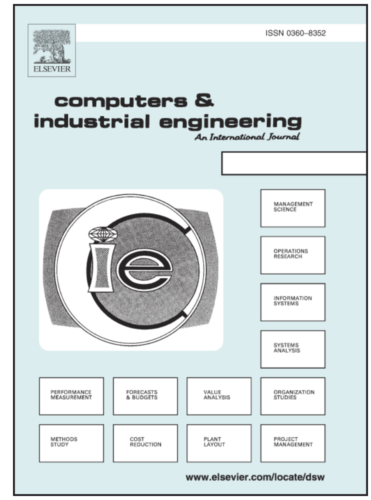
Accepted Manuscript



A quality correlation algorithm for tolerance synthesis and optimization in manufacturing operations

R.S. Ransing, Raed S. Batbooti, C. Giannetti, M.R. Ransing

PII: S0360-8352(15)00478-7
DOI: http://dx.doi.org/10.1016/j.cie.2015.12.008
Reference: CAIE 4202

To appear in: *Computers & Industrial Engineering*

Received Date: 11 May 2015
Revised Date: 1 December 2015
Accepted Date: 11 December 2015

Please cite this article as: Ransing, R.S., Batbooti, R.S., Giannetti, C., Ransing, M.R., A quality correlation algorithm for tolerance synthesis and optimization in manufacturing operations, *Computers & Industrial Engineering* (2015), doi: http://dx.doi.org/10.1016/j.cie.2015.12.008

A quality correlation algorithm for tolerance synthesis and optimization in manufacturing operations

R. S. Ransing[a*], Raed S. Batbooti[a,b], C. Giannetti[a], M.R. Ransing[c]

[a] College of Engineering, Swansea University, Swansea SA2 8PP, United Kingdom
[b] College of Engineering , University of Basra, Basra, Iraq
[c] p-matrix Ltd, Swansea, United Kingdom

Corresponding author

R. S. Ransing

College of Engineering, Swansea University Bay Campus, Swansea SA1 8EN, United Kingdom

Tel: 0044 1792 295219

Email: r.s.ransing@swansea.ac.uk

A quality correlation algorithm for tolerance synthesis and optimization in manufacturing operations

Abstract

The clause 6.1 of the ISO9001:2015 quality standard requires organisations to take specific actions to determine and address risks and opportunities in order to minimize undesired effects in the process and achieve process improvement. This paper proposes a new quality correlation algorithm to optimise tolerance limits of process variables across multiple processes. The algorithm uses reduced p-dimensional principal component scores to determine optimal tolerance limits and also embeds ISO9001:2015's risk based thinking approach. The corresponding factor and response variable pairs are chosen by analysing the mixed data set formulation proposed by Giannetti etl al. (2014) and co-linearity index algorithm proposed by Ransing et al. (2013).

The goal of this tolerance limit optimisation problem is to make several small changes to the process in order to reduce undesired process variation. The optimal and avoid ranges of multiple process parameters are determined by analysing in-process data on categorical as well as continuous variables and process responses being transformed using the risk based thinking approach.

The proposed approach has been illustrated by analysing in-process chemistry data for a nickel based alloy for manufacturing cast components for an aerospace foundry. It is also demonstrated how the approach embeds the risk based thinking into the in-process quality improvement process as required by the ISO9001:2015 standard.

Keywords: 7Epsilon, Six Sigma, No-Fault-Found product failures, in-tolerance faults, in-process quality improvement, and cause and effect analysis.

1. Introduction

ISO9001 is probably one of the most widely used quality standard today. Over 1.1 million organisations worldwide have been reported as being independently verified and certified as users of this standard. The 2015 revision of the ISO9001:2015 quality standard (ISO9001, 2015) is seen as a major change to its 2008 version. It has an overall focus on 'risk based thinking' so that industries are able to prevent undesirable outcomes in their processes. It defines risk as 'an effect of uncertainty'. In the context of this research, the definition is interpreted as follows: 'risk is an effect of uncertainty on an expected result'. Expected results or results to be achieved are set by objectives and are intended process outputs. The word 'effect' is further qualified as 'deviation from the expected' and 'uncertainty' implies 'deficiency of information related to the knowledge'.

