STATISTICAL MODELING FOR PERFORMANCE OF TEAMS IN THE CITRUS HARVEST: CLASSICAL VERSUS BAYESIAN APPROACH

Marcelo Edmundo Alves MARTINS[1]
Jorge Alberto ACHCAR[2,1]
Claudio Luis PIRATELLI[1]

- ***ABSTRACT****: This study aims to identify the main factors that contribute to the performance of different teams of workers in the citrus harvest from a production engineering viewpoint. Statistical modeling was adopted as a quantitative approach in order to analyze a dataset from a citrus company in the state of São Paulo, Brazil. The main goal of the study was to verify the relationship between these factors and the general performance indicator given by the "number of boxes". The manager in the citrus harvest area of the company indicated several variables related to the worker team performance. For the data analysis, we consider a multiple linear regression model assuming transformed responses and Poisson regression models, under a Bayesian approach. The Bayesian approach had the best adherence to the data and shows us that the fruit harvest volume was affected by factors such as the team leader, the number of pickers, the percentage of male workers, among other variables.*

- *KEYWORDS: Citrus; performance of teams; harvest; linear regression models; Poisson regression model; Bayesian analysis; Markov Chain Monte Carlo (MCMC) methods.*

1 Introduction

Fruit production is now one of the most important segments of Brazilian agriculture, accounting for 25% of national agricultural production (Carvalho et al , 2012). It involves more than 5 million people in Brazil and the country is the third-largest fruit producer in the world, surpassed only by China and India. The major producing regions in Brazil are in the Southeast, Northeast and South.

In 2010, the state of São Paulo accounted for 32.9% of the national supply of fresh fruit, according to research done for Municipal Agricultural Production (PAM) by Instituto Brasileiro de Geografia e Estatística (IBGE - Brazil's official statistics office, 2012).Table 1 shows the volume of the main fruit produced in Brazil in 2011 and 2012.

Manual harvesting is based on the use of the main human senses such as sight and touch, among many others. This method offers both advantages and disadvantages. Among the advantages, humans are complete in relation to the senses, sight, touch, smell; that is, they are best-suited to the harvest. In this process, pickers tend to cause less damage to the products.

[1] Centro Universitário de Araraquara - UNIARA, Departamento de Engenharia de Produção, S.P.,Brasil. E-mail: *clpiratelli@uniara.com.br*; *marcelo@makegestao.com.br*
[2] Universidade de São Paulo - USP, Faculdade de Medicina de Ribeirão Preto - FMRP,Departamento de Medicina Social, CEP: 14049-900, Ribeirão Preto, SP, Brasil. E-mail: *achcar@fmrp.usp.br*

Table 1 - Main Fruit Production in Brazil - estimates of production(in tons)

Fruit	2011	2012
Orange	19,655,469	18,030,413
Banana	7,023,396	6,980,192
Pineapple*	1,519,881	1,455,056
Coconut*	1,899,355	1,786,498
Grape	1,463,481	1,387,830
Apple	1,364,953	1,208,658

Source: IBGE(January 2012)-* Thousand units. Conversion: 1 unit =2.5 Kg(South-Southeast Region except PR(1.6 kg) and SC(1.67 kg)), 2.1 kg (Center-West Region) and 1.8kg (for other regions).

Selection and packaging can be performed in the field, with fewer steps. Among the possible disadvantages are the high costs of labor in some regions. In addition, this labor is often not trained which can cause many problems. The seasonality involved in the supply of labor can also be a challenge for many regions (Cortez, 2002).

A harvest team size is established by the capacity of the vehicle used, which ranges from 45 to 50 people. An identification number is given to each picker, who is given the necessary material to do the job. Typically, each team has a leader who oversees the operation. The leader may have an important impact on team productivity (the number of boxes of fruit).

There are other covariates that could promote variability on the response variable y_i (number of boxes of fruits), such as: the number of crops harvested by the team; the sex of the team leader; the age of the team leader; the marital status of the team leader; the leader's educational level; the region where the team operates; the number of pickers in the team; the percentage of male workers in the team; the average age of the workers; the percentage of married workers in the team; the average absenteeism in the team; the average daily harvest; the average distance traveled to the harvest site, and the percentage of more experienced workers in the team.

The main objective of this paper is to identify the covariates that affect team performance in the citrus harvest. To this end, we first analyzed the data using a multiple linear regression technique considering the response variable y_i (number of boxes of fruit) given by a Box-Cox transformation (Box and Cox, 1964) . As a second analysis, we analyzed the data using two Poisson regression models under a Bayesian approach considering the responses (the numbers of boxes of fruits) in the original scale. The posterior summaries of interest were obtained via Markov Chain Monte Carlo (MCMC) simulation methods, such as the popular Gibbs sampling algorithm (Gelfand and Smith,1990) or the Metropolis–Hastings algorithm, when the conditional posterior distributions required for the Gibbs sampling algorithm do not have standard parametrical forms (see, for example, Chib and Greenberg, 1995). The secondary goal of this paper is to present and to discuss the best statistical modeling for this case.

Regression analysis of count data has been studied by Hausman et al. (1984), Zeger (1988), Blundell et al. (1995), Martz and Piccard (1995), Gurmu et al. (1999), Freeland and McCabe (2004a and 2004b). Bayesian estimation has been proposed in count regression models by Harvey and Fernandes (1989), Albert (1992), Chib et al. (1998),

Settimi and Smith (2000), Martz and Hamada (2003), McCabe and Martin (2005) and Zheng (2008).

Bayesian methods have been used extensively in many applied areas, such as business administration, economics and industrial engineering. Some examples taken from the SciELO scientific database: Quinino and Bueno Neto (1997) used Bayesian methods to evaluate the accuracy of a quality inspector; Pongo and Bueno Neto (1997) and Droguett and Mosleh (2006) proposed Bayesian inference to evaluate the reliability of products in development projects; Cavalcante and Almeida (2005) used multi-criteria method and Bayesian analysis to determine preventive maintenance intervals; Kalatziz et al. (2006) used Bayesian approach in investiment decisions; Moura et al. (2007) used Bayesian methods to evaluate the efficiency of maintenance; Ferreira et al. (2009) used a Bayesian approach in a portfolio selection problem; Barossi-Filho et al. (2010) used Bayesian analysis to estimate the volatility of financial time series; Freitas et al. (2010) used a Bayesian approach to estimate the wearing out of train wheels and Che and Xu (2011) used a Bayesian data analysis for agricultural experiments. Achcar et al. (2013) used a Bayesian approach to identify covariates affecting daily counting of units that arrive for quality inspection at a food company.

