Computers & Industrial Engineering 101 (2016) 70–80

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Risk based uncertainty quantification to improve robustness of manufacturing operations

Cinzia Giannetti [a,*], Rajesh S. Ransing [b]

[a] College of Science, Swansea University, Singleton Campus, Swansea SA2 8PP, UK
[b] College of Engineering, Swansea University, Bay Campus, Crymlyn Burrows, Swansea SA1 8EN, UK

ARTICLE INFO

Article history:
Received 2 April 2016
Received in revised form 15 July 2016
Accepted 8 August 2016
Available online 9 August 2016

Keywords:
Tolerance synthesis
ISO9001:2015
7Epsilon
Six Sigma
Industry 4.0
Digital manufacturing

ABSTRACT

The cyber-physical systems of Industry 4.0 are expected to generate vast amount of in-process data and revolutionise the way data, knowledge and wisdom is captured and reused in manufacturing industries. The goal is to increase profits by dramatically reducing the occurrence of unexpected process results and waste. ISO9001:2015 defines risk as effect of uncertainty. In the 7Epsilon context, the risk is defined as effect of uncertainty on expected results. The paper proposes a novel algorithm to embed risk based thinking in quantifying uncertainty in manufacturing operations during the tolerance synthesis process. This method uses penalty functions to mathematically represent deviation from expected results and solves the tolerance synthesis problem by proposing a quantile regression tree approach. The latter involves non parametric estimation of conditional quantiles of a response variable from in-process data and allows process engineers to discover and visualise optimal ranges that are associated with quality improvements. In order to quantify uncertainty and predict process robustness, a probabilistic approach, based on the likelihood ratio test with bootstrapping, is proposed which uses smoothed probability estimation of conditional probabilities. The mathematical formulation presented in this paper will allow organisations to extend Six Sigma process improvement principles in the Industry 4.0 context and implement the 7 steps of 7Epsilon in order to satisfy the requirements of clauses 6.1 and 7.1.6 of the ISO9001:2015 and the aerospace AS9100:2016 quality standard.

1. Introduction

Industry 4.0, also called the fourth industrial revolution, has already started to take place and it will involve a complete digital transformation of many manufacturing activities. This revolution will break the existing boundaries of manufacturing operations to deliver a new generation of intelligent, co-operating and interconnected manufacturing systems capable of monitoring system performance real time to control costs, reduce downtime and prevent faults (Foresight, 2013). The new manufacturing systems will be characterised by cyber-physical systems able to interoperate via networked connections and interact with humans in complex smart factory environments. These systems will make extensive use of data and predictive analytics to manage manufacturing processes more efficiently and allow production of customised products with increased profitability and energy efficiency (Deloitte, 2015; Germany Trade & Invest, 2015; Manyika, 2012; Rockwell Automation, 2014). As new technologies are starting to be deployed as part of the fourth industrial revolution, one of the biggest challenges manufacturing companies are facing is to develop capabilities to timely access and reuse the sheer volume of data and information scattered across diverse business functions to gain new insights and to create knowledge and value for the enterprise (Foresight, 2013). As part of this digital transformation new predictive analytics tools will need to be developed to access, integrate and use the vast, multi-faceted and heterogeneous data sets that will become available, including machine and human generated data collected through sensors and other interconnected IT systems.

In the context of continual improvement, undoubtedly the new generation of manufacturing systems represent an important opportunity for leveraging existing continual improvement capabilities by exploiting the potential to create new knowledge from in-process data and enabling real-time decision making capabilities. Continual improvement is defined by the ISO9001:2015 standard as a "recurring activity to enhance performance", and the one that generally leads to a corrective or preventive action (International Standard Organisation ISO, 2014, p. 16). This

* Corresponding author.
E-mail addresses: c.giannetti@swansea.ac.uk (C. Giannetti), r.s.ransing@swansea.ac.uk (R.S. Ransing).

http://dx.doi.org/10.1016/j.cie.2016.08.002
0360-8352/

typically involves reducing variation in production processes to satisfy customer requirements. According to Stricker and Lanza (2014), the robustness of a production system should aim for both a target value of the process outcome and a stable or consistent performance with minimum deviation or variation. In multiprocess manufacturing achieving process robustness is a challenging activity because the quality of the final product is often influenced by hundreds of factors as well as part specific quality constraints (Giannetti et al., 2014, 2015; Ransing & Ransing, 2014; Roshan, Giannetti, Ransing, & Ransing, 2014). Production processes in foundries are a typical example of multiprocess manufacturing as they consists of many sub-processes (i.e. patternmaking, molding, coremaking, melting and pouring, heat treatment, welding and finishing), with their quality determined by the effect and interactions of many process inputs. For these processes, quality of the final product cannot be simply achieved by limiting process variability to predefined thresholds determined according to the customer requirements. In fact, despite working within specifications, a process may still exhibit a large amount of variance in its output target value. Process knowledge is often necessary to implement changes which will lead to enhanced performance and achieve process robustness. Recently a novel methodology, called 7Epsilon (2015), has been developed which promotes the use of risk based analysis of in-process data to create new product specific process knowledge and evaluate opportunities that will lead to improvement of manufacturing processes, as required by the ISO9001:2015 standard. (Giannetti et al., 2015; Ransing, Batbooti, Giannetti, & Ransing, 2016; Roshan et al., 2014). Ransing et al. (2016) have shown that new product specific process knowledge can be created from in-process data by means of tolerance synthesis. In the literature process tolerance synthesis is defined as the problem of allocating tolerances of process variables to achieve a specified quality at a minimum costs (Ding, Jin, Ceglarek, & Shi, 2000). Extending this definition to the context of multiprocess manufacturing, tolerance synthesis is the study of variability in all process inputs (including interactions among process inputs) in order to discover optimal regions that correlate with the occurrences of expected process outputs (results) (Ransing et al., 2016). Owing to its definition, tolerance synthesis involves developing a sound understanding of how variability of process factors (i.e. process input settings) affects the expected target value and the variability of responses (i.e. process outputs). Process robustness is then achieved by selecting optimal tolerance limits of process variables that will reduce variation of responses (Ransing et al., 2016). One approach to solve the tolerance synthesis problem and predict process robustness is to attempt to model the relationships between process factors and responses from in-process data. In the literature data driven predictive methods have been used and applied to several industrial sectors, including manufacture of fabricated metal products, computers and electronic goods (Köksal, Batmaz, & Testik, 2011). The influence of design and process parameters has also been studied via numerical simulation methods (Lewis, Manzari, Ransing, & Gethin, 2000; Lewis & Ransing, 2000; Pao, Ransing, Lewis, & Lin, 2004; Postek, Lewis, Gethin, & Ransing, 2005), decision trees (Bakır et al., 2006) and Bayesian networks (Lewis & Ransing, 1997). Typically these methods attempt to model the complex relationships between process inputs and outputs to characterise or, sometimes, predict process behaviour and find improvement opportunities. However, for complex manufacturing processes, these relationships are not easily captured due to several reasons. First of all, in multiprocess manufacturing operations, the quality of the final product is often influenced by a combination of large number of product and process variables, including both categorical and continuous variables. Secondly, relationships between inputs and quality characteristics are related not only to some physical phenomena but also to interactions of different process settings. Trying to model these relationships can become very cumbersome with the risk of including variables with little effect on the final quality output (Giannetti et al., 2014). Traditional data driven approaches, such as regression analysis, tend to fail due the inability to model complex interactions and overfitting problems due to the presence of noise. Unless some prior knowledge about the underlying model is available, fitting the data with simple models, such as a linear model, would fail to capture the complex interactions (Bakır et al., 2006). On the other hand, using more complex models (e.g. polynomials) would lead to overfitting because of the presence of noise and small amount of observations. Overfitting will then produce a model that performs very well on the available data but has very poor predictive performance. In order for process knowledge to be learnt robustly, there is the need to analyse weak patterns in noisy and heterogeneous datasets. Furthermore, because of the presence of noisy data, uncertainty of the model results need to be quantified to overcome the lack of process knowledge.

