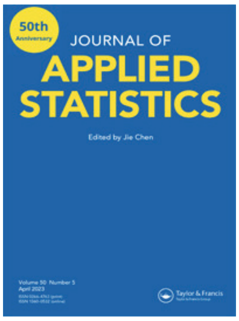


Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cjas20

A random effect regression based on the odd log-logistic generalized inverse Gaussian distribution

J. C. S. Vasconcelos, G. M. Cordeiro, E. M. M. Ortega & G. O. Silva

To cite this article: J. C. S. Vasconcelos, G. M. Cordeiro, E. M. M. Ortega & G. O. Silva (2023) A random effect regression based on the odd log-logistic generalized inverse Gaussian distribution, Journal of Applied Statistics, 50:5, 1199-1214, DOI: 10.1080/02664763.2021.2024515

To link to this article: https://doi.org/10.1080/02664763.2021.2024515

Published online: 06 Jan 2022.

JOURNAL OF APPLIED STATISTICS
2023, VOL. 50, NO. 5, 1199–1214
https://doi.org/10.1080/02664763.2021.2024515

APPLICATION NOTE

A random effect regression based on the odd log-logistic generalized inverse Gaussian distribution

J. C. S. Vasconcelos [a], G. M. Cordeiro [b], E. M. M. Ortega [a] and G. O. Silva [c]

[a]ESALQ, Universidade de São Paulo, Piracicaba, Brazil; [b]UFPE, Universidade Federal de Pernambuco, Recife, Brazil; [c]UFBA, Universidade Federal da Bahia, Salvador, Brazil

ABSTRACT

In recent decades, the use of regression models with random effects has made great progress. Among these models' attractions is the flexibility to analyze correlated data. In various situations, the distribution of the response variable presents asymmetry or bimodality. In these cases, it is possible to use the normal regression with random effect at the intercept. In light of these contexts, i.e. the desire to analyze correlated data in the presence of bimodality or asymmetry, in this paper we propose a regression model with random effect at the intercept based onthe generalized inverse Gaussian distribution model with correlated data. The maximum likelihood is adopted to estimate the parameters and various simulations are performed for correlated data. A type of residuals for the new regression is proposed whose empirical distribution is close to normal. The versatility of the new regression is demonstrated by estimating the average price per hectare of bare land in 10 municipalities in the state of São Paulo (Brazil). In this context, various databases are constantly emerging, requiring flexible modeling. Thus, it is likely to be of interest to data analysts, and can make a good contribution to the statistical literature.

ARTICLE HISTORY
Received 10 October 2020
Accepted 27 December 2021

KEYWORDS
Bimodal data; correlated data; hectare price data; regression model; simulation study

1. Introduction

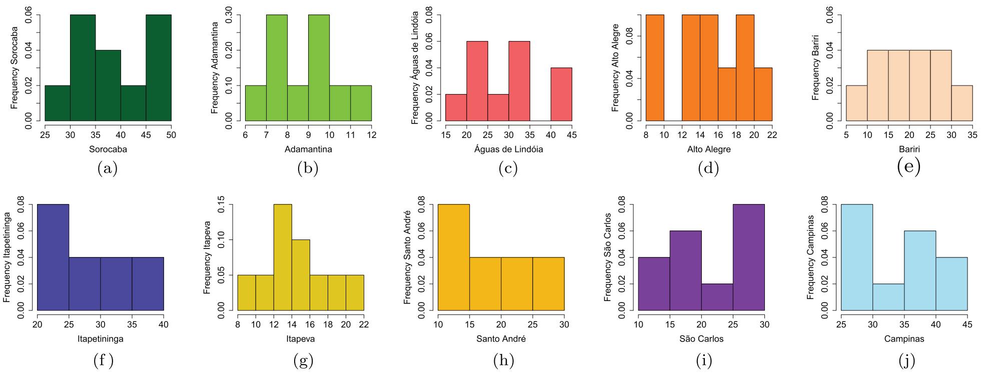
Many studies in the fields of public health, economics, agronomy, medicine, biology and the social sciences, among others, involve repeated observations of a response variable. The expression 'repeated measures' is used to designate measures obtained for the same variable or in the same experimental unit on more than one occasion; see [2,3]. Various experimental designs with repeated measures are common, such as split-plot, crossover and longitudinal. These types of investigations are referred to as correlated data studies, and they play a fundamental role in the analysis of results where it is possible to characterize alterations in the characteristics of an individual by associating these variations with a set of covariables. Due to their nature the repeated measures have a correlation structure that plays an important role in the analysis of these types of data. Besides, the distribution of the response variable can present asymmetry or bimodality.

CONTACT J. C. S. Vasconcelos ✉ juliocezarvasconcelos@hotmail.com

Recently, some works in this area were developed. For example, [8] developed random effect parametric and nonparametric regressions for analyze cognitive test data, [1] reported advances in statistical modeling in linguistics based in linear mixed-effects regressions, [6] presented an analysis of microbiome relative abundance data using a zero-inflated beta GAMLSS and meta-analysis across studies using random effects models, [5] introduced a random effect log-Burr XII regression and [4] presented host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation.

Figures 1 and 2 display the average price per hectare of bare land in 10 cities in the state of São Paulo, Brazil, where bare land consists of the soil and its surface with the respective vegetation, such as forest or pasture. These data were obtained from the website of the Instituto de Economia Agrícola (IEA)[1] and the Coordenadoria de Assistência Técnica Integral (CATI) referring to 2015. In this case, each city is interpreted as a group, and the data within each city are correlated, while between cities they are considered independent.

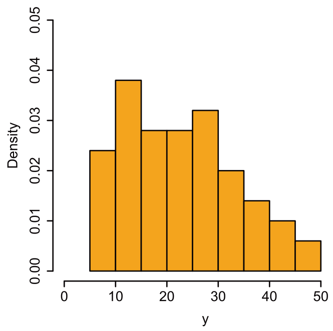
Figure 1 presents the distribution of the average price per hectare of land in each city. Note that in the cities of Sorocaba, Adamantina, Águas de Lindóia, São Carlos and Campinas, the data have bimodal distribution. Figure 2 covers the complete sample, and the data have bimodal and asymmetric distribution.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 1. Frequencies of the average price per hectare in each city.



(a)

Figure 2. Histograma of the average price per hectare of raw land.

