Supplementary Material. State-space models with HMM from: Harvesting trajectories in large-scale coffee plantations: ecological and managment drivers and implications

Emilio Mora Van Cauwelaert, Denis Boyer, Estelí Jiménez-Soto and Mariana Benítez

0. Introduction to State-space models with HMM

State space models coupled with hidden Markov models (HMMs) are models in which the distribution that generates an observation Z_t depends on the state S_t of an underlying and unobserved Markov process (Fig.1) (Zucchini 2016). In this sense, the observations Z_t can be retrieved from one to multiple distributions related to the number of states. In the context of animal movement, the state S_t is interpreted as a proxy for the behavioral state of the animal (e.g. foraging, exploring; (Michelot, Langrock, and Patterson 2016)). The observations Zt are bivariate time series ($Zt = (l_t, \phi_t)$) where l_t is the step length (the Euclidean distance) and ϕ_t the turning angle, between two succesive locations Zucchini (2016). Biologically speaking, these models try to include the process where the movement of agents (e.g. short or large steps) depends on its behavior. With these models, we can define the different originary distributions and then decode the most likely sequence of states along the trajectory of an agent, its average time within a state, and the number of switches between states.

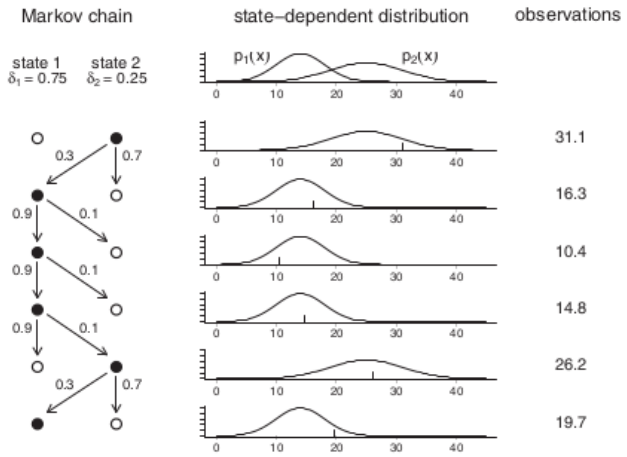

Figure 1: Process generating the observations in a two-state HMM. The chain followed the path 2, 1, 1, 1, 2, 1, as indicated on the left. The corresponding state-dependent distributions are shown in the middle. The observations are generated from the corresponding active distributions. Taken from (Zucchini, 2016)

Here we used the *movehmm* library (Michelot, Langrock, and Patterson (2019)) to i) fit the most likely distributions and HMM to the harvester movement data, ii) describe the different states and iii) decode and compare the sequence of states in the two farms (Ecological and Conventional) with the Viterbi algorithm.

1. Fitting of HMM

1.1. Preparation of data

We loaded *ggplot2*, *dplyr*, *tidyverse* and *movehmm* libraries. We then loaded the data of the trajectories and did some punctual modifications to it. We had in total 12 trajectories, with (x, y) coordinates (Fig. 2).

```
## Movement data for 12 tracks:
## E_1 -- 100 observations
## E_2 -- 80 observations
## E_3 -- 107 observations
## E_4 -- 76 observations
## C_1 -- 77 observations
## C_2 -- 70 observations
## C_3 -- 115 observations
## C_4 -- 63 observations
## C_5 -- 97 observations
## E_5 -- 104 observations
## C_6 -- 88 observations
## E_6 -- 158 observations
## No covariates.
```

A first assumption for the analysis was made: we took these time irregular trajectories (where each point represents one different tree, but where the time between two trees is variable) and treated them as a regular trajectories. This aimed to analyze the change in the movement of the workers during a day, generated by the underlying pattern of trees or by the differences in the fruit charge and ripening synchronicity. We convert this database into a *movehmm* object where the step distance l_t and the relative angle ϕ_t between the (x, y) coordinates is calculated (table 1). Now, for the following analysis, we only took the step distance and treated both farms as one. We decided to exclude the angles because they didn't have a biological meaning for irregular trajectories. The input for the analysis is then the histogram presented in Fig. 3. The values of the steps lengths ranges from 1 to 117, with a mean of 5.14 and a median equal to 4.