In this paper a novel algorithm is proposed to predict the robustness of a process by quantifying uncertainty in manufacturing operations. The main motivation of this work is to develop a robust and general purpose method for tolerance synthesis to quantify the combined effects of process variables on the quality output without making distributional assumptions and overcome the linearity assumption of previous algorithms for risk based tolerance synthesis (Giannetti et al., 2014; Ransing et al., 2016). This is achieved by introducing a novel mathematical formulation of the tolerance synthesis problem in terms of conditional quantiles of response variables and a robust algorithm based on quantile regression to find optimal tolerance limits. The method improves the previous quality correlation algorithm for tolerance synthesis (Ransing et al., 2016) by using the concept of likelihood ratio for probabilistic estimation of the effects of the new tolerance limits on the quality output. Uncertainty quantification of the newly developed hypotheses is performed using the bootstrap method to predict process robustness and aid development of new product specific process knowledge.

The paper is organised as follows. Section 2 reviews regression trees methods and their industrial applications, including traditional least square and quantile regression approaches. Section 3 introduces the tolerance synthesis problem, its mathematical formulation and the proposed algorithm. The latter includes a probabilistic approach for hypotheses validation based on calculation of likelihood ratio with bootstrap method. The method is illustrated using test data from the UCI machine learning repository. In Section 4 the proposed algorithm is applied to an industrial case study to show its application for uncertainty quantification in multiprocess manufacturing systems. The paper is concluded in Section 5.

2. Related methods: regression trees and quantile regression

Decision tree learning is a common method used for classification and regression problems, owing its popularity to easiness of interpretation and the ability to visually and explicitly represent decision making rules (Bakır et al., 2006). The general method builds a tree shaped structure to predict or classify a dependent variable (often called response variable) by recursive partitioning the data set into groups of observations with similar values of the dependent variable (Breiman, Friedman, Stone, & Olshen, 1984). One main advantage of decision tree learning is that it can deal simultaneously with continuous and categorical predictor variables, without the need of further transformations and making distributional assumptions (Francke, López-Tarazón, & Schroder, 2008). Regression trees are particular types of decision tree designed to work with continuous response variables, while

classification trees deal with categorical response variables. Some of these are described in Loh (2011). In broad terms a decision tree is built according to some rules which determine: (a) a way to select a split (splitting criterion); (b) a rule for determining when a node is terminal and (c) a rule for assigning a value (or a class) of the predicted variable at each terminal node (Breiman et al., 1984). Several algorithms have been developed for classification and regression trees. Among these, a well known algorithm is CART (Breiman et al., 1984). Other methods are CRUISE (Kim & Loh, 2001, 2003), QUEST (Loh & Shih, 1997) and GUIDE (Loh, 2009). Despite their popularity, regression tree methods with single base classifiers tend to have low accuracy and instability due to high sensitivity to small changes in the datasets (Breiman et al., 1984). This is usually overcome by using ensemble methods (Breiman, 1996, 2001; Kotsiantis, 2011) or boosting (Freund & Schapire, 1997). Random Forest is an example of ensemble method where different weak learners are trained on bootstrap samples and on a random subset of the variables. It is argued that this further randomisation procedure both increases prediction accuracy and reduces bias (Breiman, 2001). Quantile Regression Forest is an extension of Regression Forest to infer conditional quantiles of the response variables, in addition to the mean (Meinshausen, 2006).

In the literature, regression trees are widely used to study the behaviour of real world systems and processes in many different areas of application. Bakır et al. (2006) state that regression tree learning outperform traditional linear regression for defect cause modelling in casting processes, due to the ability to capture complex interactions among process variables. Random Forest and Quantile Regression Forests are also used as predictive models to estimate suspended sediment concentration and yield in streams (Francke et al., 2008). An advantage of Quantile Regression Forest is also that it allows to quantify uncertainty of the model by providing estimates of the accuracy of prediction (Francke et al., 2008). In the aerospace sector, Random Forest is applied to aircraft engine fault diagnosis (Yan, 2006). This is a challenging classification problem due to inherent characteristics associated with aircraft engines (Yan, 2006). In chemical processing, a method that combines decision trees and support vector machine is used to improve process operations of a manganese extraction plant (Jemwa & Aldrich, 2005). The authors argue that this method is useful to find process improvement opportunities, despite sparse and unreliable sparse data (Jemwa & Aldrich, 2005), with the advantage of easiness of interpretation and the ability to translate results in simple rules. Decision trees are applied to experimental data to extract useful knowledge and support the development of quality products (Shao, Rowe, & York, 2007), while random forest is also used to real time monitoring and control of faults (Auret & Aldrich, 2010). A comparative study, with benchmark data from the Tennessee Eastman Process, shows that the method based on Random Forest outperform traditional fault detection methods based on linear and non-linear PCA, while it is only slightly worst than other statistical process control approaches based on Kernel PCA (Auret & Aldrich, 2010). This is due to the fact that both Random Forest and Kernel PCA manage to account for non-linear patterns.

Quantile regression is a type of regression analysis used to predict the conditional quantiles of a dependent variable Y, given values of predictor variables $X_1, X_2, \ldots X_n$ (Koenker & Bassett, 1978). While the traditional least squares regression method estimates the conditional mean of the predicted variable, quantile regression studies the effects of covariates on the full conditional distribution (Chaudhuri & Loh, 2002). Quantile regression methods are appropriate to study how covariates affect, not only the centre of the distribution, but also lower and upper tail, becoming very useful when covariates have different effects on different parts of the conditional distribution of the predicted response (Chaudhuri & Loh, 2002). In particular, quantile regression is a robust alternative to least squares estimator in the presence of outliers or non-Gaussian distributions (Koenker & Bassett, 1978). In the literature, quantile regression has been applied to different fields, including several applications in econometrics (Bassett & Chen, 2002; Coad & Rao, 2008; Dimelis & Louri, 2002; Koenker, 2005), ecology (Cade & Noon, 2003; Francke et al., 2008; Scharf, Juanes, & Sutherland, 1998) and medicine (Geraci & Bottai, 2007; Wei, Pere, Koenker, & He, 2006). However quantile regression has not been applied yet to continual process improvement and more in general to modelling relationships between process variables to predict robustness in manufacturing processes.

Let Y be a continuous response variable, $F(y) = P(Y \leqslant y)$ its cumulative distribution function. For $0 < \alpha < 1$, the α-th quantile of Y is defined as:

$$Q_\alpha = F^{-1}(\alpha) = \inf\{y : F(y) \geqslant \alpha\}. \tag{1}$$

Given a covariate vector X, the conditional α-th quantile is defined as:

$$Q_\alpha(x) = F^{-1}(\alpha|X = x) = \inf\{y : F(y|X = x) \geqslant \alpha\}, \tag{2}$$

where $F(y|X = x) = P(Y \leqslant y|X = x)$ is the conditional distribution function. The purpose of quantile regression is to estimate the conditional quantile of Y from available data. Similarly to traditional regression, quantile regression can be set up as an optimisation problem involving minimisation of a loss function (Koenker & Bassett, 1978; Meinshausen, 2006). Traditional regression estimates the conditional mean from available data by minimising the expected error squared loss. As discussed in Meinshausen (2006), estimation of quantiles can be achieved by minimising the expected error of a tiled loss function which measures weighted absolute deviations between observations and quantiles:

$$Q_\alpha(x) = \arg\min_q E\{L(Y, q)|X = x\}, \tag{3}$$

where $L(y, q)$ is defined as:

$$L(y, q) = \begin{cases} \alpha|y - q| & if \quad y > q \\ (1 - \alpha)|y - q| & if \quad y \leqslant q \end{cases} \tag{4}$$

In the particular case when $\alpha = 0.5$, the expected value of the error function is the Least Absolute Deviation (LAD), leading to estimation of the median. Given available data, quantile regression can be set up as an optimisation problem and be solved with linear programming (Koenker & Bassett, 1978; Meinshausen, 2006). Non parametric estimation of quantiles can also be performed using regression trees (Chaudhuri & Loh, 2002). The algorithm proposed in Chaudhuri and Loh (2002) extends the GUIDE algorithm and provides piecewise constant or linear estimation of quantiles. The main advantage of this method is that it provides a simple summary of the covariates interactions and their effects on the distribution of the response variable (Chaudhuri & Loh, 2002). Another approach to non parametric estimation of quantiles is using Regression Forest to estimate the full distribution of the response variable (Meinshausen, 2006). In this case, the estimated quantiles are calculated from ensembles generated by means of the Regression Forest algorithm and does not involve minimisation of the loss function defined in Eq. (4).

