Predicting Product Quality from Operating Conditions Based on Multinomial Logistic Regression

Pedro Vaz[1](✉), Ana Cristina Braga[1], M. Sameiro Carvalho[1], and Gerardo Menezes[2]

[1] ALGORITMI Centre, University of Minho, 4800-058 Guimarães, Portugal
{acb,sameiro}@dps.uminho.pt
[2] Bosch Car Multimedia, Braga, Portugal

Abstract. The main objective of this study is to evaluate the influence of several industrial operating conditions in the occurrence of the defect (air bubbles) in a wet bonding process of two pieces (a glass and an electronic display). The study of the influence of operating conditions in business targets is of great interest to the community of process engineers, to ensure the long-term stability of the process, as this involves complex factors. It is also relevant for the development of Data-driven approaches to support Real-Time System Monitoring.

In this work, we propose to build a multinomial logistic regression model that can assess the probability of a bonded kit (glass and electronic display) having a defect in the filling material, taking into account process operating conditions. The study was conducted in a retrospective evaluation and consisted of 659 bonded kits randomly selected from a reference period. We evaluated multiple factors in the construction of a multinomial logistic regression model with 3 logit functions. The technique used to select variables to be included in the model, was the stepwise technique by choosing the smallest p-value for the variable entering in the model.

Of the modelling process through multinomial logistic regression have resulted five operating conditions that have statistical significance and that can contribute as an auxiliary for the prediction of a defect.

Keywords: Multinomial logistic regression · Logit · Process improvement · Process monitoring

1 Introduction

In today's competitive business environment, many companies still use conventional process evaluation strategy (also named process inspection strategy), based on Sorting Out of Failure (SOF). The massive usage of this process evaluation strategy is, in many cases, associated with the incapacity to stabilize and control industrial processes, in a mass-production context, to allow the adoption of other competitive strategies different from SOF, like Process Monitoring supported by industrial statistics. The conventional SOF strategy, besides giving the safety sensation of controlling the process, in many

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12952, pp. 539–551, 2021.
https://doi.org/10.1007/978-3-030-86973-1_38

cases, it masks the incapacity to improve the process in order to stabilize it and distracts the company, neglecting efforts to reduce variability in processes. The conventional strategy entails the costs related to 100% inspection method and is frequently associated with the higher scrap rate, which results in higher Internal Defect Costs (IDC). These costs in some cases could be higher than the ones involved in process improvements activities conducing to lower process variation.

Although, in recent years, companies are producing data exponentially, which could help in reducing variability in processes, they continue facing challenges in dealing with big data issues for rapid decision-making to achieve faster improved productivity [1]. Many manufacturing systems are not ready to manage big data due to the lack of smart analytic tools. However, the Industry 4.0 proposing the predictive manufacturing in the future industry [1] will push harder by data-driven methods.

In the era of Industry 4.0, Industrial Cyber-Physical Systems (ICPSs) are the cornerstone research subject and has become a worldwide research focus [2]. ICPSs integrate physical entities with cyber networks to build systems that can work more harmoniously, benefiting from integrated design and system-wide optimization. ICPS is a transdisciplinary subject in the infancy stage [2]. "Real-Time System Monitoring" is one of the directions of ICPS research, which in turn is also a transdisciplinary subject in its early stages. Data-driven approaches [3] are promising alternatives to simplify the design for ICPSs. According with Yin et al. [2] there are two key dimensions for data-driven approaches. The first dimension is the assumption of data structures. The second dimension is that of the estimation of undetermined parameters of data structures, using here statistical models.

The case study is a wet bonding process, which is divided in four operations. The first operation is the immobilization of a kit with display in a conveyor. The second operation is the picking up of a glass from a base. The third operation is the measurement, with a laser device integrated in the robot, of fiducial coordinates allocated at the display and the kit. Then, the measurement of distances between those fiducial and the arm of the robot, grabbing a mask with a glass. The fourth operation is the robot making the necessary movements in order to achieve the best position to assemble, through bonding, the glass with the display. Finally, at the end of the process a qualified operator controls all assembled (bonded) parts for quality reasons. In case of defect (air bubble between display, filling and glass), the operator registers the occurrence in a database. For the purpose of recording defects, process experts divided the area of the display in ten zones (A, B, C, D, 1, 2, 3, 4, 5 and 6). In Fig. 1 it is possible to observe a personal illustration of those zones in a display.

