Modeling nematode population dynamics using a multivariate poisson model with spike and slab variable selection

Gill Giese\textsuperscript{a}, Dayna P. Saldana Zepeda\textsuperscript{b}, Jacquelin Beacham\textsuperscript{c} and Ciro Velasco Cruz\textsuperscript{d}

\textsuperscript{a}Agricultural Science Center at Los Lunas, NMSU, Los Lunas, NM, USA; \textsuperscript{b}Universidad de Colima, Escuela de Mercadotecnia, Villa de Alvarez, Mexico; \textsuperscript{c}Entomology, Plant Pathology, and Weed Science (EPPWS), NMSU, Las Cruces, NM, USA; \textsuperscript{d}Agriculture and Extension, NMSU, Las Cruces, NM, USA

ABSTRACT
Model-based learning of organism dynamics is challenging, particularly when modeling count correlated data. In this paper, we adapt the multivariate Poisson distribution to model nematode dynamics. This distribution relaxes the mean-equal-variance property of the univariate Poisson distribution and allows recovery of the correlation among nematode genera. An observational dataset with 68 soil samples, 11 nematode genera, and 12 soil parameters is analyzed. The Spike and Slab Variable Selection procedure is adapted to obtain parsimonious models for the nematode occurrence. Nematode genus to genus interaction is assessed through the correlation matrix of the model. A simulation study validated the model's implementation. As a result, the model determined the most important covariates for each nematode and classified pairs of nematodes as: sympathetic, antagonistic or neutral, based on their estimated correlations. The model is useful for researchers and practitioners interested in studying population dynamics. In particular, the current results are important inputs when planning strategies for improving or managing soil health regarding nematodes.

ARTICLE HISTORY
Received 30 August 2020
Accepted 22 May 2021

KEYWORDS
Nematode correlation; organism–habitat relationship; parsimonious models; multivariate Poisson lognormal; NUTS algorithm

1. Introduction
The Aitchison and Ho [2]’s multivariate Poisson Lognormal is a flexible and robust model for multivariate count data. Modeling count correlated data, although numerically challenging, can help to understand and learn about population dynamics of organisms such as soil inhabiting nematodes. Two types of relationships intrigue nematologists: the relationship between organisms, and the relationship between the organisms and their environment. The first type enables scientists to learn about the organisms' interaction, via the correlation between nematodes; and the second helps to describe the organism–habitat interaction and can be studied by the mean function of the model. With the multivariate Poisson Lognormal, one can simultaneously model both relationships to develop a
more comprehensive understanding of the organisms’ behavior and to guide research and management decisions.

Nematodes are microscopic, multicellular, non-segmented, roundworms that are common to most soils and rarely occur as a single-species community [16]. Some feed on plant roots (non-beneficial, plant-parasitic, herbivorous), while others feed on bacteria and fungi (beneficial, bacterivorous/fungivorous). Bacterivore and fungivore nematodes benefit the soil ecosystem via their role in the decomposition of organic matter. They feed on bacteria and fungi, and subsequently release minerals and nutrients bound in bacterial and fungal bodies, back to the soil.

Plant parasitic nematodes cause $\sim$ $100$ billion damage yearly to agriculture worldwide. Nematodes are economic pests of many crops, including Vitis vinifera grape vineyards [6,10,11]. Premature vineyard decline, reduced vine vigor, and increased fungal infection and virus transmission as well as yield losses ( $> 60\%$) have all been attributed to plant parasitic nematode infestations [23]. Due to the substantial and long-term investment necessary for grape production, researchers and practitioners seek to understand the abiotic and biotic nature of soil microbial ecosystems and possible relationships between different nematode genera within the soil.

Nematodes have been a recurrent topic of interest in agricultural research. Murray et al. [17] and Zbylut et al. [24] analyzed the abundance of nematodes assuming a number of models; the Hurdle model, Generalized Poisson, Negative Binomial, or Zero Inflated models [1]. However, the abundance metric is a univariate variable that prevents studying genus to genus interaction together with the genus to environment relationship. In order to learn about nematode dynamics, we implement the multivariate Poisson model as an appropriate alternative to more comprehensively study organism dynamics. This generalizes the approaches based on univariate variables in other studies.

Parsimonious models are preferred over complex ones to explain nematode–environment relationships. Indeed, within the regression analysis context, this could be considered a universal rule. George and McCulloch [5] pointed out that parsimonious models simplify interpretations, inferences and conclusions. However, when several covariates are available, a parsimonious model is not straightforward, making it necessary to use a variable selection technique to objectively identify the most important covariates. In this paper, our proposed model combines the multivariate Poisson distribution and the Spike and Slab variable selection [13] technique to select the most important soil-related covariates that explain variations of the mean number of nematodes. Additionally, we use the Huang and Wand [12] prior distribution to model the covariance matrix; the component that explains the interaction between nematode genera.

The dataset is from an observational study with soil samples collected at 25 different New Mexico vineyards. Each sample included occurrence and frequency of different nematodes present, identified to the genus level. Soil parameters, such as pH, and amounts of various mineral nutrients were also measured.

This paper is organized as follows, a brief description of the dataset, and a detailed introduction of the model are given in Section 2; that includes the likelihood and prior distributions. Section 3 includes details and results of a simple simulation study designed to validate the model’s implementation. The analysis and results of the nematode dataset is presented in Section 4, and conclusions are given in Section 5.
2. Materials and methods

2.1. Data

To study the interrelationship among different species of nematodes, soil samples were collected from 25 New Mexico vineyards during 2018. A composite sample consisting of ten 3.2 cm diameter × 45 cm deep soil cores was collected from within 40 cm of 10 individual vine trunks within each of three randomly selected zones per vineyard. Nematodes were extracted from a 100 cm$^3$ subsample of soil from each vineyard zone by elutriation, wet sieving and centrifugal flotation [4,14]. A 1 mL aliquot of each 10 mL soil extract was placed on a chambered counting slide and examined on an inverted Olympus IX51 compound microscope. Observed nematodes were counted and categorized into trophic groups as bacterivores, fungivores, omnivores, carnivores or herbivores. Herbivorous nematodes were further classified to genus.

