DEP AGREEMENT NO. CZ529

Evaluation of sample size to assess managed-area-level trends in oyster habitats

John Handley, PhD

Task 1. Extend the base model workflow to other covariates and SEACAR parameters and assess their performance;

Deliverable 1b: A report summarizing the covariate testing results, including recommendations for which model(s) are likely to perform best in the oyster sampling planning workflow.

First, we review the set up from DEP CZ325 Task 3 report.

The basic statistical model treats sampled sites as given and the variables measured like shell height, percent live and density at each site as random. Familiar examples include permanent weather stations that repeatedly measure an atmospheric variable, such as wind speed, or wells used to monitor pollutant concentrations in groundwater. Elaborations of this model can include linear models or time series (growth models) where the parameters of the model, like slopes, are allowed to vary spatially.

Mathematically, given a set of sites, $S = \{s_1, \dots, s_N\}$, there is a jointly distributed random variable $\boldsymbol{Y} = (Y(s_1), \dots, Y(s_N))$. At each site, there are possible covariates, $\boldsymbol{X} = (X(s_1), \dots, X(s_N))$. The $Y(s_i)$ are jointly distributed as a multinomial random variable in the following way:

$$\boldsymbol{Y} \sim MNV(\boldsymbol{\mu}, \Sigma) \tag{1}$$

where $\boldsymbol{\mu} = \boldsymbol{X^T \beta}$ and Σ is a function of the distances $\|s_i - s_j\|$ between points. This is an example of a two-dimensional Gaussian process (GP). The covariance matrix Σ is called a kernel. For example, consider the exponential model,

$$(\Sigma)_{i,j} = \alpha^2 e^{-[\|s_i - s_j\|/\rho]^2} \tag{2}$$

where α^2 is called the "partial sill" and ρ is a decay parameter. It models a condition wherein random variables that are sampled closer together have higher covariance than those sampled farther apart, with increasing distance eventually driving covariance to zero. The partial sill defines how large the covariances/variances are at a given distance and the decay parameter defines how quickly influence decreases with increasing distance. GP's are implemented in the probabilistic programming language Stan, which increases speed and stability of model fitting.

It is customary to break these components down into a mean process, $\mu(s)$, a spatial random effects process, $w(s)$, and a pure error process $\epsilon(s)$, which captures all the variation not explained by the model and unaffected by position,

$$Y(s) = \mu_G + \mu(s) + w(s) + \epsilon(s)$$

$$\epsilon(s) \sim N(0, \sigma_\epsilon^2)$$

The mean process, $\mu(s)$, is simply the mean at each site. The spatial random effects model $w(s)$ captures the spatial variation in the mean process from site to site, $w(s) \sim MNV(0, \Sigma)$. It is possible, for example, that there are measurements associated with each site. The “fixed effect” μ_G is the “global mean” and represents the average of the parameter value over all the sampled sites.

In our setting, we have multiple observations per site. For example, Estero Bay Aquatic Preserve has 36 sites with the number of observations ranging from 3 to 1792. The model above is modified to allow multiple observations per site,

$$Y_i(s) = \mu_G + \mu(s) + w(s) + \epsilon(s)$$

Managers want to know the average parameter value within a managed area. For example, what is the average shell height in Estero Bay?

