VAST user manual

James Thorson

Purpose of document:

This document is intended to document the model structure and user-options available in package VAST. For guidance and examples of how to use the model, please see the Rmarkdown tutorials in the GitHub "/examples" directory. In the following, I try to use notation similar to the TMB code: 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). Feel free to change notation when describing the model to suite your purposes. For further details regarding terminology, motivation, and statistical properties, please read the papers listed on the GitHub main page.

Model description:

Linear predictors

I use a delta-model that includes two linear predictors. The linear predictor for encounter probability:

$$p_1(i) = \beta_1(c_i, t_i) + \sum_{f=1}^{n_{\omega 1}} L_{\omega 1}(c_i,\ f)\omega_1(s_i, f\ ) + \sum_{f=1}^{n_{\varepsilon 1}} L_{\varepsilon 1}(c_i,\ f)\varepsilon_1(s_i, f, t_i)$$

$$+ \sum_{f=1}^{n_{\delta 1}} L_{\delta 1}(v_i, f)\delta_1(v_i, f) + \sum_{p=1}^{n_p} \gamma_1(c_i,\ t_i, p)X(x_i, t_i, p) + \sum_{k=1}^{n_k} \lambda_1(k)\, Q(i, k)$$

where $p_1(i)$ is the predictor for observation i, $\beta_1(c_i, t_i)$ is an intercept for category c_i and year t_i, $\omega_1(s_i, f\ )$ represents spatial variation at location s_i for factor f and $L_{\omega 1}(c_i,\ f)$ is the loadings matrix that generates spatial covariation among categories for this linear predictor, $\varepsilon_1(s_i, f, t_i)$ is spatio-temporal variation and $L_{\varepsilon 1}(c_i,\ f)$ is the loadings matrix that generates spatio-temporal covariation for this predictor, $\delta_1(v_i, f)$ is random variation in catchability among a grouping variable (tows or vessels) and $L_{\delta 1}(v_i, f)$ is a loadings matrix that generates covariation in catchability among categories for this predictor, $X(x_i, t_i, p)$ are measured density covariates that explain variation in density and $\gamma_1(c_i,\ t_i, p)$ is the estimated impact of density covariates, and $Q(i, k)$ are measured catchability covariates that explain variation in catchability and $\lambda_1(k)$ is the estimated impact of catchability covariates for this linear predictor. Similarly, the linear predictor for positive catch rates:

$$p_2(i) = \beta_2(c_i, t_i) + \sum_{f=1}^{n_{\omega 1}} L_{\omega 2}(c_i,\ f)\omega_1(s_i, f\ ) + \sum_{f=1}^{n_{\varepsilon 1}} L_{\varepsilon 2}(c_i,\ f)\varepsilon_2(s_i, f, t_i)$$

$$+ \sum_{f=1}^{n_{\delta 1}} L_{\delta 2}(v_i, f)\delta_2(v_i, f) + \sum_{p=1}^{n_p} \gamma_2(c_i,\ t_i, p)X(x_i, t_i, p) + \sum_{k=1}^{n_k} \lambda_2(k)\, Q(i, k)$$

where all variables and parameters are defined similarly except using different subscripts (Thorson and Barnett In press, Thorson et al. In press). The loadings matrices are designed such that $\mathbf{L}^T\mathbf{L}$ is the covariance among categories for a given spatial or spatio-temporal process (Thorson et al. 2015a), and when there is only one category **L** is a 1x1 matrix (i.e. a scalar) such that its absolute value is the standard deviation for a given process. This model therefore reduces to a single-species spatio-temporal model (e.g., Thorson et al. 2015b) when only one category is available.

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

```
# Control number of factors
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. The user controls the number of catchability factors used for each component via input:

```
# Control number of spatial and spatio-temporal factors
OverdispersionConfig = c("Delta1"=0, "Delta2"=0)
```

where `OverdispersionConfig[1]` controls $n_{\delta 1}$, and `OverdispersionConfig[2]` controls $n_{\delta 2}$, and a value of zero again "turns off" that component of random covariation in catchability. For example, if the user inputs:

```
# Control number of spatial and spatio-temporal factors
OverdispersionConfig = c("Delta1"=1, "Delta2"=1)
```

then there will be one random effect estimated for each unique level of `Data_Geostat$Vessel` for both the first and second linear predictors.

Link functions

There are different user-controlled options for link-functions that calculate expected encounter probability and positive catch rates given these two linear predictors.

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

where the 2nd element of this vector controls the link functions.

1. `ObsModel[2]=0` corresponds to a conventional delta-model:

$$r_1(i) = logit^{-1}\big(p_1(i)\big)$$

   where $r_1(i)$ is the predictor encounter probability and $logit^{-1}(p_1(i))$ is the logistic function of $p_1(i)$, and:

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

   where $r_2(i)$ is the predicted biomass density for positive catch rates, $log^{-1}(p_2(i))$ 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. Alternatively, `ObsModel[2]=1` corresponds to a "Poisson-process" link function that approximates a Tweedie distribution:

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

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

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

where $r_2(i)$ is the predicted biomass given that the species is encountered. In this "Poisson-process" link function, $\exp\big(p_1(i)\big)$ is interpreted as the density in number of individuals per area such that $a_i \times \exp\big(p_1(i)\big)$ is the predicted number of individuals encountered, and $\exp\big(p_2(i)\big)$ 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 In review).

Observation models:

There are different user-controlled options for observation models for positive catch rates.

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

VAST then calculates the probability of data as:

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

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.

Settings regarding spatial domain

VAST approximates spatial and spatio-temporal variation as being piecewise-constant. To do so, the user specifies `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:

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

where $\mathbf{H}$ is a two-dimensional linear transformation representing geometric anisotropy (with a determinant of 1.0), ν is the Matern smoothness (fixed at 1.0), and κ_1 governs the decorrelation distance for that first linear predictor (κ_2 is also separately estimated for the second linear predictor). By default, the two degrees of freedom in $\mathbf{H}$ are estimated as fixed effects, but the user can specify isotropy (i.e., $\mathbf{H} = \mathbf{I}$) by specifying:

