Genomic Bayesian Prediction Model for Count Data with Genotype × Environment Interaction

Abelardo Montesinos-López[1], Osval A. Montesinos-López[2], José Crossa[2*], Juan Burgueño[2], Kent M. Eskridge[3], Esteban Falconi-Castillo[4], Xinyao He[2], Pawan Singh[2], and Karen Cichy[5]

[1] Departamento de Estadística, Centro de Investigación en Matemáticas (CIMAT), Guanajuato, Guanajuato, 36240, México.

[2] International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, México, D.F., México.

[3] University of Nebraska, Statistics Department, Lincoln, Nebraska, 68583-0963, USA.

[4] Instituto Nacional Autónomo de Investigaciones Agropecuarias (INIAP), Panamericana Sur Km 1, Quito, Ecuador.

[5] Sugar beet and Bean Research Unit, USDA-ARS, East Lansing, MI, 48824, USA.

*****Corresponding author (j.crossa@cgiar.org)

Abstract

Genomic tools allow the study of the whole genome and are facilitating the study of genotype-environment combinations and their relationship with the phenotype. However, most genomic prediction models developed so far are appropriate for Gaussian phenotypes. For this reason, appropriate genomic prediction models are needed for count data, since the conventional regression models used on count data with a large sample size (*n*) and a small number of

parameters (p) cannot be used for genomic-enabled prediction where the number of parameters (p) is larger than the sample size (n). Here we propose a Bayesian mixed negative binomial (BMNB) genomic regression model for counts that takes into account genotype by environment ($G \times E$) interaction. We also provide all the full conditional distributions to implement a Gibbs sampler. We evaluated the proposed model using a simulated data set and a real wheat data set from the International Maize and Wheat Improvement Center (CIMMYT) and collaborators. Results indicate that our BMNB model is a viable alternative for analyzing count data.

Keyword: Bayesian model; Count data; Genome enabled prediction; Gibbs sampler.

Introduction

A phenotype is the result of genotype (G), environment (E) and the genotype by environment interactions ($G \times E$) in most living organisms. Garrod (1902) observed that the effect of genes on phenotype could be modified by the environment (E). Similarly, Turesson (1922) demonstrated that the development of a plant is often influenced by its surroundings. He postulated the existence of a close relationship between crop plant varieties and their environment, and stressed that the presence of a particular variety in a given locality is not just a chance occurrence; rather, there is a genetic component that helps the individual adapt to that area.

For these reasons, today the consensus is that $G \times E$ is useful for understanding genetic heterogeneity under different environmental exposures (Kraft *et al*., 2007; Van Os and Rutten, 2009) and for identifying high-risk or productive subgroups in a population (Murcray *et al*.,

2009); it also provides insight into the biological mechanisms of complex traits such as disease resistance and yield (Thomas, 2011), and improves the ability to discover resistance genes that interact with other factors that have little marginal effects (Thomas, 2011). However, finding significant $G \times E$ interactions is challenging. Model misspecification, inconsistent definition of environmental variables, and insufficient sample sizes are just a few of the issues that often lead to low-power and non-reproducible findings in $G \times E$ studies (Jiao *et al*., 2013; Winham and Biernacka, 2013).

Genomics and its breeding applications are developing very quickly with the goal of predicting yet-to-be observed phenotypes or unobserved genetic values for complex traits and inferring the underlying genetic architecture utilizing large collections of markers (Goddard and Hayes, 2009; Zhang *et al*., 2014). Also, genomics is useful when dealing with complex traits that are multi-genic in nature and have major environmental influence (Perez-de-Castro *et al*., 2012). For these reasons, the use of whole genome prediction models continues to increase. In genomic prediction, all marker effects are fitted simultaneously on a model and simulation studies promote the use of this methodology to increase genetic progress in less time. For continuous phenotypes, models have been developed to regress phenotypes on all available markers using a linear model (Goddard and Hayes, 2009; de los Campos *et al.*, 2013). However, in plant breeding, the response variable in many traits is a count (y=0,1,2,…), for example, number of panicle per plant, number of seed per panicle, weed count per plot, etc. Count data are discrete, non-negative, integer-valued, and typically have right-skewed distributions (Yaacob *et al*., 2010).

Poisson regression and negative binomial regression are often used to deal with count data. These models have a number of advantages over an ordinary linear regression model,

including a skewed, discrete distribution (0,1,2,3,…,) and the restriction of predicted values for phenotypes to non-negative numbers (Yaacob *et al*., 2010). These models are different from an ordinary linear regression model. First, they do not assume that counts follow a normal distribution. Second, rather than modeling y as a linear function of the regression coefficients, they model a function of the response mean as a linear function of the coefficients (Cameron and Trivedi, 1986). Regression models for counts are usually nonlinear and have to take into consideration the specific properties of counts, including discreteness and non-negativity, and are often characterized by overdispersion (variance greater than the mean) (Zhou *et al*., 2012).

However, in the context of genomic selection, it is still common practice to apply linear regression models to these data or to transformed data (Montesinos-López *et al*., 2015a,b). This does not take into account that: (a) many distributions of count data are positively skewed, many observations in the data set have a value of 0, and the high number of 0's in the data set does not allow a skewed distribution to be transformed into a normal one (Yaacob *et al*., 2010); and (b) it is quite likely that the regression model will produce negative predicted values, which are theoretically impossible (Yaacob *et al*., 2010; Stroup, 2015). When transformation is used, it is not always possible to have normally distributed data and many times transformations not only do not help, they are counterproductive. There is also mounting evidence that transformations do more harm than good for the models required by the vast majority of contemporary plant and soil science researchers (Stroup, 2015). To the best of our knowledge, only the paper of Montesinos-López *et al*. (2015c) is appropriate for genomic prediction for count data under a Bayesian framework; however it does not take into account $G \times E$ interaction.

In this paper, we extend the NB regression model for counts proposed by Montesinos-López *et al*. (2015c) to take into account $G \times E$ by using a data augmentation approach. A

Gibbs sampler was derived since all full conditional distributions were obtained, which allows drawing samples from them to estimate the required parameters. In addition, we provide all the details of the efficient derived Gibbs sampler so it can be easily implemented by most plant and animal scientists. We illustrate our proposed methods with a simulated data set and a real data set on wheat Fusarium head blight. We compare our proposed models (NB and Poisson) with the Normal and Log-Normal models that are commonly implemented for analyzing count data. We also provide R code for implementing the proposed models.

Materials and Methods

The data used in this study were taken from a Ph.D. thesis (Falconi-Castillo, 2014) aimed at identifying sources of resistance to Fusarium head blight (FHB), caused by *Fusarium graminearum* and identify genomic regions and molecular markers linked to FHB resistance through association analysis.

Experimental data

Phenotypic data

A total of 297 spring wheat lines developed by the International Maize and Wheat Improvement Center (CIMMYT) was assembled and evaluated for resistance to *F. graminearum* in México over two years (2012 and 2014) and Ecuador for one year (2014). In this paper we used only 182 spring wheat lines since only for these lines we have complete marker information.

Genotypic data

DNA samples were genotyped using an Illumina 9K SNP chip with 8,632 SNPs (Cavanagh *et al*., 2013). SNP markers with unexpected genotype AB (heterozygous) were recoded as either AA or BB based on the graphical interface visualization tool of the software GenomeStudio® (Illumina). SNP markers that did not show clear clustering patterns were excluded. In addition, 66 simple sequence repeats (SSR) markers were screened. After filtering the markers for the minor allele frequency (MAF) of 0.05 and deleting markers with more than 10% of no calls, the final set of SNPs was of 1,635 SNP.

Data and software availability

The phenotypic (FHB) and genotypic (marker) data used in this study as well as basic R codes (R Core Team, 2015) for fitting the models can be directly downloaded from the repository at http://hdl.handle.net/11529/10575

Statistical Models

We used y_{ijt} to represent the count response for the tth replication of the jth line in the ith environment with $i = 1, \dots, I;\ j = 1,2, \dots, J, t = 1,2, \dots, n_{ij}$ and we propose the following linear predictor that takes into account $G \times E$:

$$\eta_{ij} = E_i + g_j + gE_{ij} \qquad (1)$$

where E_i represents the environment i, g_j is the marker effect of genotype j, and gE_{ij} is the interaction between markers and environment; $I = 3$, since we have three environments (Batan 2012, Batan 2014, and Chunchi 2014), $J = 182$, since it is the number of lines under study, and n_{ij} represents the number of replicates of each line in each environment (the minimum and maximum n_{ij} found per line were 10 and 20). The number of observations in each environment

i is $n_i = \sum_{j=1}^{J} n_{ij}$, while the total number of observations is $n = \sum_{i=1}^{I} n_i$. IJ is the product of the number of environments and number of lines. Four models were implemented using the linear predictor given in expression (1).

