Statistical models for incidence of Coccidiosis parasites in mink

Spooner, Max Peter; Stockmarr, Anders; Petersen, Heidi Huus; Gram-Nielsen, Sanne

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Statistical models for incidence of Coccidiosis parasites in mink

Max Spooner[a], Anders Stockmarr[a], Heidi Huus Petersen[b], Sanne Gram-Nielsen[b]

[a] Section for Statistics and Data Analysis, DTU Compute
[b] Division for Diagnostics & Scientific Advice, National Veterinary Institute, DTU

July 2019

DTU Compute
Department of Applied Mathematics and Computer Science

Contents

Appendices

1 Introduction

In this report we present statistical models that describe the incidence of Coccidiosis parasites in mink in thirty mink farms in Denmark during the period April to October 2016. The dataset was obtained by asking the farmer at each mink farm to select 5 sows and collect a fecal sample from the sows on 16 dates during the period April to October. Once the sows give birth to kits (usually at the end of april), a pooled fecal sample from the kits belonging to each sow was also collected on the same dates as the sows. All the fecal samples were then analysed for presence of two types of Coccidiosis-causing parasites: Eimeria and Isospora. If Eimeria was present, the sample was examined for the presence of three different types of Eimeria, called tyk, tynd and lille. In addition, the fecal sample was classified by the farmer according to a feces score (hereafter F score) ranging from 1 to 5.

Various characteristics of the mink and the farms were also included in the dataset, including from which of three feed suppliers the farm obtains the feed, birth date of the kits, litter size, age of the sow. The aim of this project was to use statistical models to investigate which factors can account for the incidence of Eimeria (and its three subtypes), and Isospora, as well as the F score, observed during the study. This means there were 6 response (independent) variables to investigate: Eimeria, tynd, tyk, lille, Isospora and F score.

The main aims were to determine how the six response variables depend on:

1. Time
2. Animal status (sow or kit)
3. Feed supplier
4. Outdoor temperature

An additional aim was to determine how F score depends on Eimeria and/or Isospora outcome.

Two different modelling approaches were used to investigate the impact of animal status. In the first approach the aim was to investigate whether there is a difference of parasite incidence between sows and kits. Therefore, the models were fit to the entire dataset, and animal status was treated as an independent varaible.

In the second approach, the aim was to investigate whether the sows might infect their kits with the parasites. In this approach, the dataset was restructured so that only the outcome for kits was modelled, but now the parasite status of each kit's mother at the previous measurement round entered the model as an independent variable.

2 Description of the Data

In this section, the dataset is described and exploratory plots are presented. First, brief definitions of the main variables of interest are given:

Response Variables:

- E_pos: Is the sample positive for Eimeria (0 or 1)
- I_pos: Is the sample positive for Isospora (0 or 1)
- tynd: Was the tynd Eimeria type observed in the sample (0 or 1)
- tyk: Was the tyk Eimeria type observed in the sample (0 or 1)
- lille: Was the lille Eimeria type observed in the sample (0 or 1)
- F_score: Feces score of the sample (0,1,2,3,4 or 5)

As well as the binary outcome of Eimeria and Isospora, cell counts (E_opg and I_opg) were also provided.

Explanatory Variables:

- Sample time
  - sample_round: ranging from 1 to 16
  - sample_date: Calender date of the sample
  - sample_week: Calender week of the sample (ranging from 15 to 40)
- Farm_id: Identifier for the 30 farms in the study (A to AB)
- Foder_central: Feed supplier (A, B or C)
- mink_id: Identifier for each mink
- T_H: Animal status (T for tæve/sow or H for hvalpe/kit)
- H_no_born: Number of kits in the litter
- H_born: date of birth of the litter
- T_age: Age of the sow
- temperature: Median outdoor temperature in Denmark during the preceding week (calculated using data from DMI).

The original dataset contained 4799 rows. However, the following observations were removed:

- 658 osbservations for which all of the response variables had missing values
- A further 60 observations for which H_no_born, H_born, T_age and several other variables were all missing values. Probably, most of these cases were for sows that never had kits and so were replaced with another sow during the study.

This resulted in a dataset of 4081 observations. Table 2.1 shows the number of observations for each sample_round, both for sows and for kits. For the early sample rounds there are no or few observations for kits because they are not yet born or otherwise so young that obtaining a fecal sample is not practical. Note that sample_rounds 1 to 5 are spaced 2 weeks apart, rounds 5 to 12 are spaced 1 week apart, rounds 12 to 15 are spaced 2 weeks apart and round 16 is 4 weeks after round round 15. Hereafter, we mostly use sample_week (ranging from 15 to 40) to identify the sample round and as the unit of measurement of time. A visualisation of how the samples are spaced in time for each sow-kit pair, in relation to the birth date of the kit litter, is shown in Fig. 2.1.

	sample_date	sample_round	sample_week	N_T	N_H	N_total
1	2016-04-12	1	15	122	0	122
2	2016-04-26	2	17	128	0	128
3	2016-05-10	3	19	141	4	145
4	2016-05-24	4	21	145	82	227
5	2016-06-07	5	23	149	141	290
6	2016-06-14	6	24	144	137	281
7	2016-06-21	7	25	143	138	281
8	2016-06-28	8	26	148	141	289
9	2016-07-05	9	27	148	141	289
10	2016-07-12	10	28	147	146	293
11	2016-07-19	11	29	148	145	293
12	2016-07-26	12	30	148	147	295
13	2016-08-09	13	32	142	142	284
14	2016-08-23	14	34	143	142	285
15	2016-09-06	15	36	148	147	295
16	2016-10-04	16	40	143	141	284

Table 2.1: Number of observations per sample_round for sows (N_T), kits (N_H) and in total (N_total)

Animal characteristics There were 300 unique mink_ids in the dataset: 149 kits (litters) and 151 sows. In summarising the animal characteristics we treat each animal as 1 observation (even though the dataset contains up to 16 sample_rounds per animal). Fig. 2.2 shows that the kits were born between 22 April and 13 May except for one kit litter born much earlier than the rest on 2 March. This very early birth date may be due to a data entry error. Litter size ranged from 2 to 13 kits with a mode at 7 kits per litter (Fig. 2.3). Table 2.2 shows the ages of the sows, with most sows being 1 year old. The sows had quite a variety of coat colors (Fig. 2.4)

Defining outdoor temperature The outdoor temperature was included in the dataset by first obtaining records of low, median and high temperature in Denmark for each day

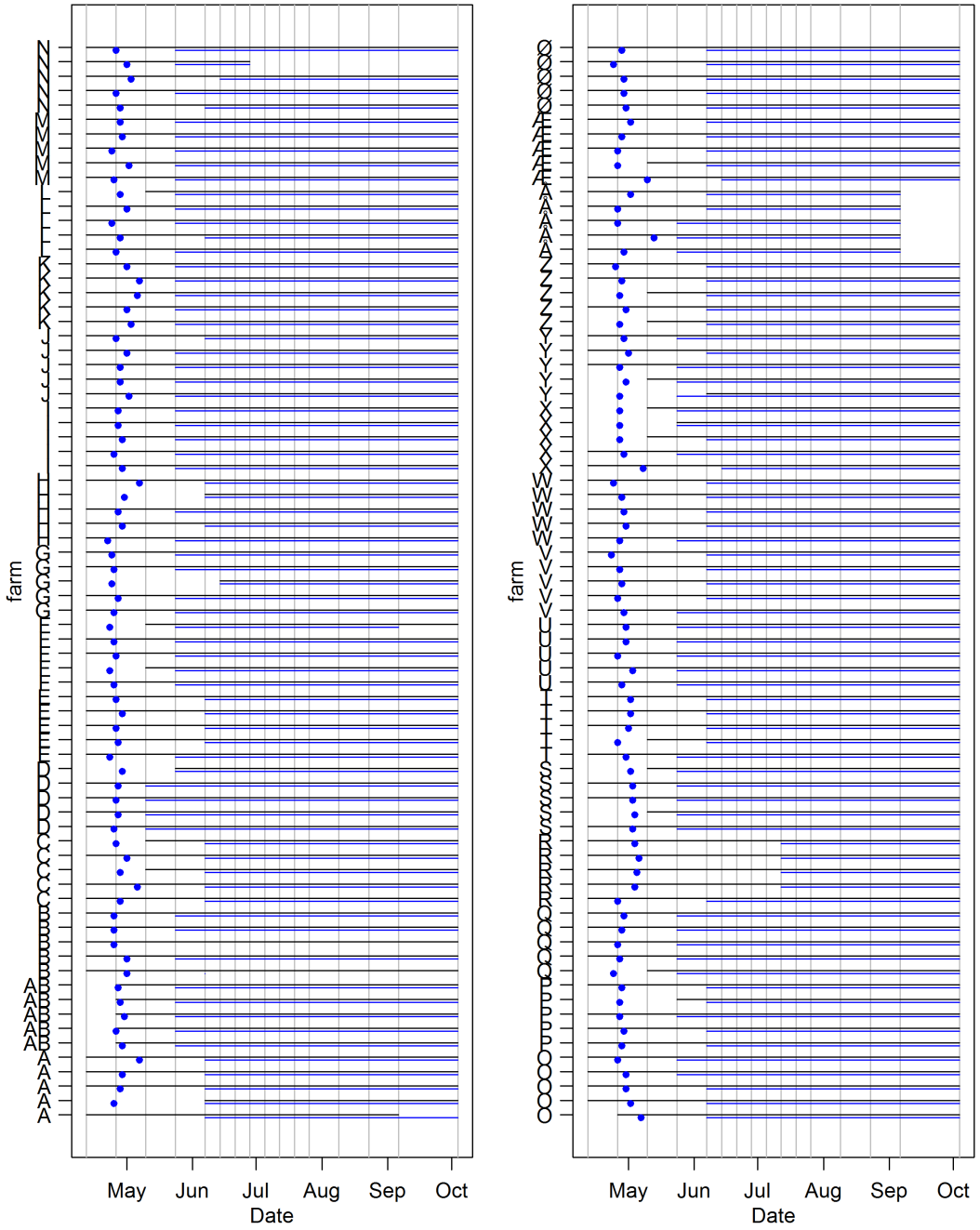

Figure 2.1: For each sow-kit pair, horizontal lines show the periods covered by the first and last sample in the data for the sow (black line) and the kit litter (blue line) as well as the birth date of the kit litter (blue dot). The kit litter for the bottom pair in farm A was born outside the plotting region (on 2 March). The vertical gray lines show the 16 sample_dates

during the period from DMI (https://www.dmi.dk/vejrarkiv/). This data is shown in Fig. 2.5. Then, for each sample_date, the median of the seven daily median temperatures during the preceding week was calculated, resulting in the values shown in Fig. 2.6.

