SELEÇÃO DE COVARIÁVEIS PARA AJUSTE DE REGRESSÃO LOGÍSTICA NA ANÁLISE DA ABUNDÂNCIA DE INVERTEBRADOS EDÁFICOS EM DIFERENTES AGROECOSSISTEMAS

Luciane da Silva Oliveira[1]
Paulo Roberto Cecon[2]
Sebastião Martins Filho[3]
Ivo Jucksch[4]

RESUMO

A regressão logística é o método estatístico de análise utilizado com a finalidade de verificar a relação entre uma variável resposta dicotômica e variáveis explicativas de interesse. Este trabalho teve como objetivo realizar um estudo sobre os fatores que influenciam a abundância de invertebrados, indicadores do tipo de uso e qualidade do solo, sob diferentes formas de manejo utilizando a Regressão Logística. Para seleção de covariáveis foi utilizada a proposta de Collett (1994) e foram apresentados estimadores dos parâmetros envolvidos em cada modelo e suas interpretações, propriedades estatísticas e critérios para se julgar a adequabilidade dos modelos selecionados. A metodologia foi aplicada a dois conjuntos de dados reais (período seco e chuvoso). No modelo final ajustado para o conjunto de dados analisado no período seco verificou-se que as covariáveis *Tipo de Sistema, Cálcio em Serapilheira*, *Matéria Orgânica do Solo*, *Potássio em Serapilheira* e a iteração entre *Cálcio e Potássio em Serapilheira* foram importantes para explicar a presença de mais de 9 indivíduos, em média, no solo. No modelo final ajustado para o período chuvoso, as covariáveis significativas para explicar a presença de 101 indivíduos, em média, no solo foram *Magnésio em Serapilheira*, *Carbono Orgânico Total na Serapilheira*, *Matéria Orgânica da Serapilheira* e *Temperatura Ambiente*.

Palavras-chave: Análise de Regressão, Regressão, Solos – Manejo.

1 INTRODUÇÃO

A regressão logística é o método estatístico usual de análise utilizado com a finalidade de verificar a relação entre uma variável resposta dicotômica e variáveis explicativas de interesse. A análise logística controla grande número de variáveis simultaneamente, permitindo que os dados sejam utilizados mais eficientemente.

Na regressão logística a variável resposta (*Y*), geralmente binária ou dicotômica, apresenta duas possibilidades de resposta (sucesso ou fracasso). Ao "sucesso", resultado mais importante ou aquele que se relaciona o acontecimento de interesse, geralmente atribui-se o valor 1 (y = 1), e ao resultado complementar "fracasso" o valor 0 (y = 0).

[1] Professora e coordenadora do Curso de Matemática – UEMG – Unidade Carangola.
[2] Professor e Chefe do Departamento de Estatística – UFV
[3] Professor e Coordenação da Comissão de Ensino do Departamento de Estatística – UFV
[4] Professor do Departamento de Solos - Centro de Ciências Agrárias - UFV

Um modelo de regressão logística prevê a probabilidade direta de um evento ocorrer e tem sido amplamente aplicado em importantes áreas como Agronomia, Biologia, Engenharia, Economia, Mineração, Transportes, Farmacologia, Medicina e nas Ciências Sociais.

Os modelos de regressão logística fazem parte da classe dos modelos lineares generalizados, ou seja, daqueles que se tornam lineares por meio da aplicação de algum tipo de transformação.

Nesse estudo verificou se a presença de invertebrados no solo é mais ou menos abundante, considerando alguns fatores como o tipo de manejo agrícola em cultivos de café, a associação entre grupos da fauna edáfica e os atributos físicos, químicos e abióticos do solo e da serapilheira.

A diversidade e a abundância da fauna invertebrada do solo, assim como a presença de determinados grupos de organismos em um sistema, podem ser usadas como indicadores eficientes da qualidade dos solos (PAOLETTI (1999); BARROS et al., 2003), entretanto, podem ser afetadas por vários fatores edáficos (tipo de solo, minerais predominantes, temperatura, pH, matéria orgânica, umidade, textura e estrutura), eventos históricos (antropogênico e geológico), topográficos e climáticos (MELO *et al.* 2009).

Os invertebrados edáficos atuam em vários processos fundamentais para a manutenção da fertilidade e qualidade dos solos de agroecossistemas e ecossistemas naturais, exercem papel central na decomposição da matéria orgânica do solo e resíduos vegetais, influenciando a disponibilidade de nutrientes (BROWN et al. 1998, HENDRIX et al., 2006 *apud* SOUZA, 2010). São capazes de melhorar a estrutura do solo pelo estabelecimento de relações com os microorganismos ou de forma direta, pela digestão, transporte e incorporação de partículas orgânicas (SILVA, 2010).

