Analysis of the trajectories of the harvesters using HMM

In this document I will present some results and interpretations of the analysis of the spatial trajectories of the harvesters using Hidden Markov Models (HMM).

0. Introduction to HMM

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 (that satisfies the Markov property; Figure 1). 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). 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 most likely sequence of states along the trajectory of an agent, its average time within a state, and the number of switches between states.

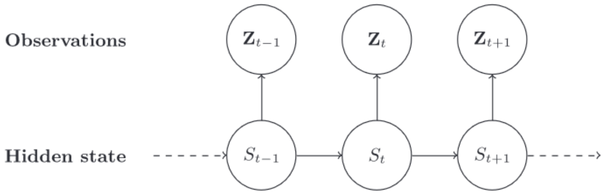

Figure 1: Representation of the HMM. taken from (Michelot, Langrock and Patterson, 2019)

Here we used the *movehmm* library (Michelot, Langrock, and Patterson (2019)) to i) fit the most likely HMM to the harvester movement data and ii) describe the different states, the sequence of states and the difference between farms.

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:
## I_Ger1 -- 100 observations
## I_Ger2 -- 80 observations
## I_Mig3 -- 107 observations
## I_MigSam4 -- 76 observations
## H_Fran5 -- 77 observations
## H_Fran6 -- 70 observations
## H_Fran7 -- 115 observations
## H_Fran8 -- 63 observations
## H_Fran9 -- 97 observations
## I_Sam10 -- 104 observations
## H_Fran11 -- 88 observations
```

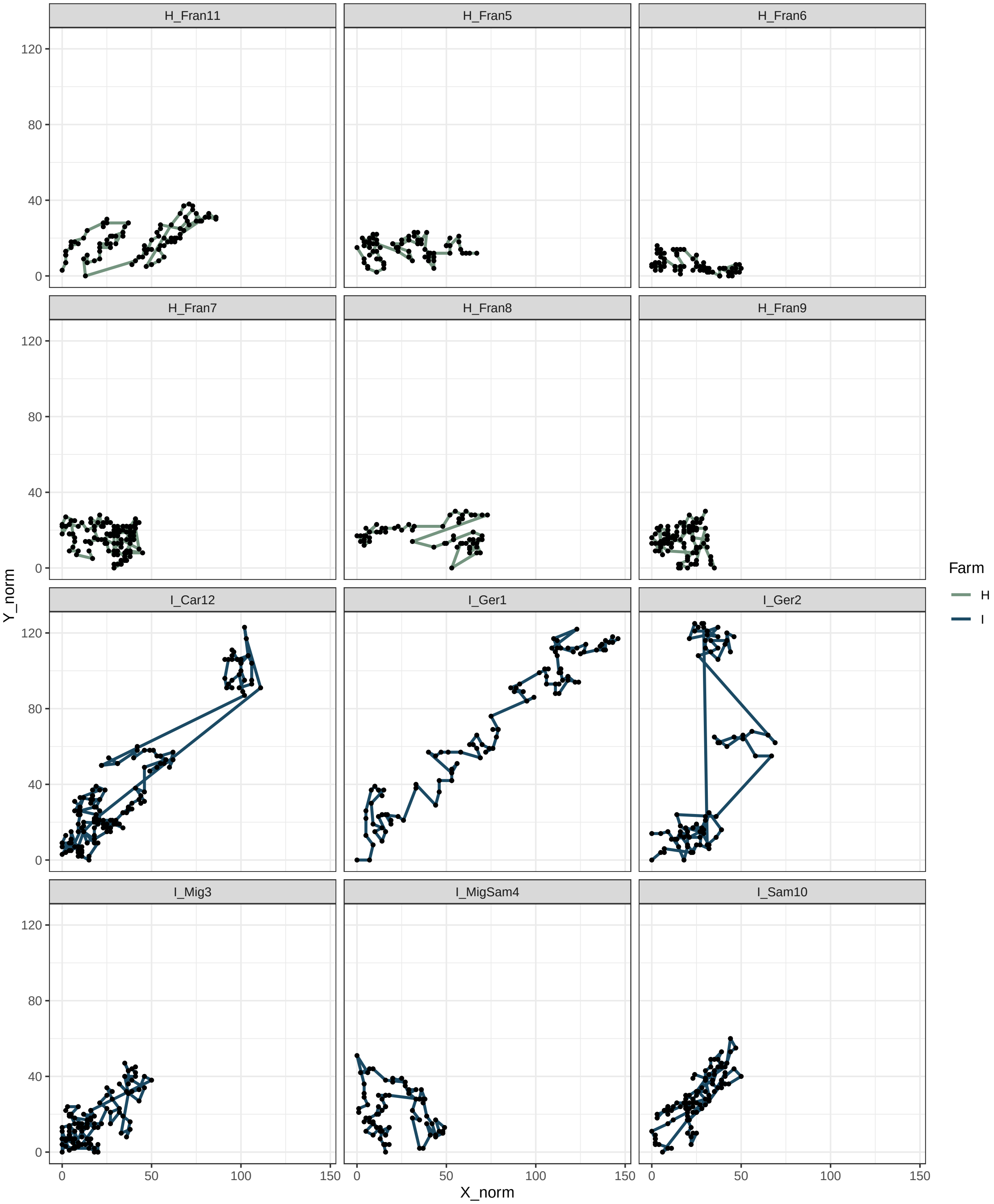

Figure 2: 12 trajectories of harvesters. The green line represent the Farm Hamburgo (H) and the blue line the Farm Irlanda (I).

