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) (ref*, fig.). 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. 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.

Here we will use the *movehmm* library (ref*) 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 have in total 12 trajectories, with (x, y) coordinates:

A first assumption for the analysis is made: we took these time irregular trajectories (where each point represents one different tree, but where the time between two trees is variable) and will treat them as a regular trajectories. We only want to analyze if the movement of the workers across the plots has different states (taking all the trajectories into account). This also implies that the underneath pattern of trees is relatively homogeneous and is not the main determinant of the trajectories. We convert this database into a movehmm object:

```
## 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
## I_Car12 -- 158 observations
## No covariates.
```

```
head(dataCosecha)
```

```
##       ID     step     angle   x   y
## 1 I_Ger1 3.605551        NA 146 117
```

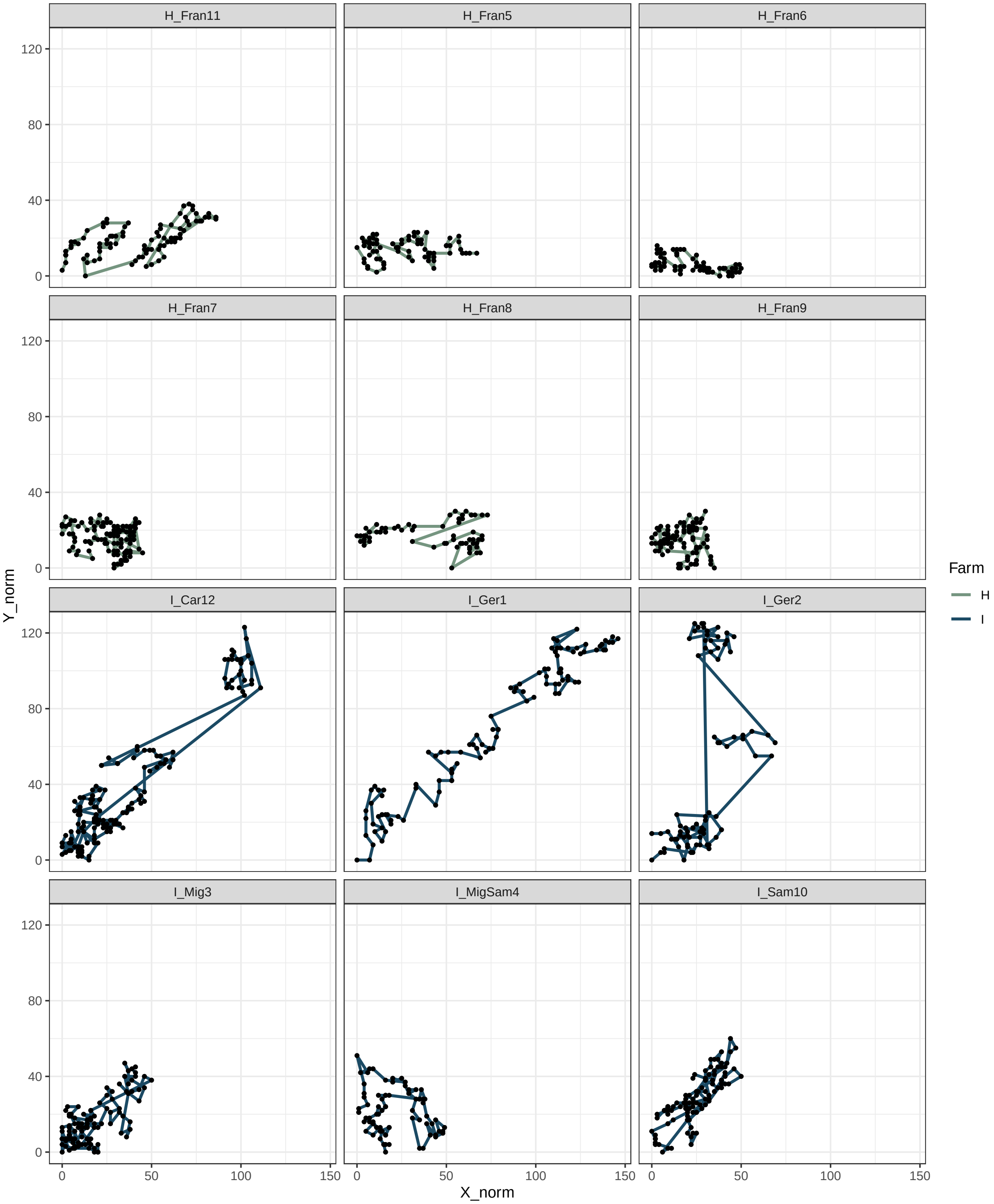

Figure 1: 12 trajectories of harvesters. The green line represent the Farm Hamburgo (H) and the blue line the Farm Irlanda (I).

```