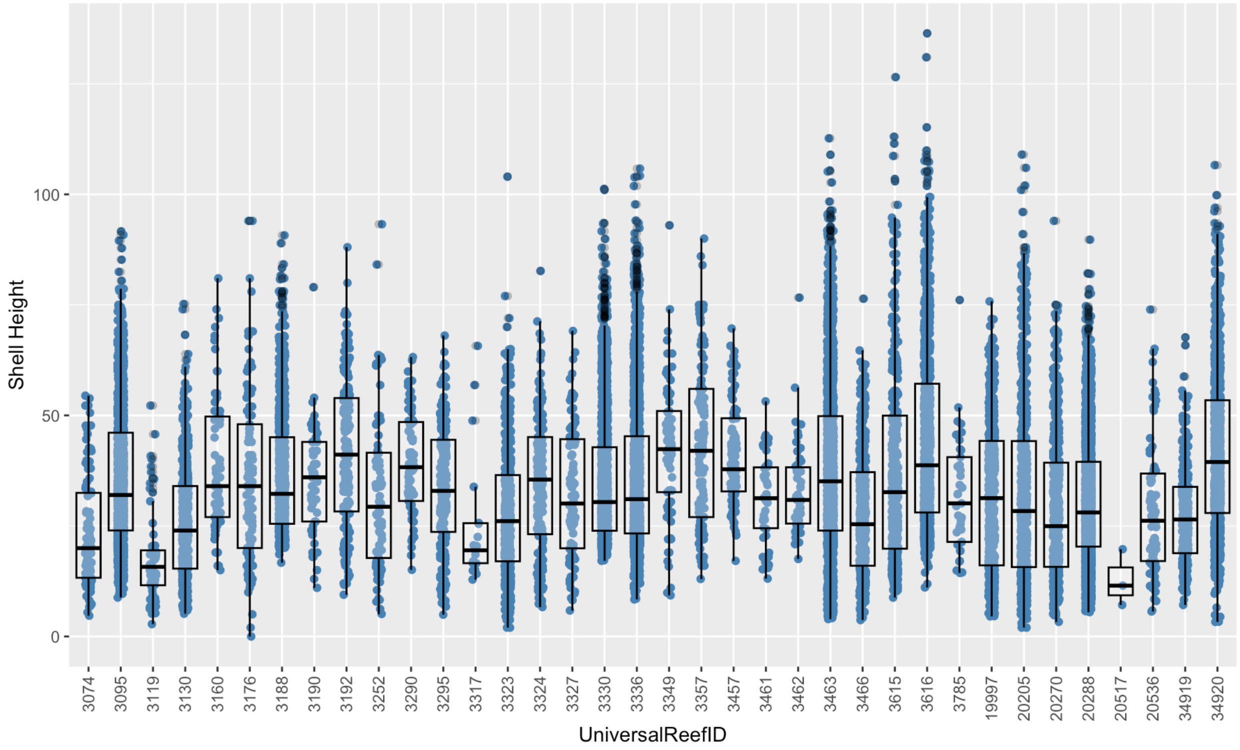

Fig 1. Shell Heights in Estero Bay with box plots showing 25% and 75% ranges.

From Figure 1, we observe that there is considerable reef-to-reef variation in values and sample sizes. This is a classic issue in statistics about estimating a mean in a heterogeneous sample. Does it make sense to simply pool the values and take the average? Doing so gives an average of 34.954 with a standard error of 0.136. We should be concerned about the variation. Is this estimate dominated by reefs with larger sample sizes? Could they bias the estimate? Should the reefs with smaller sample sizes have a greater contribution? Fortunately, this issue is solved by “random effects”, which we have implemented in a spatial context. Our estimate using a GP spatial model fitted using a Bayesian method produces a mean of 32.258 with standard error of 1.974. This is a slightly smaller estimate with a greater and more conservative error estimate. This estimation procedure has taken reef-to-reef variation into account, in contrast to the pooled approach, which produced a biased estimate with greater precision, which could be misleading.

Covariates

It is plausible that site-by-site covariates like salinity, dissolved oxygen or depth could be related to parameters of interest. We introduce covariates into the model by regressing on the site-to-site means $\mu(s)$.

$$Y_i(s) = \mu_G + \mu(s) + w(s) + \epsilon(s)$$

$$\mu(s) = \beta x(s)$$

Successful covariate regression depends on the spatial effect being improved by the covariate. There is an interplay between the spatial component and the covariate component. For example, one could regress the response on the covariate without any spatial model and then check that the quality of that effect is improved by adding the spatial component back into the model.

