# Using Random Forest to Interpret Out-of-Control Signals

**Esteban Alfaro-Cortés[1], José-Luis Alfaro-Navarro[2], Matías Gámez[1], Noelia García[2]**

[1] Quantitative Methods and Socio-Economic Development Group, Institute for Regional Development (IDR), University of Castilla-La Mancha (UCLM), Albacete, Spain; esteban.alfaro@uclm.es; matias.gamez@uclm.es

[2] Faculty of Economics and Business Administration, University of Castilla-La Mancha, Albacete, Spain; joseluis.alfaro@uclm.es; noelia.garcia@uclm.es

*Abstract: Statistical quality control procedures have become essential practices to ensure competitiveness in any manufacturing process. Since the quality of manufactured goods usually depends on several correlated characteristics, statistical multivariate techniques are needed to detect and analyze out-of-control situations. The difficulties in the interpretation of those out-of-control observations in multivariate control charts have motivated the development of different techniques in order to determine the variable or variables that have motivated the changes in the process and, in case of more than one variable as responsible of the change, to evaluate their contribution. Specifically, these techniques are mainly based in two alternatives, one that considers the* $T^2$ *decomposition and other related to the application of classification techniques. The application of this latest techniques includes increasingly sophisticated methods, being the most usual alternative based on the application of Artificial Neural Networks. In this paper, we propose Random Forest as a powerful classification technique in statistical process control, considering a wide range of different situations in the function of the type of change and the magnitude of the correlation coefficient between variables. Moreover, the performance of Random Forest is analyzed in comparison with the results obtained from the application of Artificial Neural Networks to try to find out in which cases the superiority of Random Forest can be supported.*

*Keywords: Hotelling* $T^2$*; out-of-control; signals interpretation; Random Forest; Artificial Neural Networks*

## 1 Introduction

The development of the industrial procedures has caused quality to play a crucial role as an aspect to be considered by consumers, even more, important than the price of a product. Nowadays the differences in prices between products with

similar characteristics are smaller than before and, therefore, quality has become the main criterion for consumers in their decision processes.

Thus, quality control has increased its importance in production processes and the application of statistical techniques has emerged as the main method to carry it out. In addition, it is necessary taking into account that the quality of manufactured goods depends usually on several correlated characteristics and, therefore, multivariate techniques are needed to detect out-of-control situations. Among the range of statistical multivariate techniques, Hotelling's $T^2$ is one of the most widely one used in the industrial process due to the ease of its implementation and the good results it provides when the changes in the quality characteristics are not small [1-2]. Assuming the data are independent and normally distributed, the Hotelling's $T^2$ statistic for the sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is calculated, when the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the normal distribution are known, as:

$$T_i^2 = (\mathbf{x_i} - \mu)\Sigma^{-1}(\mathbf{x_i} - \mu)^t \tag{1}$$

where $\mathbf{x}_i$ represents a *p*-dimensional vector of measurements made on a process at time period *i*.

Statistical process control based on control charts relies on showing the statistics calculated from equation 1 together with the control limits that allow the detection of out-of-control observations. In this case, the upper control limit of the $T^2$ control chart is obtained as:

$$UCL(T^2) = \frac{p(n+1)(n-1)}{n^2 - np} F_{\alpha,p,n-p} \tag{2}$$

where α is the probability of false alarm for each point plotted on the control chart and $F_{\alpha,p,n-p}$ is the percentile (1- α) of the F distribution with *p* and *n-p* degrees of freedom. The lower control limit is usually set to zero. However, if the sample size (*n*) is higher than 100, the upper control limit is usually approximated by:

$$UCL(T^2) = \frac{p(n-1)}{n-p} F_{\alpha,p,n-p} \tag{3}$$

Thus, the $T^2$ statistics are plotted together with the control limits on the $T^2$ chart and if one or more than one of the *n* points are out of the boundaries, the process is said to be out of control and the specific causes of such variation should be investigated.

If an out-of-control signal is detected, the next step would be to look for the variable or variables that are responsible for the anomaly so that the necessary corrective procedures can be undertaken.

But it is precisely at this point that the main limitation in the implementation of Hotelling's $T^2$ control charts arises, becoming one of the reasons that have limited the use of this technique in industrial processes.

To solve this problem, several alternatives have been developed in the specialized literature, based mainly on the $T^2$ decomposition and the application of classification techniques. Both procedures allow measuring the contribution of each variable and, therefore, determine the variable or variables that have motivated the changes in the process [3-5]. Moreover, it is necessary to highlight that the use of univariate control charts would lead to losing the multivariate point of view and not considering the correlation between the variables that in some cases is the key in the out of control situation.

Since the first proposal of [6] on the use of classification techniques based on discriminant analysis to detect the cause(s) of an out-of-control signal, this task can be addressed as a classification problem where the output is the variable or the variables responsible of that signal and the inputs are the values of the variables and the $T^2$ statistic. This initial proposal has triggered a prolific line of research on the use of different classification techniques, which has also been driven by the development of data mining techniques in recent years. In this sense, we should emphasize the works [7-10] that use artificial neural networks as an effective tool to interpret out-of-control signals in multivariate control charts. Moreover, [11-14] uses neural networks for pattern recognition in control charts as another kind of out of control situation; [15] uses neural networks as a statistical process control procedure; and [16] proposes an ensemble of neural networks to improve the diagnosis of out of control signals. On the other hand, decision trees [17-18] or ensemble trees [19-21] have been also used in the out-of-control signals interpretation. Finally, [22] compares linear discriminant analysis, classification trees, neural networks, and boosting trees as classification techniques to determine the cause of change in out of control situations detected by the Hotelling's $T^2$ control chart, concluding that the best performance is achieved with the ensemble trees using boosting.