## 2 I_Ger1 3.000000 -2.158799 143 115
## 3 I_Ger1 3.605551  2.553590 143 118
## 4 I_Ger1 2.236068 -1.446441 141 115
## 5 I_Ger1 5.000000  2.034444 139 116
## 6 I_Ger1 3.605551 -2.158799 139 111
```

The distance step and the relative angle is calculated for each of the 12 trajectories. We show the histograms of both the steps and the angles.

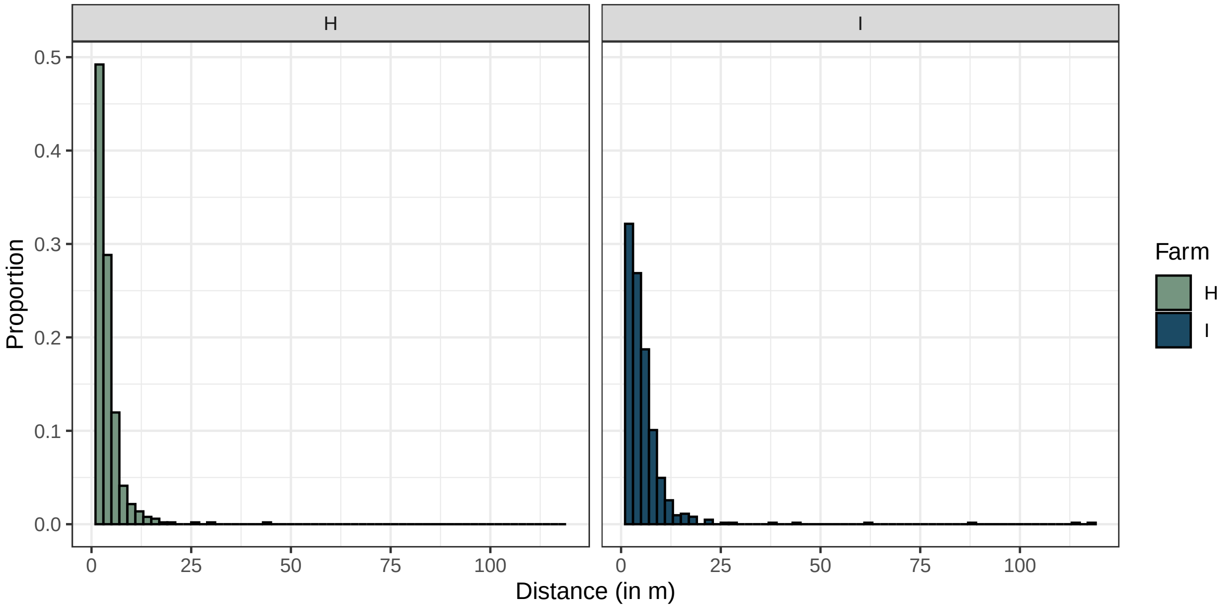

Figure 2: Histogram of the steps lengths per farm

Now, for the following analysis, we will only use the step distance and treat both farms as one. We decide to not include the angles to simplify the analysis and interpretation. In this sense, the input for the analysis is the following histogram:

1.2. Fitting of the data The fitting algorithm (ref*) of the *movehmm* library uses:

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

b) a defined family of distributions (*gamma*, *lnorm*, or *weibull*). Here we will try to fit each of the families.

c) prior parameters for each of the distribution, 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 its respective parameters (within the predefined family of distributions) that fit the data. During the fitting process, the algorithm uses the maximum-log likelihood and the *forward algorithm* (ref) using the prior parameters. In this sense, the correct choice of these prior parameters is crucial to avoid local maxima.

