# Bayesian correlated binary models for assessing the prevalence of viruses in organic agriculture

*Elena Lázaro*[1], *Carmen Armero*[1], *Luis Rubio*[3]

[1]Elena.Lazaro@uv.es, [1]Carmen.Armero@uv.es, Departament d'Estadística i Investigació Operativa, Universitat de València

[2]lrubio@ivia.es, Centro de Protección Vegetal y Biotecnología, Instituto Valenciano de Investigaciones Agrarias

**Abstract**

Bayesian correlated binary models were formulated to assess the prevalence of three viruses in organic and non organic agriculture. Markov chain Monte Carlo (MCMC) methods have been used to approximate the posterior distribution of the uncertainties in the model. Sensitivity analysis to prior assumptions of the random effects is examined through a measure based on the Hellinger distance, calibrated with respect to the standard normal distribution.

**Keywords:** Calibration; Hellinger distance; Random effects; Sensitivity analysis.

## 1. Introduction

Agriculture is evolving towards a sustainable productivity in conjunction with the protection of the environment and the human health [1]. Organic farming has suffered a strong development during the last decade but susceptibility to diseases caused by viruses is, in comparison to conventional agriculture, poorly studied. Research in the field of virus epidemiology under new growing conditions will probably become an important line of future research. Nevertheless, this study requires a characterization of agroecosystem balance [6] where the application of robust statistical techniques will be essential. Hierarchical Bayesian models are a very suitable choice due to its ability to capture and model a great quantity of uncertainties in a study, in particular correlation structures among the data.

## 2. Plots, viruses and data

In the summer of 2012, a total of 30 plots, 18 organic and 12 non organic, were selected in order to compare their susceptibility to virus infection. For this purpose, eight tomato or pepper plants were randomly selected in each plot to analyse the presence or absence of three relevant virus: Cucumber mosaic virus (CMV), Tomato mosaic virus (ToMV), and Tomato spotted wilt virus (TSWV). Virus presence in each plot was defined when the specific virus was detected in at least one of the eight selected plants. Data also included information of the altitude of the plots and a binary greenhouse factor for the non-organic plots.

## 3. A Bayesian correlated binary model

The main scientific question addressed in the study was to compare the probability of virus infection under organic and conventional conditions, with a special interest of detecting a possible organic effect. Agroecosystem state was kept in mind in the model through the inclusion of a set of generic covariates. Furthemore, an individual random effect which depicts plot susceptibility to be infected was also introduced to capture intra-plot variability and correlated prevalence among the different viruses

We construct a Generalized Linear Mixed Model (GLMM) for the Bernoulli random variables $Y_{ij}$ which describe the presence or absence of virus $j$ ($j = 1$ corresponds to ToMV, $j = 2$ with CMV and $j = 3$ with TSWV) in the plot $i$ with $i = 1, \dots, 30$.

$$\begin{aligned} (Y_{ij} \mid \theta_{ij}) &\sim \text{Bernoulli}(\theta_{ij}) \\ logit(\theta_{ij}) &= \boldsymbol{x}_i^T \boldsymbol{\beta}_{pj} + b_i \end{aligned} \tag{1}$$

where $\theta_{ij}$ is the probability that virus $j$ was detected in plot $i$; $\boldsymbol{\beta}_{pj}$ the regression coefficients associated to greenhouse and organic factors, and to the altitude of the plot in logaritmic scale; $b_i$ is a normal random effect for plot $i$ with zero mean and standard deviation $\sigma_b$ (or precision $\tau_b$).

The Bayesian model is completed with the specification of the prior distribution for the subsequent parameters and hyperparameters. We consider prior independence and normal distributions, $\text{N}(0, \sigma^2 = 1000)$, for the regression coefficients, and a Uniform distribution, $\text{Un}(0, 100)$, for the standard deviation of the random effect. The posterior distribution is approximated by means of MCMC through WinBUGS software [3]. The MCMC algorithm have ran for three Markov chains with 1 000 000 iterations after a burn-in period with 100 000 iterations. The effective iterations were thinned by storing every 10th iteration in order to decrease autocorrelation in the sample.