3. Risk based tolerance synthesis and uncertainty quantification

By definition, tolerance synthesis seeks to adjust ranges of process parameters in order to reduce the variability of a given response, so that process variation is kept "close" enough to the optimal values. One main requirement to achieve tolerance synthesis is to develop and test hypotheses regarding how variability

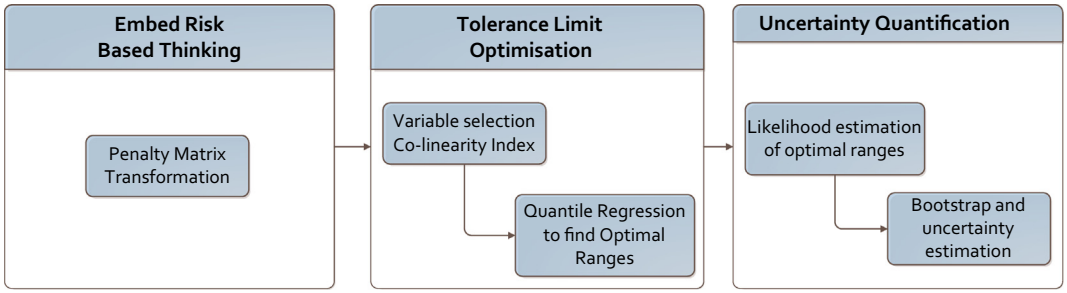


Fig. 1. The proposed algorithm consists of a three steps: (a) embed risk based thinking; (b) tolerance synthesis and (c) uncertainty quantification.

of factors affects process responses. In other words, predict process robustness. Given the complexity of multiprocess manufacturing, attempting to discover the relationship between process factor and responses using traditional methods like least squares regression is very challenging and would lead to models that are sub-optimal. In this section a novel mathematical formulation of the tolerance synthesis problem is described and it is shown how the tolerance synthesis problem can be solved with quantile regression. The main steps of the proposed algorithm for tolerance synthesis and uncertainty quantification are summarised in Fig. 1.

3.1. Mathematical formulation of the tolerance synthesis problem

The risk based approach introduced by the latest standard of ISO9001:2015 requires organizations to categorise process outputs as either acceptable or unacceptable outputs and take specific actions to determine and address risks and opportunities in order to minimise undesired effects and achieve process improvement (Ransing et al., 2016). Mathematically, a quality objective can be represented by a bound response variable (i.e. a response variable being within a specific range). In a manufacturing environment a quality objective is, for instance, keeping the percentage of rejected parts below a nominal threshold. Ransing, Giannetti, Ransing, and James (2013) have introduced the concept of penalty functions which represent deviations from desired process responses. If lower values of response correspond to a desirable outcome, a penalty value of 1 is given to response values above a certain threshold T_{max} and penalty value 0 to response values below a certain threshold T_{min}. Vice versa applies if higher values correspond to desirable outcomes. Heuristic rules for the choice of the thresholds are discussed in Giannetti et al. (2014). After this data transformation, a quality output can be quantified using a penalty value Y, with $0 \leqslant Y \leqslant 1$, which indicates deviation from desired response (0 being desirable response and 1 undesirable response). Optimal and avoid outcomes of the response variables are then defined as being $Y = 0$ and $Y = 1$.

Let's assume that there are n covariates (i.e. process factors) $X_1, X_2, \ldots, X_n$, the tolerance synthesis problem can be reduced to discover optimal ranges (or regions) that correlate with the occurrences of expected process outputs (results). Starting from an initial quality level $\alpha_0 = P(Y \leqslant 0) = P(Y = 0)$ and an expected quality output $\alpha > \alpha_0$, an optimal range is found if the conditional α-th quantile for that range $Q_\alpha(R_{ij})$ is zero. In such case the following condition is satisfied:

$$P(Y = 0|X_i \in R_{ij}) = P(Y \leqslant 0|X_i \in R_{ij}) \geqslant \alpha > \alpha_0 = P(Y = 0) \quad (5)$$

In practical terms, this means that a tolerance limit is optimal if the probability of obtaining good responses will increase when the predictor variable is bound to that range. By applying quantile regression to the data it is possible to find optimal ranges of factors that are associated with the expected quality output, namely low penalty values of response. Similarly to the penalty thresholds, the value α can be set by process engineers. Typically several quantile levels will be studied through quantile regression to achieve tolerance limit optimisation. When studying pairwise interactions between variables, this framework extends naturally by considering the conditional quantiles given the two variables being simultaneously in both ranges.

3.2. A novel risk based quantile regression algorithm for tolerance synthesis problems

The study of the single and combined effects of process variables on the process outcome is typically very cumbersome in multiprocess manufacturing systems as it involves the simultaneous analysis and comparison of the effects of many variables on one or more process outcomes (i.e. typically 30–50 variables). In this work, quantile regression trees are used to study the effect of different ranges of factors and their interactions on quantiles of penalty values. Quantile regression trees are grown for each pairwise combination of predictor variables (e.g. input factors) to study their single and interaction effects over the response variable. This is better than growing a single tree on all the input variables because it can capture also weaker interactions. Instead of choosing all the possible binary splits of the predictor variable, splits are based on quantiles following the same rationale as (Ransing et al., 2013). The input space is divided in four possible splits with eight ranges as described in Table 1. This splitting method is more appropriate since decreases computational cost and ensure a better distribution of observations over quantile ranges. Once interaction trees have been built for all pairwise combinations of variables, the algorithm selects all the single and interaction ranges with α-quantile zero. The algorithm for quantile regression used in this paper is a generalisation of LAD regression trees introduced by Breiman (1996) to estimate any arbitrary α-quantile level. The particular case for $\alpha = 0.5$ leads to the LAD regression tree method to predict the median. Given a set of possible splits at each node S, a tree is grown by choosing the split s that maximises the decrease of re-substitution error defined as:

Table 1
Ranges of factors.