Table 1: First lines of the movehmm object with harvester data

ID	step	angle	x	y
E_1	3.605551	NA	146	117
E_1	3.000000	-2.158799	143	115
E_1	3.605551	2.553590	143	118
E_1	2.236068	-1.446441	141	115
E_1	5.000000	2.034444	139	116
E_1	3.605551	-2.158799	139	111

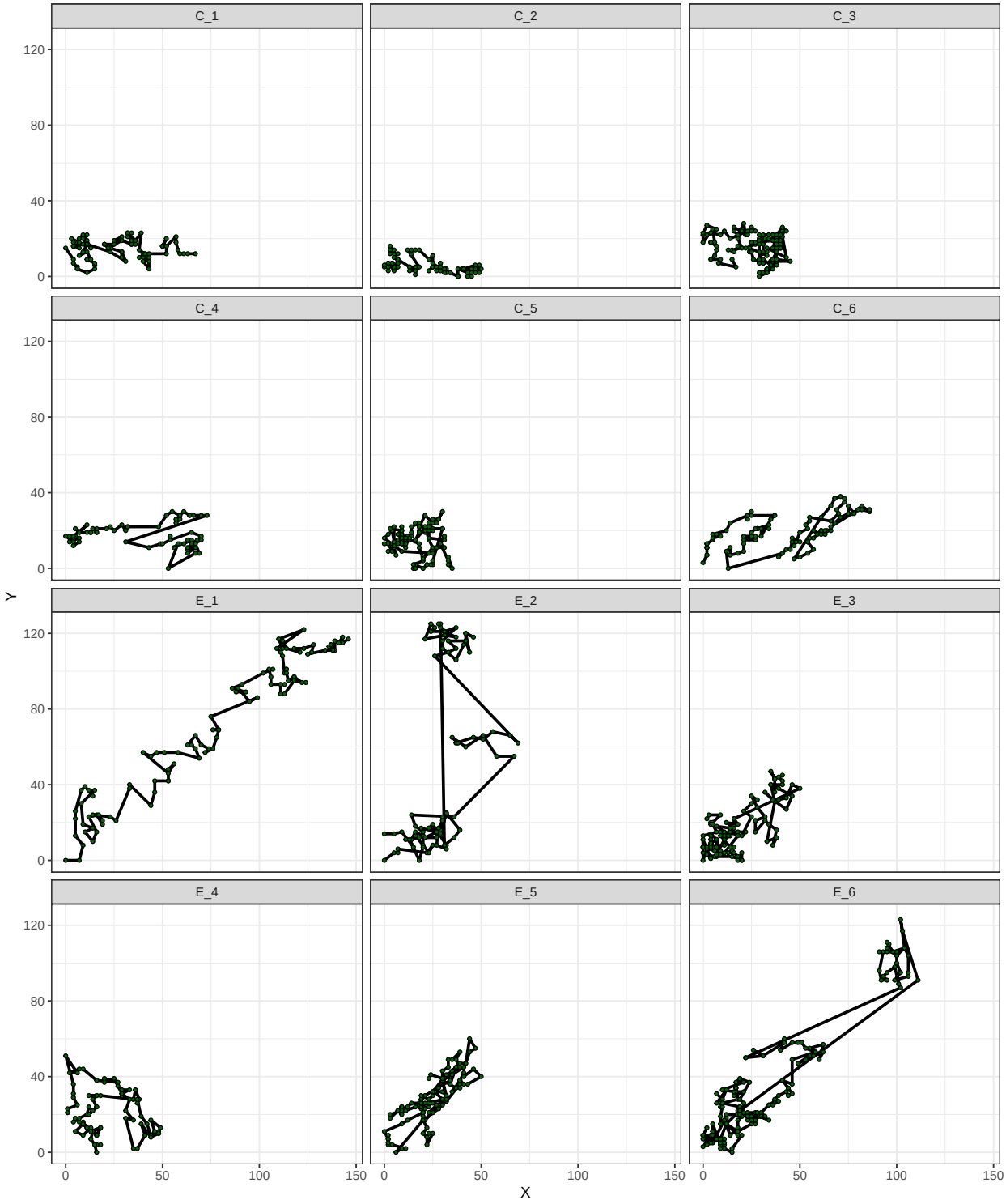

Figure 2: 12 trajectories of harvesters for both farms (E: Ecological, C: conventional). Each green dot represent a tree