We ran a loop to fit the three different families (*weibull*, *lnorm* and *gamma*) and in order to avoid local maximal likelihoods, we swapped across a range of prior parameters for each model (1000) (ref*).The min and max values of each range were chosen to encompass the full distributions (*gamma* and *lnorm*). For the *weibull* distribution, the maximum shape considered the left-skew and the scale was chosen to include the 120 m steps (see all ranges in table 1. 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 will converge to the same final parameters.

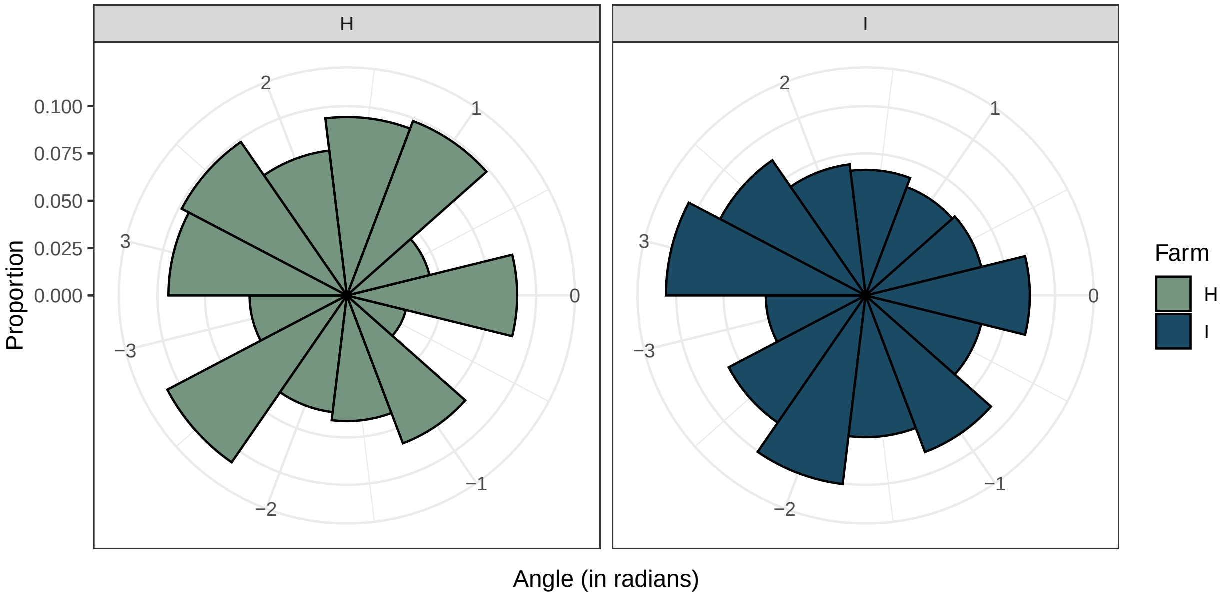

Figure 3: Radial histogram of the relative angles, per farm

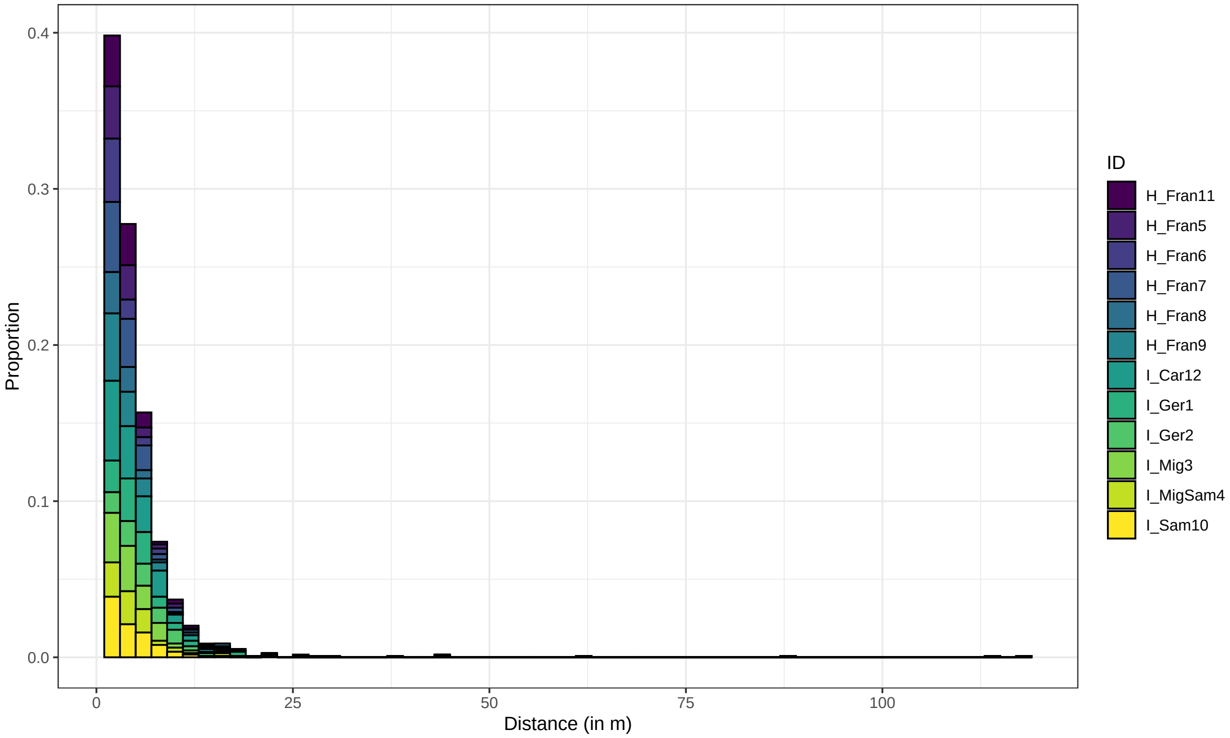

Figure 4: Histogram of the steps lengths

Table 1: Minimun and maximum values of the parameters of the distribution

value	gamma.mean	gamma.sd	weibull.shape	weibull.scale	lnorm.location	lnorm.scale
min	0.1	0.1	0.0	0.1	-1	0.001
max	20.0	20.0	2.7	15.0	20	3.000

For each combination of parameters we registered the final parameters and the minimum negative log-likelihood (table 2). The mininumm negative log-likelihood is equivalent to the maximum likelihood but works better in the fast forward algorithm (ref.).

Table 2: Examples of the results of the algorithm

model	pr__par0__st1	pr__par0__st2	pr__par1__st1	pr__par1__st2	minNegLik	AIC__model	st1__par0	st1__par1	st2__par0	st2__par1