So, to analyze correlated data in the presence of bimodality and asymmetry, and based on the studies described, it is introduced a regression with normal random intercepts based on the *odd log-logistic generalized inverse Gaussian* (OLLGIG) distribution for the purpose of considering the possible presence of heterogeneity among the cities.

The remainder of this paper is divided into four sections. In Section 2, the random effect OLLGIG regression is defined. In Section 3, the maximum likelihood estimators (MLEs) are obtained via numerical integration method, some simulations are performed and the quantile residuals are defined. In Section 4, a real data set is analyzed for illustrative purposes. Finally, some conclusions are offered in Section 5.

2. The random effect OLLGIG regression

Many generalized *inverse Gaussian* (IG) distributions aim to provide better fits to certain data sets than the traditional two or three parameter IG models. The *generalized inverse Gaussian* (GIG) distribution with three parameters [7] presents several properties and applications of this distribution. Good properties and being more flexible, the GIG distribution still did not have the appropriateness to model bimodal data. In this context, [9] introduced a new generalization of the GIG distribution called the *odd log-logistic generalized inverse Gaussian* (OLLGIG) distribution with four parameters. The most important feature of this OLLGIG distribution is that it can model bimodal data.

The cumulative distribution function (cdf) of the OLLGIG model is

$$F(y) = F(y; \mu, \sigma, \nu, \tau) = \frac{G_{\mu,\sigma,\nu}(y)^{\tau}}{G_{\mu,\sigma,\nu}(y)^{\tau} + [1 - G_{\mu,\sigma,\nu}(y)]^{\tau}}, \quad y > 0, \tag{1}$$

where

$$G_{\mu,\sigma,\nu}(y) = \int_0^y \left(\frac{b}{\mu}\right)^{\nu} \frac{t^{\nu-1}}{2K_\nu(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{bt}{\mu} + \frac{\mu}{b\,t}\right)\right] \mathrm{d}t \tag{2}$$

is the cdf of the GIG distribution, $\mu > 0$ represents the location, $\sigma > 0$ is a scale parameter, and $\nu \in \mathbb{R}$ and $\tau > 0$ are shape parameters,

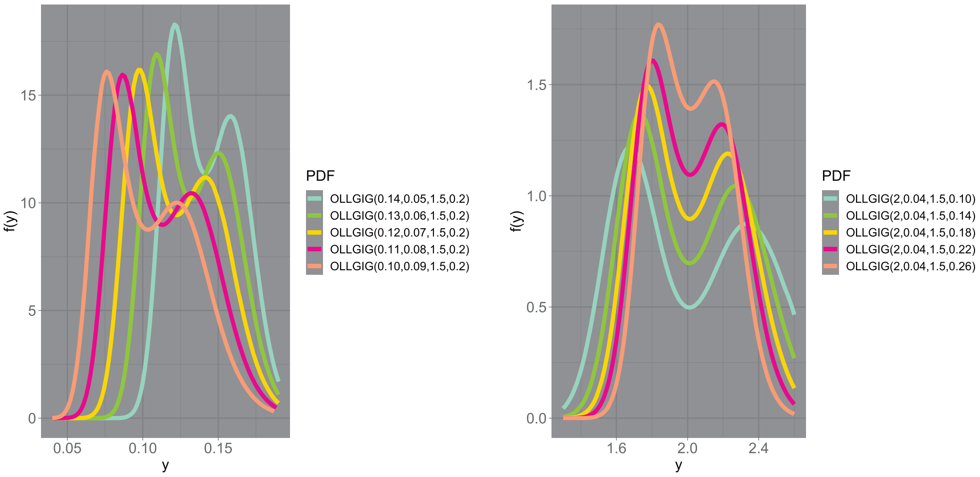
$$b = \frac{K_{\nu+1}(\sigma^{-2})}{K_\nu(\sigma^{-2})} \quad \text{and} \quad K_\nu(t) = \frac{1}{2}\int_0^\infty u^{\nu-1} \exp\left[-\frac{1}{2}t(u + u^{-1})\right] \mathrm{d}u \tag{3}$$

is the modified Bessel function of the third kind and index ν. Clearly, $G_{\mu,\sigma,\nu}(y)$ follows from (1) if $\tau = 1$. Further details and properties of the GIG distribution can be found in [7].

If $\eta(y) = G_{\mu,\sigma,\nu}(y)$, the OLLGIG density function (for $y > 0$) can be expressed as

$$f(y) = f(y; \mu, \sigma, \nu, \tau) = \left(\frac{b}{\mu}\right)^{\nu} \frac{\tau y^{\nu-1}}{2K_\nu(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{by}{\mu} + \frac{\mu}{by}\right)\right]$$
$$\times \{\eta(y)[1 - \eta(y)]\}^{\tau-1}\{\eta(y)^{\tau} + [1 - \eta(y)]^{\tau}\}^{-2}. \tag{4}$$

Figure 3 displays plots of the density function 4 for some parameter values thus showing that the OLLGIG distribution could be very flexible for modeling bimodal data.



(a) (b)

Figure 3. Plots of the OLLGIG density.

The quantile function (qf) corresponding to (1) has the simple form

$$y = Q_{GIG}\left(\frac{u^{1/\tau}}{u^{1/\tau} + [1-u]^{1/\tau}}\right), \tag{5}$$

where $Q_{GIG}(u) = G^{-1}_{\mu,\sigma,\nu}(u)$ is the qf of the GIG distribution and $u \sim \text{Uniform}(0, 1)$.

Consider a sample divided into N groups and Y_{ij} (for the jth individual in the ith group, $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$) be independent random variables having the OLLGIG distribution. Each group has random effects W_i represented by independent and identically distributed random variables with density $g(w_i; \mathbf{V})$ and variance σ_w^2, where $\mathbf{V}$ is a vector of unknown parameters. By assuming that the random effects are unobserved random variables, the regression for correlated data can be expressed as

$$\mu_{ij} = \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta} + w_i), \tag{6}$$

where $\mathbf{x}_{ij}^T = (x_{ij1}, \ldots, x_{ijp})$ is the $p \times 1$ vector of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown parameters, and w_i represents the random effects associated with the ith group.

