Citizen science and dataset combination: two examples in ecology and environment

Camille Coron[1]

1 Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.

Keywords:
Citizen science; opportunistic data; dataset combination; species distribution; air quality monitoring;

Abstract: In this presentation we will focus on how to combine datasets of different nature, or resulting from different protocols, when estimating biological parameters. This question will be studied in two different situations, one coming from biodiversity monitoring (based on Giraud et al 2016 and Coron et al 2018), and the other coming from air quality assessment and monitoring (work in progress with Jean-Michel Poggi and Benjamin Auder, from Paris-Saclay University).
We will first consider different datasets of several bird species observations in Aquitaine (region of the South-West of France), thanks to which we wish to create, for each of the observed species, a relative abundance map (i.e. to be able to compare the expected number of individuals of the considered species at different locations or at different times). These observation datasets result either from accurate protocols (the dataset is then called "standardized"), in particular imposing to observers the location and duration of observations, or from citizen science programs through which professional or amateur ornithologists share their observations without location or time constraint (the dataset is then called "opportunistic"). The opportunistic dataset, due to the absence of constraint imposed to observers, provides much more data with a finer spatial coverage, but suffers from important and unknown bias, due to observers' behaviour. Our approach consists in jointly modeling these two datasets, in order to benefit both from the calibration brought by the standardized dataset, and from the abundance of the opportunistic dataset. Our main result states that, as soon as the opportunistic dataset is large enough, the estimated relative abundance map is more precise when combining both datasets than when using only the standardized dataset.
The second question of interest, which is a work in progress, deals with air quality assessment and monitoring, and more precisely with the production of $N0_2$ or particulate matters concentration maps, at different moments. To produce these maps, we can use a physico-chemical model (like SIRANE or CHIMERE) outputs, as well as some real concentration measures, made first by a few expensive fixed stations and second by a larger number of cheaper micro-sensors. We assume concentration measures provided by the fixed stations to be centered in the real concentration that we wish to estimate, whereas we expect measures provided by micro-sensors to be biased. Our approach consists in modeling the mean error of the physico-chemical model using some explanatory variables, and to estimate the

parameters of this mean error model using real concentration measures provided by both types of devices. We expect that micro-sensors will improve concentration maps precision.

References:

1. Giraud, C., Calenge, C., Coron, C., Julliard, R. Capitalizing on opportunistic data for monitoring species relative abundances. Biometrics 72 (2), 649-58 (2016).
2. Coron, C., Calenge, C., Giraud, C., Julliard, R. Estimation of species relative abundances and habitat preferences using opportunistic data. Environmental and Ecological Statistics **25**, pages71–93 (2018)