Response Variables The 6 response variables were binary (for E_pos, I_pos, tynd, tyk, lille) and ordinal (for F_score). Table 2.3 shows the number of observations for each

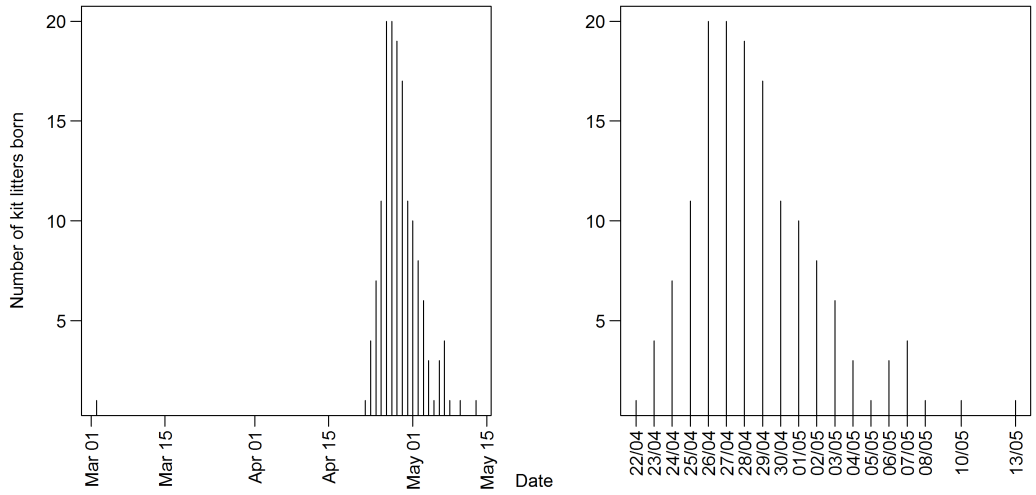

Figure 2.2: Distribution of birthdates of all the kits (left) and all the kits except the one born on 02-03-2016 (right)

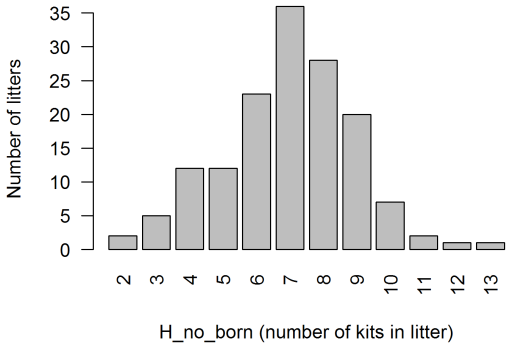

Figure 2.3: Distribution of number of kits in each litter

T_age	Number of sows
0.5	10
1	92
2	38
3	6
4	2
5	3

Table 2.2: Distribution of age of sows (in years)

outcome of the binary response variables. It was possible for more than 1 Eimeria subtype to be observed in the same sample. Table 2.4 shows the number of observations for each possible combination of the three Eimeria subtypes. Not that there was 1 observation where E_pos$= 1$ but no Eimeria subtype was observed (reflected in the first cell of Tables 2.3 and 2.4 - 3558 vs. 3559). Counts for the F_score responses are shown in Table 2.5. There were 91 observations for which F_score was missing. The most common F_score

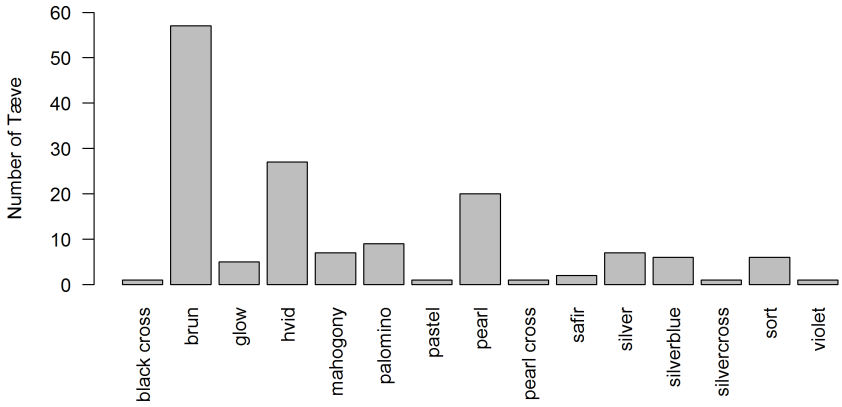

Figure 2.4: Distribution of sow coat color

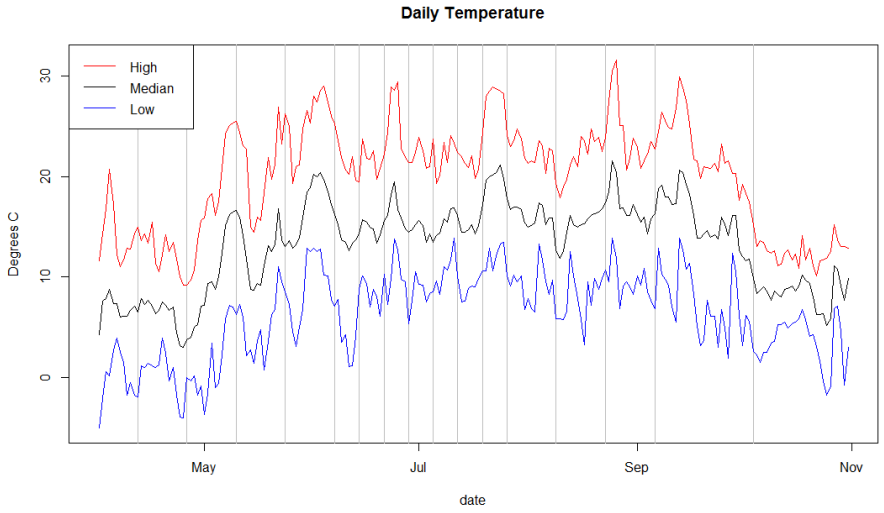

Figure 2.5: The daily temperature data from DMI. The vertical grey lines show the sample_dates

was 3.

	E_pos	I_pos	tynd	tyk	lille
0	3558	3619	3856	3975	3860
1	523	462	225	106	221

Table 2.3: Number of observations for each outcome of the binary response variables (there were no missing values for these variables)

The binary response variables can be visualised by looking at the proportion of positive outcomes at each sample_round. For the parasite counts (_OPG), and for F_score, the mean of the responses at each round can be visualised.

In Fig. 2.7, the mean parasite counts can be compared to the proportion positive. The incidence patterns over time for kits and sows are more similar for Isospora than for Eimeria. Note that the high Eimeria proportion positive for kits in sample_week 19 is based on a much smaller sample size than the other sample_weeks of 4 kit litters (see

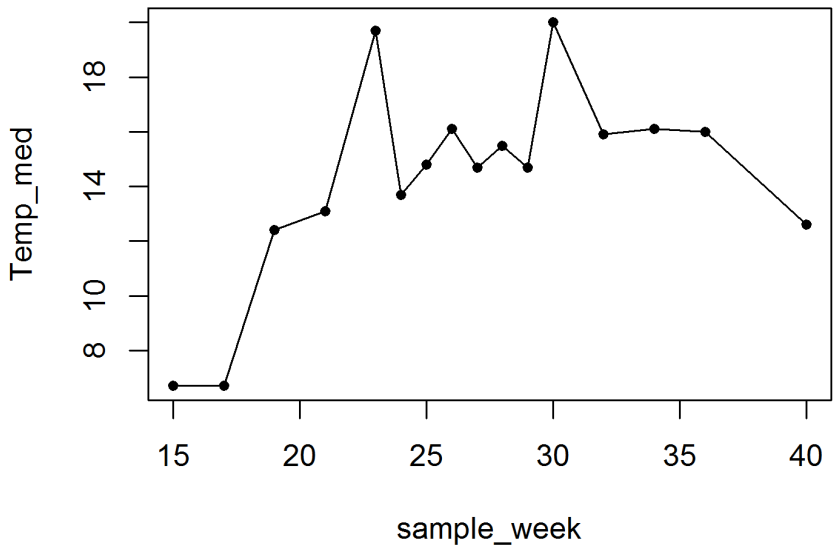

Figure 2.6: The processed temperature variable (the median daily temperature during the preceding week) at each sample date

Eimeria Types	Count
None	3559
lille only	204
tynd only	200
tyk only	89
tynd + lille	12
tyk + lille	4
tyk + tynd	12
tyk + tynd + lille	1

Table 2.4: Number of observations showing each combination of Eimeria subtypes.

F_score	Count
Missing	91
0	2
1	115
2	783
3	2520
4	546
5	24

Table 2.5: Number of observations for each F_score response.

Table 2.1).

