Statistical analysis of coral cover data from the AIMS Long-term Monitoring Program – Description of ensemble of models for the 2020-2021 Annual Condition Report

Murray Logan

14 July 2021

AIMS LTMP data has been collected using robust, widely accepted standard methods for 35 years[1].

The large Manta Tow time series dataset is complex, for example due to sampling design changes over time and logistical constraints. Estimating the abundance of hard coral across large scales from this time series is a complex modelling challenge with numerous statistical considerations. Firstly, Manta Tow data are collected on a categorical scale. Such data are typically converted to percentage cover by mapping the categories to their associated percentage midpoints and the resulting percentage data are used as inputs into statistical models. Alternatively, it is possible to model the data as ordered categories and thus defer any conversions to percentages until after the models have been fit and summarised.

Secondly, to provide large scale estimates (e.g., percentage coral cover in the Central Great Barrier Reef), it is necessary to aggregate Manta Tow data from the level of individual tows up to the level of reefs and then finally, up to the higher spatial scales. The most basic approach is to simply average the tow-level coral covers (per reef per monitored year) and then calculate the average of these averages for all monitored reefs each year. Whilst simple, this approach (Raw means) implicitly assumes that the data are drawn from Gaussian (normal) distributions. Since it is unlikely that Manta Tow coral cover data conform to a Gaussian distribution the resulting estimates are potentially misleading.

Thirdly, estimates of coral cover should be accompanied by some measure of confidence in the estimates that take into consideration measurement uncertainties as well as uncertainties that might arise from relying on a subset of reefs to represent the full population of reefs in an area. Simple means of means (Raw means) are incapable of adequately and robustly propagating such uncertainties.

Statistical models can yield robust estimates of confidence by propagating distributional properties up the levels of a model. Nevertheless, it is necessary to nominate the general form of the underlying distributions. For some data types (such as count data), identifying a sensible modelling distribution is relatively straightforward. However, in addition to the option of modelling the measured data as ordinals (categories), percentage cover data can be modelled against numerous possible candidate distributions (including binomial and beta).

In a seminal paper that described a 50% decline in coral cover over the proceeding 30 years, De'ath et al (2012)[2] modelled trends in coral cover via quasi-binomial generalised additive mixed effects models (GAMMs) on data aggregated (means) to the level of individual reefs per monitored year.

---

[1] For more information see

https://www.aims.gov.au/docs/research/monitoring/reef/sampling-methods.html

[2] De'ath, G., Fabricius, K.E., Sweatman, H. and Puotinen, M., 2012. The 27–year decline of coral cover on the Great Barrier Reef and its causes. Proceedings of the National Academy of Sciences, 109(44), pp.17995-17999. https://doi.org/10.1073/pnas.1208909109

The published trends of these analyses are familiar to many in the marine science community and thus, there is an argument for continuing to provide similar analyses into the future. Nevertheless, subsequent statistical advances have broadened the scope and complexity of available analytical tools and there is also a desire to underpin any reported trends by the best available techniques available and address a number of potential shortcomings of these analyses.

Of particular note has been the enhancement of Bayesian (STAN) and fast Bayesian approximation (INLA) tools. We identified the following broad modelling frameworks for fitting generalised mixed effects models in the R statistical and graphical environment:

- **glmmTMB**: a frequentist approach based on Template Model Builder (TMB)
- **BRMS**: a STAN (Bayesian) model building interface
- **INLA**: a Bayesian approximation framework that provides very fast estimates.

The beta distribution can be useful for data that are bound at both lower and upper ends (as percentage cover data are). However, it is also a diverse distribution that takes on very different forms (shapes) over a range of parameter values. Consequently, one single beta distribution parameterisation might not be appropriate across all reefs and years. Some modelling routines (STAN and glmmTMB) now allow us to relax the assumption of a single beta dispersion parameter across the entire model.

Given the variety of possible modelling approaches, we elected to explore, compare and contrast each approach and present an ensemble:

- **Raw means**: simple means of means (no measure of uncertainty available)

- **Original**: binomial Bayesian mixed model (STAN) on reef level data - similar to De'ath et al 2021, yet in a Bayesian framework and modelling year as a fixed categorical effect rather than with splines.

```
stan_glmer(Cover ~ Year + (Year|Reef),
           family = binomial,
           iter = 500,
           warmup = 2500,
           chains = 3, cores = 3,
           data = data)
```

- **glmmTMB Beta**: beta mixed model on reef-level percent coral cover data. Comparing this model to the Original allows us to explore the impact of beta vs binomial modelling.

```
glmmTMB(Cover ~ Year + (1|Reef),
        family = beta_family(),
        data = data)
```

- **glmmTMB Beta disp**: beta mixed model on tow level percent coral cover data in which the dispersion beta parameter is permitted to vary each year. Comparing this model to the glmmTMB Beta model allows us to explore the impact of tow-level vs reef-level data.

```
glmmTMB(Cover ~ Year + (1|Reef/Reef*Year),
        dispformula = ~Year,
        family = beta_family(),
        data = data)
```

- **INLA Beta**: beta Bayesian approximation mixed model on tow level percent coral cover data. Comparing this model to the glmmTMB Beta model allows us to explore the impact of the Bayesian approximation vs the full Bayesian engine.

```
inla(Cover ~ Year + f(Reef, model = 'iid') +
         f(ReefYear, model = 'iid'), family = 'beta',
     data = data)
```

- **BRMS Beta disp**: beta Bayesian (STAN) mixed model on tow level percent coral cover data in which the beta dispersion parameter is permitted to vary each year. Comparing this model to the glmmTMB Beta disp model allows us to explore the impact of the glmmTMB vs STAN approach to 'random effects' modelling.

```
brm(bf(Cover ~ Year + (1|Reef/ReefYear),
       phi = ~0+Year,
       family = Beta(link = 'logit')),
    iter = 1e4, warmup = 5e3,
    thin = 5, chains = 4, cores = 4,
    data = data,
    prior = prior(normal(0, 3), class = "b") +
        prior(normal(0, 3), class = "Intercept") +
        prior(gamma(2, 1), class = "sd") +
        prior(gamma(2, 1), class = "sd",
              group = "REEF_NAME") +
        prior(gamma(2, 1), class = "phi")
    )
```

- **BRMS Cumulative**: ordinal Bayesian (STAN) mixed model on tow level categorical data. Comparing this model to the glmmTMB Beta and INLA Beta models allows us to explore the impacts of the cumulative vs beta distribution.

```
brm(bf(oLIVE_CORAL|weights(n) ~ Year+(1|REEF_NAME)),
    data = data,
    family = cumulative("logit",
                        threshold = 'equidistant'),
    iter = 1e4, warmup = 5e3,
    thin = 5, chains = 4, cores = 4,
    prior = prior(normal(0, 3), class = "b") +
        prior(gamma(1, 0.5), class = "sd") +
        prior(normal(0, 3), class = "Intercept")
    )
```

The glmmTMB Beta disp modelling approach was found to be the best performing model (from various residual based diagnostics). Importantly however, all models resulted in very similar (parallel) trends in coral cover, differing mainly in absolute value. Comparisons suggest that the original model deviated the greatest from the other models and that its estimates were consistency lower than the estimates of the other models.