




# A reliability engineering case study of sugarcane harvesters

*Um estudo de caso de engenharia de confiabilidade de colhedoras de cana-de-açúcar*

Diego Carvalho do Nascimento<sup>1</sup> , Pedro Luiz Ramos<sup>1</sup> , André Ennes<sup>1</sup>, Camila Cocolo<sup>1</sup>, Márcio José Nicola<sup>1</sup>, Carlos Alonso<sup>1</sup>, Luiz Gustavo Ribeiro<sup>1</sup>, Francisco Louzada<sup>1</sup> 

<sup>1</sup>Universidade de São Paulo – USP, Instituto de Ciências Matemáticas e de Computação, São Carlos, SP, Brasil.  
E-mail: pedrolramos@usp.br

**How to cite:** Nascimento, D. C., Ramos, P. L., Ennes, A., Cocolo, C., Nicola, M. J., Alonso, C., Ribeiro, L. G., & Louzada, F. (2020). A reliability engineering case study of sugarcane harvesters. *Gestão & Produção*, 27(4), e4569. <https://doi.org/10.1590/0104-530X4569-20>

**Abstract:** The present study aimed to analyze factors associated with the equipment failures of the sugarcane harvester, whose machineries has high importance in the harvest process and cost involved. Part of the data was originally provided by a company located in the countryside of Sao Paulo State, from two machines, collected from January 2015 to August 2017, corresponding to 2.5 crops. The overall dataset was obtained from three different sources: a stop-tracking system, which provides the track of a preventive and corrective maintenance historical of the analyzed equipment; telemetry data of the equipment, captured through embedded computer systems, installed in the machine' type under study, which provide information on its operation; and meteorological data from the Brazilian National Institute of Meteorology. Multivariate analyzes were used such as principal components and multiple regression models, therefore creating a model for prediction considering the next equipment' break, then pointing to causes of process failures. Thus, the results point to some improvements concerned with individualized reliability scheme in order to reduce the number of corrective stops given the equipment.

**Keywords:** Reliability; Multivariate analysis; Optimization in maintenance planning.

**Resumo:** O presente estudo teve como objetivo analisar os fatores associados às falhas dos equipamentos da colheitadeira de cana-de-açúcar, cujas máquinas têm grande importância no processo de colheita e nos custos envolvidos. Parte dos dados foi originalmente fornecida por uma empresa localizada no interior de São Paulo, de duas máquinas, coletadas de janeiro de 2015 a agosto de 2017, correspondendo a 2,5 culturas. O conjunto geral de dados foi obtido de três fontes diferentes: um sistema de rastreamento de parada, que fornece o rastreamento de um histórico de manutenção preventiva e corretiva do equipamento analisado; dados de telemetria do equipamento, capturados através de sistemas de computador embarcados, instalados no tipo de máquina em estudo, que fornecem informações sobre sua operação; e dados meteorológicos do Instituto Nacional de Meteorologia do Brasil. Análises multivariadas foram usadas, como componentes principais e modelos de regressão múltipla, criando, assim, um modelo de previsão considerando a próxima interrupção do equipamento e apontando as causas de falhas no processo. Assim, os resultados apontam para algumas melhorias relacionadas ao esquema de confiabilidade individualizado, a fim de reduzir o número de paradas corretivas dadas ao equipamento.

**Palavras-chave:** Confiabilidade; Análise multivariada; Otimização no planejamento de manutenção.

Received Jan. 16, 2018 - Accepted Oct. 17, 2018

Financial support: São Paulo State Research Foundation (FAPESP Proc. 2017/25971-0), CNPq and FAPESP.



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 1 Introduction

Since the Brazilian colonial period, the sugarcane industry is a relevant factor in the economic development. As such, the first exportation product was the sugarcane (Goes et al., 2011), and its industry emerged/structured itself on a large scale production focused on the global market (Furtado, 1986). Despite centuries of stagnation, this industry has not collapsed, in fact, was benefited from limited periods of growth due to foreign investments and government initiatives, as its reemergence in the 19th century due to ethanol production.

The Gross Domestic Product (GDP) originated from the agribusiness sector reached the amount of R\$ 896 billion in 2016 (USP, 2017), In 2017, from January till August the GDP of sugarcane industry chain reached an important record of R\$ 156 billion (USP, 2017). Another important fact was its contribution to the economy itself which its behavior was positive given the farming industry.

The number of sugarcane mills in Brazil grew by 171% between 2000 and 2013 with a total processing capacity of 3.6 million metric tons of sugarcane per day, said by Sant'Anna et al. (2016). From the economy perspective, the sugarcane production generates nearly 700,000 direct jobs and 200,000 indirect jobs (Almeida et al., 2007; Sant'Anna et al., 2016). The mechanization process in the sugarcane industry accelerated early in the 70's, as the result of the ethanol' emergence, as an energetic alternative to fossil fuels and, later, the increasing of the international demand for sugar market during the 90's. As a result of this development, between 2000 and 2012, Brazil and United States were responsible for over 85% of the world's annual supply of ethanol (Sant'Anna et al., 2016). According to the Center for Advanced Studies on Applied Economics (CEPEA-Esalaq), the sugar prices in the international market indicated a grew of 5% per year in the last 14 years (USP, 2017). The land costs raised and strong growth of ethanol and sugar markets made it imperative to increase investments in productivity and cost reduction.

Additionally, Brazilian' governmental initiatives as the Environment Protocol for the Sugar and Energy Sector (Green Ethanol), which was adopted by main sugarcane producers in Sao Paulo State, created a framework for better practices regards to the mechanization implementation in replacement of crop burning and manual harvesting. Currently, 92% of sugar and ethanol production in Sao Paulo State is the signatory of the Protocol.