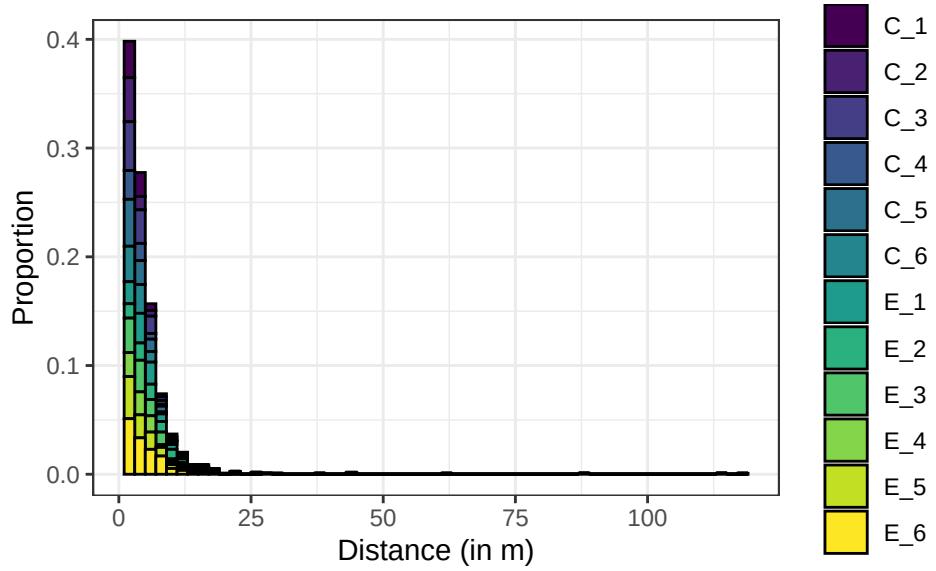

Figure 3: Histogram of the steps lengths divided by the different trajectories.

1.2. Fitting of the data

The algorithm of the *movehmm* library uses:

a) a predefined number of states. Here we will use two. But the algorithm also explores the case when both states are equal.

b) a defined family of distributions. Here we will try to fit two different exponential families (*gamma* and *weibull*).

c) prior parameters for each of the distributions, for each state.

d) The observations (step-lengths) (the data of harvesters)

With these inputs, the algorithm extracts the most likely two-state-distributions with their respective parameters (within the predefined family of distributions) from the data. During the fitting process, the algorithm uses the maximum-log likelihood and the *forward algorithm* (a recursive algorithm starting with the prior distributions; (Zucchini 2016)). We ran a loop to fit the data to two different families (*weibull*, and *gamma*) and in order to avoid local maximal likelihoods, we swapped 1000 combinations across a range of prior parameters for each model (Michelot, Langrock, and Patterson 2019). For the *gamma* distribution, the minimum and maximum values of each range were chosen to encompass the full distribution (including values below and over the mean). For the *weibull* distribution, the maximum shape considered the left-skew shape and the scale was chosen to include the 120m long steps (see the ranges in table 2). Many of these combinations of prior parameters converged to the same final distributions.

Table 2: Minimun and maximum values of the prior parameters for each distribution

value	gamma.mean	gamma.sd	weibull.shape	weibull.scale
min	0.1	0.1	0.0	0.1
max	20.0	20.0	2.7	15.0

Some combinations resulted in distributions with zero variance. We decided to remove those cases as they did not make any biological sense. We also removed states with means higher than the ranges, for visualization

(this does not change the results as they had maximal AIC that would be removed anyways). We then plotted the minimum negative log of the likelihood without these outliers (Fig.4). The minimum negative log-likelihood is equivalent to the maximum likelihood (Zucchini 2016). Almost all the combinations of prior parameters converge to two models with equivalent minimum likelihood for both distributions (Fig. 4). In this sense, the parameters of the models with minimal likelihoods are robust to the initial parameters. In particular gamma distribution resulted in models with a lower min neg likelihood (and lower AIC as show in table 3.)

