IMR workshop on sdmTMB and spatial modelling, 23-25 May

Questions??

Please edit this in real time during the workshop. We'll do our best to answer in real time too!

1. **Where is the website for the workshop?**
    a. https://pbs-assess.github.io/sdmTMB-teaching/imr-2023/

2. **What do you do when your choice of mesh affects model results a lot?**
    a. Things to consider:
        i. Are predictions being made beyond the range of the data? [this is not something that should be done – if so, it might be useful to increase the border mesh – or increase the curvature making it more convex]
        ii. Are vertices << number of data points? [this should be true]. If your mesh with sdmTMB is called `mesh`, you can look at `mesh$mesh$n`
        iii. Are results converging to similar solutions as the mesh resolution increases? [this is something we expect]

3. **Should the cutoff of the mesh be the same as the range, or is it better that it is smaller?**
    a. Great question – I think you want to have the cutoff distance to be smaller than the spatial range

4. **How will the spatial temporal field be affected by varying sample site locations and amount of sample sites for every year?**
    a. The spatiotemporal fields will be generally unaffected by sample site locations or amount of data – years with less data will have wider CIs of estimated fixed effects (if present)

5. **What do you do if your data spans multiple UTM zones (say, 3)?**
    a. There's 2 options with sdmTMB. First, you could choose the zone a priori and specify manually for the entire domain. Or (2) you can add the add_utm_columns() function to add UTM X / Y coordinates (this would pick th UTM zone with the most data, and convert based on that)
        i. Great! Sorry, follow up: does that mean the error when using an adjacent UTM zone is negligible/acceptable?
            1. Eric: I think you'd probably want to check this post-hoc. Most of the areas I work in have 1 zone with a tiny sliver of a second, so we assume it to be negligible. But I have colleagues in Alaska working across 4-5 UTM zones – so there may errors in doing this. Two options: (1) use something like an Albers projection, which might help, or (2) use INLA's `INLA::inla.mesh.create()` function which takes a `globe` argument

6.One of the advantages of using random factors vs fixed factors is to save degrees of freedom. How are these types of spatial or spatiotemporal fields performing in this respect, i.e. how many replicates are they "eating"?

a. Good question – seems like the question is trying to look at the estimated degrees of freedom. This will depend on the amount of spatial autocorrelation, number of data points, etc. In general, these approximate df will be much smaller than if you were to say create a spatial grid a priori and estimate fixed effects for each stratum (grid cell)

7. is the default option for tweedie() in sdmTMB to automatically estimate the power parameter (as in tw() in mgcv)?

a. Yes, good question – by default it is estimated. It can also be fixed though

8. Can AR(1) parameters in the spatiotemporal fields be fixed, rather than estimated?

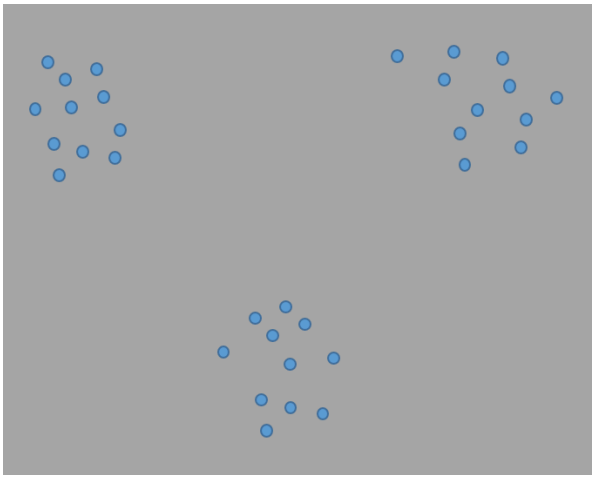
a. Yes! Anything can be fixed

9. How does this method perform compared to including correlation structures (in gamm or glmm) such as: "correlation = corGaus(form =~ SOUTH + WEST|Year,nugget=TRUE)". E.g. It is often a struggle to account for both spatial and temporal correlations in the same model (one hopes that either of them is negligible). I see this method as a way to solve the problem, would you agree?

a. Agree this solves the problem of accounting for temporal and spatial correlations. I'm not 100% sure about the model being fit with that corGaus() structure – but it seems like it's only fitting IID spatiotemporal fields, not spatial ones, correct?