The mechanical process in the sugarcane harvesting uses basically 2 types of equipment: harvesters and tractors for transportation. Among them, harvesters are the most important equipment not only because of its embedded high technology and cost but also due to its sensitive function through the whole production chain. The production flow must be stable and constant since the factory amount of production is settled at the beginning of each day infeasible of changes. The unavailability of the harvester due to a corrective maintenance breaks, creates production' fluctuation given a sugarcane required for manufacture, jeopardizing plant's productivity as well as rising production costs. Silva et al. (2011) discusses the importance of planning and operations' management in the agricultural area as fundamental to guarantee the supply of raw material to the industrial unit.

Based on the following scenario, the development of a predictive model for mechanical failure' reduction will optimize the availability of the harvester, contributing for the machines' productivity increase and stability of its supply of raw material (sugarcane) for productions plants. The basic principle is to keep the harvester in operation for the longest period possible and the main objective of this work is to

evaluate the variables related to corrective maintenance and develop a predictive model to mitigate those maintenances. As for the development of this predictive model, multivariate techniques such as Principal Components Analysis and Multiple Regression were used.

This article is distributed as; the following section presenting the theoretical foundation considering the principal topics listed in this work supported upon the applied field and brief descriptions of the used techniques. Dataset and Methodology will be later depicting, in details, the origin, and formation of the database. Finally, sections designated to empirical results and conclusion will be also presented in order to synthesize the contributions of this work.

## **2 Theoretical foundation**

### **2.1 Agribusiness**

Brazil is a global power in the agribusiness industry. Currently, it is the third largest exporter of agriculture products with an estimated volume of 80 billion dollars (WTO, 2016). Sugar accounts for 10 billion dollars of exports and is the second most important exported product followed by animal protein (6 billion dollars). The dynamism of the Brazilian agribusiness is noticed not only on its share of the global commodities markets but its importance to the Brazilian GDP. The agribusiness sector is responsible for 20% of the Brazilian GDP or 1,25 trillion of reais in 2016 (USP, 2017).

The Brazilian agribusiness success in the international market is due to strong competitive advantages as the vast fertile soil and adequate climate for extensive crops. Besides natural conditions, the agribusiness sector has been investing for decades on mechanization and technological improvements on existing crops. According to the U.S Department of Agriculture (USDA), Brazil has one of the highest productivity growth rates in this sector, 4.28% between 2006 and 2010 (Gasques, 2017), followed by China (3.25%), Chile (3.08%) e Japan (2.86%).

### **2.2 Preventive and corrective maintenance**

During the harvesting period, the sugarcane harvesters are submitted to two types of maintenance: the preventive (or revision) and the corrective. Both actions have the objective of prolonging the useful life and providing good working conditions to the machines, in view of their critical working conditions.

The preventive maintenance is related to a set of actions that aim to reduce the probability of machines failures for a certain period of time and also to restore the ideal conditions of their operation. This planning helps to reduce unforeseen events' occurrence and improves equipment's operational quality. On the other hand, the corrective maintenance occurs after any of the machine's components has failed. Such intervention is performed to ensure that the harvester is unavailable in the shortest possible time, but it is subject to logistic restrictions such as the existence of spare parts in stock, the need for long-distance movement, and the availability of maintenance personnel. It is also worth mentioning that this work is usually performed in full time since the harvesters operate on a 7×24 regime (twenty-four hours a day, seven days a week; constantly).

Harvesters are still subject to another type of maintenance: a general overhaul that occurs in the off-season. It is characterized as a type of corrective maintenance that aims to make the machine completely renewed. Although it is not addressed in this paper, this type of maintenance is important to understand the life cycle of agricultural machinery.

### 2.3 Regression analysis

The regression analysis plays an important role in statistical methods. Such analysis try to find the best relationship between a dependent variable and the predictors. A multiple linear regression analysis is commonly described as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon \tag{1}$$

where:  $Y$  is the dependent variable;  $\alpha$  is the intercept;  $\beta$  are the slopes;  $X_i$  the  $i$ -th predictor variable; and  $\epsilon$  is a random variable (usually assumed to be normally distributed).

A simple application of this approach is to predict future values by replacing the predictor variables into equation (1). For an overview of regression techniques, the reader is referred to Draper & Smith (2014).

Here, there is the presence of categorical variables that need to be included in the model. Such characteristic can be easily considered during the modeling as they are already implemented in standard statistical software such as R (see Faraway, 2002) for a detailed discussion).

The regression analysis is sensible to the number of variable in the model. The main goal is to find a parsimonious model that has the better predictive response without too many parameters. In this case, a common approach is to consider a technique known by stepwise (see Bendel & Afifi, 1977), such approach sequentially adds variables to the model until find the best model to represent the response. Finally, in many cases the assumption that the errors follow a normal distribution are not satisfied. For these cases, generalized linear models can be easily considered (see McCullagh, 1984) and the references therein).

### 2.4 Principal component analysis

The Principal Component Analysis (PCA) is an important multivariate technique that aims to reduce the redundancy of many of the variables without losing information. Following Morrison (1976) the principal components  $Y = (Y_1, \dots, Y_p)$  from a data set of  $p$  variables  $X' = (x_1, x_2, \dots, x_p)$  are defined as the linear combination  $Y_i = a_{i1}x_1 + a_{i2}x_2 + \dots + a_{ip}x_p$ , where the coefficients  $a_{ik}$  are the elements of the eigenvector  $a_i$  associated to the covariance matrix's eigenvalue from variables  $X'$ , the linear combination can also be represented by equation (2):

$$\begin{cases} Y_1 = a'_1 X = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p \\ Y_2 = a'_2 X = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p \\ \dots \\ Y_p = a'_p X = a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pp}x_p \end{cases} \tag{2}$$

where:  $Var(Y) = a_i'Var(X)a_i = \lambda$ ,  $Cov(Y_i, Y_j) = 0$  then  $\forall i = j$ .