Fig. 2.8 suggests that there may be some difference between feed suppliers, and that kits generally have a higher incidence than sows (especially of Isospora).

Fig. 2.9 shows the incidence rates for each Eimeria subtype, per feed supplier. The Tynd subtype shows the greatest difference between feed suppliers, with higher incidence rate in feed supplier B than A and C.

Fig. 2.10 shows that the mean F_score is not differ greatly between kits and sows, but appears higher for feed supplier A than for feed suppliers B and C.

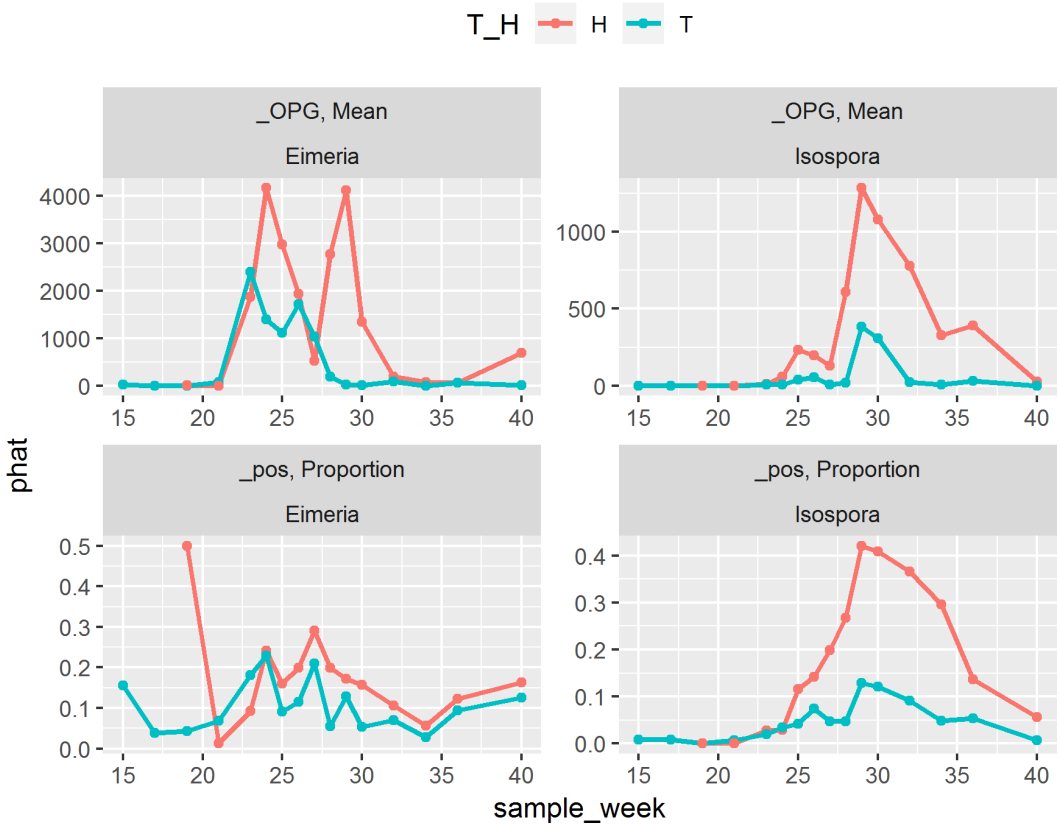

Figure 2.7: Mean counts (top row) and Proportion positive (bottom row) for Eimeria (left) and Isospora (right) over time, for kits (H) and sows (T).

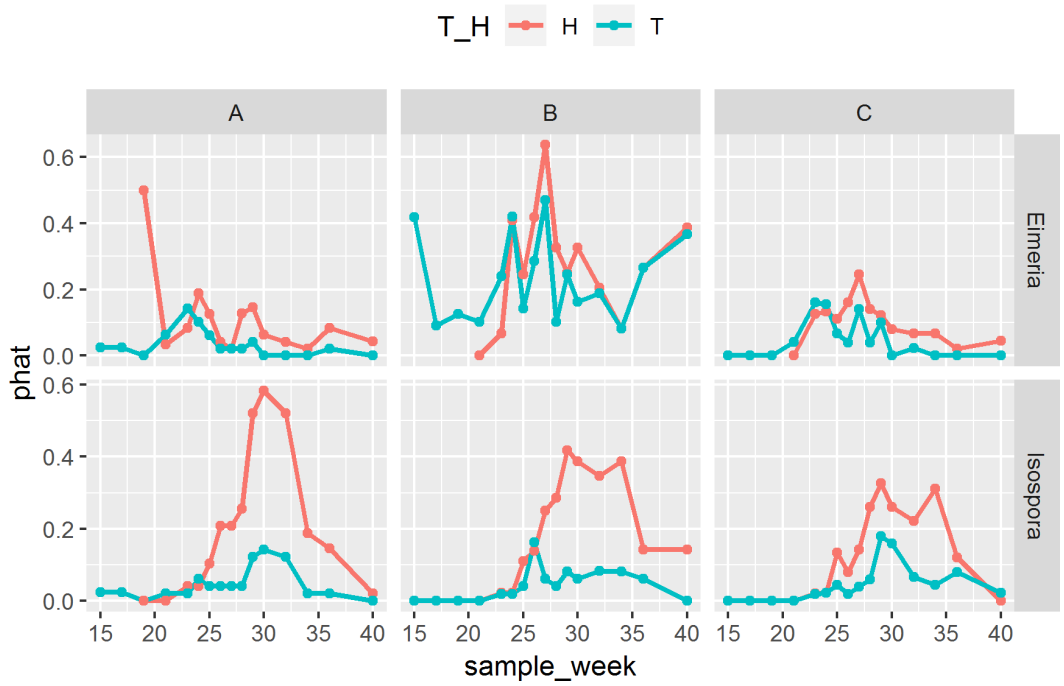

Figure 2.8: Proportion positive for Eimeria (top) and Isospora (bottom) per feed supplier (columns) for kits and sows

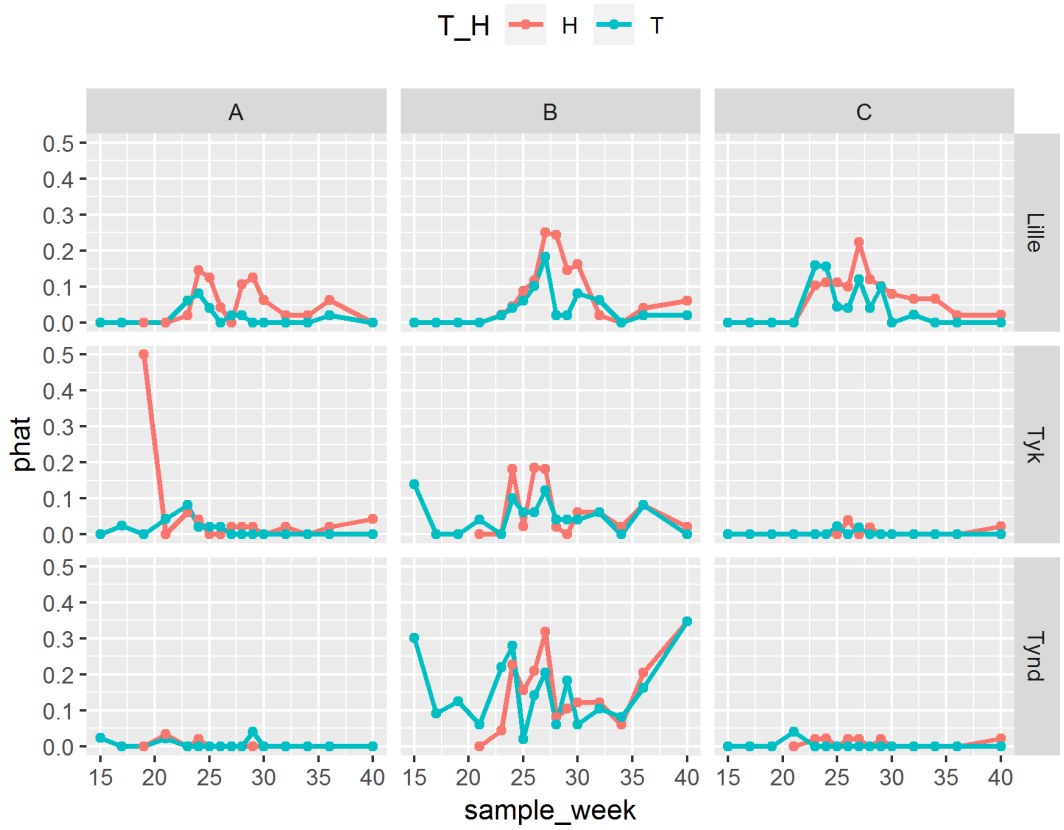

Figure 2.9: Proportion positive for kits and sows per feed supplier (columns) and per Eimeria subtype (rows)

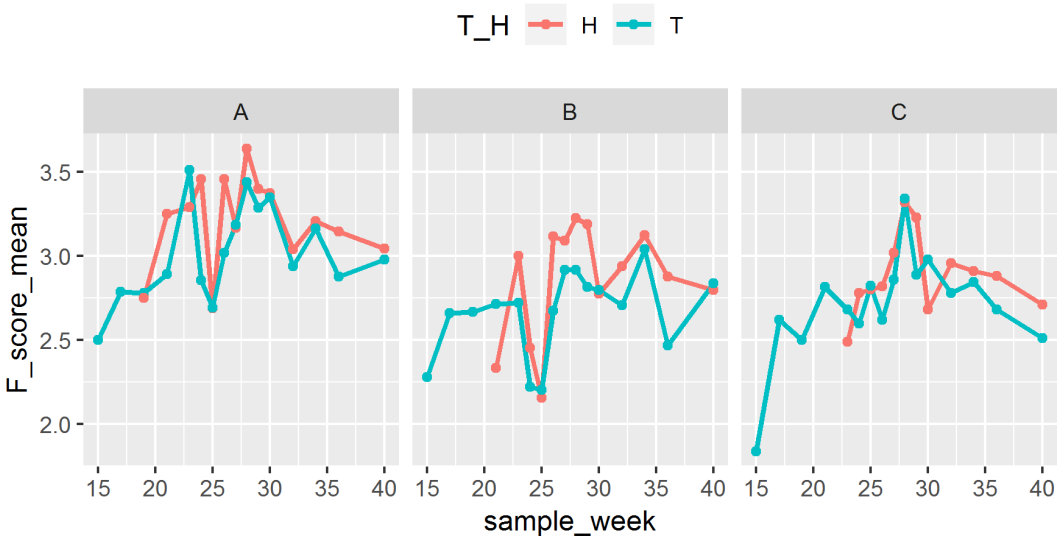

Figure 2.10: Mean F_score for kits and sows for each feed supplier

3 Methods

The aim was to fit statistical models to the data in order to determine which factors have a significant effect on the 6 response variables. The following challenges were identified:

- The response variables were binary (or ordinal for the F_score)
- There appears to be a non-linear relationship between parasite incidence and time (resulting in the irregularly shaped "incidence profiles" shown in the previous section)
- The samples are not independent - samples from the same farm/family/individual are likely to be correlated.