Hence, any observed and unexplained deviation from expected results is inferred as a manifestation of deficiency of knowledge about the process. The '7Epsilon for ISO9001:2015' training course (Ransing and Ransing, 2015) defines risk as 'the effect of deficiency of knowledge (or uncertainty) that manifests itself as deviation(s) from desired process response values (or expected results)'. It quantifies the effect or the deviation from the desired response values with a penalty value. A zero penalty value is assigned for desired response values and 100 penalty value is denotes unacceptable process response (Ransing et al. 2013). Penalty values are scaled between zero and 100 for the remaining (or intermediate) process response values. In other words, it attempts to quantify the effect of uncertainty using penalty values. In a process based approach advocated by the ISO 9000 family of standards, the deficiency of knowledge relates to uncertainty on whether there exist any regions of values for process inputs that are likely to produce expected process outputs (results). In this context the process of evaluating tolerance limits in order to discover optimal regions of process values from in-process data is referred to as 'tolerance limit optimization'. As shown in Fig. 10 and discussed later in the paper, the optimal tolerance limits are discovered by analysing the process variance on mixed data sets. A tolerance limit for a process variable (factor or process input) is assumed to be robust if neither optimal nor avoid region is discovered.

1.1 Literature Review

A variety of data mining methods for clustering, classification, prediction and optimisation are reported in the literature. Early studies proposed Bayesian and Neural Network based models (Lewis & Ransing, 1997, Ransing & Lewis 1997). Recently, Koksal et al. (2011) have presented a detailed review of traditional data mining techniques for quality improvement in manufacturing industry. Colledani et al. (2014) have reviewed around 326 publications on manufacturing systems for production quality. However, both reviews have not discussed the in-process quality improvement solutions for analysing unexplained variation among production batches. The most relevant discussion in the review by Colledani et al. (2014) is on the multi-stage quality correlation analysis, Stream of Variations Analysis (SOVA) (Ceglarek et al. 2004) and No-Fault-Found product failures in service (Prakash et al. (2009). The quality correlation concept assumes that the quality of the product is dependent on the quality of the output at specific upstream processes. The similarity between the proposed approach and the SOVA is that both approaches associate product quality to the covariance of process variables and are used for making process adjustments that prevent products from falling into in-tolerance fault regions. However, the SOVA method is designed for multistage assembly process relating key product and control characteristics represented in CAD/CAM models with information about process layout and sequence of operations. The proposed approach is designed for a production environment where the unexplained variation in quality among product batches needs to be associated with the corresponding process variability.

Striker and Lanza (2014) define a robust process as a process that is resilient to either disturbances or differing conditions. They have used the same interpretation of risk as given by the ISO 9000 standard that it is the effect of uncertainty on objectives. They have also advocated relating risk to process robustness. The disturbance is set to have negative influence on at least one of the three dimensions: cost, time and quality. According to authors the existing risks are the disturbances that the production system is exposed to and the robustness of the production system is increased by reducing the influence of the existing disturbances. The authors argue that an optimal solution should correspond to a production system configuration that includes all machinery, equipment, staff and organisational processes. However, the paper does not propose a methodology for addressing process improvement opportunities and risks.

The novelty and originality of this research lies in the following:

- Definition and quantification of undesired process outputs (Ransing et al. 2013) and its relation to ISO9001:2015's risk based thinking.
- Extension of the mixed data formulation for co-linearity indices and data pre-treatment methods proposed by Giannetti et. al. (2014) to calculate the corresponding lower dimensional principal component scores.
- Propose a new methodology based on using the lower dimensional principal component scores to determine potential optimal and avoid regions as well as process interactions. These regions are not constrained to the pre-determined

quartile limits as previously suggested (Ransing et al. 2013, Giannetti et al. 2014) and can be considered as opportunities for process improvement as defined in clause 6.1 of ISO9001:2015 quality standard.

In this paper the proposed algorithm is presented with reference to a foundry case study. Section 2 gives background information on the foundry case study and relates the previous work to the problem under discussion. The main algorithm for discovering optimal tolerance limits for one or more process factors is described in Section 3. The results on a foundry in-process quality improvement case study are discussed in Section 4. Finally the paper is concluded in Section 5.

2. Background information

Foundry or metal casting process is a complex process with many sub-processes such as pattern making, mould and core making, melting and pouring process etc. For an investment casting process, the mould (or shell) making process may have further sub-processes such as coating and drying processes. Investment casting foundries produce very complex shaped and alloyed components such as turbine blades. Sometimes, it takes months to produce a turbine blade from the initial wax processing stage to the final casting. A typical continual process improvement study may have over 60-70 measurable process inputs that govern the quality of the final turbine blade.