From a statistical exploratory analysis of operating conditions´ historical data in the wet bonding process, it was possible to observe that defect occurrence can vary with the type of robot where it is assembled, the spatial position of the robot, and if, during the assembly process, an error occurs at the work stations "Automatic Optical Inspection" (AOI) or "FLC". The last one is a shuttle to take out the parts that did not assemble or had some problem during the assembling.

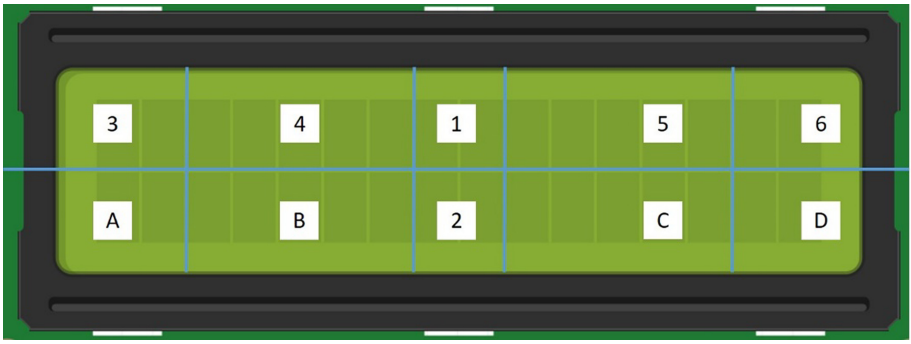

Fig. 1. Illustration of zones of the defect classification. (Source: Adapted fictitious image, for illustration purposes only).

Better understanding the operational conditions and the structures associated with them, may help the process engineers to diagnose and prevent more accurately the occurrence of operating conditions associated with the defect, thus obtaining successful stability.

In statistics, logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable (a dependent variable that can take on a limited number of categories) based on one or more predictor variables. Multinomial logistic regression is defined for a response variable with three or more discrete outcomes. It is an extension of logistic regression based on the binomial distribution (i.e., where the response has only two outcomes).

Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership [4].

In this work, multinomial logistic regression is used to predict the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale).

2 Methodology

This section presents the statistical method, the sample data, variables and model that support this work.

2.1 Statistical Method

In this work, we propose to build a multinomial logistic regression (MLR) model that could assess the association of the defect occurrence (air bubble in filling material) with process operational conditions measures and some discrete characteristics (errors occurrence in work stations “AOI” or “FLC”).

Often logistic regression is referred to as a binary classifier, since there are only two outcomes [5]. In cases, where the target variable can have multiple outcomes, as the case study of this work, we face a multi-level classification problem. In these cases, MLR is often the preferred method of estimation.

MLR compares multiple groups through a combination of binary logistic regressions. The group comparisons are equivalent to the comparisons for a dummy-coded dependent variable, with a group used as the reference. MLR provides a set of coefficients for each of the two comparisons.

Adapting the example of Braga et al. [4] if the dependent variable Y has four categories, 0 is the code for the reference category, 1, 2 and 3 the remaining categories. For a model with four categories, we have three logit functions, one for Y = 1 vs Y = 0, other for Y = 2 vs Y = 0 and for Y = 3 vs Y = 0.

In other words, in the multinomial logistic regression, we develop k-1 models for k classes and a set of independent binary regression, also called logits. Essentially, in context of ordinal logistic regression we have two classes, so we develop one model.

When we extend to a multi-class problem, it means that for k classes, we can develop k-1 models, using multinomial logistic regression. We develop our logit functions according to Braga et al. [4].