   Well you can choose different options, like corAR1, corSphere…depending on how you want to model the correlation (sort of equivalent to choosing between IID, AR1 etc etc). However, a major limit is that it's not possible to model a temporal correlation simultaneously. Which I am not sure it is in sdmTMB either

10. Question about the mesh. How homogenous data does it require - can it be fitted to this type of spatial data?



Yes. I think it will create a somewhat finer “triangles” where you have the cluster of data and you will have bigger ones between. You can also create a mesh for simpler types of survey design (e.g. stations along a linear gradient)

11. Mesh discussion: can you elaborate on some instances when the mesh might not be appropriate?

a. Spatial non-stationarity: if spatial autocorrelation is changing over space (or even time) this could be a problem. Similar happens if the spatial variance is changing over time or space
b. Patchily distributed species shifting distribution year to year (e.g. mackerel, other pelagics)
c. Species that are changing their spatial distribution in response to the environment amd / or abundance (and the survey samples a different fraction of the population in each year)
d. Surveys non-overlapping in space year to year, or shifting distribution

12. When is it appropriate to include both spatial and spatio-temporal fields in the same model vs. using only spatio-temporal fields?

Great question – this will vary based on the dataset. I think a reasonable workflow is to generally try to (1) fit a spatiotemporal model, (2) fit a model with spatial and spatiotemporal fields. Did model # 2 converge? Are the variances for both estimated to be >> 0? If one is really tiny, such as the spatial variance, then it means the spatiotemporal variance is totally dominating the data – so probably ok to try a model without the spatial field

13. There is a technical problem:

> mesh <- make_mesh(predictor_dat, c("X", "Y"), cutoff = 0.05)
Error in fmesher.read(prefix, "manifold") :
File 'C:/Users/hirokos/AppData/Local/Temp/Rtmpait18S/fmesher2894ed45da3.manifold' does not exist.

How to solve this error?

Install newest R version, reinstall INLA, TMB and sdmTMB. See Brian's email from 22. May.

Brian’s email pasted here: Believe this is an error on IMR computers + the default directory for R to install packages. The following fix has worked on multiple computers (both windows).

Try to run:

install.packages("sdmTMB", dependencies = TRUE)

library(sdmTMB)

mesh <- make_mesh(pcod, xy_cols = c("X", "Y"), cutoff = 10)

If you get a confusing error message similar to “fmesher.read(prefix, "manifold") : file manifold' does not exist”, you can try the following steps. If you aren’t on windows, I think this arises with INLA and where R defaults to install packages… hopefully that’s enough for you to figure it out on your system.

1. Close R. Open a text editor (notepad) and create a file with the line (if at IMR, substitute your employee number. If not IMR, just point to your Documents folder).

R_LIBS_USER=C:/Users/a37789/Documents/R/win-library/%v

2. Save the file as “.Renviron” in My Documents. Make sure there is no extra file extension (deviously notepad will append .txt by default and windows explorer doesn’t show file extensions by default… that makes it not work!)

3. Open R and type “.libPaths()”. You should see something like C:/Users/a37789/Documents/R/win-library/4.3 (or 4.2). **If you see /appdata/local/ that’s no good**.

4. Create the C:/Users/a37789/Documents/R/win-library/4.3 folder (i.e. go to my documents, make an “R” folder, open it and create a “win-library” folder, etc.)

5. Install INLA and sdmTMB again. Confirm that R installs them and all the dependencies in the folder you just made.

6. Try the basic vignette up to “make_mesh” again.

15. How do you recommend diagnosing spatial autocorrelation in your model? Visualizing residuals in space from a non-spatial model, or run tests (e.g., DHARMa::testSpatialAutocorrelation)?
I think both are a good idea. There are statistical tests like the ***Moran’s I*** statistic in R, https://stats.oarc.ucla.edu/r/faq/how-can-i-calculate-morans-i-in-r/ (and a number of other extensions that weight data differently)