The common procedure in these works can be seen as a combination of multivariate control charts with classification techniques. First, a multivariate control chart is used and once the chart provides an out-of-control signal, the classification technique is used to determine which variable or variables have changed. This procedure allows a clearer interpretation of the out-of-control observations.

In this paper, we propose the application of random forest as an alternative to the most widely applied technique to this problem so far that is, artificial neural networks. Since the first appearance of the random forest method in 2001 [23], this tree ensemble method has grown in popularity, and this is currently the classification technique implemented by default in massive data processing systems (Big Data Analysis) due to its good behavior both in terms of speed and ability to handle large samples of data.

The superiority of an ensemble of trees, such as random forest, over single trees could be explained focusing on two of the problems derived from using individual

trees, stability, and accuracy. When minor modifications to the training set lead to important changes in a classifier, it is said to be unstable. According to [24] classification trees and neural networks are unstable methods. Methods such as decision trees have a high variance, but on average they are right that is, they are quite unbiased. Therefore, the correct class is usually the winner if the majority vote is applied for the aggregation of several of them.

Secondly, [25] proved that if the average error rate for one observation is less than fifty percent and the classifiers used in the ensemble are independent in producing their errors, the expected error of that observation can be reduced to zero when the number of combined classifiers increases. On the other hand, the ideal combination is to use very accurate classifiers, but they disagree as many times as possible since the combination of identical classifiers does not bring any benefit. In random forests, which try that the trees are not closely related to each other, randomness is introduced in the generation of these trees, so that each tree will be a function of the training set, but also of a random vector, which will influence the development of the forest.

To analyze the behavior of random forest in our problem, the results obtained will be compared with those achieved through artificial neural networks. Thus, Section 2 presents the random forest classification technique. Section 3 shows the simulation and analysis procedure in which a wide range of combinations of types of shift and correlation levels between variables is considered. The discussion of results for simulated data can be seen in Section 4. Finally, our concluding remarks and future lines of research are outlined in Section 5.

# 2 Random Forest

[23] defines a random forest as a classifier consisting of a collection of tree-structured classifiers $\{C(\boldsymbol{x}, \Theta_i), i=1, 2, \ldots\}$ where the $\{\Theta_i\}$ are independent and identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$.

Random forest using a random selection of features involves the joint use of two ensemble methods, bagging, and random input selection. The training sets are bootstrap samples of the same size as original drawn, with replacement, from the original data set. Then, a new tree is built for each one of the training data set using random input selection. That is to say, in each node. a small subset of features is randomly selected to split on. Then, the tree is grown to maximum size without being pruned. The number of variables, *F*, for the selected group must be set up previously.

As Breiman claimed, the error of the forest depends on the diversity and the accuracy of the individual trees. The optimal ensemble is made up of individual

classifiers as much accurate and diverse as possible, but these features move in opposite directions. The higher the *F* value, the higher the strength or accuracy, but the lower the diversity between the individual trees. On the other hand, the lower the *F* value, the lower the strength and the correlation among the individual trees. Therefore, this is the most important parameter to be tuned in a random forest. Breiman tried two values of *F*. The first value was 1, so only one variable was used. The second took the first integer less than $\log_2 p+1$, where *p* is the number of inputs. Later on, the same author advised setting the F value as the square root of *p*, although according to him, the results were not sensitive to the number of features selected to split each node. From his experiments over twenty data sets commonly used in automatic learning, Breiman found surprisingly that using a single random input variable the results were only slightly worse or even better than selecting a group.

A random selection of features makes the procedure faster since the number of input variables for which the gain of information has to be calculated is reduced. So, building a random forest in this way will be faster than other ensemble methods such as bagging or boosting, for instance.

The algorithm for building random forests can be summarized as follows:

1) Set the number of trees to grow.
2) For each tree:
    a) Draw a random subset ($T_k$) of the training set *T* (*N* observations with replacement) to train each tree. The elements in *T*, but not in $T_k$ are called out-of-bag (*oob*).
    b) Set *F* (number of variables to make a split) << *p* (number of input variables) and choose the best split among the *F* randomly selected variables for each node in each tree.
    c) Grow the tree to maximum size.
    d) Use *oob* training data to estimate error and variable importance.
3) Assign a class to new data as the majority vote among all the trees.
4) Use *oob* data to estimate the classification accuracy (or error) for the random forest and the importance measure for each input variable.

Although random forest is seen as a promising technique, it also has some drawbacks. Among them we can highlight two. First, as any ensemble method its interpretation is not as easy as that of a single tree. Second, random forests are biased to categorical variables with a high number of levels.

# 3 Method

## 3.1 Simulation Procedure

As mentioned before, the main goal of this paper is to check the better performance of random forests in comparison to artificial neural networks in the interpretation of out-of-control signals. We must remember that the classification techniques are used here as a complement to the $T^2$ control chart, i.e., control charts are used to detect the presence of out-of-control observations and, after that, the classification techniques are implemented to try to determine which variable or variables have caused this situation. In order to study the behavior of both classification techniques under different circumstances related to the magnitude of the shift and to the degree of correlation between variables, it is necessary to consider a wide range of situations, covering the most interesting cases.

