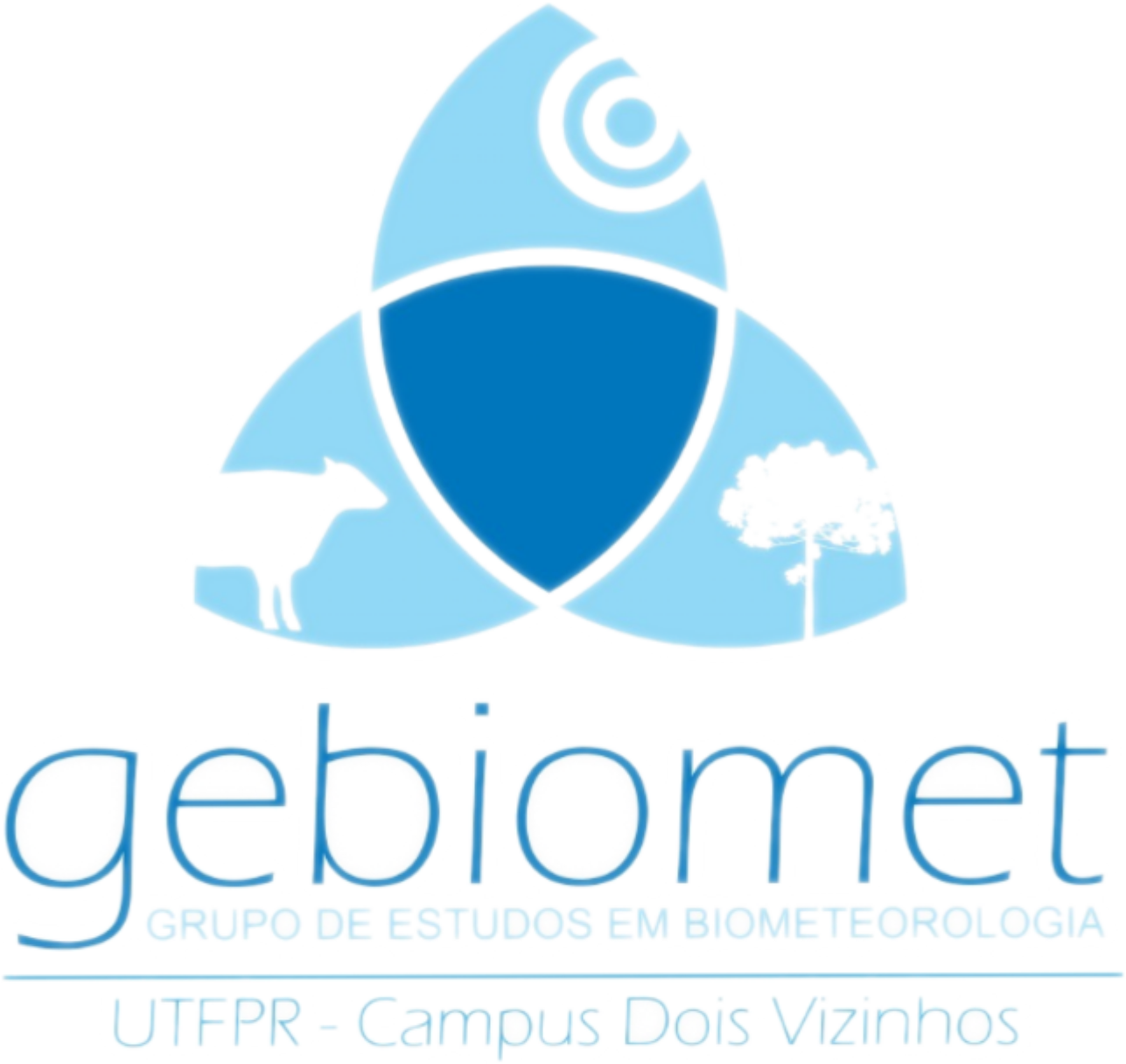

Analisys INTA

Edgar de S. Vismara e Frederico M. C. Vieira

October 9, 2023

Table of Content

1 Introduction

2 Background

2.1 Bayesian framework

The Bayesian framework, is a powerful and flexible statistical framework that provides a systematic and coherent approach to uncertainty and inference. At its core, Bayesian statistics is founded on the principles of probability theory and Bayes' theorem, which allow us to update our beliefs in light of new evidence.