We also explored the differences in the estimated parameters between the resulting states and their AIC. This corroborates that the models with the minimal AIC have two clearly distinct states (purples dots in Fig. 5). Besides, we ran all the analysis with one-state distributions for both families and compared them with our two-state distribution models using the minimal AIC (table 3). As we note in table 3, two-state distribution have the minimal AIC, and particularly with the gamma family.

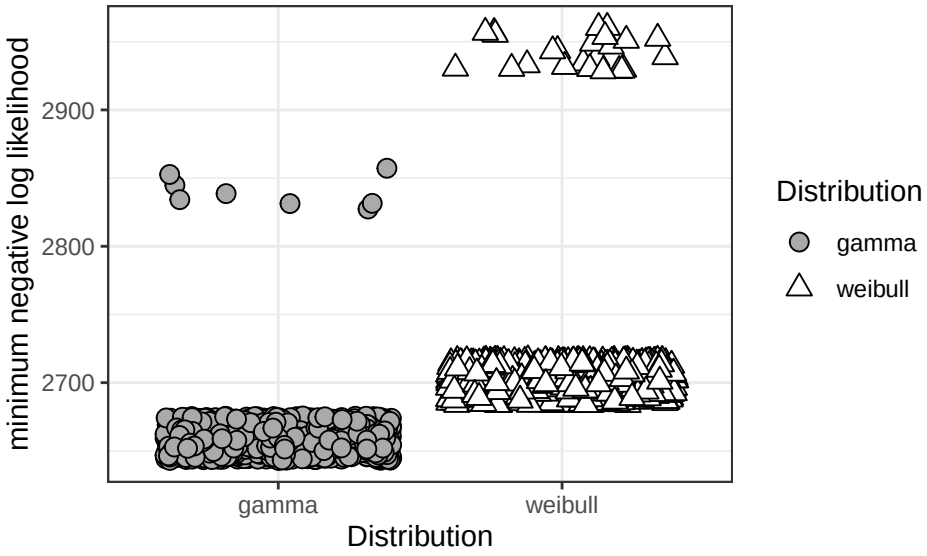

Figure 4: Minimun Negative Log-likelihood per combination of prior parameters for two different distributions

Table 3: Best models for each of the distributions. We added the prior parameters, the likelihood, AIC criteria and the final parameters. pr:prior, st1: state 1, p1: parameter 1 (shape for weibull and mean for gamma, p2: parameter 2 (scale for weibull and sd for gamma).

States	model	pr_st1_p1	pr_st1_p2	pr_st2_p1	pr_st2_p2	minNegL	AIC	s1_p1	s1_p2	s2_p2	s2_p2
2	gamma	1.94	5.61	6.46	1.15	2659.24	5332.48	17.34	15.18	4.28	2.25
2	weibull	2.20	0.80	5.10	13.20	2700.20	5414.39	9.66	6.74	15.17	14.81
1	weibull	1.43	4.18	NA	NA	2944.47	5892.95	1.22	5.57	NA	NA
1	gamma	17.54	19.98	NA	NA	2843.43	5690.86	5.14	3.54	NA	NA