The literature contains a rapidly growing number of published papers using the Bayesian paradigm in almost every applied area, such as medicine, economics, environmental sciences or engineering, since there has been a huge advance in computer hardware and software in the last twenty years. According to Fildes (2006), Bayesian methods have been gaining in prominence in the number of their citations in important journals in recent years and chief editors see them as a hot topic in forecasting (comprising counting problems). Armstrong and Fildes (2006) argue that many forecasting areas have developed methods, but few of them have been adopted in organizations in practice. It is important to recognize that the best statistical model for the presented problem will help managers achieve better performance forecasts for manual harvests, that is, it would be helpful as a tool in putting together new harvest teams.

The paper is organized as follows: in section 2 we present some justifications for the study; in section 3 we present the methodology used; in section 4, the multiple linear regression technique and the Poisson regression model under a Bayesian approach are presented; in section 5 we describe the object of study then we apply and analyze the techniques from section 4 to the collected data; finally in section 6 we present a discussion on the results obtained and make some concluding remarks.

2 Justification for the study

The worldwide competition many companies have faced has led them to continuously improve their results. Great importance has therefore been placed on looking into factors that impact productivity. Another important point that justifies this study is the presence of sectors that also use manual labor. This has resulted in rural mechanization because of environmental problems and a lack of labor. In Brazil, with mechanized sugar cane harvesting at around 68% in the last years, some of this labor was expected to migrate to the citrus harvest. This has not happened, as most of the workers migrated to the construction industry. Harvest costs have thus increased greatly in the last decade (CEPEA, 2012).

According to Hsu and Wang (2007), modeling industry data sets is a challenge, given some important points in the area:

- A large data sample is not always available;
- Data is usually missing;
- There are interfering outliers;
- The predictor variables are not correlated with the responses, among many other reasons.

The use of Bayesian inference is justified by the possibility of greater flexibility in the interpretations of the obtained inference results. In particular, one of the main advantages of using Bayesian inference is the use of prior information in the choice of prior distributions, especially for data obtained in industrial applications. When we do not have good prior information, we can use non-informative priors in the model's parameters. Another great advantage is the use of standard existing Markov Chain Monte Carlo (MCMC) simulation methods. Hence, we do not need to use asymptotical results based on the asymptotical normality of the maximum likelihood estimators or standard asymptotical likelihood ratio tests, which depend on large sample sizes.

3 Methodology

The methods to be employed in scientific research can be selected from the identification of the problem, the formulation of hypotheses, or the delimitation of the universe, or sample. A selection of these steps will depend on several factors related to the study, such as the nature of the phenomena, the object or research, the resources, the research approach (qualitative or quantitative, or a combination of both), among others (Lakatos and Marconi, 2008).

As Miguel (2007), methodologically this study can be classified as an applied, objective and descriptive quantitative approach. Bertrand and Fransoo (2002) define quantitative research in production engineering as that which models a problem whose variables have causal and quantitative relationships. In this way, it becomes possible to quantify the behavior of dependent variables under a particular domain, allowing the researcher to make predictions. In general, quantitative research may use mathematical, statistical or computational (simulation) modeling - specifically, in this work statistical modeling will be adopted.

For this study, we considered a data set related to the responses of n = 605 orange picker teams selected from different regions of São Paulo state, in Brazil. Among the several existing responses related to this study, we consider in this paper the following: the daily harvest production in number of boxes of fruit picked, for different teams of workers.

Among several possible covariates associated with each class of workers (the samples related to the problem), the manager advised us to select the following: the number of crops harvested by the team, the sex of the team leader, age of the team leader, the marital status of the team leader, the educational level of the team leader, the region where the team operates, the number of pickers, the percentage of male workers in the team, the average age of workers in the team, the percentage of married workers in the

team, average absenteeism for the team, the average daily harvest, the average distance to the harvest site, and the percentage of more experienced workers in the team.

Note that the formulation of appropriate statistical models for the data can lead to great gains for the companies in the fruit sector in terms of identifying key factors that control the variability of responses and forecasts.

4 Statistical modeling

To analyze the impact of some covariates on counting problems, such as team productivity, characterized as number of boxes of fruit, we can proceed with a classical approach (multiple linear regression assuming transformed responses) or a Bayesian approach assuming Poisson regression models for the counting data in the original scale.

4.1 A classical approach assuming a linear regression model

In statistical analysis, when the main goal is to verify the combined effect of the covariates on a response Y, we usually employ multiple linear regression techniques (see, for example, Draper and Smith, 1981; Seber and Lee, 2003, or Montgomery and Runger, 2011).

In this way, to analyze the productivity data in the citrus fruit sector, we assume a multiple linear regression model considering the covariates introduced in Section 3 and a transformed response Y, given by:

$$\begin{aligned} y_i = {} & \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \\ & \beta_7 x_{7i} + \beta_8 x_{8i} + \beta_9 x_{9i} + \beta_{10} x_{10i} + \beta_{11} x_{11i} + \beta_{12} x_{12i} + \beta_{13} x_{13i} + \beta_{14} x_{14i} + \varepsilon_i \end{aligned} \quad (1)$$

where i = 1,2, ..., 605; the random errors ε_i are assumed to be independent, distributed with a normal distribution, with zero mean and constant variance σ^2;the covariates are respectively given by:

- x_{1i} denotes the number of crops harvested by the team;
- x_{2i} denotes the sex of the team leader;
- x_{3i} denotes the age of the team leader;
- x_{4i} denotes the marital status of the team leader;
- x_{5i} denotes the team leader's educational level;
- x_{6i} denotes the region where the team operates;
- x_{7i} denotes the number of pickers in the team;
- x_{8i} denotes the percentage of male workers in the team;
- x_{9i} denotes the average age of the workers;
- x_{10i} denotes the percentage of married workers in the team;
- x_{11i} denotes the average absenteeism in the team ;

- x_{12i} denotes the average daily harvest;
- x_{13i} denotes the average distance traveled to the harvest site;
- x_{14i} denotes the percentage of more experienced workers in the team.