The implementation of classification techniques requires both training and testing processes so once 1,000 out-of-control observations are generated for each different case, 900 observations will be used for the training process and 100 for testing the results. For the simulation process a bivariate normal distribution with the in-control parameters, $\mu$ and $\Sigma$, known is assumed for the quality inputs[1].

As it has been stated before, a wide range of cases is considered through the combination of different shifts in the mean, both in type and magnitude, with different correlation levels. Specifically, shifts of magnitude equal to 1, 2, and 3 standard deviations have been considered for each one of the input variables separately and in both variables at the same time. Moreover, two possible directions are considered for each shift, an increase (positive change) or a decrease (negative change) in the mean. With relation to the correlation level a range from -0.9 to 0.9 by 0.1 is considered and additionally the values 0.95 0.97 and 0.99, with both positive and negative signs, are included.

To sum up, there are eight possible changes depending on whether the change affects, increasing or decreasing, the mean of one or both variables. Additionally, twenty-five different values for the correlation coefficient and three levels of changes, 1, 2, or 3 standard deviations are considered. This wide range of situations has led us to have a total amount of 600 cases. Therefore, and taking into account that 1,000 observations have been simulated for each different case, the entire sample contains a total of 600,000 observations.

Some preliminary tests have shown us that cases in which both variables change in the same direction are equivalent regardless of whether these changes increase

[1] In this work, only two quality inputs are considered. The inclusion of more than two variables constitutes a future line of research.

or decrease the mean; similar results have been found when one variable increases and the other decreases regardless of which one is increasing and which one is decreasing; and finally, changes in only one variable regardless of which variable has changed and whether the change is an increase or a decrease, could also be considered equivalent. In this way, the possible scenarios are reduced to three cases without loss of generality. The first one assumes that only one variable changes; the second one considers the case where both variables change, but they do it in opposite direction; and finally, the third case with both variables changes in the same direction.

Then, using as inputs the $T^2$ statistic and the values of the two variables, and being the output the type of the change in the mean detected by the Hotelling's $T^2$ control chart *(Shift in one variable, Shift in same sense in $X_1$ and $X_2$ and Shift in different sense in $X_1$ and $X_2$)*, we have used random forest or neural network to out-of-control diagnosis.

## 3.2 Classification Procedure

The neural network (NN) model selected in this work is the well-known multilayer perceptron. The number of nodes in the input and the output layers has been set by the structure of our analysis, that is, the number of explanatory variables and the number of classes, respectively. On the other hand, several experiments were carried out to find the number of layers and hidden elements that gave the greatest accuracy in the prediction of the test data set. The resulting architecture is a feedforward network with a hidden layer that includes eight nodes. The training algorithm chosen is quasiNewton.

With regard to the random forest model (RF), the number of trees used in each iteration is 500 and the *F* parameter (number of variables randomly sampled as candidates at each split) is 2. The maximum number of terminal nodes in each tree of the forest is 5.

The data simulation and analysis processes have been developed using the R software [26]. Specifically, the packages to generate multivariate normal data and carry out the classification task are: mvtnorm [27], MASS and nnet [28] and randomForest [29]. All these packages are available on the R project website (http://rprojects.org) and the specific R code is available upon request to the authors of this article.

# 4 Results and Discussion

Tables 1 to 3 show the results obtained for changes of magnitude 1, 2, and 3 standard deviations, combined with the 25 correlation coefficient values.

Specifically, these tables show the classification error of RF and NN along with the difference in this error between both models. In addition, the cases where the application of RF provides an advantage over NN are highlighted. More specifically, RF has been considered better than NN when the difference in the classification error is greater than 1%.

In general, both RF and NN show good performance, in many cases without significant differences but we will try to draw some general comments taking into account the correlation structure and the type of changes. For example, it can be seen that the results are better for the largest change considered (3 standard deviations) and for correlation levels greater than 0.9. The first pattern of behavior is quite logical since the important changes are easier to detect and therefore, it is easier to determine the variable or variables that have motivated the change. However, with regard to the second statement, the results are not as obvious as in the previous case. Although higher correlation levels are not usual in statistical process control that is, values greater than 0.9, in these cases, the RF behavior is better regardless of the type and magnitude of the change.

To deepen the analysis of results, we begin with the most difficult case to be solved. These are the smallest changes, of magnitude equal to a standard deviation. The results, displayed in Table 1, show a good performance of both classification methods for high correlation levels. However, when the correlation level is medium or small, more feasible situations in statistical process control, the performance of both methods worsens. Specifically, when there is a positive correlation. and both variables change in the same direction, RF shows better performance than NN, although these differences are not too important. RF also works better than NN when the correlation is negative and the two variables change in opposite direction. In addition, when only one variable changes and the correlation level is small, RF is also better. To sum up, when the change in the variable is small, the most common but most challenging case in statistical process control, the use of RF is advantageous in terms of the classification error, which means a better diagnosis of out-of-control situations.

The results in Table 2 (changes of magnitude equal to two standard deviations) show similar behavior as in Table 1 but with less noticeable differences for positive correlation and both variable changing in the same direction. Finally, Table 3 (change of three standard deviations) shows that the cases where RF could be said that improves the behavior of NN is when only one variable shifts and there is a little correlation.