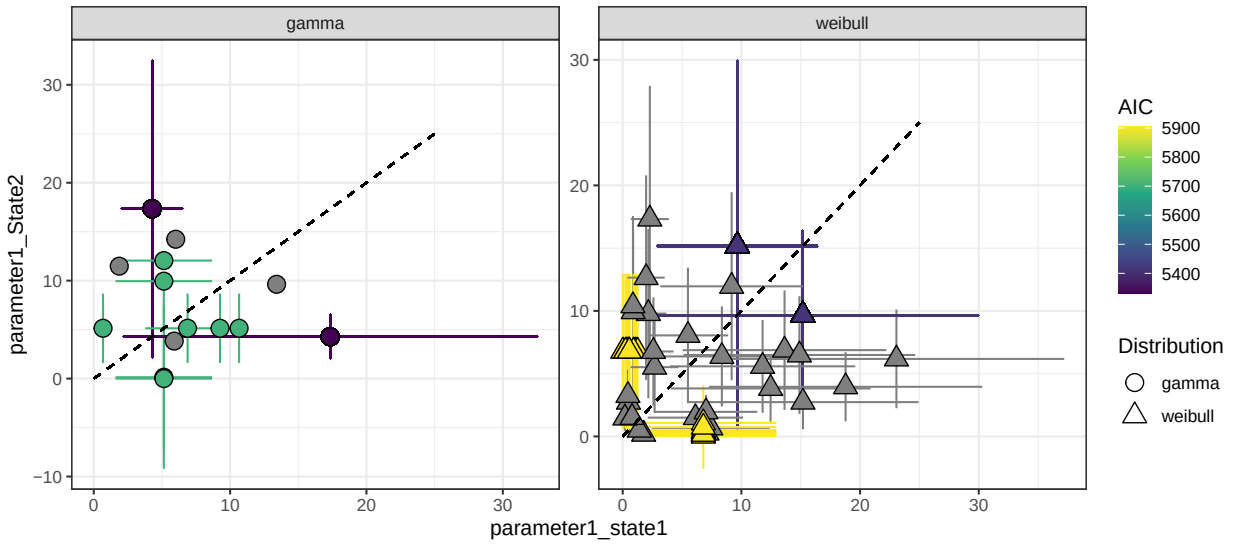

Figure 5: Parameters space for both states for each of the distributions. The (x,y) coordinate of each dot represent the mean of state 1 and 2, and the horizontal and vertical errorbar their standard deviation. The dotted line represent the identity, where both states are equivalent. The color represent the AIC value, where gray values show infinitely big values. The purple dots represent the minimal AIC

2. Description of the states, for each studied distribution.

We plotted the two state distributions for *gamma* and *weibull* families with the obtained final parameters (see figure. 6 and table 4). For both families, one of the states produced a distribution that encompass most of the highly frequent short steps and falls abruptly for size steps bigger than 13 m. The other state generates a long tailed distribution with a lower probability of short steps (compared to the other distribution) and a non zero probability for steps bigger than 13 m (Fig. 6 and table 4). This results show that from the movement of the harvesters (observations Z_t) we can extract two hidden markov states (S_t). The first state generate distributions with short steps (we can interpret this as harvesting the closest tree) and the second state generate longer steps (when the worker has to move to another part of the plot). This second state is highly unlikely (3.7% of steps in gamma distribution, 5.2% for weibull). Now, for the following analysis, we took only the gamma distribution as it presented a lower AIC.

Table 4: Final parameters for both gamma and weibull distributions.
For the gamma distribution $\alpha = \mu^2/\sigma^2$; $\beta = \mu/\sigma^2$; $\theta = 1/\beta$.

Parameter	Gamma_St1	Gamma_St2	Weibull_St1	Weibull_St2
μ (mean)	17.30	4.28	–	–
σ (sd)	15.20	2.25	–	–
α (shape)	1.30	3.62	2.05	1.05
θ (scale)	13.40	1.18	4.71	14.5
β (rate)	0.07	0.84	–	–