| **Variable** | **mean** | **sd** | $Q_{2.5\%}$ | $Q_{50\%}$ | $Q_{97.5\%}$ |
|---|---|---|---|---|---|
| $P(ToMV \mid org)$ | 0.125 | 0.089 | 0.012 | 0.105 | 0.345 |
| $P(ToMV \mid noorg \cap green)$ | 0.214 | 0.198 | 0.005 | 0.152 | 0.726 |
| $P(ToMV \mid noorg \cap nogreen)$ | 0.056 | 0.085 | 0.000 | 0.023 | 0.309 |
| $P(CMV \mid org)$ | 0.095 | 0.075 | 0.007 | 0.077 | 0.287 |
| $P(CMV \mid noorg \cap green)$ | 0.102 | 0.133 | 0.001 | 0.049 | 0.495 |
| $P(CMV \mid noorg \cap nogreen)$ | 0.158 | 0.152 | 0.004 | 0.109 | 0.565 |
| $P(TSWV \mid org)$ | 0.032 | 0.040 | 0.000 | 0.018 | 0.145 |
| $P(TSWV \mid noorg \cap green)$ | 0.134 | 0.162 | 0.001 | 0.070 | 0.602 |
| $P(TSWV \mid noorg \cap nogreen)$ | 0.197 | 0.175 | 0.006 | 0.146 | 0.648 |

Table 1: Descriptive of the posterior distribution of the probability that a reference plot at an altitude of 151 metres is infected with the virus CMV, ToMV, and TSWV

The estimated model shows a strong association between all the covariates regarding to the probability of infection. Table 1 shows a descriptive for the posterior distribution of the probability of infection for each viruses and reference plot populations: organic, non organic-greenhouse and non organic-non greenhouse plots for an altitude of 151 metres. Organic plots were less susceptible than non organic plots for TSWV and CMV infections. With regard to ToMV, the organic effect was weaker and the tendency changed. Altitude increase has an inverse effect in probabilty of infection as we can see in Figure 1. This effect is the same for the rest of viruses.

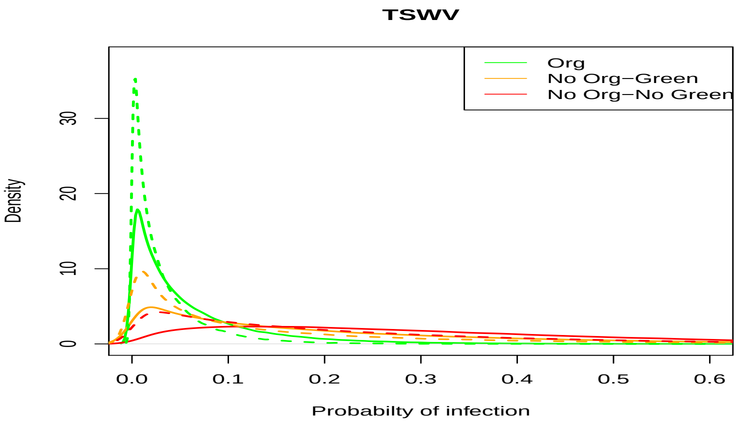

Figure 1: Posterior distribution of the probability of infection with TSWV for each reference plot with regard to two altitudes: 34 (solid) and 151 metres (dashed).

## 4. Sensitivity analysis

Prior robustness is a relevant issue in applied statistics, mainly when dealing with models with random effects. We have conducted a analysis for comparing influence of the hyperprior distribution of the random effect based on the proposals in [5]. We have considered different random effect scale parameters assumptions (see Table 1), and have quantified sensitivity to these choices comparing hyperparamter marginal posterior distribution and also the fixed effects (regression coefficients) ones. We have used a general local sensitivity measure [4]

$$S(\pi_1, \pi_2) = \frac{H(\pi_1(\gamma \mid data), \pi_2(\gamma \mid data))}{H(\pi_1(\gamma), \pi_2(\gamma))}, \tag{2}$$

where $H()$ is the Hellinger distance [2] between the posterior (prior) distributions $\pi_1(\gamma \mid data)$ and $\pi_2(\gamma \mid data)$ ($\pi_1(\gamma)$ and $\pi_2(\gamma)$). The Hellinger distance is a symmetric and invariant measure of discrepancy between two probability distributions which maximal value is 1 and is equal to 0 when both distributions are equal.

