1)$$

$$222 \quad \boldsymbol{\omega}_2(\cdot, f) \sim MVN(\mathbf{0}, \mathbf{R}_2)$$

$$223 \quad \boldsymbol{\varepsilon}_1(\cdot, f, t) \sim MVN(\mathbf{0}, \mathbf{R}_1)$$

$$224 \quad \boldsymbol{\varepsilon}_2(\cdot, f, t) \sim MVN(\mathbf{0}, \mathbf{R}_2)$$

$$225 \quad \boldsymbol{\xi}_1(\cdot, c, p) \sim MVN(\mathbf{0}, \mathbf{R}_1)$$

$$226 \quad \boldsymbol{\xi}_2(\cdot, c, p) \sim MVN(\mathbf{0}, \mathbf{R}_2)$$

227 where  $\boldsymbol{\omega}_1(\cdot, f)$  is the vector formed when subsetting  $\omega_1(s, f)$  for a given  $f$ . Specifying a  
 228 variance of 1.0 ensures that the covariance among categories is defined by the loadings  
 229 matrix for that term.

230

### 231 **Structure on parameters among years:**

232 There are different user-controlled options for specifying structure for intercepts or spatio-  
 233 temporal variation across time, using input:

```
234 RhoConfig = c("Beta1"=0, "Beta2"=0, "Epsilon1"=0, "Epsilon2"=0)
235
```

### 236 *Temporal structure on intercepts*

237 By default (when  $\text{RhoConfig}[1]=0$  and  $\text{RhoConfig}[2]=0$ ) the model specifies that each  
 238 intercept  $\beta_1(t)$  and  $\beta_2(t)$  is a fixed effect. However, other settings specify the following  
 239 structure:

$$\beta_1(t+1) \sim \text{Normal}(\rho_{\beta_1}\beta_1(t), \sigma_{\beta_1}^2)$$

$$\beta_2(t+1) \sim \text{Normal}(\rho_{\beta_2}\beta_2(t), \sigma_{\beta_2}^2)$$

where `RhoConfig[1]` controls the specification of  $\rho_{\beta_1}$ :

1. *Independent among years* – `RhoConfig[1]=1` specifies  $\rho_{\beta_1} = 0$
  2. *Random walk* – `RhoConfig[1]=2` specifies  $\rho_{\beta_1} = 1$
  3. *Constant intercept* – `RhoConfig[1]=3` specifies  $\rho_{\beta_1} = 0$  and  $\sigma_{\beta_1}^2 = 0$  (i.e.,  $\beta_1(t)$  is constant for all  $t$ )
  4. *Autoregressive* – `RhoConfig[1]=4` estimates  $\rho_{\beta_1}$  as a fixed effect
- and settings are defined identically for `RhoConfig[2]` specifying  $\rho_{\beta_2}$ .

249

250 *Temporal structure on spatio-temporal variation*

251 By default (when `RhoConfig[3]=0` and `RhoConfig[4]=0`), the model specifies that each spatio-  
 252 temporal random effect  $\varepsilon_1(s, f, t)$  and  $\varepsilon_2(s, f, t)$  is independent among years. However,  
 253 other settings specify the following structure

$$\varepsilon_1(s, f, t+1) \sim \text{MVN}(\rho_{\varepsilon_1}\varepsilon_1(s, f, t), \mathbf{R}_1)$$

$$\varepsilon_2(s, f, t+1) \sim \text{MVN}(\rho_{\varepsilon_2}\varepsilon_2(s, f, t), \mathbf{R}_2)$$

256 where `RhoConfig[3]` controls the specification of  $\rho_{\varepsilon_1}$ :

- 257 1. *Random walk* – `RhoConfig[3]=2` specifies  $\rho_{\varepsilon_1} = 1$
  - 258 2. *Autoregressive* – `RhoConfig[3]=4` estimates  $\rho_{\varepsilon_1}$  as a fixed effect
- 259 and settings are defined identically for `RhoConfig[4]` specifying  $\rho_{\varepsilon_2}$ .