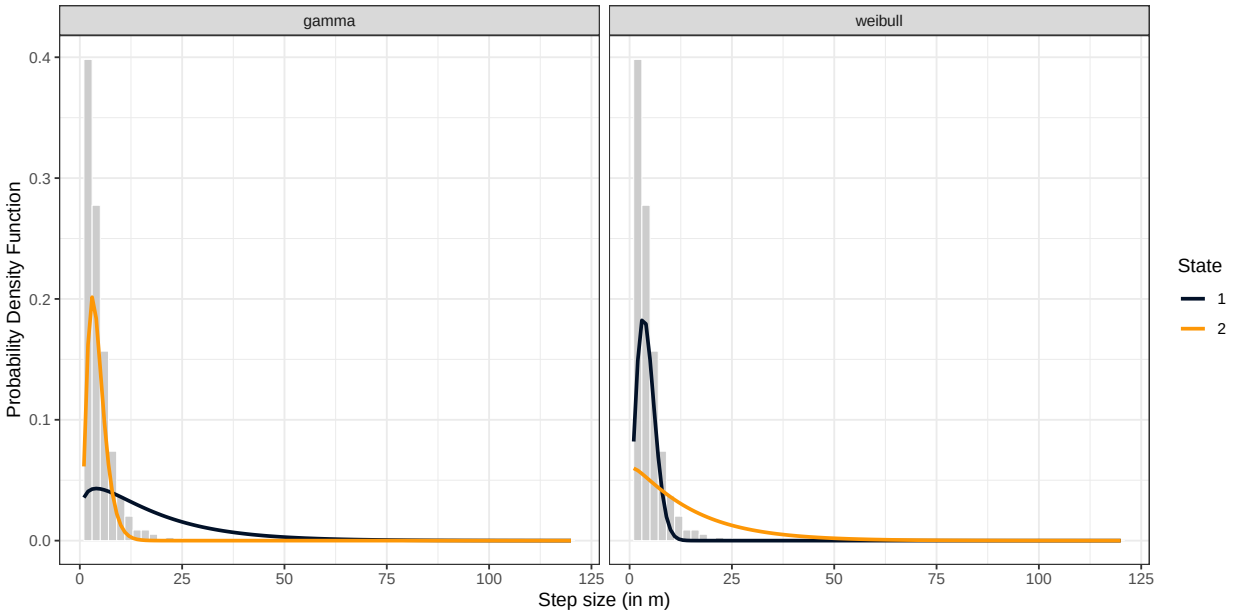

Figure 6: Two state model distributions. The used parameters are shown in Table 4

```
## Decoding states sequence... DONE
```

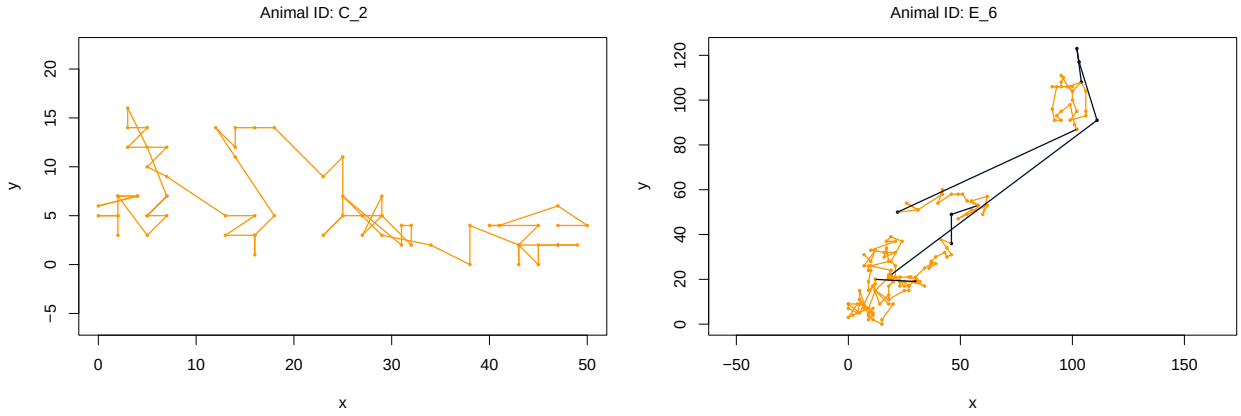

Figure 7: Distribution of states St along two trajectories. Each color follow the code of colors of Fig.6 for the gamma distribution)

3. Sequence of states and relation with the farms.

With the Viterbi algorithm, the package *movehmm* decodes the most likely sequence of states (assuming a markov chain) and the transitions matrix between hidden states. This takes into account the conditional probabilities between the observations and hidden states $P(Z_t|S_t)$) (Zucchini 2016). In this sense, *movehmm* can estimate the state with the highest probability for each step but this might not be the same as the state in the most probable sequence returned by the Viterbi algorithm (Fig. 8, second and third row vs first row). This is because the Viterbi algorithm performs "global decoding", whereas the state probabilities are "local decoding"Michelot, Langrock, and Patterson (2019).

```
## Decoding states sequence... DONE
## Computing states probabilities... DONE
```