Model NB

Distributions: $y_{ijt}|g_j, gE_{ij} \sim \text{NB}(\mu_{ij}, r)$, with r being the scale parameter, $\mu_{ij} = \exp(\eta_{ij})$, $\boldsymbol{g} = (g_1, \ldots, g_J)^{\boldsymbol{T}} \sim N(\mathbf{0}, \boldsymbol{G_1}\sigma_g^2)$, $\boldsymbol{gE}_i = (gE_{i1}, \ldots, gE_{iJ})^{\boldsymbol{T}} \sim N(\mathbf{0}, \boldsymbol{G_2}\sigma_{gE}^2)$. Note that the NB distribution has expected value μ_{ij} and is smaller than the variance $\mu_{ij} + \frac{\mu_{ij}^2}{r}$. $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$ were assumed known, with $\boldsymbol{G_1}$ computed from marker $\boldsymbol{X}$ data (for k = 1, … , p markers) as $\boldsymbol{G_1} = \frac{\boldsymbol{XX}^T}{p}$; this matrix is called the Genomic Relationship Matrix (GRM) (VanRaden, 2008). While $\boldsymbol{G_2}$ is computed as $\boldsymbol{G_2} = \boldsymbol{I_I} \otimes \boldsymbol{G_1}$ of order IJxIJ and $\otimes$ denotes the Kronecker product, $\boldsymbol{I_I}$ means that we assume independence between environments.

Model Pois

This model is the same as **Model NB**, except that $y_{ijt}|g_j, gE_{ij} \sim \text{Poisson}(\mu_{ij})$. Since according to Zhou *et al.* (2012) and Teerapabolarn and Jaioun (2014) the $\lim_{r\to\infty} NB(\mu_{ij}, r) = Pois(\mu_{ij})$, **Model Pois** was implemented using the same method as **Model NB**, but fixing r to a large value, depending on the mean count. We used $r = 1000$, which is a good choice when the mean count is less than 100.

Model Normal

Model Normal is similar to **Model NB,** except that $y_{ijt}|g_j, gE_{ij} \sim \text{N}(\eta_{ij}, \sigma_e^2)$ with identity link

function.

Model Log-Normal

Model Log-Normal is similar to **Model NB,** except that $\log(y_{ijt}+1)\big|g_j, gE_{ij} \sim N(\eta_{ij}, \sigma_e^2)$ with identity link function.

When $p > n$, implementing **Models NB and Pois** is challenging. For this reason, we propose a Bayesian method for dealing with situations when $p > n$. The **Models Normal** and **Log-Normal** were implemented in the package BGLR of de los campos *et al.* (2014).

Bayesian mixed negative binomial regression

Rewriting the linear predictor (1) as $\eta_{ij} = \boldsymbol{x}_i^T\boldsymbol{\beta} + \sum_{h=1}^{2} b_{hij}$, with $\boldsymbol{x}_i^T = [\, x_{i1}, x_{i2}, x_{i3}]$, where x_{ik} is an indicator variable that takes the value of 1 if it is observed in environment i and 0 otherwise, for $k = 1,2,3$; $\boldsymbol{\beta}^T = \left[\, \beta_{1,}\beta_{2,}, \beta_{3,}\right]$, since three is the number of environments under study, $b_{1ij} = g_j$ and $b_{2ij} = gE_{ij}$. Note that in a sequence of independent Bernoulli (π_{ij}) trials, the random variable y_{ijt} denotes the number of successes before the r*th* failure occurs. Then

$$\Pr\left(Y_{ijt} = y_{ijt}\big|g_j, gE_{ij}\right) = \binom{y_{ijt}+r-1}{y_{ijt}}\left(1-\frac{\mu_{ij}}{r+\mu_{ij}}\right)^r\left(\frac{\mu_{ij}}{r+\mu_{ij}}\right)^{y_{ijt}} \quad \text{for } y_{ijt} = 0,1,2,\ldots$$

$$= \frac{\Gamma(y_{ijt}+r)}{y_{ijt}!\Gamma(r)}\frac{[\exp(\eta_{ij}^*)]^{y_{ijt}}}{[1+\exp(\eta_{ij}^*)]^{y_{ijt}+r}}, \quad y_{ijt} = 0,1,2,\ldots \qquad (2)$$

Since $\pi_{ij} = \frac{\mu_{ij}}{r+\mu_{ij}} = \frac{r\mu_{ij}}{1+r\mu_{ij}} = \frac{r\exp(\eta_{ij})}{1+r\exp(\eta_{ij})} = \frac{\exp(\eta_{ij}^*)}{1+\exp(\eta_{ij}^*)}$, where $\eta_{ij}^* = \boldsymbol{x}_i^T\boldsymbol{\beta}^* + \sum_{h=1}^{2} b_{hij}$, $\boldsymbol{\beta}^* = [\, \beta_1^*, \beta_2^*, \beta_3^*]$, with $\beta_i^* = \beta_i - \log(r)$ since $\boldsymbol{x}_i^T$ is composed of three indicator variables. We can rewrite (Eq 2) as:

$$\Pr\left(Y_{ijt}=y_{ijt}\middle|g_j,gE_{ij}\right) = \frac{\Gamma(y_{ijt}+r)}{y_{ijt}!\Gamma(r)}2^{-y_{ijt}-r}\exp\left(\frac{y_{ijt}-r}{2}\eta_{ij}^{*}\right)\int_0^{\infty}\exp\left[-\frac{\omega_{ijt}\left(\eta_{ij}^{*}\right)^2}{2}\right]f(\omega_{ijt},y_{ijt}+r,0)d\omega_{ijt} \quad (3)$$

Expression (3) was obtained using the equality given by Polson *et al*. (2013): $\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b}e^{\kappa\psi}\int_0^{\infty}e^{-\frac{\omega\psi^2}{2}}f(\omega;b,0)d\omega$, where $\kappa = a-b/2$ and $f(.,b,0)$ denotes the density of $PG(b,c=0)$, the PG Pólya-Gamma distribution with parameters b and $c=0$ (see Definition 1 in Polson *et al*., 2013).

From here, conditional on $\omega_{ijt}\sim PG(y_{ijt}+r,c=0)$,

$$\Pr\left(Y_{ijt}=y_{ijt}\middle|g_j,gE_{ij},\omega_{ijt}\right) = \frac{\Gamma(y_{ijt}+r)}{y_{ijt}!\Gamma(r)}2^{-y_{ijt}-r}\exp\left(\frac{y_{ijt}-r}{2}\eta_{ij}^{*}\right)\exp\left[-\omega_{ijt}\left(\eta_{ij}^{*}\right)^2/2\right] \quad (4)$$

To be able to get the full conditional distributions, we provide the prior distributions, $f(\boldsymbol{\theta})$, for all the unknown model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^{*}, \sigma_{\beta}^2, \boldsymbol{b}_1, \sigma_{b1}^2, \boldsymbol{b}_2, \sigma_{b2}^2, r)$. We assume prior independence between the parameters, that is,

$$f(\boldsymbol{\theta}) = f(\boldsymbol{\beta}^{*})f\left(\sigma_{\beta}^2\right)f(\boldsymbol{b}_1)f(\sigma_{b1}^2)f(\boldsymbol{b}_2)f(\sigma_{b2}^2)f(r).$$

We assign conditionally conjugate but weakly informative prior distributions to the parameters because we have no prior information. Prior specification in terms of $\boldsymbol{\beta}^{*}$ instead of $\boldsymbol{\beta}$ is for convenience. We adopt proper priors with known hyper-parameters whose values we specify in model implementation to guarantee proper posteriors. We assume that $\boldsymbol{\beta}^{*}\middle|\sigma_{\beta}^2\sim N_p\left(\boldsymbol{\beta}_0,\sum_0\sigma_{\beta}^2\right)$, $\sigma_{\beta}^2\sim\chi^{-2}\left(\nu_{\beta},S_{\beta}\right)$ where $\chi^{-2}\left(\nu_{\beta},S_{\beta}\right)$ denotes a scaled inverse chi-square distribution with shape ν_{β} and scale S_{β} parameters, $\boldsymbol{b}_1|\sigma_{b1}^2\sim N_{nb1}(\boldsymbol{0},\boldsymbol{G_1}\sigma_{b1}^2)$, $\sigma_{b1}^2\sim\chi^{-2}(\nu_{b1},S_{b1})$, $\boldsymbol{b}_2|\sigma_{b2}^2\sim N_{nb2}(\boldsymbol{0},\boldsymbol{G_2}\sigma_{b2}^2)$, $\sigma_{b2}^2\sim\chi^{-2}(\nu_{b2},S_{b2})$ and $r\sim G(a_0,1/b_0)$. Next we combine (Eq 4)

using all data with priors to get the full conditional distribution for parameters $\boldsymbol{\beta}^*$, σ_β^2, $\boldsymbol{b}_1$, σ_{b1}^2, $\boldsymbol{b}_2$, σ_{b2}^2 and r.