According to McElreath (ano), Bayesian statistics emphasizes the importance of assigning probabilities to uncertain events or parameters. Unlike traditional frequentist statistics, where parameters are considered fixed and unknown, Bayesian statistics treats them as random variables with probability distributions. This paradigm shift allows us to incorporate prior knowledge or beliefs into our analyses, making Bayesian methods particularly well-suited for handling complex and uncertain problems.

One central concept in the Bayesian framework is the predictive posterior distribution. This distribution captures our updated beliefs about a parameter or the outcome of an event after observing new data. It combines our prior beliefs (expressed as a prior distribution) with the likelihood of the observed data given the parameter (expressed as a likelihood function) to produce a posterior distribution. This posterior distribution represents our updated uncertainty about the parameter or event, accounting for both prior knowledge and new evidence.

The predictive posterior distribution is especially valuable because it allows us to make predictions about future observations or events. By sampling from the posterior distribution, we can generate a range of possible outcomes, each weighted by its posterior probability. This approach not only provides point estimates but also quantifies uncertainty, making Bayesian statistics a powerful tool for decision-making, forecasting, and modeling in a wide range of fields, from science and engineering to finance and machine learning.

In summary the Bayesian framework emphasizes the integration of prior knowledge and observed data through probability theory, resulting in a coherent and principled approach to inference. The predictive posterior distribution, a fundamental concept in Bayesian statistics, encapsulates our updated beliefs and serves as a valuable tool for making informed decisions and predictions in the face of uncertainty.

2.2 Prior selection

According To Gelman (2006) we characterize a prior distribution as weakly informative if it is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available. In general any problem has some natural constraints that would allow a weakly-informative model. For example, for regression models on the logarithmic or logit scale, with predictors that are binary or scaled to have standard

deviation 1, we can be sure for most applications that effect sizes will be less than 10, etc.

The Regularization priors are a fundamental concept in Bayesian modeling. Their primary purpose is to prevent overfitting and stabilize model behavior by imposing penalties on parameter estimates. Regularization priors encourage parameter estimates to be close to zero or within a specific range, promoting model simplicity and generalization.

McElreath's approach emphasizes the importance of regularization priors in making Bayesian models more robust, especially in situations with limited data or complex model structures. Regularization priors help stabilize parameter estimates, reduce model complexity, and enhance model interpretability.

For example, suppose you're predicting housing prices based on features like square footage, bedrooms, and neighborhoods. You employ a Bayesian linear regression model with regularization priors (on coefficients ($\beta_i \sim \mathcal{N}(0, \sigma^2)$). The regularization strength parameter, σ^2, is set to 0.5, indicating a moderate level of regularization.

The regularization encourages coefficients to be small, preventing erratic model behavior even with limited data. This approach balances data fitting and model stability, yielding a more robust and interpretable predictive model for housing prices.

3 Data analysis

3.1 Descriptive analysis

3.2 Non gaussian models

3.2.1 Behavior model

$$
\begin{aligned}
y_{ijkl} &\sim \text{Poisson}(\lambda_{ijkl}) \\
\log(\lambda_{ijkl}) &= \alpha_{ijl} + C_j + D_k + L_l + (CDL)_{jkl} \\
\alpha_{ijl} &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2) \\
\mu_\alpha &\sim \text{Normal}(1, 0.3) \\
\sigma_\alpha^2 &\sim \text{Half-Cauchy}(0, 25) \\
C_j &\sim \text{Normal}(0, 0.5) \\
D_k &\sim \text{Normal}(0, 0.5) \\
L_l &\sim \text{Normal}(0, 0.5) \\
(CDL)_{jkl} &\sim \text{Normal}(0, 0.5)
\end{aligned}
$$

Where:

- y_{ijkl} represents the Poisson-distributed outcome for observation i within the combination of levels of four factors j, k and l.
- λ_{ijkl} represents the Poisson rate parameter for observation i within the combination of levels of three factors.
- The log of the Poisson rate parameter ($\log(\lambda_{ijkl})$) is modeled as the sum of the

random intercept (α_{ijl}) and the fixed effects (C_j, D_k, L_l), as well as the three-way interaction term ($(CDL)_{jkl}$).