Esse trabalho teve como objetivo verificar os fatores que podem influenciar a abundância de invertebrados no solo sob diferentes formas de manejo, utilizando a Regressão Logística, aspectos de inferência e metodologia para seleção de covariáveis.

1.1 Modelo de Regressão Logística

Os métodos de regressão têm como objetivo descrever as relações entre a variável resposta (*Y*) e a variável explicativa (*X*). Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. A variável dependente *Y* assume apenas dois possíveis valores (1 ou 0), sendo $\pi_i = P(Y=1 \mid X=x_i)$ a probabilidade de "sucesso" e $1-\pi_i = P(Y=0 \mid X=x_i)$ a probabilidade de "fracasso".

O modelo de regressão logístico binário é um caso particular dos modelos lineares generalizados, mas especificamente dos modelos *logit*, nos quais a variável dependente é associada a uma variável aleatória Bernoulli. Segundo os estudos de Cox apud Hosmer e Lemeshow (1989), muitas das funções distribuições têm sido propostas, porém a função ideal para o caso da variável resposta ser dicotômica é a função logito (*logit*), pois é extremamente flexível e fácil de ser usada e interpretada.

Assim, baseada no modelo *logit,* a forma do modelo de regressão logística é dada como:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_1)}{1+\exp(\beta_0 + \beta_1 x_1)'} \tag{1}$$

No modelo de regressão logística múltipla a probabilidade de sucesso é dada por:

$$\begin{aligned} \pi_i = \pi(x_i) = P(Y=1 \mid X=x_i) &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \mathrm{K} + \beta_p x_{ip})}{1+\exp(\beta_0 + \beta_1 x_{i1} + \mathrm{K} + \beta_p x_{ip})} \\ &= \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)} \end{aligned} \tag{2}$$

e a probabilidade de fracasso por:

$$\begin{aligned} 1-\pi_i = 1-\pi(x_i) = P(Y=0 \mid X=x_i) &= \frac{1}{1+\exp(\beta_0 + \beta_1 x_{i1} + \mathrm{K} + \beta_p x_{ip})} \\ &= \frac{1}{1+\exp(x_i^T \beta)} \end{aligned} \tag{3}$$

Assume-se que Y_i tem uma distribuição de Bernoulli com parâmetro de sucesso π_i e que o "*logit*" para o modelo de regressão logística múltipla é dada pela equação:

$$g(x_i) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = x_i^T \beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{4}$$

Para a estimação dos parâmetros β no modelo de regressão logística usa-se o método dos mínimos quadrados, no qual são determinados valores minimizam a soma dos quadrados de desvios de valores observados de y_i dos valores preditos $\hat{y}_i$ (HOSMER e LEMESHOW, 1989).

Segundo Meyer (1978), o método de máxima verossimilhança conduz a estimativas razoáveis para os dados dicotômicos. Então, a função de verossimilhança é dada por:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_i) + \ln\left(\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right) \right] \tag{5}$$

$$= \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_i) - \ln(1 + \exp(\beta_0 + \beta_1 x_i)) \right]$$

Para determinar os valores de β que maximizam $l(\beta)$, deriva-se a função (5) em relação aos elementos do vetor β, que por sua vez são funções dos logaritmos presentes na equação, e assim obtendo duas equações:

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \exp(\beta_0 + \beta_1 x_i) \right] \tag{6}$$

$$\frac{\partial l(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \left[y_i x_i - \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \exp(\beta_0 + \beta_1 x_i) x_i \right] \tag{7}$$

Para o estudo em que as variáveis são dicotômicas, a interpretação dos coeficientes se dá pela razão de chance (*odds ratio*), que é a razão das proporções para os dois resultados possíveis, isto é, a razão entre sucesso (π_i) e fracasso ($1 - \pi_i$).

Devido a fácil interpretação, a razão de chance é uma medida de associação muito utilizada e possui propriedades estatísticas que a tornam fundamental em muitos tipos de estudos.

Após o ajuste do modelo (estimação dos parâmetros $\beta_i's$) deve-se testar a significância das variáveis decorrentes no modelo. Nesse processo está envolvido o teste de hipóteses estatísticas, o qual determina se as variáveis independentes no modelo estão “significantemente” relacionadas com a variável resposta.

Essa estatística é chamada de *deviance* (desvio) e avalia o valor ajustado na regressão logística, desempenhando o mesmo papel que a soma de quadrados residuais tem na regressão linear.

Considerando o modelo com as proporções estimadas $\hat{\pi}_i$, a *deviance* pode ser escrita como:

$$D = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right) \right]$$

A *deviance* sempre é positiva e quanto menor, melhor é o ajuste do modelo.

Para testar o ajuste dos coeficientes também podem ser utilizados outros métodos estatísticos semelhantes ao anterior, como: Teste de Wald e Teste de Escore.