The first challenge was addressed by using logistic regression models which can model the probability of a positive outcome when the response variable is binary. For the F_score, ordinal logistic regression was used.

The second challenge was addressed by augmenting the data with B-spline basis functions as additional independent variables, allowing the model greater flexability to estimate the incidence profiles over time.

The third challenge was addressed by including a random farm effect in the model, so that the model can adjust to the varying "baseline" incidence rates in each farm.

Finally, backwards model selection was used to identify the significant effects to keep in the final model. By combining these techniques, a single model was obtained for each response variable.

3.1 Choice of B-spline basis functions

To model non-linear relationships between a response variable, y and independent variable x, one common approach is to use polynomial regression which entails including higher powers of x in the model (e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 ...$). However, if y is changing a lot over the range of x then it can be necessary to include very high powers of x to explain the fit adequately. An alternative method involves dividing the range of x into smaller sections and fitting lower order polynomials within each section. B-spline basis functions provide a way to do this[1]. For K "knots" (dividing points between the sections) and for polynomials of degree D, then $K + D$ B-spline basis functions can be defined which result in $K + D$ new variables, say $time_1, time_2, ..., time_{K+D}$ that are included as new terms in the model and allow the model to fit to local patterns over the range of x. The method used for selecting K and D was based on visual assessment of

the fit of the following two simple logistic models for different values of K and D:

$$\text{logit}(\Pr(E\_pos = 1)) = \beta_0 + \beta_1 time_1 + \beta_2 time_2 + ... + \beta_{K+D} time_{K+D} \quad (3.1)$$
$$\text{logit}(\Pr(I\_pos = 1)) = \beta_0 + \beta_1 time_1 + \beta_2 time_2 + ... + \beta_{K+D} time_{K+D} \quad (3.2)$$

The original independent variable used to calculate the splines was sample_week. The position of the knots was determined by the $K + 1$ quantiles of sample_week. Fitting of the splines was done using splines package in R[2].

As shown in section 4, values of $K = 3$ knots, and polynomial degree $D = 2$ were selected based on the above method. This resulted in 5 time variables to be considered in the more advanced models: $time_1, time_2, ..., time_5$.

3.2 Further Data Processing

There were only 4 samples for kits in sample_week 19 (see Table 2.1). These 4 samples would influence the model disproportionately and were therefore removed. As the birth dates of the kits was also of interest, the samples for the kit litter registered as born on 2 March (see Fig. 2.2) were also removed as this date was suspected to be an error, and such an extreme value would greatly influence the model. The resulting dataset contained 4050 observations.

Computational estimation of mixed logistic regression models is more robust when the variables are of a similar range in magnitude. Therefore, the numerical independent variables were transformed as follows (the centring and scaling parameters were chosen as the closest integer to the mean and standard deviation respectively):

- H_born was scaled as the number of days since 29-04-2016 (the median birth date) divided by 4. I.e. a birthdate of 30-04-2016 becomes 0.25.
- H_no_born was scaled by subtracting 7 and dividing by 2.
- Temp was scaled by subtracting 15 and dividing by 3.
- T_age was not scaled

Due to few samples with an F_score of 0 or 5 (see Table 2.5), the F_score variable was processed by combing outcome 0 with outcome 1, and combining outcome 5 with outcome 4. Therefore, the procesed F_score variable had 4 levels: F_score = "≤ 1", "2", "3", or "≥ 4".

3.3 Model specification

3.3.1 Binary Responses: E_pos, I_pos, tyk, tynd, lille

For each of the 5 binary response variables ($y = E\_pos, I\_pos, tynd, tyk$, and $lille$), the following mixed-effect, logistic regression model was specified as the initial model:

$$\begin{aligned}\text{logit}(\Pr(y=1)) &= u_{farm} \\ &+ \beta_0 + \beta_1 time_1 + \beta_2 time_2 + \beta_3 time_3 + \beta_4 time_4 + \beta_5 time_5 \\ &+ \beta_6 H\_no\_born + \beta_7 H\_born + \beta_8 Temp + \beta_9 T\_age + \beta_{Foder\_central} + \beta_{T\_H} \\ &+ \beta_{T\_H\_1} time_1 + \beta_{T\_H\_2} time_2 + \beta_{T\_H\_3} time_3 + \beta_{T\_H\_4} time_4 + \beta_{T\_H\_5} time_5 \\ &+ \beta_{T\_H\_6} H\_no\_born + \beta_{T\_H\_7} H\_born + \beta_{T\_H\_8} Temp + \beta_{T\_H\_9} T\_age + \beta_{T\_H\_Foder\_central} \\ &+ \epsilon \quad (3.3)\end{aligned}$$

where u_{farm} is the random intercept for each of the 30 farms, $u_{farm} \sim \mathcal{N}(0, \sigma^2_{farm})$, and ϵ is the model error. All the other terms in the model are fixed effects, where lines 2 and 3 in Eq. 3.3 contain the main effects for all the variables of interest, and line 3 and 4 the interaction of those same variables with T_H.

The initial model was fitted, and then backwards model selection was applied where the least significant term (based on likelihood ratio test p-values) was removed from the model until all remaining terms in the model had a significance of $p < 0.05$. During this procedure, $\beta_1 time_1 + \beta_2 time_2 + \beta_3 time_3 + \beta_4 time_4 + \beta_5 time_5$ were treated collectively as a single term in the model, and similarly for $\beta_{T\_H\_1} time_1 + \beta_{T\_H\_2} time_2 + \beta_{T\_H\_3} time_3 + \beta_{T\_H\_4} time_4 + \beta_{T\_H\_5} time_5$. This is appropriate because these 5 time variables were defined collectively by the B-spline functions based on the single sample_week variable, and so they should not be treated individually.

The models were fitted using the lme4 package[3] in R which estimates the parameters via maximum likelihood.

3.3.2 Ordinal Response: F_score

In order to model F_score which was not binary but rather had four ordered outcomes, ordinal logistic regression was used. The following initial model was specified:

$$\begin{aligned}\text{logit}(\Pr(F\_score > j)) &= u_{farm} - \theta_j \\ &+ \beta_1 time_1 + \beta_2 time_2 + \beta_3 time_3 + \beta_4 time_4 + \beta_5 time_5 + \beta_{Foder\_central} + \beta_{T\_H} \\ &+ \epsilon \quad (3.4)\end{aligned}$$

For $j = 1, 2$ and 3. Again, a random farm effect, u_{farm}, is included in the model. For different j, the model predicts $\Pr(F\_score > j)$, by using a different intercept, θ_j, but otherwise the model assumes the same relationship between $\Pr(F\_score > j)$ and the independent variables for all j.

Due to the greater complexity of ordinal logistic regression, it was computationally necessary to start with a more simple initial model than for the binary outcome variables. Then, forwards model selection was applied to check for significance of all the additional terms in Eq. 3.3, but also the 5 binary variables (E_pos, I_pos, tyk, tynd, lille) were treated as potential explanatory variables and tested for significance. The most significant candidate term was added to the model until no new terms had a significance below 0.05.

The model was fitted using the ordinal package[4] in R to obtain maximum likelihood estimates of the parameters.

3.3.3 Autoregressive Models to investigate sow to kit transmission effect

To investigate the hypothesis that kits are more likely to test positive for the parasites if their sows were positive, an autoregressive component was included in the model structure.

First, the subset of samples for kits was extracted. Then, for each kit sample, the parasite outcomes for E_pos and I_pos for the kit itself, and the kit's mother (identified based on the mink_id numbers), from the preceding sample_round, was appended to the dataset as a new explanatory variable. These variables were renamed E_pos_H, I_pos_H, E_pos_T and I_pos_T.

Finally the following initial model was fitted for $y = E\_pos$ and $y = I\_pos$:

$$\begin{aligned}\text{logit}(\Pr(y_{round=j} = 1)) &= u_{farm} \\ &+ \beta_0 + \beta_1 time_1 + \beta_2 time_2 + \beta_3 time_3 + \beta_4 time_4 + \beta_5 time_5 \\ &+ \beta_6 H\_no\_born + \beta_7 H\_born + \beta_8 Temp + \beta_9 T\_age + \beta_{Foder\_central} \\ &+ E\_pos\_T_{round=j-1} + I\_pos\_T_{round=j-1} + E\_pos\_H_{round=j-1} + I\_pos\_H_{round=j-1} \\ &+ \epsilon \quad (3.5)\end{aligned}$$

where u_{farm} is the random farm effect. The same spline basis functions from the previous models, time_1 to time_5 are used.