In summary, the results allow us to verify how, in the most common situations in statistical process control, the application of RF is advantageous compared to NN. Specifically, for small or moderate correlation levels and change in only one of the two variables, the behavior of RF is better. It is also better when the correlation is positive and the two variables change in the same direction or when the correlation is negative and the two variables change in the opposite sense,

these being the most feasible cases taking into account the correlation structure. This is pointed out in Figure 1.

Table 1
Error with shift of one standard deviation

| | ***Shift in one variable*** | | | ***Shift in different sense in $X_1$ and $X_2$*** | | | ***Shift in same sense in $X_1$ and $X_2$*** | | | **Total** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** |
| **-0.99** | 0.009 | 0.003 | 0.006 | 0.043 | 0.001 | 0.042 | 0.001 | 0.000 | 0.001 | 0.018 | 0.001 | 0.016 |
| **-0.97** | 0.054 | 0.037 | 0.017 | 0.058 | 0.033 | 0.025 | 0.042 | 0.037 | 0.005 | 0.051 | 0.036 | 0.016 |
| **-0.95** | 0.125 | 0.109 | 0.016 | 0.050 | 0.043 | 0.007 | 0.107 | 0.097 | 0.010 | 0.094 | 0.083 | 0.011 |
| **-0.9** | 0.230 | 0.262 | **-0.032** | 0.068 | 0.072 | -0.004 | 0.359 | 0.268 | 0.091 | 0.219 | 0.201 | 0.018 |
| **-0.8** | 0.529 | 0.451 | 0.078 | 0.080 | 0.089 | -0.009 | 0.197 | 0.210 | **-0.013** | 0.269 | 0.250 | 0.019 |
| **-0.7** | 0.592 | 0.435 | 0.157 | 0.076 | 0.101 | **-0.025** | 0.151 | 0.220 | **-0.069** | 0.273 | 0.252 | 0.021 |
| **-0.6** | 0.519 | 0.431 | 0.088 | 0.096 | 0.121 | **-0.025** | 0.183 | 0.201 | **-0.018** | 0.266 | 0.251 | 0.015 |
| **-0.5** | 0.509 | 0.473 | 0.036 | 0.109 | 0.122 | **-0.013** | 0.184 | 0.186 | -0.002 | 0.267 | 0.260 | 0.007 |
| **-0.4** | 0.458 | 0.480 | **-0.022** | 0.156 | 0.164 | **-0.008** | 0.251 | 0.201 | 0.050 | 0.288 | 0.282 | 0.007 |
| **-0.3** | 0.395 | 0.431 | **-0.036** | 0.121 | 0.139 | **-0.018** | 0.253 | 0.196 | 0.057 | 0.256 | 0.255 | 0.001 |
| **-0.2** | 0.487 | 0.507 | **-0.020** | 0.138 | 0.146 | **-0.008** | 0.240 | 0.202 | 0.038 | 0.288 | 0.285 | 0.003 |
| **-0.1** | 0.313 | 0.478 | **-0.165** | 0.314 | 0.176 | 0.138 | 0.287 | 0.189 | 0.098 | 0.305 | 0.281 | 0.024 |
| **0.0** | 0.384 | 0.541 | **-0.157** | 0.237 | 0.155 | 0.082 | 0.288 | 0.187 | 0.101 | 0.303 | 0.294 | 0.009 |
| **+0.1** | 0.432 | 0.437 | **-0.005** | 0.254 | 0.200 | 0.054 | 0.164 | 0.171 | -0.007 | 0.283 | 0.269 | 0.014 |
| **+0.2** | 0.495 | 0.531 | **-0.036** | 0.274 | 0.176 | 0.098 | 0.112 | 0.148 | **-0.036** | 0.294 | 0.285 | 0.009 |
| **+0.3** | 0.444 | 0.430 | 0.014 | 0.294 | 0.224 | 0.070 | 0.099 | 0.148 | **-0.049** | 0.279 | 0.267 | 0.012 |
| **+0.4** | 0.408 | 0.428 | **-0.020** | 0.246 | 0.211 | 0.035 | 0.114 | 0.114 | 0.000 | 0.256 | 0.251 | 0.005 |
| **+0.5** | 0.512 | 0.471 | 0.041 | 0.198 | 0.180 | 0.018 | 0.096 | 0.135 | **-0.039** | 0.269 | 0.262 | 0.007 |
| **+0.6** | 0.519 | 0.430 | 0.089 | 0.185 | 0.222 | **-0.037** | 0.113 | 0.146 | **-0.033** | 0.272 | 0.266 | 0.006 |
| **+0.7** | 0.544 | 0.473 | 0.071 | 0.149 | 0.186 | **-0.037** | 0.091 | 0.118 | **-0.027** | 0.261 | 0.259 | 0.002 |
| **+0.8** | 0.450 | 0.373 | 0.077 | 0.236 | 0.267 | **-0.031** | 0.088 | 0.090 | -0.002 | 0.258 | 0.243 | 0.015 |
| **+0.9** | 0.229 | 0.241 | -0.012 | 0.326 | 0.273 | 0.053 | 0.073 | 0.076 | -0.003 | 0.209 | 0.197 | 0.013 |
| **+0.95** | 0.115 | 0.104 | 0.011 | 0.117 | 0.091 | 0.026 | 0.075 | 0.054 | 0.021 | 0.102 | 0.083 | 0.019 |
| **+0.97** | 0.042 | 0.033 | 0.009 | 0.024 | 0.020 | 0.004 | 0.063 | 0.043 | 0.020 | 0.043 | 0.032 | 0.011 |
| **+0.99** | 0.013 | 0.002 | 0.011 | 0.000 | 0.000 | 0.000 | 0.061 | 0.001 | 0.060 | 0.025 | 0.001 | 0.024 |