1.1.1 Variáveis *dummy* no Modelo

Quando as variáveis explicativas categóricas possuem mais de duas categorias é necessário reparametrizá-las antes de incluí-las no modelo.

Admitindo que seja *p* o número de variáveis independentes e se a *j*-ésima variável independente da equação x_j assume k_j níveis, são feitas $k_j - 1$ variáveis indicadoras (*dummy*) para representá-la. Será denotado como D_{ju} as variáveis dummy e os coeficientes dessas variáveis como β_{ju}, onde $u = 1, 2, \mathrm{K}, k_j - 1$. E assim, a equação da transformação logarítmica assume a seguinte forma:

$$g(x_i) = \beta_0 + \beta_1 x_1 + \mathrm{K} + \sum_{u=1}^{k_j - 1} \beta_{ju} D_{ju} + \beta_p x_{p_1}$$

Existem outros métodos de codificação, é importante ressaltar que o método escolhido para codificar os dados influencia na interpretação dos resultados obtidos. Isto é, as maneiras como serão codificadas as variáveis qualitativas não afetam a probabilidade final do modelo, mas interferem na interpretação dos coeficientes obtidos.

1.1.2 Teste de Ajuste do Modelo

Para saber se o modelo selecionado explica razoavelmente bem o comportamento da variável resposta deve-se testar a qualidade do ajuste do modelo.

Pra testar esta qualidade utilizou-se o teste de Qui-Quadrado, onde as hipóteses de interesse são:

$$\begin{cases} H_0 : \pi(x) = \dfrac{e^{(\beta_0+\beta_1 x_1+\mathrm{K}+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\mathrm{K}+\beta_p x_p)}} \\ H_1 : \pi(x) \neq \dfrac{e^{(\beta_0+\beta_1 x_1+\mathrm{K}+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\mathrm{K}+\beta_p x_p)}} \end{cases}$$

Este teste consiste em agrupar, em classes, os valores ajustados $\hat{\pi}_i$ que são similares com, aproximadamente, o mesmo número de indivíduos em cada classe.

A estatística do teste será dada por:

$$\chi^2 = \sum_{j=1}^{c}\left[\sum_{k=0}^{1}\frac{\left(O_{jk}-E_{jk}\right)^2}{E_{jk}}\right]$$

onde $n=\sum_{j=1}^{c} n_j$ é tamanho da amostra, $O_{j1}=\sum_{j=1}^{n} y_j$ e $O_{j0}=\sum_{j=1}^{n}\left(1-y_j\right)=n_j-O_{j1}$ são número de casos observados e $E_{j1}=\sum_{j=1}^{c}\hat{\pi}\left(x_j\right)$ e $E_{j0}=\sum_{j=1}^{n}\left[1-\hat{\pi}\left(x_j\right)\right]$ são número de casos esperados.

A estatística segue aproximadamente uma distribuição χ^2 com $c-2$ graus de liberdade, onde *c* é o número de classes. A regra de decisão para as alternativas testadas com nível de significância igual a α é:

$$\begin{cases} \text{Rejeito } H_0, \quad \text{se } \chi^2_{cal} > \chi^2\left(1-\alpha; c-2\right) \\ \text{Não rejeito } H_0, \quad \text{se } \chi^2_{cal} \leq \chi^2\left(1-\alpha; c-2\right) \end{cases}$$

1.1.3 Análise de Resíduos do Modelo

Uma vez que o modelo foi construído, torna-se importante a análise dos resíduos dessa modelagem. Isso porque os resíduos indicam o quanto o modelo está se ajustando à amostra de dados utilizados, dado que ele é medido pela diferença entre o valor esperado e o valor observado. Apesar de não fornecer informação suficiente para que se conclua que o modelo está ajustando conforme o esperado pode detectar falha na modelagem e nos resultados encontrados.

Foi utilizada a análise do resíduo de Pearson que é dado por:

$$R_i = \sqrt{\frac{\hat{\pi}(x_j)}{1-\hat{\pi}(x_j)}}$$

Através da análise de resíduos no modelo é possível verificar as suposições de homocedasticidade e independência e a presença de pontos discrepantes.

2. MATERIAL E MÉTODOS

2.1 Área de estudo/Procedência dos dados

Os dados utilizados nesse trabalho são provenientes de um estudo realizado no Município de Araponga, Zona da Mata de Minas Gerais, dentro da microrregião de Viçosa, cedidos pelo Departamento de Solos da Universidade Federal do Ceará . A coleta de dados foi realizada em quatro propriedades rurais de agricultores familiares e comerciais desse município, onde foram coletadas amostras de solo em diferentes agroecossistemas e sistemas naturais, sendo assim descritos:

- Sistemas convencionais (SC) de café (*Coffea arábica L.*) – cultivo de café solteiro a pleno sol com uso de fertilizantes e agrotóxicos.
- Sistemas de manejo agroecológico (AGRO) - cultivo de café solteiro a pleno sol com o surgimento e manutenção de vegetação espontânea, sem uso de agrotóxicos.
- Sistemas agroflorestais (SAF) - cultivo de café consorciado com árvores frutíferas ou não, com o surgimento e manutenção de vegetação espontânea, sem uso de agrotóxicos.
- Mata nativa próximas as propriedades rurais.