The preceding parasite status for the kit itself was included, in order to determine whether a possible transmission effect from the sow exists, even when controlling for "self" transmission from the kit itself.

Backwards model selection was applied in the same way as described in section 3.3.1.

The models were fitted using the lme4 package[3] in R which estimates the parameters via maximum likelihood.

4 Results

4.1 Defining B-spline basis functions

The results of fitting the models described in Section 3.1 are shown in Fig. 4.1 and Fig. 4.2. Based on these plots, $D = 2$ and $K = 3$ were selected as sufficient. The resulting B-spline basis functions are plotted against sample_week in Fig. 4.3.

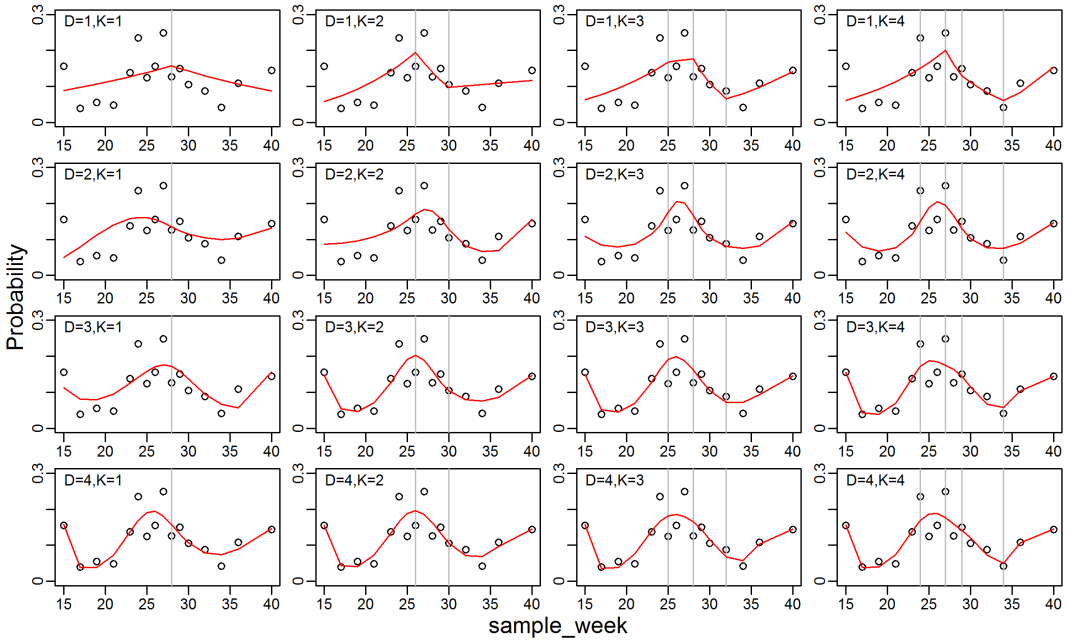

Figure 4.1: For E_pos, observed proportion positive (black points) and fitted probabilities of positive response (red line) using the spline regression models with different values of D and K. Vertical grey lines indicate the knot locations

4.2 Model Coefficients and odds ratios

The estimated coefficients for the first 6 final models are all summarised in Table 4.1. Note that the following variables were not found to be significant in any of the models, so do not appear in the table:

- H_born
- T_age
- Temp

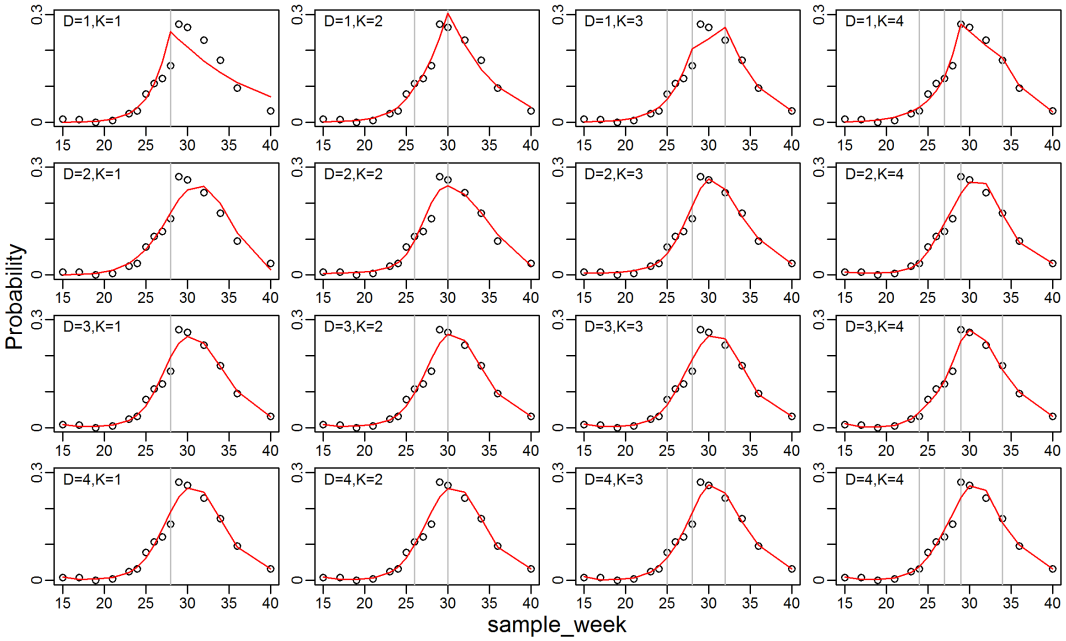

Figure 4.2: For I_pos, observed proportion positive (black points) and fitted probabilities of positive response (red line) using the spline regression models with different values of D and K. Vertical grey lines indicate the knot locations

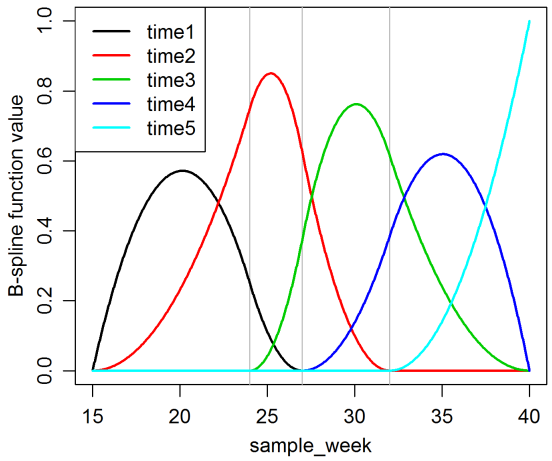

Figure 4.3: B-sline basis functions resulting from using $D = 2$ and $K = 3$. Vertical grey lines indicate the knot locations

- H_born:T_H
- T_age:T_H
- Temp:T_H
- H_no_born:T_H.

In addition, non of the binary variables (E_pos, I_pos, tyk, tynd, lille) were found to be significant in the model for F_score.

The number of observations used to fit the first 5 models in Table 4.1 was 4050. As the value of F_score was missing for 91 samples, 3959 samples were used to fit the F_score model.

It is worth noting that $p > 0.1$ for all of time_1, time_2, time_3, time_4 and time_5 in the model for lille. However, when tested collectively with the likelihood ratio test, these 5 terms combined were highly significant and it is therefore appropriate to keep them in the model.

	Eimeria	Isospora	tyk	tynd	lille	F_score
(Intercept)	-28.337***	-33.169***	-3.749***	-4.765***	-58.139***	$\hat{\theta}_1$ =-3.169***
						$\hat{\theta}_2$ =-0.691**
						$\hat{\theta}_3$ =2.842***
time_1	24.529***	27.907**	-2.517**	-1.989**	54.716***	2.303***
time_2	26.428***	30.890***	0.736	-0.200	55.433***	1.260***
time_3	25.862***	33.268***	-1.354*	-1.252**	55.745***	3.137***
time_4	24.428***	31.567***	-0.363	-1.179*	52.931***	1.094***
time_5	25.824***	30.111***	-0.972	0.339	54.055***	1.495***
H_no_born	0.127*		0.192*			
FdCtB	1.735***		1.376***	3.746***	0.824*	-1.249***
FdCtC	0.393		-1.238**	0.295	0.717*	-1.112***
T_HT	24.977***	27.999**			7.132	-0.462***
time_1:T_HT	-25.838***	-28.554**			-5.138	
time_2:T_HT	-25.547***	-28.756**			-7.899	
time_3:T_HT	-26.583***	-30.148***			-8.453	
time_4:T_HT	-25.314***	-29.340**			-8.122	
time_5:T_HT	-25.678***	-29.776***			-8.579	
FdCtB:T_HT	0.556'					
FdCtC:T_HT	0.037					
$\hat{\sigma}_{farm}$	0.327	0.583	0.157	0.128	0.575	0.405

Table 4.1: Model Coefficients for all 6 models with significance codes based on Wald p values. Foder_center has been abbreviated to FdCt.

"***"	$p < 0.001$
"** "	$0.001 \leq p$ 0.01
" * "	$0.01 \leq p < 0.05$
" ' "	$0.05 \leq p < 0.1$
" "	$0.1 \leq p \leq 1$

More detailed tables of coefficients and p-values for each model are included in Appendix A.

Estimates of the odds ratios, with 95% confidence intervals, for differences between feed suppliers, and between kits and sows, are shown in Tables 4.2 and 4.3.

4.3 Model checking and visualisations

To validate the models, the assumption that the random effect is normally distributed can be checked visually. Fig. 4.4 shows that the assumption is fairly reasonable for most of the models.

