Urbanization and Kelp Coverage in Southern Californa, a Basian Approach

Isaac Rosenthal

May 16, 2017

**Introducton and Background**

Kelp forests are an amazing ecological treasure. As a natural resource they provide many important ecological and economic services, including (but not limited to) food, fertilizer, fish habitat, and a number of useful compounds and chemicals (Edgar, Samson, & Barrett, 2005; Edyvane, 2003; Leet, Dewees, Klingbeil, & Larson, 2001; Steneck & Erlandson, 2002). Giant kelp (Macrocystis pyrifera) is a globally distributed species and is the foundation of the traditional canopy forming "kelp forest". M. pyrifera is found in cold, nutrient rich water, typically near the coast (Johnson et al., 2011). M. pyrifera has a narrow range of environmental tolerance and can be sensitive to even small changes (Johnson et al., 2011).

Kelp die offs on localized or regional scales are not unheard of, and in some areas the drivers of this loss are reasonably well understood to be driven by temperature increases or other environmental phenomena. (Halpern, Walbridge, Selkoe, & Kappel, 2008; Lotze, 2006; Reed et al., 2016). Kelp is also vulnerable to anthropogenic activities resulting in sedimentation, eutrophication, and other coastal pollution (Steneck & Erlandson, 2002; Valentine & Johnson, 2004). Many of these stressors are associated with urbanized areas (although eutrophication is also strongly linked to agriculture, but that is beyond the scope of this particular project)

Because kelp patches are ephemeral by nature, it can be difficult to distinguish between natural variation and long-term trends (Reed et al., 2016). In order to make claims about large scale spatial or temporal trends, it is critical to utilize datasets large enough to allow differentiation between short and long term changes. The USGS landsat database provides satellite imagery of the entire planet, back to 1985 (USGS, 2016). This dataset should be sufficient to address my hypothesis: urbanized areas have less kelp than non-urbanized areas.

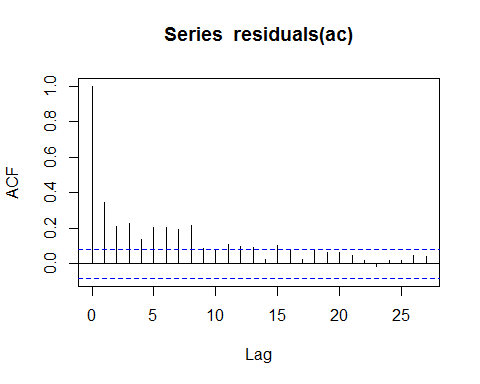
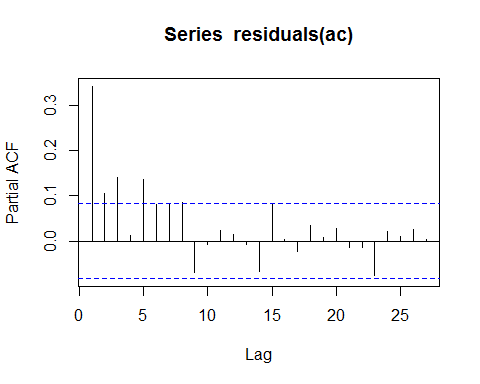
**Data**

Data was adapted from Reed et al. 2016 (cite). As provided by the authors, this dataset contained kelp biomass for California, from San Francisco south. The dataset was pre-filtered to pixels that contained kelp at some point in the dataset. For this analysis, data was trimmed down to all kelp within a 50 kilometer square of 3 urbanized and 3 non-urbanized control sites. Urbanized sites consist of Santa Barbara, Los Angeles, and San Diego, with non-urbanized controls being 50 kilometers south (or, in the case of Santa Barbara, east) of each city. While admittedly rough, modeling the zone of urban influence was outside the scope of this project. Biomass was then converted to presence/absence of kelp in a given landsat pixel, and then expressed as percent cover of the maximum cover for each site. This was then summarized as quarterly averages for analysis in order to pull out any contributions of seasons. See "02\_cleanup.R" in this GitHub repository for processing scripts.