A coleta do solo foi efetuada na camada de 0 – 10 cm, realizada em dois períodos do ano: – seco (Junho – Setembro) e chuvoso (Dezembro – Março), com 60 amostras em cada período (15 amostras de cada sistema).

As amostras foram selecionadas em áreas demarcadas aleatoriamente em pontos distanciados entre dez e quinze metros um do outro. Para a avaliação da macro e mesofauna edáfica, foram coletados blocos de solo de 20 x 20 x 10 cm, sendo coletada, primeiramente, a serapilheira da superfície, considerando todo resíduo vegetal sobre a superfície do solo.

Do material coletado foram extraídos os invertebrados edáficos para mensuração do número total de indivíduos por amostra e realizadas as análises físicas e químicas do solo e da serapilheira.

2.2 Construção do modelo

Buscou-se construir um modelo para determinar a quantidade média de indivíduos (invertebrados edáficos) por m^2 em diferentes condições climáticas (época seca e época chuvosa) e, bem como, analisar os fatores que podem influenciar esse resultado. Para determinar os fatores ambientais responsáveis pela ocorrência de mais ou menos indivíduos por m^2 no solo, foi utilizada a análise de regressão logística.

A variável resposta (variável dependente) analisada foi denominada como o "*número médio de indivíduos por m^2 encontrados no solo*". Foi usada a seguinte codificação para tornar a variável resposta dicotômica:

Em época seca:

- 0 para representar a presença de 9 ou menos indivíduos por m^2 no solo
- 1 para representar a presença de mais de 9 indivíduos por m^2 no solo

Em época chuvosa:

- 0 para representar a presença de 101 ou menos indivíduos por m^2 no solo
- 1 para representar a presença de mais de 101 indivíduos por m^2 no solo

Os valores 9 e 101 referem-se ao valor médio de indivíduos encontrados por m^2, nas amostras coletadas, em épocas diferentes e, serviram como valor de referência para categorização da variável resposta.

As covariáveis utilizadas na análise são apresentadas na Tabela 2 (ver matrizes de dados completas no apêndice A e B). Dentre elas, a covariável "*Tipo de Sistema*" foi recodificada antes de ser incluída no modelo por ser uma covariável categórica. Assim, a Tabela 1 ilustra a codificação através de covariáveis *dummy*.

Tabela 1 - Codificação de covariáveis *dummy* utilizadas na análise realizada no solo e em serapilheira em período seco e chuvoso

Tipo de Sistema	Variáveis *Dummy*		
	Vd1	*Vd2*	*Vd3*
Convencional	0	0	0
Mata	1	0	0
Agroecológico	0	1	0
Agro florestal	0	0	1

Tabela 2 – Código, descrição e tipo de variáveis utilizadas na análise realizada no solo e em serapilheira em período seco e chuvoso

Código	Descrição	Tipo
Vd1	*Dummy* sistema 1 (0: Convencional; 1: Mata)	Categórica
Vd2	*Dummy* sistema 2 (0: Convencional; 1: Agroecológico)	Categórica
Vd3	*Dummy* sistema 3 (0: Convencional; 1: Agroflorestal)	Categórica
V4	Nitrogênio total em serapilheira	Contínua
V5	Fósforo em serapilheira	Contínua
V6	Potássio em serapilheira	Contínua
V7	Cálcio em serapilheira	Contínua
V8	Magnésio em serapilheira	Contínua
V9	Manganês em serapilheira	Contínua
V10	Zinco em serapilheira	Contínua
V11	Ferro em serapilheira	Contínua
V12	Carbono orgânico total na serapilheira	Contínua
V13	Relação carbono/nitrogênio na serapilheira	Contínua
V14	Matéria orgânica da serapilheira	Contínua
V15	Porcentagem de umidade do solo	Contínua
V16	Temperatura ambiente	Contínua
V17	Temperatura do solo	Contínua
V18	Peso seco da serapilheira em t/há	Contínua
V19	Peso seco da serapilheira em g/kg	Contínua
V20	Microporosidade do solo	Contínua
V21	Macroporosidade do solo	Contínua
V22	Porosidade total do solo	Contínua
V23	Densidade do solo	Contínua
V24	Ph do solo	Contínua
V25	Fósforo no solo	Contínua
V26	Potássio no solo	Contínua
V27	Cálcio no solo	Contínua
V28	Magnésio no solo	Contínua
V29	Alumínio no solo	Contínua
V30	Soma de base	Contínua
V31	CTC (capacidade de troca de cátions do solo) efetiva	Contínua
V32	CTC (capacidade de troca de cátions do solo) total	Contínua
V33	Saturação de bases do solo	Contínua

V34	Saturação por alumínio	Contínua
V35	Matéria orgânica do solo	Contínua
V36	Fósforo remanescente do solo	Contínua

Para a seleção das covariáveis foi utilizado o método derivado da proposta de Collett executado com o auxílio do pacote estatístico R (*R Development Core Team*), versão 2.11.1.