The nematode quantification revealed 11 different nematode genera present in the soil samples; 7 were Herbivorous or Non-Beneficial, and 4 Beneficial. As mentioned above, the Non-Beneficial were classified to genera level, while the Beneficial to trophic group. However, for all purposes in this study, the nematodes listed in Table 1 are hereafter named only by genera. Not all genera were found in every soil sample; however, in every sample, at least two genera were present. For example, in the first soil sample only genera *Helicotylenchus*, *Tylenchus* and *Omnivores Dorylamoid* were found. This led us to hypothesize that a relationship among the nematodes exists that may explain why some genera coexist more often than others.

The dataset includes 12 soil parameters (Table 2) that provide a partial description of the environment occupied by the nematodes. These soil parameters might influence the presence of some types of nematodes. For example, soil pH might impact a given nematode genus’s presence. Therefore, the proposed model may provide new insight about the nematode–environment interaction, explain some of the interaction among nematodes and help understand the organismal dynamics.

| Table 1. List of nematode genera in the composite soil samples (B: Beneficial, H: Herbivorous or non-beneficial) collected from New Mexico vineyards, 2018. |
|---|---|
| Type | Nematode | Type | Nematode |
| H | Pratylenchus | H | Paratylenchus |
| H | Tylenchorhynchus | B | Tylenchus |
| H | Meloidogyne | B | Bacterivores |
| H | Mesocriconema | B | Fungivores |
| H | Helicotylenchus | B | Omnivores Dorylamoid |
| H | Xiphe americanum | |

| Table 2. List of soil parameters analyzed from New Mexico soil samples, 2018. |
|---|
| Soil parameters |
| Nitrate | Sulfate | Zinc | Manganese |
| Iron | Copper | Boron | Chloride |
| pH | SAR | Moisture | Saturation |
2.2. Model

The Poisson distribution for the $i$th soil sample is

$$P(Y_i = y_i|\lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!},$$

where $y_i$ is the number of nematodes of one kind or genus, and $\lambda_i$ is the mean parameter of the distribution.

It is well known that the Poisson distribution’s mean equals its variance, a strong condition that in practice is not commonly met. In order to relax this condition, and make the Poisson an appropriate model to a wide variety of data, some modifications are typically applied. For example, one could include a mixing parameter, so that marginally $y_i$ follows a Negative Binomial distribution. In this paper, we assume a multivariate Poisson distribution that, as shown in Section 2.3, relaxes the mean equal variance condition and, at the same time, helps to model the correlations between pairs of nematodes as explained in the next section.

2.2.1. Modeling interrelationship among organisms

In order to learn about the relationship of nematodes with themselves and their surroundings, we adapt the Poisson Lognormal model proposed by [2] that generalizes (1) into a multivariate distribution to model an $m$ dimensional response vector.

Let $Y_i = (Y_{i1}, \ldots, Y_{im})^T$ be the response vector for the $i$th sample. In the [2] model, the corresponding $m$ dimensional vector of parameters $\lambda_i$ is assumed random with a Lognormal distribution. Equivalently, we define a random vector $W_i = (W_{i1}, \ldots, W_{im})^T$ so that $\log(\lambda_i) = W_i$. The elements of the response vector $Y_i$ are conditionally independent given $W_i$, as follows

$$P(Y_i = y_i|W_i = w_i) = \prod_{j=1}^{m} \frac{e^{-w_{ij}(e^{w_{ij}})^{y_{ij}}}}{y_{ij}!};$$

and

$$W_i \sim N_m(\alpha + \mu_i, \Sigma),$$

where $\Sigma$ is the variance–covariance matrix of size $m \times m$. Both parameters, the intercept ($\alpha$) and $\mu_i$, define the mean function of the distribution, as in a regular regression analysis. The definition of $\mu_i$ is application dependent, typically $\mu_i = X_i^\ast^T \beta$, where $X^\ast$ is the design matrix $X$ with transformed columns, and $\beta$ is a vector of regression coefficients of size $q$. More detailed information on how $X^\ast$ is constructed for this analysis, is given in the Appendix.

Thus, the likelihood function or sampling distribution of a random sample of size $n$, provided that $\text{Cov}(W_i, W_{i'}) = 0 \forall i \neq i', i = 1, \ldots, n$, is as follows

$$P(y|w) = \prod_{i=1}^{n} P(y_i|w_i),$$

where $P(y_i|w_i)$ is given in (2).


2.3. Marginal moments

The model in (4) is appropriate to model Poisson correlated data, and to an extent, it is straightforward mathematically, allowing the derivations of marginal moments of $Y_i$, analytically. For instance, the basic marginal moments such as mean, variance, covariance, and correlation, have the following expressions.

\[
\begin{align*}
E(Y_{ij}) &= \exp(\mu_{ij} + \sigma_{jj}/2) \\
\Var(Y_{ij}) &= E(Y_{ij}) + E(Y_{ij})^2 \left( \exp(\sigma_{jj}) - 1 \right) \\
\Cov(Y_{ij}, Y_{ij'}) &= E(Y_{ij})E(Y_{ij'}) \left( \exp(\sigma_{jj'}) - 1 \right) \\
\Corr(Y_{ij}, Y_{ij'}) &= \frac{\Cov(Y_{ij}, Y_{ij'})}{\sqrt{\Var(Y_{ij})\Var(Y_{ij'})}},
\end{align*}
\]

where $\sigma_{jj'}$ is the $(j, j')$ element of $\Sigma$; and $j(\neq j') = 1, \ldots, m$.