The response variable y_i (number of boxes of fruit) is given by a Box-Cox transformation (Box and Cox, 1964), namely:

$$y_i = [v_i^{\gamma} - 1]/\gamma \quad (2)$$

where γ is a parameter to be estimated from data and v_i is the total volume (number of boxes) collected by the workers. Note that if $\gamma = 0$, the above equation reduces to,

$$y_i = \log(v_i)$$

where log(.) is the natural logarithm. Note that the Box-Cox transformation for the response was considered to have approximately, the usual assumptions needed for the proposed regression model (normality of errors and constancy of the variance for the errors).

Under the classical approach the estimates for the regression parameters are obtained using a minimum squares approach.

4.2 A Bayesian approach assuming a Poisson regression model

As an alternative, we could analyze the data using Bayesian methods assuming the responses in the original scale not considering the Box-Cox transformation for the total harvested. For this case, we consider Poisson regression models under a Bayesian approach.

Let Y_i (denoting the number of boxes of fruit) be a random variable with a Poisson distribution given by,

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad (3)$$

where $y_i = 0, 1, 2, \ldots$ denotes the total harvest of fruit by the i^{th} team of workers, i = 1, 2, , ..., 605. Note that theoretically the mean and variance of the Poisson distribution (3) should both be equal to λ_i.

To relate the parameter λ_i with the same covariates introduced in (1), let us consider the following regression model:

$$\begin{aligned}\log(\lambda_i) = {} & \beta_0 + \beta_1(x_{1i} - 9.2595) + \beta_2 x_{2i} + \beta_3(x_{3i} - 43.1587) + \beta_4 x_{4i} + \\ & \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7(x_{7i} - 47.5223) + \beta_8 x_{8i} + \beta_9(x_{9i} - 35.8777) + \\ & \beta_{10} x_{10i} + \beta_{11} x_{11i} + \beta_{12}(x_{12i} - 170.891) + \beta_{13} x_{13i} + \beta_{14} x_{14i}\end{aligned} \quad (4)$$

Note that some covariates were centered at their means for better stability of the simulation procedure used to generate samples from the posterior distribution of interest. In this way, 9.2595 is the arithmetic mean of the values x_{1i} ; 43.1587 is the arithmetic mean of the values x_{3i} ; 47.5223 is the arithmetic mean of the values x_{7i} ; 35.8777 is the arithmetic mean of the values x_{9i} and 170.891 is the arithmetic mean of the values x_{12i}, i

=1,…,605. Let us denote the model given by (3) and (4) as "Model 1." The formulation (4) ensures that λ_i be positive, for i = 1, 2, ..., n. Assuming the model (4), the likelihood function for the vector $\boldsymbol{\theta}$ of parameters associated with the model is given by,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{605} f(data/\boldsymbol{\theta}) \tag{5}$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})$ and f(data/$\boldsymbol{\theta}$) is the Poisson probability distribution (3) for the data.

Combining the joint prior distribution for $\boldsymbol{\theta}$ (a product of normal distributions) with the likelihood function L($\boldsymbol{\theta}$) given in (5), the posterior distribution for $\boldsymbol{\theta}$ is determined from the Bayes formula (see for example, Box and Tiao, 1973).

The posterior summaries of interest can be obtained using Markov Chain Monte Carlo (MCMC) methods (see for example, Gelfand and Smith, 1990, or Chib and Greenberg, 1995). A great simplification in the generation of samples from the posterior distribution for $\boldsymbol{\theta}$ is obtained by using the software Open Bugs (Spiegelhalter et al, 2003), which only requires the specification of the distribution for the data and a prior distribution for the parameters of the model.

As an alternative, we may propose a second model (6) in the presence of a random effect w_i, i = 1, ..., 605 which captures the extra-Poisson variability (see, for example, Albert and Chib, 1993; Cruchley and Davies, 1999; Dunson, 2000, 2003, or Henderson and Shimakura, 2003).

In this way, we consider the Poisson regression model given by,

$$\begin{aligned}\log(\lambda_i) = {} & \beta_0 + \beta_1(x_{1i} - 9.2595) + \beta_2 x_{2i} + \beta_3(x_{3i} - 43.1587) + \beta_4 x_{4i} \\ & + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7(x_{7i} - 47.5223) + \beta_8 x_{8i} + \beta_9(x_{9i} - 35.8777) + \\ & \beta_{10} x_{10i} + \beta_{11} x_{11i} + \beta_{12}(x_{12i} - 170.891) + \beta_{13} x_{13i} + \beta_{14} x_{14i} + w_i\end{aligned} \tag{6}$$

where w_i is a random effect or non-observed latent variable assumed with a normal distribution N $(0,1/\zeta_w)$.With this regression model, we capture the extra-Poisson variability. Denote this model "Model 2." Observe that “model 2” differs of “model 1” by the inclusion of the random effect or latent variable w_i , i =1,…,n, which captures the extra-variability of the Poisson distribution. The “model 1” does not have this random effect.

The extra-Poisson variability of “model 2” defined by (3) and (6) is given as follows. Since Y_i has a Poisson probability distribution function (3),we have,

$$E(Y_i/\lambda_i) = var(Y_i/\lambda_i) = \lambda_i,$$

that is,

$$E(Y_i/\boldsymbol{\beta}, \boldsymbol{x}_i, w_i) = exp(\boldsymbol{\beta x}_i + w_i),$$

and

$$var(Y_i/\boldsymbol{\beta}, \boldsymbol{x}_i, w_i) = exp(\boldsymbol{\beta x}_i + w_i).$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{14})$ and $\mathbf{x}_i = (1, x_{1i}, x_{2i}, \ldots, x_{14i})$ for i =1,…,605.

Since E(Yi/β,xi) = E[E(Yi/β,xi,wi)], we have, E(Yi/β,xi) = exp(βxi)E[exp(wi)].