Range name	Description
B25% (Bottom25%)	$x \leqslant 0.25$-quantile
T75% (Top75%)	$x > 0.25$-quantile
B50% (Bottom50%)	$x \leqslant 0.5$-quantile
T50% (Top50%)	$x > 0.5$-quantile
M50% (Middle50%)	0.25-quantile $< x <$ 0.75-quantile
B25%T25% (Bottom25% and Top25%)	$x \leqslant 0.25$-quantile and $x \geqslant 0.75$-quantile
B75% (Bottom75%)	$x <$ 0.75-quantile
T25% (Top25%)	$x \geqslant 0.75$-quantile

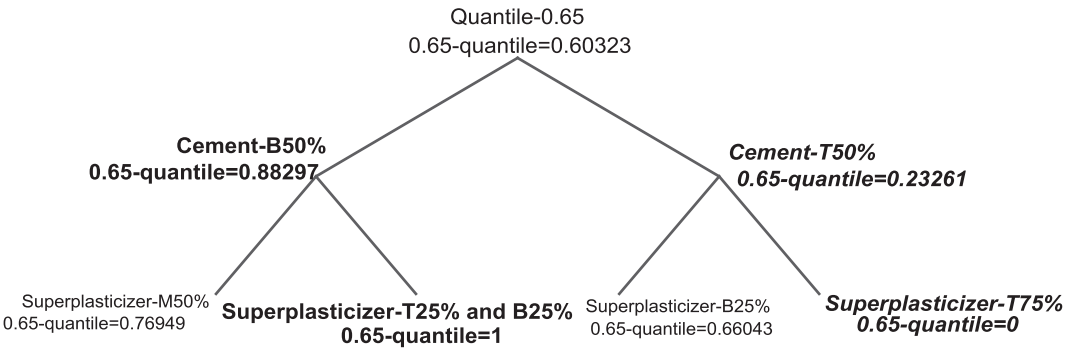


Fig. 2. An example of a tree obtained using quantile regression tree algorithm. Optimal ranges are those corresponding to response quantile equal to zero, avoid ranges are those corresponding to response quantile equal to one.

$$\frac{1}{n}\left(\sum_{x_n\in t} L(y_n, q(t)) - \sum_{x_n\in t_L} L(y_n - q(t_L)) - \sum_{x_n\in t_R} L(y_n - q(t_R))\right) \tag{6}$$

where $q(t)$ is the quantile at node t, y_n are the values of the response functions corresponding to the n^{th} observation and t_L and t_R are respectively the left and right nodes of the split. In the particular case of 0.5-quantile, the function $L(y_n, q(t))$ can be replaced by $|y_n - median(t)|$, leading to the LAD regression algorithm described in Breiman (1996).

Algorithm 1. Quantile Regression Tree to find Optimal Ranges.

```
foreach combination of factors X_i and X_j do
    stop = FALSE;
    while stop == FALSE do
        for each quantile range of X_i and X_j do
            Find best range to maximise decreases of ResubError defined in
             Eq. 6;
            if #obs < MinObs || no ResubError decrease then
                stop = TRUE
            end
        end
    end
end
foreach Regression Tree do
    foreach Terminal Node do
        if Quantile == 0 then
            Select Range as Optimal
        end
    end
end
```

By construction, regression trees will find ranges of factors that better separate instances of desirable/undesirable penalty values of responses. An example of a regression tree for $\alpha = 0.65$ can be seen in Fig. 2. The quantile regression method was applied to the Concrete dataset downloaded from the UCI Machine Learning library (UCI Machine Learning Repository, 2016). The dataset contains 1030 measurements of compressive strength for a given mixture at a specific age and it includes one quantitative output variable (compressive strength) and eight quantitative input variables. The regression tree in the picture shows interactions between two variables, namely Cement and Super-plasticiser. The interaction range with Cement-T50% and Super-plasticizer-T75% is identified as optimal because it is associated with low penalty values, being the 0.65-quantile zero. This is an improvement compared with the initial 0.65-quantile of 0.6. An avoid interaction range is also found when Cement is in B50% and Super-plasticiser in T25% and B25%.

Compared to regression trees based on ordinary least squares, the quantile regression tree method is more suitable to study in-process data because it is more robust to the presence of noise and outliers. Using the criteria based on the loss tiled function provides a better way to separate good and bad observations in the presence of noisy data. In order to increase robustness of the method, prior to the analysis with quantile regression trees, a variable selection step is performed using the Co-linearity index concept described in Giannetti et al. (2014). This is needed to select the most important variables. The Co-linearity index is an approximation of noise free correlations between a process response and process variables in a reduced dimensional space that accounts for the majority of the variance. Noise reduction is achieved by projecting the data in a lower dimensional space through Principal Component Analysis (Giannetti et al., 2014; Ransing et al., 2013).

3.3. Likelihood ratio test for new tolerance limits

The risk based thinking introduced by ISO9001 requires organisation to make improvements by assessing risks and opportunities. The optimal ranges found with regression trees can be used to develop hypotheses about possible adjustments to tolerance limits that are likely to lead to quality improvements. In order to test these hypotheses the Likelihood Ratio (*LR*) is used to mathematically quantify the strength of improvement and hence compare the effects of several proposed tolerance limits on the quality output. The Likelihood Ratio (*LR*) is commonly used to evaluate the accuracy of diagnostic testing and it is defined as:

$$LR = \frac{P(T+|D+)}{P(T+|D-)} \tag{7}$$

where $P(T+|D+)$ is the probability of a person testing positive given that the person has the disease and $P(T+|D-)$ is the probability of a person testing positive given that the person does not have the disease. A *LR* greater than 1 indicates that the test result is associated with the disease while a value less than 1 indicates that the test is associated with absence of disease. The concept of LR can be easily extended to tolerance limit optimisation where the $D+ = Optimal, D- = Avoid$ and $T+ = Range$ with the following definition:

$$LR = \frac{P(Range|Optimal)}{P(Range|Avoid)} \tag{8}$$

Similarly to diagnostic testing, $LR > 1$ indicates that the range of a variable (or combined range of variables) is associated with optimal process outcomes while $LR < 1$ indicates association with avoid process outcomes. In other words, the LR indicates how more likely is that process observations with optimal values are in the new tolerance limit (i.e. range) compared to process observations with avoid values. Another important property of the *LR* is that it also measures the ratio between pre-intervention (i.e. original tolerance limit) and post-intervention (i.e. new tolerance limits) odds:

$$LR = \frac{Odds(Optimal|Range)}{Odds(Optimal)}. \tag{9}$$

Hence the *LR* quantifies the extent of improvement of the new tolerance limit compared to the previous tolerance limit. In other words, it helps to quantify the robustness of the process. If the extent of improvement is small the process is said to be robust with respect to the input factor. The process improvement study should then focus on identifying new input factors that may explain the undesired deviation from expected results. If the extent of the improvement is large, a confirmation trial needs to be conducted to validate the hypothesis and create new product specific process knowledge. The 7Epsilon case studies give detailed information on implementing process improvement steps (Arjunwadkar, Ransing, & Ransing, 2015; Ransing & Ransing, 2014; Roshan et al., 2014) and embedding organizational knowledge management techniques (Giannetti et al., 2015).

The calculation of likelihood ratio involves estimation of conditional probabilities, namely $P(Range|Optimal)$ and $P(Range|Avoid)$. The latter can be calculated from available data using the maximum-likelihood estimate. Let c be a class (either *Optimal* or *Avoid*), the conditional probability can be estimated as:

$$P(Range|c) = \frac{n_{rc}}{n_c} \tag{10}$$

where n_{rc} is the number of observations in the given range and belonging to class c and n_c is the number of observations belonging to class c. For sparse data the calculation of the conditional probability may become problematic when only few observations for a given class exist. This often happens when analysing in-process data where the number of avoid observations might be limited. In this case estimation of the conditional probability using frequency counts might not me appropriate. In order to overcome this problem, it is suggested in the literature to use a smoothed estimation of probability (Cestnik & Bratko, 1991; Jiang & Li, 2011), defined as:

$$P(Range|c) = \frac{n_{rc} + m * p}{n_c + m}. \tag{11}$$

The term $m * p$ is a smoothing parameter that takes into account the prior probability p and a weighting factor m. How much weight to give to the initial probability is chosen by changing the values of the parameter m. The m-estimator is equivalent to Laplace estimator when $m = C$ and $p = 1/C$, where C is the number of classes. The Laplace estimators assumes that classes are distributed with uniform probability. In the current paper, the prior probability is chosen as the maximum likelihood estimation of $P(Range)$. The rationale behind this choice of smoothing is to correct the probability of each node taking into account the prior probability of the range. The parameter m is typically chosen by the analyst. Since m gives a weight to the prior probability of the class, the choice of m is determined by how much trust is placed on this initial estimation. Higher values of m indicate more confidence in the prior probability and typically lead to more conservative estimation of *LR*. In the industrial case study presented in this paper the value of m is chosen via numerical simulations that are further described in Section 4.