The marginal model derived from (2) makes the variance and mean different. In fact, $E(Y_{ij}) < \Var(Y_{ij})$ proves that this model relaxes the constraint of the univariate Poisson distribution, as it is often desired. Notice that because $|\Corr(Y_{ij}, Y_{ij'})| < |\Corr(W_{ij}, W_{ij'})|$, the range of possible correlation values is not as wide as that of the corresponding Normal distribution [2].

The equations in (5) allow the computation of marginal moments. However, unlike the normal distribution case, the variance and covariance are sample specific because both depend on $\mu_{ij}$, and are not equal for all $i$. Thus, the multiple values of variance and covariance lead to a cumbersome marginal analysis. Instead, it is more convenient to work with $\mu$ and $\Sigma$ for inferences.

2.4. Simpler models

We evaluate two types of simpler models. In both types, the nematode correlation is relaxed; however, one follows the univariate version of the [2] model, while the others are traditional models for count data.

2.4.1. Independent model

A simplified model of interest results if $\Sigma$, the covariance matrix, is set equal to the identity matrix. This simplification implies that the nematodes do not interact with each other, an assumption that contradicts our intuition about the social behavior of nematode populations. However, if proved more appropriate for the data, this model would greatly simplify interpretation. Additionally, this model provides an opportunity to evaluate the significance of the effect of the nematode’s interaction, when compared with the model given in (3).

In particular, the simplification implies that

\[
W_i^{(ni)} \sim N_m(\alpha^{(ni)} + \mu_i^{(ni)}, I),
\]

The mean structure in (6) is preserved as in (3), but the covariance structure changes, as it is set equal to the identity matrix, $I$. The superscripts $(ni)$ in (6) are labels that denote ‘no nematode interaction’.
2.4.2. Univariate models

It is of interest to evaluate the ability of three univariate models to model the nematode dataset. In essence, they are similar to the independent model above, however, their sampling distributions, \( P(Y_i | \theta) \), are Poisson, Generalized Poisson, and Negative Binomial. The first two are defined as in [7], while the latter is defined in [8]. Given their definitions, \( \theta \) is a parameter of parameters that completely specify the distributions. For instance, \( \theta = \mu \) for the Poisson distribution, \( \theta = [r, \mu] \) for the Negative Binomial, and \( \theta = [\delta, \mu] \) for the Generalized Poisson.

In practice, a large number of covariates are often available. Models with several covariates although useful, are not easy to use for inference or interpretation. Practitioners are typically more interested in parsimonious models, with only the most important covariates. To determine the subset of most important soil parameters to explain the mean number of nematodes, we adapt the Spike and Slab variable selection procedure [13].

2.5. Variable selection

It is of interest to have a simple mean model for each nematode genus (i.e. a parsimonious mean function, \( \mu_j \), for \( j = 1, \ldots, m \)). Such a model will facilitate inference, interpretation and understanding of the relationship between nematodes and soil parameters. We adapt the Spike and Slab (SS) variable selection procedure of [13] to find the subset of the most important covariates to be in \( \mu_j \). The SS procedure assumes that the \( k \)th regression coefficient follows a normal distribution conditioned on a binary variable, \( \gamma_k \), that drives the significance of the coefficient through a mixture distribution, as follows.

\[
\beta_k | \gamma_k \sim N(0, \gamma_k v_k), \quad \text{for } k = 1, \ldots, q,
\]

\[
\gamma_k = \begin{cases} 
1, & \text{with probability } \xi \\
\nu_0, & \text{with probability } 1 - \xi.
\end{cases}
\]

(7)

The parameter \( \xi \in (0, 1) \), and \( \nu_0 > 0 \) is a user-defined small quantity that serves as a reference value for the variance of the non-significant regression coefficients; \( v_k \) is the variance parameter. Two prior distributions for \( \beta_k \) are in play according to (7). The two distributions in the mixture have the same mean but different variances. When \( \beta_k \) is important to explain the response variable, then with high probability \( \gamma_k = 1 \) and \( \beta_k \sim N(0, \nu_k) \); depending on the magnitude of \( \nu_k \), the profile of this density could resemble a slab if \( \nu_k \) is large. However, when \( \beta_k \) is not important, \( \gamma_k = \nu_0 \) and the prior distribution for \( \beta_k \) collapses at its mean value, resembling a spike at 0.

As a closely related topic to the variable selection step, we want to mention that the following simple transformation of the columns of the design matrix, \( X \), is numerically convenient. Let \( X^* \) be the transformed design matrix, its elements are \( x_{il}^* = \frac{(x_{il} - \min_{i=1,\ldots,n} x_{il})}{\text{range}_{x_l}} \), for \( l = 1, \ldots, q \), where range \( x_l = (\max_{i=1,\ldots,n} x_{il} - \min_{i=1,\ldots,n} x_{il}) \). Furthermore, given that \( \mu \) is a function of \( X^* \), the posterior distributions for the regression coefficients are for the transformed covariates. Thus, the following transformation, \( \beta_l = \beta^*_l \times \frac{1}{\text{range}_{x_l}} \) recovers the regression coefficients for the untransformed covariates, where \( \beta^*_l \) is the regression coefficient for \( x_{il}^* \), and \( \beta_l \) is for \( x_{il} \).

To learn about unknown parameters involved in the multivariate Poisson and the SS models presented above, prior distributions are needed. In the next section, we give the
priors for these parameters, together with the numerical algorithms needed to estimate their posterior distributions.

### 2.6. Prior distributions definition and algorithms for posterior distributions estimation

Learning about the unknown parameters in equations (4) and (3) requires prior distributions and numerical algorithms such as MCMC and Gibbs sampling to estimate their posterior distributions. Given the prior for \( W \) in (3), its full conditional distribution results analytically intractable, which entails the use of the NUTS\(^1\) algorithm to efficiently simulate samples from \( W \)'s full conditional distribution.