260

## 261 **Parameter estimation**

262 Parameters are estimated using maximum likelihood, where the maximum likelihood of fixed  
 263 effects is obtained by integrating a joint likelihood function with respect to random effects

(Searle et al. 1992, Gelman and Hill 2007, Thorson and Minto 2015). This integral is approximated using the Laplace approximation (Skaug and Fournier 2006), as implemented in Template Model Builder (Kristensen et al. 2016). The likelihood is then optimized in the R statistical environment (R Core Team 2017), and standard errors are obtained using a generalization of the delta method (Kass and Steffey 1989). Derived quantities calculated via a nonlinear transformation of random effects can be bias-corrected using the epsilon-method (Tierney et al. 1989, Thorson and Kristensen 2016). Depending upon user-specified options, different parameters will be either fixed (estimated via maximizing the log-likelihood) or random (integrated across when calculating the log-likelihood). Please use R function ``ThorsonUtilities::list_parameters( Obj )`` to see a list of estimated parameters (where ``Obj`` is the compiled VAST object), including which are fixed or random.

## Combining multiple data types

VAST can be used to combine presence-absence, count, and biomass-sampling data. This involves specifying a Poisson-link delta model which predicts each data type from numbers density  $\exp(p_1(i))$  and biomass-per-individual  $\exp(p_2(i))$ , see Grüss and Thorson (In press) for details. This approach is specified by associating each observation with a given error distribution using input `e_i` where e.g. `e_i[1]` is the error-distribution for the 1<sup>st</sup> observation. The user then specifies multiple observation errors via input `ObsModel_ez`:

```
# Control observation error
ObsModel_ez = cbind( "PosDist"=c(13,14,2), "Link"=c(1,1,1) )
```

In this specification, `e_i[1]==1` indicates that the first observation follows a Bernoulli distribution for presence-absence data, `e_i[1]==2` indicates that this observation follows a lognormal-Poisson distribution for count data, and `e_i[1]==3` indicates that it follows a

gamma distribution for biomass-sampling data. This specification can be modified to include different combinations of these same data types.

## **Relationship to other named models**

VAST can be configured to be identical to (or closely mimic) many models that have previously been published in ecology and fisheries:

1. *Spatial Gompertz model*: If intercepts are constant across years, spatio-temporal variation follows an autoregressive process, and only one category is modelled, then VAST is identical to a spatio-temporal Gompertz model (Thorson et al. 2014).
2. *Spatial factor analysis*: If only one year is analysed and multiple categories are modelled, VAST is similar to spatial factor analysis (Thorson et al. 2015a), although it permits the use of a delta-model (i.e., separate analysis of encounters and positive catch rates).
3. *Spatial dynamic factor analysis*: If intercepts are constant among years, spatio-temporal variation follows an autoregressive process, and multiple categories are modelled, then VAST is similar to spatial dynamic factor analysis (Thorson et al. 2016a), although VAST allows separate estimates of spatial vs. spatio-temporal covariation and also the use of a delta-model.

## **Settings regarding derived quantities**

After a nonlinear minimizer has identified the value of fixed effects that maximizes the Laplace approximation to the marginal likelihood, Template Model Builder predicts the value of random effects that maximizes the joint likelihood conditional on these fixed effects. Estimated values of fixed and random effects are then used to predict density  $d(x, c, t)$  as follows:

$$d(x, c, t) = r_1^*(x, c, t) \times r_2^*(x, c, t)$$

314 where  $r_1^*(x, c, t)$  and  $r_2^*(x, c, t)$  are identical to the values specified previously, except that  
 315 catchability variables are excluded from their computation (i.e.,  $\eta_1(v, f) = 0$  and  $\lambda_1(k) = 0$ ,  
 316 etc.)

317 By default, density is used to predict total abundance for the entire domain (or a  
 318 subset of the domain) for a given species:

$$319 \quad I(c, t, l) = \sum_{x=1}^{n_x} (a(x, l) \times d(x, c, t))$$

320 where  $a(x, l)$  is the area associated with extrapolation-cell  $x$  for index  $l$ ; and  $n_x$  is the  
 321 number of extrapolation-cells (Shelton et al. 2014; Thorson et al. 2015b). The user can also  
 322 specify additional post-hoc calculations via the Options vector:

```
323 Options = c("SD_site_density"=0, "SD_site_logdensity"=0, "Calculate_Range"=0,  
324 "Calculate_evenness"=0, "Calculate_effective_area"=0, "Calculate_Cov_SE"=0,  
325 'Calculate_Synchrony'=0, 'Calculate_Coherence'=0)  
326
```

327 1. *Distribution shift* – RhoConfig[3]=1 turns on calculation of the centroid of the  
 328 population's distribution:

$$329 \quad Z(c, t, m) = \sum_{x=1}^{n_x} \frac{(z(x, m) \times a(x, 1) \times d(x, c, t))}{I(c, t, 1)}$$

330 where  $z(x, m)$  is a matrix representing location for each knot (by default  $z(x, m)$  is the  
 331 location in Eastings and Northings of each knot), representing movement North-South  
 332 and East-West). This model-based approach to estimating distribution shift can account  
 333 for differences in the spatial distribution of sampling, unlike conventional sample-based  
 334 estimators (Thorson et al. 2016b).