On average precision foundries lose 3-5% of their revenue in rejected or reworked castings. In a foundry environment such a process is referred to as a stable process. Many foundries have a higher internal rejection rate. The challenge for foundry process engineers is to be able to make small adjustments to several process parameters (e.g. slight adjustments to the tolerance limits of various parameters such as alloy compositions at various states of melting and pouring process, pouring temperature, moulding parameters etc.). Undertaking one change at a time is not sufficient. Even for experts, it is not easy to choose the top 3-4 critical process variables that can be shown as being responsible for causing a 3-5% rejection rate. The proposed formulation is applicable for similar industrial conditions. A sample in-process data on production batches is shown in Table 1 and explained in the next subsection.

2.1 Risks, penalty values and bubble diagrams: a foundry case study

The continual process improvement example presented in an earlier publication (Ransing et al. 2013) is used here to illustrate how the penalty matrix approach offers a natural way of quantifying changes in risk as defined by the ISO 9001:2015 standard and the opportunities for creating additional knowledge resulting from the deficiency of knowledge (or uncertainty) that creates the risk.

An investment casting foundry manufacturing nickel based super alloy castings has variability in castings rejected due to shrinkage related defects per melt (i.e. casting produced per fixed amount of molten metal). Chemical composition readings and other processing parameters were noted and are shown in Table 1. This section highlights similarity between the risk based thinking and the penalty matrix approach.

Table 1 Sample in-process data relating percentage of casting rejected due to shrinkage related defects per melt (process output). The corresponding process inputs on chemical composition for the nickel based alloy are shown. Two process inputs Niobium and Tungsten are discussed further in the text

Shrinkage %	Carbon	Aluminium	Boron	Cobalt	Chromium	Iron	Molybdenu	Niobium	Tungsten	Tantalum	Titanium	Zirconium	Aluminiu	Nitroger	Oxygen	Ta/Ti
0.12	0.101	3.23	0.009	7.857	15.2	0.086	1.663	0.846	2.556	1.587	3.23	0.037	6.46	33.25	6.65	0.492
0	0.093	3.145	0.009	7.971	15.295	0.086	1.644	0.798	2.594	1.558	3.211	0.05	6.365	11.4	6.65	0.486
0.15	0.107	3.249	0.009	7.781	15.248	0.152	1.691	0.893	2.423	1.653	3.278	0.031	6.527	38	10.45	0.505
0	0.103	3.249	0.008	8.028	15.096	0.105	1.653	0.865	2.489	1.568	3.211	0.035	6.46	22.8	7.6	0.489
0	0.105	3.183	0.008	7.781	15.001	0.124	1.682	0.808	2.423	1.52	3.107	0.032	6.289	21.85	5.7	0.49
0	0.107	3.107	0.008	7.8	15.295	0.19	1.663	0.808	2.442	1.615	3.145	0.022	6.251	20.9	8.55	0.514
0	0.109	3.145	0.01	7.866	15.267	0.095	1.691	0.77	2.451	1.653	3.192	0.024	6.337	20.9	9.5	0.518
0	0.112	3.287	0.009	7.743	15.305	0.19	1.663	0.817	2.461	1.672	3.211	0.023	6.498	26.6	5.7	0.521
0.02	0.106	3.145	0.009	7.838	15.352	0.095	1.644	0.808	2.461	1.596	3.164	0.023	6.308	38.95	11.4	0.505
0	0.106	3.249	0.008	7.809	15.276	0.095	1.634	0.817	2.48	1.558	3.173	0.024	6.422	30.4	3.8	0.492
0	0.108	3.097	0.008	7.781	15.286	0.095	1.653	0.836	2.432	1.625	3.202	0.023	6.299	29.45	7.6	0.508
0	0.108	3.183	0.008	7.828	15.286	0.095	1.634	0.798	2.48	1.577	3.145	0.026	6.327	20.9	7.6	0.502
0	0.106	3.24	0.008	7.857	15.02	0.143	1.663	0.865	2.518	1.615	3.173	0.03	6.413	32.3	5.7	0.509
0	0.108	3.268	0.009	7.895	15.267	0.171	1.672	0.865	2.47	1.568	3.211	0.03	6.489	32.3	7.6	0.489
0	0.102	3.306	0.008	7.885	15.248	0.114	1.634	0.817	2.489	1.492	3.145	0.027	6.451	36.1	11.4	0.475
0.07	0.102	3.306	0.009	7.942	15.229	0.067	1.663	0.865	2.47	1.549	3.183	0.031	6.489	20.9	7.6	0.487