A random effects model without the spatial component is,

$$Y_i(s) = \mu_G + \mu(s) + \epsilon(s)$$

$$\mu(s) = \beta x(s)$$

The spatial model is called Bayesian kriging which is essentially a way to smooth and interpolate across points in space. The farther a given location is from sites with existing data for the parameter of interest, the greater the uncertainty in prediction. But that prediction is based on interpolation among known values from sampled locations. Bayesian kriging is a statistical approach that differs from other forms of kriging that fit a surface to values at location points because it can handle multiple observations per site and estimates uncertainty. In this application, both the repeated observations per site and the spatial arrangement contribute to model uncertainty. Having multiple observations per site means we estimate means at each site, but the variance of observations per site also contributes to the uncertainty. Because kriging is a form of interpolation, it can fit the means of the data quite well. When one introduces a covariate, the expectation is that it might improve the fit if that covariate predicts the means even better. Figure 1 shows a typical result of adding a covariate to one of the spatial oyster parameter models developed for this project, where the spatial model without a covariate follows the empirical means closely, while the spatial model with a covariate estimates the general pattern in the parameter of interest as a function of the covariate (i.e., increase in percent live as a function of dissolved oxygen in the case of Figure 1), but does not estimate the empirical means well. The reason is that the spatial model with a covariate is essentially doing a regression but taking into account spatial dependency. Thus, the fact that the pure random effects model produced a similar result without the spatial dependency, suggests that the relationship between the covariate and the spatial component is much less influential on the model fit than the relationship between the response and the covariate. So, although the use of spatial dependencies theoretically should improve the estimate of the relationship between a covariate and response parameter (mean), the inverse is not true, meaning the addition of the covariate does not improve the model's predictions of response values across space.

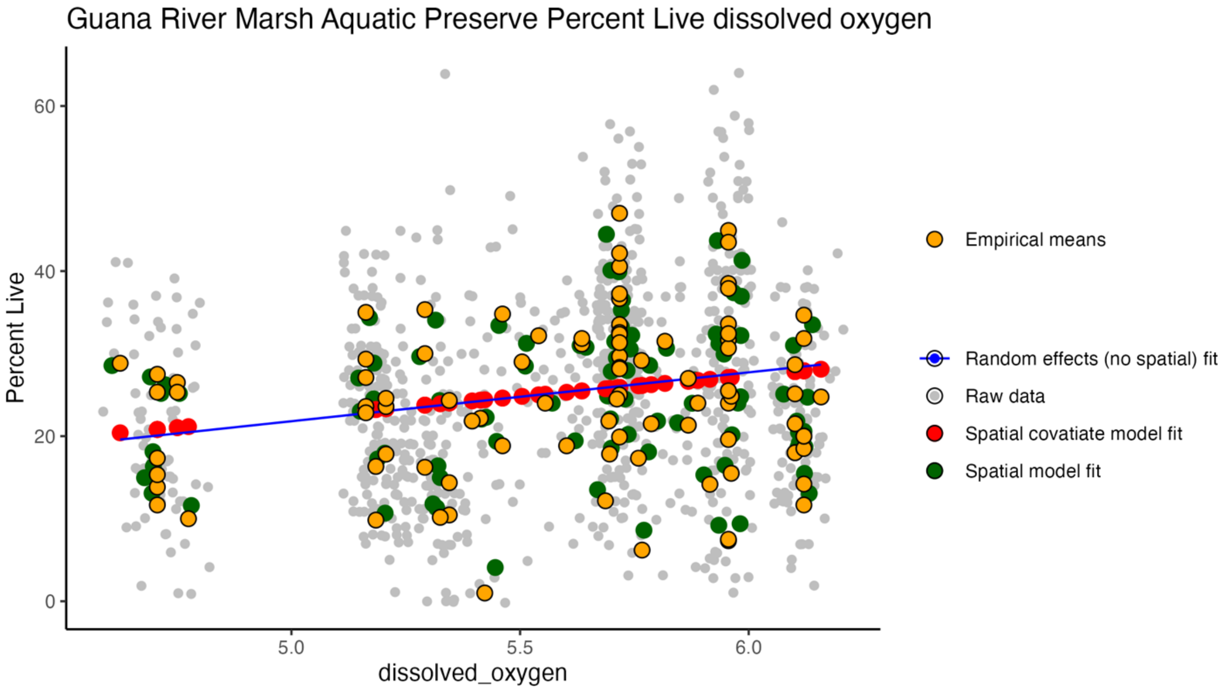

Figure 1. Fits of spatial model with covariate (red), random effects no spatial (blue), and spatial model (Bayesian kriging) (green). $\hat{\rho} = 2.00 \pm 0.21$, $\hat{\beta} = 5.03 \pm 0.96$; $\hat{\beta} = 5.92 \pm 1.02$; $\hat{\rho} = 2.00 \pm 0.21$. R2 for both regression models is 0.173.