Next, the fitted values (probabilities for the binary variables, and mean for F_score) from the models can be visually compared with the observed proportions. These plots are also useful for interpretation of the models, and to visualise the effect of the different coefficients and odds ratios in tables 4.1/4.2/4.3.

Variable	Comparison	E_pos
Foder_central	B vs. A	$5.67_{T\_H=H}$ [3.55, 9.05]
		$9.88_{T\_H=T}$ [5.85, 16.70]
	C vs. A	$1.48_{T\_H=H}$ [0.90, 2.44]
		$1.54_{T\_H=T}$ [0.85, 2.79]
T_H	H vs. T	$3.77_{FdrCent=A}$ [1.98, 7.17]
		$2.16_{FdrCent=B}$ [1.36, 3.44]
		$3.63_{FdrCent=C}$ [2.02, 6.52]

Table 4.2: Odds Ratios for E_pos, for comparing different feed suppliers, and kits to sows, with 95% confidence limits in square brackets. Due to the interaction between T_H and time, the T_H odds ratios assume that sample_week = 30. As the model for E_pos contains an interaction between Foder_central and T_H, the odds ratio for one of them depends on the level of the other, as specified in the table.

	I_pos	tyk	tynd	lille	F_score
B vs. A	1	3.96 [2.38, 6.6]	42.4 [18.6, 96.6]	2.28 [1.2, 4.34]	0.29 [0.19, 0.42]
C vs. A	1	0.29 [0.12, 0.73]	1.34 [0.46, 3.91]	2.05 [1.08, 3.9]	0.33 [0.22, 0.49]
H vs. T	6.67 [4.51, 9.86]	1	1	3.39 [1.88, 6.09]	1.59 [1.38, 1.82]

Table 4.3: Odds Ratios for the other response variables, for comparing different feed suppliers, and kits to sows, with 95% confidence limits in square brackets. It is assumed that sample_week = 30 for the H vs. T odds ratio for I_pos and lille as these models contain interactions between T_H and time.

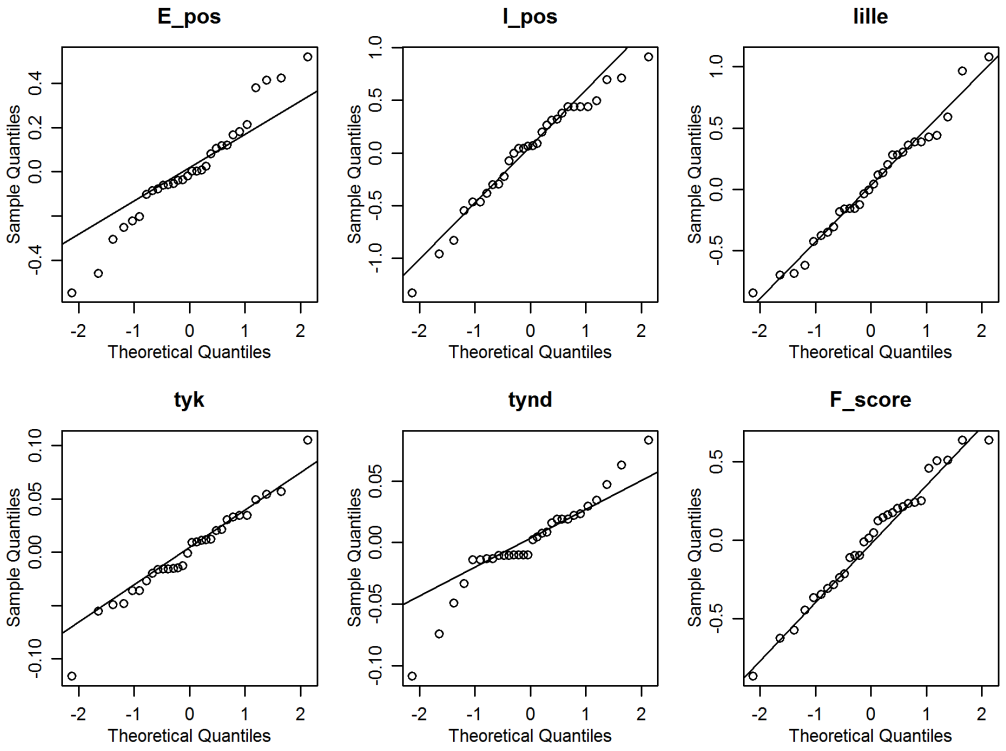

Figure 4.4: Normal qq-plot of the random intercepts, u_{farm}, for the six models.

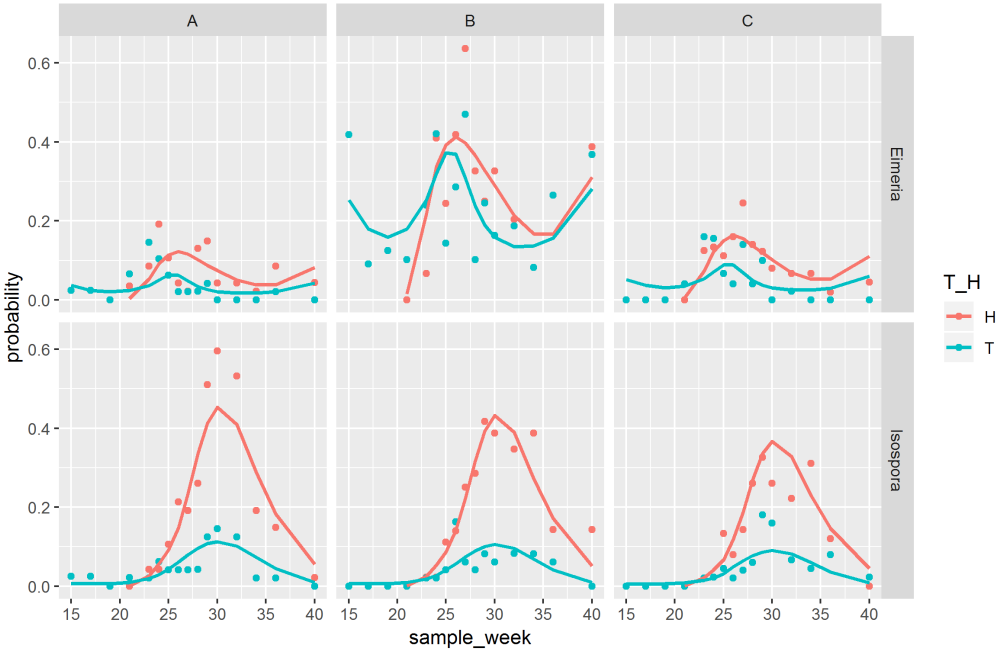

Figure 4.5: Fitted probabilities (lines) and observed proportions (points) for E_pos (top row) and I_pos (bottom row) for Feed supplier A, B and C (columns)

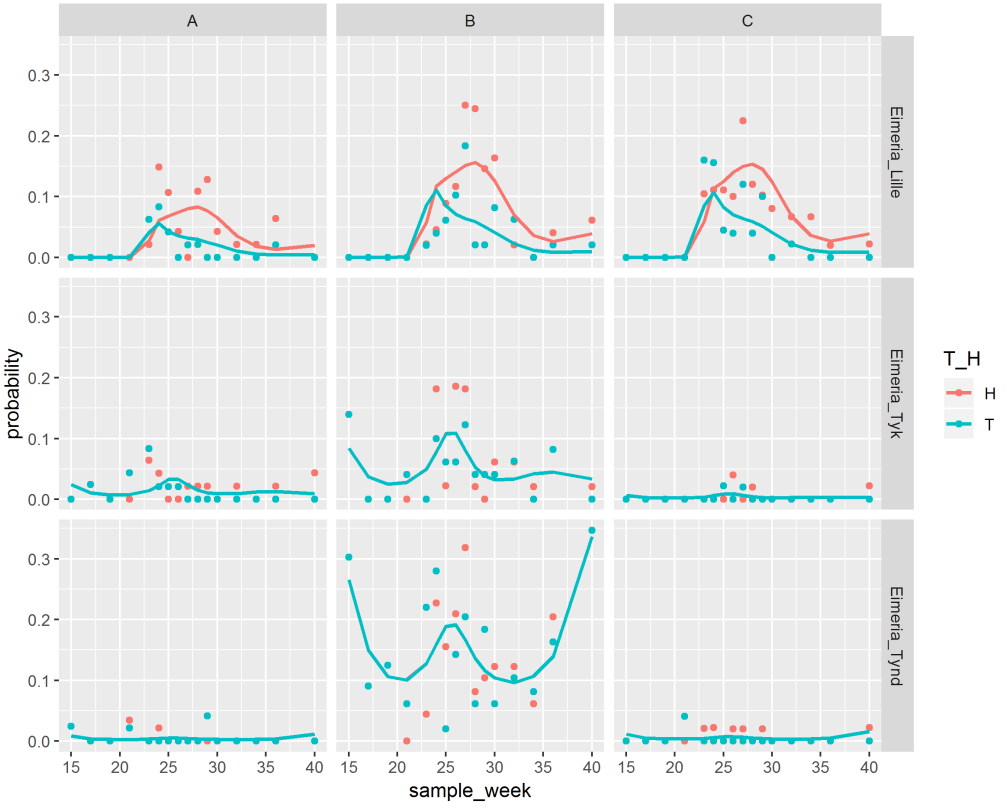

Figure 4.6: Fitted probabilities (lines) and observed proportions (points) for the three Eimeria subtypes (rows) for Feed supplier A, B and C (columns)

4.4 Autoregressive Models

The coefficients for the final autoregressive models for predicting E_pos and I_pos for kits based on past parasite status of the kits and sows, are summarised in Table 4.4, and