Full conditional distributions

The full conditional distribution of $\boldsymbol{\beta}^*$ is given as:

$$f(\boldsymbol{\beta}^*|\boldsymbol{y}, ELSE) \sim N\left(\widetilde{\boldsymbol{\beta}}_0, \widetilde{\boldsymbol{\Sigma}}_0\right) \tag{5}$$

where $\widetilde{\boldsymbol{\Sigma}}_0 = (\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \boldsymbol{X}^T\boldsymbol{D}_{\boldsymbol{\omega}}\boldsymbol{X})^{-1}$, $\widetilde{\boldsymbol{\beta}}_0 = \widetilde{\boldsymbol{\Sigma}}_0(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \boldsymbol{X}^T\boldsymbol{D}_\omega\sum_{h=1}^2\boldsymbol{Z}_h\boldsymbol{b}_h + \boldsymbol{X}^T\boldsymbol{\kappa})$, $\boldsymbol{y}_{ij} = \left[y_{ij1}, \dots, y_{ijn_{ij}}\right]^T$, $\boldsymbol{y}_i = \left[\boldsymbol{y}_{i1}^T, \dots, \boldsymbol{y}_{iJ}^T\right]^T$, $\boldsymbol{y} = [\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_I^T]^T$, $\boldsymbol{\kappa}_{ij} = \frac{1}{2}\left[y_{ij1} - r, \dots, y_{ijn_{ij}} - r\right]^T$, $\boldsymbol{\kappa}_i = \left[\boldsymbol{\kappa}_{i1}^T, \dots, \boldsymbol{\kappa}_{iJ}^T\right]^T$, $\boldsymbol{\kappa} = [\boldsymbol{\kappa}_1^T, \dots, \boldsymbol{\kappa}_I^T]^T$, $\boldsymbol{X}_{ij} = [\boldsymbol{1}_{n_{ij}}^T \otimes \boldsymbol{x}_i]^T$, $\boldsymbol{X}_i = [\boldsymbol{X}_{i1}^T, \dots, \boldsymbol{X}_{iJ}^T]^T$, $\boldsymbol{X} = [\boldsymbol{X}_1^T, \dots, \boldsymbol{X}_I^T]^T$, $\boldsymbol{D}_{\omega ij} = \mathrm{diag}\left(\omega_{ij1}, \dots, \omega_{ijn_{ij}}\right)$, $\boldsymbol{D}_{\omega i} = \mathrm{diag}\left(\boldsymbol{D}_{\omega i1}, \dots, \boldsymbol{D}_{\omega iJ}\right)$, $\boldsymbol{D}_\omega = \mathrm{diag}(\boldsymbol{D}_{\omega 1}, \dots, \boldsymbol{D}_{\omega I})$, $\boldsymbol{b}_{hi} = [b_{hi1}, \dots, b_{hiJ}]^T$, $\boldsymbol{b}_h = [\boldsymbol{b}_{h1}^T, \dots, \boldsymbol{b}_{hI}^T]^T$,

$$\boldsymbol{Z}_{1i} = \begin{bmatrix} \boldsymbol{1}_{n_{1i1}} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1}_{n_{1i2}} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{1}_{n_{1iJ}} \end{bmatrix}, \boldsymbol{Z}_1 = [\boldsymbol{Z}_{11}^T, \dots, \boldsymbol{Z}_{1I}^T]^T \text{ and } \boldsymbol{Z}_2 = \boldsymbol{Z}_1 * \sim \boldsymbol{X},$$

where $*\sim$ indicates the horizontal Kronecker product between $\boldsymbol{Z}_1$ and $\boldsymbol{X}$. The horizontal Kronecker product performs a Kronecker product of $\boldsymbol{Z}_1$ and $\boldsymbol{X}$ and creates a new matrix by stacking these row vectors into a matrix. $\boldsymbol{Z}_1$ and $\boldsymbol{X}$ must have the same number of rows, which is also the same number of rows in the result matrix. The number of columns in the result matrix is equal to the product of the number of columns in $\boldsymbol{Z}_1$ and $\boldsymbol{X}$. When the prior for $\boldsymbol{\beta}^* \propto$ constant, the posterior distribution of $\boldsymbol{\beta}^*$ is also normally distributed, $N\left(\widetilde{\boldsymbol{\beta}}_0, \widetilde{\boldsymbol{\Sigma}}_0\right)$, but we set the term $\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}$ to zero in both $\widetilde{\boldsymbol{\Sigma}}_0$ and $\widetilde{\boldsymbol{\beta}}$.

The fully conditional distribution of ω_{ijt} is

$$f(\omega_{ijt}|\boldsymbol{y}, ELSE) \sim PG(y_{ijt} + r, \boldsymbol{x}_i^T\, \boldsymbol{\beta}^* + \sum_{h=1}^{2} b_{hij}) \quad (6)$$

Defining $\boldsymbol{\eta}^h = \boldsymbol{X}\,\boldsymbol{\beta}^* + \boldsymbol{Z}_h\boldsymbol{b}_h$, with $h = 1,2$, the conditional distribution of $\boldsymbol{b}_h$ is given as

$$f(\boldsymbol{b}_h|\boldsymbol{y}, ELSE) \sim N\left(\widetilde{\boldsymbol{b}}_h,\ \boldsymbol{F_h}\right) \quad (7)$$

If $\boldsymbol{\eta}^1 = \boldsymbol{X}\,\boldsymbol{\beta}^* + \boldsymbol{Z}_2\boldsymbol{b}_2$, then $\boldsymbol{F_1} = \left(\sigma_{b_1}^{-2}\boldsymbol{G}_1^{-1} + \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{Z}_1\right)^{-1}$, $\widetilde{\boldsymbol{b}}_1 = \boldsymbol{F_1}(\boldsymbol{Z}_1^T\boldsymbol{\kappa} - \boldsymbol{Z}_1^T\boldsymbol{D}_\omega\boldsymbol{\eta}^1)$ and then $\boldsymbol{b}_1|\boldsymbol{y}, ELSE \sim N\left(\widetilde{\boldsymbol{b}}_1,\ \boldsymbol{F_1}\right)$. Similarly, by defining $\boldsymbol{\eta}^2 = \boldsymbol{X}\,\boldsymbol{\beta}^* + \boldsymbol{Z}_1\boldsymbol{b}_1$, we arrive at the full conditional of $\boldsymbol{b}_2$ as $\boldsymbol{b}_2|\boldsymbol{y}, ELSE \sim N\left(\widetilde{\boldsymbol{b}}_2,\ \boldsymbol{F_2}\right)$, where $\boldsymbol{F_2} = \left(\sigma_{b_2}^{-2}\boldsymbol{G}_2^{-1} + \boldsymbol{Z}_2^T D_\omega\boldsymbol{Z}_2\right)^{-1}$, $\widetilde{\boldsymbol{b}}_2 = \boldsymbol{F_2}(\boldsymbol{Z}_2^T\boldsymbol{\kappa} - \boldsymbol{Z}_2^T\boldsymbol{D}_\omega\boldsymbol{\eta}^2)$.