From the normality of the random effects w_i, that is, $\sim N(0,\sigma_w^2), \sigma_w^2 = 1/\zeta_w$ we observe that $\exp(w_i)$ has a log-normal distribution (see for example, Lawless,1982) with mean,

$$E[exp(w_i)] = exp(\sigma_w^2/2),$$

and variance

$$var[exp(w_i)] = [exp(\sigma_w^2) - 1] exp(\sigma_w^2).$$

That is,

$$var[exp(wi)] = [exp(\sigma_w^2) - 1] exp(\sigma_w^2). \quad (7)$$

In the same way, as var(Yi/ β, xi) = E[var(Yi/ β, xi ,wi)] + var[E(Yi/ β, xi ,wi)],we have, var(Yi/ β, xi) = exp(β'xi)E[exp(wi)] + exp(2β'xi)var[exp(wi)].

That is,

$$var(Yi/\beta,xi) = exp(\beta xi)\, exp(\sigma_w^2/2) + exp(2\beta xi)\, [exp(\sigma_w^2) - 1]\, exp(\sigma_w^2). \quad (8)$$

From (7) and (8), we observe that the mean is different from the variance for Yi/ β, xi ; that is, we have an extra-variability given by the term,

$$exp(2\beta'xi)\, [exp(\sigma_w^2) - 1]\, exp(\sigma_w^2). \quad (9)$$

Let us assume a hierarchical Bayesian analysis for this model with the same prior distributions for the regression parameters considered for "model1" and a uniform prior distribution in the interval (0,10) for the parameter ζ win the second stage of the hierarchical Bayesian analysis.

5 The goal of the study and statistical modeling

The study was conducted in a fruit processing company, one of the world's largest juice producers. Its supply chain starts from planting seedlings, planting, fruit production, harvesting (the company owns farms in the states of São Paulo and Minas Gerais) and processing juice.

The most important juice consumer markets are the US and Europe - the juice is transported in bulk by ships. It is one of the most important crops and it is expensive to harvest as it is done manually, requiring a large number of workers.

From a preliminary data analysis, Figures 1 and 2 show the effects of some covariates (described in section 3) on the harvest (number of boxes). To confirm if these effects are significant in the responses we will proceed with the statistical modeling presented in section 4.

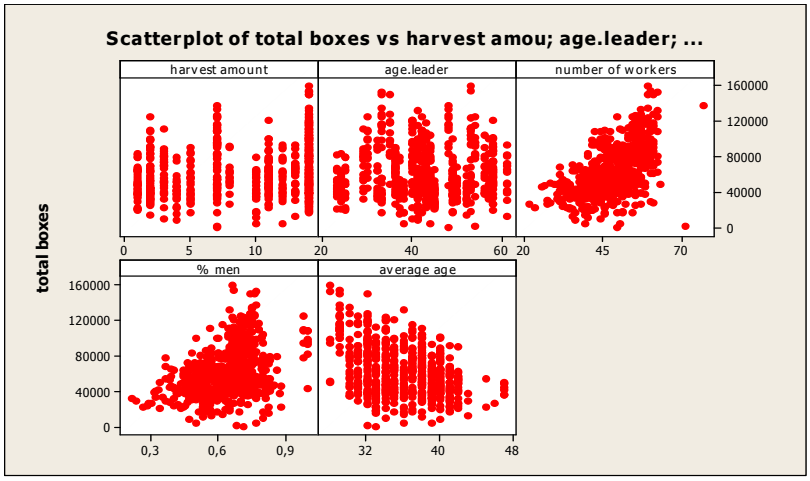

Figure 1 – Scatterplot of total boxes against covariates.

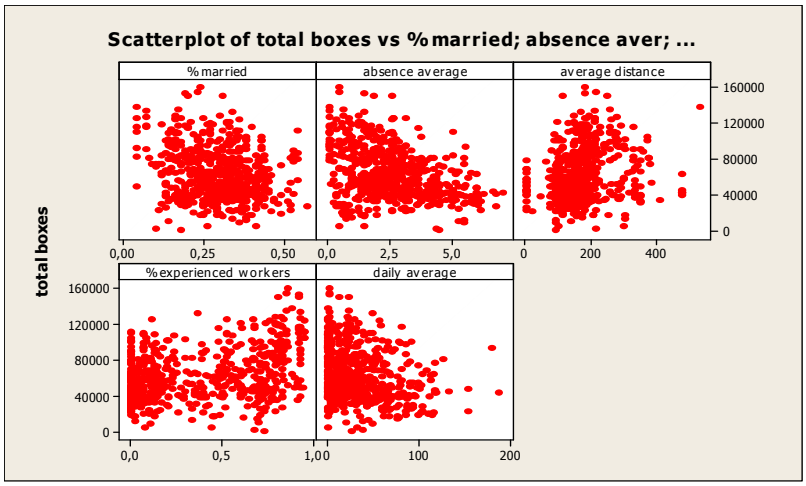

Figure 2 – Scatterplot of total boxes against covariates.

5.1 Classical statistical analysis

Assuming the regression model (1) we have the following fitted regression model obtained by least squares using MINITAB® version 16 software:

$$
\begin{aligned}
y \;=\; & 1999 \;+\; 19.7255\; harvest.count \;+\; 6.505 leader.sex(1 = male) \\
& -\; 0.0425939\; leader.age \\
& -\; 89.3966\; leader.marital.status\; (1 = married) \\
& -\; 98.931\; leader.school\; (1 = primary) \\
& -\; 37.5849\; region(north = 1) \\
& +\; 59.6954\; number.workers \;+\; 1144.75\; \%men \\
& -\; 49.5175\; average.age \;+\; 1336.19\; \%married \\
& -\; 175.135\; average.absences \;-\; 12.8122\; daily.average \\
& +\; 0.219866\; average.distance \\
& -\; 175.383\; \%\; workers.more.experience.)
\end{aligned}
\tag{10}
$$

The Box-Cox transformation of the response was considered with the estimated γ given in (2) (maximum likelihood estimator) equals to 0.736278 with a 95% confidence interval for γ given by (0.632778, 0.843778). The rounded value 0.736278 was used in the regression analysis. Table 2 shows the summaries of the inferences obtained for this model:

Table 2- Estimation byleast squares and p-values for the regression coefficients considering the transformed response (Box-Cox)

Predictor	LSE	SE	T	P
constant	1999.00	633.479	31.556	0.002
harvest.count	19.73	8.314	23.726	0.018
leader.sex(1=male)	6.51	109.611	0.0593	0.953
leader.age	-0.04	4.813	-0.0088	0.993
leader.mar.stat(1=married)	-89.40	76.270	-11.721	0.242
school.level (1=primary)	-98.93	69.844	-14.165	0.157
region (north=1)	-37.58	83.087	-0.4524	0.651
number.workers	59.70	4.703	126.942	<0.001
%men	1144.75	328.663	34.831	0.001
average.age	-49.52	15.398	-32.159	0.001
%married	1336.19	419.539	31.849	0.002
average.absences	-175.13	24.758	-70.738	<0.001
daily.average	-12.81	1.036	-123.711	<0.001
average.distance	0.22	0.475	0.4626	0.644
%experienced.workers	-175.38	181.665	-0.9654	0.335
S = 771.010 R-Sq = 54.00%	R-Sq(adj) = 52.91%			

(LSE: Least squares estimator; SE: Standard Error; T:Student-t statistics; P:p-value)
Source: Author´s own.

From the results in Table 2, we observe that the significant covariates in the number of boxes of fruit (p-values smaller than 0.05) are the following:

- x1 :the number of crops harvested by the team(harvest.count);
- x11 : the average absenteeism in the team(average.absences);
- x7 :the number of pickers in the team(number.workers);
- x8 : the percentage of male workers in the team(%men);
- x9 : the average age of the workers(average.age);
- x10: the percentage of married workers in the team(%married);
- x12: the average daily harvest(daily.average).

All the other covariates were not significative (p-value > 0.05); that is, the covariates,

- x2 : the sex of the team leader(leader.sex(1=male));
- x3: the average age of the team leader(leader.age);
- x4 : the marital status of the team leader(leader.mar.stat(1=married));
- x5 : the team leader's educational level(school.level (1=primary));
- x6 : the region where the team operates(region (north=1));
- x13 : the average distance traveled to the harvest site(average.distance);
- x14 : the % of more experienced workers in the team(%experienced.workers).

To check the validity of the models, Figure 3 shows the graphs for the residuals for the fitted model.

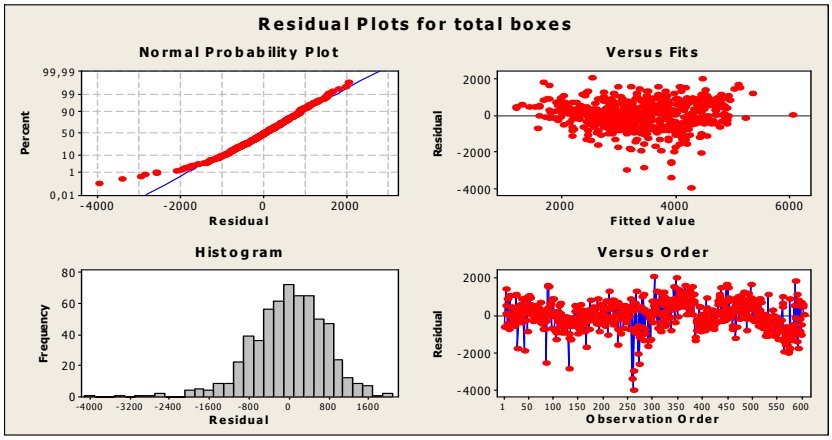

Figure 3 – Residual graphs.

From the graphs in Figure 3, we see that the necessary assumptions for the validity of the statistical model (normality of the residuals, constant variance of errors) are only

approximately observed for the model, assuming the response with a Box-Cox transformation. Overall, the residuals do not satisfy the needed model assumptions, which may affect the significance of the hypothesis tests on the regression parameters even considering the Box-Cox transformation, usually used to guarantee standard ANOVA (analysis of variance) assumptions such as normality and constant variance of the residuals.

5.2 Bayesian statistical analysis

Assuming the Poisson regression model defined by (3) and (4) and using the Open Bugs software (Spiegelhalter et al, 2003), we simulated a initial sample ("burn-in-sample") size of 5,000 which was discarded to eliminate the effect of the initial values used in the Gibbs Sampling algorithm, and another 50,000 samples from where it was chosen every 50^{th} sample to have approximately uncorrelated samples. In this way, we obtained a final sample size of 1,000 used to generate values for β_r and $\beta_0$with r = 1,2, ..., 14. The obtained posterior summaries (posterior means, standard deviations and subsequent credibility intervals with probability equals to 0.95) are shown in Table 3. The convergence of the algorithm was monitored using graphical methods (see, for example, Paulino et al, 2003).

From the results in Table 3, we observe that all covariates have significative effects on the total daily production of fruit, since zero is not included in all 95% credibility intervals of the regression parameters.

Table 3. Posterior summaries for the Poisson regression model(total amount of fruit) model1

Parameter	Mean	SD	95%	Cred. interval
β_0	10.84	0.001648	10.84	10.84
β_1	0.005654	0.0433	0.00557	0.005739
β_{10}	0.6236	0.002101	0.6198	0.628
β_{11}	-0.07274	0.133	-0.07301	-0.07248
β_{12}	-0.0464	0.00257	-0.0514	-0.0417
β_{13}	-0.07442	0.966	-0.07626	-0.07256
β_{14}	-0.005321	0.00586	-0.005333	-0.005309
β_2	0.02674	0.562	0.0256	0.02784
β_3	0.594	0.0248	0.543	0.640
β_4	-0.03691	0.389	-0.0377	-0.03617
β_5	-0.03225	0.352	-0.03289	-0.03152
β_6	0.006695	0.442	0.005848	0.007541
β_7	0.02472	0.0245	0.02467	0.02477
β_8	0.578	0.001758	0.5745	0.5814

β_9	-0.02054	0.0793	-0.0207	-0.02038

Source: Author´s own (SD: posterior standard deviation)

Figure 4 shows graphs for the total amount of fruit observed in each team and the mean estimated by "Model 1" against samples. The sum of the absolute values of the differences (estimated mean minus observed fruit amount) using "Model 1" is given by 8.439,433.