**Statistical approach**

Statistical analyses were done primarily with the BRMS package in R. BRMS is a flexible tool for writing multi-level Bayesian models. A Bayesian approach is fitting for this work for a number of reasons. One advantage of this approach is the ability to incorporate prior information into my analysis. Uninformative priors were used in this analysis, but moving forward with this research the ability to adjust prior beliefs will become valuable as I begin to tease apart the finer details of this system. The primary reason I have chosen a Bayesian approach for this work is because it will allow me to discuss degree of belief in my results. A frequentist approach would constrain inferences to p-value interpretation, or some similar metric. While this hard line can be useful in many applications, in this case it will be more valuable to discuss these results in a less restrictive framework. A similar "degree of belief" style of inference can be seen in other modeling projects designed to influence policy making, such as the International Panel on Climate Change (IPCC, 2013).

**The Model**  
Because this is a time series, it will be important to accurately model any temporal autocorrelation in the system. This dataset has an additional level of complexity from spatial autocorrelation because there are three sites, each with a paired control. Using a rough linear model we can generate ACF and partial ACF plots to get an idea of the temporal autocorrelation structure.

As we can see, there appears to be autocorrelation out to a lag of 8 time points. Because each time point represents quarters, this is equivelent to a two year lag. This two year lag is not an unreasonable estimation, and was even expected; giant kelp in southern California are known to require between two and five years to recover from large desturbance events (Grahm et al., 1997). By allowing temporal autocorrelation to vary by site, I am accounting for some of the difference between geographic regions. Specifically, I suspect that ARMA autocorrelation structure where I have a two year lag varying by site will be a good fit. The ACF plot shows a long decay, which is consistant with an AR structure, but the PACF shows a long decay as well, which would suggest an ARMA structure. Reed et al. showed a consistant decrease in kelp canopy cover over the same 15 year period as this analysis, lending support to a moving-average structure (Reed et al., 2016)

The following models were run. kelp\_brm\_qtr incorporates an 8 time point lag with an AR structure, whereas kelp\_brm\_yr incorporates an explicit two year lag. Finally, kelp\_brm\_yr\_arma incorporates an autoregressive moving average autocorrelation structure. In all models, percent coverage is the response variable, with fixed effects of treatment (urban or nonurban) and quarter (representing season).

kelp\_brm\_qtr <- brm(percent\_coverage ~ treatment + qtr, data=kelp, chains=3, autocor = cor\_ar(~time\_point | site, p = 8), iter = 6000, warmup = 2000, cores = 2)

**Diagnostics**

Using summary(), we can inspect rhat to see if we achieved convergance, and neff to see our number of effective samples. Only one model is shown in order to save space, but all three models converged properly with rhats of 1 and large numbers off effective samples. See file final\_paper.RMD in this github repositoryfor other summary tables.

## Family: gaussian(identity)   
## Formula: percent\_coverage ~ treatment + qtr   
## Data: kelp (Number of observations: 560)   
## Samples: 4 chains, each with iter = 6000; warmup = 2000; thin = 1;   
## total post-warmup samples = 16000  
## ICs: LOO = Not computed; WAIC = 4947.91  
##   
## Correlation Structure: arma(~year|site, 2, 2, 0)  
## Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat  
## ar[1] 0.43 0.20 0.08 0.88 2382 1  
## ar[2] 0.44 0.17 0.04 0.74 2310 1  
## ma[1] -0.15 0.20 -0.60 0.18 2436 1  
## ma[2] -0.46 0.13 -0.67 -0.15 2482 1  
##   
## Population-Level Effects:   
## Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat  
## Intercept 17.62 4.09 9.79 26.00 13463 1  
## treatmenturban -4.23 5.18 -14.92 5.59 13911 1  
## qtr 1.13 0.70 -0.24 2.51 12269 1  
##   
## Family Specific Parameters:   
## Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat  
## sigma 19.88 0.59 18.76 21.09 13962 1  
##   
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample   
## is a crude measure of effective sample size, and Rhat is the potential   
## scale reduction factor on split chains (at convergence, Rhat = 1).

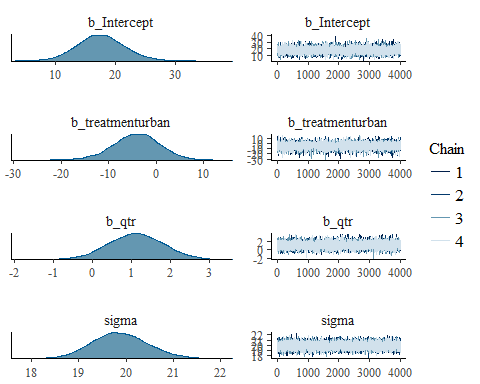
In order to select the best fitting model, LOOAIC can be compared using the loo()function.

