**Report on experimental infection statistical data analysis**

**Study design**

We conducted an experimental infection protocol on 75 goats divided in 11 batches of 6-7 individuals. Within each batch, one animal (seeder) was previously infected under controlled conditions with the strain of the Senegal 2020 PPRV lineage IV. Each goat was fitted with an Ultra-Wideband (UWB) sensor (BeSpoon – STMicroelectronics, France) set to record their precise (≈ 10-30 cm) 3D position with a frequency f = 1 Hz. Shortly before the beginning of each experiment, the six naïve animals were introduced in an open 6 x 4 metres experimental facility under natural photoperiodic conditions and with ad libitum access to feed and water. Each experimentation started from the introduction of the seeder and were contrasted in terms of the length of exposure After recovery, the animals were separated in individual pens in a 171 m² open area and regularly monitored for 5 up to 8 weeks to detect PPRV symptoms of infection. As validation tests, RT-PCR on ocular/nasal secretions or, in the case of death, on lung tissues were performed. Spatial proximities between individuals were computed from their individual 3D positioning records.

**Data analysis**

We have analyzed the distribution of the distance between the individuals and the seeder for each of the experiments, and the boxplots show that for most individuals, and in most experiments, the distance exceeds 2 m. Here's an example of the distribution for individuals in Experiment 2, with a contact duration of 44h.

A graph with different colored rectangles

Description automatically generated

Figure 1 : Boxplots representing the distribution of the "Distance/sec" characteristic for each individual (Captors) in the "experiment\_2\_44h" experiment. Boxes encompass the interquartile range (IQR), the thick horizontal line indicates the median, and whiskers extend to 1.5 times the IQR. Isolated points represent outliers.

**Models:**

***Model 1: Binomial model***

The main objective of this study was to quantify the number of positive cases (successful transmissions) following each experiment. To model this situation, we used a binomial model, where the probability of success (transmission) is assumed to vary with the duration of each experiment, according to the following functional relationship: .

***Model 2: Segmented model***

To refine our understanding of the transmission dynamics of PPRV viral lineage IV, we developed an approach that considers the distance between pairs of animals (susceptible animal + seeder) as well as the time of exposure to this distance. This method enabled us to model the probability of transmission as a function of spatial and temporal variables specific to each animal. For each animal, we calculated the cumulative exposure time as a function of three predefined distances from the infected animal: 1 m, 1.5 m, and 2 m. These distances were chosen to reflect a variety of exposure scenarios within the experimental environment. The individual probability of transmission was modelled as follows: , where represents the cumulative exposure time for the animal under consideration. For each distance, a separate model was built and run, using the Bernoulli distribution to model transmission (infected/uninfected) as an observed binary variable.

***Model 3: Envelopes model***

The third approach uses recordings of distances per minute, enabling a more granular and dynamic analysis of animal interactions. Unlike the second approach, which is based on fixed distances (1 m, 1.5 m and 2 m) and cumulative exposure time, this method captures the minute variations in distances between animals over time. Instead of considering predefined distances, the third approach identifies groups where the distance between animals stabilizes. For each group, the median distance and total stabilization time are calculated. The transmission probability formula in the third approach: introduces an exponential component that takes both distance and time into account in a more sophisticated way.

**Parameter estimation:**

All models’ parameters were estimated via Bayesian sampling using PyMC, a Python library for Bayesian statistical modeling based on Theano. The No-U-Turn Sampler (NUTS) was used to draw samples from the posterior parameter distributions using the Markov Chain Monte Carlo (MCMC) sampling algorithm. Parallel computation on CPU resources was accelerated by Blackjax, a numerically optimized computation library. The MCMC simulations were run on an Intel Core i5 10th generation processor with the following configuration: 10 parallel chains, 1000 tuning iterations with a target accept rate of 0.95, followed by 10,000 sample draws per chain for final parameter estimation. To study chain quality and convergence, we used Gelman-Rubin diagnostics and Effective Sample Size (ESS).

|  |  |  |
| --- | --- | --- |
| Models | Parameters | Prior distributions |
| Model 1 |  |  |
| Model 2 |  |  |
| Model 3 |  |  |
|  |  |