- α_{ijl}, represents a random effect for log counts within the individual i, which is nested within the category j, which is further nested within locality l
- μ_α represents the mean of the random effects for the intercepts.
- σ^2_α represents the variance of the random effects for the intercepts.
- C_j, D_k, L_l are fixed effects for factors j, k, and l.
- $(CDL)_{jkl}$ represents a three-way interaction term.

3.2.1.1 Respiratory rate model

the same but with $\mu_\alpha \sim \text{Normal}(3.5, 0.5)$ instead of $\mu_\alpha \sim \text{Normal}(1, 0.3)$

3.2.1.2 Higyene and lameness Score model

$$
\begin{aligned}
y_{ijkl} &\sim \text{Ordinal}(\theta_{ijkl}) \\
\text{logit}(\theta_{ijkl}) &= \alpha_{ijl} + C_j + D_k + L_l + (CDL)_{jkl} \\
\alpha_{ijl} &\sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) \\
\mu_\alpha &\sim \text{Normal}(0, 1.5) \\
\sigma^2_\alpha &\sim \text{Half-Cauchy}(0, 25) \\
C_j &\sim \text{Normal}(0, 0.5) \\
D_k &\sim \text{Normal}(0, 0.5) \\
L_l &\sim \text{Normal}(0, 0.5) \\
(CDL)_{jkl} &\sim \text{Normal}(0, 0.5)
\end{aligned}
$$

where:

- y_{ijkl} represents the ordinal-distributed outcome for observation i within the combination of levels of three factors: j, k, and l.
- θ_{ijkl} represents the cumulative log-odds of the ordinal response categories for observation i within the combination of j categories, k localities, and l days.
- The logit link function ($\text{logit}(\theta_{ijkl})$) is modeled as the sum of the random intercept (α_{ijl}) and the coefficients for the main effects (C_j, D_k, L_l) and the triple interaction ($(CDL)_{jkl}$) between the three factors.
- α_{ijl} represents a random effect for the cumulative log-odds within the individual i, which is nested within the category j, which is further nested within locality l.
- μ_α represents the mean of the random effects for the cumulative log-odds.
- σ^2_α represents the variance of the random effects for the cumulative log-odds.
- C_j, D_k, L_l are coefficients for the main effects of the three factors.
- $(CDL)_{jkl}$ is a coefficient for the triple interaction between the three factors.

3.3 Gaussian models

3.3.1 Temperature of the bed model

$$
\begin{aligned}
y_{ijkl} &\sim \text{Normal}(\mu_{ijkl}, \sigma^2) \\
\mu_{ijkl} &= \alpha + M_j + D_k + L_l + (CDL)_{jkl} \\
\alpha &\sim \text{Normal}(25, 10) \\
M_j &\sim \text{Normal}(0, 5) \\
D_k &\sim \text{Normal}(0, 5) \\
L_l &\sim \text{Normal}(0, 5) \\
(MDL)_{jkl} &\sim \text{Normal}(0, 5) \\
\sigma^2 &\sim \text{Half-Cauchy}(0, 25)
\end{aligned}
$$

Where:

- y_{ijkl} represents the gaussian-distributed outcome for observation i within the combination of levels of three factors j, k and l.
- μ_{ijkl} represents the parameter mean of Gaussian distributions and is modeled as the sum of the intercept (α) and the predictors (M_j, D_k, L_l), as well as the three-way interaction term ($(MDL)_{jkl}$).
- M_j, D_k, L_l are the effects for factors j, k, and l.
- $(MDL)_{jkl}$ represents a three-way interaction term.
- σ^2 represents the variance of the outcome variable.

3.3.2 Milking production model

$$
\begin{aligned}
y_{ijkl} &\sim \text{Normal}(\mu_{ijkl}, \sigma^2) \\
\mu_{ijkl} &= \alpha_{ijk} + C_j + D_k + L_l + (CDL)_{jkl} \\
\alpha_{ijk} &\sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) \\
\mu_\alpha &\sim \text{Normal}(35, 10) \\
\sigma^2_\alpha &\sim \text{Half-Cauchy}(0, 25) \\
C_j &\sim \text{Normal}(0, 5) \\
D_k &\sim \text{Normal}(0, 5) \\
L_l &\sim \text{Normal}(0, 5) \\
(CDL)_{jkl} &\sim \text{Normal}(0, 5) \\
\sigma^2 &\sim \text{Half-Cauchy}(0, 25)
\end{aligned}
$$

Where:

- y_{ijkl} represents the gaussian-distributed outcome for observation i within the combination of levels of three factors j, k and l.
- μ_{ijkl} represents the parameter mean of Gaussian distributions and is modeled as the sum of the random intercept (α_{ijl}) and the fixed effects (C_j, D_k, L_l), as well as the three-way interaction term ($(CDL)_{jkl}$).
- α_{ijl} represents a random effect for the intercept within the individual i, which is nested within the category j, which is further nested within locality l.

