Supervised vs. non-supervised hidden semi-Markov modeling for inferring behavioral modes from movement paths

R. Joo[a], **S. Bertrand**[b] and **R. Fablet**[c]

[a]Instituto del Mar del Perú
Callao, Peru
kaoridrummer@gmail.com

[b]Institut de Recherche pour le Développement
Research Unit 'EME' UMR212
Sète, France
Sophie.Bertrand@ird.fr

[c]Institut Telecom/Telecom Bretagne
UMR 6285 LabSTICC, TOMS (Statistical Signal Processing and Remote Sensing)
Brest, France
ronan.fablet@telecom-bretagne.eu

Keywords: movement ecology; modelling of animal movement; Markov models; machine learning; trajectories; Vessel Monitoring System; fishermen; *Engraulis ringens*

Abstract: In movement ecology, the inference of behavioral modes of individuals from their trajectories is commonly addressed using hidden Markov models (HMMs). Recently, hidden semi-Markov models (HSMMs), extensions of HMMs, have been applied to animal and human movement, showing better fit and better inference of behavioral modes, respectively, than HMMs. HSMMs are usually fitted within a non-supervised framework, using an EM algorithm. Availability of tracking data for which the true behavioral modes are known, i.e. ground-truthed data, would allow for model validation, as well as for a supervised setting for HSMMs fitting. One of the few predators for whom we can have access to true foraging behavioral modes in a natural environment is the fisherman. Here, using ground-truthed data collected on fishermen in the Humboldt Current System, we are able to fit HSMMs within a supervised setting and perform an independent validation of the model. For testing whether if a supervised setting improves inference accuracy, we compare the performance of supervised and non-supervised HSMMs. Results support the use of supervised setting therefore highlighting the importance of ground-truthed data for model validation and higher inference accuracy. Further perspectives are presented, including the use of supervised and semi-supervised approaches for other animals' movement and the estimation of the resources (i.e. number of observations) to be allocated for gathering an optimal ground-truthed dataset.