## LOOIC SE  
## kelp\_brm\_qtr 4949.06 57.07  
## kelp\_brm\_yr 4970.93 55.24  
## kelp\_brm\_yr\_arma 4947.94 56.33  
## kelp\_brm\_qtr - kelp\_brm\_yr -21.87 13.66  
## kelp\_brm\_qtr - kelp\_brm\_yr\_arma 1.12 6.35  
## kelp\_brm\_yr - kelp\_brm\_yr\_arma 22.99 11.32

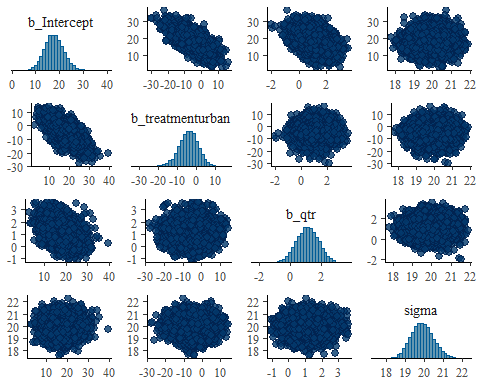
As indicated by loo(), the model incorporating a two year lagged ARMA autocorrelation structure is the best fit (although the 8 time point lag AR structure is also a reasonably good fit). The two year lagged AR structure was a considerably worse fit. The LOOIC reported here is similar to the WAIC and can be interpreted in the same way - the two criteria are asymptotically equal and give essentially the same estimate (see documentation for R package LOO v. 1.1.0)

Of note here is that BRMS treats the first level of treatment (in this case nonurban) as the intercept, and presents the estimate for the urban treatment as a measure of deviation from that mean. In other words, a simple inspection of this summary table hints that urban sites may have less kelp than nonurban sites. Post hoc inspections of condtional and marginal distributions of differences between treatments can be done for a more quantititative inference, but first there are several other diagnostics we must do check.

We can manually inspect the chains for convergance using plot(). This is good practice, as BRMS uses hamiltonian monte-carlo sampling and any divergences in the chains here are diagnostic of issues with locating the typical set, potentially requiring some form of data transformation to remedy

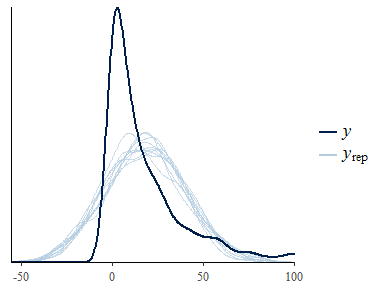


Note that the above output was truncated; for full inspection of chains see source code. This plot also shows us the posterior distributions of these parameters. All chains converged, and posterior distributions look normal, perhaps b\_Intercept and sigma appear slightly skewed.

We can also see these posterior distributions through a pairs plot. 

Again, the above plot was truncated for presentation. Here we see normally distributed posteriors, and a negative correlation between treatments. The treatments are mutually exclusive, so this is not particuarly concerning.

Another useful diagnostic is the PPcheck. ppcheck() creats a plot that compares the observed values of the response variable (in this case percent kelp coverage) to simulated datsets drawn from the posterior distribution.



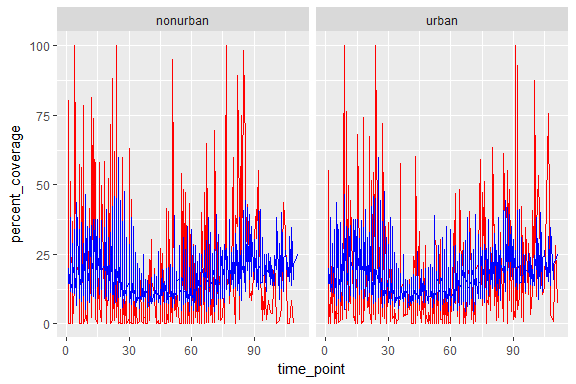
This is not a very good fit. Ideally, the simulated distributions would line up with the observed values. THis may be a problem with my likelihood function - I was unable to get a better fit in time to submit this project, but this is something I am actively working on improving. It could also be due to missing parameters such as temperature.