**Results:**

**Parameter estimation results:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Distance cut off | Parameters | Enveloppe Threshold | Mean | SD | 95% HDI  Lower – upper |
| Model 1 | - |  | - | 0.00023 | 0.00006 | 0.00013 – 0.00034 |
| Model 2 | 1 m |  | - | 0.0049 | 0.0012 | 0.0027 – 0.0073 |
| 1.5 m |  | - | 0.0031 | 0.0008 | 0.0017 – 0.0046 |
| 2 m |  | - | 0.002 | 0.0005 | 0.0011 – 0.003 |
| Model 3 | 1 m |  | 0.1 | 0.0078 | 0.0020 | 0.0041 – 0.0116 |
| 0.2 |  |  |  |
| 0.3 |  |  |  |
|  | 0.1 | 0.3708 | 0.0374 | 0.2994 – 0.4448 |
| 0.2 |  |  |  |
| 0.3 |  |  |  |
| 1.5 m |  | 0.1 | 0.0044 | 0.0011 | 0.0024 – 0.0066 |
| 0.2 | 0.0045 | 0.0011 | 0.0024 – 0.0066 |
| 0.3 | 0.0054 | 0.0013 | 0.0029 **–** 0.0080 |
|  | 0.1 | 0.3710 | 0.0374 | 0.3002 – 0.4455 |
| 0.2 | 0.3711 | 0.0374 | 0.3002 – 0.4452 |
| 0.3 | 0.3710 | 0.0374 | 0.2992 **–** 0.4445 |
| 2 m |  | 0.1 | 0.0034 | 0.0008 | 0.0019 – 0.0051 |
| 0.2 | 0.0035 | 0.0009 | 0.0019 – 0.0052 |
| 0.3 | 0.0036 | 0.0009 | 0.0019 – 0.0053 |
|  | 0.1 | 0.3712 | 0.0375 | 0.3002 – 0.4463 |
| 0.2 | 0.3712 | 0.0374 | 0.2997 – 0.4456 |
| 0.3 | 0.3712 | 0.0374 | 0.2987 – 0.4445 |