If you’re interested in the ***spatial variogram***, there are a number of ways – e.g. these FAQs. Packages like gls() are one way to fit these –
https://gsp.humboldt.edu/olm/R/04_01_Variograms.html
https://stats.oarc.ucla.edu/r/faq/how-do-i-generate-a-variogram-for-spatial-data-in-r/

Spatial spline correlograms of residuals
see example Fig A6: https://cdnsciencepub.com/doi/full/10.1139/cjfas-2018-0281#fa1
“Spatial spline correlograms of residuals from the (A) binomial and (B) positive components of the delta-models. Positive and negative values indicate positive and negative spatial autocorrelation (Moran’s I), respectively, and y = 0 is the expected value under the null hypothesis of no spatial autocorrelation. The distance where the autocorrelation intersects (y = 0) is the decorrelation distance, which is roughly 40 km for the Hawaii longline species and 5 km for the West Coast groundfish species. Calculations were made by package “ncf” version 1.2-5 (Bjornstad and Falck 2001).”

16. Presence only models?
see bonus lectures at
https://github.com/ericward-noaa/sdmTMB-teaching/tree/main/imr-2023

built slides:
https://pbs-assess.github.io/sdmTMB-teaching/dfo-tesa-2023/
https://pbs-assess.github.io/sdmTMB-teaching/dfo-tesa-2023/13-presence-only.html#1

17. Is there an easy way to convert mesh into a grid for predict? Say we want to plot spatial model results for even-sized grid covering the region of the original data. In the examples on the website (e.g. here) you load qcs_grid when you demonstrate how to plot the spatial model. In most applications, the grid does not exist by default.

Agree that converting the mesh would be nice. Uncertain whether that is possible, but it is relatively straightforward to make a projection grid using the sf package. E.g.

```
library(tidyverse);library(sf);library(sdmTMB)

tmp <- dt %>%
  st_as_sf(coords = c("X", "Y"), crs = suppressWarnings(sdmTMB::get_crs(dt, ll_names =
c("lon", "lat")))) %>%
  summarise(geometry = st_combine(geometry)) %>%
  st_convex_hull()

proj_dt <- tmp %>%
  st_make_grid(cellsize = rep(100,2), square = TRUE, what = "centers") %>%
  st_sf() %>%
  st_filter(tmp) %>%
  st_coordinates() %>%
  as_tibble()
```

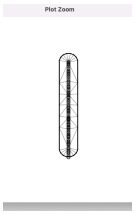
We'll talk about this today/tomorrow! Some options for making a prediction grid:

- cover all the data locations at a specified resolution, eg 10 km. Currently the day 3 ex does this. Code: https://github.com/pbs-assess/sdmTMB-teaching/blob/main/imr-2023/exercises/coastal-survey-ex/coastal-survey-index-south.R#L205. Output plot: https://github.com/pbs-assess/sdmTMB-teaching/blob/main/imr-2023/exercises/coastal-survey-ex/plots/south/preds_2years.png
- for a survey with strata, create a regular grid and then clip to strata polygons
- use mesh vertices: https://github.com/pbs-assess/sdmTMB/discussions/230

18. Can sdmTMB be used for spatial interpolation when the primary sampling unit is a transect, rather than a point in space?

The short answer is yes. Using the simulation function we were working through, we can create a new data frame where 'X' is constant, but the transect varies in the `Y` direction

```
predictor_dat <- data.frame(
  X = 0, Y = runif(300),
  a1 = rnorm(300), year = rep(1:6, each = 50)
)
mesh <- make_mesh(predictor_dat, xy_cols = c("X", "Y"), cutoff = 0.1)
plot(mesh)
```



19.If we want to use sdmTMB/tmbstan in a bayesian way, is it possible to bring in previous distribution ranges or other spatial data in as priors, for example from a previous study?

Yes - it is possible to add priors – see more on Bayesian estimation here:

https://pbs-assess.github.io/sdmTMB/articles/bayesian.html

In particular, and even for maximum likelihood models, we've found it's often really helpful to put vague priors on the Matern parameters

20. How to test the significance of covariates in a sdmTMB model? Is it possible to do model comparison (via AIC or other metric) between nested models, both to check for significance of covariates as well as to see whether a spatial field improve the model fit?

