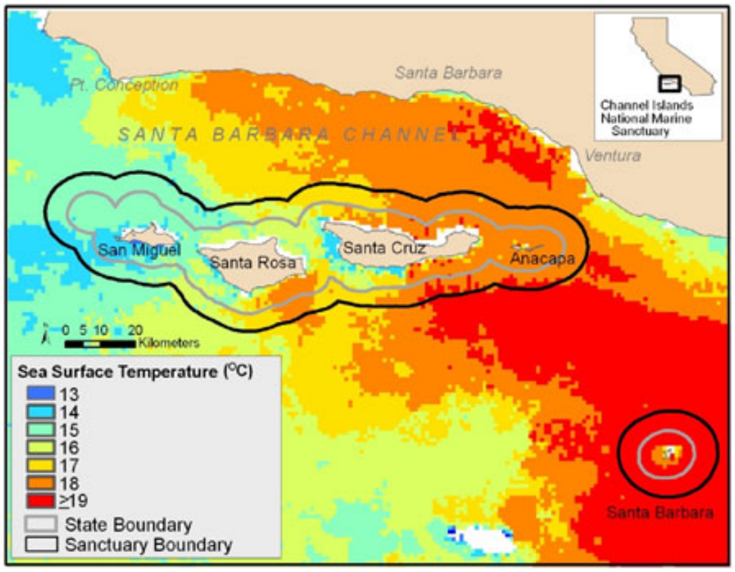
Homework 6 (30 pts)

This is going to use the Channel Islands fish dataset that we've encountered previously. The file "oxyjulis_subset.csv" contains the data to use. We're going to look at whether a particular species, *Oxyjulis californica* (adults only), tends to be more or less abundant as a function of temperature. Temperature varies greatly across the islands (and across time too), because they lie in a region where colder northern current and warmer southern currents meet.



© Keoki Stender

It seems simple to ask whether a particular fish is more abundant at colder or warmer sites, but it will be tricky to model this properly with this highly overdispersed data. There are different ways one could combine the count data with temperature data; what I've done is to just take each count, and associated it with the temperature measured on the same day at the same site.

Make some exploratory plots of temperature vs Site, the distribution of all counts, and count vs temperature. What are your thoughts so far?

This is count data, so starting with a Poisson GLM seems like a good idea. There are many samples per site, so we should include Site as a predictor (make sure to make it a factor!). Make a Poisson GLM with temperature and Site as predictors, and quantify how overdispersed the data is. In principle it would be a good idea to account for the fact that there are multiple measurements per year, and that the time series at each site may be autocorrelated over time, but we're going to focus on how to model the distribution of the data.

Based on the shape of the data, it seems like a good idea to consider zero inflation. But we'll want to compare a zero-inflated model to a non-inflated model, so first fit a negative binomial GLM with temperature and site as predictors. Plot the fitted effects, and report likelihood ratio tests for the predictors. What's the estimate for theta for the negative binomial? What does this model say about the relationship between abundance and temperature? Plot the deviance residuals vs 1) the fitted values, 2) vs the predictors. Does a linear model for temperature (on the link scale) seem reasonable? Later we'll use simulation to ask whether these residuals look weird or not.

The data has plenty of zeros, so let's consider whether a zero-inflated model yields a better fit and/or different patterns. We have two predictors (temperature and site), and either of these can be included in a zero-inflated model for predicting 1) the count part and 2) the extra zeros. So, there are a lot of possible models that could be constructed and compared. For now let's keep it somewhat simple and make 5 models to compare: 1) the negative binomial model with temperature + site; 2) a zero-inflated poisson with temperature + site as predictors only for the count model; 3) a zero-inflated negative binomial with temperature + site as predictors only for the count model; 4) a zero-inflated poisson with temperature + site as predictors both for the count model and also for the zero-inflation (binomial) model; 5) a zero-inflated negative binomial with temperature + site as predictors both for the count model and also for the zero-inflation (binomial) model. Use the pscl package.

Calculate AIC for all five models. Which is the 'best' model (lowest AIC)? What does it mean that this is the best model, compared to the other models?

What is the effect(s) of temperature in the best model? How do you interpret this result, statistically and biologically, compared to the negative binomial model (i.e. non-inflated)?

Do you notice anything a little funny about the standard errors for the model coefficients for the best model? We're probably on the verge of breaking the algorithm by fitting so many parameters for the different sites. In general fitting this many levels is better achieved using random effects, but we'll get to that later.

Unfortunately the nice effects visualization from the effects package isn't implemented for the zero-inflated model software. But we would still like to visualize the fitted results. For model #5 in the above list, let's visualize the fitted effect of temperature on the probability of getting an 'extra' zero. Because the model has a factor for Site, that means the different sites will have different intercepts (but the same slope for temperature). So pick four sites and use curve() to plot the fitted logistic curve for those sites, all on the same plot, with the x-axis having the same range as the range of temperature in the dataset. Roughly how much does temperature change the proportion of extra zeros?