Table 2
Error with shift of two standard deviations

| | **Shift in one variable** | | | **Shift in different sense in $X_1$ and $X_2$** | | | **Shift in same sense in $X_1$ and $X_2$** | | | **Total** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** |
| **-0.99** | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.005 | 0.001 | 0.000 | 0.001 | 0.002 | 0.000 | 0.002 |
| **-0.97** | 0.002 | 0.000 | 0.002 | 0.006 | 0.000 | 0.006 | 0.000 | 0.000 | 0.000 | 0.003 | 0.000 | 0.003 |
| **-0.95** | 0.008 | 0.003 | 0.005 | 0.010 | 0.002 | 0.008 | 0.000 | 0.002 | -0.002 | 0.006 | 0.002 | 0.004 |
| **-0.9** | 0.030 | 0.022 | 0.008 | 0.011 | 0.010 | 0.001 | 0.012 | 0.011 | 0.001 | 0.018 | 0.014 | 0.003 |
| **-0.8** | 0.074 | 0.090 | **-0.016** | 0.021 | 0.024 | **-0.003** | 0.129 | 0.080 | 0.049 | 0.075 | 0.065 | 0.010 |
| **-0.7** | 0.190 | 0.192 | **-0.002** | 0.028 | 0.044 | **-0.016** | 0.204 | 0.139 | 0.065 | 0.141 | 0.125 | 0.016 |
| **-0.6** | 0.300 | 0.215 | 0.085 | 0.048 | 0.054 | **-0.006** | 0.123 | 0.143 | **-0.020** | 0.157 | 0.137 | 0.020 |
| **-0.5** | 0.267 | 0.232 | 0.035 | 0.052 | 0.060 | **-0.008** | 0.155 | 0.149 | 0.006 | 0.158 | 0.147 | 0.011 |
| **-0.4** | 0.199 | 0.233 | **-0.034** | 0.077 | 0.065 | 0.012 | 0.182 | 0.122 | 0.060 | 0.153 | 0.140 | 0.013 |
| **-0.3** | 0.269 | 0.259 | 0.010 | 0.069 | 0.069 | 0.000 | 0.135 | 0.130 | 0.005 | 0.158 | 0.153 | 0.005 |
| **-0.2** | 0.244 | 0.270 | **-0.026** | 0.065 | 0.074 | **-0.009** | 0.190 | 0.131 | 0.059 | 0.166 | 0.158 | 0.008 |
| **-0.1** | 0.211 | 0.247 | **-0.036** | 0.084 | 0.082 | 0.002 | 0.225 | 0.112 | 0.113 | 0.173 | 0.147 | 0.026 |
| **0.0** | 0.215 | 0.249 | **-0.034** | 0.172 | 0.106 | 0.066 | 0.099 | 0.100 | -0.001 | 0.162 | 0.152 | 0.010 |
| **+0.1** | 0.242 | 0.237 | 0.005 | 0.159 | 0.130 | 0.029 | 0.071 | 0.069 | 0.002 | 0.157 | 0.145 | 0.012 |
| **+0.2** | 0.226 | 0.249 | **-0.023** | 0.174 | 0.113 | 0.061 | 0.077 | 0.088 | **-0.011** | 0.159 | 0.150 | 0.009 |
| **+0.3** | 0.232 | 0.240 | **-0.008** | 0.149 | 0.129 | 0.020 | 0.086 | 0.074 | 0.012 | 0.156 | 0.148 | 0.008 |
| **+0.4** | 0.207 | 0.234 | **-0.027** | 0.128 | 0.109 | 0.019 | 0.099 | 0.079 | 0.020 | 0.145 | 0.141 | 0.004 |
| **+0.5** | 0.268 | 0.231 | 0.037 | 0.144 | 0.157 | **-0.013** | 0.033 | 0.045 | **-0.012** | 0.148 | 0.144 | 0.004 |
| **+0.6** | 0.284 | 0.221 | 0.063 | 0.106 | 0.147 | **-0.041** | 0.040 | 0.043 | -0.003 | 0.143 | 0.137 | 0.006 |
| **+0.7** | 0.217 | 0.159 | 0.058 | 0.168 | 0.146 | 0.022 | 0.031 | 0.042 | **-0.011** | 0.139 | 0.116 | 0.023 |
| **+0.8** | 0.068 | 0.089 | **-0.021** | 0.121 | 0.084 | 0.037 | 0.030 | 0.028 | 0.002 | 0.073 | 0.067 | 0.006 |
| **+0.9** | 0.038 | 0.024 | 0.014 | 0.021 | 0.013 | 0.008 | 0.016 | 0.008 | 0.008 | 0.025 | 0.015 | 0.010 |
| **+0.95** | 0.010 | 0.001 | 0.009 | 0.001 | 0.001 | 0.000 | 0.012 | 0.001 | 0.011 | 0.008 | 0.001 | 0.007 |
| **+0.97** | 0.006 | 0.001 | 0.005 | 0.000 | 0.000 | 0.000 | 0.014 | 0.001 | 0.013 | 0.007 | 0.001 | 0.006 |
| **+0.99** | 0.002 | 0.000 | 0.002 | 0.044 | 0.000 | 0.044 | 0.011 | 0.000 | 0.011 | 0.019 | 0.000 | 0.019 |