The risk based approach requires organizations to categorise process outputs as acceptable and unacceptable outputs. This is best done by studying the variation in process outputs plotted as a scatter diagram Fig. 1. Note that the number of castings produced for the same product from each melt are grouped and characterized as a production batch. The percentage of castings rejected in each batch is shown in Fig. 1. Majority of casting batches had zero percentage rejections. However, some batches had unexplained deviation from expected results of zero percentage rejections. This deviation is interpreted as an effect of the possible deficiency of process knowledge (or uncertainty). The deficiency in the process knowledge is linked to tolerance limit optimization where the hypothesis is the existence of optimal regions within the tolerance limit (or minimum and maximum values) observed for process inputs for all values of process outputs. Few examples of potential hypotheses that suggest deficiency of process knowledge are given below. These examples are illustrated with respect to the tolerance limit of process parameter %Niobium (Fig. 2). Few examples of uncertainty are:

a. Should the target value of %Niobium be 0.75, 0.8 or 0.85?[1]
b. Is the range robust? Or should it be changed?
c. Should the lower or upper limit be changed simultaneously or individually?
d. In terms of quartiles, is the top 25%, top 50%, bottom 50%, bottom 25% or middle 50% quartile any better than the current range?
e. Is there any other optimal or avoid range within observed values?

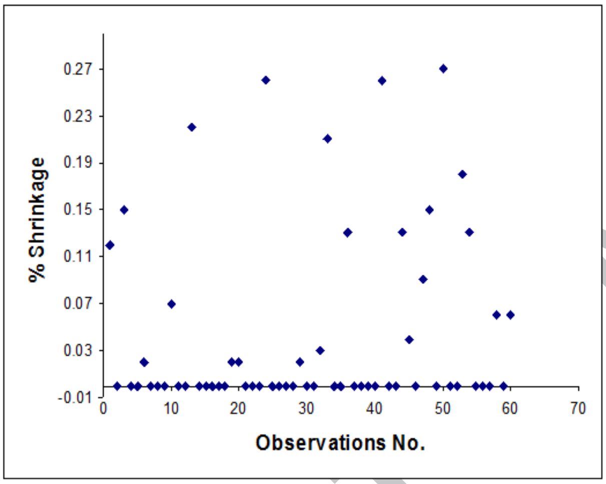


Fig. 1 Deviation from expected results (or desired response values) for the process output casting rejected due to the shrinkage defect per melt

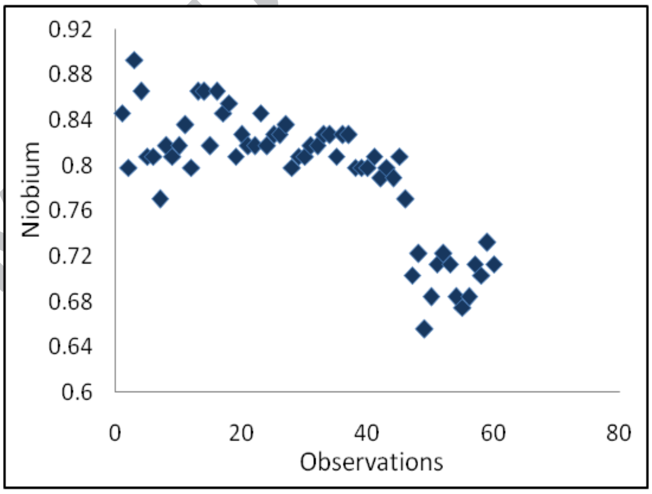


Fig. 2 Scatter diagram for factor Niobium. The uncertainty or the deficiency of knowledge is the hypothesis that optimal regions may exist within the tolerance limits of process inputs

A process drift for %Niobium is clearly observed in subsequent melts. However, it is noted that the Niobium percentage was always maintained within the required process specification. Over last twenty years, the first author has worked with numerous foundries across the world and it is not uncommon in the foundry world that two process experts from the same foundry will fine-tune the process differently. It is often a challenging task for the top management to find a common ground. The proposed formulation offers an evidence based solution to discover the existence of robust and optimal tolerance limits. Although most of the process