Consequently, covariates are not generally helpful for prediction unless the correlation between the parameter and covariate is nearly one. Indeed, estimated slopes from the parameter-specific regression models with and without a spatial component mostly agree (Figure 2) but there are exceptions that we investigate more closely next.

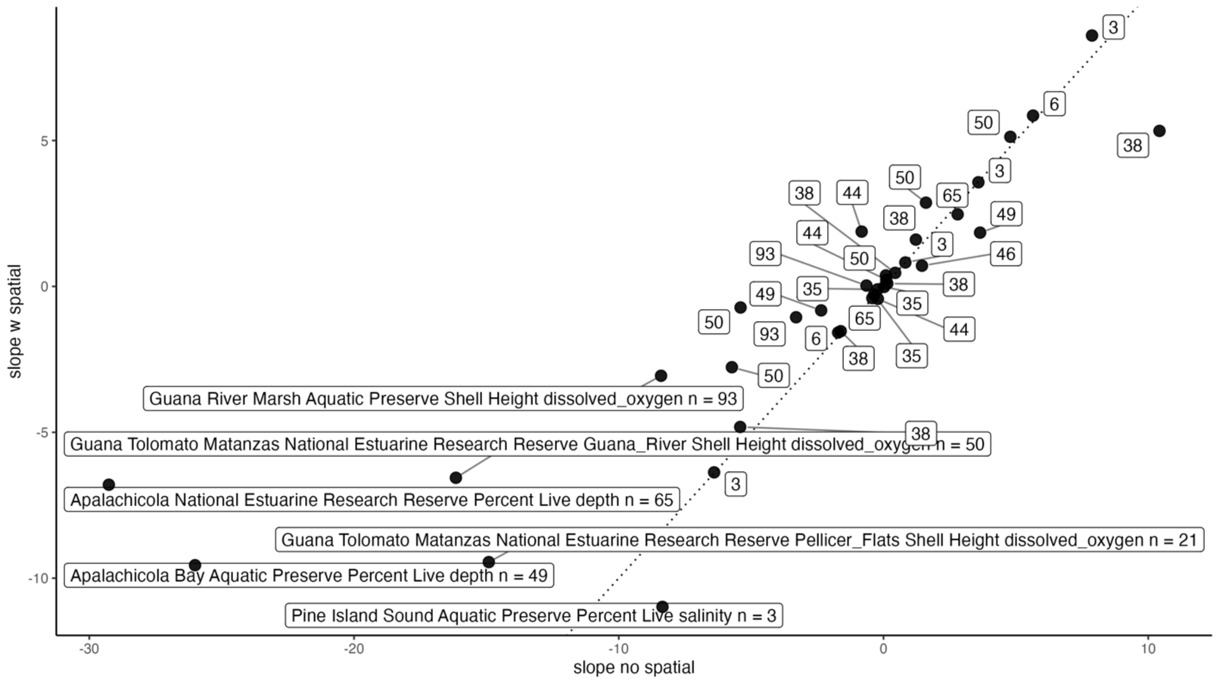

Figure 2. Plot of slopes for regression models for each managed area with and without a spatial component. The slopes generally agree but when they do not, slopes for the spatial regression model are generally smaller. Point labels show the number of sites for the relevant managed area x parameter combination.

Figure 3 shows one of the cases where the slopes do not agree. In this case, the spatial regression model is similar to the spatial only model but is not as accurate.