The PCA transform the available variables of a certain database to a combination of  $p$  non-correlated variables. This technique allows the selection of  $k$  variables ( $k \leq p$ ) that explain most of the data variability.

The following relationship holds  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_p$ , that is,  $Var(Y_1) \leq Var(Y_2) \leq \dots \leq Var(Y_p)$ . Therefore, first principal component ( $Y_1$ ) is associated with the highest eigenvalue of the covariance matrix. This is followed by the second principal component ( $Y_2$ ) whose eigenvectors are associated to the second highest eigenvalue, then followed by  $n-2$  principal components that explain the total variability of the analyzed data. Therefore, the first principal component is considered the most important variable  $Y$  variable as it represents a major share of the total variation of the data set.

From a geometric point of view, the chosen linear combinations make the principal components represent a new set of coordinates obtained from data rotation to the direction of greater variability. The rotation process provides clearer covariance structure of the analyzed data. In practice, usually two or three principal components are selected as they represent a major portion of the variability.

## 2.5 Survival analysis

Survival analysis is an area of the statistical methods that focus on estimating the lifetime of the event of interest. In this case, the event of interest is the failure of mechanical and electronic components of the harvesting machine. An important characteristic in the survival analysis is the possibility to include partial observations related to the lifetime of the components. This occurs commonly in practice as in many cases we cannot perform the study until all components have failed. Such characteristic is usually referred as censoring (or censored data) and removing such data may include unnecessary bias in the lifetime estimation.

One of the major interests in survival analysis is the survival function, often called  $S(\cdot)$ . It is defined as the probability of the event (death) occurring after a given point in time and is represented by equation (3):

$$S(t) = P(T > t) \quad (3)$$

The survival function is assumed to start at 1 at time zero, i.e.,  $S(0) = 1$  and decreases as the time increase tending to zero at some point. In our context, every component is working at the beginning of the experiment and they are expected to fail after some period. For an overview of survival analysis, the reader is referred to Lawless (2011) and Tableman & Kim (2016).

Another important information is obtained from the hazard function, the hazard rate is defined as the rate of occurrence of the event (death or failure) conditioned to the survival until a given time  $t$ , i.e., it is the instantaneous risk in  $t$  that an individual will not survive further. The hazard rate is obtained from equation (4):

$$h(t) = -\frac{d}{dt} \log(S(t)) \quad (4)$$

The hazard function can also be represented in its cumulative form, where the value at time  $t$  represents the sum of the risk over the entire period of time prior to  $t$ .

### 3 Materials and methods

The dataset available contains nearly 60,000 occurrences, during the collection period of 30 months, corresponding to 2.5 harvests. The crops start in April 2015 until March 2017. The harvest period occurs between April and November, approximately. During the harvest' months, the harvesters work twenty-four hours over the seven days weekly (24x7). During the off-season, from December to March, the equipment has a general revision, considering renewable of the machines. Even though, in this work the defaults will be considered time-invariant. Therefore, this paper evaluated data coming from the following three-part datasets totalizing in 22 variables. The first informative part of the dataset contains the equipment's maintenance history (stop-tracking system), presenting the maintenance dataset as corrective and preventive historical data. The second part holds the equipment's telemetry data which, captured through embedded computer systems, installed in the harvesters providing information about its operation. Also, a meteorological data provided by the Brazilian National Institute of Meteorology (INMET), was added to verify relations between clime information and maintenance occurrence. Tables 1-3 show what data was available for evaluation respectively

**Table 1.** Variables related to maintenance and failures.

Variable	Explanation
Equipment	Equipment code
Problem	Problem category
Failure start	Date and time of failure occurrence
Unavailable Time	Unavailability time

**Table 2.** Telemetry variables in database.

Variable	Explanation
Vehicle	Number of equipments
Begin	Beginning of the data recording
Length	Time measured during data recording
Activity	Activity executed by the machine
Operator	Conductor of the machine
KM	Kilometer
VM	Average Speed
REV	Number of motor rotation
RPM	Rotation per minute
TRE	Time - reverse mode
TER	Time - elevator in reverse operation
TCB	Time - cutting in operation
TEL	Time - elevator in operation
TML	Time - engine in operation
TMO	Time - engine idle

**Table 3.** Variables related to meteorological data.

Variable	Explanation
Rainfall	Daily quantity of rain (in $mm^3$ )
Temperature	Daily maximum temperature
Relative humidity	Daily average relative humidity

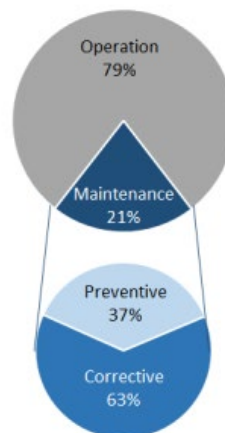
The evaluation process followed the following work-flow for each new cycle accomplished: 1) categorize the failures cause; 2) raking the failures category (duration, occurrences and recurrences); 3) compare the equipment performance; 4) identify the relationship between failures and variables; 5) calculate the time between failures. This work-flow is plotted in Figure 1.

**Figure 1.** Working flow process chain.

The following session will discuss the statistical analysis and empirical results under the adopted methodology.

#### 4 Empirical results

Initially, a descriptive analysis was performed in order to structure the data analysis. During 30 months (2.5 years' harvests), harvesters A and B had operated between 8,600 and 9,000 hours per equipment. Both machines had a total maintenance dedication time, including corrective and preventive, corresponded to 26% to 28% compared with the total time in operation. Figure 2 shows the relationship between total time under maintenance and operation. Hence, the first analysis was related to the unavailable versus operation time. Around 63% of the stop-tracking time was related to corrective maintenance, and only 37% was a programmed stop (preventive maintenance).

**Figure 2.** Preventive and corrective maintenance proportion between working time relation.

On average, the preventive stops are scheduled every 7 days, they present a pattern in interval and duration. The occurrences of the preventive versus corrective stops' distribution can be verified in Figure 3. Moreover, it is possible to verify visually a greater occurrence of non-programmed interventions (correctives) then preventives stop.