```
## I_Car12 -- 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). We plotted the histograms of both the steps and the angles (Fig. 3 and Fig. 4).

Table 1: First lines of the movehmm object

ID	step	angle	x	y
I_Ger1	3.605551	NA	146	117
I_Ger1	3.000000	-2.158799	143	115
I_Ger1	3.605551	2.553590	143	118
I_Ger1	2.236068	-1.446441	141	115
I_Ger1	5.000000	2.034444	139	116
I_Ger1	3.605551	-2.158799	139	111

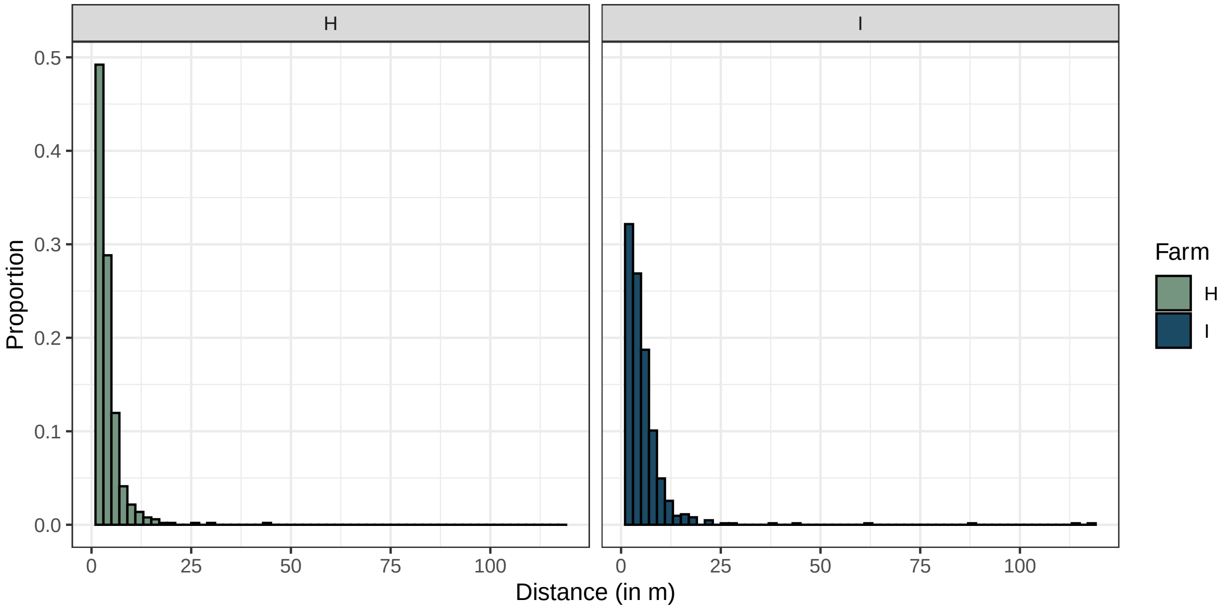

Figure 3: Histogram of the steps lengths per farm

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. 5.

1.2. Fitting of the data

The algorithm of the *movehmm* library uses:

a) a predefined number of states. Here we will use two.