Considering $\mathbf{x}$ to be the vector of covariates of length $p + 1$ with $x_0 = 1$ for entering with the constant term. We will designate the three logit functions by:

$$g_1(\mathbf{x}) = \ln\left(\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{1p}x_p = \left(1, \mathbf{x}'\right)\boldsymbol{\beta}_1 \quad (1)$$

$$g_2(\mathbf{x}) = \ln\left(\frac{P(Y=2|\mathbf{x})}{P(Y=0|\mathbf{x})}\right) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \ldots + \beta_{2p}x_p = \left(1, \mathbf{x}'\right)\boldsymbol{\beta}_2 \quad (2)$$

$$g_3(\mathbf{x}) = \ln\left(\frac{P(Y=3|\mathbf{x})}{P(Y=0|\mathbf{x})}\right) = \beta_{30} + \beta_{31}x_1 + \beta_{32}x_2 + \ldots + \beta_{3p}x_p = \left(1, \mathbf{x}'\right)\boldsymbol{\beta}_3 \quad (3)$$

It follows that the four conditional probabilities for each category of dependent variable, given the vector of covariates, are:

$$P(Y=0|\mathbf{x}) = \frac{1}{1 + \exp(g_1(\mathbf{x})) + \exp(g_2(\mathbf{x})) + \exp(g_3(\mathbf{x}))} \quad (4)$$

$$P(Y=1|\mathbf{x}) = \frac{\exp(g_1(\boldsymbol{x}))}{1 + \exp(g_1(\mathbf{x})) + \exp(g_2(\mathbf{x})) + \exp(g_3(\mathbf{x}))} \quad (5)$$

$$P(Y=2|\mathbf{x}) = \frac{\exp(g_2(\mathbf{x}))}{1 + \exp(g_1(\mathbf{x})) + \exp(g_2(\mathbf{x})) + \exp(g_3(\mathbf{x}))} \quad (6)$$

$$P(Y=3|\mathbf{x}) = \frac{\exp(g_3(\mathbf{x}))}{1 + \exp(g_1(\mathbf{x})) + \exp(g_2(\mathbf{x})) + \exp(g_3(\mathbf{x}))} \quad (7)$$

Following the convention given to the binary model, we can do $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$ for j = 0, 1, 2, 3 each being a function of a vector with $3(p + 1)$ parameters, $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \boldsymbol{\beta}'_3\right)$.

A general expression for the conditional probability for a model whose dependent variable has four categories, can be given by:

$$P(Y = j|\mathbf{x}) = \frac{\exp\left(g_j(\mathbf{x})\right)}{\sum_{k=0}^{3} \exp(g_k(\boldsymbol{x}))} \tag{8}$$

where the vector $\boldsymbol{\beta}_0 = 0$ and $g_0(\mathbf{x}) = 0$.

To construct the likelihood function, it should create four binary variables coded by "0"0 and "1"1 to indicate the group member of an observation. The variables are coded as follows:

- If $Y = 0$ then $Y_1 = 0$, $Y_2 = 0$ and $Y_3 = 0$;
- If $Y = 1$ then $Y_1 = 1$, $Y_2 = 0$ and $Y_3 = 0$;
- If $Y = 2$ then $Y_1 = 0$, $Y_2 = 1$ and $Y_3 = 0$;
- If $Y = 3$ then $Y_1 = 0$, $Y_2 = 0$ and $Y_3 = 1$;

Independently of the value that Y takes, $\sum Y_i = 1$. The conditional likelihood function for an independent sample of n observations is:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\pi_0(\boldsymbol{x}_i)^{y_{0i}} \pi_1(\boldsymbol{x}_i)^{y_{1i}} \pi_2(\boldsymbol{x}_i)^{y_{2i}} \pi_3(\boldsymbol{x}_i)^{y_{3i}}\right] \tag{9}$$

Making the logarithm and using the fact that the $\sum y_i = 1$ for each i, the logarithm of the likelihood function is [4]:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_{1i} g_1(\boldsymbol{x}_i) + y_{2i} g_2(\boldsymbol{x}_i) + y_{3i} g_3(\boldsymbol{x}_i) - ln\left[1 + e^{g_1(\boldsymbol{x}_i)} + e^{g_2(\boldsymbol{x}_i)} + e^{g_3(\boldsymbol{x}_i)}\right]\} \tag{10}$$

The likelihood equations are the result of the first partial derivatives of $L(\boldsymbol{\beta})$ to the $3(p + 1)$ unknown parameters. Using the notation $\pi_{ji} = \pi_j(\mathbf{x}_i)$, the form of these equations is:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}\left(y_{ji} - \pi_{ji}\right) \tag{11}$$

for $j = 1,2,3$ and $k = 0, 1, 2,\ldots, p$; with $x_{0i} = 1$ for each case.