2.3 Seleção de Covariáveis

Nesse estudo optou-se por utilizar uma estratégia de seleção de covariáveis derivada da proposta de Collett (1994), citado por Colosimo e Giolo (2006), em que as informações do pesquisador podem ser incluídas no processo de decisão, o que envolve uma participação mais ativa do estatístico e pesquisador em cada passo do processo de seleção, podendo, por exemplo, incluir covariáveis relevantes no estudo independente de significância estatística.

Os passos utilizados no processo de seleção são descritos como se segue:

1. Primeiramente ajustar todos os modelos contendo uma única covariável. Em seguida, incluir todas as covariáveis significativas ao nível de 0,10. Nesse passo, utilizar o teste da razão de verossimilhanças.
2. Ajustar conjuntamente as covariáveis significativas no passo 1. Em seguida ajustar modelos reduzidos, excluindo uma única covariável de cada vez, pois na presença de certas covariáveis, outras podem deixar de ser significativas. Verificar quais as covariáveis que provocaram um aumento significativo na estatística da razão de verossimilhanças. Somente aquelas que atingiram a significância devem permanecem no modelo.
3. Com as covariáveis que ficaram retidas no passo 2, ajustar um novo modelo e as covariáveis que foram excluídas no passo 2 retornaram ao modelo para confirmar se não são estatisticamente significativas.
4. Incluir ao modelo as eventuais covariáveis significativas no passo 3 juntamente com aquelas do passo 2. Neste passo retornam-se com as covariáveis excluídas no passo 1 para confirmar se elas não são estatisticamente significativas.

5. Ajustar um modelo incluindo as covariáveis significativas no passo 4 e testar se alguma delas pode ser retirada do modelo.
6. Com as covariáveis que “sobreviveram” ao passo 5, ajusta-se então o modelo final para os efeitos principais. Deve-se verificar a possibilidade de inclusão de termos de interação dupla entre as covariáveis incluídas no modelo. O modelo final será composto pelos efeitos principais identificados no passo 5 e os possíveis termos de interação significativos nesse passo.

2.4 Medidas de qualidade do ajuste

Para saber se o modelo selecionado explica razoavelmente bem o comportamento da variável resposta deve-se testar a qualidade do ajuste do modelo. Neste estudo utilizou-se o Teste de Hosmer e Lemeshow correspondente a um teste Qui-quadrado,

A estatística de teste sob a hipótese nula foi dada por:

$$\chi_{HL}^{\ 2} = \sum_{j=1}^{g} \frac{\left(o_j - \mathrm{e}_j\right)^2}{\mathrm{e}_j\left(1 - \frac{\mathrm{e}_j}{n_j}\right)} = \sum_{j=1}^{g} \frac{\left(o_j - \overline{p}_j\right)^2}{n_j \overline{p}_j\left(1 - \overline{p}_j\right)} \sim \chi^2_{g-2}$$

em que

n_j é o número de observações pertencentes ao grupo *j*, verificando-se $n = \sum_{j=1}^{g} n_j$

o_j é a frequência observada de sucesso no grupo *j*, onde $o_j = \sum_{j=1}^{n_j} y_{ij}$ e y_{ij} é a i-ésima observação do grupo *j*.

e_j é a frequência esperada de sucesso no grupo *j*, onde $\mathrm{e}_j = n_j \overline{p}_j$ e $\overline{p}_j = \frac{\sum_{j=1}^{n_j} \hat{p}_{ji}}{n_j}$

$\hat{p}_j$ é a probabilidade predita correspondente à i-ésima observação do grupo *j*.

A um nível de significância estabelecido, busca-se não rejeitar a hipótese de que não existem diferenças entre os valores preditos e observados. O critério de avaliação se distingue um pouco do convencional, pois geralmente o que se pretende é rejeitar a hipótese nula. Nesse caso, se houver diferenças significativas entre as classificações preditas pelo modelo e as observadas, então o modelo não

representa a realidade de forma satisfatória. Em tal situação, o modelo não seria capaz de produzir estimativas e classificações muito confiáveis (HOSMER e LEMESHOW, 1989).

A estatística do teste de Hosmer e Lemeshow tem distribuição qui-quadrado com g – 2 graus de liberdade, em que g = 10 grupos.