Further, assume that $\text{Cov}(W_i, Y_{ij}) = 0$ and that, conditioned on the random effects W_i, the response variables within the group i are independent with variance σ_w^2. So, the regression can be reduced to the classical regression when $\sigma_w^2 = 0$. For the random effect regression (6), the following assumptions hold:

- $y_{ij} \mid w_i \sim \text{OLLGIG}(\mu_{ij}, \sigma, \nu, \tau)$, and marginal pdf

$$f(y_{ij}|w_i) = \left(\frac{b}{\mu_{ij}}\right)^{\nu} \frac{\tau y_{ij}^{\nu-1}}{2K_\nu(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{by_{ij}}{\mu_{ij}} + \frac{\mu_{ij}}{by_{ij}}\right)\right]$$
$$\times \{\eta(y_{ij})[1-\eta(y_{ij})]\}^{\tau-1}\{\eta(y_{ij})^\tau + [1-\eta(y_{ij})]^\tau\}^{-2}, \tag{7}$$

where

$$\eta(y_{ij}) = \int_0^{y_{ij}} \left(\frac{b}{\mu}\right)^{\nu} \frac{t^{\nu-1}}{2K_\nu(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{bt}{\mu} + \frac{\mu}{bt}\right)\right] \mathrm{d}t.$$

- The random variables $W_i \sim \mathrm{N}(0, \sigma_w^2)$ (for $i = 1, \ldots, N$) have density

$$g(w_i; \mathbf{V}) = \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{w_i^2}{2\sigma_w^2}\right), \quad w_i \in \mathbb{R}. \tag{8}$$

The variance of W_i is $Var(W_i) = \sigma_w^2$. In this case, the parameter vector is $\mathbf{V} = \sigma_w^2$.

3. Estimation, simulations and residuals

The estimates of the parameters of the random effect OLLGIG regression are calculated via maximum likelihood. For each group i, the vector of the response variable is represented by $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{in_i})^{\mathrm{T}}$. The likelihood function conditional on the random effects (independence within the group) for the individuals of the ith group is

$$L_i(y_{ij} \mid w_i) = \prod_{j=1}^{n_i} f(y_{ij}|w_i), \tag{9}$$

where $f(y_{ij}|w_i)$ is the density (7). By assuming that the terms W_i and Y_{ij} are independent random variables, the contribution of the ith group to the marginal likelihood function is

$$\int L_i(y_{ij} \mid w_i) g(w_i; \sigma_{w^2})\, \mathrm{d}w_i,$$

where $g(\cdot)$ is the random effect density (8) and $L_i(y_{ij}|w_i)$ is given by (9).

Hence, under the assumption of independence between the N groups, the marginal likelihood function for the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma, \nu, \tau, \sigma_{w^2})^\top$ reduces to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij} \mid w_i) g(w_i, \sigma_{w^2})\, \mathrm{d}w_i. \tag{10}$$

Let $(y_{11}, \mathbf{x}_{11}), \ldots, (y_{1n_1}, \mathbf{x}_{1n_1}), \ldots, (y_{N1}, \mathbf{x}_{N1}), \ldots, (y_{Nn_N}, \mathbf{x}_{Nn_N})$ be $n = n_1 + \cdots + n_i$ observations, where y_{ij} is the response variable and $\mathbf{x}_{ij}$ is the vector of covariates associated with the jth observation of the ith group. Then, assuming the normal distribution (8) for the random effects and that Y is a random variable having the OLLGIG density (7), the logarithm of the marginal likelihood function (10) can be expressed as

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log\left\{\int_{-\infty}^{+\infty} \prod_{j=1}^{n_i} \left(\frac{b}{\mu_{ij}}\right)^{\nu} \frac{\tau y_{ij}^{\nu-1}}{2K_\nu(\sigma^{-2})} \exp\left[-\frac{1}{2\sigma^2}\left(\frac{by_{ij}}{\mu_{ij}} + \frac{\mu_{ij}}{by_{ij}}\right)\right]\right.$$
$$\left. \times \prod_{j=1}^{n_i} \frac{\{\eta(y_{ij})[1-\eta(y_{ij})]\}^{\tau-1}}{\{\eta(y_{ij})^\tau + [1-\eta(y_{ij})]^\tau\}^2 \sqrt{2\pi}\sigma_w} \exp\left(-\frac{1}{2}\frac{w_i^2}{\sigma_w^2}\right) \mathrm{d}w_i\right\}. \tag{11}$$

The MLE $\widehat{\boldsymbol{\theta}}$ of the vector of parameters can be calculated by maximizing the log-likelihood (11) using the `gamlss` package [10] in **R** software. Initial values for $\boldsymbol{\beta}$, σ, ν and σ_w can be obtained from the fit of the GIG(μ,σ, ν) regression model.

3.1. Simulation study

The quality of the MLEs of the parameters for the random effect OLLGIG regression is investigated via Monte Carlo simulations. One thousand replicates were performed for two groups ($N = 10$ and $N = 20$) with different sizes ($n_i = 5, 25$ and $70, i = 1, \ldots, N$). A sample size n_i was generated for each replication from the OLLGIG($\mu_{ij}, \sigma, \nu, \tau$) distribution with $\nu = 1.5$ fixed under the configurations: $\sigma = 0.3$ and $\tau = 0.6$. For the parameters in μ_{ij}, the following values were taken: $\beta_0 = 0.15$ and $\beta_1 = 0.6$, and for the variance component $\sigma_w = 0.2$ (for $N = 10$) and $\sigma_w = 0.5$ (for $N = 20$). So, the parameter μ_{ij} has the systematic component $\mu_{ij} = \exp[(\beta_0 + w_i) + \beta_1 x_{ij1}]$.

The response variable, the random effects and the explanatory variable were generated as:

- $y_{ij} \sim$ OLLGIG($\mu_{ij}, \sigma, \nu, \tau$);
- $W_i \sim$ Normal($0, \sigma_w^2$);
- $x_{ij} \sim$ Bernoulli(0.5).

Based on the results of the two scenarios ($\sigma_w = 0.2$, $N = 10$) and ($\sigma_w = 0.5$, $N = 20$) given in Table 1, it is noted that the MSEs decrease when n increases (as expected).