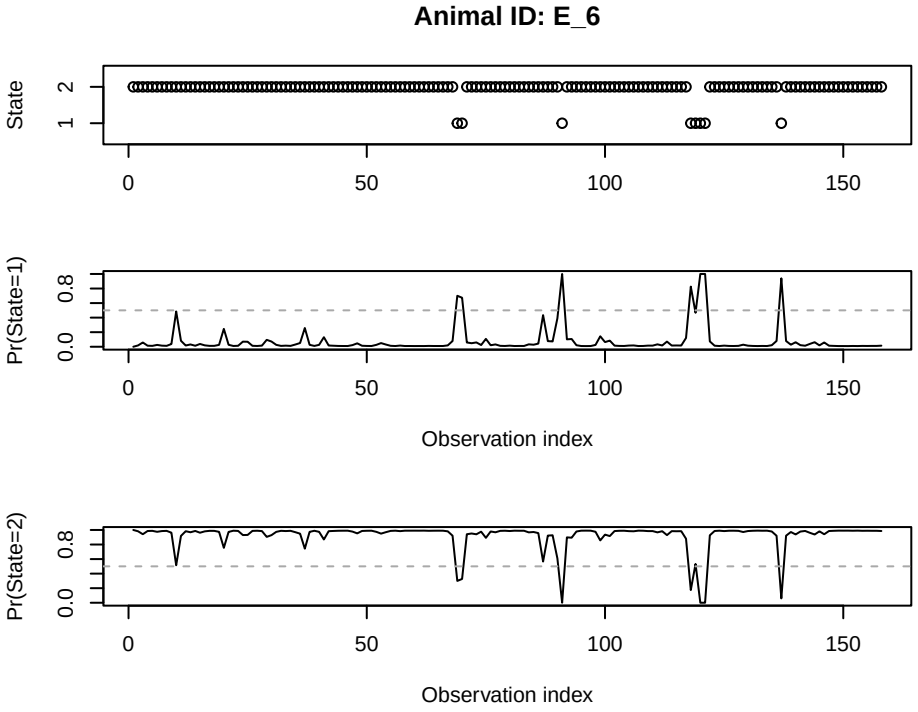

Figure 8: State probabilities for each step of E_6 trajectory and result of the Viterbi algorithm

We finally plotted the sequence of the most likely states along each of the trajectories (Fig. 9) and estimate the probability of the most unlikely state (state 1, that represent the relocalization of the workers in the plot)- according to the farm the trajectory belonged to (Fig. 10). We note that the percentage of steps from state 1 correlates with the identity of the farm (Fig. 10): 4 out of 6 trajectories that belong to the Ecological Farm, presented a higher than the mean percentage of longer steps, compared to the Conventional Farm, where 5 out of 6 were below the mean. (**note**: in the main text the number of the states was inverted for practicality)).