3. RESULTADOS E DISCUSSÃO

Para o conjunto de dados da serapilheira e do solo no período seco, foram ajustados todos os modelos utilizando a estratégia de seleção derivada da proposta de Collett (1994). Na etapa final chegou-se a três modelos que não apresentaram muita discrepância nos valores da estatística do Teste da Razão da Verossimilhança. Os modelos observados mostraram ter alguma influência sobre a característica avaliada, que nesse caso, referia-se a ocorrência de mais de 9 indivíduos por m^2, em média, no solo das áreas estudadas:

- Modelo 1:

$$P_{(\text{Mais de 9 indivíduos})} = \frac{e^{(\beta_0+\beta_1 Vd1+\beta_2 Vd2+\beta_3 Vd3+\beta_4 V7+\beta_5 V35+\beta_6 V6+\beta_7 Vd2{*}V6+\beta_8 V7{*}V6)}}{1+e^{(\beta_0+\beta_1 Vd1+\beta_2 Vd2+\beta_3 Vd3+\beta_4 V7+\beta_5 V35+\beta_6 V6+\beta_7 Vd2{*}V6+\beta_8 V7{*}V6)}}$$

- Modelo 2:

$$P_{(\text{Mais de 9 indivíduos})} = \frac{e^{(\beta_0+\beta_1 Vd1+\beta_2 Vd2+\beta_3 Vd3+\beta_4 V7+\beta_5 V35+\beta_6 V6+\beta_7 Vd2{*}V6)}}{1+e^{(\beta_0+\beta_1 Vd1+\beta_2 Vd2+\beta_3 Vd3+\beta_4 V7+\beta_5 V35+\beta_6 V6+\beta_7 Vd2{*}V6)}}$$

- Modelo 3:

$$P_{(\text{Mais de 9 indivíduos})} = \frac{e^{(\beta_0+\beta_1 Vd1+\beta_2 Vd2+\beta_3 Vd3+\beta_4 V7+\beta_5 V35+\beta_6 V6+\beta_7 V7{*}V6)}}{1+e^{(\beta_0+\beta_1 Vd1+\beta_2 Vd2+\beta_3 Vd3+\beta_4 V7+\beta_5 V35+\beta_6 V6+\beta_7 V7{*}V6)}}$$

Para avaliar se os modelos finais foram bem ajustados e então decidir qual deles deveria ser usado, utilizou-se o Teste Hosmer e Lemeshow, que testaram a qualidade do ajuste, avaliando a capacidade preditiva dos modelos.

A Tabela 3 exibe o resultado do teste para os três modelos e ao nível de significância de 5%, não foi possível rejeitar a hipótese nula de que não houve diferenças significativas entre os valores preditos e observados para os modelos 1 e

3, o que indicou que esses modelos foram capazes de produzir classificações confiáveis.

Tabela 3 – Teste de Hosmer e Lemeshow para o conjunto de dados da serapilheira e do solo no período seco

Modelo	Qui-quadrado	g.l.	Valor *p*
1	66,001	8	0,580
2	169,000	8	0,031
3	64,149	8	0,601

Assim, o modelo mais adequado para a análise da quantidade média de indivíduos no solo em época seca, foi o modelo 3, uma vez que ele é bem ajustado e é mais parcimonioso. O modelo final ficou composto pelas covariáveis: *Dummy* sistema 1 - Sistema Convencional/Mata (Vd1)*, Dummy* sistema 2 - Sistema Convencional/ Agroecológico (Vd2), *Dummy* sistema 3 - Sistema Convencional/ Agroflorestal (Vd3), Cálcio em serapilheira (V7), Matéria orgânica do solo (V35), Potássio em serapilheira (V6) e a iteração entre Cálcio e Potássio em serapilheira (V7*V6).

Além de obter um modelo, testar a significância de seus parâmetros e verificar a acurácia e eficiência desse modelo encontrado, outra análise interessante de ser feita é a da razão das chances, calculada por $\exp\left(\hat{\beta}\right)$. A Tabela 4 mostra os valores dessas razões para o modelo final.

Tabela 4 – Razão de chance do modelo final ajustado para o conjunto de dados da serapilheira e do solo no período seco

Variáveis	$\hat{\beta}$	Erro padrão	Valor *p*	Razão de Chance $\exp\left(\hat{\beta}\right)$
Constante	0,852	4,270	0,842	
Vd1	3,800	1,501	0,011	44,688
Vd2	0,190	1,265	0,880	1,210
Vd3	-0,022	1,258	0,986	0,979
V7	-0,730	0,323	0,024	0,482
V35	0,431	0,169	0,011	1,538
V6	-0,903	0,858	0,293	0,405
V7*V6	0,118	0,070	0,092	1,125

