Analysing opportunist data in Citizen Sciences: statistical modelling for loose protocols

P. Monestiez [ab], J. Chadœuf [a], D. Pinaud [b], K. Le Rest [b], O. Filippi-Codaccioni [c], L. Couzi [c], V. Bretagnolle [b]

[a] INRA, BioSP Biostatistique et Processus Spatiaux,
Domaine Saint-Paul, Site Agroparc,
84914 Avignon, France
monestiez@avignon.inra.fr, joel@avignon.inra.fr

[b] Centre d'Etudes Biologiques de Chizé
UMR 7372, CNRS – Université La Rochelle
79360 Villiers-en-Bois, France
pinaud@cebc.cnrs.fr, breta@cebc.cnrs.fr

[c] LPO Aquitaine
433 Chemin de Leysotte,
33140 Villenave d'Ornon, France
ondine.filippi-codaccioni@lpo.fr, laurent.couzi@lpo.fr

Keywords: citizen sciences, bigdata, species distributions, spatial hierarchical model, INLA.

Abstract: Volunteers networks involved in citizen science (SC) programs provide a real opportunity to address conservation and species distributions questions on broad temporal and spatial scales. In the last ten years development of new technologies – smart-phones, friendly user websites – has dramatically increased the volume of collected data and the number of SC programs, but in most cases, with weakly defined and heterogeneous data collection protocols. Opportunist data in our case are data collected by a large number of different observers, whose spatial and temporal distribution is greatly heterogeneous, the effort is usually unknown, zeros are generally unreported, and finally positive count may be reported differently or even censured according to species. To analyse such data we propose a multivariate hierarchical model with latent spatial – spatiotemporal – fields for relative abundances of each considered species. Its specificity is to account for different types of observation and for observer characteristics in distribution or behaviour. First results show that it seems possible to correct several main biases, to model count positive-only data and to infer fairly well relative density maps in a multi-species context, using a Bayesian framework and INLA R-package tools. We analysed a case study data set of several thousand observations from the French Ligue pour la Protection des Oiseaux (LPO, Birdlife France) to show the feasibility of such approaches and we checked the inference quality and limits on smaller simulated examples.