The maximum likelihood estimator (MLE), $\widehat{\boldsymbol{\beta}}$ is the result of the resolution of this set of equations equal to 0 in order to $\boldsymbol{\beta}$.

2.2 Data and Sample

To analyse factors, or operating conditions that favour the absence of defects, or deliver products with the contracted customer requirements, we test aggregate data of 659 assembled units, within the three shifts in operation in the device named "AOI10_005". The sample pertains to the time interval from 20.03.2018 until 02.05.2018. In this period the internal rejection rate (IRR) of the process stayed low and with lower variation (see Fig. 2).

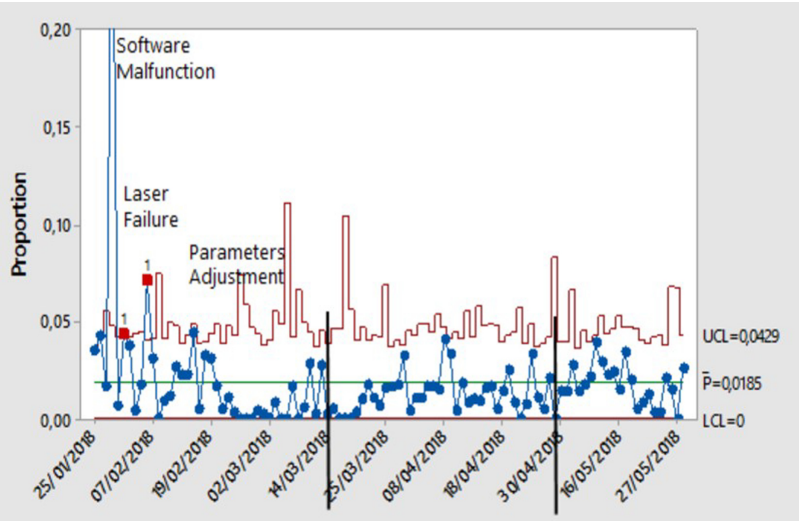

Fig. 2. AOI10_005 Minitab screen shot of Internal Rejection Rate (IRR) lowest and stable period.

2.3 Variables

In this study, the occurrence of the defect (air bubble between the sandwich composed by the display, filling and glass) was treated as the output variable. The available monitoring variables (input variables) are shown, in Table 1.

Table 1. Model variables definition.

Item	Variable	Definition	Code
1	Bubble in zone 1	Dummy variable that reflects defect (bubble) occurrence at zone 1 of the display	0 - no bubble 1 - buble
2	Bubble in zone A	Dummy variable that reflects defect (bubble) occurrence at zone A of the display	0 - no bubble 1 - buble
3	Bubble in zone D	Dummy variable that reflects defect (bubble) occurrence at zone D of the display	0 - no bubble 1 - buble
4	*AOI10_004*	The variable that reflects the unit was processed at a robot named AOI10_004	0 - no 1 - yes
5	*AOI10_005*	The variable that reflects the unit was processed at a robot named AOI10_005	0 - no 1 - yes
6	Rework	Variable indicating if the unit was submitted to recursion	0 - no 1 - yes

(*continued*)

Table 1. (*continued*)