The Gibbs sampling algorithm is used to learn about \( \Sigma \), whose prior and full conditional distributions are given in [12]. The hyper-parameters of the prior and hyper-prior distributions involved in the full conditional for \( \Sigma \) are defined as follows: an inverse Wishart is included as an intermediate distribution, with the degrees of freedom parameter set equal to 2, and a varying scale matrix. The main diagonal elements of the scale matrix follow independent Inverse Gamma, other intermediate distributions, with rate parameter equal to the square of 1e6. Huang and Wand [12] show that marginally, the standard deviations of \( \Sigma \) follow Haft \( t \)-distributions, and that the correlations follow Uniform distributions in the interval \((-1,1)\). See Huang and Wand [12] for more detailed information about the prior distribution for \( \Sigma \).

The Gibbs sampling is also used to learn about \( \alpha \) and \( \beta \). The prior distribution for \( \alpha \) is non-informative, of the form \( \pi(\alpha) \propto 1 \), and its full conditional is given in the Appendix. The distribution for \( \beta \) is the Spike and Slab prior given in (7), with a Uniform(0,1) prior distribution for \( \xi \), a Gamma(shape = 5, rate = 50) prior distribution for \( 1/\nu_k \) \((k = 1, \ldots q)\), and as for \( \nu_0 \), it is set equal to 5e-15.

### 3. Simulation study

To evaluate the implementation of the model, a simulation study is carried out under the following conditions.

A dataset is simulated using Equations (2) and (3). The simulated sample size is \( n = 80 \), \( m = 4 \), and \( \Sigma \) is given in (8). For each of the \( m = 4 \) simulated genera, the vector of regression coefficients, \( \beta \) of size \( q = 13 \), is equal to \((-1.0, 2.0, -2.0, 0.0, 0.0, 1.2, -1.3, 1.9, 0.0, -1.9, 0.0, -1.5, 1.5)\). Components of the data matrix, \( X_D \), are simulated from a \( N_q(0,0.25) \), \( i = 1, \ldots, n \).

\[
\Sigma = \begin{pmatrix}
1 & -0.5674 & 0.1385 & 0.3629 \\
-0.5674 & 1 & -0.1829 & -0.0669 \\
0.1385 & -0.1829 & 1 & 0.7986 \\
0.3629 & -0.0669 & 0.7986 & 1
\end{pmatrix}
\] (8)

#### 3.1. Sensitivity study

The effect of sample size on the parameter estimations is evaluated in a sensitivity study. Five sample sizes \( n \) are evaluated \( \{60, 70, 80, 90, 100\} \). The mean square error (MSE) is
computed for the regression coefficients and the elements of the correlation matrix as the summarizing statistic. The MSE of the regression coefficients is defined as

\[ \text{MSE}(\beta) = \frac{1}{m} \sum_{j=1}^{m} (\beta_j - \hat{\beta}_j)^2, \]

where \( \hat{\beta}_j \) is the median of the samples generated from the parameter’s posterior distribution. In order to account for variability in the simulation process, we repeat the computation of the \( \text{MSE}(\beta) \) \( R \) times. The average MSE computed as \( \text{MSE}(\beta) = \frac{1}{R} \sum_{r=1}^{R} \text{MSE}_r(\beta) \), and the empirical 68% interval \( (\text{MSE}(\beta) \pm sd_{\text{MSE}(\beta)}) \) where \( sd_{\text{MSE}(\beta)} = \frac{1}{R-1} \sum_{r=1}^{R} (\text{MSE}_r(\beta) - \text{MSE}(\beta))^2 \) are reported. Similar summaries are computed for the correlation matrix.

### 3.2. Simulation results

The algorithm used for model fit was coded in Armadillo C++ [21] and RStudio [19]. The following summarizations are based on the last 200,000 samples taken every 10th of the 400,000 iterations. Convergence of the chains was graphically verified based on trace plots, which are the most common convergence diagnostic method, according to [20]. However, for brevity, we are not showing such graphics (available upon request).

Table 3 shows the 95% Credible Intervals (CI) for the unique elements of the estimated \( \Sigma \). Every unique element of \( \Sigma \) is within its CI. Also, the maximum at posteriori (MaP) of each entry is close to its corresponding element in \( \Sigma \) (data shown as the last column in the table).

The biplot [3] graphic displays multivariate relationships in two dimensions. Figure 1 presents the biplot of the MaP values of the correlation matrix. This graphic shows items ‘Nema4’ and ‘Nema3’ as positively correlated (their arrows are pointing in the same direction), while items ‘Nema1’ and ‘Nema2’ are negatively correlated (their arrows point in opposite directions). Therefore, visual understanding of the items’s relationships can be effectively and easily assessed using biplots, and can prompt further research questions; for instance, one can focus attention to study ‘Nema4’ and “Nema3”, or ‘Nema1’ and ‘Nema2’ in more detail to explore biological reasons for those correlation values.

Estimates of the regression coefficients are shown in Figure 2. The asterisks indicate the actual values used to simulate the data. The vertical lines are the 95% Credible intervals.

| Entry | 2.5% | MaP | 97.5% | \( \Sigma(i,j) \) |
|-------|------|-----|-------|-----------------|
| 1     | 0.2836 | 0.6233 | 0.3073 | 0.5674 |
| 2     | 0.407 | 0.4195 | 0.6952 | 0.1385 |
| 3     | 0.5919 | 0.2617 | 0.1222 | 0.1829 |
| 4     | 0.0174 | 0.3501 | 0.6379 | 0.3629 |
| 5     | 0.483 | 0.1493 | 0.2223 | 0.0669 |
| 6     | 0.3002 | 0.643 | 0.8525 | 0.7986 |

Note: All the entries are within their CIs.
Figure 1. Biplot representation of the MaPs of the estimated correlation matrix $\Sigma$.