Yes, check the SE and confidence intervals for the parameters in the model fit. More formal model selection with cross-validation is coming soon, lecture 3 today: https://pbs-assess.github.io/sdmTMB-teaching/imr-2023/06-comparing-models.html#1 (this morning!). We can also do some model averaging with stacking

21. In the case where we fit a model with the aim of making predictions into the future (i.e., extrapolating in time), would it make sense to include spatiotemporal fields? If so, how are they dealt with in the new time points (e.g., years)?

Great point – I will make sure to mention this. The short answer is it depends on the time horizon. If you're making predictions a year in advance, and you're using spatiotemporal fields that are not IID (they are either random walks, or AR(1) fields) then the fields can be projected. For much more than 3 time steps, you're right that any spatiotemporal fields can blow up (variance increases) – so this is not a good approach if you wanted to do say ~ 20 or 50 year projections

22. If you fit a model with spatial and spatiotemporal fields, and saw spatial patterning persistent in the residuals, why might that be?

Spatial patterns might remain in the residuals in a number of cases. First, it could be that there's spatial non-stationarity, and the assumption of constant spatial parameters is flawed. Second, it could be that the mesh resolution is too coarse – a finer mesh might be able to better capture nuanced spatial patterns.

23. **A model passes all residual tests and other diagnostics (imagine you have spatial + spatio-temporal fields + several covariates modeled as splines) but at the time of making prediction over the prediction grid, some estimates (est) explodes (few magnitude higher than what is reasonable and the next highest value). it appears that it is due to the values of the covariate at that prediction grid (e.g. imagine SST and oxygen concentration and both had exponential shape form (in log-link scale) with higher values leading to higher response. And one prediction grid had both high SST and oxygen). What to do?**

I can try… when you say the prediction estimate isn't reasonable, that says to me that you think the model isn't reasonable or you are predicting beyond your observed data (in space, SST, O2, or combination). If you think the prediction for high SST should be different if O2 is high/low, then maybe you should have an SST:O2 interaction term in the model. Or if SST and O2 are linear effects, maybe these should be more complex (quadratic or spline)?

Adding a bivariate smoother could also be a possibility to explore. It can show that these high SST and O2 might rarely happen in the data thus, change the combined effect of SST and O2 for example for these "out-of sample" prediction grid.

24. Is there a straight forward way in sdmtmb to do aggregated/stacked species distribution models like shown here (https://onlinelibrary.wiley.com/doi/10.1111/geb.12102)?

25. How could you make `est_non_rf` only include a single linear covariate

In a model with multiple linear covariates, if each was z-scored (or held at their means if not the case), then you could create a prediction grid with different values of the single variable of focus - and then est_non_rf would be the impact of that covariate

26. I understand that higher elpd is better. But what is a meaningful difference in elpd?

Depends in part on the precision of the elpd estimates. If the elpd for two models is quite variable by CV fold, for instance, then not likely to be meaningfully different. See https://avehtari.github.io/modelselection/CV-FAQ.html#se_diff

sdmTMB doesn't calculate se(elpd) though. This is something we should build in in the future!

27. Rule of thumb for a "good" number of folds to use in CV?

k = 10 is often used. Results more sensitive to blocking (eg in space, time) than number of folds. We recommend checking out the `blockCV` or `spatialsampling` packages, and see Valavi et al. 2019, https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13107

28. Why use block cross validation?

If you don't, could get biased estimates of model predictive ability, incorrect inference on model selection, etc.

Couple good references:
https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.02881
https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13107
https://cran.r-project.org/web/packages/blockCV/index.html

29. Multiple ages?

paste age_year as the time index in a spatiotemporal model, then each interaction of age and year will get a spatial field.

30. Follow-up question to 29.: what would be the implication of using AR1 on age_year as spatial fields? Would that imply an AR1 correlation across ages within a year and across years?
If `age_year` is passed in as the `time` argument, then the interpretation is that each age in each year has a unique spatiotemporal distribution. Another way to set this model up would be to use "year" as the time index of the spatiotemporal field, but include age effects as spatially varying coefficients. These models would then assume a unique spatial distribution for each age class – as well as shared spatiotemporal deviations by year that are experienced similar by each age

31. The tidy(model) function seems to not output anymore estimates of smoothers in the latest version, neither in the "fixed", "ran_pars" or "ran_vals". Is this a bug or intentional?

