Modélisation statistique des données d'observation issues des Sciences Participatives

Pascal Monestiez [1] & Groupe de Travail CiSStats

[1] *BioSP, INRA, Site Agroparc, 84914 AVIGNON. pascal.monestiez@inra.fr*

Résumé. Le nombre de programmes d'observation participative a fortement augmenté ces dernières années, portant sur des groupes d'espèces de plus en plus diversifiés, et avec des volumes de données en forte croissance du fait des saisies directes sur sites web ou smart-phones. En volume, ces données dépassent désormais largement les capacités des études scientifiques institutionnelles, et deviennent en conséquence incontournables tant sur le plan de la connaissance scientifique que pour les gestionnaires d'espaces naturels. Ce développement rapide n'empêche cependant pas une grande hétérogénéité des protocoles, ni l'absence de planification dans la plupart des cas, engendrant potentiellement de nombreux problèmes. De fort biais peuvent apparaitre dans les résultats même pour des données en très grand nombre. Après une analyse des types de données rencontrés dans différents programmes existants, nous présenterons globalement les solutions actuellement développées sur le plan statistique pour en extraire des distributions spatio-temporelles par espèces, ainsi que des éléments caractérisant la biodiversité. Des approches basées sur des modèles hiérarchiques bayésiens avec des champs spatiaux latents et des modèles d'observation permettent de travailler sur des sources de données hétérogènes et de qualités variables. Les perspectives dans le cadre multivariable avec des ensembles d'espèces sont abordées. Nous montrons qu'une bonne connaissance du processus d'observation et des observateurs bénévoles eux-mêmes est centrale dans la validation et la valorisation de ces données. Dans le cadre des plateformes intégratives au niveau national ou européen, il devient essentiel de ne pas agréger les données de manière irréversible ni d'omettre ce qui caractérise leur source.

Mots-clés. Analyse des données, fouille de données, Méthodes bayésiennes, Grande dimension, données massives, Statistique spatiale, spatio-temporelle