| Hyperparameter | Hyperprior distributions |
|---|---|
| $\tau_b$ | (1): $\text{Ga}(0.001, 0.001)$, (2): $\text{Ga}(0.01, 0.01)$ |
| $\sigma_b$ | (3): $\text{Unif}(0, 100)$, (4): $\text{Unif}(0, 10)$ |
| $\sigma_b$ | (5): $\text{HN}(0, 2025)$, (6): $\text{HN}(0, 25)$ |

Table 2: Hyperprior distributions for the precision and scale parameters associated to the random effects in the model. $\text{HN}(0, \sigma^2)$ represents a half-normal distribution with scale parameter $\sigma$.

Table 3 shows the Hellinger distance of the marginal posterior distribution of the regression coefficients associated with virus ToMV and sensitivity of the random effects hyperparameter.

| Parameter | $H(1,2)$ | $H(3,4)$ | $H(5,6)$ | $C(H(1,2))$ | $C(H(3,4))$ | $C(H(5,6))$ |
|---|---|---|---|---|---|---|
| $\beta_{01}$ | 0.039 | 0.01 | 0.011 | 0.109 | 0.027 | 0.032 |
| $\beta_{11}$ | 0.031 | 0.007 | 0.009 | 0.088 | 0.020 | 0.027 |
| $\beta_{21}$ | 0.02 | 0.006 | 0.008 | 0.057 | 0.017 | 0.024 |
| $\beta_{31}$ | 0.045 | 0.01 | 0.01 | 0.129 | 0.030 | 0.030 |
| Hyperparameter | $S(1,2)$ | $S(3,4)$ | $S(5,6)$ | $C(S(1,2))$ | $C(S(3,4))$ | $C(S(5,6))$ |
| $\sigma_b$, $\tau_b$ | 0.503 | 0.012 | 0.026 | 0.492 | 0.011 | 0.025 |

Table 3: Hellinger distance and its calibration between the posterior marginal distributions of the regression coefficients, and sensitivity and its calibration of the random effects precision for the different hyperpriors introduced in Table 2.

Note that prior distribution of regression coefficientes is fixed in all models due to sensitivity is quantified through the Hellinger distance. Calibration of the differences is also provided. The marginal posterior distribution for the regression coefficients are very similar in all models. However, Gamma hyperpriors always produce the greatest discrepancies. In the case of sensitivity of the random effects precision, we also obtain that Gamma and Uniform hyperpriors provide the greatest and smallest sensitivity values and calibrations, respectively.

Calibration was made with respect to the unit invariance normal distribution. For instante, a value $C(H(1,2))$= 0.192 means that the difference between marginal posterior ditributions of the paramater $\beta_{01}$ is comparable with the difference between a N$(0,1)$ and N$(0.192,1)$. In the case of the hyperparameter the value $C(S(1,2))$= 0.492 means that two priors whose difference is comparable with the difference between a N$(0,1)$ and N$(1,1)$ generate two posteriors whose difference is comparable with the difference between N$(0,1)$ and N$(0.492,1)$. These tendencies were also observed in the rest of parameters.

## 5. Acknowledgments

This research has been partially supported by research grants MTM2012-38067-C02-01 and FPU/02042/2013 from the Spanish Goverment and RTA2013-00047-C02-01 from INIA.

## 6. Bibliography

[1] Bengtsson, J., Ahnstromn J. and Weibull, A.-C. (2005). The effects of organic agriculture on biodiversity and abundance: a meta-analysis. *Journal of Applied Ecology*, 42, 261-269.

[2] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag.

[3] Lunn, D. J., Thomas, A., Best, N. and D. Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

[4] McCulloch, R. (1989). Local model influence. *Journal of the American Statistical Association*, 84(406): 473-478.

[5] Roos, M. and Held, L. (2001). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6, 2, 259-278.

[6] Serra, J., Ocón, C., Jiménez, A., Arnau, J., Malagón, J., and Porcuna, J. L. (1999). Epidemiología de las virosis en la Comunidad Valenciana: el caso del “virus de la cuchara” del tomate. *Comunidad Valenciana Agraria*, 14: 47-53.