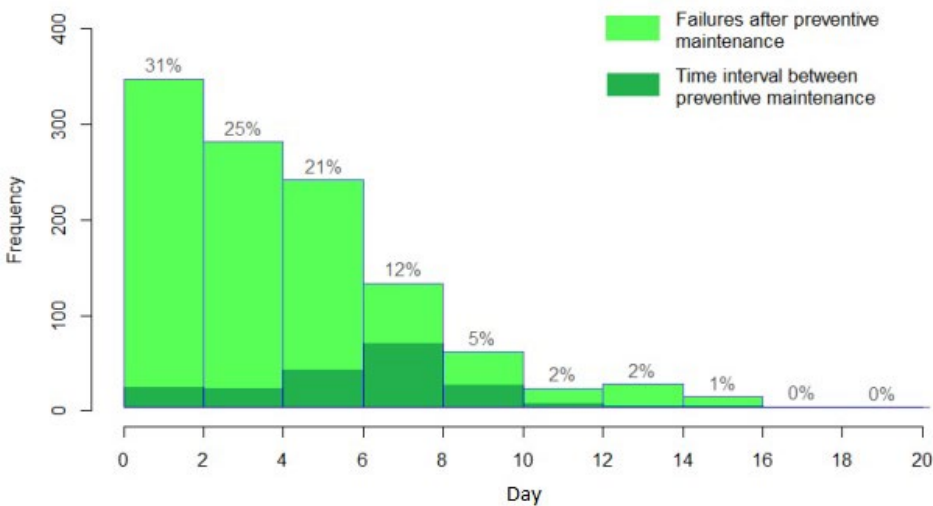


Figure 3. Histogram - preventive and corrective maintenance.

Based on the assumption of the machines become new after submitted to preventive maintenance, new corrective maintenance may not be expected close to the preventive action. Figure 4 also presents the defaults' frequency, in both machines, which events occur between 83% and 89% before 7 days. Corrective maintenances occurred frequently within a week before the preventive maintenance comes.

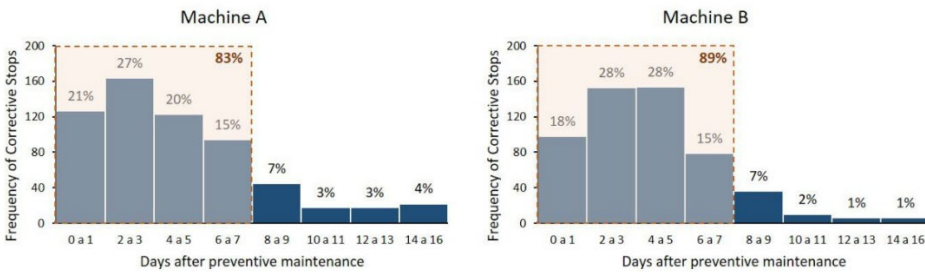


Figure 4. Occurrence of corrective maintenance.

Broadening previous analysis, the accumulated time of corrective maintenance was performed to evaluate principal defaults currently presented during the harvests cycle. Table 4 summarizes the principal failures by its number of occurrences, resulting in a ranking of the most important failures. The top five types of failure correspond more than 60% of the total number of occurrences and will be the main focus of this study.

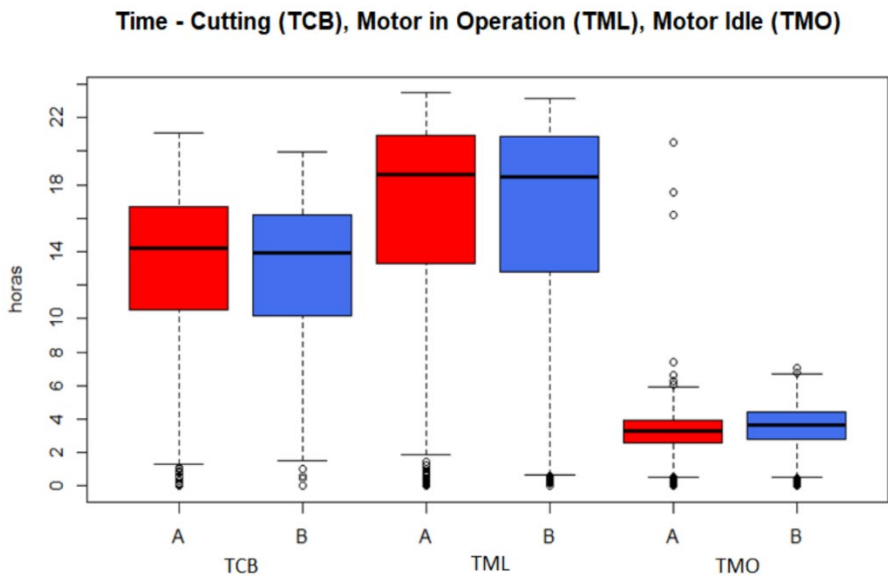


**Table 4.** Main Failures Rank - Time and Number of Occurrences.

Component	Failure	Count	Time	% Count(cum.)	% Time(cum.)
1	Crop Divider	150	344.7	16.1	15.1
2	Diesel Motor	135	492.9	30.5	36.8
3	Elevator	116	239.9	42.9	47.3
4	Pricker	113	241.1	55.0	57.9
5	Transmission	84	189.91	64.0	66.2
6	Electric	79	207.9	72.5	75.4
7	Roller	74	170.4	80.4	82.9
8	Final Drive	60	156.0	86.8	89.7
9	Base Cutters	53	118.3	92.5	94.9
10	Primary Extractor	43	80.2	97.1	98.4
11	Air Conditioning	27	35.9	100.0	100.0

Cum. represents cumulative.