[1] Note that such small adjustments to various process inputs are outside the scope of traditional design of experiments. A typical design of experiment study is more likely to choose a much wider range of % Niobium values e.g. 0%, 0.5%, 0.75% and 1%.

variables are likely to follow a Gaussian distribution, the analysis of in-process foundry data has consistently shown an existence of occasional skewed distribution. Giannetti et al. (2014) have suggested data pre-treatment method and a median based covariance PCA formulation for mixed datasets that increase variance contributions for skewed in-process data. As a result, such process variables receive increased weighting and are put under the microscope. This technique has also been implemented in the proposed formulation.

It is hypothesized that the effect of this uncertainty is the resulting deviation from expected results. However, if the process engineer believes that the proposed niobium range for his or her foundry is robust then there should not be any statistically significant correlation between one or more regions of niobium values and shrinkage values. However, care needs to be taken as correlation does not always imply causation. The domain knowledge on how various chemical elements for nickel based alloys influence shrinkage related defects and subsequently mechanical properties is necessary. The numerical simulation software tools can also be used to optimize the design and process parameters to achieve the reduction of defects (Lewis & Ransing, 2000; Lewis et al., 2004; Pao et al., 2004; Postek et al. 2005).

The penalty matrix approach has suggested various transformations on raw process input and output data (Ransing et al. 2013, Giannetti et al. 2014) in order to balance mixed data sets that combine process inputs, continuous and discrete values. This is briefly summarized in Section 2.2.

The changes in the risk are quantified by identifying regions of acceptable and unacceptable variations in process outputs in Fig. 1. These regions are associated with a penalty value. The acceptable and unacceptable results are given a zero and hundred penalty value respectively and a penalty value between zero and hundred is assigned for all process output values between the acceptable and unacceptable thresholds Fig. 3. The corresponding penalty values are transferred onto factor (process input) scatter diagram as shown in Fig. 2 to transform it to a bubble diagram Fig. 4. The penalty values can be further transferred on multiple process inputs. An example of interaction with two process inputs (%Niobium and %Tungsten) is shown in Fig. 5.

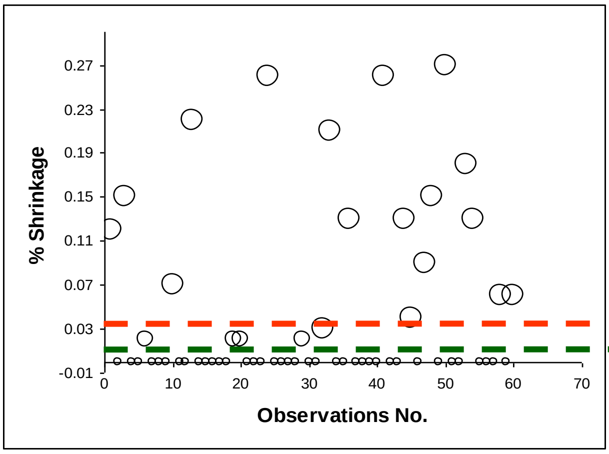


Fig. 3: The deviation from expected results (or desired response values) is penalized. ISO 9001:2015 terms this as changes in risks and requires organizations to address risk to achieve improvement.

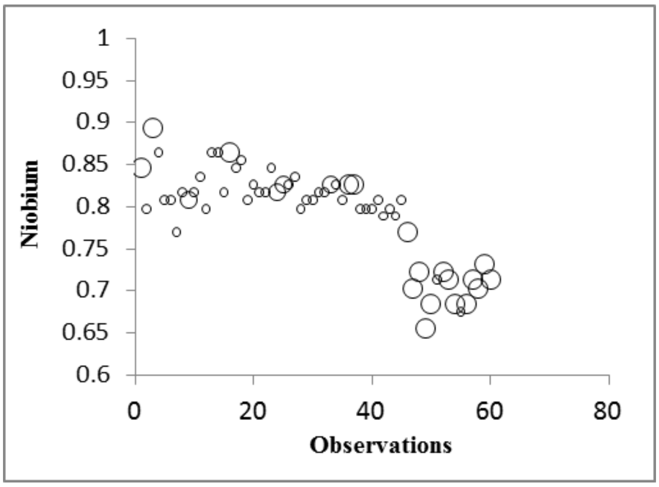


Fig. 4: The corresponding penalty values are transferred on all relevant factor (process input) scatter diagrams in order to generate hypotheses for taking potential actions as required by the clause 6.1 of ISO 9001:2015