335 2. *Range expansion* – RhoConfig[5]=1 turns on calculation of effective area occupied. This  
 336 involves calculating biomass-weighted average density:

$$337 \quad D(c, t, l) = \sum_{x=1}^{n_x} \frac{a(x, l) \times d(x, c, t)}{I(c, t, l)} d(x, c, t)$$

Effective area occupied is then calculated as the area required to contain the population at this average density:

$$A(c, t, l) = \frac{I(c, t, l)}{D(c, t, l)}$$

This effective-area occupied estimator can then be used to monitor range expansion or contraction or density-dependent range expansion (Thorson et al. 2016c).

## List of features

I next provide a list of “features” organized as decisions that can be made by the analyst. Although this is somewhat redundant with the explanations provided above, this list might be useful for some readers to provide a high-level overview of different options that are available.

### *Basic features in a generalized linear model (GLM)*

1. Specifying one of several possible distributions for data;
2. Specifying one of several possible link functions for predicting data given linear predictors;
3. Including dynamic habitat covariates or not;
4. Including catchability covariates or not;

### *Basic features in a spatio-temporal generalized linear mixed model (GLMM)*

5. Specify an “extrapolation grid” using input `SpatialDeltaGLMM::Prepare_Extrapolation_Data_Fn(..., Region)`, which is used to calculate the area associated with each knot  $a_x$ . This can be a user-specified extrapolation grid if `SpatialDeltaGLMM::Prepare_Extrapolation_Data_Fn(..., Region="User", input_grid=Input)`, where `Input` is a data frame supplied by the user.
6. Specifying a method for defining “knots”;
7. Specifying the number of “knots”;

- 363 8. Spatial variation being estimated (“turned on”) or ignored (“turned off”) for either linear  
364 predictor #1 or #2;
- 365 9. Spatio-temporal variation being estimated (“turned on”) or ignored (“turned off”) for  
366 either linear predictor #1 or #2;
- 367 *Derived quantities*
- 368 10. Specifying strata for use when calculating derived quantities;
- 369 11. Calculating one of many possible “derived quantities”, including range shift, effective  
370 area occupied, abundance indices, covariance among categories within a multivariate  
371 model, or synchrony among categories.
- 372 *Non-standard decisions regarding temporal structure*
- 373 12. Annual intercepts being estimated as fixed effects in every year, fixed at the same value  
374 for all years, or estimated as a random effect with independent deviations in each year, a  
375 first-order autoregressive structure, or a random-walk structure.
- 376 13. Spatio-temporal variation being estimated as independent deviations in each year,  
377 following a first-order autoregressive structure over time, or following a random-walk  
378 structure over time.
- 379 *Multivariate analysis*
- 380 14. Including a “multivariate” structure with multiple responses that covary due to a specified  
381 number of “factors” for spatial and spatio-temporal terms;
- 382 15. Rotate results prior to interpretation, using either principle components rotation or  
383 varimax rotation;
- 384 *Unusual circumstances and spatial cases*
- 385 16. Specifying separate distributions for different data sets (e.g., when multiple surveys are  
386 available);



17. Specifying that some data are predicted based on summing linear predictors across multiple variables (e.g., when modelling density for different size classes, and specifying that some data are aggregated measurements of multiple sizes-classes);

18. Specifying multiple “seasons” (e.g., when modelling data with both annual and monthly spatio-temporal variation).

## Common problems

There are two basic problems that are often encountered during spatio-temporal delta-GLMMs:

1. *Encounter rates*: Some combination of categories and year has 0% or 100% encounter rate. If there is 100% encounter rate for category  $c$  in year  $t$ , then  $\beta_p(c, t) \rightarrow \infty$  and/or  $\varepsilon_p(s, c, t) \rightarrow \infty$  for that year. If there is 0% encounter rate in year  $t$ , then  $\beta_p(c, t) \rightarrow -\infty$  and/or  $\varepsilon_p(s, c, t) \rightarrow -\infty$  and there is no information to estimate  $\beta_r(c, t)$  or  $\varepsilon_r(s, c, t)$  for that category  $c$  and year  $t$ ;
2. *Bounds*: Some parameter(s) hits a bound;