Table 3
Error with shift of three standard deviations

| | **Shift in one variable** | | | **Shift in different sense in $X_1$ and $X_2$** | | | **Shift in same sense in $X_1$ and $X_2$** | | | **Total** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** | **RF** | **NN** | **RF-NN** |
| **-0.99** | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **-0.97** | 0.001 | 0.000 | 0.001 | 0.002 | 0.000 | 0.002 | 0.011 | 0.000 | 0.011 | 0.005 | 0.000 | 0.005 |
| **-0.95** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **-0.9** | 0.004 | 0.002 | 0.002 | 0.004 | 0.000 | 0.004 | 0.000 | 0.001 | -0.001 | 0.003 | 0.001 | 0.002 |
| **-0.8** | 0.017 | 0.014 | 0.003 | 0.007 | 0.003 | 0.004 | 0.012 | 0.010 | 0.002 | 0.012 | 0.009 | 0.003 |
| **-0.7** | 0.040 | 0.036 | 0.004 | 0.010 | 0.008 | 0.002 | 0.038 | 0.025 | 0.013 | 0.029 | 0.023 | 0.006 |
| **-0.6** | 0.063 | 0.056 | 0.007 | 0.012 | 0.016 | -0.004 | 0.074 | 0.049 | 0.025 | 0.050 | 0.040 | 0.009 |
| **-0.5** | 0.104 | 0.085 | 0.019 | 0.019 | 0.020 | -0.001 | 0.079 | 0.059 | 0.020 | 0.067 | 0.055 | 0.013 |
| **-0.4** | 0.095 | 0.118 | **-0.023** | 0.033 | 0.032 | 0.001 | 0.105 | 0.047 | 0.058 | 0.078 | 0.066 | 0.012 |
| **-0.3** | 0.089 | 0.095 | **-0.006** | 0.044 | 0.036 | 0.008 | 0.084 | 0.061 | 0.023 | 0.072 | 0.064 | 0.008 |
| **-0.2** | 0.095 | 0.103 | **-0.008** | 0.049 | 0.036 | 0.013 | 0.095 | 0.067 | 0.028 | 0.080 | 0.069 | 0.011 |
| **-0.1** | 0.096 | 0.105 | **-0.009** | 0.051 | 0.042 | 0.009 | 0.061 | 0.044 | 0.017 | 0.069 | 0.064 | 0.006 |
| **0.0** | 0.082 | 0.113 | **-0.031** | 0.088 | 0.053 | 0.035 | 0.074 | 0.052 | 0.022 | 0.081 | 0.073 | 0.009 |
| **+0.1** | 0.088 | 0.106 | **-0.018** | 0.071 | 0.056 | 0.015 | 0.053 | 0.038 | 0.015 | 0.071 | 0.067 | 0.004 |
| **+0.2** | 0.111 | 0.113 | **-0.002** | 0.063 | 0.049 | 0.014 | 0.045 | 0.038 | 0.007 | 0.073 | 0.067 | 0.006 |
| **+0.3** | 0.109 | 0.114 | **-0.005** | 0.081 | 0.058 | 0.023 | 0.033 | 0.031 | 0.002 | 0.074 | 0.068 | 0.007 |
| **+0.4** | 0.110 | 0.093 | 0.017 | 0.080 | 0.062 | 0.018 | 0.034 | 0.020 | 0.014 | 0.075 | 0.058 | 0.016 |
| **+0.5** | 0.116 | 0.081 | 0.035 | 0.065 | 0.053 | 0.012 | 0.021 | 0.020 | 0.001 | 0.067 | 0.051 | 0.016 |
| **+0.6** | 0.050 | 0.049 | 0.001 | 0.062 | 0.044 | 0.018 | 0.008 | 0.009 | -0.001 | 0.040 | 0.034 | 0.006 |
| **+0.7** | 0.033 | 0.030 | 0.003 | 0.046 | 0.029 | 0.017 | 0.013 | 0.009 | 0.004 | 0.031 | 0.023 | 0.008 |
| **+0.8** | 0.016 | 0.007 | 0.009 | 0.010 | 0.007 | 0.003 | 0.006 | 0.004 | 0.002 | 0.011 | 0.006 | 0.005 |
| **+0.9** | 0.007 | 0.001 | 0.006 | 0.000 | 0.000 | 0.000 | 0.003 | 0.001 | 0.002 | 0.003 | 0.001 | 0.003 |
| **+0.95** | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 | 0.002 | 0.001 | 0.000 | 0.001 |
| **+0.97** | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.003 | 0.000 | 0.003 | 0.001 | 0.000 | 0.001 |
| **+0.99** | 0.004 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 | 0.003 | 0.000 | 0.003 | 0.002 | 0.000 | 0.002 |