3.4. Uncertainty quantification with bootstrap

Uncertainty of *LR* estimations obtained from in-process data and associated risks are also further evaluated with the bootstrap method to overcome the lack of process knowledge and give confidence in the results before implementing costly confirmation trials. Bootstrap is a well know technique that is generally used for estimating the distribution of sample statistics by re-sampling the original data set (Efron, 1982; Efron & Tibshirani, 1986). Bootstrapping simulates new experimental data by randomly choosing observations (with replacement) from the available data to create a fixed number n of samples. This is used to estimate the empirical distribution, confidence intervals and standards errors of the unknown parameter. Bootstrap can be applied to evaluate uncertainty of different statistical parameters including correlation coefficient or regression models parameters (Efron & Tibshirani, 1986). Platt, Hanley, and Yang (2000) have also demonstrated the use of bootstrap for estimating confidence interval in diagnostic testing.

In this particular case, bootstrap is applied to the likelihood ratio estimation of in-process data. From the empirical distribution of the likelihood ratio process engineers can quantify uncertainty and reject hypotheses with high uncertainty before running expensive confirmation trials. It also gives an indication on the level of confidence in the results. The method is particularly useful in the absence of additional data to test the hypotheses. By using bootstrap it is possible to calculate confidence intervals of the likelihood ratio as shown in Fig. 3. The dashed lines indicate respectively the 5% and 95% confidence limits and the solid line is the mean of likelihood ratio for each range or interaction between ranges. Form the plots it can be inferred (with high level of confidence) that Age 25% is associated with optimal values of responses, being the 95% confidence limit of likelihood ratio about 10. Furthermore interaction between Cement-T25% and Water-B50% is also contributing to higher Compressive Strength since the mean likelihood ratio of interaction is higher than the respective main effect. The hypothesis that this interaction contributes to improvement of response values is supported by evidence from the data, being the 95% confidence limit of this interaction slightly less

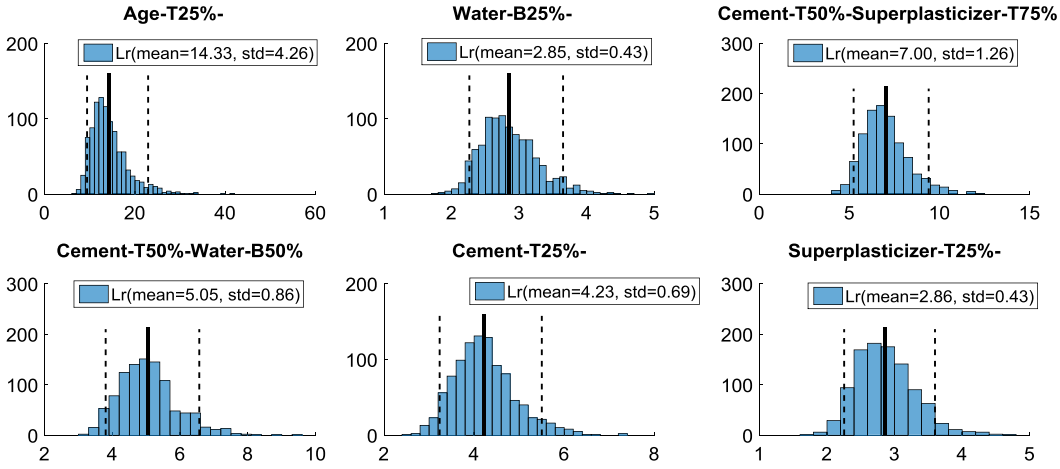


Fig. 3. The histograms of bootstrapped likelihood ratio ($n_{samples} = 1000$) can help to validate new hypothesis for process improvement identified with quantile regression. High values of likelihood ratio indicates that the suggested changes to the tolerance limits are likely to contribute towards process improvement targets. In order to accept an hypothesis the 95% confidence limit should be greater than 1, with higher values giving more confidence to include the selected ranges in a confirmation trial.

than 4. Under the assumption that the dataset is a good representation of the population, this implies that there is a 95% confidence that the likelihood ratio of the interaction will be higher than one, hence that the range will contribute to improve process robustness.

4. Application to an industrial case study

The proposed methodology has been applied to a real foundry case study for tolerance synthesis of a steel casting process. The data set has 16 process parameters related to chemical composition (input) and one response variable, occurrence of shrinkage defects. The response represents percentage of defective castings produced for a given volume of molten metal used. This is normally identified as a "heat" in the foundry domain. For the corresponding time period, the average value was considered for input factors with higher sampling rates. The data sets was used in previous studies and a full description can be found in Ransing et al. (2013, 2016). The main aim of the analysis is to achieve a reduction of shrinkage defects by changing the tolerance limits of the input parameters. All the process variables are continuous variables. A scatter plot of percentage of shrinkage defects is shown in Fig. 4, showing variability of shrinkage defects. From a process improvement perspective, process engineers aim to achieve process robustness by adjusting tolerance limits of the process inputs. Thresholds for penalty values are set to $T_{max} = 0.03$ and $T_{min} = 0$, respectively indicating undesirable and desirable values. A quantile analysis was performed for $\alpha = 0.75$ and results for optimal ranges are included in Table 2. For instance, Niobium-M50% range and interaction between Iron-75% and Aluminium-T50% are identified as optimal ranges, leading to a decrease of shrinkage defects. A visualisation of NiobiumT50% and interaction of Iron-T25% and Cobalt-T50% via bubble diagrams is also shown in Fig. 5. The size of the bubbles is proportional to penalty values and the colour scale indicates variation between optimal and avoid values (from blue to red).

It can be seen how the rectangular regions found by quantile trees separates observations with high and low penalty values. Quantile regression is used to identify possible regions associated with high or low penalty values. These regions are similar to those found with previous methods (Ransing et al., 2013, 2016). In addition to fining optimal regions the current methodology uses an additional bootstrap step with calculation of likelihood ratio which