Pode-se observar que dentre os fatores que influenciam a presença de mais ou menos indivíduos por m^2 no solo nas áreas estudadas, o *cálcio* e o *potássio em*

serapilheira, atuaram de forma negativa, isto é, quando a quantidade desses elementos foi alta na serapilheira, as chances de aumentar o número de indivíduos no solo diminuíram. Assim, o aumento de uma unidade (em gkg-1) de cálcio e de potássio, separadamente, diminui em aproximadamente 52% e 60% respectivamente, as chances de ocorrência de mais de 9 indivíduos por m^2 no solo. Porém, a interação entre esses dois elementos no solo atuou de forma positiva. Verificou-se que com o aumento de uma unidade da iteração entre cálcio e potássio, existe a possibilidade de se aumentar em aproximadamente 13% as chances de ocorrência de mais de 9 indivíduos por m^2 no solo.

Com o aumento de uma unidade de *matéria orgânica no solo*, as chances de ocorrência de mais de 9 indivíduos por m^2 no solo aumentam em 54% aproximadamente. Em relação ao Sistema Convencional, categoria de referência utilizada na codificação das variáveis *dummies*, o Sistema "Mata" aumenta aproximadamente 45 vezes a chance de se encontrar mais de 9 indivíduos por m^2 no solo das áreas estudadas, em época seca. Os Sistemas Agroflorestal e Agroecológico não apresentaram significância estatística em relação ao Sistema Convencional.

Para o conjunto de dados da Serapilheira e do solo no período chuvoso, no último passo da seleção, quatro covariáveis foram selecionadas para o modelo final. Para completar a modelagem foi verificada a possibilidade de inclusão de termos de iteração dupla entre as covariáveis já incluídas no modelo. Nenhuma iteração foi significativa ao nível de 0,10.

Desta forma, o modelo final para a estimativa da probabilidade de ocorrência de mais de 101 indivíduos por m^2 no solo, nas áreas estudadas foram:

$$P_{(\text{Mais de } 101 \text{ indivíduos})} = \frac{e^{(\beta_0+\beta_1 V8+\beta_2 V12+\beta_3 V14+\beta_4 V16)}}{1+e^{(\beta_0+\beta_1 V8+\beta_2 V12+\beta_3 V14+\beta_4 V16)}}$$

Pelo teste de Hosmer e Lemeshow, ao nível de significância de 5%, não foi possível rejeitar a hipótese nula de que não houve diferenças significativas entre os valores preditos e observados, indicando que o modelo foi capaz de produzir classificações confiáveis, isto é, o grau de acurácia do modelo é bom (Tabela 5).

Tabela 5 – Teste de Hosmer e Lemeshow no modelo ajustado para o conjunto de dados da serapilheira e do solo no período chuvoso

Qui-quadrado	g.l.	Valor *p*
5,7390	8	0,676

Como foi verificado, o modelo final foi bem ajustado e ficou composto pelas covariáveis: Magnésio em serapilheira (V8), Carbono orgânico total na serapilheira (V12), Matéria orgânica da serapilheira (V14) e Temperatura ambiente (V16).

Pela análise da razão das chances **(**Tabela 6), observou-se dentre os fatores que influenciam a presença de mais ou menos indivíduos por m^2 no solo nas áreas estudadas, o *magnésio* e o *carbono orgânico total na serapilheira*, atuaram de forma positiva, isto é, quando a quantidade desses elementos for alta na serapilheira, as chances de aumentar o número de indivíduos no solo também aumentam. Assim, o aumento de uma unidade de magnésio e de carbono orgânico total na serapilheira aumenta em aproximadamente 16 e 98 vezes, respectivamente, as chances de ocorrência de mais de 101 indivíduos por m^2 no solo.

Tabela 6 – Razão de chance do modelo final ajustado para o conjunto de dados da serapilheira e do solo no período chuvoso

Variáveis	$\hat{\beta}$	Erro Padrão	Valor *p*	Razão de Chance $\exp(\hat{\beta})$
Constante	25,507	8,908	0,004	
V8	2,781	0,878	0,002	16,135
V12	4,588	2,449	0,061	98,268
V14	-2,736	1,379	0,047	0,065
V16	-0,834	0,271	0,002	0,434

Com o aumento de uma unidade de *matéria orgânica na serapilheira*, as chances de ocorrência de mais de 101 indivíduos por m^2 no solo diminuem em 94% aproximadamente, e se houver o aumento de um grau na *temperatura ambiente* em relação à temperatura ambiente registrada na época da pesquisa, pode ocorrer uma diminuição de aproximadamente 57% da chance de se encontrar mais de 101 indivíduos por m^2 no solo das áreas estudadas, em época chuvosa.