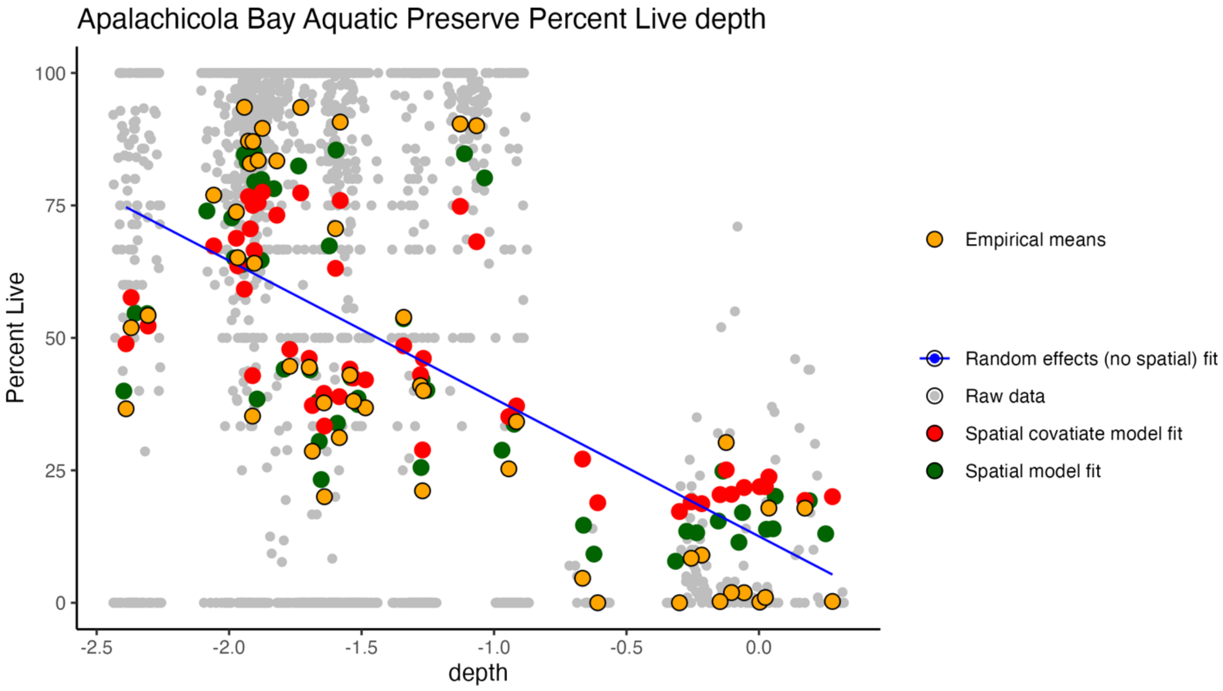

Figure 3. Estimated slope for the spatial regression model: $\hat{\beta} = -9.56 \pm 3.36$; estimated slope for the no-spatial model: $\hat{\beta} = -26.00 \pm 1.63$.

When the spatial component does not improve the model, the estimated value of the decay parameter ρ tends to be very small, indicating the spatial component is negligible (see Equation 2) and we are effectively fitting a basic simple regression model. Such models do not provide guidance for the GRTS site selection because the estimation uncertainty is essentially the same for all sample locations. In contrast, when the decay parameter is close to 2, the model is essentially the spatial model. Figure 4 shows plots when the R2 is >= .90 and there is no improvement beyond the pure spatial model.