3.2. Residual analysis

For the new random effect regression, the quantile residuals (qrs) have the form

$$\widehat{qr}_{ij} = \Phi^{-1}\left\{\frac{\hat\eta^{\hat\tau}(y_{ij})}{\hat\eta^{\hat\tau}(y_{ij}) + [1 - \hat\eta(y_{ij})]^{\hat\tau}}\right\}, \tag{12}$$

where

$$\eta(y_{ij}) = \int_0^{y_{ij}} \left(\frac{\hat b}{\hat\mu_{ij}}\right)^{\hat\nu} \frac{t^{\hat\nu-1}}{2\hat K_\nu(\hat\sigma^{-2})} \exp\left[-\frac{1}{2\hat\sigma^2}\left(\frac{\hat b\, t}{\hat\mu_{ij}} + \frac{\hat\mu_{ij}}{\hat b t}\right)\right] \mathrm{d}t,$$

$$\hat b = \hat K_{\nu+1}(\hat\sigma^{-2})/\hat K_\nu(\hat\sigma^{-2}) \quad \text{and} \quad \hat K_\nu(t) = \frac{1}{2}\int_0^\infty u^{\hat\nu-1} \exp\left[-\frac{1}{2}t(u + u^{-1})\right] \mathrm{d}u,$$

and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cdf.

Table 1. Results of the simulation study: Scenario 1: ($\sigma_w = 0.2, N = 10$). Scenario 2 ($\sigma_w = 0.5, N = 20$).

		$n_i = 5$			$n_i = 25$			$n_i = 70$		
	Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
Scenario 1	β_0	0.232	0.082	0.024	0.139	−0.011	0.004	0.156	0.006	0.002
	β_1	0.624	0.024	0.019	0.595	−0.005	0.003	0.596	−0.004	0.001
	σ	0.465	0.165	0.129	0.301	0.001	0.019	0.269	−0.031	0.004
	τ	0.964	0.364	0.536	0.619	0.019	0.095	0.548	−0.052	0.023
	σ_w	0.179	−0.021	0.008	0.202	0.002	0.001	0.192	−0.008	0.000
Scenario 2	β_0	0.376	0.226	0.060	0.385	0.235	0.057	0.264	0.114	0.013
	β_1	0.600	0.000	0.008	0.598	−0.002	0.002	0.600	0.000	0.001
	σ	0.348	0.048	0.046	0.288	−0.012	0.003	0.278	−0.022	0.002
	τ	0.714	0.114	0.212	0.585	−0.015	0.023	0.561	−0.039	0.009
	σ_w	0.284	−0.216	0.049	0.597	0.097	0.010	0.556	0.056	0.003

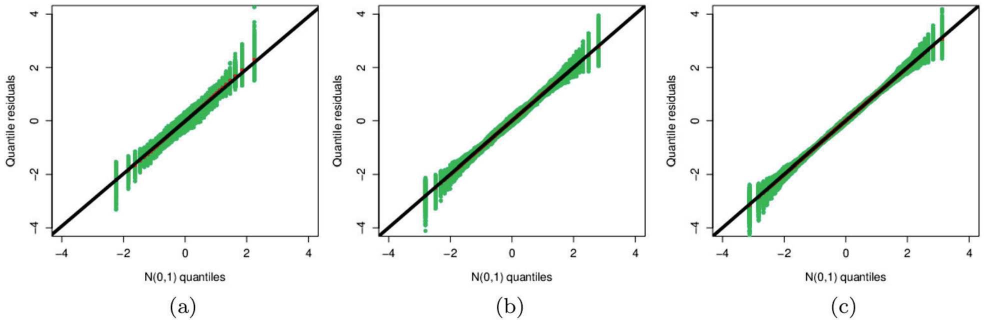


Figure 4. Normal probability plots for the qrs ($N = 10$) with $n_i = 5$, $n_i = 25$ and $n_i = 70$.

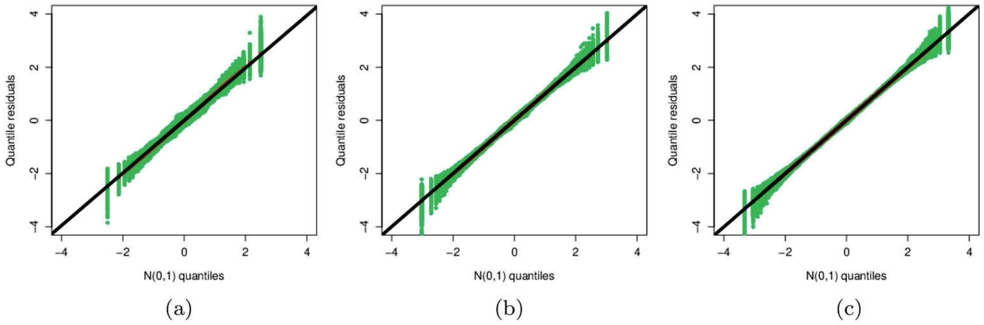


Figure 5. Normal probability plots for the qrs ($N = 20$) with $n_i = 5$, $n_i = 25$ and $n_i = 70$.

Envelopes can be constructed from these residuals to provide better interpretation of the probability normal plots. The majority of the residuals will be randomly distributed within these bands if the regression is well-fitted.

Simulation study for the residuals. A simulation study is conducted to investigate the behavior of the empirical distribution of the residuals in (12). On thousand samples are generated via the algorithm described in Section 3.1. The normal probability plots are obtained for testing the normality of the residuals.

The residuals $\widehat{qr}_{ij}$ in (12) are calculated for each fitted regression. Figures 4 and 5 display the plots of the ordered residuals versus the expected values of the normal order statistics. These plots reveal that the empirical distribution of the qrs agrees with the standard normal distribution when n_i increases.

4. Application: hectare price data

An application is now provided using the `gamlss` package [10] in the **R** software to explain the average price per hectare of bare land in ten cities in the state of São Paulo (Brazil). These data come from the website of the Instituto de Economia Agrícola (IEA)[2] and the Coordenadoria de Assistência Técnica Integral (CATI) and refer to the two halves of 2015. The bare land price is defined as the commercial value of the land after deducting the value of structures, installations and other improvements, such as: buildings, storage sheds, barns and worker housing; stables, corrals, water pipes/hoses, aviaries, pigpens and other installations for shelter or management of animals; yards and similar areas for drying of agricultural products; rural electrification equipment; groundwater catchment and other installations for supply or distribution of water, including dams and tanks; fences; other improvements not related to rural activity; as well as perennial and temporary crops, cultivated and improved natural pastures; and planted forests.