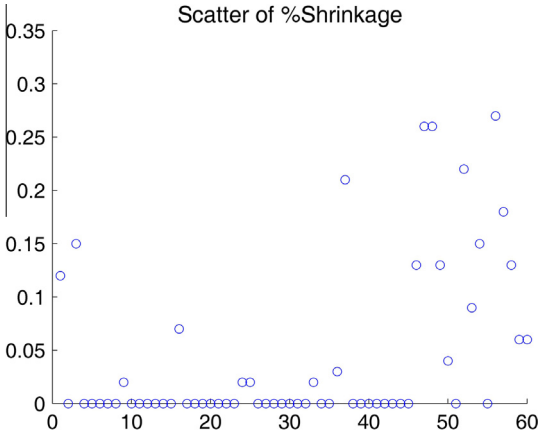


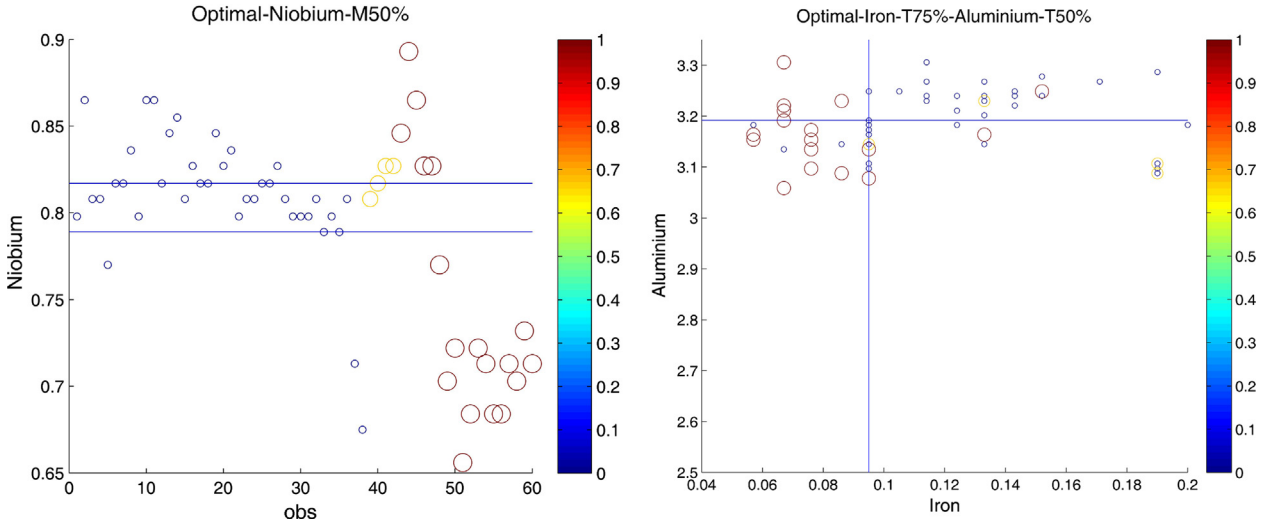
Fig. 4. The scatter graph of Shrinkage defects show variabilities of values.

Table 2
An extract of optimal ranges found by quantile analysis.

Range	Min	Max	Range	Min	Max	ResubError
Nb-M50%	0.79	0.82				0.016667
Fe-T75%	0.10	0.20	Co-T50%	7.86	8.03	0.017045
Al-T25%	3.24	3.31				0.025
Fe-T75%	0.10	0.19	Al-T50%	3.19	3.31	0.028409
Fe-T75%	0.10	0.19	Al+Ti-T50%	6.37	6.53	0.028409
W-T50%	2.46	2.59	Co-T25%-B25%	7.74	8.03	0.028846

can help process engineers to confirm or reject hypotheses. The result following the bootstrap simulation are shown in Fig. 6.

By studying the bootstrap plots, process engineers can choose which optimal ranges can be included in the confirmation trial. From the bootstrap analysis, it can be seen that there is some uncertainty concerning these results due to large variance of the distributions. However, there is some evidence that Niobium-M50% range will increase the probability of low penalty values. Similarly interaction of Iron-T25% and Aluminium-T50% is likely to decrease the occurrence of shrinkage defects. Following a study of bootstrap plots, a confirmation trial was designed as shown in Table 3. The confirmation trial include optimal ranges due to main effect (Niobium-T50%) and interactions (Iron-T75%-Aluminium-T50% and Iron-T75%-Tungsten-T50%).



(a) 1a (b) 1b

Fig. 5. Optimal ranges found with quantile regression trees leaning are shown in the scatter plot. The colour scale of observations (from blue to red) indicates the penalty values associated with these observations (0 representing optimal values and 1 representing avoid values). The quantile regression algorithm is able to identify regions of inputs parameters associated with a large number of optimal values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

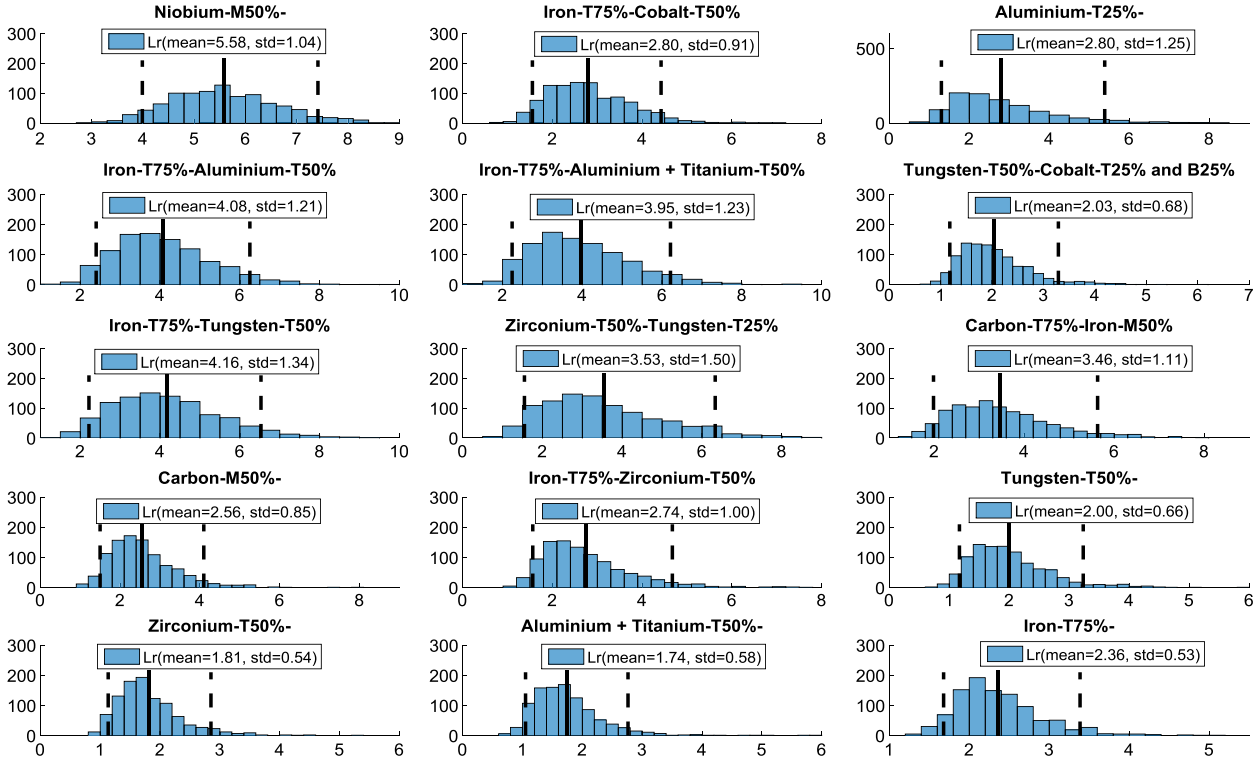


Fig. 6. Bootstrap analysis of optimal ranges with $n_{samples} = 1000$. The plots show large variance but several ranges such as Niobium-50% and interaction range Iron-T75%-Aluminium-T50% have the 95% likelihood ratio lower limit greater that 1, indicating that they are likely to contribute to improvement of response values.

Table 3
Optimal ranges chosen for confirmation trial.

Id	Factor-range	Min	Max	Reason
*R*1	Niobium-M50%	0.79	0.82	Main effect
*R*2	Iron-T75%	0.10	0.20	Interaction
*R*3	Aluminium-T50%	3.19	3.31	Interaction
*R*4	Tungsten-T50%	2.46	2.53	Interaction

For this analysis a value of $m = 6$ was used for smoothing the probability. The value m is chosen from experimental results to achieve a trade-off between variance and bias. Fig. 7 shows the mean likelihood ratio estimation and error bars from bootstrap with sample size 1000 for varying values of m. As m increases, the variance of likelihood ratio estimation decreases. An optimal choice of m is the one that corresponds to the point where the error

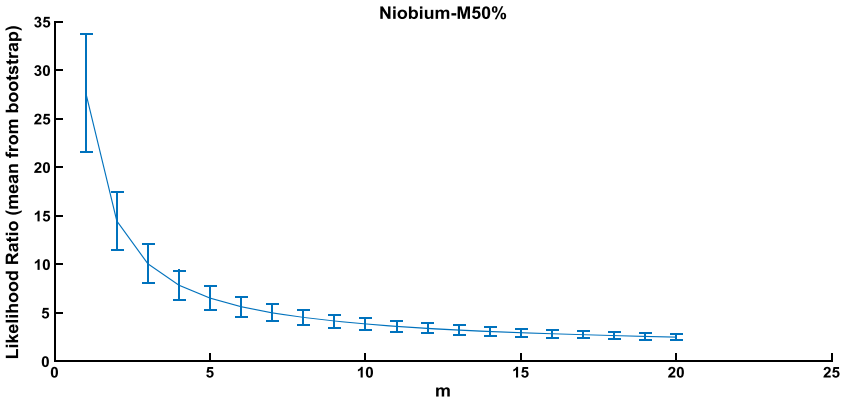


Fig. 7. Bootstrap simulations to calculate the likelihood ratio with varying values of the smoothing parameter m. The value $m = 6$ is chosen where the error bars (which are related to the variance) show a slow decrease.

does not decrease significantly. Choosing values $m > 6$ would lead to a small reduction of variance at the expense of higher bias. From the plot it can also be noted that the LR decreases as m increases. This is due to the fact that, having a small number of avoid observations, when calculating $P(Range|Avoid)$ the smoothing parameter m introduces new avoid observations which will increase the value of $P(Range|Avoid)$ and hence decrease the LR. This estimation of LR can be thought as a conservative estimation or worst case scenario.