A graph of infection

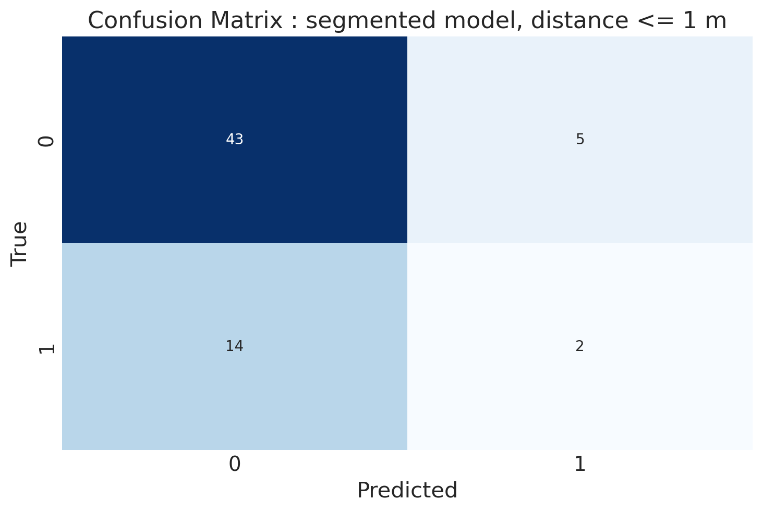
Description automatically generated

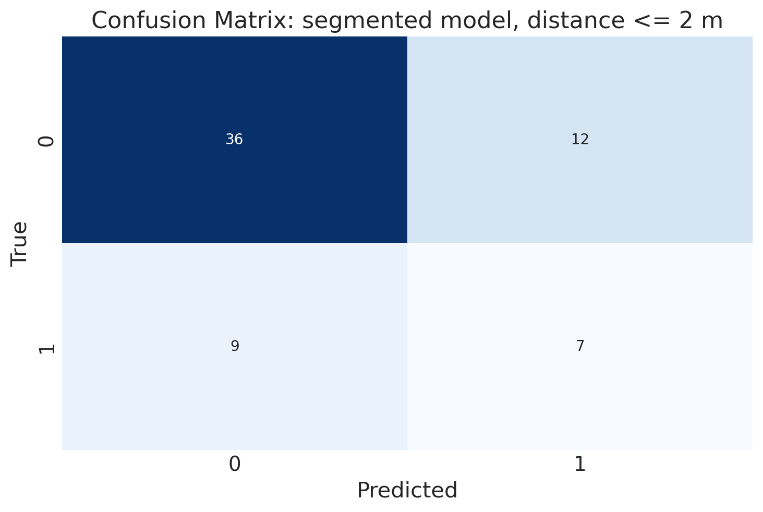
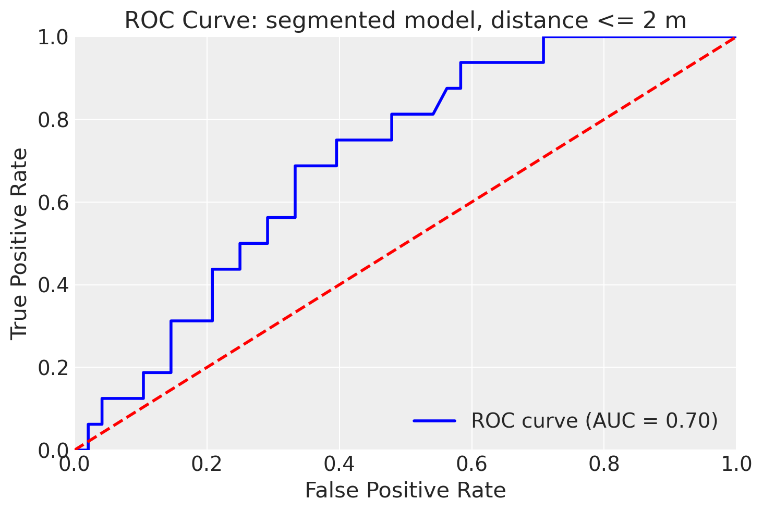
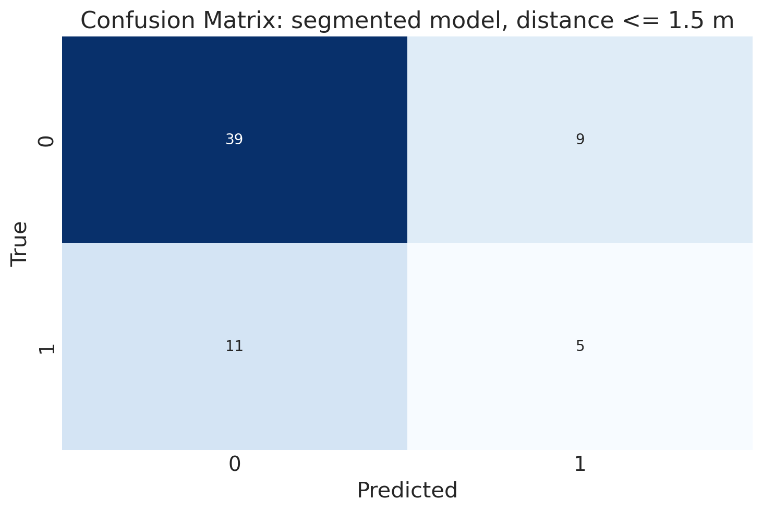
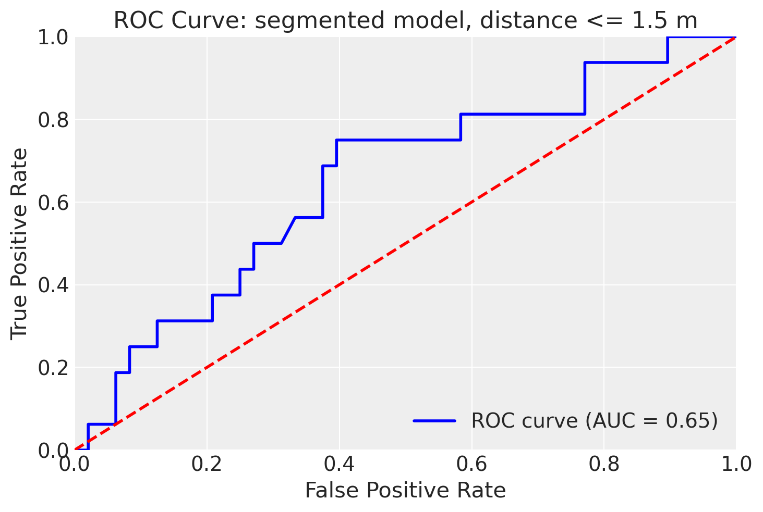
Figure 2 : This figure represents the estimated number of new positives for each experiment, blue dots represent the real numbers observed and the red dots are the mean of estimated ones. The dark red area represents the CI [25%, 75%] and the lighter area represents the CI [2.5%, 97.5%]