weibull	2.20	0.80	5.10	13.20	2700.2	5414.39	2.05	4.71	1.05	14.45
weibull	0.55	0.43	5.03	3.43	2700.2	5414.39	2.05	4.71	1.05	14.45
weibull	1.85	0.45	5.02	9.64	2700.2	5414.39	2.05	4.71	1.05	14.45
weibull	2.13	1.15	3.73	6.19	2700.2	5414.39	2.05	4.71	1.05	14.45
weibull	0.37	0.16	5.09	8.39	2700.2	5414.39	2.05	4.71	1.05	14.45
weibull	2.04	2.34	13.02	2.42	2700.2	5414.39	1.05	14.45	2.05	4.71

Some results registered combinations where one of the final parameters created distributions with zero variance. We decided to remove those cases as they did not make any biological sense. We then plot the min negative log of the likelihood without these outliers (fig*)

We then chose, for each family of distributions the model that had the minimun AIC value (table 3).

Table 3: Best models (minimal AIC) for each of the distribution

model	pr__par0__st1	pr__par0__st2	pr__par1__st1	pr__par1__st2	minNegLik	AIC__model	st1__par0	st1__par1	st2__par0	st2__par1
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
lnorm	-0.54	-0.71	1.11	1.31	2625.26	5264.52	1.75	0.72	1.22	0.52

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

Now that we have the best two state model for each of the distributions, we will describe the states S_t they produced, their sequence and the observed differences between the farms.