4.1. Likelihood ratio estimation of overall confirmation trial

After selecting the ranges to include in the confirmation trial, the joint effect of the selected ranges over the response variable can be simulated with an additional bootstrap step and the overall likelihood ratio of the confirmation trial is calculated. The overall likelihood ratio can be calculated as a product of the single likelihood ratio in case the variables are conditionally independent with respect to the response variable. For instance, Niobium-T50% ($R1$) does not show any interaction with the other factors ($R2, R3, R4$). In this case it is safe to assume conditional independence based on evidence from the data and the overall likelihood ratio can be calculated as:

$$LR(R1 \cap R2 \cap R3 \cap R4) = LR(R1) * LR(R2 \cap R3 \cap R4). \quad (12)$$

Conditional independence cannot be assumed for $R2, R3$ and $R4$. In this case the likelihood ratio is calculated as follow:

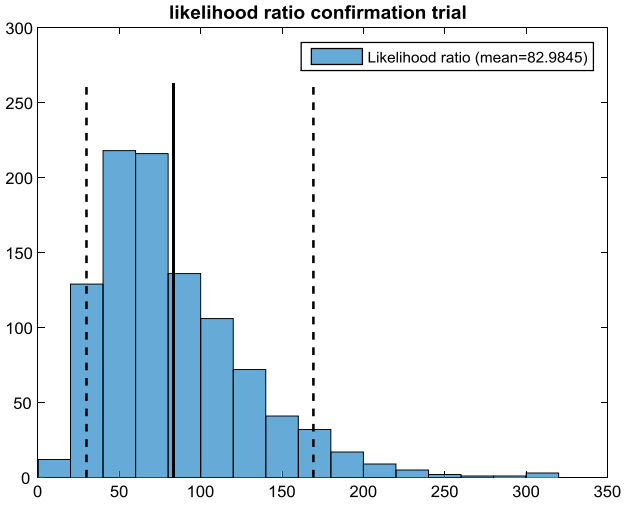


Fig. 8. The total likelihood of confirmation trial calculated with a further bootstrap step with 1000 samples.

$$LR(R2 \cap R3 \cap R4) = LR(R2) * LR(R3|R2) * LR(R4|R2 \cap R3). \quad (13)$$

An histogram of the predicted likelihood ratio of confirmation trial is shown in Fig. 8, with smoothing parameter $m = 6$. The mean likelihood ratio is estimated to be 83 with LR being greater than 32 in 95% of the cases. This indicates that the suggested changes will lead to significant improvement of response penalty values. Based on the available evidence, these results give a high degree of confidence that the new tolerance limits chosen will lead to significant improvements in the quality output and hence they can be included in a confirmation trial where these hypotheses can be further tested with new data.

5. Conclusion

Continual improvement is typically achieved by reducing variation in production processes to satisfy customer requirements. For complex manufacturing operations, consisting of several manufacturing process and with large number of process variables, keeping process variability within customers tolerance limits may not be enough to obtain high quality goods. Typically these processes have some intrinsic variability that only exists for a specific process and at a specific time. This may lead to situations where, despite the process being within the agreed tolerance limits, large variations are still present leading to sub-optimal operations, which will then lead to defective parts and high costs due to waste and rework. Industry 4.0 is likely to generate additional streams of in-process data where traditional Six Sigma based approaches become inapplicable. In these situations traditional statistical analysis techniques, such as regression analysis, are not sufficient to discover process improvement opportunities. Product specific process knowledge is often necessary to discover and validate hypotheses for improvements and make the necessary adjustments to enhance process performance (Giannetti et al., 2014, 2015; Ransing et al., 2013). Issues such as noisy and sparse data can also affect the statistical analysis leading to situation where process engineers may not have entire trust in the results, hence missing out in process improvement opportunities.

In the context of multiprocess manufacturing operations the concept of tolerance synthesis has been introduced in a recent publication (Ransing et al., 2016). This is defined as the study of variability of process inputs in order to discover optimal regions that correlate with the occurrences of expected process outputs (results). Tolerance synthesis allows process engineers to discover new and improved tolerance limits by developing and testing hypotheses regarding how variability of factors affects process responses. In this paper a novel mathematical formulation of the

tolerance synthesis problem for multiprocess manufacturing operations is presented. This new mathematical formalisation describes process improvement objectives in terms of quantiles of response variables. A quantile regression analysis with decision trees is also proposed to discover new tolerance limits by studying the single and combined effects of several partitions of the input parameters on variability of response values. The main advantage of the quantile regression tree algorithm is that it does not make distributional assumptions and does not require any prior knowledge about the type of relationship among input and output variables. A probabilistic estimation of the effectiveness of the proposed changes, using the likelihood ratio, is proposed. The likelihood ratio, typically used in diagnostic testing, quantifies the extent of improvement of the newly discovered tolerance limits on the quality output and provides a quantitative measure which can help process engineer to test hypotheses to overcome the lack of process knowledge. Since the likelihood ratio involves the calculation of conditional probabilities, in case of sparse data a smoothed probability function is used. Uncertainty associated with the likelihood ratio estimates is also quantified through a bootstrap step to give a degree of confidence in the results before running expensive confirmation trials. The method has been successfully applied to a real industrial scenario in the foundry industry, showing its effectiveness to predict process robustness. Because there is no assumption on the type of relationships and data distributions, the method is generic and can also be extended to other manufacturing processes. The proposed methodology will allow the organizations to extend Six Sigma principles for process improvements for the Industry 4.0 environment and implement 7Epsilon steps in order to satisfy various requirements of the newly proposed ISO9001:2015 quality standard.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) CHERISH-DE Centre (EP/M022722/1); and the European Regional Development Fund (ERDF) ASTUTE project. The authors would like to thank Dr. Meghana Ransing, CEO, p-matrix Ltd for sharing her experience on 7Epsilon and providing industrial context for this research. Credit is given to Mr. Phil George, Solution Architect Team Leader, Rockwell Automation for his interpretation of 7Epsilon as Six Sigma + IIoT. No new data were created during this study. Full description of the method and results are included in the manuscript.

References

7Epsilon Website. (2015). What is 7Epsilon?, retrieved from <http://www.7epsilon.org/what-is-7Epsilon.html>. Accessed 20/02/16.

Arjunwadkar, S., Ransing, M., & Ransing, R. (2015). Seven steps to energy efficiency for foundries. *Technical Paper, Foundry Management and Technology, 143*(3), 24–29.

Auret, L., & Aldrich, C. (2010). Unsupervised process fault detection with random forests. *Industrial & Engineering Chemistry Research, 49*(19), 9184–9194.

Bakır, B., Batmaz, I., Güntürkün, F., İpekçi, İ., Köksal, G., & Özdemirel, N. (2006). Defect cause modeling with decision tree and regression analysis. *World Academy of Science, Engineering and Technology, 24*, 1–4.