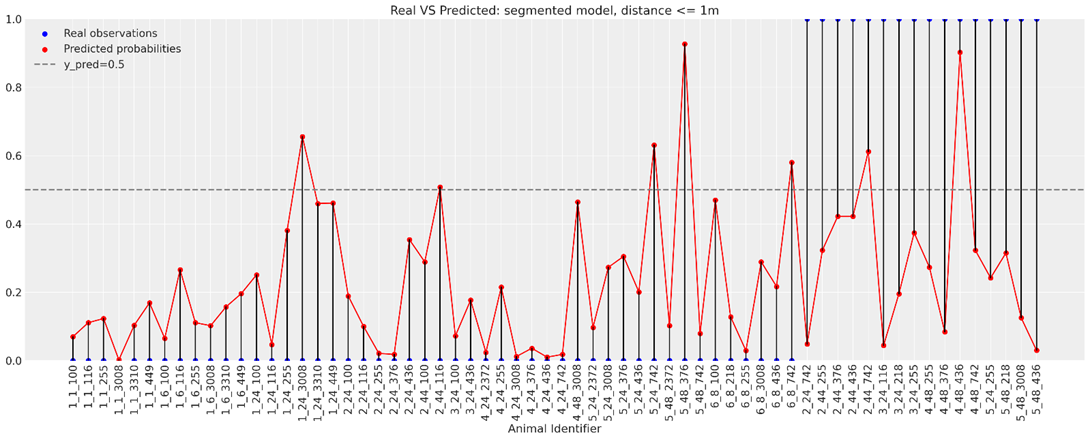
The figure shows that the number of positive cases estimated for each experiment follows the same pattern as actual observations.