The fully conditional distribution of $\sigma_{b_h}^2$, for $h = 1,2$, is

$$f(\sigma_{b_h}^2|\boldsymbol{y}, ELSE) \sim \chi^{-2}(\tilde{\nu}_b = \nu_{b_h} + n_{b_h}, \tilde{S}_b = (\boldsymbol{b}_h^T\boldsymbol{G}_h^{-1}\boldsymbol{b}_h + \nu_{b_h}S_{b_h})/\nu_{b_h} + n_{b_h}) \quad (8)$$

with n_{b_1}=J and n_{b_2}=IJ.

The conditional distribution of $\sigma_{\beta^*}^2$ is

$$f(\sigma_{\beta^*}^2|\boldsymbol{y}, ELSE) \sim \chi^{-2}(\tilde{\nu}_{\beta^*} = \nu_{\beta^*} + I, \tilde{S}_\beta = [(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \nu_{\beta^*}S_{\beta^*}]/\nu_{\beta^*} + I) \quad (9)$$

Taking advantage of the fact that the NB distribution can also be generated using a Poisson representation (Quenouille, 1949) as $Y = \sum_{l=1}^{L} u_l$, where $u_l \sim Log(\pi)$, $\pi = \frac{\mu}{r+\mu}$ and is independent of $L \sim Pois(-r\log(1-\pi))$, where Log and $Pois$ denote logarithmic and Poisson distributions, respectively. Then we infer a latent count L for each $Y \sim NB(\mu, r)$ conditional on Y and r. Therefore, following Zhou *et al.* (2012), we obtain the full conditional of r by alternating

$$f(r|\boldsymbol{y}, ELSE) \sim G\left(a_0 - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}} \log\left(1-\pi_{ijt}\right), \frac{1}{b_0+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}} L_{ijt}}\right) \quad (10)$$

$$f(L_{ijt}|\boldsymbol{y}, ELSE) \sim CRT\left(y_{ijt}, r\right) \quad (11)$$

where $CRT\left(y_{ijt}, r\right)$ denotes a Chinese restaurant table (CRT) count random variable that can be generated as $L_{ijt} = \Sigma_{l=1}^{y_{ijt}} d_l$, where $d_l \sim Bernoulli\left(\frac{r}{l-1+r}\right)$. For details of the CRT random variable derivation, see Zhou and Carin (2012, 2015).

Gibbs sampler

The Gibbs sampler for the latent parameters of the NB with $G \times E$ can be implemented by sampling repeatedly from the following loop:

1. Sample ω_{ijt} values from the Pólya-Gamma distribution in (6).
2. Sample $L_{ijt} \sim CRT\left(y_{ijt}, r\right)$ from (11).
3. Sample the scale parameter (r) from the gamma distribution in (10).
4. Sample the location effects ($\boldsymbol{\beta}^*$) from the normal distribution in (5).
5. Sample the random effects $(\boldsymbol{b}_h)$ with $h = 1,2$, from the normal distribution in (7).
6. Sample the variance effects $(\sigma_{b_h}^2)$ with $h = 1,2$, from the scaled inverted χ^2 distribution in (8).
7. Sample the variance effect $(\sigma_{\beta^*}^2)$ from the scaled inverted χ^2 distribution in (9).
8. Return to step 1 or terminate when chain length is adequate to meet convergence diagnostics.

Model implementation

The Gibbs sampler described above for the BMNB model was implemented in R-Core Team (2015). Implementation was done under a Bayesian approach using Markov Chain Monte Carlo (MCMC) through the Gibbs sampler algorithm, which samples sequentially from the full conditional distribution until it reaches a stationary process, converging with the joint posterior distribution (Gelfand and Smith, 1990). To decrease the potential impact of MCMC errors on prediction accuracy, we performed a total of 60,000 iterations with a burn-in of 30,000, so that 30,000 samples were used for inference. We did not apply thinning of the chains following the suggestions of Geyer (1992), MacEachern and Berliner (1994) and Link and Eaton (2012), who provide justification of the ban on subsampling MCMC output for approximating simple features of the target distribution (e.g., means, variances, and percentiles). We implemented the prior specification given in the section Bayesian mixed negative binomial regression with $\boldsymbol{\beta}^*|\sigma^2_\beta \sim N_p(\boldsymbol{\beta}_0 = \mathbf{0}^T_3, \boldsymbol{I}_3 \times 10{,}000)$, $\boldsymbol{b}_1|\sigma^2_{b1} \sim N_{nb1}(\mathbf{0}^T_{nb1}, \boldsymbol{G}_1\sigma^2_{b1})$, where $\boldsymbol{G}_1$ is the GRM, that is, the covariance matrix of the random effects, $\sigma^2_{b1} \sim \chi^{-2}(\nu_{b1} = 3, S_{b1} = 0.001)$, $\boldsymbol{b}_2|\sigma^2_{b2} \sim N_{nb2}(\mathbf{0}^T_{nb2}, \boldsymbol{G}_2\sigma^2_{b2})$, $\boldsymbol{G}_2$ is the covariance matrix of the random effects that belong to the $G \times E$ term, $\sigma^2_{b2} \sim \chi^{-2}(\nu_{b2} = 3, S_{b2} = 0.001)$, and $r \sim G(a_0 = 0.01, 1/(b_0 = 0.01))$. All these hyper-parameters were chosen to lead weakly informative priors. The convergence of the MCMC chains was monitored using trace plots and autocorrelation functions. We also conducted a sensitivity analysis on the use of the inverse gamma priors for the variance components and we observed that the results are robust under different choices of priors.

Assessing prediction accuracy

We used cross-validation to compare the prediction accuracy of the proposed models

for count phenotypes. We implemented a 10-fold cross validation, that is, the data set was divided into 10 mutually exclusive subsets; each time we used 9 subsets for the training set and the remaining one for validation set. The training set was used to fit the model and the validation set was used to evaluate the prediction accuracy of the proposed models. To compare the prediction accuracy of the proposed models, we calculated the Spearman correlation (Cor) and the mean square error of prediction (MSEP), both calculated using the observed and predicted response variables of the validation set. Models with large absolute values of Cor indicate better prediction accuracy, while small MSEP indicate better prediction performance. The predicted observations, $\hat{y}_{ij}$, were calculated with M collected Gibbs samples after discarding those of the burn-in period. For **Models NB** and **Pois** the predicted values were calculated as $\hat{y}_{ij} = \frac{\sum_{s=1}^{M} \exp(x_{i1}\beta_1^{*(s)} + x_{i1}\beta_2^{*(s)} + x_{i1}\beta_3^{*(s)} + \log(\hat{r}^{(s)}) + \hat{g}_j{}^{(s)} + \widehat{gE}_{ij}{}^{(s)})}{S}$, where $\hat{r}^{(s)}, \beta_1^{*(s)}, \beta_2^{*(s)}$, $\beta_3^{*(s)}$ and $\hat{g}_j{}^{(s)}$ and $\widehat{gE}_{ij}{}^{(s)}$ are estimates of β_1^*, β_2^*, β_3^*, r, g_j and gE_{ij}, for line j in environment i obtained in the s*th* collected sample. For **Model Normal** as $\hat{y}_{ij} = \frac{\sum_{s=1}^{M}(x_{i1}\hat{\beta}_1^{(s)} + x_{i1}\hat{\beta}_2^{(s)} + x_{i1}\hat{\beta}_3^{(s)} + \hat{g}_j{}^{(s)} + \widehat{gE}_{ij}{}^{(s)})}{S}$ and for **Model LN** the predicted observations were calculated as $\hat{y}_{ij} = \frac{\sum_{s=1}^{M} \exp(x_{i1}\hat{\beta}_1^{(s)} + x_{i1}\hat{\beta}_2^{(s)} + x_{i1}\hat{\beta}_3^{(s)} + \hat{g}_j{}^{(s)} + \widehat{gE}_{ij}{}^{(s)} + \frac{\hat{\sigma}_e^{2(s)}}{2})}{S} - 1$, using the corresponding estimates of each model.