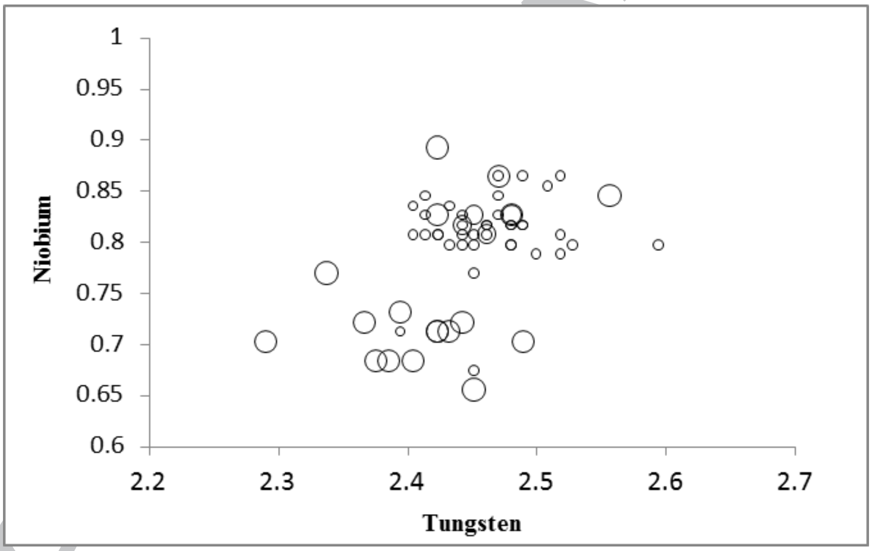


Fig. 5: The penalty values can be further transferred onto values for two or more factors (process inputs) to discover hypotheses related to interactions.

It is not always practical to visualize in-process data using bubble diagrams as shown in Fig's 3-5 with overlapping observations and also when number of observations increase. The algorithm presented in Section 3 has an ability to discover optimal ranges e.g. an optimal range for %Niobium is identified as 0.77-0.865 and shown in Table 6, Section 4.2. Similarly, the algorithm also discovered an optimal interaction for %Niobium in range 0.77-0.865 and %Tungsten in range 2.413 – 2.594 as shown in Fig. 5. These results are discussed in detail in Section 4.

2.2 Data pre-treatment and calculation of co-linearity index in a reduced *p*-dimensional space

The data pre-treatment transformations are normally performed before undertaking principal component analysis (PCA). Data centring by extracting the mean value and standardising to a unit standard deviation is a widely used method. Various scaling techniques such as range scaling, pareto scaling, log and power transformations, level scaling and max scaling are used depending on the goal of a PCA analysis (Bécue-Bertaut & Pagès, 2008; Parente & Sutherland, 2013; Van den Berg et al. 2006). It was pointed out by Parente & Sutherland (2013) that the scaling can be used to boost the correlation between variables in PCA. In the present work, three different pre-treatment methods are used depending on the variable classification: response scaling, categorical variables scaling, and quantitative variables scaling. Response scaling is achieved using penalty values, as described in Section 2.1, where response values are transformed in the range [0 1] according to maximum and minimum threshold values (T_{max} and T_{min}). These thresholds quantify acceptable and unacceptable process response and are chosen by the analyst. The choice is normally achieved by penalising the worst 10-15% observations corresponding to at least 5-10 bad points and use zero penalty values for the best 30-40% of observations corresponding to at least 10-20 good points (Gianneti et al., 2014).

The categorical variables scaling is mainly based on the principles of Multiple Factor Analysis (MFA) (Escofier and Pages, 1994). Categorical variables are transformed into continuous variables by introducing an indicator variables z_i which takes a value of one when the original category has occurred and zero otherwise. A further transformation, based on equivalence of Multiple Correspondence Analysis (MCA) and weighted PCA, is applied to categorical data so they can be treated as continuous data for the purpose of PCA, as described in Giannetti et al (2014)and Bécue-Bertaut et al. (2008).