The random effect OLLGIG regression is utilized to explain how the land category of ten cities (Sorocaba, Adamantina, Águas de Lindóia, Alto Alegre, Bariri, Itapetininga, Itapeva, Santo André, São Carlos and Campinas) influence the average land price per hectare. A brief description of each city is now described. Sorocaba is one of the best cities in Brazil for investment in new start-ups as well as to live. It is near the capital, São Paulo. Adamantina has an area of approximately 411 square kilometers and is known for farming and stock breeding, and rural life in general. Águas de Lindóia is the capital of Brazil regarding hot springs, the reason why it is a cornerstone of the 'Paulista Waters Circuit'. Alto Alegre was served by Companhia Telefônica Brasileira (CTB) until 1973, and after that it was absorved into the Telecomunicações de São Paulo (TELESP), which constructed a central switching building that is still used today. In 1998, the company was sold to Telefônica as part of the privatization program, and in 2012 the company adopted the Vivo brand for fixed and cellular telephone operations. Bariri has a mixed industrial and agricultural base, in the latter case mostly sugarcane growing. Itapetininga is a large producer of corn, soybeans, oranges, milk and beef, as well as resins. Itapeva is an important producer of ores, especially phyllite, and also is among the leading municipalities in the state in the production of grain crops, besides having extensive reforested areas. Santo André has predominantly Atlantic Forest vegetation, mainly in parks and environmental preservation areas. São Carlos is an important regional industrial center. Finally, Campinas is the state's largest city other than the capital, with a strong base of high-tech companies and educational institutions.

For this study, the variables are:

- y_{ij}: average price (R$) of a hectare of bare land (this variable was divided by 1000);
- x_{ij1}: land categories (field land, primary cropland, secondary cropland, pasture land, reforestation land) (for $i = 1, \ldots, 10, j = 1, \ldots, n_i$).

Table 2 lists the averages and standard deviations (SDs) of the prices per hectare of bare land for each land category. The maximum price refers to the primary cropland, whereas the minimum price refers to the field land. The histogram of the average price per hectare (y_{ij}) in Figure 1 (Section 1) shows the presence of bimodality. So, for the marginal analysis, the OLLGIG distribution is capable to model these data.

Table 2. Averages and SDs for hectare price data.

Land category	Average	SD
Field land	16.846	7.691
Primary cropland	29.971	11.829
Secondary cropland	25.656	10.803
Pasture land	22.810	9.003
Reforestation land	18.754	7.689

Table 3. Results from the fitted densities.

Distribution	$\log(\mu)$	$\log(\sigma)$	ν	τ
OLLGIG	21.132	3.700	23.118	0.324
	(0.936)	(0.931)	(13.348)	(0.144)
GIG	22.806	0.632	3.374	1
	(1.091)	(0.249)	(2.142)	(–)
IG	22.807	0.109	−0.5	1
	(1.192)	(0.008)	(–)	(–)

Table 4. Some statistical measures.

Distribution	AIC	CAIC	BIC	HQIC	A^*	W^*	KS
OLLGIG	**744.634**	**745.055**	**755.054**	**748.851**	**0.238**	**0.031**	**0.050**
GIG	748.182	748.432	755.998	751.345	0.554	0.087	0.076
IG	749.077	749.201	754.288	751.186	0.863	0.138	0.093

Table 3 gives the MLEs of the parameters and their standard errors (SEs) (in parentheses) from the fitted OLLGIG, GIG and IG distributions to the hectare prices. Table 4 lists the values of AIC (Akaike Information Criterion), CAIC (Consistent Akaike Information Criterion), BIC (Bayesian Information Criterion), HQIC (Hannan-Quinn information criterion), A^* (Anderson–Darling), W^* (Cramér-von Misses) and KS (Kolmogarov–Smirnov) for the fitted distributions. The results reveal that the OLLGIG distribution has the lowest values for these statistics, among the three. So, it could be chosen as the best distribution to explain the current data.

Likelihood ration (LR) tests for comparing the OLLGIG distribution with two special cases are reported in Table 5. The figures in this table (specially the p-values) reveal that the OLLGIG distribution provides a better fit to these data than the other two special cases.

The empirical and estimated cdfs of the OLLGIG, GIG and IG distributions are given in Figure 6(a). The histogram of the data and the fitted OLLGIG, GIG and IG densities are displayed in Figure 6(b). These plots also reveal that the OLLGIG distribution provides the best fit to the hectare price data.

Regression analysis with systematic components. Four dummy variables $d_{ij1},\ldots,d_{ij4}$ are defined since the covariable x_{ij1} has five categories of land. According to the previous analysis, the random effect OLLGIG regression has the form (for $i=1,\ldots,10$, $j=1,\ldots,n_i$)

$$\mu_{ij}=\exp(\beta_0+\beta_1 d_{ij1}+\beta_2 d_{ij2}+\beta_3 d_{ij3}+\beta_4 d_{ij4}+w_i).$$

The random effect OLLGIG, GIG and IG regressions are compared via the AIC, BIC and GD (Global Deviance) statistics in Table 6. The wider regression outperforms the GIG and IG regressions irrespective of the criteria and then it can be used effectively in the analysis of these data.

Table 5. LR tests.

Models	Hypotheses	LR statistic	p-value
OLLGIG vs GIG	$H_0:\tau=1$ vs $H_1:H_0$ is false	5.559	0.018
OLLGIG vs IG	$H_0:\tau=1,\nu=-0.5$ vs $H_1:H_0$ is false	8.444	0.015

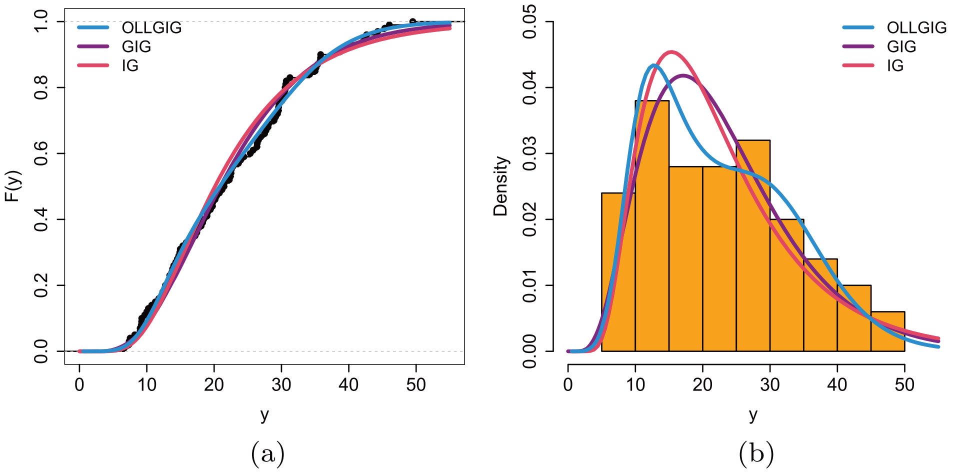


Figure 6. (a) Three estimated cdfs and empirical cdf. (b) Three estimated densities.