Simulation study

To show the performance of the proposed Gibbs sampler for count phenotypes that takes into account $G \times E$, we performed a simulation study under model (1) in two scenarios (S1 and S2). Scenario 1 had three environments ($I = 3$), 20 genotypes ($J = 20$), $\boldsymbol{G_1} = \boldsymbol{I}_{60}$, $\boldsymbol{G_2} = \boldsymbol{I_I} \otimes \boldsymbol{G_1}$ and $\sigma_{b_1}^2 = \sigma_{b_2}^2 = 0.5$, with four different numbers of replicates of each genotype in each

environment, $n_{ij} = 5, 10, 20$ and 40. Scenario 2 is equal to scenario 1, except that $\boldsymbol{G_2} = 0.7\boldsymbol{I}_{60} + 0.3\boldsymbol{J}_{60}$, where $\boldsymbol{J}_{60}$ is a square matrix of ones of order 60×60. In this second scenario, we imitated the correlation between lines of real data available in genomic selection. The priors used for the simulation study in both scenarios (S1 and S2) were approximately flat for all parameters: for $\boldsymbol{\beta}|\sigma_\beta^2 \sim N(\boldsymbol{\beta}_0^T = [0,0,0], \boldsymbol{I}_3 \times 10000)$, for $r \sim G(0.001, 1/0.001)$, for $\sigma_{b_1}^2$ and $\sigma_{b_2}^2$ a $\sim \chi^{-2}(0.50002, 4.0002)$, while for $\boldsymbol{b}_1|\sigma_{b1}^2 \sim N(\boldsymbol{0}, \boldsymbol{G_1})$, and for $\boldsymbol{b}_2|\sigma_{b2}^2 \sim N(\boldsymbol{0}, \boldsymbol{G_2})$. We computed 20,000 MCMC samples; Bayes estimates were computed with 10,000 samples since the first 10,000 were discarded as burning. We report average estimates obtained by using the proposed Gibbs sampler along with standard deviations (SD) (Table 1). All the results in Table 1 are based on 50 replications.

Results

Given in Table 1 are the results of the simulation study in both scenarios (S1 and S2). The bias when estimating the parameters is a little larger in S1 compared to S2. Also, parameter β_0 is the parameter with larger bias (underestimated). Both variances (σ_1^2, σ_2^2) are overestimated in scenario 1, but only σ_1^2 is overestimated in scenario 2. Also, with a sample size of $n_{ij} = 5$, parameter r had a larger SD; however, for larger sample sizes ($n_{ij} = 20,40$), the SD were considerably reduced. In general, there was not a large reduction in SD when the sample size increased from 5 to 10, 20 and 40, the exception being the estimation of r in both scenarios and the estimation of β_0 in scenario 1, where there was a large reduction in SD when the sample size increased. Although estimations do not totally agree with the true values of the parameters, the proposed Gibbs sampler for count data that takes into account $G \times E$ did a good job of

estimating the parameters, since the estimates are close to the true values with a SD of reasonable size.

Using the real data set, we compared four scenarios (given in Table 2) for each model. Table 2 shows that in the linear predictor, scenarios 1 and 2 do not take into account interaction effects, only main effects. Also, scenarios 1 and 3 do not use marker information. These four scenarios were studied to investigate the gain in model fit and prediction ability taking into account the interaction effects and using the marker information available.

The posterior means (Mean), posterior standard deviation (SD) of the scalar parameters, and posterior predictive checks for each scenario of the proposed models are given in Table 3. For the four models, the posterior means of the beta regression coefficients, variance components, and over-dispersion parameters (r) are similar between scenarios 1 and 2 and between scenarios 3 and 4. In terms of goodness of fit measured by the loglikelihood posterior mean (loglink), the scenarios rank as follows: scenario 3, rank 1; scenario 4, rank 2; scenario 1, rank 3; and scenario 2, rank 4, for the four proposed models, with the exception of **Model Pois** where the ranking was scenario 3, rank 1; scenario 4, rank 2; scenario 2, rank 3; and scenario 1, rank 4. Therefore, there is evidence that with the four proposed models in terms of goodness of fit, the best scenario is S3. Of the four models under study, Table 3 shows that **Model LN** reports the best fit since it has the largest Loglik.

In Table 4 we present the mean and standard deviation of the posterior predictive checks (Cor and MSEP) for each location (Batan 2012, Batan 2014 and Chunchi 2014) resulting from the 10-fold cross-validation implemented for the four models and four scenarios. The predictive checks given in Table 4 were calculated using the testing set. In **Model NB**, according to the Spearman Correlation, the ranking of scenarios was as follows: in Batan 2012 and Batan 2014,

1 for scenario 4, 2 for scenario 3, 3 for scenario 1, and 4 for scenario 2. In Chunchi 2014, the ranking was 1 for scenario 3, 2 for scenario 2, 3 for scenario 4, and 4 for scenario 4. With the MSEP, the ranking for **Model NB** in Batan 2012 was 1 for scenario 3, 2 for scenario 4, 3 for scenario 1, and 4 for scenario 2. In Batan 2014, the ranking was 1 for scenario 2, 2 for scenario 1, 3 for scenario 3, and 4 for scenario 4. In Chunchi 2014, the ranking in terms of MSEP was 1 for scenario 3, 2 for scenario 2, 3 for scenario 4, and 4 for scenario 1. Under **Model Pois,** the ranking of the 4 scenarios in each locality was exactly the same as the ranking reported for **Model NB**. For **Model Normal** in terms of the Spearman correlation, scenario 1 was the best in prediction accuracy in Batan 2012 and Chunchi 2014, while scenario 4 was the worst in all three locations. In terms of MSEP, the best scenario was 3 in Batan 2014 and Chunchi 2014, and the worst was scenario 4 in Batan 2014 and Chunchi 2014. For **Model LN** in terms of the Spearman correlation, the best scenarios were scenarios 1 and 2, and the worst was scenario 3 in Batan 2012. In Batan 2014, the best scenario was 1, then scenario 3 and the worst was scenario 4. In Chunchi 2014, the best scenario was scenario 3, then scenario 2 and the worst was scenario 2. In terms of MSEP for Batan 2012, the best scenario was 3, then scenario 1 and the worst was scenario 4. In Batan 2014, the best scenario was 1, then 2 and the worst was scenario 4. Finally, in Chunchi 2014, the best scenario was 3, then 2 and the worst was scenario 1.

Table 5 gives the average of the ranks of the two posterior predictive checks (Cor and MSEP) that were used. Since we are comparing four scenarios for each model, the values of the ranks range from 1 to 4, and the lower the values, the better the scenario. For ties we assigned the average of the ranges that would have been assigned had there been no ties. Table 5 shows that the best scenarios were scenarios 3 and 4 under **Model NB** and **Pois** in Batan 2012. In Batan 2014 under **Models NB** and **Pois,** the best scenario was 3, while in Chunchi 2014, the

best scenarios were 3 and 1**.** Under **Model Normal**, the best scenario was scenario 3 in Batan 2014 and Chuchi 2014, while in Batan 2012, the best scenarios were 2 and 3. Finally, under **Model LN,** the best scenario was 3 in Chunchi 2014, and scenario 1 in Batan 2012 and Batan 2014.

Results in Tables 4 and 5 indicate that the best models in terms of prediction accuracy are **Models NB** and **Pois**, since they had better predictions in the validation set based on both the posterior predictive checks (Cor and MSEP) implemented, although in terms of goodness of fit, **Model LN** was the best. These results are in partial agreement with the findings of Montesinos-Lopez *et al.* (2015), who came to the conclusion that **Models NB** and **Pois** are good alternatives for modeling count data, although in this study, the best predictions were produced by **Model LN**. However, this model did not take into account the $G \times E$ interaction.

Discussion

Developing specific methods for count data for genome-enabled prediction can help to improve the selection of candidate genotypes early in time when the phenotypes are counts. However, currently in genomic selection, phenotypic data (dependent variable) are not taken into account before deciding on the modeling approach to be used, mainly due to the lack of genome-enabled prediction models for non-normal phenotypes. The Bayesian regression models proposed in this paper aim to fill this lack of genome-enabled prediction models for non-normal data.

The first advantage of our proposed methods for count data is that they take into account the nonlinear relationship between responses and consider the specific properties of counts, including discreteness, non-negativity, and over-dispersion (variance greater than the mean);

this guarantees that the predictive response will not be negative, which makes no sense for count data. In addition, our methods take into account $G \times E$, which plays a central role when selecting candidates genotypes in plant breeding.

Another advantage of our proposed method is that the proposed Gibbs sampler has an analytical solution since we were able to obtain all the full conditional distributions required analytically. This was possible because we constructed our Gibbs sampler using the data augmentation approach proposed by Polson *et al*., (2013) for count data. For this reason, we believe it is an attractive alternative for fitting complex multilevel data for counts because, in addition to its simplicity, it can generate samples from a high dimensional probability distribution.