2.1. Gamma model

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

2.2. lnorm model

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

2.3. Weibull model

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

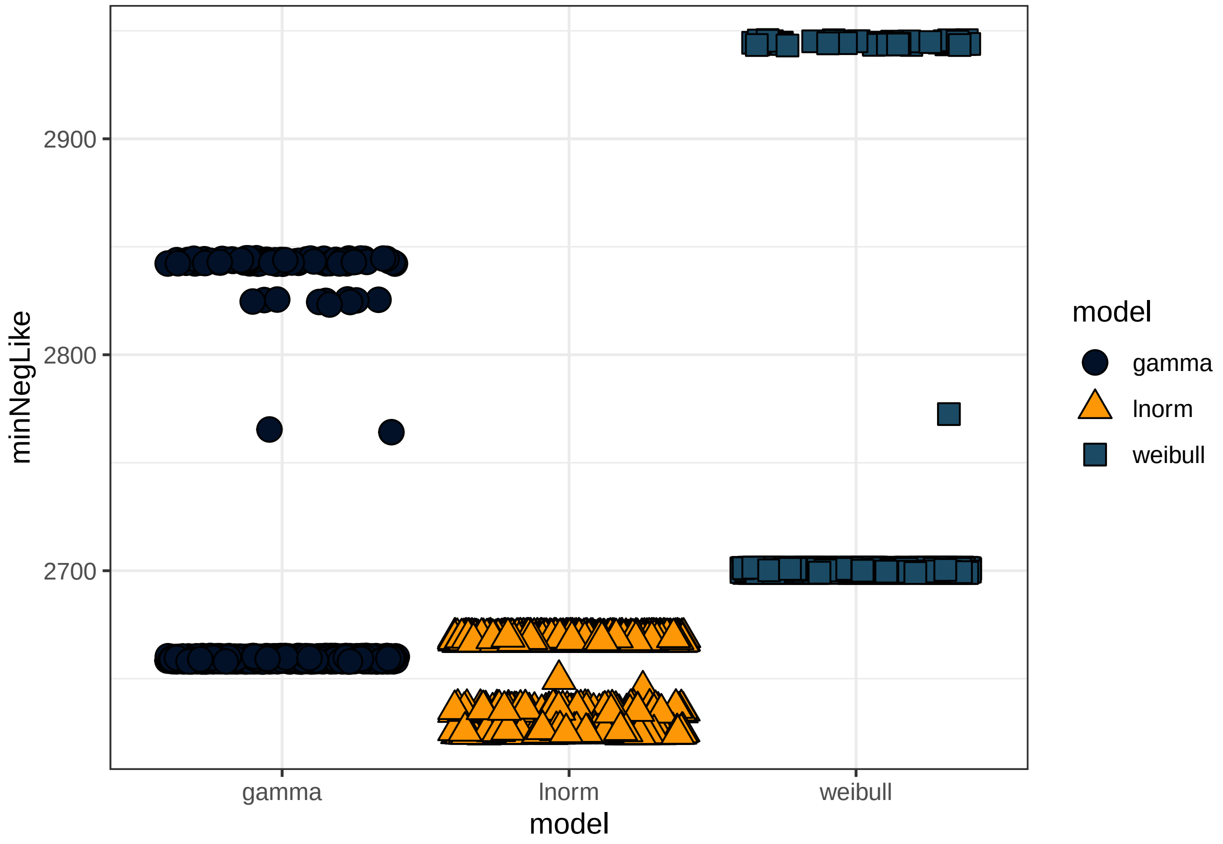

Figure 5: Minimun Negative Log-likelihood per combination of prior parameters for three different distributions