Operation variables (telemetry) analysis was performed in order to evaluate the homogeneity of both machines. Figure 5 demonstrates some equipment not only similar performance but the analyzed data points to similarity of its variations. Therefore, it should be evaluated the possible implications of those variations on the overall failure study.

**Figure 5.**BoxPlot - Some Telemetry Sample Variables.

In order to support the assumption that both machines operate in a similar manner, and can be considered just as one equipment, the operation variables' averages obtained were initially submitted to the t-test. Considering a significance level higher than 10%, there is no statistical evidence that machine A and B operate differently from each other.

### 4.1 Multivariate analysis of defective parts

Initially, the distributions associated with the top three main stops were analyzed individually, as well as their relationships with the use of the machines. The database considering the corrective stops was divided by its type, calculating the intervals between the stops, and the given sensors (measurements to determine the use of the machines). For these intervals, the operational data were added, with the exception of the variables “Average Speed” and “RPM” whose mean is a better indicator.

Discussing empirical distribution, variables related to operation time and kilometers have exponential distribution decay. Thus, this perception reinforces the equipment’ operation limit preceding the next maintenance. Moreover, other variables like Average Speed and RPM have normal distributions, presenting low variance (but fat tails), as depicted in Figure 6.

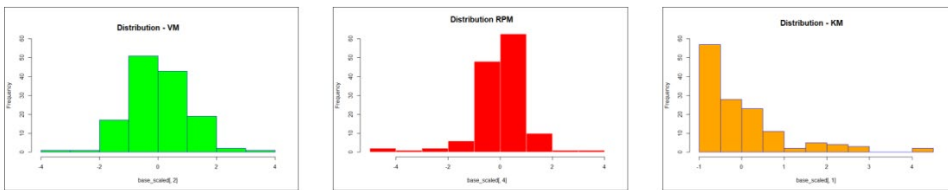


Figure 6. Frequency Distribution of some telemetry variables.

Figure 7 shows a strong correlation between most of the telemetry variables, presented on the dataset. The visual analysis of the variables exhibits linear correlations between 7 out of 10 independent variables. Most of these variables were directly, or indirectly, related to the operating time of the equipment.

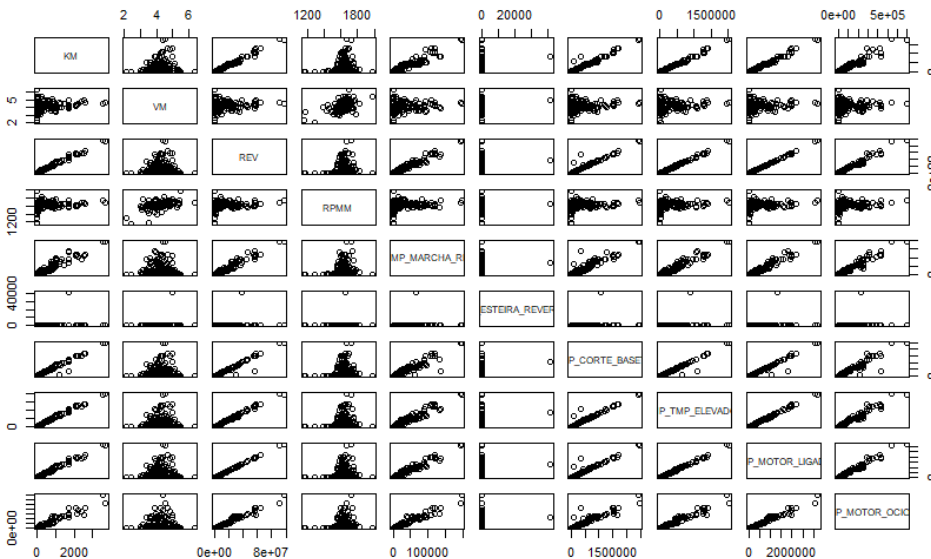
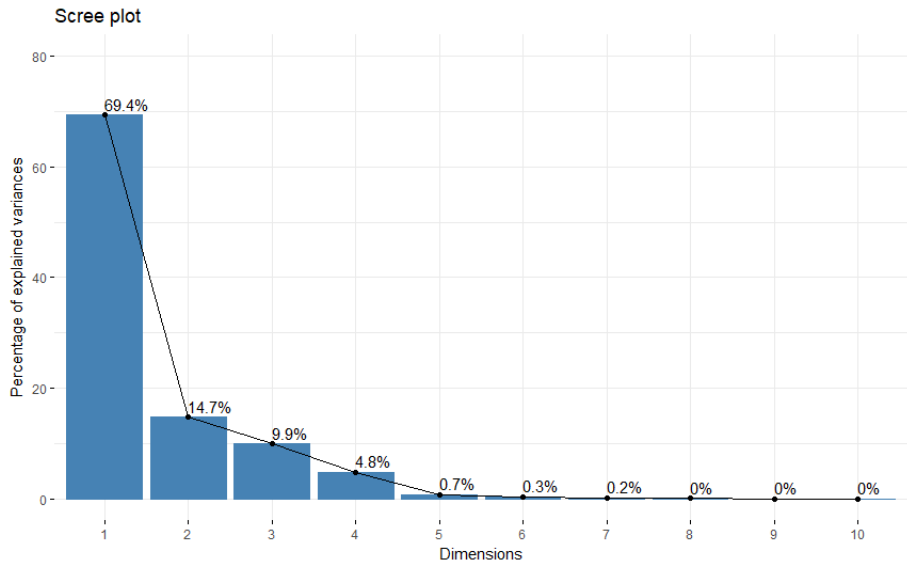


Figure 7. Scatter Plot - Relation among Independent variables.

This feature relates directly to the possibility of the data reduction under multivariate technique such as Principal Component Analysis. Figure 8 bring forward the explanatory concentration in the first three principal components, responsible for about 94% of the total dataset variance, which only the first component is responsible for 64% of the total variance.