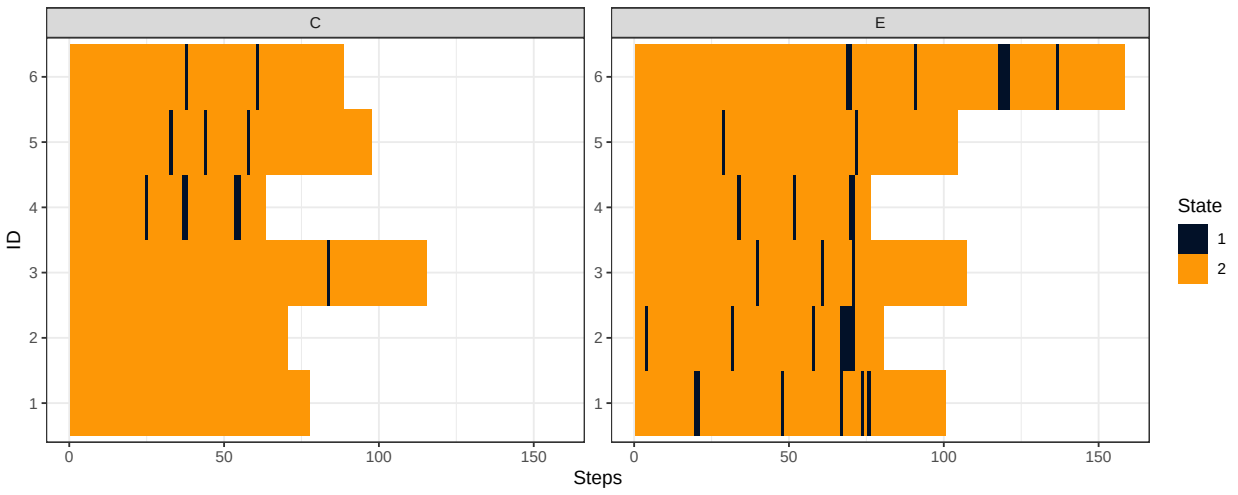

Figure 9: Sequence of states along the trajectory in two different farms (C: Conventional, E: Ecological)

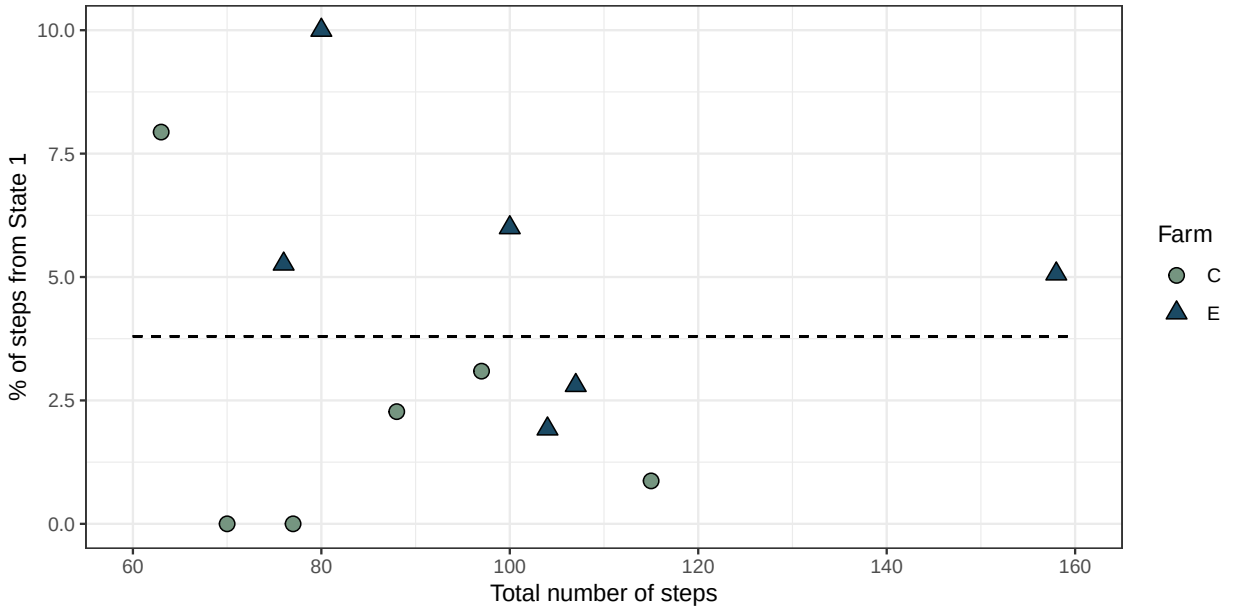

Figure 10: Relation between the probability of state 1, the total number of steps and the farm (C: Conventional, E: Ecological). The dashed line represent the average percentage of steps that belong to state 1 (3.8%).

4. References

Michelot, Théo, Roland Langrock, and Toby Patterson. 2019. "moveHMM: An r Package for the Analysis of Animal Movement Data." *ArXiv Preprint*, 1–24.

Michelot, Théo, Roland Langrock, and Toby A Patterson. 2016. "moveHMM: An r Package for the Statistical Modelling of Animal Movement Data Using Hidden Markov Models." *Methods in Ecology and Evolution* 7 (11): 1308–15.

Zucchini, MacDonald, W. 2016. *Hidden Markov Models for Time Series: An Introduction Using r*. Chapman; Hall/CRC Press.