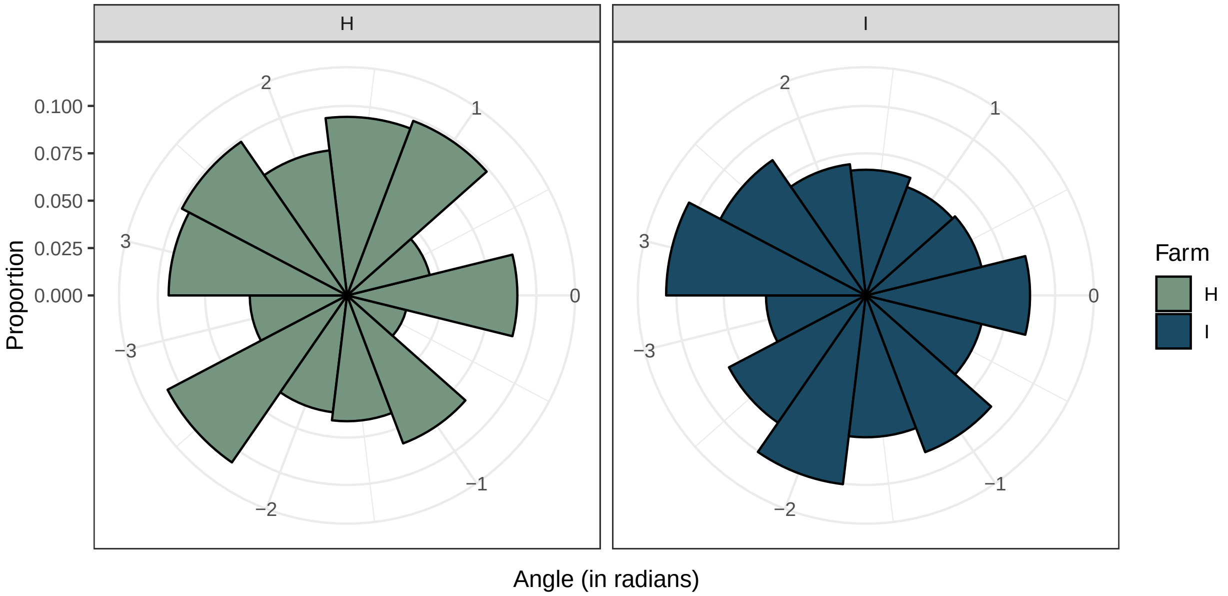

Figure 4: Radial histogram of the relative angles, per farm

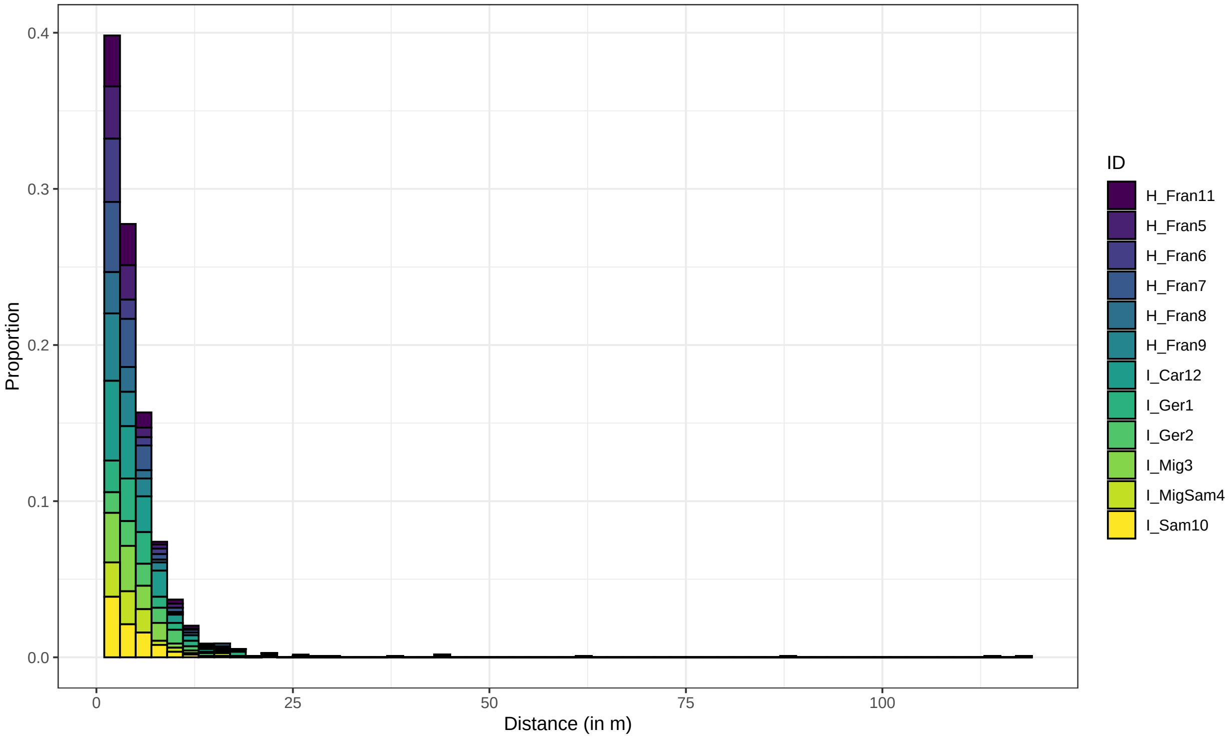

Figure 5: Histogram of the steps lengths

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) that fit the data and the most likely sequence of states (assuming a markov process). 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 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. For the *weibull* distribution, the maximum shape considered the left-skew shape and the scale was chosen to include the 120m long steps (see ranges in table 2). It is important to note that these are only priors guesses, and the real parameters can outbound these limits. 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

For each combination of parameters we registered the final parameters and the minimum negative log-likelihood. The minimum negative log-likelihood is equivalent to the maximum likelihood.

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