- μ_α represents the mean of the random effects for the intercepts.
- σ^2_α represents the variance of the random effects for the intercepts.
- C_j, D_k, L_l are fixed effects for factors j, k, and l.
- $(CDL)_{jkl}$ represents a three-way interaction term.
- σ^2 represents the variance of the outcome variable.

You may refer to this equation using `\ref{eq:label}`, e.g., see Equation **??**

4 Results

Table 1 is an example when using `knitr::kable()` to generate the table and *kableExtra* functions to modify it:

Table 1: A table produced with knitr and kableextra

	Group 5				**Group 6**	
	Group 1		Group 2		Group 3	Group 4
	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21.0	6	160	110	3.90	2.620
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875
Datsun 710	22.8	4	108	93	3.85	2.320
Hornet 4 Drive	21.4	6	258	110	3.08	3.215
Hornet Sportabout	18.7	8	360	175	3.15	3.440

Note:
Your comments go here.

4.1 Figures

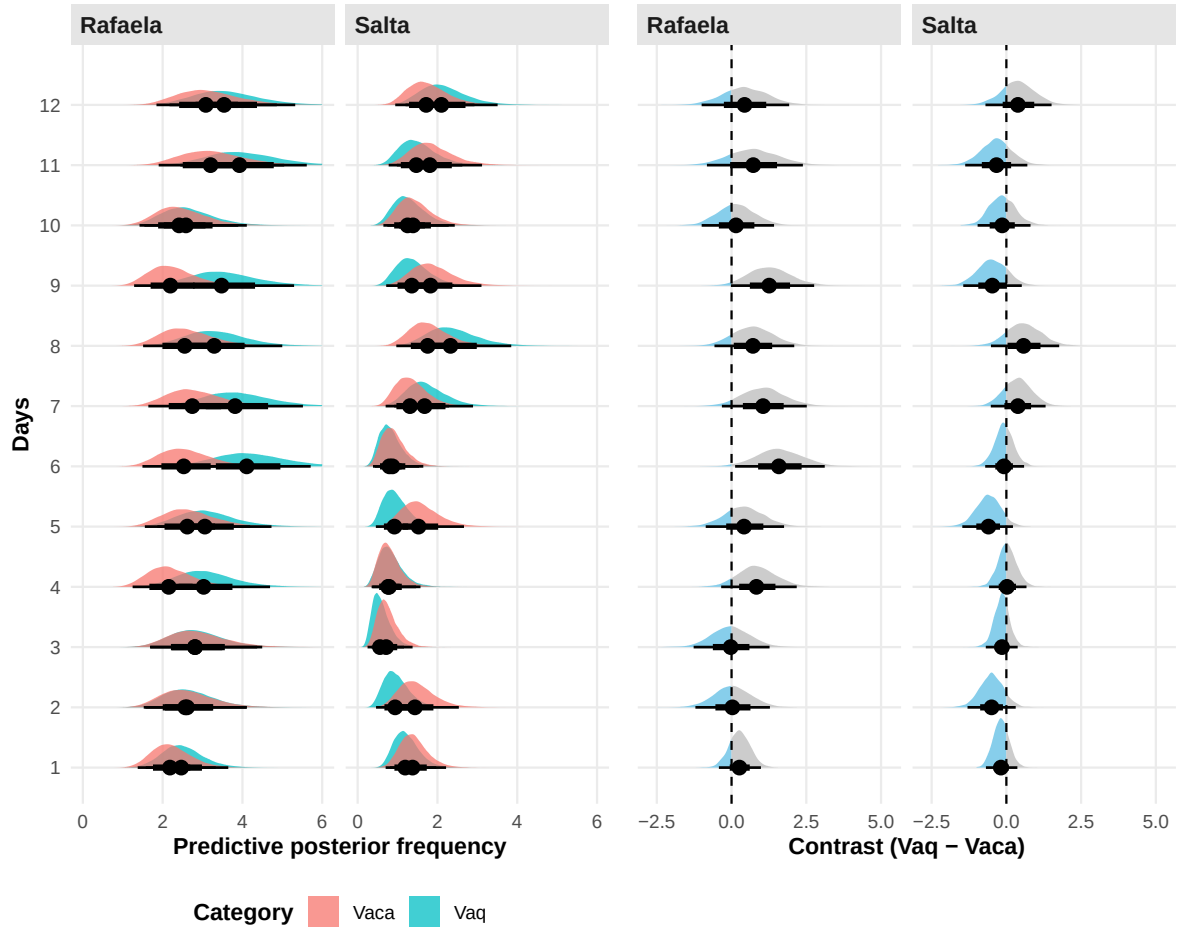

Figure 1: Comparing predictive posterior distribution of 'OCIOP' of categories across days on both localities.

5 Adding citations and bibliography

Link a `.bib` document via the YAML header and the bibliography will be printed at the very end (as usual). The default bibliography style is provided in the `bib.bst` file (do not delete), which adopts the SAGE Harvard reference style.

References can be cited directly within the document using the R Markdown equivalent of the LaTeXcitation system `[@key]`, where key is the citation key in the first line of the entry in the .bib file. Example: (Taylor and Green, 1937). To cite multiple entries, separate the keys by semicolons, e.g. (Knupp, 1999; Kamm, 2000).

There is also the package citr, which I highly recommend: *citr* provides functions and an RStudio add-in to search a BibTeX-file to create and insert formatted Markdown citations into the current document. If you are using the reference manager Zotero the add-in can access your reference database directly.

5.1 Software

If you want to include a paragraph on the software used, here is some example text/code to get the current R and package versions. The code to create a separate bibiliography file named 'packages.bib' with all package references has already been added at the beginning of this script (code chunk 'generate-package-refs').