These problems can be solved by:

1. *Encounter rates*: constraining terms that vary among years (e.g., intercept  $\beta$  and spatio-temporal variation  $\varepsilon(s, t, p)$ ). This can be done in many different ways that are each idiosyncratic and require some special justification. The easiest options are:
  - a. If there is a small number of years with 100% encounter rate, try `ObsModel[2]=3`. This indicates that VAST should check for species-years combinations with 100% encounter rates and fix corresponding intercepts for encounter probability to an extremely high value.

b. If there is a small number of years with either 100% of 0% encounter rate, add temporal structure to intercepts and spatio-temporal terms using `RhoConfig` options.

c. Four other options are listed on the [wiki](#).

2. *Bounds*: Please try running the model without estimating standard errors or a final newton step:

```
# Specify derived quantities to calculate
TMBhelper::Optimize( ..., getsd=FALSE, newtonsteps=0 )
Then check what parameters are being estimated near an upper or lower boundary.
```

## How to implement basic model changes

There are a few basic model types that users often want to fit using VAST. I briefly describe how these can be done here.

1. *Fitting encounter/non-encounter data*: If the user wishes to use only the first component of a delta-model, i.e., to fit a binomial model to simply predict encounter probabilities, then, the `ObsModel` vector should be set to `c("PosDist"=[Make Choice], "Link"=0)`, where [Make Choice] can be any option for continuous data (i.e., 0, 1, or 2). The user should then turn off the last two elements of the `FieldConfig` vector (i.e., `FieldConfig[3]=0` and `FieldConfig[4]=0`) such that there is no spatial or spatio-temporal variability in positive catch rates, and also turn off annual variation in the intercept for positive catch rates (i.e., `RhoConfig[2]=3`). Finally, the user should “jitter” their presence observations by a very small amount (i.e., add a random normal deviation with a very small standard deviation, `rnorm(n=1, mean=0, sd=0.001)`, to each observation for which `b_i=1`). This will result in VAST estimating a logistic regression model for encounter/non-encounter data, except with one additional parameter estimated ( $\sigma_M$ ), plus one additional parameter per category ( $\beta_2(c)$ ), where these additional parameters have no

436 impact on other parameters, are not meant to be interpreted statistically or biologically,  
437 and are an artefact of using VAST (which is designed to fit a delta-model) to  
438 encounter/non-encounter data. This feature has been used to estimate species  
439 distributions for use in ecosystem models (Grüss et al. In press, 2017).