Some combinations resulted in distributions with zero variance. We decided to remove those cases as they did not make any biological sense. We then plotted the min negative log of the likelihood without these outliers (Fig. 6). We note that almost all the combinations of prior parameters converge to two minimum likelihoods. We chose, for each family of distributions the model that had the minimum AIC value (table 3).

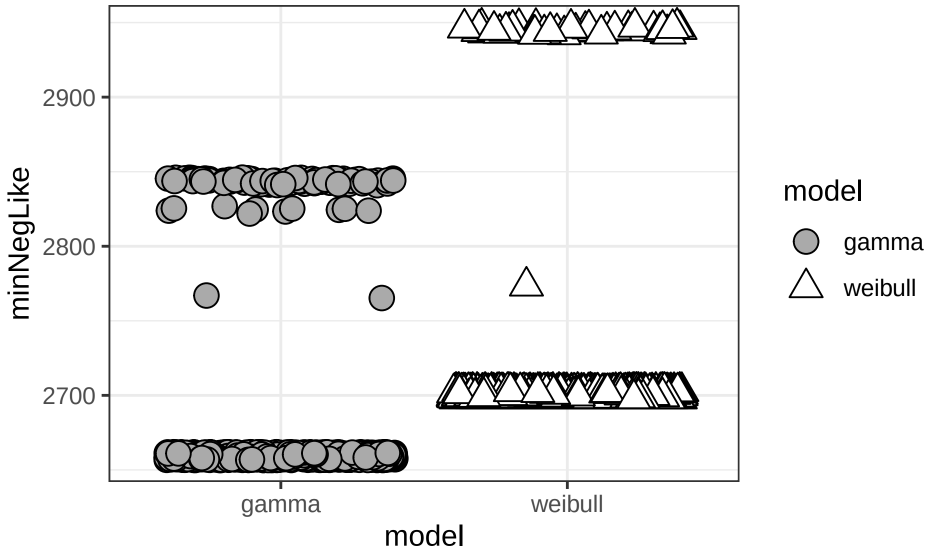

Figure 6: Minimun Negative Log-likelihood per combination of prior parameters for two different distributions

Table 3: Best models for each of the distributions. 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).

model	pr_st1_p1	pr_st1_p2	pr_st2_p1	pr_st2_p2	minNegL	AIC	s1_p1	s1_p2	s2_p2	s2_p2
weibull	2.20	0.80	5.10	13.20	2700.20	5414.39	2.05	4.71	1.05	14.45
gamma	1.94	5.61	6.46	1.15	2659.24	5332.48	17.34	15.18	4.28	2.25

We plotted the two state distributions for *gamma* and for *weibull* families (see figure. 7 and table 4 for the final parameters). For both families, one of the states produced a distribution that encompass all 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. 7 and table 4).