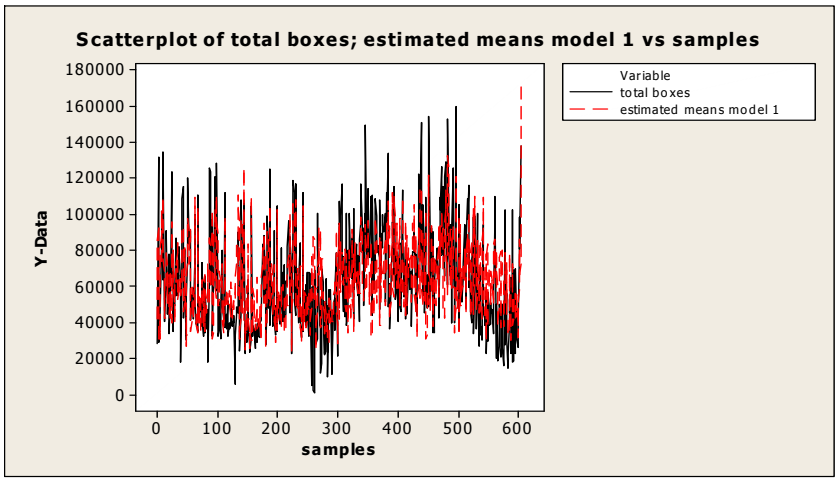

Figure 4 – Graphs of the observed and estimated means against samples(model 1).

One problem observed with this model: the sample mean (equals to 63,402) is very different from the sample variance (equals to 797,028), which violates the basic assumption of the Poisson distribution. That is, we have an extra-Poisson variability in the model. For this situation, we will consider the model (6) in the presence of a random effect w_i, i = 1, ..., 605 which captures the extra-Poisson variability.

Assuming the Poisson regression model defined by "Model 2" in the presence of a random effect with a normal distribution, we also used the Open Bugs software with a simulated “burn-in-sample" size of 100,000 discarded to eliminate the effect of the initial values used in the Gibbs sampling simulating algorithm and another 200,000 samples from where we selected one sample in every 200th simulated samples to have approximately uncorrelated samples. In this way, we obtained a final generated sample of size 1,000 for all parameters to be used to find the Monte Carlo estimates for the parameters of the model. The obtained posterior summaries (posterior mean, posterior standard deviation and 95% credibility intervals) are shown in Table 4.

From the results in Table 4, we observe significative effects of the following covariates on the response (zero is not included in the 95% credibility interval for each related regression coefficient):

- x1 :the number of crops harvested by the team;

- x11 : the average absenteeism in the team;
- x14 : the percentage of more experienced workers in the team.
- x2 : the sex of the team leader;
- x3: the age of the team leader;
- x7 :the number of pickers in the team;
- x8 : the percentage of male workers in the team.

All the other covariates were not significative (zero is included in the 95% credibility interval for each related regression coefficient):

- x4: the marital status of the team leader;
- x5: the team leader's educational level;
- x6: the region where the team operates;
- x9: the average age of the workers;
- x10: the percentage of married workers in the team;
- x12: the average daily harvest;
- x13: the average distance traveled to the harvest site;

Table 4. Posterior summaries for the Poisson regression model - "model 2"

parameter	mean	SD	95%	Cred. interval
β_0	11.1	0.08037	10.93	11.21
β_1	0.006785	0.003108	0.001101	0.01376
β_{10}	0.07592	0.1618	-0.1164	0.3914
β_{11}	-0.09603	0.01734	-0.1235	-0.06519
β_{12}	0.419	0.253	-0.0564	0.887
β_{13}	-0.006338	0.1099	-0.1695	0.181
β_{14}	-0.005112	0.001022	-0.006861	-0.003682
β_2	0.07945	0.04003	0.00976	0.1669
β_3	-0.003435	0.001715	-0.007848	- 0.906
β_4	-0.007533	0.0364	-0.07016	0.06508
β_5	-0.004285	0.02353	-0.07135	0.03585
β_6	0.01261	0.04165	-0.06482	0.0869
β_7	0.02449	0.00198	0.01966	0.0281

parameter	mean	SD	95%	Cred. interval
β_8	0.2474	0.04693	0.1762	0.3433
β_9	-0.004132	0.006285	-0.01583	0.006843

Source: Author´s own (SD: posterior standard deviation)

Figure 5 shows the graphs of observed total amount of fruit boxes in each team and the means estimated by "Model 2". We observe a very good fit for “model 2” for the dataset.

We also note that the sum of the absolute values of the differences (estimated mean minus observed amount of fruit boxes) considering "model 2" is given by 5.594,90. That is, "model 2" is much better than "Model 1," and has much better predictions than those obtained by "model 1."

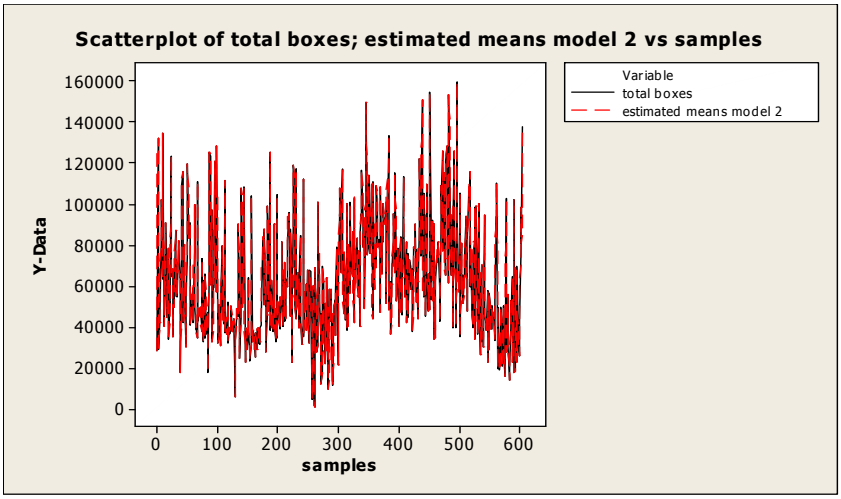

Figure 5 -Graphs of the observed and estimated means against samples(model 2).

