Integrating massive and heterogeneous spatio-temporal data in environmental science.

In all the fields of environmental sciences, massive and heterogeneous datasets are progressively becoming available. These datasets open possibilities to infer spatio-temporal processes at a fine spatio-temporal resolution, and urge to address methodological challenges needed for this integration.

First, sampling is not always standardized ; sampled locations can be concentrated in areas of higher process intensity. This is often referred to as preferential sampling and it can lead to biased predictions if not accounted for in inference. Second, the data can be aggregated at various spatial scales. This is usually called change of support (COS), and no approach exists to handle complex data such as highly tailed or zero-inflated data in COS. Finally, the combination of these heterogeneous data sources requires a careful insight of their relative contribution to inference so as to avoid model misspecification.

In this presentation, I will present a generic framework that tackled these methodological issues. The approach is illustrated through applications in marine ecology, geography and climate science.

The spatial preference of sampling agents can depend on many factors. Consequently, our model accounts for the preference to the variable under study as well as additional covariates that can be included explicitly or implicitly in the model. The model we introduce allows to unbias model outputs while being generic and parsimonious.

The COS component of the model allows to account for complex data such as zero-inflated or highly tailed data. This is done by considering that aggregated data are a convolution of zero-inflated individual observations. Such component is crucial to estimate properly the spatial predictions and the relationship between the variable under study and additional covariates that are used to describe that variable (e.g. fish distribution can be described by sediment or bathymetry).

Finally, the combination of the distinct data sources is realized within a hierarchical framework. We assessed the consistency of the different data sets through a consistency statistical test. Typically, if COS is not accounted for properly in the model, the statistical test will reject consistency which emphasizes that one model component is misspecified.