**Figure 8.** Variance Proportion corresponding to the Principal Component explanation.

According to PCA results, the first principal component relates the variables from the operating time and kilometers, responsible for most of the variance of the data. The second principal component composed of RPM and Average Speed, whose distributions diverge from the other variables. Finally, the third principal component is composed of a single variable, which has very few observations, known as TER (Reversing Elevator time). There is a strong evidence of the relationship between principal components and variables' features in practice. Table 5 presents the % of each variable per eigenvector (component).

**Table 5.** Principal Components - Breakdown (in %).

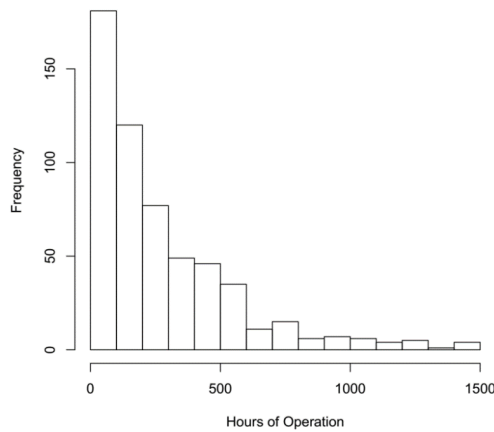
Variables	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
KM	14.2	0.0	0.0	0.6	4.9
VM	0.4	51.0	0.0	46.6	1.4
REV	14.4	0.1	0.0	0.0	0.4
RPM	0.8	45.6	3.6	49.6	0.3
TRE	13.7	0.8	0.2	0.8	23.5
TER	0.3	1.8	96.0	1.9	0.1
TCB	13.8	0.1	0.0	0.3	48.2
TEL	14.3	0.1	0.0	0.2	1.6
TML	14.4	0.2	0.0	0.0	0.3
TMO	13.9	0.5	0.1	0.1	19.3
TOTAL	100.0	100.0	100.0	100.0	100.0

After obtaining good results with space rotation and problem dimension reduction, conditions were created to develop a regression model since the dependent variables related to telemetry are now independent. This result is guaranteed by the definition of PCA in which its components are orthogonal, that is, independent.

### 4.2 Regression analysis

Once the space rotation was proceeding, a regression model was conducted to evaluate the dependence of the machines working hours (the time) between failures' type and the predict variables such as telemetry data transformed with PCA technique, the accumulated failure number (of the same case and of the all cases), the meteorological data (rainfall, temperature and relative humidity), the failure cause, the equipment, and the crop. The chosen regression method was the multivariate regression method.

For the multiple regression analysis, it was calculated only the five most common failure types. The failures times does not follow a normal distribution as can be checked in the Figure 9. This is expected as the failure times are only positive values, a powerful alternative is to consider a regression model that follows a Gamma distribution.



**Figure 9.** Hours of operation equipment's variable distribution.

According to the empirical distribution linear regression models are not suitable, therefore, the generalized linear models (GLM) are considered, in particular the dependent variable is assumed as a Gamma distribution, with an inverse link function, using an implemented package in R. The theoretical model is represented by  $Y \sim \text{Gamma}$  and the link function  $g(\cdot) = \text{Inverse}$  with.

$$\begin{aligned}
 Y = & g(\beta_0 + \beta_1 \log(\# \text{Failures}) + \beta_2 \log(\# \text{Failure.Cause}) + \beta_3 \text{PC1} + \beta_4 \text{PC2} + \beta_5 \text{PC3} + \\
 & + \beta_6 \text{Occurrence.Type} + \beta_7 \text{Equipment} + \beta_8 \text{Occurrence.Type} * \text{Equipment} + \beta_9 \text{Crop} + \\
 & + \beta_{10} \text{Occurrence.Type} * \text{Crop} + \beta_{11} \text{Occurrence.Type} * \text{Failures.Cause} + \beta_{12} \text{Rainfall} + \\
 & + \beta_{13} \text{Relative.Humidity} + \beta_{14} \text{Temperature})
 \end{aligned}
 \tag{5}$$

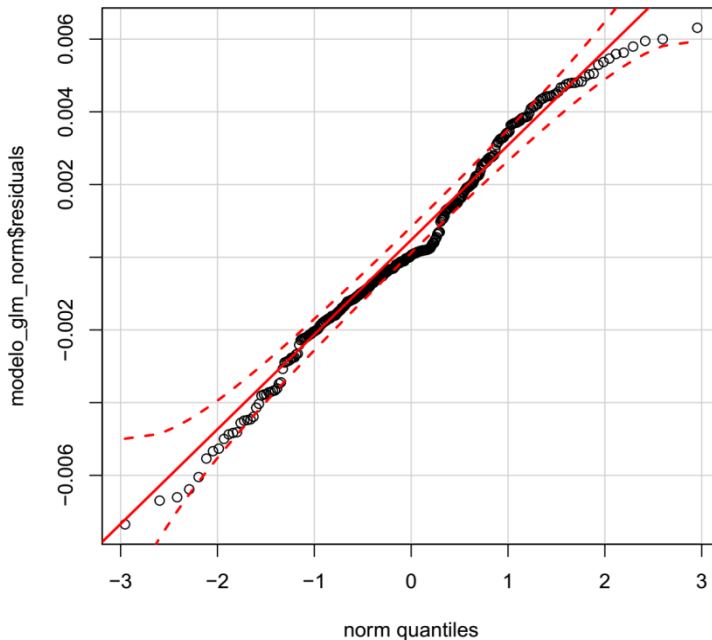
The estimates of the regression parameters are presented in Table 6, and the selected parameters can be seen in final model available in Equation 5. The significance of the parameters estimates helps us to identify the best explanatory variables. The principal components related to operating time and kilometers, as well as RPM and Average Speed showed to be statistical significant. It seems to have a difference, captured by the model, among the defaults' type, additionally among them considering the crops and its cumulative numbers of stops per crop. Additionally, the meteorological variables showed to be significant for Precipitation and Maximum Temperature averaged in the working time affecting the performance of the machines.