6 Concluding remarks

In this paper, we identified the covariates that have a significant impact on the performance (volume harvested) of different citrus harvest teams of a citrus company in São Paulo state.

In this way, we have used different modeling approaches for the counting data: standard multiple regression models assuming a Box-Cox transformation for the counting data and Poisson regression models for the counting data in the original scale in the presence or not of a random effect which captures extra-Poisson variability.

From the observed results of the study, we observe that the standard multiple linear regression (MLR) presented results not totally different from the obtained results using the Poisson regression model (“Model 2”) under a Bayesian approach. However, the MLR results could be put in check when we have dubious assumptions about the behavior of residuals (discussed in section 4) as we have seen in our study.

For the analysis of counting data, the Poisson regression model under a Bayesian approach has great advantages when compared with the standard multiple regression model assuming transformed data; in this way we have a direct modeling for the count data without the need for transformations.

When comparing the Bayesian models("model 1" against "model 2"), the second model is shown to be more adherent to the input data due to the inclusion of a random effect which captures much of the variability of each sample(teams). In other words, the model is more sensitive to the effects of the covariates. Moreover, it is a model that captures the extra-Poisson variability, something that cannot be discarded given the difference between the sample mean and sample variance.

It is therefore evident that the use of Poisson regression models can be a good alter native for analyzing industrial data, especially under a Bayesian approach using MCMC simulation methods to obtain the summaries of interest. In general these data show great variability and the usual assumptions of traditional models cannot be verified. The introduction of latent variables can lead to results and predictions with great accuracy, as noted in this case study.

As a final conclusion, we observe from the obtained results, that important covariates have great impact on the better performance of the workers in the citrus company as the number of crops harvested by the team, the average absenteeism in the team, the percentage of more experienced workers in the team, the sex of the team leader, the age of the team leader, the number of pickers in the team and the percentage of male workers in the team.

These results could be of great interest for the citrus companies.

MARTINS, M. E. A.; ACHCAR, J. A.; PIRATELLI, C. L. Modelagem estatística para o desempenho de equipes de colheita de laranjas: métodos clássicos versus métodos Bayesianos, *Rev. Bras. Biom*., São Paulo, v.32, n.4, p.525-543, 2014.

- RESUMO: O objetivo deste estudo é identificar os principais fatores que contribuem para o desempenho de diferentes equipes de colhedores de laranja, sob a ótica da engenharia de produção. A técnica de modelagem estatística foi utilizada como abordagem quantitativa, com o objetivo de analisar um conjunto de dados de uma empresa no Estado de São Paulo. Particularmente, estamos interessados em estudar as relações entre algumas variáveis e o indicador de desempenho de colheita "número de caixas". As variáveis relacionadas com as equipes de colheita foram indicadas pelo gestor desta operação. Os dados foram modelados através de regressão linear múltipla, assumindo uma transformação na variável resposta e através de modelos de regressão de Poisson, sob o enfoque Bayesiano. O enfoque Bayesiano mostrou-se mais aderente aos dados e permitiu a conclusão de que variáveis relacionadas ao líder das equipes, ao número de colhedores, ao percentual de colhedores do sexo masculino, dentre outras, afetam o volume de caixas colhidas.

- PALAVRAS-CHAVE: Desempenho de equipes; colheita; regressão linear múltipla; modelos de regressão de Poisson;análise Bayesiana;método de Monte Carlo em cadeia de Markov.

References

ACHCAR J.A.; PIRATELLI, C.L.; SANDRIM, R.R. Daily counting of manufactured units sent for quality control. *Pesquisa Operacional,* v.33, n.2, p.185-198, 2013.

ALBERT, J. H. Bayesian analysis of a Poisson random-effects model. *American Statistician*, v.4, p.246-253,1992.

ALBERT, J. H. ; CHIB, S. Bayesian analysis of binary and polychotomus response data. *Journal American Statistical Association*, v.88, n.422, p.669-679,1993.

ARMTRONG, J.S.;FILDES, R. Making process in forecasting. *International Journal of Forecasting*, v.22, p.433-441, 2006.

BAROSSI-FILHO, M.; ACHCAR J. A.; SOUZA, R. M. Modelos de volatilidade estocástica em séries financeiras: uma aplicação para o IBOVESPA. *Economia Aplicada***,** v.14, n.1, p.25-40, 2010.

BERTRAND, J.W.M.; FRANSOO, J.C. Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, v.22, n.2, p.241-261, 2002.

BLUNDELL, R.; Griffith R ; Van Reenen J. Dynamic count models of technological Innovation. *Economic Journal***,** v.105, p.333-344,1995.

BOX, G.E.P. ; COX, D.R. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, v.26, p.211–252,1964.

BOX, G.E.P.; TIAO, G. Bayesian inference in statistical analysis. New York: Addison-Wesley,1973. 608 p.

CAVALCANTE, C. A. V.; ALMEIDA, A. T. Modelo multicritério de apoio a decisão para o planejamento de manutenção preventiva utilizando PROMETHEE II em situações de incerteza. *Pesquisa Operacional*, v.25,n.2,p.279-296, 2005.

CARVALHO, C.; VENCATO, A. Z.; KIST, B. B.; SANTOS, C.; SILVEIRA, D.; REETZ, E. R.; BELING, R. R.; CORREA, S. *Brazilian fruit yearbook.* Santa Cruz do Sul, RS, Brazil: Editora Gazeta Santa Cruz, 2012.

CEPEA. Centro de Estudos Avançados em Economia Aplicada: harvesting costs 2003/2012. Universidade de São Paulo, ESALQ,2012.

CHE, X.; XU, S. Bayesian data analysis for agricultural experiments. *Canadian Journal of Plant Science*,v.91, n.4,p.599-601, 2011.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, v.49, p.327-335,1995.

CHIB, S.; GREENBERG, E. ; RAINER,W. Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics*, v.86, p.33-54,1998.