Our proposed methods showed superior performance in terms of prediction accuracy compared to **Models Normal** and **Log-Normal**. Also, we observed that in **Models NB** and **Pois** taking into account the $G \times E$ increase considerable the prediction accuracy which is expected since there is enough scientific evidence that including the $G \times E$ interaction improve prediction accuracy. Finally, more research is needed to study the proposed methods using real data sets and to extend the proposed genomic-enabled prediction models to deal with so many zeros in count response variables and for modeling multiple traits.

Acknowledgments

We very much appreciate CIMMYT field collaborator, laboratory assistants, and technicians who collected the phenotypic and genotypic data used in this study.

References

Cameron, A.C., and P.K. Trivedi. (1986). Econometric models based on count data. Comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1(1): 29-53.

Cavanagh, C.R., Chao, S., Wang, S. *et al.* (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proceedings of the National Academy of Sciences. 110(20) 8057-8062.

de los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genetics 9: e1003608.

de los Campos G, Pataki A, Pérez P (2014). The BGLR (Bayesian Generalized Linear Regression) R-Package. [http://bglr.r-forge.r-project.org/BGLR-tutorial.pdf

Falconi-Castillo, E. (2014). Association mapping for detecting QTLs for Fusarium head blight and Yellow rust resistance in bread wheat. Ph.D. Dissertation. Michigan State University, East Lansing, Michigan, USA.

Garrod, A.E. (1902). The incidence of alkatonuria: a study in chemical individuality. Lancet 160:1616-1620.

Gelfand, A.E., and A.F. Smith. (1990). Sampling-based approaches to calculating marginal densities. J. Am. Statist. Assoc. 85: 398-409.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo. Stat Sci 473-483.

Goddard, M.E., and B.J. Hayes. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10: 381-391.

Jiao, S, L. Hsu, S. Bézieau, H. Brenner, A.T. Chan, *et al*. (2013). SBERIA: Set Based Gene-Environment Interaction test for rare and common variants in complex diseases. Genet. Epidemiol. 37:452-64.

Kraft, P., Y.C. Yen, D.O. Stram, J. Morrison, and W.J. Gauderman. (2007). Exploiting gene environment interaction to detect genetic associations. Hum. Hered. 63: 111-119.

Link, W.A., and M.J. Eaton (2012). On thinning of chains in MCMC. Methods Ecol Evol 3: 112-115.

MacEachern, S.N., and L.M. Berliner. (1994). Subsampling the Gibbs sampler. Am. Stat. 48:188-190.

Montesinos-López, O.A., A. Montesinos-López, P. Pérez-Rodríguez, G. de los Campos, K.M. Eskridge *et al*. (2015a). Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. G3|Genes|Genomes|Genetics 5: 1-10.

Montesinos-López, O.A., A. Montesinos-López, J. Crossa, J. Burgueño, and K.M. Eskridge. (2015b). Genomic-enabled prediction of ordinal data with Bayesian logistic ordinal regression. G3|Genes|Genomes|Genetics 5: 2113-2126.

Montesinos-López, O.A., A. Montesinos-López, P. Pérez-Rodríguez, K.M. Eskridge, X. He, P. Juliana, P. Singh, and J. Crossa. (2015c). Genomic prediction models for count data. Journal of Agricultural, Biological, and Environmental Statistics (JABES). 20: 533–554, DOI: 10.1007/s13253-015-0223-4.

Murcray, C.E., J.P. Lewinger, and W.J. Gauderman. (2009). Gene-environment interaction in genome-wide association studies. Am. J. Epidemiol. 169: 219-226.

Pérez-de-Castro, A.M., S. Vilanova, J. Cañizares, L. Pascual, J.M. Blanca, M.J. Diez, and B. Picó. (2012). Application of genomic tools in plant breeding. Current Genomics 13(3): 179.

Polson, N.G., J.G. Scott, and J. Windle. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. J. Am. Statist. Assoc. 108: 1339-1349.

Quenouille, M.H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. Biometrics 5: 162-164.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Stroup, W.W. (2015). Rethinking the analysis of aon-Normal data in alant and soil science. Agron. J. 107: 811-827.

Teerapabolarn, K., and K. Jaioun. (2014). An improved Poisson approximation for the Negative binomial bistribution. Applied Mathematical Sciences 8(89): 4441-4445.

Thomas, D. (2011). Response to 'Gene-by-environment experiments: a new approach to finding the missing heritability' by Van Ijzendoorn *et al*. Nature Rev. Genet. 12: 881.

Turesson, G. (1922). The genotypical response of the plant species to the habitat. Hereditas 3: 211-350.

Van Os, J., and B. Rutten (2009). Gene-environment-wide interaction studies in psychiatry. Am J Psychiatry 166: 964-966.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414-4423.

Windle, J., C.M. Carvalho, J.G. Scott, L. Sun. (2013). Polya-Gamma Data Augmentation for Dynamic Models. arXiv preprint arXiv:1308.0774.

Winham, S.J., and J.M. Biernacka. (2013). Gene–environment interactions in genome-wide association studies: current approaches and new directions. J. Child Psychol. Psychiatry 54: 1120-1134.

Yaacob, W.F.W., M.A. Lazim, and Y.B. Wah. (2010). A practical approach in modelling count data. In Proceedings of the Regional Conference on Statistical Sciences (pp. 176-183).

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao, *et al*. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PloS One 9: e93017.

Zhou, M., L. Li, D. Dunson, and L. Carin. (2012). Lognormal and gamma mixed negative binomial regression. In Machine Learning: Proceedings of the International Conference on Machine Learning (vol. 2012. p. 1343). NIH Public Access.

Zhou, M., and L. Carin. (2012). Augment-and-conquer negative binomial processes. In Advances in Neural Information Processing Systems (pp. 2546-2554).

Zhou, M., and L. Carin. (2015). Negative binomial process count and mixture modeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 37(2): 307-320.

Appendix A

Derivation of full conditional distribution for all parameters.

Full conditional for β^*

$$f(\boldsymbol{\beta}^*|\mathbf{y}, \text{ELSE}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{n_{ij}} \Pr\left(Y_{ijt} = y_{ijt}\middle|\mathbf{x}_i^T, r, \omega_{ijt}, b_{1i}, b_{2ij}\right) f(\boldsymbol{\beta}^*)$$

$$\propto \exp\left(\boldsymbol{\kappa}^T\mathbf{X}\,\boldsymbol{\beta}^* + \boldsymbol{\kappa}^T\sum_{h=1}^{2}\mathbf{Z}_h\mathbf{b}_h - \frac{1}{2}\left(\mathbf{X}\,\boldsymbol{\beta}^* + \sum_{h=1}^{2}\mathbf{Z}_h\mathbf{b}_h\right)^T\mathbf{D}_\omega\left(\mathbf{X}\,\boldsymbol{\beta}^* + \sum_{h=1}^{2}\mathbf{Z}_h\mathbf{b}_h\right)\right.$$

$$\left. - \frac{1}{2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)\right)$$

$$\propto \exp\left(-\frac{1}{2}[\boldsymbol{\beta}^{*T}\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \mathbf{X}^T\mathbf{D}_\omega\mathbf{X}\right)\boldsymbol{\beta}^* - 2\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \mathbf{X}^T\mathbf{D}_\omega\sum_{h=1}^{2}\mathbf{Z}_h\mathbf{b}_h + \mathbf{X}^T\boldsymbol{\kappa}\right)^T\boldsymbol{\beta}^*]\right)$$

$$\propto \exp\left(-\frac{1}{2}[(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}_0)^T\tilde{\boldsymbol{\Sigma}}_0^{-1}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}_0)]\right) \propto N(\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\Sigma}}_0)$$

where $\tilde{\boldsymbol{\Sigma}}_0 = (\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2} + \mathbf{X}^T\mathbf{D}_\omega\mathbf{X})^{-1}$, $\tilde{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\Sigma}}_0\left(\boldsymbol{\Sigma}_0^{-1}\sigma_\beta^{-2}\boldsymbol{\beta}_0 - \mathbf{X}^T\mathbf{D}_\omega\sum_{h=1}^{2}\mathbf{Z}_h\mathbf{b}_h + \mathbf{X}^T\boldsymbol{\kappa}\right)$.