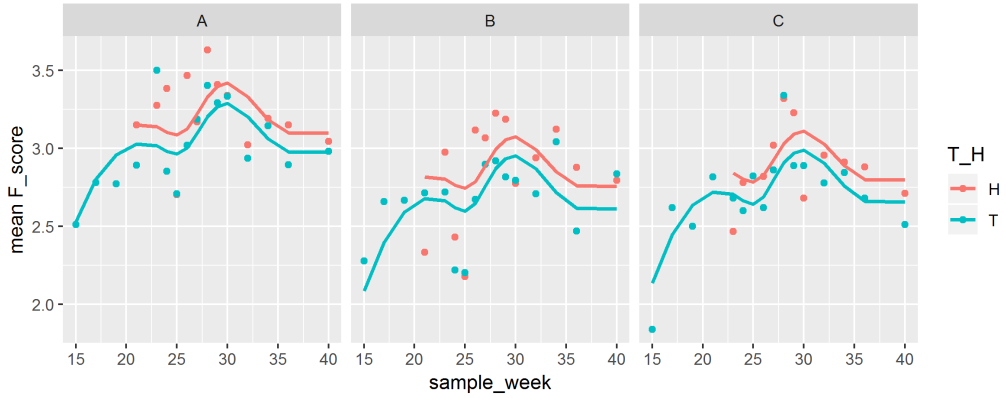

Figure 4.7: Expected mean F_score from the model (lines) and observed mean F_score (points) for Feed supplier A, B and C (columns)

more detailed tables of the coefficients are included in Appendix A. Odds ratios for the parasite status explanatory variables are included in Table 4.5

As only the outcome for kits is modelled, and only samples for which the preceding kit and sow outcomes exist can be used, the number of observations used to fit these models was 1598.

Besides time_1 to time time_5, Foder_central, and parasite status in preceding round, no other explanatory variables were found to be significant in these two models.

	E_pos	I_pos
(Intercept)	-89.512***	66.620*
time_1	87.087**	-73.282*
time_2	87.376**	-68.792*
time_3	86.922**	-66.955*
time_4	85.688**	-68.482*
time_5	86.952**	-69.781*
Foder_centralB	1.528***	
Foder_centralC	0.281	
E_positiv_T	0.487*	
E_positiv_H	0.452*	
I_positiv_T	0.600*	0.593*
I_positiv_H		0.711***
$\hat{\sigma}_{farm}$	0.2674	0.4249

Table 4.4: Model coefficients for the autoregressive models (same significance codes as in Table 4.1)

.

Comparison	E_pos	I_pos
E_pos_T: 1 vs. 0	1.63 [1.11, 2.38]	1
E_pos_H: 1 vs. 0	1.57 [1.10, 2.24]	1
I_pos_T: 1 vs. 0	1.82 [1.05, 3.15]	1.81 [1.12, 2.93]
I_pos_H: 1 vs. 0	1	2.04 [1.50, 2.76]

Table 4.5: Odds ratios for autoregressive models

5 Discussion and Conclusion

The first models for E_pos and I_pos confirmed that animal status had a significant effect on parasite outcome, with odds of being positive for Eimeria at least doubled for kits compared to sows (Table 4.2). For Isospora, the effect was even greater, with odds of being positive over 6 times greater for kits compared to for sows (Table 4.3).

Feed supplier was also found to have a significant effect on Eimeria, as well as it's three subtypes, with Feed supplier B always having the biggest odds of a positive outcome (Tables 4.2 and 4.3, and Figs. 4.5 and 4.6). However, Feed supplier was not found to be significant for Isospora outcome.

The F_score model found that Feed supplier B and C had lower odds of a higher F_score, compared to feed supplier A.

The spline basis functions to model the time relationship were found to be highly significant in all of the models considered. However, besides the explanaotry variables just mentioned (Animal status, feed supplier, time), only litter size was found to be significant in the models for E_pos, and tyk. Otherwise, none of the other independent variables (sow age, temperature, birth date) were found to have a significant effect in any of the models.

It was also notable that no parasite status (E_pos, I_ops, tyk, tynd, lille) was found to be significant for predicting F_score.

Finally, the autoregressive models provide support for the hypothesis that the sows transmit the parasites to the kits. Even when controlling for the previous parasite status of the kit itself, odds of a kit litter being positive for Eimeria were 1.63 times greater if the litter's sow was positive for Eimeria in the preceding measurement round (Table 4.5). Curiously, the odds of a kit litter being positive for Eimeria were 1.82 times greater if the litter's sow was positive for *Isospora* in the previous round. Odds of a kit litter being positive for Isospora were 1.81 times greater if the litter's sow was positive for Isospora in the preceding measurement round.

Bibliography

[1] Hastie, T. J. (1992) Generalized additive models. Chapter 7 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

[2] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. URL https://www.R-project.org/.

[3] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[4] Christensen, R. H. B. (2019). ordinal - Regression Models for Ordinal Data. R package version 2019.3-9. http://www.cran.r-project.org/package=ordinal/.

A Model Coefficient Tables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-28.3367	6.5643	-4.3168	0.0000
time_1	24.5285	6.8623	3.5744	0.0004
time_2	26.4281	6.5303	4.0470	0.0001
time_3	25.8621	6.5843	3.9279	0.0001
time_4	24.4279	6.5551	3.7266	0.0002
time_5	25.8238	6.5659	3.9330	0.0001
H_no_born	0.1274	0.0519	2.4540	0.0141
Foder_centralB	1.7348	0.2387	7.2678	0.0000
Foder_centralC	0.3929	0.2556	1.5372	0.1242
T_HT	24.9766	6.5671	3.8033	0.0001
time_1:T_HT	-25.8379	6.8777	-3.7568	0.0002
time_2:T_HT	-25.5469	6.5336	-3.9101	0.0001
time_3:T_HT	-26.5833	6.5922	-4.0325	0.0001
time_4:T_HT	-25.3144	6.5690	-3.8536	0.0001
time_5:T_HT	-25.6776	6.5725	-3.9068	0.0001
Foder_centralB:T_HT	0.5558	0.2848	1.9515	0.0510
Foder_centralC:T_HT	0.0375	0.3334	0.1124	0.9105

Table A.1: Model 1 (E-pos) coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-33.1691	8.9291	-3.7147	0.0002
time_1	27.9068	9.1988	3.0337	0.0024
time_2	30.8896	8.9172	3.4641	0.0005
time_3	33.2679	8.9364	3.7228	0.0002
time_4	31.5674	8.9282	3.5357	0.0004
time_5	30.1109	8.9352	3.3699	0.0008
T_HT	27.9989	8.9424	3.1310	0.0017
time_1:T_HT	-28.5544	9.2816	-3.0765	0.0021
time_2:T_HT	-28.7558	8.9333	-3.2190	0.0013
time_3:T_HT	-30.1475	8.9563	-3.3661	0.0008
time_4:T_HT	-29.3397	8.9538	-3.2768	0.0010
time_5:T_HT	-29.7757	8.9846	-3.3141	0.0009

Table A.2: Model 2 (I-pos) coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.7495	0.4872	-7.6967	0.0000
time_1	-2.5173	0.9055	-2.7801	0.0054
time_2	0.7356	0.4764	1.5440	0.1226
time_3	-1.3536	0.6007	-2.2536	0.0242
time_4	-0.3626	0.6958	-0.5212	0.6023
time_5	-0.9716	0.6479	-1.4996	0.1337
H_no_born	0.1920	0.0970	1.9794	0.0478
Foder_centralB	1.3762	0.2606	5.2805	0.0000
Foder_centralC	-1.2380	0.4697	-2.6356	0.0084

Table A.3: Model 3 (tyk) coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.7654	0.5058	-9.4218	0.0000
time_1	-1.9889	0.6290	-3.1621	0.0016
time_2	-0.1997	0.3416	-0.5845	0.5589
time_3	-1.2515	0.4207	-2.9751	0.0029
time_4	-1.1793	0.4812	-2.4508	0.0143
time_5	0.3391	0.3667	0.9250	0.3550
Foder_centralB	3.7464	0.4207	8.9057	0.0000
Foder_centralC	0.2952	0.5449	0.5417	0.5881

Table A.4: Model 4 (tynd) coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-58.1386	10.4299	-5.5743	0.0000
time_1	54.7157	10.6989	5.1142	0.0000
time_2	55.4329	10.4071	5.3264	0.0000
time_3	55.7446	10.4431	5.3380	0.0000
time_4	52.9312	10.4295	5.0752	0.0000
time_5	54.0551	10.4340	5.1807	0.0000
Foder_centralB	0.8239	0.3282	2.5104	0.0121
Foder_centralC	0.7172	0.3289	2.1808	0.0292
T_HT	7.1316	10.5505	0.6759	0.4991
time_1:T_HT	-5.1381	10.9660	-0.4686	0.6394
time_2:T_HT	-7.8993	10.5117	-0.7515	0.4524
time_3:T_HT	-8.4529	10.5923	-0.7980	0.4249
time_4:T_HT	-8.1221	10.5764	-0.7679	0.4425
time_5:T_HT	-8.5790	10.5828	-0.8107	0.4176

Table A.5: Model 5 (lille) coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
1|2	-3.1690	0.2310	-13.7208	0.0000
2|3	-0.6913	0.2211	-3.1264	0.0018
3|4	2.8415	0.2234	12.7216	0.0000
time_1	2.3033	0.2997	7.6859	0.0000
time_2	1.2605	0.1733	7.2724	0.0000
time_3	3.1370	0.2080	15.0838	0.0000
time_4	1.0943	0.2212	4.9475	0.0000
time_5	1.4950	0.2011	7.4351	0.0000
Foder_centralB	-1.2493	0.2000	-6.2475	0.0000
Foder_centralC	-1.1123	0.2000	-5.5613	0.0000
T_HT	-0.4623	0.0707	-6.5378	0.0000