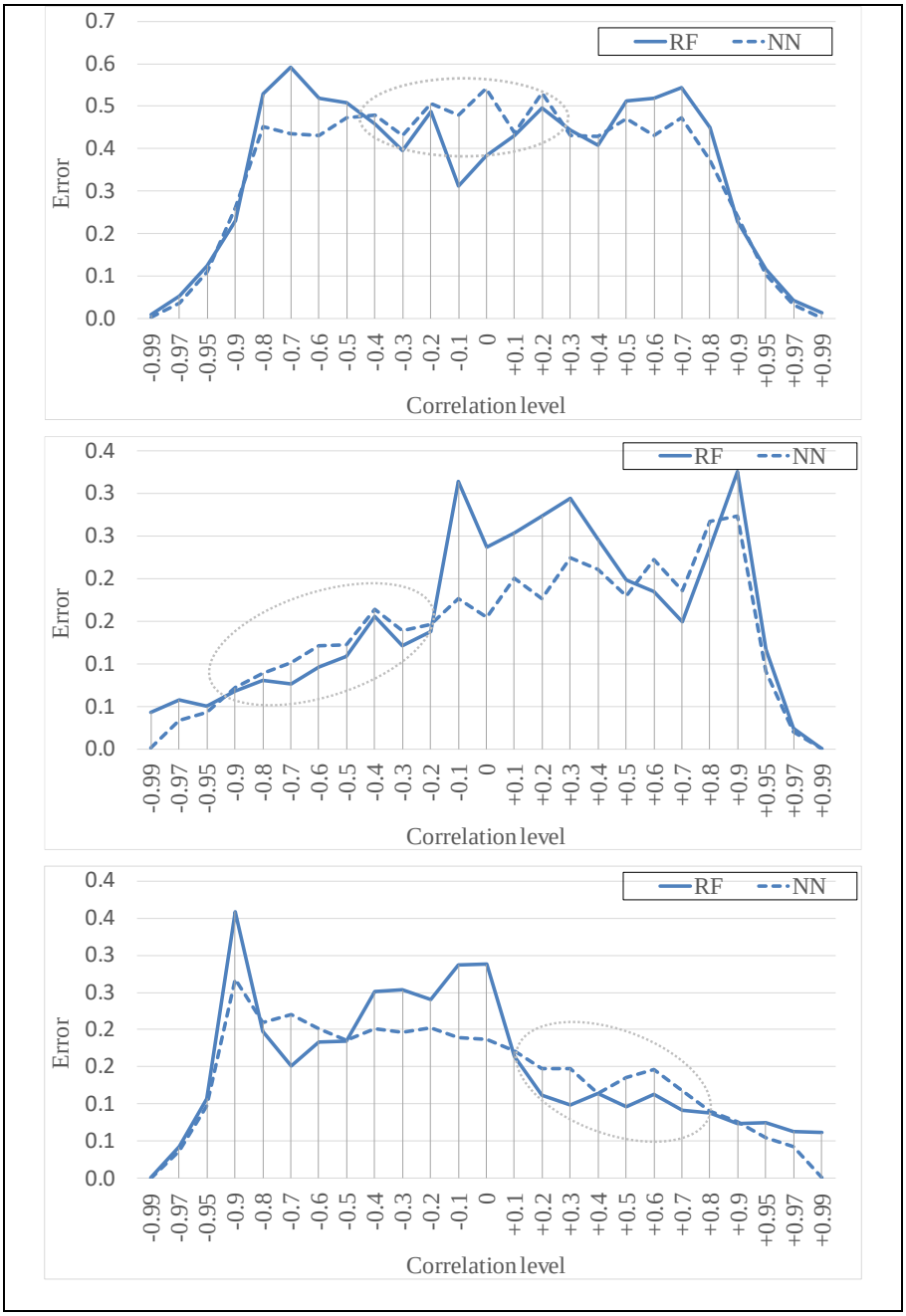


Figure 1

Error by correlation level. Shift of 1 standard deviation and, from up to bottom, shift in only one variable, shift in two variables in different senses and shift in two variables in the same sense

## Conclusions

In recent years, classification methods are useful as a complement to T2 graphs for the diagnosis of out-of-control signals, neural networks being the most used method. The proposal developed in this paper would facilitate the application of multivariate quality control in production processes where the quality of manufactured goods depends on several correlated characteristics and small changes can have big consequences, such as the medical industry. Concretely, the use of random forest as an alternative classification method improves results in certain situations.

In this sense, the performance of both NN and RF methods improves with the change magnitude and with the absolute value of the correlation level. The comparison between these methods depends on the correlation structure combined with the type of change. The results allow us to verify how in the most common situations in statistical process control, the application of RF supposes an advantage in comparison with NN. Specifically, for small or moderate correlation levels and change in only one of the two variables, RF provides better results than NN. RF performance is also better when the correlation is positive and the two variables change in the same direction or when the correlation is negative and the two variables change in different sense (most feasible cases taking into account the correlation structure). These results allow us to verify that there is not a technique with a predominant behavior over the other although, depending on the case to be treated, using one technique or another allows obtaining better results.

This work opens a new research line, currently under development, which would allow validating these methods for a higher number of variables, providing an alternative procedure to the current use of dimensionality reduction techniques.

## Acknowledgement

The research group in Applied Statistical and Classification Techniques (GITECA), to which three of the authors of this work belong, would like to thank the University of Castilla-La Mancha for funding this research through its Research Groups Support Program.