Figure 2. 95\% Credible intervals (CIs) for the simulated regression coefficients, $\beta$. The asterisks represent the simulated values. Filled dots are the maximum at posteriori (CI), and the dots are the MaPs (maximum at posteriori). In red (visible only in the electronic version) are the CIs that do not touch the reference line at zero. The CIs of Var7 of Nema2, Var2 of Nema3, and Var2 of Nema4 missed to include their corresponding simulated values by a small quantity.

3.3. Sensitivity study results

Figure 3 shows the MSEs of the regression coefficients (Reg Coeff) and of the elements of the correlation matrix (Correl Matrix) for different sample sizes, $n$. For each $n$, the
Figure 3. MSE average and its empirical 68% interval, for different sample sizes, of the regression coefficients and correlation matrix.

average and the interval of the MSE are displayed in the figure, based on $R = 5$. The MSE of the regression coefficients decreases as $n$ increases, while that of the correlation matrix remains fairly constant. The minima MSEs are reached when $n = 100$. Although they tend to decrease as $n$ increases, the regression coefficients MSE values for samples sizes 70, 80 and 90 are not so different from each other.

The results from the simulation study validates the model and its implementation. In the following section, we present the results of the nematode data analysis.

4. Nematode data analysis results

The nematode dataset as described in Section 2.1 has 12 soil parameters as covariates, listed in Table 2. The response variable is the number of nematodes found in the soil samples. Any nematode genus not found in a particular soil sample, was set equal to zero. Eleven genera of nematodes were found in the samples. To jointly model the nematodes, we use the Multivariate Poisson model given in (4).

Nematode genera occurrence tends to be discriminate and possesses a degree of exclusivity, i.e. certain genera are likely to occur together, while others are seldom found together simultaneously. One of this study’s objectives is to investigate the nematode genus to genus relationships. In order to achieve this objective, we estimate the correlations via a simple transformation of $\Sigma$ in (3). With the correlations, nematodes classification takes place as follows: if two genera of nematodes have a positive correlation then they are described as sympathetic, if their correlation is negative, they are described as antagonistic, and if their correlation is around zero, they are described as neutral.

Another objective of this study is to describe the nematode–habitat relationship, that can be achieved by selecting the subset of covariates in the mean function of the model, that better explains the mean number of nematodes. Thus, the variable selection procedure presented in Section 2.5 is implemented for the nematode dataset. The results of this objective are particularly important because learning about the subset of soil parameters that explains nematode presence can help to develop and implement strategic plans or practices to increase/decrease nematode populations in order to provide better growing conditions for a particular crop or improve soil health in general.
The response vector that represents the actual counts of the eleven different genera of nematodes found in the \( i \)th soil sample, is denoted by \( Y_i = (Y_{i1}, \ldots, Y_{i11})^T \), and their genera are listed in Table 1.

### 4.1. Model comparison

In order to compare models, we computed the Deviance Information Criterion [22] as the statistic that summarizes model fit. This criterion assists in selecting one model as best if its DIC is smallest among evaluated models. We are particularly interested in evaluating two models: one that includes the correlation of nematodes (Full model, (3)), and one that assumes that the nematodes do not interact with each other (Independent model, (6)). However, we are also interested in comparing the three univariate models, typically used to model count data: Poisson, Generalized Poisson, and Negative Binomial.

Based on the DIC shown in Table 4, the worst possible model is the Poisson, while the best is the Full model, though slightly more complex than the Independent model (since it has a greater pD). Thus, our proposed model given in (4) better explains the nematode dataset. This result supports the assumption that the nematodes interact with each other, and modeling that interaction results in a much more comprehensive and richer model. Incidentally, within the evaluated class of univariate models, the Generalized Poisson is the best for the nematode data.

In the following sections, we present summary results of the analysis of the nematode data under the Full model.

### 4.2. Correlation between pairs of nematodes

Figure 4 shows the MaPs of the posterior distributions of the elements of the correlation matrix (derived from \( \Sigma \)), and its 90% credible interval. There is a large negative correlation worth highlighting. The correlation between \( B_{\text{Omnivores\_Dorylamoid}} \) and \( B_{\text{Bacterivores}} \) genera is \(-0.34\), which is the largest negative value among all correlation values. This number suggests that these two organisms are somewhat antagonistic, even though they are both ‘Beneficial’. On the other hand, the correlation between \( B_{\text{Tylenchus}} \) and \( H_{\text{Tylenchorhynchus}} \) is 0.64, and between \( B_{\text{Tylenchus}} \) and \( B_{\text{Bacterivores}} \) is 0.71, both

|                  | DIC     | pD     |
|------------------|---------|--------|
| Full model       | 2143.6  | 398.8  |
| Independent model| 2391.9  | 394.8  |
| Poisson          | 12763.7 | 775.0  |
| Generalized Poisson| 3082.1 | 153.7  |
| Negative Binomial| 3129.0  | 205.7  |

Notes: The Full model includes the nematodes interaction. The Independent model assumes no-interaction of nematodes. The Poisson, Negative Binomial and Generalized Poisson are traditional models for count data.
Figure 4. Nematode correlation estimates. The number at the top in each cell is the Maximum at Posteriori, and at the bottom, the 90% credible intervals (in parentheses).

Values suggest sympathy. There are other examples of antagonistic and sympathetic correlations evident in the figure. However, the association can be better observed in Figure 5, that shows the biplot of the point estimate (MaPs) of the correlation matrix. It suggests that B_Bacterivores and B_Tylenchus are 'sympathetic', because their arrows point in the same direction. Likewise, H_Paratylenchus and B_Omnivores_Dorylaimoid are somewhat sympathetic. However, these two pairs have arrows pointing to opposite directions, thus we categorize them as 'antagonistic'. Finally, H_Mesocriconema, H_Helicotylenchus and H_Pratylenchus are examples of 'neutral' nematode, due to their relatively short arrows in the plot.