**Table 6.** Generalized Linear Regression Results.

	Estimate	Standart Error	t-value	
Intercept	-0.0041	0.0020	-20.405	*
Total num stop (per crop)	0.0000	0.0000	-0.1267	
Stop num given its problem (per crop)	0.0000	0.0000	0.5841	
PCA 1	0.0062	0.0005	121.715	***
PCA 2	-0.0022	0.0008	-26.014	**
PCA 3	0.0009	0.0010	0.8453	
Problem - Elevator	-0.0002	0.0007	-0.3026	
Problem - Motor	0.0014	0.0007	18.560	
Problem - Pricker	0.0000	0.0007	0.0034	
Problem - Transmission	0.0018	0.0007	26.391	**
Machine A vs B	-0.0005	0.0004	-10.202	
Crop 2016	0.0007	0.0006	11.487	
Crop 2017	0.0008	0.0006	14.346	
Problem (Elevador): Machine B	0.0000	0.0006	-0.0529	
Problem (Motor): Machine B	0.0000	0.0006	-0.0519	
Problem (Pricker): Machine B	0.0007	0.0006	12.599	
Problem (Transmission): Machine B	0.0001	0.0006	0.1569	
Problem (Elevator): Crop 2016	0.0014	0.0007	19.670	
Problem (Motor): Crop 2016	-0.0006	0.0007	-0.9446	
Problem (Pricker): Crop 2016	-0.0012	0.0008	-15.989	
Problem (Transmission): Crop 2016	-0.0008	0.0008	-0.9693	
Problem (Elevator): Crop 2017	0.0001	0.0008	0.1550	
Problem (Motor): Crop 2017	-0.0021	0.0008	-27.521	**
Problem (Pricker): Crop 2017	-0.0004	0.0008	-0.5380	
Problem (Trasmission): Crop 2017	-0.0022	0.0008	-27.002	**
Stop num: Problem (Elevator)	-0.0001	0.0000	-10.976	
Stop num: Problem (Motor)	-0.0001	0.0000	-18.756	
Stop num: Problem (Pricker)	0.0000	0.0000	-0.6498	
Stop num: Problem (Transmission)	-0.0002	0.0001	-35.968	***
Precipitation	-0.0019	0.0004	-44.227	***
Air Humidity	0.0015	0.0011	12.994	
Max Temperature	0.0042	0.0010	39.656	***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1. PCA: Principal Component Analysis.

The residuals analysis presented in Figure 10 provides evidence of the goodness of the fit as the predictive values are closed to the theoretical ones.



**Figure 10.** Model's Residuals Distribution.

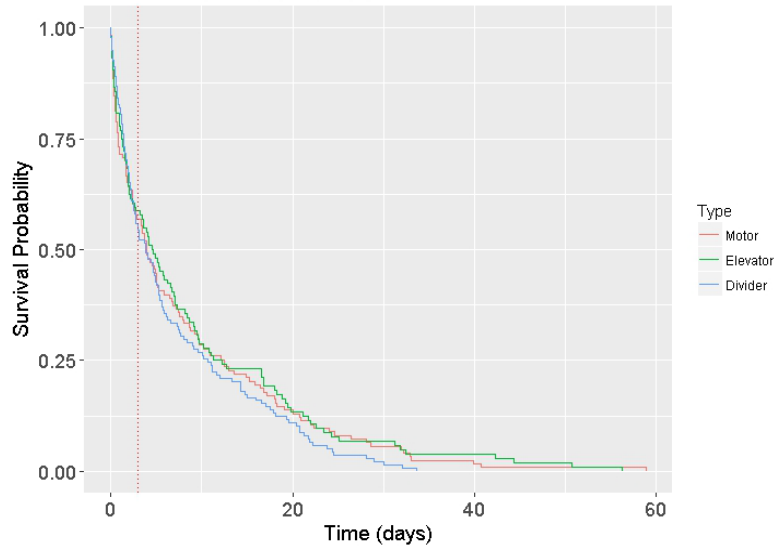
Given the presented results, the adopted generalized linear model shows satisfactory results explaining the causality in the working hours features influence. Predictions can be proceeding since the adjusted model is appropriated.

### 4.3 Survival analysis

The failure times of the three main types of corrective stops were studied under survival analysis. The components considered were: diesel engine, line divider, and elevator. Non-parametric techniques were used to study the empirical data allowing us to compare the obtained results with the ones under the regression approach.

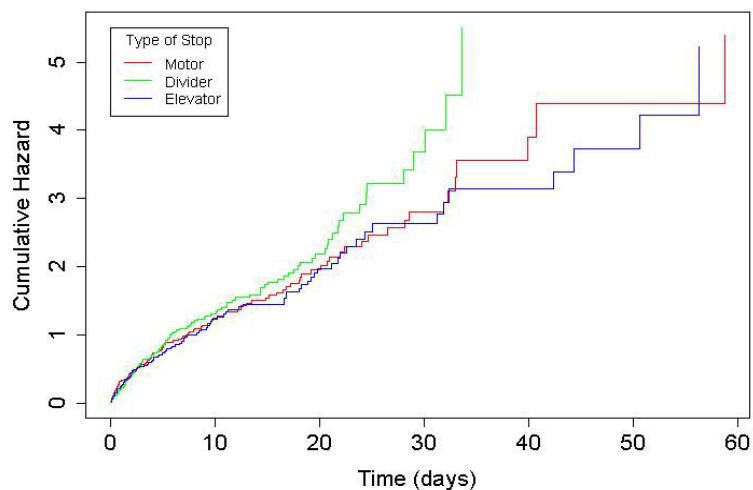
The event of interest is the occurrence of a break that leads to a corrective stop in the harvest machine. Further we assume that the corrective maintenance leaves the equipment in full working condition. Hence, the survival time is the time since the last stop that occurred due to a break of the same type of equipment. Here, we have complete data of the stops of the harvesters and no techniques were necessary to deal with censored observations. The survival curves for the three most common failures were calculated using the Kaplan-Meier estimator (see Figure 11).