Table A.6: Model 6 (F-score) coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-89.5122	26.6156	-3.3631	0.0008
time_1	87.0873	27.1199	3.2112	0.0013
time_2	87.3759	26.5577	3.2900	0.0010
time_3	86.9219	26.6448	3.2622	0.0011
time_4	85.6881	26.5889	3.2227	0.0013
time_5	86.9518	26.6154	3.2670	0.0011
Foder_centralB	1.5280	0.2326	6.5686	0.0000
Foder_centralC	0.2811	0.2477	1.1346	0.2565
E_positiv_T	0.4871	0.1944	2.5061	0.0122
I_positiv_T	0.6001	0.2796	2.1465	0.0318
E_positiv_H	0.4520	0.1807	2.5019	0.0124

Table A.7: Autoregressive Model for E_pos in kits

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	66.6200	29.5692	2.2530	0.0243
time_1	-73.2820	30.2194	-2.4250	0.0153
time_2	-68.7923	29.5296	-2.3296	0.0198
time_3	-66.9555	29.5869	-2.2630	0.0236
time_4	-68.4821	29.5576	-2.3169	0.0205
time_5	-69.7808	29.5743	-2.3595	0.0183
I_positiv_T	0.5931	0.2460	2.4111	0.0159
I_positiv_H	0.7112	0.1552	4.5810	0.0000

Table A.8: Autoregressive Model for I_pos in kits

B Extra Plots

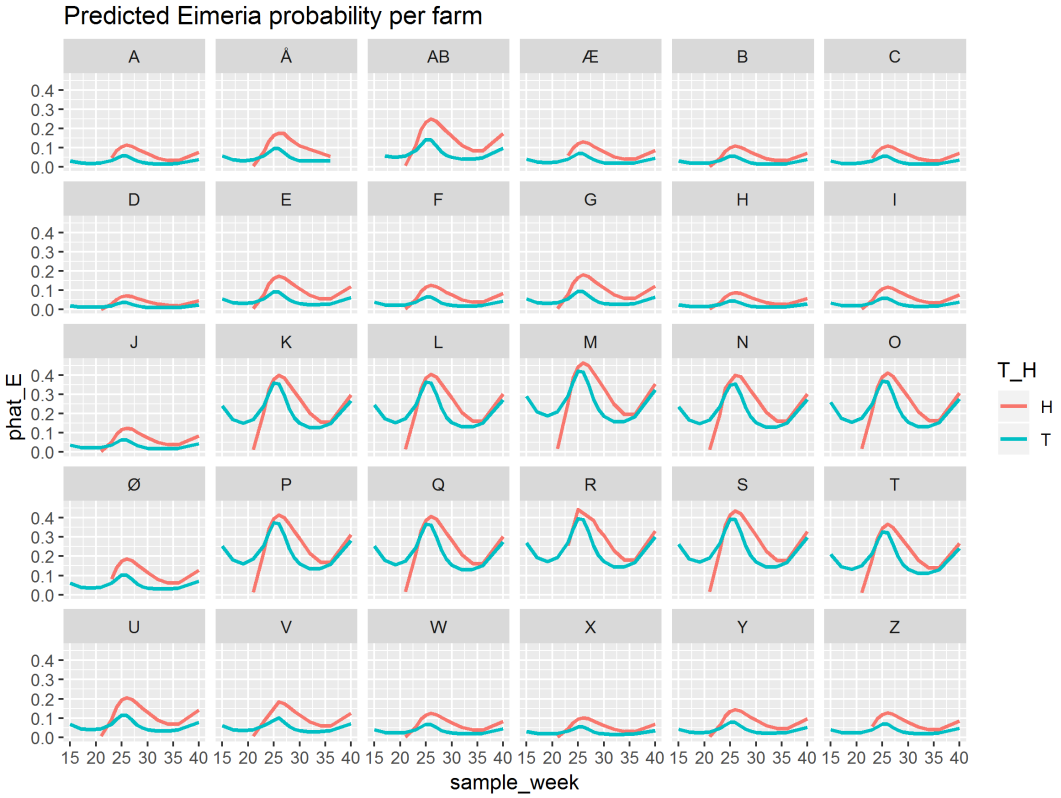

Figure B.1: Model probability that E_pos = 1 for each farm

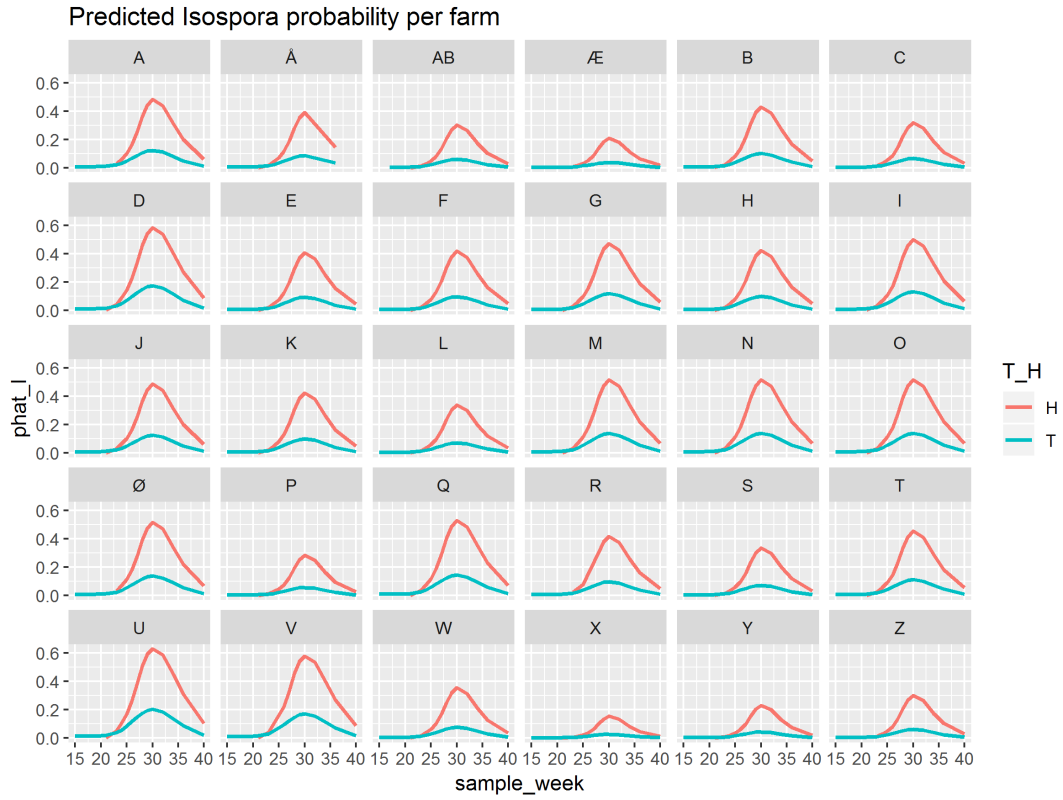

Figure B.2: Model probability that E_pos = 1 for each farm

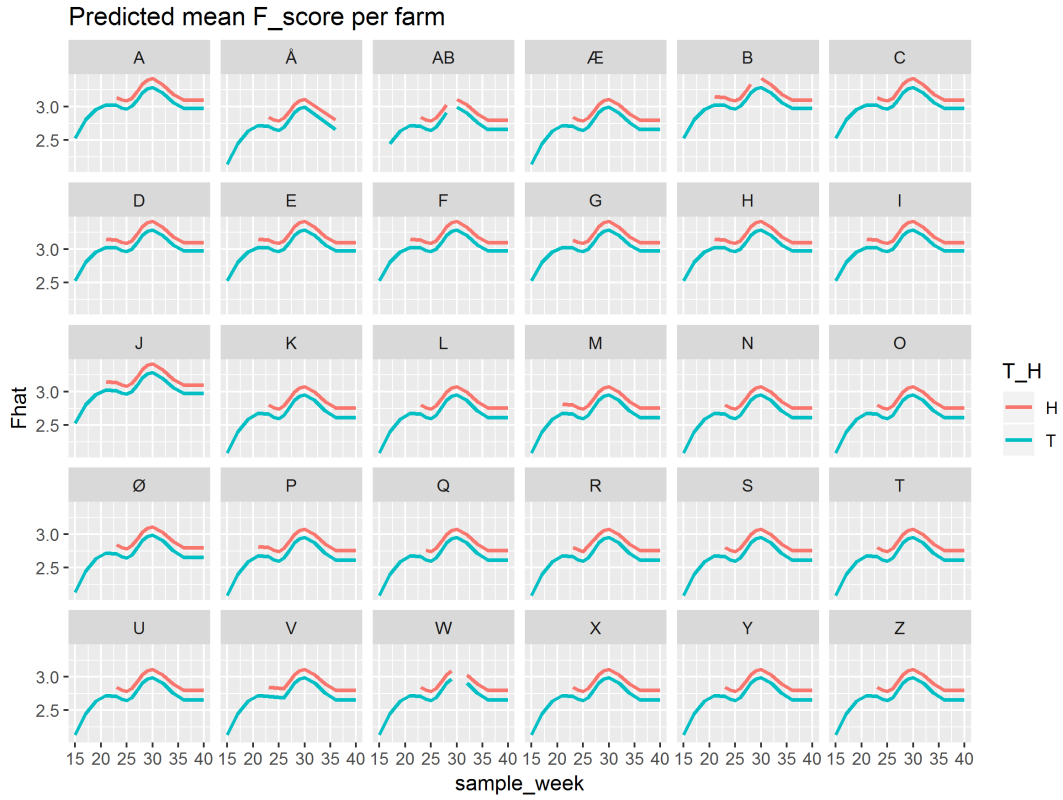

Figure B.3: Model mean F_score for each farm