### 4.3. Variable selection

In Figure 6, the regression coefficient point estimates (MaPs) and CIs are plotted with a subplot for each genus. The x-axis lists all the covariates given in Table 2, and the y-axis has the estimates of the regression coefficients and their 90% credible
Figure 5. Biplot representation of the MaPs of the estimated correlation matrix, of the nematode dataset.

 intervals. The vertical segments in red (visible only in the electronic version) indicate that they do not intersect the reference horizontal line at 0, interpreted as the covariates corresponding to those regression coefficients have a significant effect on the mean count. For instance, for H_Pratylenchus genus, ‘SAR’ (negative effect) is the single most important covariate to explain the presence of this nematode. For the B_Tylenchus nematode, ‘manganese’ (positive effect), ‘pH’ (negative effect), and ‘SAR’ (negative effect) are the important covariates to explain the mean count of this genus. Meanwhile, B_Omnivores_Dorylamoid’s important covariates are ‘manganese’ (positive effect), ‘chloride’ (small positive effect), ‘pH’ (positive effect), and ‘moisture’ (negative effect). Similar assessments can be derived for other genera, based on the figure.

It is also worth noticing that there are some covariates that on average are far from being zero; for instance, the MaP value of pH in the H_Mesocriconema genus shows a negative effect on the mean number of nematodes, however its large variability makes its CI intersect
Figure 6. The effect size of the covariates by genus. The maximum at posteriori (filled dots) and 90% credible interval (vertical bars). In red are the CIs that do not include zero.

the reference line. A closer study of this covariate, and others with a similar pattern, may reveal more information about their effects.

4.4. Model goodness of fit

A graphical depiction of the model’s goodness of fit is given in Figure 7. Each panel shows the fitted mean count versus the observed count values, and the 90% credible interval of the fit, by genus. It is clear that the width of the CI increases as the observed values increase, an expected pattern because (a) the dataset does not contain many large values, and (b) as the values increase in magnitude, so do their variability. For instance, with respect to point (a), genus *H_Paraylenchus* has only one observation greater than ninety. Thus, from the information presented in the figure, we are confident that the model is adequately explaining the variability of the data.

To quantitatively assess goodness of fit of the model, we computed *p*-values of the Chi-squared goodness of fit test and the paired *t*-test by nematode genus. Both tests use the
observed counts and the MaPs of the fitted mean counts as inputs, hereafter referred as variables. The Chi-squared test computes the total distance between these two variables to evaluate the null hypothesis that states: no difference between the two variables. While, the paired t-test computes the distance between the averages of these two variables, to evaluate the null hypothesis that states: the variables mean difference is zero. Table 5 shows the p-values that result from the tests. The Chi-squared p-values are all equal to one. The p-values of the paired t-test are all greater than 5%. Thus, the conclusion is: do not to reject the null hypotheses, which means that the model is explaining the data adequately.

5. Concluding remarks

The model that combines the multivariate Poisson model and the Spike-Slab variable selection technique was able to identify the set of the most important soil parameters that explain the mean count of nematodes and to estimate nematode correlations of eleven nematode genera.
Table 5. Estimated $p$-values of the goodness of fit test, based on the Chi-squared and paired $t$ tests, by genus.

| Genus               | Chi-squared | Paired $t$ |
|---------------------|-------------|------------|
| H_Paratylenchus     | 1.00        | 0.95       |
| B_Tylenchus         | 1.00        | 0.67       |
| H_Mesocriconema     | 1.00        | 0.92       |
| H_Pratylenchus      | 1.00        | 0.89       |
| H_Helicotylenchus   | 1.00        | 0.87       |
| H_Meloidogyne       | 1.00        | 0.94       |
| H_Tylenchorhynchus  | 1.00        | 0.82       |
| B_Bacterivores      | 1.00        | 0.79       |
| H_Xiphe_americanum  | 1.00        | 0.52       |
| B_Omnivores_Dorylamoid | 1.00    | 0.54       |
| B_Fungivores        | 1.00        | 0.44       |

Typically, the correlation between nematodes is not modeled when analyzing the abundance of nematodes in soil samples. However, we found in Section 4.1 that the univariate analyses were outperformed by our proposed model, an approach that specifically models the correlation of nematodes. These results validate our hypothesis that nematodes are social organisms, and that when this interaction is modeled, it improves model fit.

The estimated nematode correlations are useful for further inferences. For instance, one can classify nematodes as antagonistic (those with negative correlation), neutral (those with correlation around zero), or sympathetic (those with positive correlation). Such information could aid researchers in testing, planning and/or applying some management strategies to control herbivorous, destructive nematodes; or to improve soil health in general.

The SS is a flexible and efficient variable selection technique when applied to the analysis of the multivariate Poisson data. The SS method selected the most important soil parameters that explain the mean count of nematodes, specific to each genus. This piece of information, together with the correlations of nematodes, can be of broad and practical importance as a design component of strategic plans to regenerate the health of long exploited soils.

The simulation study shows that the CIs of the elements of the correlation matrix are rather wide (see Table 3). In future work, we will evaluate other possible prior distributions for learning about the correlation matrix in order to narrow the width of the CIs. We chose the prior distribution of Huang and Wand [12] because of its simplicity and robustness. However, the procedure described by Lewandowski et al. [15] is another interesting option.

Finally, given its generality, we want to adapt the proposed model to analyze experimental data. In particular, from an ongoing experiment designed to study nematode population growth under different conditions. The adaptation is going to be able to incorporate the effect of time on the nematode populations while accounting for nematode genera correlations.

**Note**

1. No-U-Turn Sampler (NUTS) is an MCMC algorithm developed by Homan and Gelman [9], that is based on the Hamiltonian Dynamics. In the Appendix, we give a brief description of this algorithm.
Disclosure statement
No potential conflict of interest was reported by the author(s).