All analyses were performed using the statistical software R (version 4.3.1) (R Core Team, 2023). This report, including tables and figures, was generated using the packages 'rmarkdown' (version 2.24) (Allaire et al., 2023), 'bookdown' (version 0.35) (Xie, 2023a), 'UHHformats' (version 1.0.0.9000) (Otto, 2022), 'knitr' (version 1.43) (Xie, 2023b), 'kableExtra' (version 1.3.4) (Zhu, 2021), 'xtable' (version 1.8.4) (Dahl et al., 2019), and 'tidyverse' (version 2.0.0) (Wickham, 2023)

6 References

Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W and Iannone R (2023) *rmarkdown: Dynamic Documents for R*. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.24.

Dahl DB, Scott D, Roosen C, Magnusson A and Swinton J (2019) *xtable: Export Tables to LaTeX or HTML*. URL http://xtable.r-forge.r-project.org/. R package version 1.8-4.

Kamm J (2000) Evaluation of the Sedov-von Neumann-Taylor blast wave solution. Technical Report Technical Report LA-UR-00-6055, Los Alamos National Laboratory.

Knupp P (1999) Winslow smoothing on two-dimensional unstructured meshes. *Eng Comput* 15: 263–268.

Otto S (2022) *UHHformats: Templates for HTML and PDF/LaTeX Output Formats Designed for the UHH*. R package version 1.0.0.9000.

R Core Team (2023) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Taylor G and Green A (1937) Mechanism of the production of small eddies from large ones. *P Roy Soc Lond A Mat* 158(895): 499–521.

Wickham H (2023) *tidyverse: Easily Install and Load the Tidyverse*. URL https://CRAN.R-project.org/package=tidyverse. R package version 2.0.0.

Xie Y (2023a) *bookdown: Authoring Books and Technical Documents with R Markdown*. URL https://CRAN.R-project.org/package=bookdown. R package version 0.35.

Xie Y (2023b) *knitr: A General-Purpose Package for Dynamic Report Generation in R*. URL https://yihui.org/knitr/. R package version 1.43.

Zhu H (2021) *kableExtra: Construct Complex Table with kable and Pipe Syntax*. URL https://CRAN.R-project.org/package=kableExtra. R package version 1.3.4.