Full conditional for ω_{ijt}

$$f\left(\omega_{ijt}\middle|\boldsymbol{y}, \text{ELSE}\right) \propto \exp\left[-\frac{\omega_{ijt}\left(\mathbf{x}_i^T\,\boldsymbol{\beta}^* + \sum_{h=1}^{2} b_{hij}\right)^2}{2}\right] f(\omega_{ijt}; y_{ijt} + r, 0)$$

$$\propto \exp\left[-\frac{\omega_{ijt}\left(\mathbf{x}_i^T\,\boldsymbol{\beta}^* + \sum_{h=1}^{2} b_{hij}\right)^2}{2}\right] f(\omega_{ijt}; y_{ijt} + r, 0) \;\propto\; PG(y_{ijt} + r, \mathbf{x}_i^T\,\boldsymbol{\beta}^* + \sum_{h=1}^{2} b_{hij})$$

Full conditional for $\mathbf{b}_1$

Defining $\boldsymbol{\eta}^1 = \mathbf{X}\,\boldsymbol{\beta}^* + \mathbf{Z}_2\mathbf{b}_2$ the conditional distribution of $\mathbf{b}_1$ is given as

$$f(\mathbf{b}_1|\boldsymbol{y}, \text{ELSE}) \propto \exp\left(\boldsymbol{\kappa}^{\mathrm{T}}\mathbf{Z}_1\mathbf{b}_1 - \frac{1}{2}(\mathbf{Z}_1\mathbf{b}_1 + \boldsymbol{\eta}^1)^{\mathrm{T}}\mathbf{D}_\omega(\mathbf{Z}_1\mathbf{b}_1 + \boldsymbol{\eta}^1)\right) f(\mathbf{b}_1|\sigma^2_{b_1})$$

$$\propto \exp\left\{-\frac{1}{2}\left[\mathbf{b}_1^{\mathrm{T}}(\sigma_{b_1}^{-2}\mathbf{G}_1^{-1} + \mathbf{Z}_1^{\mathrm{T}}\mathbf{D}_\omega\mathbf{Z}_1)\mathbf{u} - 2\,(\mathbf{Z}_1^{\mathrm{T}}\boldsymbol{\kappa} - \mathbf{Z}_1^{\mathrm{T}}\mathbf{D}_\omega\boldsymbol{\eta}^1)^{\mathrm{T}}\,\mathbf{b}_1\right]\right\}$$

$$\propto \exp\left\{-\tfrac{1}{2}\,(\mathbf{b}_1 - \tilde{\mathbf{b}}_1)^{\mathrm{T}}\mathbf{F}_1^{-1}(\mathbf{b}_1 - \tilde{\mathbf{b}}_1)\right\} \propto f(\mathbf{b}_1|\text{ELSE}) \sim N(\tilde{\mathbf{b}}_1,\ \mathbf{F_1})$$

where $\mathbf{F_1} = (\sigma_{b_1}^{-2}\mathbf{G}_1^{-1} + \mathbf{Z}_1^{\mathrm{T}}\mathbf{D}_\omega\mathbf{Z}_1)^{-1}$ and $\tilde{\mathbf{b}}_1 = \mathbf{F_1}(\mathbf{Z}_1^{\mathrm{T}}\boldsymbol{\kappa} - \mathbf{Z}_1^{\mathrm{T}}\mathbf{D}_\omega\boldsymbol{\eta}^1)$.

Full conditional for $\sigma^2_{b_h}$

$$f(\sigma^2_{b_h}|\boldsymbol{y}, \mathbf{b}) \propto \frac{1}{(\sigma^2_{b_h})^{\frac{\nu_{b_h}+n_{b_h}}{2}+1}} \exp\left(-\frac{\mathbf{b}_h^{\mathrm{T}}\mathbf{G}_h^{-1}\mathbf{b}_h + \nu_{b_h}S_{b_h}}{2\sigma^2_{b_h}}\right)$$

$$\propto \chi^{-2}(\tilde{\nu}_b = \nu_{b_h} + n_{b_h}, \tilde{S}_b = (\mathbf{b}_h^{\mathrm{T}}\mathbf{G}_h^{-1}\mathbf{b}_h + \nu_{b_h}S_{b_h})/\nu_{b_h} + n_{b_h})$$

with n_{b_1}=J and n_{b_2}=IJ.

Full conditional for σ^2_β

$$f(\sigma^2_{\beta^*}|\boldsymbol{y}, \text{ELSE}) \propto \frac{1}{(\sigma^2_{\beta^*})^{\frac{\nu_{\beta^*}+3}{2}+1}} \exp\left(-\frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \nu_{\beta^*}S_{\beta^*}}{2\sigma^2_{\beta^*}}\right)$$

$$\propto \chi^{-2}(\tilde{\nu}_{\beta^*} = \nu_{\beta^*} + I, \tilde{S}_\beta = [(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \nu_{\beta^*}S_{\beta^*}]/\nu_{\beta^*} + I)$$

Full conditional for r

To make the inference of r, we first place a gamma prior on it as $r \sim G(a_0, 1/b_0)$. Then we infer a latent count L for each count conditional on Y and r. To derive the full conditional of r, we use the following parameterization of the NB distribution: $Y \sim NB(\pi, r)$ with $\pi = \frac{\mu}{r+\mu}$. Since $L \sim \text{Pois}(-r\log(1-\pi))$, by construction we can use the Gamma-Poisson conjugacy to update r. Therefore,

$$\mathrm{f}(\mathrm{r}|\boldsymbol{y}, \mathrm{ELSE}) \propto f(r)\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t}^{n_{ij}} f\left(y_{ijt}\middle|L_{ijt}\right)f(L_{ijt})$$

$$\propto r^{a_0-1}\exp(-rb_0)\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t}^{n_{ij}}(-r\log(1-\pi_{ij}))^{L_{ijt}}\exp(r\ \log(1-\pi_{ij}))$$

$$\propto r^{a_0+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}}L_{ijt}-1}\exp[-(b_0-\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}}\log\left(1-\pi_{ij}\right)r$$

$$\propto G(a_0-\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}}\log\left(1-\pi_{ij}\right),\frac{1}{b_0+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{n_{ij}}L_{ijt}}) \qquad \text{(A5)}$$

According to Zhou *et al*. (2012), the conditional posterior distribution of L_{ijt} is a Chinese restaurant table (CRT) count random variable. That is, $L_{ijt}\sim CRT\left(y_{ijt},r\right)$ and we can sample it as $L_{ijt}=\Sigma_{l=1}^{y_{ijt}}d_l$, where $d_l\sim Bernoulli\left(\frac{r}{l-1+r}\right)$. For details of the CRT random variable derivation, see Zhou and Carin (2012, 2015).

Table 1. Posterior mean (Mean) and posterior standard deviation (SD) of the Bayesian method with four sample sizes (n_{ij}) for Model NB. S denotes scenario.

			$n_{ij}=5$		$n_{ij}=10$		$n_{ij}=20$		$n_{ij}=40$	
S	Parameter	True	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	β_0	1.5	1.484	0.357	1.488	0.269	1.542	0.233	1.549	0.213
	β_1	-1	-0.981	0.256	-0.994	0.247	-1.075	0.250	-1.016	0.190
	β_2	1	0.997	0.270	0.985	0.223	0.994	0.268	0.949	0.223
	r	5	5.079	0.916	5.078	0.519	5.017	0.471	5.027	0.330
	σ_1^2	0.5	0.542	0.196	0.594	0.176	0.582	0.180	0.590	0.216
	σ_2^2	0.5	0.503	0.134	0.524	0.136	0.531	0.110	0.512	0.114
2	β_0	1.5	1.4808	0.5009	1.4596	0.5041	1.5611	0.6108	1.4723	0.4979
	β_1	-1	-1.0631	0.2348	-0.9975	0.2040	-1.008	0.2226	-1.025	0.1908
	β_2	1	0.9504	0.2356	1.0294	0.2167	0.9925	0.1954	0.9685	0.2018
	r	5	5.1030	0.8060	4.9901	0.5928	5.0367	0.3485	5.0275	0.2033
	σ_1^2	0.5	0.5422	0.1827	0.5650	0.2199	0.5785	0.1872	0.5296	0.1837
	σ_2^2	0.5	0.4987	0.1155	0.5084	0.1423	0.5302	0.1301	0.5123	0.1047

Table 2. Scenarios proposed to fit the real data set with **Models NB**, **Pois, Normal and LN**. E stands for Environment, L for lines, G for lines taking into account markers; EL and EG are interaction effects of E and L and E and G.

