Identification uncertainty and probabilistic classification methods: from DNA sequences to bird species identity

U. M. Camargo[a] and O. Ovaskainen[b]

[a] Department of Biosciences, University of Helsinki, Finland
ulisses.camargo@helsinki.fi

[b] Department of Biosciences, University of Helsinki, Finland
otso.ovaskainen@helsinki.fi

Keywords: survey design and analysis; big data; monitoring of biodiversity.

Abstract: Just as next-generation sequencers have fueled research in bioinformatics, autonomous recording units are stimulating the need for new statistical methods related to bioacoustics, especially in data processing, such as automated species identification tools. When collecting bird vocalization data, *e.g.* for studying Amazon bird communities, the use of numerous autonomous recorders as compared to manual listening by field observers can be compared to the revolution of massive parallel sequencing (MPS) and conventional sequencing methods. MPS produces extensive data sets with a fraction of cost compared to Sanger sequencing, but the analysis of MPS data calls for sophisticated bioinformatics methods. Similarly, autonomous audio recording produces extensive datasets, but there are major challenges to extract relevant information from data in a reliable manner. Species identification through genetic barcoding is based on comparing the similarity of a query sequence to reference sequences obtained from well-identified samples. Similarly, acoustic species identification is based on comparing the similarity of a query vocalization to those present in a reference database. Both approaches have traditionally been based on arbitrary threshold levels of similarity between samples, resulting on a poor description of identification uncertainty. In this talk I will present a Bayesian approach to assess identification uncertainty from DNA barcoding data and illustrate how I am adapting this framework to the automated identification of bird vocalizations. A systematic quantification of identification uncertainty is a crucial component when making reliable biological inferences from acoustic data. Using probability as the measure of uncertainty makes it possible to integrate the information into state-space models of community dynamics, propagating identification uncertainty through the entire analytical pipeline from the raw data to the biological inferences.