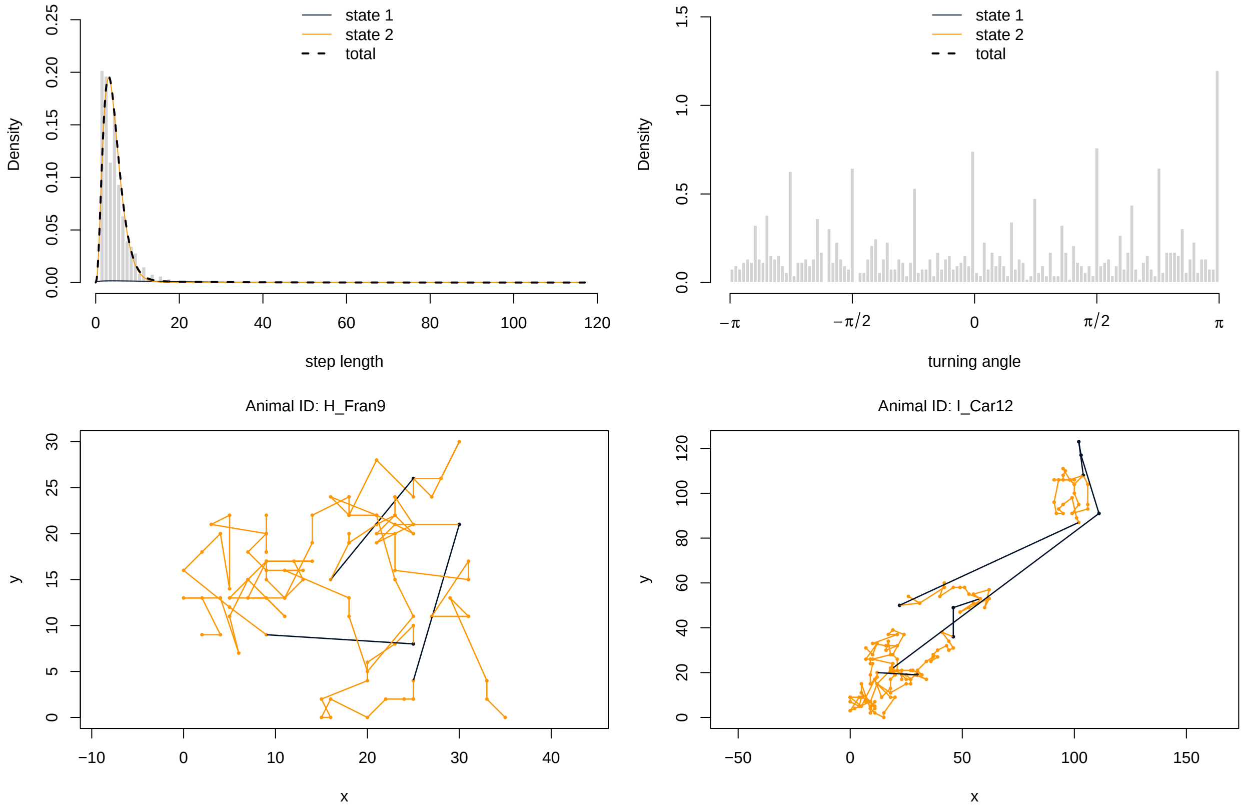

Figure 6: Gamma model

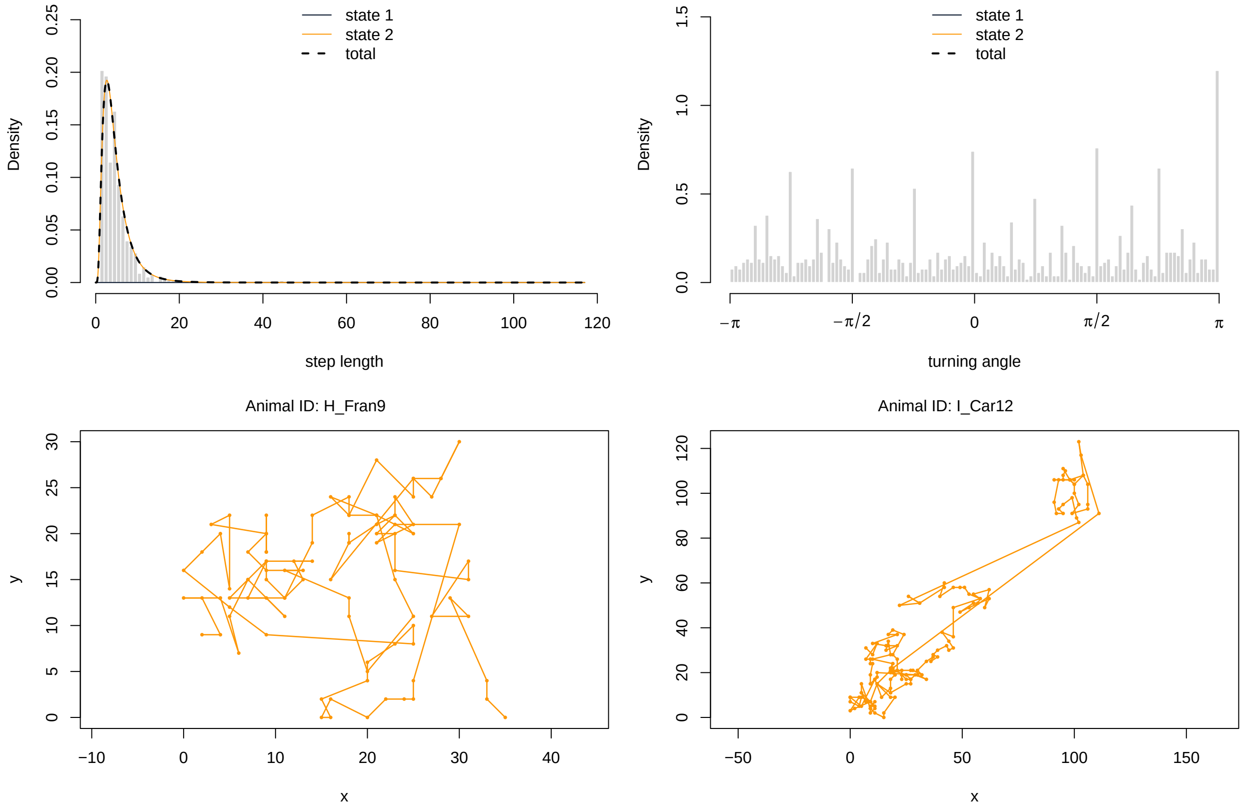