CRUCHLEY,R.;DAVIES, R.B. A comparison of population average random-effect models for the analysis of longitudinal count data with base-line information. *Journal of the Royal Statistical Society*, Series A, v.162, n.3, p.331-347,1999.

CORTEZ, L. A. B. ;BRAUNBECK, A. O.; CASTRO, L. R.; ABRAHÃO, R. F.; CARDOSO, J. L. Sistemas de Colheita para Frutas e Hortaliças: oportunidades para sistemas semi-mecanizados. *Revista Frutas e Legumes*, v.12, p.26-29,2002.

DROGUETT, E. L.; MOSLEH,A. Análise bayesiana da confiabilidade de produtos em desenvolvimento. *Gestão da Produção*, v.13, n.1,p.57-69,2006.

DUNSON, D.B. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society* B, v.62, p.355-366, 2000.

DUNSON, D.B. Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association,* v.98, p.555-563,2003.

DRAPER, N.R.;SMITH, H. Applied regression analysis. Wiley series in probability and mathematical statistics,1981.709p.

FERREIRA, R. J. P.; ALMEIDA FILHO, A.T.; SOUZA, F.M.C. A decision model for portfolio selection. *Pesquisa Operacional*, v.29, n.2, p.403-417, 2009.

FILDES, R. The forecasting journals and their contribution to forecasting research: Citation analysis and expert opinion. *International Journal of forecasting*, v. 22, p.415-432,2006.

FREELAND, R.K.; McCABE,B.P.M. Forecasting discrete valued low count time series. *International Journal of Forecasting***,** v.20, p.427–434, 2004a.

FREELAND, R.K.; McCABE,B.P.M. Analysis of low count time series data by Poisson autoregression. *Journal of Time Series Analysis*, v. 25, p.701–722,2004b.

FREITAS, M.A.;COLOSIMO, E.A.;SANTOS, T.R.; PIRES, M.C. Reliability assessment using degradation models: Bayesian and classical approaches. *Pesquisa Operacional*, v. 30, n.1, p.195-219, 2010.

GELFAND, A.E.; SMITH, A.F.M. Sampling-based approaches to calculating marginal distributions. *Journal of the American Statistical Association*, v. 85, n.410,p.398-409,1990.

GURMU, S.; RILSTONE, P.; STERN S. Semiparametric estimation of count regressionmodels. *Journal of Econometrics*, v. 88, p.123–150,1999.

HARVEY, A.C.; FERNANDES, C. Time series models for count or qualitative observations. *Journal of Business and Economic Statistics***,** v. 7,p.407–417,1989.

HAUSMAN, J.A.; HALL, B.H.; GRILICHES, Z. Econometric models for count data with applications to the patents R and D relationship**.** *Econometrica*, v. 52, p.909–938,1984.

HERDERSON, R.; SHIMAKURA, S. A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, v. 90,p.355–366,2003.

HSU, L.C.; WANG, C.H.Forecasting the output of integrated circuit industry using a greymodel improved by the Bayesian analysis. *Technological Forecasting & Social Change*, v.74, p.843–853,2007.

IBGE. Instituto Brasileiro de Geografia e Estatística. Levantamento sistemático da produção agrícola. Rio de Janeiro: IBGE, v. 25, p.1-88, 2012.

KALATZIS, A.E.G.; AZZONI, C.R.; ACHCAR, J. A. Uma abordagem bayesiana para decisões de investimentos.*Pesquisa Operacional*, v. 26, n.3,p.585-604,2006.

LAKATOS, E.M.; MARCONI, M.A. Técnicas de Pesquisa, São Paulo: Atlas, 2008.310 p.

MARTZ, H.F.; HAMADA, MS. Uncertainty in counts and operationg time in estimating Poisson occurence rates. *Reliability Engineering & System Safety*, v. 80, p.75-79, 1995.

MARTZ, H.F.; PICARD, R.R. Uncertainty in Poisson event counts andexposure time in rate estimation. *Reliability Engineering & System Safety*, v.48,p.181–90,1995.

McCABE, B.P.M.; MARTIN, GM. Bayesian predictions of low count time series. *International Journal of Forecasting***,** v. 21, p.315–330, 2005.

MIGUEL, P.A.C. Estudo de caso na engenharia de produção: estruturação e recomendações para sua condução. *Produção,* v. 17, n.1, p.216-229,2007.

MONTGOMERY, D.C.; RUNGER, G.C. Applied statistics and probability for engineeers, fifty edition. New York: Wiley,2011.768 p.

MOURA, M.C.; ROCHA, S.P.V.; DROGUETT, E.L. Avaliação Bayesiana da eficácia da manutenção via processo de renovação generalizado. *Pesquisa Operacional,* v. 27, n.3,p.569-589,2007.

PAULINO, C.D.; TURKMAN, M.; MURTEIRA, B. Estatística Bayesiana. Lisboa: Fundação Calouste Gulbenkian,2003.446 p.

PONGO, R.M.R.; BUENO NETO, P.R. Uma metodologia bayesiana para estudos de confiabilidade na fase de projeto: aplicação em um produto eletrônico. *Gestão da Produção*, v.4, n.3, p.305-320,1997.

QUININO, R.C.; BUENO NETO, P.R. Avaliação bayesiana de inspetores no controle estatístico de atributos. *Gestão da Produção*, v. 4, n.3, p.296-304,1997.

SEBER, G.A.F.; LEE, A.J. Linear regression analysis, second edition. Wiley series in probability and mathematical statistics,2003.

SETTIMI, R.; SMITH, J.Q. A comparison of approximate Bayesian forecasting methods for non-Gaussian time series. *Journal of Forecasting***,** v. 19, p.135–148,2000.

SPIEGELHALTER, D.J.; THOMAS, A.; BEST, N.G.; LUNN, D. WinBugs: user manual, version 1.4. Cambridge, U.K: MRC Biostatistics Unit, 2003.

ZEGER, S.L. A regression model for time series of counts. *Biometrika*, v. 75, p.621–629,1988.

ZHENG, X. Semiparametric Bayesian estimation of mixed count regression models. *Economic Letters*, v.100, p.435-438,2008.

Received in 14.08.2014

Approved after revised in 29.11.2014