Lecture 18: Mapping gappy data

Recap

We've been looking at objective mapping as a means to grid irregularly spaced data. But what happens when data are extremely sparse, when we need to fill big holes in our sampling scheme?

The ocean heat content problem

When data are spaced more closely than a decorrelation scale, then objective mapping will produce plausible, smoothly varying maps. But if data are sparse, we might be in trouble. Sparse data are a common problem in oceanographic research, and especially in assessing average conditions prior to the modern satellite and Argo observing error.

In class we reviewed three specific examples:

Ocean heat content. Ocean temperature data collected from long hydrographic lines are gappy in space and time. Argo did a lot to remedy this problem, staring around 2005. But prior to that, gaps pose real challenges. Boyer et al (2016) provide a nice review of strategies. Among them

1. Levitus et al, 2000, carried out a straightforward mapping, replacing data gaps with climatology. The approach is entirely consistent with best practices in objective mapping—let the map revert to the mean when update data are not available. But since climatology is invariant, and much of the ocean is unsampled, this means that many locations have mapped values equivalent to the mean. The resulting trend risks looking small, since no data implies a zero or small trend.

2. Other strategies (e.g. Ishii et al, 2006 or Gouretski et al, 2012) have tried other strategies for optimizing mapping scales to vary mapping scales and try to fill gaps in the data more completely.

3. Willis et al, 2003, used the covariance between sea surface height anomalies from altimetry and subsurface variability measured by Argo to gain global information. This allowed them to fill gaps across the entire data record, albeit only during the altimeter era.

4. PMEL followed a strategy similar to Willis et al, not using altimeter data but assuming that at any give point in time, regions without data should have an anomaly similar to the mean anomaly in regions that were sampled. This allowed them to fill in unsampled areas to produce a consistent result.

5. The UK Met Office used a damped persistence appoach, setting an intial guess equal to the monthly climatology plus a damped term eachal to a damping coefficient $\alpha = 0.9$ (chosen experimentally) multiplied by the difference between the previous month's anomaly and the previous month's climatology. This damped correction prevents the background guesses from reverting too strongly to zero.

6. Domingues et al, (2008) used altimeter data to define overall patterns of variability, defined by the first 30 Empirical Orthogonal Functions. They then projected observed variability onto these EOFs to extend the analysis back in time prior to the start of the satellite era.

While the results of these strategies are not wildly different, the differing strategies clearly account for differences in estimates of global heat content increase and global sea level rise.

These aren't the only options. Notably, in the deep ocean, Purkey and Johnson (2010) assesed warming in basins deeper than 4000 m by assuming that any observations collected in the basin were more or less represenative of the entire deep basin.

Mapping pCO_2 with a neural net. While the ocean heat content mapping strategies are aligned with fairly traditional objective mapping and regression approaches, there are other options. In recent years, a plethora of tools have emerged to use machine learning strategies to map irregularly sampled data. One prime example involves neural net based approaches (e.g. Landschützer et al, 2013) to define biogeochemical provinces and to map CO_2 concentrations in regions that are poorly sampled by other data.

Mapping O_2 with random forest. In class we looked at the random forest approach used by Giglio et al (2018) to map sparsely sampled O_2 measurements by filling gaps using more densely sampled temperature and salinity. Random forest uses a non-linear regression approach that subdivides the data based on different criteria (e.g. temperature exceeding a certain threshold, or salinity less than some threshold) in order to develop an algorithm to determine how to fill missing values. A jupyter notebook for this is posted separately from these notes. Matlab routines also exist for these, but I have not written out examples.

Bibliography

Boyer, T., and Coauthors, (2016). Sensitivity of Global Upper-Ocean Heat Content Estimates to Mapping Methods, XBT Bias Corrections, and Baseline Climatologies. *J. Climate*, **29**, 4817-4842, doi:10.1175/JCLI-D-15-0801.1.

Giglio, D., Lyubchich, V., and Mazloff, M. R. (2018). Estimating oxygen in the Southern Ocean using Argo temperature and salinity. *Journal of Geophysical Research: Oceans*, **123**, 4280-4297. doi:10.1029/2017JC013404.

Landschützer, P., N. Gruber, D. C. E. Bakker, U. Schuster, S. Nakaoka, M. R. Payne, T. Sasse, and J. Zeng (2013), A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink, *Biogeosciences*, **10**, 7793- 7815, doi:10.5194/bg-10-7793-2013.

Purkey, S. G., and G. C. Johnson, (2010). Warming of Global Abyssal and Deep Southern Ocean Waters between the 1990s and 2000s: Contributions to Global Heat and Sea Level Rise Budgets. *J. Climate*, **23**, 6336-6351, doi:10.1175/2010JCLI3682.1.