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

Parameter	Gamma_St1	Gamma_St2	Weibull_St1	Weibull_St2
μ (mean)	17.30	4.30	–	–
σ (sd)	15.20	2.25	–	–
α (shape)	1.30	3.62	2.05	1.05
θ (scale)	13.30	0.84	4.71	14.5
β (rate)	0.07	1.18	–	–

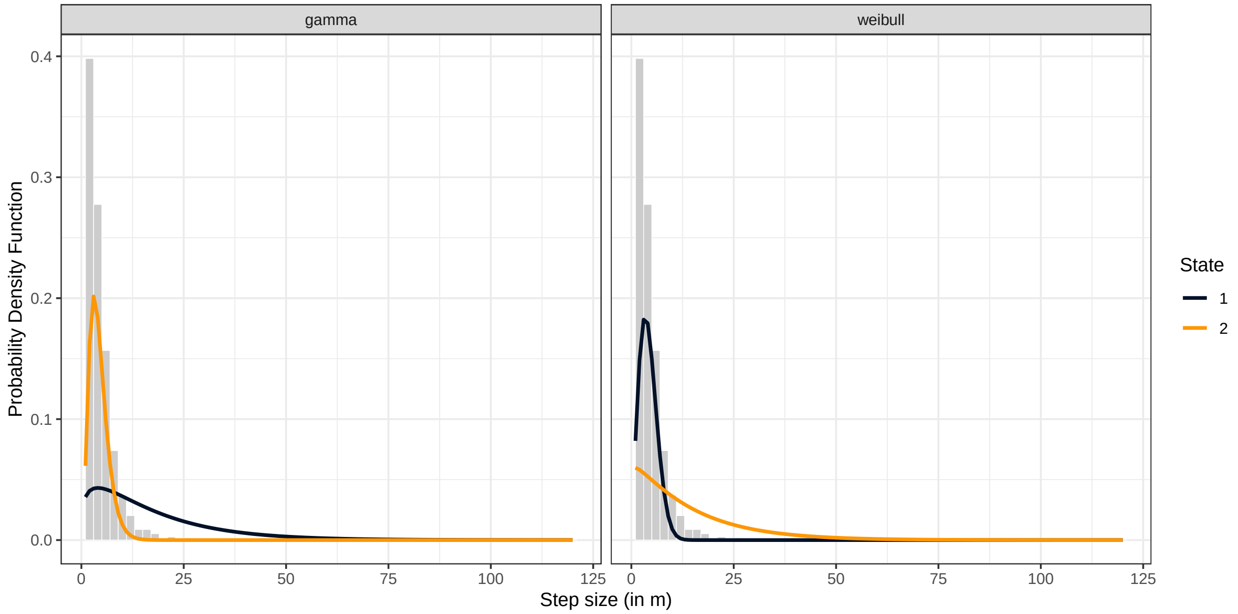

Figure 7: Two state model distributions. The used parameters are shown in Table 4

This results makes sense with our intuition, and confirms that harvesters have two behaviors that generate two significantly different trajectories: when they walk to the closest tree, or when they have to move to another part of the plot. The gamma distribution showed a lower AIC and is the best fitting for our data. We took two trajectories (Fran9 and Car12) to illustrate both states with the gamma distribution, where we can see the most likely states across the trajectories (Fig. 8).

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

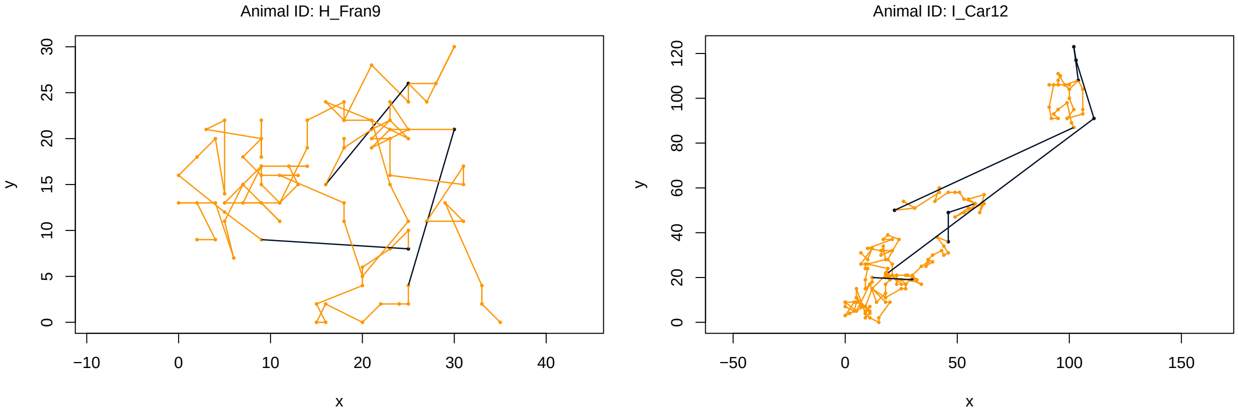

Figure 8: Distribution of states St along two trajectories. Each color follow the code of colors of Fig.7)