**Segmented:**

A graph of a curve

Description automatically generated 





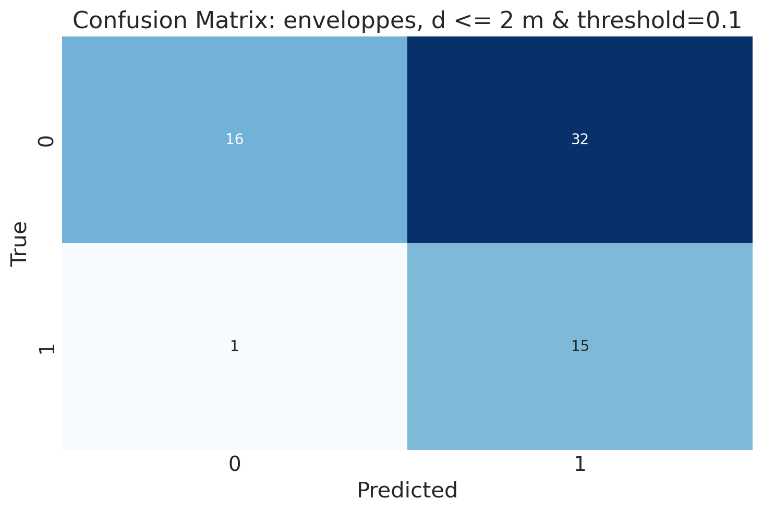
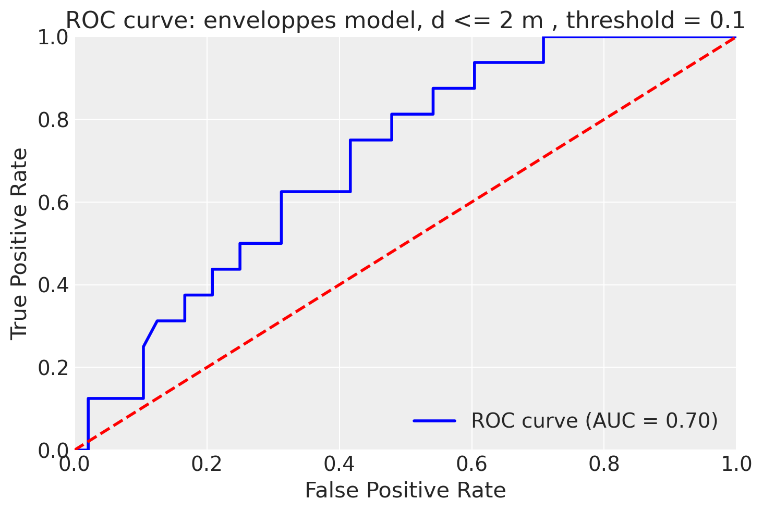
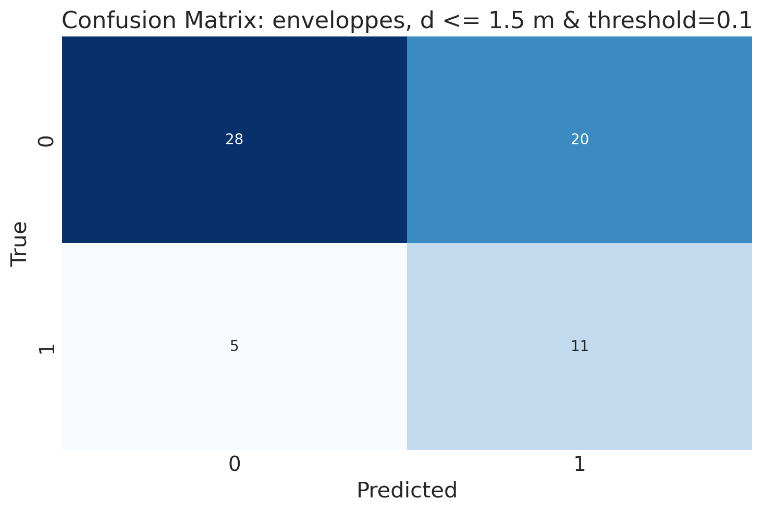
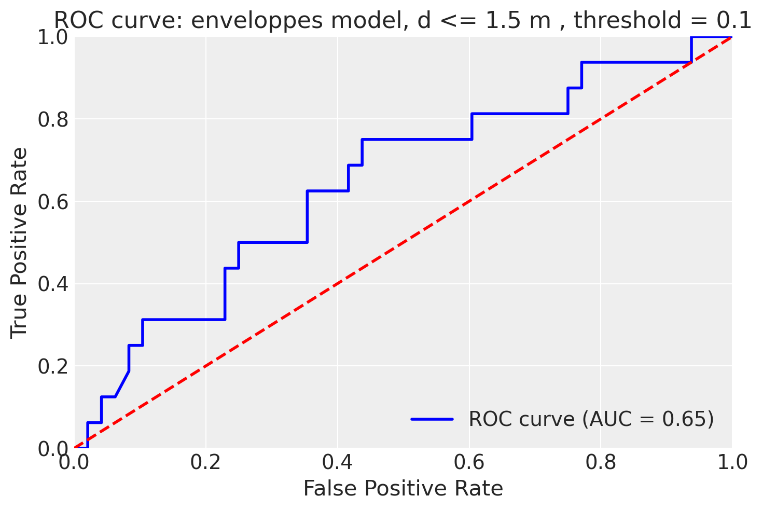
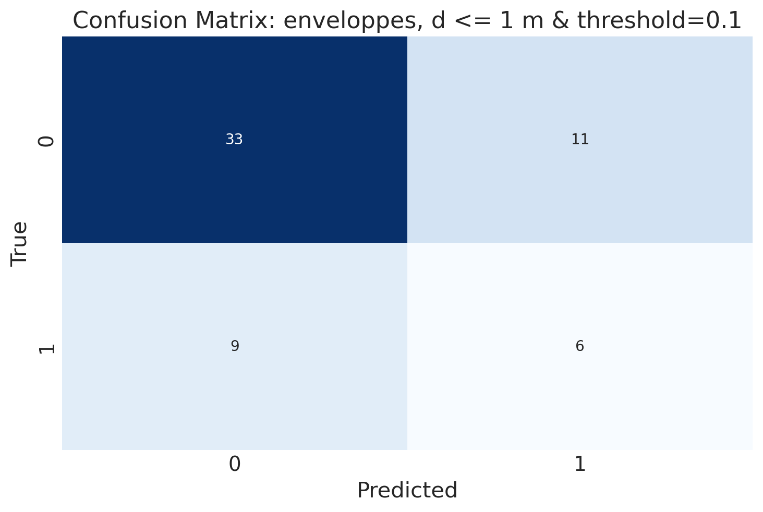
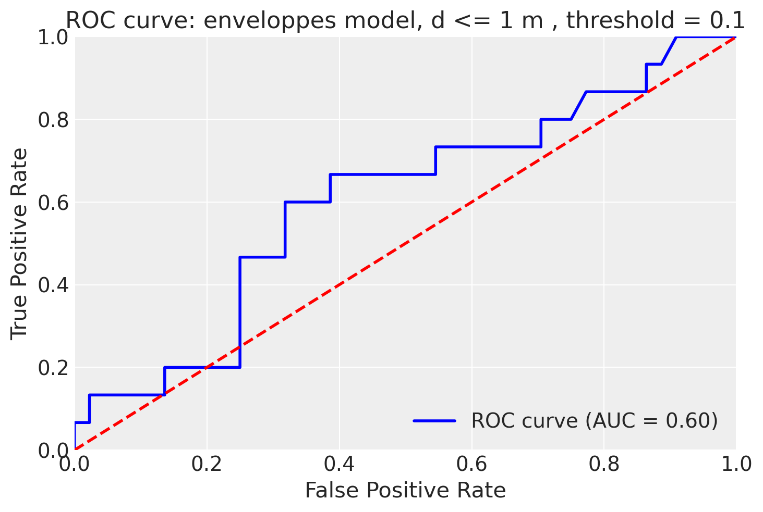
A graph with red and blue lines