Item	Variable	Definition	Code
7	Offset P1	The linear distance between point 1 reference at display and laser meter head included in the robot terminal that holds the tool, holding the glass to bond with the display	Milimeters (mm)
8	Offset P2	The linear distance between point 2 reference at display and laser meter head included in the robot terminal that holds the tool, holding the glass to bond with the display	mm
9	Offset P3	The linear distance between point 3 reference at display and laser meter head included in the robot terminal that holds the tool, holding the glass to bond with the display	mm
10	Offset P4	The linear distance between point 4 reference at display and laser meter head included in the robot terminal that holds the tool, holding the glass to bond with the display	mm
11	Alignment X	The linear displacement along the X-axis of the robot axis system	mm
12	Alignment Y	The linear displacement along the Y-axis of the robot axis system	mm
13	Alignment R	The linear displacement along the R axis of the robot axis system	
14	Distance display	The linear distance between display and laser meter head included in the robot terminal that holds the tool, holding the glass to bond with the display	Mm
15	AOI error	Dummy variable that reflects if, during the assembly process, occurred an error at the AOI (Automatic Optical Inspection) of particle inspection	0 - no 1 - yes
16	FLC error	Dummy variable that reflects if, during the assembly process, occurred an error at 2D reading code	0 - no 1 - yes

2.4 Model

Units with bubbles are the response or target that we wish to predict. As previously seen in the statistical method section, we generically refer to the response as Y. The P1 is a predictor we name as X_1. Likewise, name P2 as X_2, and so on. We can refer to the predictor vector collectively as $\mathbf{x}= (x_1x_2x_3\ldots,x_p)$.

With the logit function defined in Sect. 2.1, we can make predictions of Y (bubbles) at new points $\mathbf{x}$ (robot offsets and process discrete characteristics).

In our case, the reference category it is a normal situation, which reflects the scenario where no bubble appears in any zone of the display. For the reference category, the variable response Y has the code 0. In terms of data set structure, it corresponds to the line where we have 0, 0, 0 for the variable Y_1 (Bubble in zone 1), Y_2 (Bubble in zone 2) and Y_3 (Bubble in Zone 3) (Table 2).

Here, the response variable can have multiple outcomes. We are facing a multi-level classification problem. Specifically, the problem is a four-level classification problem where we investigate, the response variable (Y) coded 0 = no Bubbles, 1 = Bubble in zone 1, 2 = Bubble in zone A and 3 = Bubble in Zone D. Therefore, our response variable Y has four categories. The reason we checked for defects in zones 1, A and D of the kit is because they represent the top 3 in terms of defect occurrence. This information was obtained during the statistical exploratory analysis.

Actually, the software installed at the end of line quality control workstation, only permits the operator to register, in case of bubbles, one zone of occurrence. The operator assumes as the defect occurrence zone the one with the highest incidence, in case of defect in more than one zone.

Table 2. Definition of scenarios or levels.

Scenario	Y_1(Zone 1)	Y_2(Zone A)	Y_3(Zone D)	Y
No bubble/any zone	0	0	0	0
Bubble in Zone 1	1	0	0	1
Bubble in Zone A	0	1	0	2
Bubble in Zone D	0	0	1	3

In Table 2 are exposed the scenarios defined for defect occurrence according to the operator classification at the production line aligned with each of the four model categories. For a model with four categories, we will have three logit functions.

The general model equation for each logit is expressed as:

$$\begin{aligned}&\ln[P(NS\ level = i)/P(NS\ level = 0)] = intercept_i + b_{1_i} \times AOI10_{004} + b_{2_i} \times \\ &AOI10_{005} + b_{3_i} \times \text{reworked} + b_{4_i} \times \text{offsetP1} + b_{5_i} \times \\ &b_{6_i} \times \text{offsetP3} + b_{7_i} \times \text{offsetP4} + b_{8_i} \times \text{AlignX} + b_{9_i} \times \text{AlignY} \\ &b_{10_i} \times \text{AlignR} + b_{11_i} \times \text{distdisplay} + b_{12_i} \times \text{AOIerror} \\ &b_{13_i} \times \text{FLCerror}\end{aligned} \tag{12}$$

(with $1 \leq i \leq 3$ and reference NS level = 0) and b the estimates of parameters β.

3 Results

In this section, we present the results of the computational statistical analysis. All statistical analysis were carried out in R software using package *nnet* with the function *multinom*.

To analyse the accuracy of the model we use a misclassification matrix constructed in R using the command *table* with *y*_actual and *y*_predict. The results are presented on Table 3.

Table 3. Classification table

	y_predict			
y_actual	0	1	2	3
0	508	4	3	0
1	34	15	2	0
2	28	2	19	2
3	32	0	6	4

The sum of correct classifications is 546 and the total number of classifications is 659, so the measure of accuracy is 82.8%.