Figure 7: lnorm model

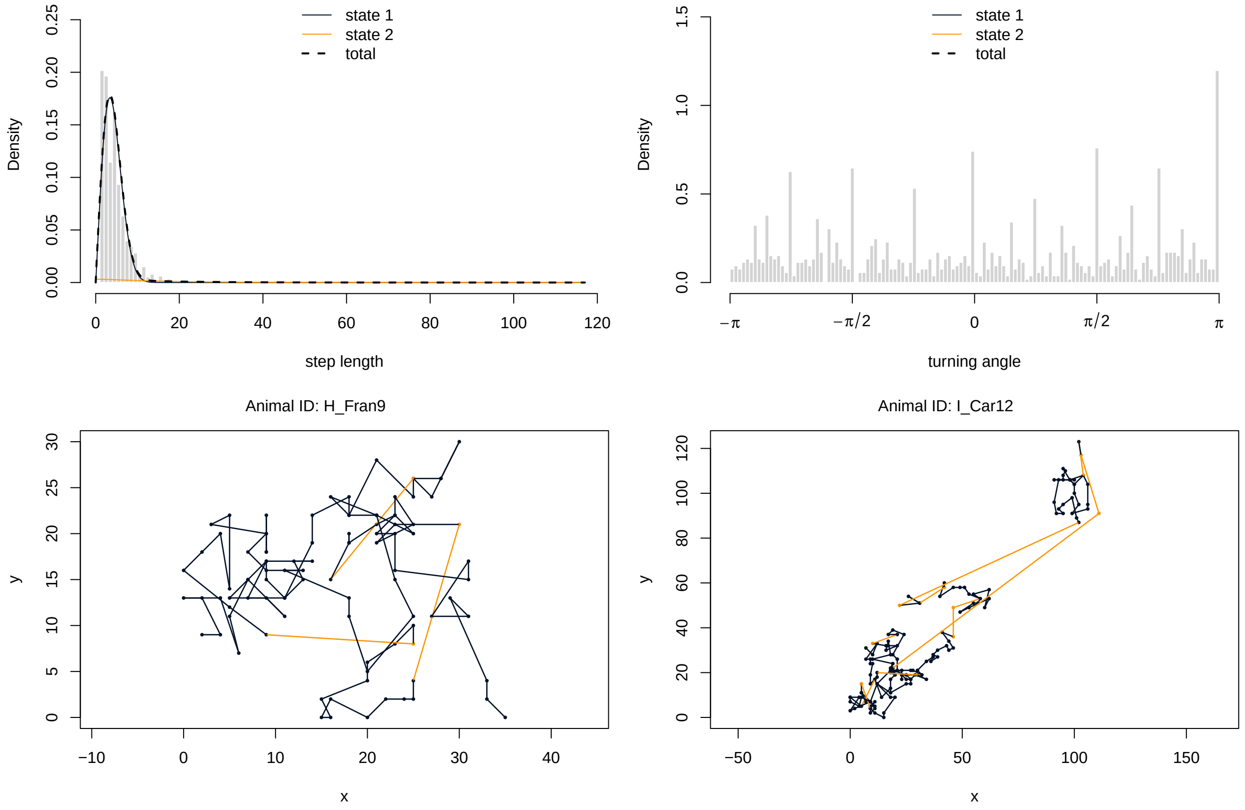

Figure 8: Weibull model