**Results**

Now that we've verfied that this model fits reasonably well (except for that pp\_check), we can begin to make some inferences. One direction we can go here is to generate predictions and see how they compare to the data. In BRMS, the workhorse here is the fitted()function. In the following chunk, we are generating predicted values based off of our model.

preds <- as.data.frame(fitted(kelp\_brm\_yr\_arma))  
#add original data back on for plotting  
preds <- preds %>%  
 mutate(site = kelp$treatment) %>%   
 mutate(time\_point = kelp$time\_point)

Now we can plot these predicted values on top of the original data and see how well the two line up.

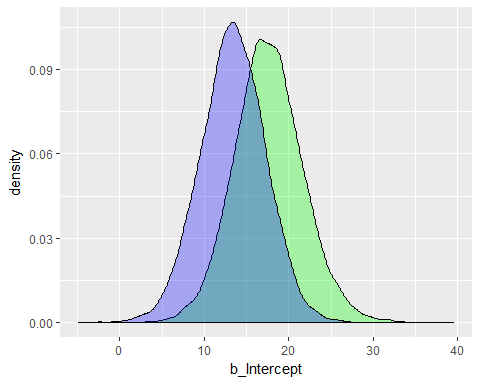


This lines up reasonably well, with red being the data and blue being the predictions. Again, we see hints of sub-par fit as the predictions seem to be clustered in the middle of the full range; the model is not predicting extreme value very well.

Despite a lack of p-values, we can still compare treatments in this baysian analysis. This is done by examining the degree to which posterior distributions overlap. In order to do this, we must first extract samples from the posterior distribution using posterior\_samples(). Thankfully, BRMS already generates and stores these samples when fitting the model, so this is a trivial step.

samps <- posterior\_samples(kelp\_brm\_yr\_arma)

After some data wrangling, we can plot the densities of each distribution in order to visualize the marginal difference between treatments, with green being the nonurban sites and blue being urban.



A more concrete treatment contrast can be done by examinining the conditional differences between treatments. If our question is "to what degree of certainty does urbanization decrease kelp coverage", we need only to determine the percentage of samples in which urbanization was greater than 0, or in other words, the number of samples in which urban sites had kelp coverage equal to or greater than non-urban sites. If we convert this to a percent and subtract it from 100, we get the degree of certainty that urban sites have less kelp than non-urban sites.

100-sum((samps$b\_treatmenturban > 0)/length(samps$b\_treatmenturban)\*100)

## [1] 80.4125

We see a 19.6% overlap in conditional difference between distributions, and we can say with 80.4% certainty that there is a negative effect of urbanization on kelp coverage in these three regions of California.

**Discussion** Clearly, this is a rough model. Both the ppcheck and the plot of fitted values over the data suggest that the model is either a missing parameter or has an incorrect autocorelation structure. Given the complexities of modeling time series data across multiple sites, this is an issue that will require deeper investigation.

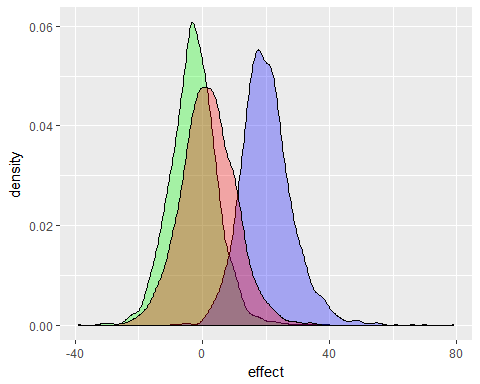
Bearing in mind this (hopefully not glaring) issue with the model, the results of this analysis are thought provoking. An 80% certainty that urbanization decreases kelp coverage is a compelling result given the complexities of the system. As I expand this work into the rest of california and then globally into Tasmania and beyond it will be interesting to see if this trend is common to other regions.

I think a major limitation of this analysis is that my only predictor is time. It would be ideal to include temperature directly to get at environmental differences between years. Reed et al. have recently shown that kelp coverage may not be strongly linked to temperature anomalies in the previous year, but nonetheless this will be a better model if it doesn't treat every year as identical (Reed et al., 2016).