## References

[1] S. Vidal-Puig and A. Ferrer A Comparative Study of Different Methodologies for Fault Diagnosis in Multivariate Quality Control, *Communications in Statistics - Simulation and Computation*, 43:5, 2014, 986-1005, DOI: 10.1080/03610918.2012.720745

[2] J. Yu, X. Zheng and S. Wang. Stacked denoising autoencoder-based feature learning for out-of-control source recognition in multivariate manufacturing process. *Quality and Reliability Engineering International*, *35*(1), 2019, 204-223

[3] C. Fuchs and R. Kenett *Multivariate Quality Control: Theory and Applications*. Marcel Dekker: New York, 1998

[4] RL. Mason, ND. Tracy and JC. Young. Decomposition of T2 for multivariate control chart interpretation. *Journal of Quality Technology* 1995; 27:109-119

[5] RL. Mason, ND. Tracy and JC. Young. A practical approach for interpreting multivariate T2 control chart signals. *Journal of Quality Technology* 1997; 29:396-406

[6] BJ. Murphy Selecting out of control variables with the T2 multivariate quality control procedure. *Journal of the Royal Statistical Society: Series D (The Statistician)* 1987; 36:571-583, https://doi.org/10.2307/2348668

[7] CS. Cheng A multi-layer neural network model for detecting changes in the process mean. *Computers and Industrial Engineering* 1995; 28(1): 51-61, https://doi.org/10.1016/0360-8352(94)00024-H

[8] STA. Niaki and B. Abassi Fault diagnosis in multivariate control charts using artificial neural networks. *Quality and Reliability Engineering International* 2005; 21: 825-840, https://doi.org/10.1002/qre.689

[9] SI. Chang and C. A. Aw A neural fuzzy control chart for detecting and classifying process mean shifts. *International Journal of Production Research* 1996; 34(8): 2265-2278, https://doi.org/10.1080/00207549608905024

[10] R-S. Guh. On-line identification and quantification of mean shifts in bivariate processes using a neural network-based approach. *Quality and Reliability Engineering International* 2007; 23: 367-385, DOI: https://doi.org/10.1002/qre.796

[11] R-S. Guh and YC. Hsieh. A neural network based model for abnormal pattern recognition of control charts. *Computers and Industrial Engineering* 1999; 36: 97-108. https://doi.org/10.1016/S0360-8352(99)00004-2

[12] R-S. Guh and JDT. Tannock. A neural network approach to characterize pattern parameters in process control charts. *Journal of Intelligent Manufacturing* 1999; 10(5): 449-462, https://doi.org/10.1023/A:1008975131304

[13] R-S. Guh and JDT. Tannock. Recognition of control chart concurrent patterns using a neural network approach. *International Journal of Production Research* 1999; 37(8), 1743-1765, https://doi.org/10.1080/002075499190987

[14] Z. Miao and M. Yang. Control chart pattern recognition based on convolution neural network. In *Smart Innovations in Communication and Computational Sciences* (pp. 97-104) Springer, Singapore, 2019

[15] F. Zorriassatine and JDT. Tannock. A review of neural networks for statistical process control. *Journal of Intelligent Manufacturing* 1998; 9(3): 209-224, DOI https://doi.org/10.1023/A:1008818817588

[16] JB. Yu and LF. Xi. A neural network ensemble-based model for on-line monitoring and diagnosis of out-of-control signals in multivariate manufacturing processes. *Expert Systems with Applications* 2009; 36(1): 909-921, https://doi.org/10.1016/j.eswa.2007.10.003

[17] S. G. He, Z. He and G. A. Wang Online monitoring and fault identification of mean shifts in bivariate processes using decision tree learning techniques. *Journal of Intelligent Manufacturing* 2013; 24(1): 25-34, https://doi.org/10.1007/s10845-011-0533-5

[18] S. He, GA. Wang, M. Zhang and DF. Cook. Multivariate process monitoring and fault identification using multiple decision tree classifiers. *International Journal of Production Research* 2013; 51(11): 3355-3371, https://doi.org/10.1080/00207543.2013.774474

[19] CS. Cheng and HT, Lee. Identifying the Source of Variance Shifts in Multivariate Statistical Process Control Using Ensemble Classifiers. In: Tan C., Goh T. (eds) *Theory and Practice of Quality and Reliability Engineering in Asia Industry*. Springer: Singapore, 2017, https://doi.org/10.1007/978-981-10-3290-5_3

[20] J. Jiang and H-M. Song, Diagnosis of Out-of-control Signals in Multivariate Statistical Process Control Based on Bagging and Decision Tree. *Asian Business Research* 2017; 2(2):1-6, DOI: https://doi.org/ 10.20849/abr.v2i2.147

[21] E. Alfaro, JL. Alfaro, M. Gámez and N. García. A boosting approach for understanding out-of-control signals in multivariate control charts. *International Journal of Production Research* 2009; 47(24): 6821-6834, https://doi.org/10.1080/00207540802474003

[22] E. Alfaro, JL. Alfaro, M. Gámez and N. García. A Comparison of Different Classification Techniques to Determine the Change Causes in Hotelling's T2 Control Chart. *Quality and Reliability Engineering International* 2015; 31: 1255-1263 doi: 10.1002/qre.1901

[23] L. Breiman. Random Forest. *Machine Learning* 2001; 45(1): 5-32

[24] L. Breiman. Bagging predictors. *Machine Learning* 1996, 24(2):123-140

[25] LK. Hansen and P. Salamon, P. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence* 1990, 12(10): 993-1001

[26] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/. 2016

[27] A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl and T. Hothorn. mvtnorm: Multivariate normal and t distributions. *R package*, version 1.0-0, 2014

[28] WN. Venables and BD. Ripley. *Modern applied statistics with S*, 4th edition, Springer, New York, 2002

[29] L. Liaw and M. Wiener. Classification and Regression by randomForest. *R News* 2002; 2(3): 18-22