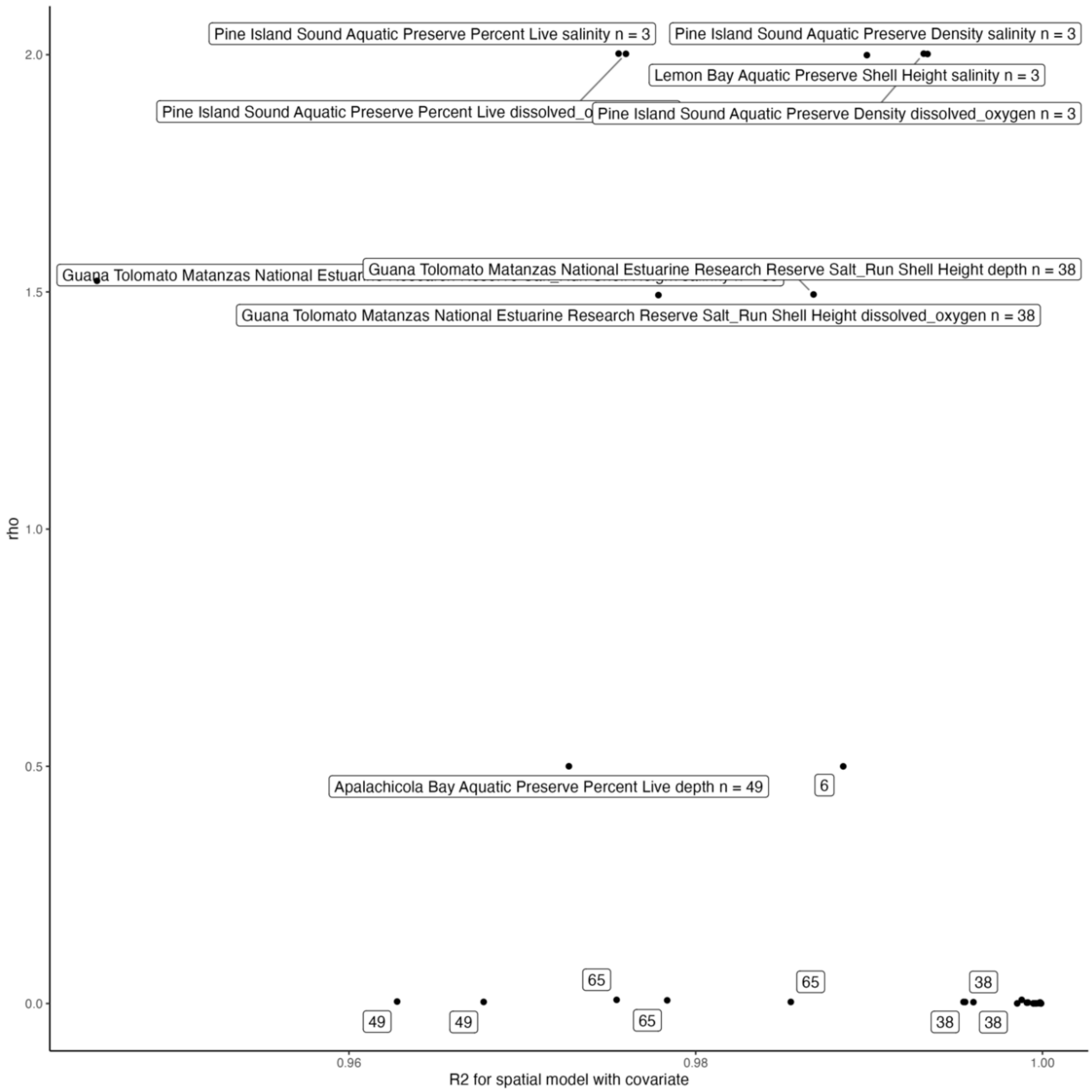

Figure 4. Models with R2 >= 0.9 showing examples where ρ is close to 2 which is essentially the spatial model and examples where ρ is close to zero meaning the model is essentially a random effects linear model.

The upshot is that the spatial model with regression does provide a better estimate of the response of a covariate than a pure random effects model that ignores spatial dependency, but it is not useful for spatial prediction. These results suggest that using the pure spatial models (i.e., without covariates) is the best option for the oyster sampling planning workflow moving forward.

In the software distribution there is a file, “covariate_results.csv” with all the mean estimates, standard deviations and R2 values for each combination of Managed Area, Parameter and Covariate. There is also a folder maintained on Google Drive called “covariate_results_plots” of all plots of the fits in the style of Figures 1 and 3. A link to the drive is a spreadsheet called “CZ529_Files_Manifest.csv”, which has a list of all the files and directories associated with agreement CZ529.