Scenario	Main effects			Interaction effects	
	E	L	G	EL	EG
1	X	X			
2	X		X		
3	X	X		X	
4	X		X		X

Table 3. Estimated beta coefficients, variance components, and posterior predictive checks for the four scenarios (S1, S2, S3, S4) for each proposed model (**Model NB, Model Pois, Model Normal and Model LN)**. Mean stands for posterior mean and SD for posterior standard deviation.

			Model NB					
	S1		S2		S3		S4	
Parameter	Mean	SD	Mean	SD	Mean	SD	Mean	SD
β_1^*	-0.933	0.600	-1.046	0.611	-2.521	0.711	-2.383	0.992
β_2^*	-0.826	0.710	-1.158	0.661	-2.273	0.577	-2.725	1.001
β_3^*	-0.026	0.480	-0.152	0.564	-1.688	0.851	-1.961	0.777
σ_1^2	0.425	0.048	1.374	0.167	0.341	0.050	1.033	0.153
σ_2^2					0.376	0.031	1.035	0.096
r	2.802	0.117	2.813	0.116	11.866	1.115	11.549	1.170
Loglik	-1526.649		-1526.882		-1268.827		-1275.253	
Cor	0.694		0.694		0.899		0.891	
MSEP	2.130		2.116		0.750		0.767	
			Model Pois					
β_1^*	-7.135	0.217	-7.211	0.388	-6.693	0.111	-6.802	0.327
β_2^*	-7.075	0.132	-7.166	0.108	-7.072	0.161	-7.266	0.185
β_3^*	-5.969	0.431	-6.463	0.293	-5.879	0.163	-6.658	0.276
σ_1^2	0.443	0.049	1.457	0.172	0.348	0.047	1.027	0.144
σ_2^2	-	-	-	-	0.381	0.031	1.045	
r	1000		1000		1000		1000	
Loglik	-1477.634		-1477.52		-1228.73		-1234.973	
Cor	0.662		0.662		0.899		0.891	
MSEP	1.866		1.860		0.743		0.758	
			Model Normal					
β_1	-12.3	5.86	7.9	4.36	13.7	3.69	9.22	3.11
β_2	-12.2	5.8	7.93	4.41	13.6	3.73	9.11	3.16
β_3	-10.4	5.87	9.66	4.36	15.5	3.69	10.94	3.1
σ_1^2	0.957	0.161	1.42	0.345	0.722	0.175	1.58	0.403
σ_2^2	-	-	-	-	1.33	0.182	1.13	0.343
r	2.75	0.136	2.91	0.147	1.67	0.109	2.23	0.172
Loglik	-1918		-1957		-1542		-1747	
Cor	0.595		0.557		0.831		0.705	
MSEP	2.405		2.600		1.073		1.679	
			Model LN					
β_1	-3.950	0.505	-6.340	3.330	1.410	0.481	3.320	1.310
β_2	-3.950	0.483	-6.330	3.320	1.410	0.487	3.320	1.290
β_3	-3.510	0.485	-5.850	3.330	1.860	0.494	3.790	1.310

σ_1^2	0.085	0.013	0.146	0.028	0.069	0.013	0.157	0.030
σ_2^2	-	-	-	-	0.081	0.011	0.053	0.018
r	0.172	0.009	0.181	0.009	0.107	0.007	0.147	0.011
Loglik	-484		-518		-125		-354	
Cor	0.709		0.679		0.882		0.789	
MSEP	2.500		2.626		1.254		1.974	

Table 4. Estimated posterior predictive checks with cross validation for **Models NB**, **Pois, Normal and LN**. () denotes the ranking of the four scenarios for each posterior predictive check. Each average was obtained as the mean of the rankings of the four posterior predictive checks for each scenario.

		Batan 2012		Batan 2014		Chunchi 2014	
		Model NB					
Scenario		Cor	MSEP	Cor	MSEP	Cor	MSEP
S1	Mean	0.426 (3)	0.977 (3)	0.427 (4)	1.388 (2)	0.182 (3)	11.733 (4)
	SD	0.331	0.723	0.327	1.351	0.401	9.471
S2	Mean	0.423 (4)	0.980 (4)	0.432 (3)	1.383 (1)	0.204 (2)	11.222 (2)
	SD	0.327	0.717	0.325	1.356	0.373	8.614
S3	Mean	0.539 (2)	0.497 (1)	0.522 (2)	1.480 (3)	0.224 (1)	8.645 (1)
	SD	0.283	0.376	0.292	2.318	0.386	5.688
S4	Mean	0.557 (1)	0.607 (2)	0.564 (1)	1.850 (4)	0.122 (4)	11.343 (3)
	SD	0.243	0.438	0.222	2.684	0.407	8.154
		Model Pois					
S1	Mean	0.426 (3)	0.977 (3)	0.427 (4)	1.388 (2)	0.182 (3)	11.733 (4)
	SD	0.331	0.723	0.327	1.351	0.401	9.471
S2	Mean	0.423 (4)	0.980 (4)	0.432 (3)	1.383 (1)	0.204 (2)	11.222 (2)
	SD	0.327	0.717	0.325	1.356	0.373	8.614
S3	Mean	0.539 (2)	0.497 (1)	0.522 (2)	1.480 (3)	0.224 (1)	8.645 (1)
	SD	0.283	0.376	0.292	2.318	0.386	5.688
S4	Mean	0.557 (1)	0.607 (2)	0.564 (1)	1.850 (4)	0.122 (4)	11.343 (3)
	SD	0.243	0.438	0.222	2.684	0.407	8.154
		Model Normal					
S1	Mean	0.358 (1)	1.096 (4)	0.367 (2)	1.788 (1)	0.148 (1)	7.425 (2)
	SD	0.280	0.883	0.397	1.701	0.318	4.151
S2	Mean	0.344 (2)	0.988 (2)	0.334 (3)	2.010 (3)	0.074 (3)	7.454 (3)
	SD	0.326	0.652	0.440	2.462	0.330	4.339
S3	Mean	0.330 (3)	0.806 (1)	0.371 (1)	1.963 (2)	0.146 (2)	7.318 (1)
	SD	0.300	0.495	0.400	2.986	0.287	4.159
S4	Mean	0.267 (4)	1.029 (3)	0.237 (4)	2.373 (4)	0.039 (4)	8.482 (4)
	SD	0.338	0.731	0.445	3.420	0.238	4.326
		Model LN					
S1	Mean	0.510 (1.5)	0.661 (2)	0.455 (1)	1.601 (1)	0.149 (2)	8.099 (4)
	SD	0.208	0.419	0.307	2.348	0.379	5.113
S2	Mean	0.510 (1.5)	0.663 (3)	0.433 (3)	1.778 (2)	0.086 (4)	7.819 (2)
	SD	0.224	0.392	0.353	2.820	0.459	5.311
S3	Mean	0.505 (3)	0.639 (1)	0.449 (2)	1.871 (3)	0.153 (1)	7.759 (1)
	SD	0.208	0.451	0.313	3.162	0.371	5.209
S4	Mean	0.428 (4)	0.721 (4)	0.427 (4)	1.951 (4)	0.087 (3)	8.038 (3)
	SD	0.246	0.415	0.327	3.148	0.413	5.187

Table 5. Rank averages for the four scenarios for each model (**Models NB**, **Pois, Normal and LN**) resulting from the 10-fold cross-validation implemented. Each average was obtained as the mean of the rankings given in Table 4 for the two posterior predictive checks (Cor and MSEP) in each scenario.

Scenario	Batan 2012	Batan 2014	Chunchi 2014	Batan 2012	Batan 2014	Chunchi 2014
	Model NB			**Model Normal**		
S1	3	3	3.5	2.5	1.5	1.5
S2	4	2	2	2	3	3
S3	1.5	2.5	1	2	1.5	1.5
S4	1.5	2.5	3.5	3.5	4	4
	Model Pois			**Model LN**		
S1	3	3	3.5	1.75	1	3
S2	4	2	2	2.25	2.5	3
S3	1.5	2.5	1	2	2.5	1
S4	1.5	2.5	3.5	4	4	3