Table 6. Statistics.

Model	AIC	BIC	GD
OLLGIG	**524.969**	**568.186**	**491.790**
GIG	528.885	570.248	497.131
IG	552.789	591.419	523.134

Table 7. Results from the fitted random effect OLLGIG regression.

	Parameter	MLE	SE	*p*-value
Intercept	β_0	2.902	0.026	< 0.001
Primary cropland	β_1	0.608	0.040	< 0.001
Secondary cropland	β_2	0.441	0.042	< 0.001
Pasture land	β_3	0.322	0.043	< 0.001
Reforestation land	β_4	0.121	0.041	0.004
	$\log(\sigma)$	−0.697	0.076	
	ν	−4.167	0.961	
	τ	3.358	0.245	
	σ_w	0.424		

The results from the random effect OLLGIG regression fitted to these data via the `gamlss` package [10] in **R** are reported in Table 7. The explanatory variable x_{ij1} at the 5% level gives a significant difference among the levels of the land category to explain the price per hectare of bare land. The estimate of the variance component σ_w is also different from zero. So, it is necessary to consider random effects in the regression. Some interpretations are addressed at the end of this section.

The random effect OLLGIG regression is compared with two special regressions via LR statistics in Table 8. The figures in this table indicate that the wider regression provides a better fit to these data than the other two regressions.

Further, the qrs are plotted in Figure 7(a). These residuals are randomized around zero, thus revealing the suitability of the regression for analyzing the hectare price data. Finally, the quality of the adjustment range of the wider regression is verified by constructing the

Table 8. LR tests.

Regressions	Hypotheses	LR statistic	*p*-value
OLLGIG vs GIG	$H_0 : \tau = 1$ vs $H_1 : H_0$ is false	5.341	0.012
OLLGIG vs IG	$H_0 : \tau = 1$ and $\nu = -0.5$ vs $H_1 : H_0$ is false	31.344	< 0.001

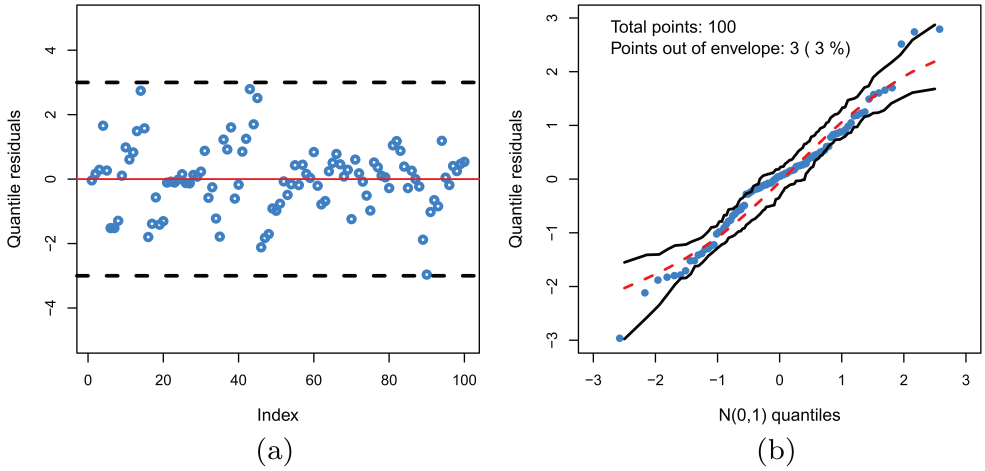


Figure 7. (a) Residual analysis of the random effect OLLGIG regression fitted to the hectare price data. (b) Normal probability plot for the qrs with envelope.

Table 9. Hypothesis testing for bare land category levels.

Hypotheses H_0	Estimate	SE	*p*-value
Primary cropland - field land = 0	0.608	0.040	< 0.001
Secondary cropland - field land = 0	0.441	0.042	< 0.001
Pasture land - field land = 0	0.322	0.043	< 0.001
Reforestation land - field land = 0	0.121	0.041	0.004
Secondary cropland - primary cropland = 0	−0.167	0.045	< 0.001
Pasture land - primary cropland = 0	−0.286	0.045	< 0.001
Reforestation land - primary cropland = 0	−0.487	0.044	< 0.001
Pasture land - secondary cropland = 0	−0.119	0.047	0.013
Reforestation land - secondary cropland = 0	−0.321	0.045	< 0.001
Reforestation land - pasture land = 0	−0.202	0.046	< 0.001

normal probability plot for the qrs with the simulated envelope. It is clear a good fitted regression shown in this envelope.

Table 9 provides the hypothesis testing to compare all levels of the covariate x_{ij1} considering the complete hectare price data. All levels of the land categories are significant (5%), and then there is strong evidence of differences between the price per hectare of bare land for each category.

The predicted random effects of the fitted regression are given in Table 10.

The hypothesis testing comparing the levels of the land category for each city separately with the associated predicted random effects are reported in the Appendix. Considering the 5% significance level, the interpretations based on these tables are addressed below.