References
[1] D. Agarwal, A. Gelfand, and S. Citron-Pousty, Zero-inflated models with application to spatial count data, Environ. Ecol. Stat. 9 (2002), pp. 341–355.
[2] J. Aitchison and C.H. Ho, The multivariate Poisson-LogNormal distribution, Biometrika 76 (1989), pp. 643–653.
[3] D. Bradu and K.R. Gabriel, The biplot as a diagnostic tool for models of two-way tables, Technometrics 20 (1978), pp. 47–68.
[4] D.W. Byrd, K.R. Barker, H. Ferris, C.J. Nusbaum, W.E. Griffin, R.H. Small, and C.A. Stone, Two semi-automatic elutriators for extracting nematodes and certain fungi from soil, J. Nematol. 8 (1976 Jul), pp. 206–12.
[5] E.I. George, R.E. McCulloch, Approaches for Bayesian variable selection, Stat. Sin. 7 (1997), pp. 339–373.
[6] G. Giese, J. Beacham, I. Guzman, M. Kersten, C. Velasco-Cruz, and S. Thomas, Nematode impact on grape compositional quality. In process, March 2019.
[7] T. Harris, Z. Yang, and J.W. Hardin, Modeling underdispersed count data with generalized Poisson regression, Stat. J. 12 (2012), pp. 736–747.
[8] J.M. Hilbe, Negative Binomial Regression, 2nd ed., Cambridge University Press, New York, 2011.
[9] M.D. Homan and A Gelman, The No-U-Turn sampler: adaptively setting path lengths in hamiltonian Monte Carlo, J. Mach. Learn. Res. 15 (2014), pp. 1593–1623.
[10] A.D. Howland, R.P. Schreiner, and I.A. Zasada, Spatial distribution of plant-Parasitic nematodes in semi-Arid vitis vinifera vineyards in Washington, J. Nematol. 46 (2014), pp. 321–330.
[11] A.D. Howland, P.A. Skinkis, J.H. Wilson, E. Riga, J.N. Pinkerton, I.A. Zasada, and R.P. Schreiner, Impact of grapevine (Vitis vinifera) varieties on reproduction of the northern Root-Knot nematode (Meloidogyne hapla), J. Nematol. 47 (2015), pp. 141–147.
[12] A. Huang and M.P. Wand, Simple marginally noninformative prior distributions for covariance matrices, Bayesian Anal. 8 (2013 Jun), pp. 439–452.
[13] H. Ishwaran and J.S. Rao, Spike and slab variable selection: Frequentist and Bayesian strategies, Ann. Stat. 33 (2005 Apr), pp. 730–773.
[14] W.R. Jenkins, A rapid centrifugal-flotation technique for separating nematodes from soil, Plant Dis. Repor. 48 (1964), pp. 692.
[15] D. Lewandowski, D. Kurowicka, and H. Joe, Generating random correlation matrices based on vines and extended onion method, J. Multivar. Anal. 100 (2009), pp. 1989–2001.
[16] M.V. McKenry and L. Bettiga, Nematodes. In Grape Pest Management 3rd ed., L. Bettiga, ed., University of California, Agriculture and Natural Resources, Oakland, CA, 2013. pp. 447–470.
[17] L. Murray, S. Thomas, J. Schroeder, S. Kreider, Z. Ou, J. Trojan, and C. Fiore, Modeling the root-Knot nematode/NUTSEDGE pst complex: perspectives from weed science, nematology and statistics, Conf. Appl. Stat. Agricul. 5 (2011), pp. 128–151.
[18] R.M. Neal, MCMC using Hamiltonian dynamics, Handbook of Markov Chain Monte Carlo, Vol. 54, (2010), pp. 113–162.
[19] RStudio Team. RStudio: Integrated Development Environment for R. RStudio, PBC., Boston, MA, 2020.
[20] V. Roy, Convergence diagnostics for Markov chain Monte Carlo, Annu. Rev. Stat. Appl. 7 (2020), pp. 387–412.
[21] C. Sanderson and R. Curtin, Armadillo: a template-Based C++ library for linear algebra, J. Open Source Softw. 1 (2016), pp. 1–2.
[22] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van der Linde, The deviance information criterion: 12 years on, J. R. Stat. Soc Ser. B Stat Methodol. 76 (2014), pp. 485–493.
[23] D. Teliz, B.B. Landa, H.F. Rapoport, F.P. Camacho, and R.M.J. Diaz, Plant-parasitic nematodes infecting grapevine in southern Spain and susceptible reaction to root-knot nematodes of rootstocks reported as moderately resistant, Plant Disease 91 (2007), pp. 1147–1154.
Appendix

A The likelihood function

Bayesian modeling requires the following inputs: the Likelihood function, also known as sampling distribution, and prior distributions for the unknown parameters. The latter were given in Section 2.6, and the likelihood is given in (4). These distributions define the potential and its gradient of the Hamiltonian, for the NUTS algorithm.

B Potential and its gradient

The model for $W$ has an intercept, denoted by $\alpha$. Typically, its prior distribution is $\pi(\alpha) \propto 1$. If one computes the distribution for $W$ marginal of $\alpha$, it results in that $W \sim N(X^t \beta, \Sigma_T)$, where $\Sigma_T = (\Sigma^{-1} - \Sigma^{-1} JJ^T \Sigma^{-1} \sigma_0)^{-1}$, $J$ is a vector of ones, and $\sigma_0 = 1/(J^T \Sigma^{-1} J)$. Then, for the NUTS algorithm [9], the potential energy ($U$), when $Y$ (an $m$ dimensional vector) follows a Poisson distribution with $n$ samples, is

$$U(W) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[ \exp(W_{ij}) - Y_{ij} W_{ij} \right] + (W - X^* \beta)^T \Sigma_T^{-1} (W - X^* \beta)/2.$$  \hspace{1cm} (A1)

The gradient vector, which is the derivative of (A1) with respect to $W_{ij}$, is

$$\nabla U = \exp(W) - Y + \Sigma_T^{-1} (W - X^* \beta).$$  \hspace{1cm} (A2)

The NUTS algorithm is an efficient extension of the Hamiltonian Monte Carlo algorithm [18].