One potential confounding factor in this model is the size of the sites. Because Los Angeles is much larger than San Diego or Santa Barbara, it is entirely possible that the control site for Los Angeles is still more urbanized than is appropriate for the control sites. This is a larger issue that I am going to need to address as I move into more detailed analyses of the Floating Forests data. To begin to dig deeper into this, I modeled the fixed effect of site to get an idea of how this could be skewing my resuts.

kelp\_brm\_site <- brm(percent\_coverage ~ site + qtr, data=kelp, chains=2,  
 cores = 2, iter =2000,  
 autocor = cor\_arma(~year | site, p = 2, q = 2))

site\_samps <- posterior\_samples(kelp\_brm\_site)   
  
#LA\_control\_vs\_SB\_control <- samps$b\_Intercept- samps$b\_treatmenturban   
ggplot(data = site\_samps) +  
 geom\_density(aes(x = b\_siteLA\_control), fill = "green", alpha = 0.3) +  
 geom\_density(aes(x = b\_siteSB\_control) ,fill = "blue", alpha = 0.3) +   
 geom\_density(aes(x = b\_siteSD\_control), fill = "red", alpha = 0.3) +  
 labs(x="effect")



This alternate model converged, and shows differences between control sites. THis warrents further investigation,but overall I have shown that, with a good deal of certainty, urban sites in california have less kelp than non-urban sites.

**References**

Edgar, G. J., Samson, C. R., & Barrett, N. S. (2005). Species Extinction in the Marine Environment: Tasmania as a Regional Example of Overlooked Losses in Biodiversity. Conservation Biology, 19(4), 1294-1300. <http://doi.org/10.1111/j.1523-1739.2005.00159.x>

Edyvane, K. (2003). Conservation Monitoring and Recovery of Threatened Giant Kelp ( Macrocystis pyrifera ) Beds in Tasmania - Final Report. Department of Primary Industries, Water & Environment, Hobart, (November), 106-177.

Grahm, M. H., Harrold, C., Lisin, S., Light, K., Watanabe, J. M., & Foster, M. S. (1997). Population dynamics of giant kelp Macrocystis pyrifera along a wave exposure gradient. Marine Ecology Progress Series, 148, 269-279.

Halpern, B. S., Walbridge, S., Selkoe, K., & Kappel, C. (2008). A Global Map of Human Impact on Marine Ecosystems. Science, 319(2008), 948-952. <http://doi.org/10.1126/science.1149345>

IPCC. (2013). Summary for Policy Makers. Climate Change 2013: THe Physical Science Basis. <http://doi.org/10.1017/CBO9781107415324>

Johnson, C. R., Banks, S. C., Barrett, N. S., Cazassus, F., Dunstan, P. K., Edgar, G. J., . Taw, N. (2011). Climate change cascades: Shifts in oceanography, species' ranges and subtidal marine community dynamics in eastern Tasmania. Journal of Experimental Marine Biology and Ecology, 400(1-2), 17-32. <http://doi.org/10.1016/j.jembe.2011.02.032>

Leet, W., Dewees, C., Klingbeil, R., & Larson, E. (2001). Giant Kelp. Lotze, H. K. (2006). Depletion, Degradation, and Recovery Potential of Estuaries and Coastal Seas. Science, 312(5781), 1806-1809. <http://doi.org/10.1126/science.1128035>

Reed, D., Washburn, L., Rassweiler, A., Miller, R., Bell, T., & Harrer, S. (2016). Extreme warming challenges sentinel status of kelp forests as indicators of climate change. Nature Communications, 7(May), 1-7. <http://doi.org/10.1038/ncomms13757>

Steneck, R., & Erlandson, J. M. (2002). Kelp Forest Ecosystems???: Biodiversity , Stability , Resilience and Future.

United States Geological Survey. (2016). Landsat Missions. Retrieved from <https://landsat.usgs.gov/>

Valentine, J. P., & Johnson, C. R. (2004). Establishment of the introduced kelp Undaria pinnatifida following dieback of the native macroalga Phyllospora comosa in Tasmania, Australia. Marine and Freshwater Research, 55, 223-230. <http://doi.org/10.1071/MF03048>