Table 10. Prediction of the random effects.

w_i	Predicted	Cities
w_1	0.618	Sorocaba
w_2	−0.833	Adamantina
w_3	0.369	Águas de Lindóia
w_4	−0.331	Alto Alegre
w_5	−0.024	Bariri
w_6	0.139	Itapetininga
w_7	−0.339	Itapeva
w_8	−0.082	Santo André
w_9	−0.027	São Carlos
w_{10}	0.508	Campinas

- Table A1 gives the results of the comparisons between the categories of land in the city of Sorocaba in relation to the hectare price; there is no significant difference between the categories: secondary cropland - primary cropland.
- Table A2 shows that there is a difference between the categories of land: primary cropland - field land, pasture land - primary cropland and reforestation land - primary cropland in the Adamantina city in relation to the hectare price.
- Table A3 shows that there is a significant difference between all categories of land in the Águas de Lindóia city in relation to the hectare price.
- Table A4 shows that there is no significant difference between the categories of land in the Alto Alegre city in relation to the hectare price.
- Table A5 shows that there is a difference between the categories: primary cropland - field land, secondary cropland - field land, pasture land - field land, reforestation land - field land, reforestation land - primary cropland and reforestation land - secondary cropland in the Bariri city in relation to the hectare price.
- Table A6 shows that there is a significant difference between all categories of land in the Itapeniniga city in relation to the hectare price.
- Table A7 shows that there is a difference between the categories: primary cropland - field land, secondary cropland - field land, pasture land - field land, secondary cropland - primary croplands and pasture land - primary cropland in the Itapeva city in relation to the hectare price.
- Table A8 shows that there is a difference between the categories: primary cropland - field land, secondary cropland - field land, pasture land - field land, reforestation land - primary cropland, reforestation land - secondary cropland and reforestation land - pasture land in the Santo André city in relation to the hectare price.
- Table A9 shows that there is no significant difference between the categories of land in the São Carlos city in relation to the hectare price.
- Table A10 shows the results of the comparisons between the categories of land in the city of Campinas in relation to the hectare price, note that there is no significant difference between the categories: reforestation land - field land secondary cropland - primary cropland reforestation land - pasture land.

5. Conclusions

This work presents a new random effect regression based on the *odd log-logistic Generalized inverse Gaussian* distribution. The new regression is a useful extension of some regressions with random effects. For different sample sizes, a simulation study was carried out to verify the consistency of the maximum likelihood estimates of the parameters. Quantile residuals are defined for the proposed regression. The proposed regression model with random effects is suitable to model correlated data involving agricultural economics, and here it is applied to investigate the existence of statistically significant differences in land categories that influence the average price per hectare in some municipalities chosen at random in the state of São Paulo, Brazil. The proposed regression model is versatile, making it suitable for a scenario where a large variety of databases is always available, requiring flexible models. The proposed model can also be a good option to analyze correlated data when the response variable has a bimodal or asymmetric distribution.

Notes

1. See: link: http://www.iea.agricultura.sp.gov.br/out/Bancodedados.php
2. See: link: http://www.iea.agricultura.sp.gov.br/out/Bancodedados.php

Acknowledgments

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and to thank the Instituto de Economia Agrícola (IEA) and Coordenadoria de Assistência Técnica Integral (CATI), Brazil, for the availability of data.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

J. C. S. Vasconcelos http://orcid.org/0000-0001-6794-3175
G. M. Cordeiro http://orcid.org/0000-0002-3052-6551
E. M. M. Ortega http://orcid.org/0000-0003-3999-7402
G. O. Silva http://orcid.org/0000-0001-5446-380X

References

[1] C. Coupé, *Modeling linguistic variables with regression models: Addressing non-Gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape*, Frontiers in Psychology 9 (2018), p. 513.

[2] M.J. Crowder and D.J. Hand, *Analysis of Repeated Measures*, Vol. 41. CRC Press, New York, 1990.

[3] P.J. Diggle, *An approach to the analysis of repeated measurements*, Biometrics 44 (1988), pp. 959–971.

[4] S. Dirmeier, C. Dächert, M. van Hemert, A. Tas, N.S. Ogando, F. van Kuppeveld, and N. Beerenwinkel, *Host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation*, PLoS Comput. Biol. 16 (2020), p. e1007587.

[5] E.M. Hashimoto, G.O. Silva, E.M. Ortega, and G.M. Cordeiro, *Log-Burr XII gamma-Weibull regression model with random effects and censored data*, J. Stat. Theory Pract. 13 (2019), pp. 1–21.

[6] N.T. Ho, F. Li, S. Wang, and L. Kuhn, *metamicrobiomeR: An R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models*, BMC Bioinformatics 20 (2019), p. 188.

[7] B. Jørgensen, *Statistical Properties of the Generalized Inverse Gaussian Distribution*, 2nd ed. Springer, New York, 1982.

[8] G. Muniz-Terrera, A.V.D. Hout, R.A. Rigby, and D.M. Stasinopoulos, *Analysing cognitive test data: Distributions and non-parametric random effects*, Stat. Methods Med. Res. 25 (2016), pp. 741–753.

[9] J.C. Souza Vasconcelos, G.M. Cordeiro, E.M. Ortega, and E.G. Araújo, *The new odd log-logistic generalized inverse Gaussian regression model*, J. Probab. Stat. 2019 (2019), pp. 1–13.

[10] D.M. Stasinopoulos and R.A. Rigby, *Generalized additive models for location scale and shape (GAMLSS) in* ***R***, J. Stat. Softw. 23 (2007), pp. 1–46.

Appendix

Table A1. Hypothesis testing for land category levels to Sorocaba city.

Primary cropland - field land = 0	0.467	0.031	< 0.001
***Secondary cropland - field land = 0	0.384	0.032	< 0.001
Pasture land - field land = 0	0.221	0.029	0.001
Reforestation land - field land = 0	0.031	0.052	0.578
Secondary cropland - primary cropland = 0	−0.084	0.028	0.032
Pasture land - primary cropland = 0	−0.247	0.027	< 0.001
Reforestation land - primary cropland = 0	−0.436	0.050	< 0.001
Pasture land - secondary cropland = 0	−0.163	0.028	0.002
Reforestation land - secondary cropland = 0	−0.353	0.051	< 0.001
Reforestation land - pasture land = 0	−0.189	0.049	0.013

Table A2. Hypothesis testing for land category levels to Adamantina city.

Primary cropland - field land = 0	0.425	0.137	0.027
Secondary cropland - field land = 0	0.271	0.135	0.101
Pasture land - field land = 0	0.176	0.136	0.252
Reforestation land - field land = 0	0.163	0.135	0.280
Secondary cropland - primary cropland = 0	−0.154	0.091	0.149
Pasture land - primary cropland = 0	−0.249	0.093	0.044
Reforestation land - primary cropland = 0	−0.262	0.091	0.035
Pasture land - secondary cropland = 0	−0.094	0.089	0.339
Reforestation land - secondary cropland = 0	−0.108	0.087	0.272
Reforestation land - pasture land = 0	−0.013	0.089	0.889

Table A3. Hypothesis testing for land category levels to Águas de Lindóia city.