Table 4 presents the results of coefficient values obtained from the application of the model to the case study.

Table 4. Coefficient estimation for multinomial logistic regression.

Logit	Component	Estimate,$\hat{\beta}$	Std. Error	*p*-value	*Signif. Level*
	Intercept	−29.608	0.786	< 2.2E−16	0.001
1	AOI10_004	−15.191	0.956	< 2.2E−16	0.001
	AOI10_005	−14.417	0.731	< 2.2E−16	0.001
	Rework	−2.118	0.685	0.002	0.01
	OffsetP1	7.254	5.932	0.221	
	OffsetP2	−10.598	6.929	0.126	
	OffsetP3	−2.468	1.614	0.126	
	OffsetP4	9.938	3.461	0.004	0.01
	AlignX	−6.136	20.870	0.768	
	AlignY	3.509	17.330	0.839	
	AlignR	108.581	48.244	0.024	0.05
	Dist. Display	2.537	0.115	< 2.2E−16	0.001

(*continued*)

Table 4. (*continued*)

Logit	Component	Estimate,$\hat{\beta}$	Std. Error	*p*-value	*Signif. Level*
	AOIError	3.266	0.812	5.76E−05	0.001
	FLCError	1,326	1.027	0.196	
	Intercept	23.182	2.602	< 2.2E−16	0.001
2	AOI10_004	12.409	1.415	< 2.2E−16	0.001
	AOI10_005	10.772	1.427	4.52E−14	0.001
	Rework	−1.226	0.586	0.036	0.05
	OffsetP1	−2.353	5.615	0.675	
	OffsetP2	−0.349	6.606	0.957	
	OffsetP3	−1.705	1.399	0.223	
	OffsetP4	3.570	3.349	0.286	
	AlignX	29.255	22.527	0.194	
	AlignY	−24.992	17.811	0.160	
	AlignR	107.561	39.133	0.006	0.01
	Dist. Display	−2.641	0.274	< 2.2E−16	0.001
	AOIError	3.866	0.650	2.75E−09	0.001
	FLCError	2.499	0.854	0.003	0.01
	Intercept	20.625	2.588	1.61E−15	0.001
3	AOI10_004	10.724	1.510	1.25E−12	0.001
	AOI10_005	9.902	1.512	5.82E−11	0.001
	Rework	−0.947	0.510	0.063	
	OffsetP1	5.761	5.483	0.293	
	OffsetP2	−5.549	6.488	0.392	
	OffsetP3	−1.711	1.641	0.297	
	OffsetP4	4.283	3.425	0.211	
	AlignX	34.742	22.407	0.121	
	AlignY	32.300	17.852	0.070	
	AlignR	16.595	41.904	0.692	
	Dist. Display	−2.067	0.266	8.30E−15	0.001
	AOIError	2.277	0.699	0.001	0.01
	FLCError	2.480	0.913	0.006	0.01

Source: *multinom* function of R software *nnet* package

In Table 4 it is possible to observe the statistical significance coefficients with a maximum value of 0.05, meaning a minimum confidence interval of 95%.

Taking into account the results obtained, for each logit or model equation we can interpret the estimated coefficients. Also important is the sign of the coefficient. It will tell us if it affects positively or negatively the log odds in analysis (defect occurrence). For this case study it is of relevance to know the coefficients that affect positively the log odds of defect occurrence in order to prevent it. Below are presented the statistically significant variables that affect positively the defect occurrence in a specific zone of the display.

For discrete variables:

- Errors in the AOIs impact the appearance of bubbles in the 3 zones under study, with a confidence interval of 99.9% for zones 1 and A and 99% for zone D;
- FLC error seems to favour the appearance of bubbles in zones A and D with 99% of confidence.

For continuous variables:

- Distance to the display seems to be strongly associated with the likelihood of bubbles appearing in all zones studied with 99.9% confidence for the 3 zones;
- Alignment in R appears to be associated with defects in zones 1 and A with 95% confidence interval and 99%, respectively;
- OffsetP4 seems to impact the occurrence of bubbles in zone 1 with a 99% confidence interval;

4 Discussion

As mentioned, we need to understand the inference in order to be able to understand the relationship between the independent variable and the dependent variable and that will help in policymaking, consequently taking better decisions.