Sendo a matéria orgânica composta por carbono orgânico, observa-se que o carbono orgânico na serapilheira contribui de forma positiva para o aumento de indivíduos no solo, enquanto que a matéria orgânica na serapilheira contribui de forma negativa para esse aumento. Essa incoerência pode ter ocorrido devido o

experimento ser realizado em ambiente não controlado e com pequeno número de repetição.

A covariável "*matéria orgânica*" esteve presente no modelo final para o período seco e chuvoso. No período seco, foi possível observar que quanto maior a quantidade de matéria orgânica no solo, maior sua contribuição para o aumento da quantidade de invertebrados edáficos, porém, no período chuvoso uma grande quantidade de matéria orgânica na serapilheira contribui para a diminuição do número de invertebrados no solo.

Sendo a matéria orgânica composta por carbono orgânico, foi observada certa incoerência no período chuvoso – o carbono orgânico na serapilheira contribui de forma positiva para o aumento de indivíduos no solo, enquanto que a matéria orgânica na serapilheira contribui de forma negativa para esse aumento, isto é, contribui para a diminuição.

Somente o modelo final para o período seco apresentou iterações significativas entre covariáveis. A interação entre Cálcio e Potássio contribui em aproximadamente 13% para o aumento do número de indivíduos por m^2 no solo em período seco.

A temperatura ambiente só foi relevante no modelo final para o período chuvoso onde é maior a concentração de indivíduos por m^2, sendo que se seu valor for elevado, aumenta-se a chance de diminuir a quantidade e a variedade de invertebrados no solo.

4. CONCLUSÕES

Das trinta e seis covariáveis iniciais utilizadas para estudar o problema, somente quatro foram importantes no modelo final, tanto para o período seco quanto para o período chuvoso. No período seco, por causa das covariáveis *Dummys* foram incluídos os outros níveis da covariável "*Tipo de sistema*" e uma interação significativa entre Cálcio e Potássio em serapilheira.

O desempenho discriminatório do modelo final para cada período analisado (seco e chuvoso), avaliado pelo teste de Hosmer e Lemeshow não apresentou diferença estatisticamente significativa entre os valores observados e preditos.

Observou-se a necessidade de se desenvolver outras pesquisas com aplicação dessa mesma metodologia, com os ajustes necessários a cada caso, em outras localidades e com um maior número de repetições, a fim de ampliar os conhecimentos e melhor compreender a relação entre a variável resposta e as demais covariáveis do modelo, buscando resultados mais representativos. Além disso, deve-se testar interações entre outras covariáveis e verificar a significância dessas interações no modelo.

REFERÊNCIAS BIBLIOGRÁFICAS

BARROS, E.; NEVES, A.; BLANCHART, E.; FERNANDES, E.C.; WANDELLI, E.; LAVELLE, P. Development of the soil macrofauna community under silvopastoral and agrosilvicultural systems in Amazonia. *Pedobiologia*, Jena, v.47, p.273-280, 2003.

BROCCO, J. B. *Ponderação de modelos com aplicação de Regressão Logística binária.* 2006. 78 f. Dissertação (Mestrado) - Universidade Federal do São Carlos. 2006.

CASELLA, G.; BERGER, R. L. *Statistical Inference*. 2nd ed. Pacific Grove: Duxbury, 2002.

COLLETT, D. *Modelling Survival Data in Medical Research.* Chapman and Hall, London, 1994.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. ABE - Projeto Fisher. São Paulo: Edgar Blücher, 2006.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. 2nd ed. Massachusetts: John Wiley & Sons, 2000.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. Massachusetts: John Wiley & Sons, 1989.

MELO, F. V. et al. A importância da meso e macrofauna do solo na fertilidade e como bioindicadores. *Boletim Informativo da Sociedade Brasileira de Ciência do Solo*, 34(01), 38-43. Jan/abr. 2009.

MEYER, P. L. *Probabilidade*: aplicações à estatística. Rio de Janeiro: Livros Técnicos e Científicos, 1978.

PAOLETTI, M.G. The role of earthworms for assessment of sustainability and as bioindicators. *Agriculture, Ecosystems and Environment*, Amsterdam, v.74, p.137-155, 1999.

R Development Core Team 2009. *R*: a language and environment for statistical computing.Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: Nov. 2010.

SILVA, J. *Invertebrados Edáficos em Sistemas de Produção com Café na Zona da Mata de Minas Gerais.* 2010. 218 f. Tese (Doutorado) – Universidade Federal do Ceará. 2010.

SOUZA, M. E. P. *Oligochaetas em solos sob sistemas de manejos a pleno sol e agroflorestal e vermicompostagem associada com pós de rochas.* 2010. 58 f. Dissertação (Mestrado) - Universidade Federal de Viçosa, 2010.

WAGNER, M. B.; MOTTA**,** V. T.; DORNELLES, C. *SPSS passo a passo*: Statistical Package for the Social Sciences. Caxias do Sul: Educs, 2004. 172p.