From Figure 11 we observe that as the time increase the probability of that the component has not fail decrease quickly, therefore, the failures occur in a small interval. Additionally, we drew a line that represent approximately the 25% of the failure of each component which correspond to 3 days.



**Figure 11.** Survival curve by type of stop. Median represented by red line points out the expected in 50% of the cases to recurrence in 3 days.

The cumulative hazard curves for the three stop types were also calculated under the Nelson-Aalen estimator. They are presented in Figure 12. This nonparametric estimator is useful for analyzing graphically the cumulative intensity of occurrence of failure without the need to assume any kind of distribution or special characteristic of the data and is especially suitable for cases where the sample size is relatively small.



**Figure 12.** Cumulative hazard curve by type of stop.

It is remarkable how the curves for survival and cumulative hazard are similar among the failure types, indicating that there is no great difference in probability of failure among the most common types. It is also notable how survival probability falls quickly, reaching 50% of survival for all types in the period between 3 and 5 days after the last break.

## 5 Discussion

In this study, we observed that during a significant amount of time the harvesters are unavailable due to corrective stops, that is, reactive actions are performed to overcome the problems of the machine breakdown. As we have many types of failure multivariate analysis was considered in which indicated a high correlation between the variables of telemetry base, indicating that some of them can be disregarded without significant losses when describing harvesters' behavior. Further, regression analysis using generalized linear models are considered to proposed a model that is capable of planning an adequate preventive maintenance, able to reduce machine unexpected failure's probability.

Also, another possible approach was to consider the maintenance efforts on the five most frequent types of stops, which correspond about to 64% of the total. In this regard, the maintenance activities aim to increase the useful life and improve the machines operating conditions with simple preventive stops. Continuing with the findings, our results corroborate Ripoli&Ripoli (2008) that discussed the participation of factors such as agronomic, environmental and management conditions in the influence exerted on the mechanized harvesting operation. These factors can compromise the quality of the raw material, productivity, as well as the longevity of the cane field.

This finding is in agreement with the result obtained in the survival model, which indicated that in a period between 3 and 5 days after the last break, the harvester presents a 50% chance of breaking again for the same reason. This points out the lack of planning during the harvester's life cycle results in operational losses. Future works should concern on the implementation of this technology, in a form of an app developed for monitoring of and equipment's failure predict. This app could implement the regression analysis model and estimate the amount of time until next failure, testing the hypothesis of reduction in the failure's probability taking into account the founds in the PCA and meteorological data.

## Acknowledgements

Pedro L. Ramos is grateful to the São Paulo State Research Foundation (FAPESP Proc. 2017/25971-0), F. Louzada is grateful to CNPq and FAPESP.

## References

- Almeida, E. F., Bomtempo, J. V., & Silva, C. M. D. S. (2007). *The performance of Brazilian biofuels*. Paris: OECD.
- Bendel, R. B., & Afifi, A. A. (1977). Comparison of stopping rules in forward\stepwise" regression. *Journal of the American Statistical Association*, 72(357), 46-53.
- Draper, N. R., & Smith, H. (2014). *Applied regression analysis*. New York: John Wiley & Sons.
- Faraway, J. J. (2002). *Practical regression and anova using R*. Vienna: R Foundation for Statistical Computing.
- Furtado, C. (1986). *Formação econômica do Brasil*. São Paulo: Editora Nacional.
- Gasques, J. G. (2017). Wachstumstreiber der brasilianischen landwirtschaft: total efaktor produktivit" at. *EuroChoices*, (16), 24-25. <http://dx.doi.org/10.1111/1746-692X.12146>.
- Goes, T., Marra, R., Araújo, M., Alves, E., & Souza, M. O. (2011). Viabilidade econômica do biodiesel em Mato Grosso. *Revista de Política Agrícola*, 20(1), 39-51.



- Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (Vol. 362). Hoboken: John Wiley & Sons.
- McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3), 285-292. [http://dx.doi.org/10.1016/0377-2217\(84\)90282-0](http://dx.doi.org/10.1016/0377-2217(84)90282-0).
- Morrison, D. (1976). *Multivariate statistical methods*. Singapura: McGraw Hill.
- Ripoli, M. L. C., & Ripoli, T. C. C. (2008). *Sistemas de colheita* (pp. 671-693). Campinas: Instituto Agronômico.
- Sant'Anna, A. C., Shanoyan, A., Bergtold, J. S., Caldas, M. M., & Granco, G. (2016). Ethanol and sugarcane expansion in brazil: what is fueling the ethanol industry? *The International Food and Agribusiness Management Review*, 19(4), 163-182. <http://dx.doi.org/10.22434/IFAMR2015.0195>.
- Silva, J., Alves, M., & Costa, M. (2011). Planejamento de turnos de trabalho: uma abordagem no setor sucroalcooleiro com uso de simulação discreta. *Gestão & Produção*, 18(1), 73-90. <http://dx.doi.org/10.1590/S0104-530X2011000100006>.
- Tableman, M., & Kim, J. S. (2016). *Survival analysis using S: analysis of time-to event data*. New York: CRC Press.
- Universidade de São Paulo – USP. Centro de Estudos Avançados em Economia Aplicada – CEPEA. Escola Superior de Agricultura “Luiz de Queiroz” – ESALQ. (2017). *PIB do agronegócio 2017 (janeiro a dezembro de 2017)*. Piracicaba.
- World Trade Organization – WTO. (2016). *World trade statistical review 2016*. Geneva: WTO.