From computational statistical analysis (SA) results were possible to derive a set of target conditions for the process variables, namely:

a. **Process discrete variables target conditions derived from SA results**:

   I. Application of a process reaction limit (stop process) equal one in case of AOI error or FLC error.

b. **Process continuous variables target conditions derived from SA results**:

   I. Analyse the concept of display base retention (immobilization) to reduce positional variation and expand the estimation of the robust variation for the other continuous variables (offsets) without statistical significance.
   II. Estimation of a combined robust variation range for the continuous variables with statistical significance (see Table 5).

III. Process monitoring with alarm in case of statistically significant variables laying outside the combined robust variation range.
IV. Standardized and controlled maintenance to comply with robust variation band.

Table 5. Combined robust variation range for the continuous variables in the model.

Limits	Variables			
	Offset P4	Alignment Y	Alignment R	Dist. display
Min	−1.004	−0.020	−0.013	15.290
Max	0.596	0.030	0.010	16.370

The developed statistical model with the necessary tuning can be applied to every process of the company when it is necessary to predict and monitor categorical response variables in a complex operating conditions environment. The model can also be the data structure of a "Real-Time System Monitoring" contributing to ICPS research.

5 Conclusion

This study had as its objective the analysis of the influence of several industrial operating conditions in the occurrence of the defect (air bubbles) in a wet bonding process of two pieces (a glass and an electronic display).

In this paper, we present the development and application of a multi-class classification statistical model, for predicting product quality from operating conditions, based on multinomial logistic regression. The study ran, in a real industrial process.

From data associated with 16 variables, we conceived a statistical model to estimate robust operating conditions.

According to the results obtained it was found that, to evaluate the probability of an assembled kit (glass and electronic display) having a defect (air bubble) in the filling material during the bonding process it is important to monitor the following process variables: occurrence of errors in workstations "AOI" and "FLC", distance to display, Alignment in R and OffsetP4.

This model contributes to better understanding the operating conditions of the bonding process associated with the occurrence of the defect and to estimate the robust operational conditions combination of these 5 variables. Therefore, we think that it will be a tool to predict more accurately the occurrence of a defect on the bonding process.

Furthermore, our model integrated in a "Real-Time System Monitoring" may represent an alternative to build a data-driven white-box soft-sensor [6–8], in this way contributing to future ICPS research.

Acknowledgements. This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

1. Lee, J., Kao, H., Yang, S.: Service innovation and smart analytics for Industry 4.0 and big data environment. ScienceDirect Procedia CIRP **16**(3), 3–8 (2014)
2. Yin, S., Rodríguez-Andina, J.J., Jiang, Y.: real-time monitoring and control of industrial cyber-physical systems: with integrated plant-wide monitoring and control framework. IEEE Ind. Electron. Mag. **13**(4), 38–48 (2019)
3. Kano, M., Nakagawa, Y.: Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry. Comput. Chem. Eng. **32**, 12–24 (2008)
4. Braga, A.C., Urzal, V., Ferreira, A.P.: Orthodontics diagnostic based on multinomial logistic regression model. In: Murgante, B., et al. (eds.) ICCSA 2013. LNCS, vol. 7971, pp. 585–595. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39637-3_46
5. James, L.: Logistic Regression in R Tutorial (2018). https://www.datacamp.com/community/tutorials/logistic-regression-R#rstats, Accessed 9 Apr 2021
6. Shang, C., Fang, Y., Huang, D., Lyu, W.: Data-driven soft sensor development based on deep learning technique. J. Process Control **24**, 223–233 (2014)
7. Ookita, K.: Operation and quality control for chemical plants by soft sensors (in Japanese). CICSJ Bull. **24**, 31–33 (2006)
8. Sharma, S., Tambe, S.S.: Soft-sensor development for biochemical systems using genetic programming. Biochem. Eng. J. **85**, 89–100 (2014)