```
# Turn of geometric anisotropy
Data = Data_Fn( …, Aniso=FALSE )
```

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

$$\omega_1(\cdot,f) \sim MVN(\mathbf{0},\sigma_{\omega 1}^2\mathbf{R}_1)$$

$$\omega_2(\cdot,f) \sim MVN(\mathbf{0},\sigma_{\omega 1}^2\mathbf{R}_2)$$

$$\varepsilon_1(\cdot,f,t) \sim MVN(\mathbf{0},\sigma_{\varepsilon 1}^2\mathbf{R}_1)$$

$$\varepsilon_2(\cdot,f,t) \sim MVN(\mathbf{0},\sigma_{\varepsilon 2}^2\mathbf{R}_2)$$

where $\omega_1(\cdot,f)$ is the vector formed when subsetting $\omega_1(s,f)$ for a given f, and $\sigma_{\omega 1}^2$ is the variance of $\omega_1(s,f)$, where other parameters are defined similarly. Specifying a variance of 1.0 ensures that the covariance among categories is defined by the loadings matrix for that term. However, VAST allows spatio-temporal variance to be specified differently as discussed in the section titled “Structure on parameters among years”.

Structure on parameters among years:

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

```
# Control autoregressive structure for parameters over time
RhoConfig = c("Beta1"=0, "Beta2"=0, "Epsilon1"=0, "Epsilon2"=0)
```

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

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

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

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 $\sigma^2_{\beta 1} = 0$ (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 $\rho_{\beta 2}$.

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

$$\varepsilon_1(s,f,t+1) \sim MVN(\rho_{\varepsilon 1}\varepsilon_1(s,f,t), \sigma^2_{\varepsilon 1}\mathbf{R}_1)$$

$$\varepsilon_2(s,f,t+1) \sim MVN(\rho_{\varepsilon 1}\varepsilon_2(s,f,t), \sigma^2_{\varepsilon 2}\mathbf{R}_2)$$

where `RhoConfig[3]` controls the specification of $\rho_{\varepsilon 1}$:

1. *Random walk* – `RhoConfig[3]=2` specifies $\rho_{\varepsilon 1} = 1$
2. *Autoregressive* – `RhoConfig[3]=4` estimates $\rho_{\varepsilon 1}$ as a fixed effect

and settings are defined identically for `RhoConfig[4]` specifying $\rho_{\varepsilon 2}$.

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 category are modelled, VAST is similar to spatial factor analysis (Thorson et al. 2015a), although it permits the use of a delta-model (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 category 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 user 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 $d(x,c,t)$ for :

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

where $r_1^*(x,c,t)$ and $r_2^*(x,c,t)$ are identical to the values specified previously, except that catchability variables are excluded from their computation (i.e., $\delta_1(v,f) = 0$ and $\lambda_1(k) = 0$, etc.)

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

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

where $a(x,l)$ is the area associated with extrapolation-cell x for index l (Shelton et al. 2014, Thorson et al. 2015b). The user can also specify additional post-hoc calculations via input:

```
# Control observation error
RhoConfig = c("SD_site_density"=0, "SD_site_logdensity"=0, "Calculate_Range"=0,
"Calculate_evenness"=0, "Calculate_effective_area"=0, "Calculate_Cov_SE"=0,
'Calculate_Synchrony'=0, 'Calculate_Coherence'=0)
```

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

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

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

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

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

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

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

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

Works cited

Lindgren, F. 2012. Continuous domain spatial models in R-INLA. ISBA Bull. **19**(4): 14–20.

Lindgren, F., Rue, H., and Lindström, J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. **73**(4): 423–498. doi:10.1111/j.1467-9868.2011.00777.x.

Shelton, A.O., Thorson, J.T., Ward, E.J., and Feist, B.E. 2014. Spatial semiparametric models improve estimates of species abundance and distribution. Can. J. Fish. Aquat. Sci. **71**(11): 1655–1666. doi:10.1139/cjfas-2013-0508.

Thorson, J.T. In review. Three problems with the conventional delta-model for biomass sampling data, and a computationally efficient alternative.

Thorson, J.T., and Barnett, L.A.K. In press. Comparing estimates of abundance trends and distribution shifts using single- and multispecies models of fishes and biogenic habitat. ICES J. Mar. Sci. doi:10.1093/icesjms/fsw193.

Thorson, J.T., Ianelli, J.N., and Kotwicki, S. In press. The relative influence of temperature and size structure on fish distribution shifts: a case study on walleye pollock in the 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., 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.