C Hamiltonian Monte Carlo (HMC)

In physics, the Hamiltonian, $H(W, p)$, is defined as

$$H(W, p) = U(W) + K(p),$$

where $U(\cdot)$ (see (A1)) is the potential energy (its input, also known as the position, is typically the vector of unknown parameters of dimension $m$, that in this case the position vector is $W$), and $K(p)$ is known as the kinetic energy ($p$ is the momentum vector of size $m$). According to the Hamiltonian, the dynamics of an ‘object’ in a $m$ dimensional space, is described with these two energies. Given its position and momentum, the object will likely move, after a given number of jumps, to a new position with direction defined by the gradient (see (A2)), following the frictionless trajectory. The object’s new position, determined by the LeapFrog method, whose inputs are: the current $W, p$, the number of jumps and their height, is proposed. The proposed position is accepted/rejected using the Markov Chain Monte Carlo ratio.

As described by Neals, the algorithm needs the definition of the number of jumps and their height, two very sensitive parameters that require tuning. As a result, the algorithm is an option but a non-efficient one, even though it has some other attractive properties, such as fast convergence of the chains and less autocorrelated samples once the jumps and heights are appropriately tuned up. The NUTS algorithm, which uses HMC, sets the number of jumps to 1, and automatically calibrates the more appropriate jump’s height, making it a very attractive algorithm, since there is no need for tuning, and it preserves the HMC properties: easy to implement, fast convergence of the chains to their stationary states, and low autocorrelation of samples of the vector positions.

To learn about the other parameters, their full conditional distributions are needed. They are defined as follows.
D  Full conditionals

On each iteration, the algorithm updates \( W \), and the unknown parameters using the following full conditionals.

- The full conditional for \( \beta \) is
  \[
  \beta \sim \text{Normal} \left( \Sigma_\beta^{-1} X^* \Sigma_\beta^{-1} W, \Sigma_\beta^{-1} \right),
  \]
where \( \Sigma_\beta = (X^* \Sigma_\beta^{-1} X^* + \Delta^{-1}) \), and \( \Delta = \text{diag}(\gamma_k \nu_k) \).
- For \( k = 1, \ldots, q \) (where \( q \) is the number of regression coefficients)
  1. The full conditional for \( \gamma_k \) is
     \[
     \gamma_k \sim \xi_{1,k} \frac{\xi_{2,k}}{\xi_{1,k} + \xi_{2,k}} \delta_1(\cdot)
     \]
     where \( \xi_{1,k} = (1 - \xi) v_0^{-1/2} \exp(-\beta_k^2/(2v_0v_k)) \) and \( \xi_{2,k} = \xi \exp(-\beta_k^2/(2v_0v_k)) \), and \( \delta_1(\cdot) \) is a measure concentrated at \( x \).
  2. For \( \nu_k \), it is
     \[
     1/v_k \sim \text{Gamma}(5 + 1/2, \text{rate} = 50 + \beta_k^2 \gamma_k).
     \]
- For \( \xi \), its full conditional is
  \[
  \xi \sim \text{Beta}(1 + \#\{k : \gamma_k = 1\}, 1 + \#\{k : \gamma_k = \nu_0\}).
  \]
- The full conditional for \( \alpha \) is
  \[
  \alpha \sim \text{N}(\sigma_0 \Sigma^{-1} W, \sigma_0).
  \] (A3)
- Following [12], the full conditional for \( \Sigma \) is
  \[
  \Sigma \sim \text{Inverse-Wishart}(df_1, S),
  \]
where \( df_1 = df + n + m - 1 \), and \( S = \sum_{i=1}^m (W_i - \alpha - X^* \beta)(W_i - \alpha - X^* \beta)^T + 2 \ast df \ast \text{diag}(1/a_1, \ldots, 1/a_m) \), and \( df = 2 \).
- The full conditional for \( a_k \) for \( k = 1, \ldots, m \) is
  \[
  a_k \sim \text{Inverse-Gamma}\left(\frac{df + m}{2}, df \ast \text{diag}(\Sigma^{-1})_{kk} + 1/(1e6)^2\right)
  \]

E  Observation vector and design matrix

Let \( Y_i = (Y_{i1}, \ldots, Y_{im})^T \) be the vector of actual number of nematodes found in the \( i \)th sample. Then, the observation vector of size \( nm \) is

\[
Y = \begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix}
\]

Similarly, the construction of the matrix of covariates is carried out as follows. We assume that the sample-specific observed covariates values are in \( X_D \) of size \( n \times q \). The transformed matrix, denoted by \( X^*_D \), is computed from \( X_D \) directly applying the transformation laid out in Section 2.5.

For the \( m \) nematodes and sample \( i \), the design sub-matrix of size \( (m \times mq) \) is

\[
X^*_i = \begin{bmatrix}
x_i^* & 0 & 0 & \ldots & 0 \\
0 & x_i^* & 0 & \ldots & 0 \\
0 & 0 & \ddots & \ldots & 0 \\
\vdots & \vdots & \ddots & \ldots & \vdots \\
0 & 0 & 0 & \ldots & x_i^*
\end{bmatrix}
\]
where $x_i^*$ is the $i$th row vector of matrix $X_D^*$, of size $q$. Finally, for the complete data set

$$X^* = \begin{bmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{bmatrix},$$

The dimension of $X^*$ is $mn \times mq$. 