The tidy does not give estimates on the smoothers (yet) – this is an issue and something we haven't had time to fix – but will soon

32. How would you analyze line transect data (i.e. effort along lines, detections as points) with sdmTMB? Should the effort lines be segmentized (equal length, using centroid long/lat as reference point) or gridded for this purpose?

There's probably some interesting work to be done about how to construct the mesh on line-transect data – a similar question was asked above, and a simple mesh example created. But I don't think the effort lines need to be segmented of equal length – but maybe downsampled to represent some regular sampling frequency? By 100m or 1km intervals for example? Maybe this is what the question is asking

33. How would you estimate the uncertainty across an integration grid for a spatial-only model? get_index() provides this functionality (95% CIs) if there is a time component (at least for certain model types). What is the best way to produce the CIs without the time component or if the response is not abundance/biomass but e.g. proportions or probabilities?

You might want to do this manually, depending on what you're trying to generate. If the model is presence/absence, are you interested in a metric of total cells occupied? Or maybe average occurrence? Either way, I think this would need to be done manually – make the predictions (and standard errors) at the prediction grid, and then manipulate the output yourself on the R side

Follow-up question: Assuming we would like to get for instance average occurrence and confidence intervals, how should the standard errors then be aggregated/weighted across the integration grid to get an appropriate representation of uncertainty? Would it be possible to use instead the simulation function in bootstrap approach, simulating a large number of spatial predictions and deriving the uncertainty from the distribution of average occurrences across x simulations?

If you wanted to fully propagate all the uncertainties, it seems like you could

(1) fit the model and predict to a regular grid, getting your SEs out as well
(2) Draw random normal numbers using the mean and SE from each grid cell
(3) Back transform them with plogis() to get proportions
(4) Draw 1 random bernoulli / binomial draw from each grid cell and calculate the average occurrence across cells
(5) Repeat 2-4 1000 times

34. We saw examples of models with a spatial field, a spatiotemporal field, and a spatial+spatiotemporal field. For the sake of curiosity, would it be possible to include only a temporal field (no spatial correlation assumed)?

Yes – you can do this in a couple ways. First, make sure 'spatial' and 'spatiotemporal' are turned off. Then you could model the temporal correlation as:

(a) with a smooth , e.g. s(year)
(b) with a time varying effect (this is a DLM), e.g. time_varying = ~ 1,

You can also change the `time_varying_type` to make the time varying model a random walk, AR(1) process, etc

35. A very basic question, what is the advantage/ difference of including spatial and temporal variation as random fields rather than fixed effects e.g. as covartiaes lat/lon, time in a GAM?

This is a good question. One way to think about this is that all GAM models can basically be re-written as GLMMs (this is what `brms` does, and also see this very accessible paper: https://ro.uow.edu.au/cgi/viewcontent.cgi?article=3526&context=eispapers) – so a GAM is estimating the temporal or spatial smooths as random effects, just as the sdmTMB / SPDE models are.

see also David Miller's paper: https://link.springer.com/article/10.1007/s13253-019-00377-z

36. What is the formula for the confidence intervals used by get_index()?

37. On the topic of **survey overlap + combining data**, please consult also the ICES report from WKUSER2 (https://www.ices.dk/community/groups/Pages/WKUSER2.aspx)

38. What if you want to combine 2 surveys from different seasons/months?

Eric said something and I didn't start typing soon enough…
perhaps could include season_year as a spatially-varying effect (similar to including age_year above #s 29-30)

39. In the spatially variant coefficient example, you plot slope deviations in space. Can you explain how to interpret these in terms of how large the effect is, for example in the log-length vs age example? I.e., how to get from slope deviations to the actual slope of the local linear regression, need to exponentiate, etc.

40. What exactly are you wanting us to do for this exercise? It all seemed very unclear to me. And are we supposed to be writing a report about it afterwards?