Bassett, G. W., Jr., & Chen, H.-L. (2002). Portfolio style: Return-based attribution using quantile regression. In *Economic applications of quantile regression* (pp. 293–305). Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees*. CRC Press.

Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment, 1*(8), 412–420.

Cestnik, B., & Bratko, I. (1991). On estimating probabilities in tree pruning. In *Machine learning-EWSL-91* (pp. 138–150). Springer.

Chaudhuri, P., & Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 561–576.

Coad, A., & Rao, R. (2008). Innovation and firm growth in high-tech sectors: A quantile regression approach. *Research Policy, 37*(4), 633–648.

Deloitte, A. G. (2015). *Industry 4.0 challenges and solutions for the digital transformation and use of exponential technologies*. McKinsey Global Institute. retrieved from. Accessed 13/01/16 <http://www2.deloitte.com/content/dam/Deloitte/ch/Documents/manufacturing/ch-en-manufacturing-industry-4-0-24102014.pdf>.

Dimelis, S., & Louri, H. (2002). Foreign ownership and production efficiency: A quantile regression analysis. *Oxford Economic Papers*, 449–469.

Ding, Y., Jin, J., Ceglarek, D., & Shi, J. (2000). *Process-oriented tolerance synthesis for multistage manufacturing systems. Manufacturing science and engineering-2000*. ASME.

Efron, B. (1982). Computer intensive methods in statistics. *Some Recent Advances in Statistics*, 173–181.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.

Foresight (2013). *The future of manufacturing: A new era of opportunity and challenge for the UK*. The Government Office for Science. retrieved from. Accessed 13/01/16 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/255922/13-809-future-manufacturing-project-report.pdf.

Francke, T., López-Tarazón, J., & Schroder, B. (2008). Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. *Hydrological Processes, 22*(25), 4892.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Geraci, M., & Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics, 8*(1), 140–154.

Germany Trade and Invest (2015). *Industry 4.0 smart manufacturing for the future*. Germany Trade and Invest. retrieved from. Accessed 13/01/16 <http://www.gtai.de/GTAI/Content/EN/Invest/_SharedDocs/Downloads/GTAI/Brochures/Industries/industrie4.0-smart-manufacturing-for-the-future-en.pdf>.

Giannetti, C., Ransing, R., Ransing, M., Bould, D., Gethin, D., & Sienz, J. (2014). A novel variable selection approach based on co-linearity index to discover optimal process settings by analysing mixed data. *Computers & Industrial Engineering, 72*, 217–229.

Giannetti, C., Ransing, R. S., Ransing, M. R., Bould, D. C., Gethin, D. T., & Sienz, J. (2014). Knowledge management and knowledge discovery for process improvement and sustainable manufacturing: A foundry case study. In *Proceedings of the 1st sustainable design and manufacturing conference, SDM14, Cardiff, UK*. .

Giannetti, C., Ransing, M. R., Ransing, R. S., Bould, D. C., Gethin, D. T., & Sienz, J. (2015). Organisational knowledge management for defect reduction and sustainable development in foundries. *International Journal of Knowledge and Systems Science (IJKSS), 6*(3), 18–37.

International Standard Organisation ISO (2014). *Draft BS EN ISO 9001 quality management systems - requirements*. Geneva, Switzerland: International Standard Organization.

Jemwa, G. T., & Aldrich, C. (2005). Improving process operations using support vector machines and decision trees. *AIChE Journal, 51*(2), 526–543.

Jiang, L., & Li, C. (2011). An empirical study on class probability estimates in decision tree learning. *Journal of Software, 6*(7), 1368–1373.

Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association, 96*(454).

Kim, H., & Loh, W.-Y. (2003). Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics, 12*(3), 512–530.

Koenker, R. (2005). *Quantile regression* (vol. 38). Cambridge university press.

Koenker, R., & Bassett, G. Jr., (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications, 38*(10), 13448–13467.

Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review, 35*(3), 223–240.

Lewis, R., Manzari, M., Ransing, R., & Gethin, D. (2000). Casting shape optimisation via process modelling. *Materials & Design, 21*(4), 381–386.

Lewis, R., & Ransing, R. (1997). A semantically constrained bayesian network for manufacturing diagnosis. *International Journal of Production Research, 35*(8), 2171–2178.

Lewis, R., & Ransing, R. (2000). The optimal design of interfacial heat transfer coefficients via a thermal stress model. *Finite Elements in Analysis and Design, 34*(2), 193–209.

Loh, W.-Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, 1710–1737.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14–23.

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica, 7*(4), 815–840.

Manyika, J. (2012). *Manufacturing the future: The next era of global growth and innovation*. McKinsey Global Institute. retrieved from. Accessed 13/01/16 <http://www.mckinsey.com/insights/manufacturing/the_future_of_manufacturing>.

Meinshausen, N. (2006). Quantile regression forests. *The Journal of Machine Learning Research, 7*, 983–999.

Pao, W., Ransing, R., Lewis, R., & Lin, C. (2004). A medial-axes-based interpolation method for solidification simulation. *Finite Elements in Analysis and Design, 40*(5–6), 577–593.

Platt, R. W., Hanley, J. A., & Yang, H. (2000). Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test. *Statistics in Medicine, 19*(3), 313–322. http://dx.doi.org/10.1002/(SICI)1097-0258(20000215)19:3<313::AID-SIM370>3.0.CO;2-K. URL.

Postek, E., Lewis, R., Gethin, D., & Ransing, R. (2005). Influence of initial stresses on the cast behaviour during squeeze forming processes. *Journal of Materials Processing Technology, 159*(3), 338–346.

Ransing, R., Batbooti, R., Giannetti, C., & Ransing, M. (2016). A quality correlation algorithm for tolerance synthesis in manufacturing operations. *Computers & Industrial Engineering, 93*, 1–11.

Ransing, R. S., Giannetti, C., Ransing, M. R., & James, M. W. (2013). A coupled penalty matrix approach and principal component based co-linearity index technique to discover product specific foundry process knowledge from in-process data in order to reduce defects. *Computers in Industry, 64*(5), 514–523.

Ransing, M., & Ransing, R. (2014). If only my foundry knew what it knows: A 7Epsilon perspective on root cause analysis and corrective action plans for ISO9001:2008. In *61st annual technical conference on investment casting, Covington, KY, USA*. .

Rockwell Automation. (2014). The connected enterprise maturity model. <http://www.rockwellautomation.com/global/innovation/connected-enterprise/maturity-model.page>. Accessed 03/03/16.

Roshan, H., Giannetti, C., Ransing, M., & Ransing, R. (2014). If only my foundry knew what it knows: A 7Epsilon perspective on root cause analysis and corrective action plans for ISO9001:2008. In *Proceedings of the 71st world foundry congress (WFC14), Bilbao, Spain*. .

Scharf, F. S., Juanes, F., & Sutherland, M. (1998). Inferring ecological relationships from the edges of scatter diagrams: Comparison of regression techniques. *Ecology, 79*(2), 448–460.

Shao, Q., Rowe, R., & York, P. (2007). Comparison of neurofuzzy logic and decision trees in discovering knowledge from experimental data of an immediate release tablet formulation. *European Journal of Pharmaceutical Sciences, 31*(2), 129–136.

Stricker, N., & Lanza, G. (2014). The concept of robustness in production systems and its correlation to disturbances. *Procedia CIRP, 19*, 87–92. 2nd CIRP Robust Manufacturing Conference (RoMac 2014).

UCI Machine Learning Repository. (2016). Concrete compressive strength. <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine, 25*(8), 1369–1382.

Yan, W. (2006). Application of random forest to aircraft engine fault diagnosis. *IMACS multiconference on computational engineering in systems applications* (Vol. 1, pp. 468–475). IEEE.