3. Probabilities of states per farm

We now estimate, once the given distribution of states per trajectories, the

```
## `summarise()` has grouped output by 'ID'. You can override using the `.groups`
## argument.
```

refe:

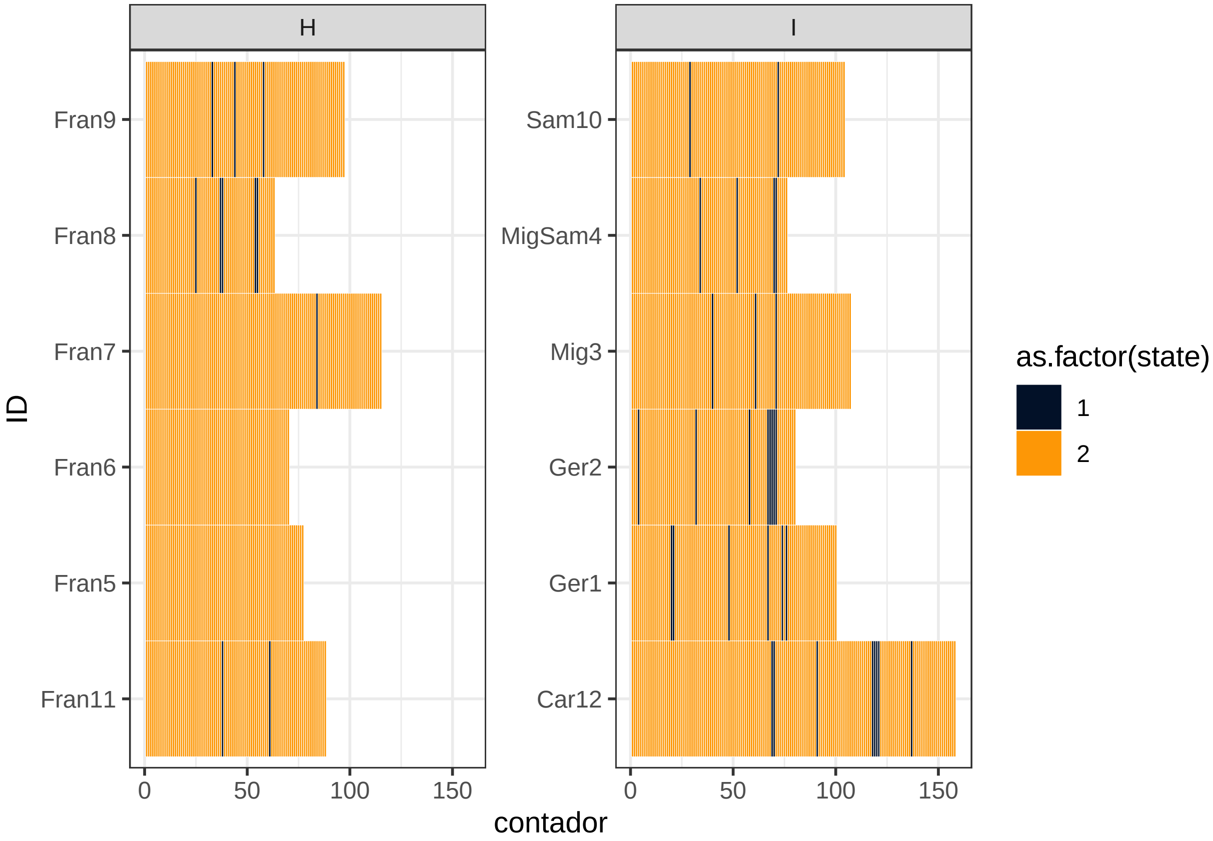

Figure 9: DDAII7)

which in a very effective way considers all possible state sequences that might have given rise to the observed time series, exploiting the dependence structure of the model. (In fact, the availability of this and other recursion techniques is probably the key reason for the popularity of HMMs.) This makes numerical maximization of the (log-)likelihood, and hence maximum likelihood estimation, feasible in most cases. For a fitted model, one can easily obtain the most likely sequence of states and also examine various variables of interest such as the average rate of movement in each state. For a comprehensive account of the HMM machinery, we recommend Zucchini, MacDonald & Langrock (2016).

https://cran.r-project.org/web/packages/moveHMM/vignettes/moveHMM-guide.pdf

https://theomichelot.github.io/software/

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

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