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

**Envelopes:**



A graph with red lines

Description automatically generated

A graph with red lines

Description automatically generated

A graph with red and blue lines

Description automatically generated

The Receiver Operating Characteric (ROC) curves show similar results between both segmented and envelopes models in terms of the values of Area Under the ROC curve (AUC), but the difference rises in the prediction of negative and positive classes as represented by confusion matrices and the last figures. We can clearly see that envelope models predict the positive class (infected) very well as the distance reaches 2 m, in the other hand segmented models tend to classify the majority of individuals as negative class (non-infected). As for our purpose, we are interested to predict the positive class, and the envelope model outperforms the segmented model.

# Exposition time estimation using envelopes model:

We used the optimal estimated values of parameters of the best model: enveloppes model. We fixed the distance and then we used this formula to simulate over time to estimate the minimum exposure time window required to have a successful contamination.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Distance |  |  | Exposure time | P | Model |
| d = 1 m | 0.0078 | 0.3708 | 555 (401 – 984) | 0.95 | enveloppes |
| d = 1 m | 0.0078 | 0.3708 | 128 (93 – 228) | 0.50 | enveloppes |
| d = 1.5 m | 0.0044 | 0.3710 | 1186 | 0.95 | enveloppes |
| d = 1.5 m | 0.0044 | 0.3710 | 274 | 0.50 | enveloppes |
| d = 2 m | 0.0034 | 0.3712 | 1850 (1433 – 2873) | 0.95 | enveloppes |
| d = 2 m | 0.0034 | 0.3712 | 428 (331 – 665) | 0.50 | enveloppes |

A graph of different colored lines

Description automatically generated

The curves presented show the evolution of probability versus time as a function of distance (d ≤ 1m, d ≤ 1.5m and d ≤ 2m). Each curve represents the mean probability and the associated 95% high-density interval. The smaller the distance (d ≤ 1m), the faster the curve reaches high probabilities as a function of exposure time.