Primary cropland - field land = 0	0.759	0.059	< 0.001
Secondary cropland - field land = 0	0.555	0.033	< 0.001
Pasture land - field land = 0	0.404	0.034	< 0.001
Reforestation land - field land = 0	0.171	0.018	< 0.001
Secondary cropland - primary cropland = 0	−0.202	0.066	0.028
Pasture land - primary cropland = 0	−0.354	0.067	0.003
Reforestation land - primary cropland = 0	−0.586	0.060	< 0.001
Pasture land - secondary cropland = 0	−0.151	0.044	0.018
Reforestation land - secondary cropland = 0	−0.384	0.033	< 0.001
Reforestation land - pasture land = 0	−0.232	0.034	0.001

Table A4. Hypothesis testing for land category levels to Alto Alegre city.

Hypotheses H_0	Estimate	SE	p-value
Primary cropland - field land = 0	0.418	0.192	0.081
Secondary cropland - field land = 0	0.299	0.204	0.203
Pasture land - field land = 0	0.284	0.209	0.232
Reforestation land - field land = 0	0.025	0.171	0.891
Secondary cropland - primary cropland = 0	−0.119	0.237	0.636
Pasture land - primary cropland = 0	−0.133	0.241	0.604
Reforestation land - primary cropland = 0	−0.393	0.209	0.119
Pasture land - secondary cropland = 0	−0.014	0.251	0.957
Reforestation land - secondary cropland = 0	−0.274	0.220	0.269
Reforestation land - pasture land = 0	−0.259	0.225	0.301

Table A5. Hypothesis testing for land category levels to Bariri city.

Primary cropland - field land = 0	0.954	0.176	0.003
Secondary cropland - field land = 0	0.778	0.131	0.002
Pasture land - field land = 0	0.623	0.139	0.007
Reforestation land - field land = 0	0.268	0.094	0.036
Secondary cropland - primary cropland = 0	−0.177	0.201	0.419
Pasture land - primary cropland = 0	−0.331	0.207	0.169
Reforestation land - primary cropland = 0	−0.687	0.179	0.012
Pasture land - secondary cropland = 0	−0.155	0.170	0.405
Reforestation land - secondary cropland = 0	−0.509	0.136	0.013
Reforestation land - pasture land = 0	−0.355	0.144	0.057

Table A6. Hypothesis testing for land category levels to Itapetininga city.

Primary cropland - field land = 0	0.572	0.027	< 0.001
Secondary cropland - field land = 0	0.404	0.009	< 0.001
Pasture land - field land = 0	0.301	0.006	< 0.001
Reforestation land - field land = 0	0.101	0.008	< 0.001
Secondary cropland - primary cropland = 0	−0.167	0.028	0.002
Pasture land - primary cropland = 0	−0.271	0.027	< 0.001
Reforestation land - primary cropland = 0	−0.471	0.027	< 0.001
Pasture land - secondary cropland = 0	−0.104	0.008	< 0.001
Reforestation land - secondary cropland = 0	−0.304	0.009	< 0.001
Reforestation land - pasture land = 0	−0.199	0.013	0.004

Table A7. Hypothesis testing for land category levels to Itapeva city.

Primary cropland - field land = 0	0.699	0.061	< 0.001
Secondary cropland - field land = 0	0.474	0.056	< 0.001
Pasture land - field land = 0	0.339	0.071	0.005
Reforestation land - field land = 0	−0.157	0.280	0.599
Secondary cropland - primary cropland = 0	−0.225	0.065	0.018
Pasture land - primary cropland = 0	−0.359	0.079	0.006
Reforestation land - primary cropland = 0	−0.489	0.909	0.614
Pasture land - secondary cropland = 0	−0.134	0.074	0.127
Reforestation land - secondary cropland = 0	−0.614	0.268	0.071
Reforestation land - pasture land = 0	−0.094	0.904	0.921

Table A8. Hypothesis testing for land category levels to Santo André city.

Primary cropland - field land = 0	0.889	0.182	0.004
Secondary cropland - field land = 0	0.598	0.076	0.001
Pasture land - field land = 0	0.548	0.070	0.001
Reforestation land - field land = 0	0.141	0.121	0.298
Secondary cropland - primary cropland = 0	−0.291	0.183	0.172
Pasture land - primary cropland = 0	−0.341	0.181	0.117
Reforestation land - primary cropland = 0	−0.749	0.206	0.015
Pasture land - secondary cropland = 0	−0.049	0.073	0.527
Reforestation land - secondary cropland = 0	−0.457	0.123	0.014
Reforestation land - pasture land = 0	−0.408	0.119	0.019

Table A9. Hypothesis testing for land category levels to São Carlos city.

Primary cropland - field land = 0	0.393	0.253	0.182
Secondary cropland - field land = 0	0.278	0.251	0.318
Pasture land - field land = 0	0.199	0.239	0.443
Reforestation land - field land = 0	0.147	0.224	0.540
Secondary cropland - primary cropland = 0	−0.115	0.272	0.691
Pasture land - primary cropland = 0	−0.194	0.261	0.491
Reforestation land - primary cropland = 0	−0.245	0.248	0.367
Pasture land - secondary cropland = 0	−0.079	0.259	0.772
Reforestation land - secondary cropland = 0	−0.131	0.245	0.617
Reforestation land - pasture land = 0	−0.052	0.233	0.834

Table A10. Hypothesis testing for land category levels to Campinas city.

Primary cropland - field land = 0	0.418	0.057	< 0.001
Secondary cropland - field land = 0	0.312	0.054	0.002
Pasture land - field land = 0	0.171	0.053	0.024
Reforestation land - field land = 0	0.040	0.079	0.629
Secondary cropland - primary cropland = 0	−0.106	0.050	0.088
Pasture land - primary cropland = 0	−0.247	0.049	0.004
Reforestation land - primary cropland = 0	−0.378	0.076	0.004
Pasture land - secondary cropland = 0	−0.141	0.046	0.028
Reforestation land - secondary cropland = 0	−0.272	0.074	0.014
Reforestation land - pasture land = 0	−0.131	0.074	0.136