## 442 Works cited

- 443 Edwards, A.M., and Auger-Méthé, M. 2019. Some guidance on using mathematical notation  
444 in ecology. *Methods Ecol. Evol.* **10**(1): 92–99. doi:10.1111/2041-210X.13105.
- 445 Gelman, A., and Hill, J. 2007. *Data analysis using regression and multilevel/hierarchical*  
446 *models*. Cambridge University Press, Cambridge, UK.
- 447 Grüss, A., and Thorson, J. In press. Developing spatio-temporal models using multiple data  
448 types for evaluating population trends and habitat usage. *ICES J. Mar. Sci.*
- 449 Grüss, A., Thorson, J.T., Babcock, E.A., and Tarnecki, J.H. In press. Producing distribution  
450 maps for informing ecosystem-based fisheries management using a comprehensive  
451 survey database and spatio-temporal models. *ICES J. Mar. Sci.*  
452 doi:10.1093/icesjms/fsx120.
- 453 Grüss, A., Thorson, J.T., Sagarese, S.R., Babcock, E.A., Karnauskas, M., Walter, J.F., and  
454 Drexler, M. 2017. Ontogenetic spatial distributions of red grouper (*Epinephelus*  
455 *morio*) and gag grouper (*Mycteroperca microlepis*) in the U.S. Gulf of Mexico. *Fish.*  
456 *Res.* **193**(Supplement C): 129–142. doi:10.1016/j.fishres.2017.04.006.
- 457 Kass, R.E., and Steffey, D. 1989. Approximate bayesian inference in conditionally  
458 independent hierarchical models (parametric empirical bayes models). *J. Am. Stat.*  
459 *Assoc.* **84**(407): 717–726. doi:10.2307/2289653.
- 460 Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B.M. 2016. TMB: Automatic  
461 Differentiation and Laplace Approximation. *J. Stat. Softw.* **70**(5): 1–21.  
462 doi:10.18637/jss.v070.i05.
- 463 Lindgren, F. 2012. Continuous domain spatial models in R-INLA. *ISBA Bull.* **19**(4): 14–20.
- 464 Lindgren, F., Rue, H., and Lindström, J. 2011. An explicit link between Gaussian fields and  
465 Gaussian Markov random fields: the stochastic partial differential equation approach.  
466 *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(4): 423–498. doi:10.1111/j.1467-  
467 9868.2011.00777.x.
- 468 Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre,  
469 A.J., and Possingham, H.P. 2005. Zero tolerance ecology: improving ecological  
470 inference by modelling the source of zero observations. *Ecol. Lett.* **8**(11): 1235–1246.
- 471 R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R  
472 Foundation for Statistical Computing, Vienna, Austria. Available from  
473 <https://www.R-project.org/>.
- 474 Searle, S.R., Casella, G., and McCulloch, C.E. 1992. *Variance components*. John Wiley &  
475 Sons, Hoboken, New Jersey.
- 476 Shelton, A.O., Thorson, J.T., Ward, E.J., and Feist, B.E. 2014. Spatial semiparametric models  
477 improve estimates of species abundance and distribution. *Can. J. Fish. Aquat. Sci.*  
478 **71**(11): 1655–1666. doi:10.1139/cjfas-2013-0508.
- 479 Skaug, H., and Fournier, D. 2006. Automatic approximation of the marginal likelihood in  
480 non-Gaussian hierarchical models. *Comput. Stat. Data Anal.* **51**(2): 699–709.
- 481 Thorson, J.T. 2018. Three problems with the conventional delta-model for biomass sampling  
482 data, and a computationally efficient alternative. *Can. J. Fish. Aquat. Sci.* **75**(9):  
483 1369–1382. doi:10.1139/cjfas-2017-0266.
- 484 Thorson, J.T., and Barnett, L.A.K. 2017. Comparing estimates of abundance trends and  
485 distribution shifts using single- and multispecies models of fishes and biogenic  
486 habitat. *ICES J. Mar. Sci.* **74**(5): 1311–1321. doi:10.1093/icesjms/fsw193.
- 487 Thorson, J.T., Ianelli, J.N., and Kotwicki, S. In press. The relative influence of temperature  
488 and size structure on fish distribution shifts: a case study on walleye pollock in the  
489 Bering Sea. *Fish Fish.*

- Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C., and Zipkin, E.F. 2016a. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Glob. Ecol. Biogeogr.* **25**(9): 1144–1158. doi:10.1111/geb.12464.
- Thorson, J.T., and Kristensen, K. 2016. Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. *Fish. Res.* **175**: 66–74. doi:10.1016/j.fishres.2015.11.016.
- Thorson, J.T., and Minto, C. 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. *ICES J. Mar. Sci. J. Cons.* **72**(5): 1245–1256. doi:10.1093/icesjms/fsu213.
- Thorson, J.T., Pinsky, M.L., and Ward, E.J. 2016b. Model-based inference for estimating shifts in species distribution, area occupied and centre of gravity. *Methods Ecol. Evol.* **7**(8): 990–1002. doi:10.1111/2041-210X.12567.
- Thorson, J.T., Rindorf, A., Gao, J., Hanselman, D.H., and Winker, H. 2016c. Density-dependent changes in effective area occupied for sea-bottom-associated marine fishes. *Proc R Soc B* **283**(1840): 20161853. doi:10.1098/rspb.2016.1853.
- Thorson, J.T., Scheuerell, M.D., Shelton, A.O., See, K.E., Skaug, H.J., and Kristensen, K. 2015a. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods Ecol. Evol.* **6**(6): 627–637. doi:10.1111/2041-210X.12359.
- Thorson, J.T., Shelton, A.O., Ward, E.J., and Skaug, H.J. 2015b. Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. *ICES J. Mar. Sci. J. Cons.* **72**(5): 1297–1310. doi:10.1093/icesjms/fsu243.
- Thorson, J.T., Skaug, H.J., Kristensen, K., Shelton, A.O., Ward, E.J., Harms, J.H., and Benante, J.A. 2014. The importance of spatial models for estimating the strength of density dependence. *Ecology* **96**(5): 1202–1212. doi:10.1890/14-0739.1.
- Tierney, L., Kass, R.E., and Kadane, J.B. 1989. Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Am. Stat. Assoc.* **84**(407): 710–716.