Scaling of continuous variables is performed using a data pre-treatment that uses median (med) and interquartile (iqr) range as shown in Table (2). This table also includes all data transformations used in the present work with all necessary descriptions.

Table 2: Data transformations applied to raw data (Giannetti et al, 2014).

Data Set	Description	Transformation	Scaling factor
Response (lower the better) (Resp)	Penalty value Scaling	$x_{ij}^{(k)}=\begin{cases} 0 & \text{if } x_{ij} \le T_{min} \\ 1 & \text{if } x_{ij} \ge T_{max} \\ \frac{(x_{ij}^{(k)} - T_{\min})}{T_{\max} - T_{\min})} & \text{otherwise} \end{cases}$	$\sqrt{\frac{1}{\lambda_{\max,\text{Resp}}}}$
Response (higher the better) (Resp)	Penalty value Scaling	$x_{ij}^{(k)}=\begin{cases} 0 & \text{if } x_{ij} \ge T_{max} \\ 1 & \text{if } x_{ij} \le T_{min} \\ \frac{(x_{ij}^{(k)} - T_{\max})}{T_{\min} - T_{\max})} & \text{otherwise} \end{cases}$	$\sqrt{\frac{1}{\lambda_{\max,\text{Resp}}}}$
Categorical variables (*CVar*)	MFA transformation	$x_{ij} = \frac{z_{ij}-w_j}{w_j}$, $w_j = \sum_{i=0}^{m} \frac{z_{ij}}{m}$	$\sqrt{\frac{w_j}{\lambda_{\max,CVar}}}$
Quantitative Variables (*QVar*)	Robust standardization	$x_{ij} = \frac{x_{ij} - med_j}{iqr_j}$	$\sqrt{\frac{w_j}{\lambda_{\max,QVar}}}$

2.2.1 Co-linearity index

As explained by Giannetti et al. (2014), the data matrix X_T is created after applying data pre-treatment transformations as suggested in Table 1 on the sample data matrix (X) shown in Table 1. The PCA is performed on covariance matrix *Cov* as follows:

$$Cov = \frac{1}{m-1} X_T^t X_T$$

The loading matrix is calculated based on the following equation

$$L_s = D_s^{-1} V D_e$$

Where: L_s is the standardized loading matrix, V is the matrix of eigenvectors arranged as column vectors in descending order of eigenvalues, D_s is the diagonal matrix of the standard deviations of the columns of X_T and D_e is the diagonal matrix containing the square roots of eigenvalues.

The first p significant principal components are identified using a scree plot (Cattell, 1996). In the reduced dimensional space spanned by the first p components, the inner product of i^{th} and j^{th} row vectors of $L_{s,p}$ represents the correlation between variable i and j. After that co-linearity index can be plotted by plotting angles and length of the loading vectors. The co-linearity index plot is then divided into five regions: a region with no correlation when the co-linearity index value is between -0.2 to 0.2, two regions that identify weak correlations with co-linearity index values -0.5 to -0.2 and 0.2 to 0.5 respectively and the two strong correlation regions which include co-linearity index between -1 to -0.5 and 0.5 to 1 for negative and positive correlation respectively.

The weakly correlated variables, in addition to strongly correlated variables, are chosen for further investigation for exploring interactions and the corresponding number is referred to as n_c. The p-dimensional principal component space is chosen for projecting scores on the corresponding loading vectors based on their co-linearity index value. The details of these concepts are explained in the next section.

3. A quality correlation algorithm for robust tolerance limit optimization

Principal component scores represent the original data matrix X_T in the principal component space. For a number of components p, the dimensions of the scores matrix (T_p) is $m \times p$, where m is the number of observations and p is the number of principal components. The score matrix (T_p) in a p-dimensional principal component space can be expressed as:

$$T_p = X_T \, L_{s,p}$$

Geometrically, each row of the scores matrix is the projection of the respective observations in the p-dimensional space. In the literature, scores plots have been used to discover similarities between observations (Begam and Kumar, 2014). The similarity and clustering behaviour of scores depends on the contribution of the observations on principal components. The importance of an observation for a component is determined by the magnitude of the squared score value (Abdi & Williams, 2010). Chen et al. (1998) used PCA to analyse the Fourier transform infrared spectroscopy spectra of cotton fiber based on separating the spectra into different groups according to its PCA scores. It was shown that the scores of spectra are divided into two groups separated by the first two principal components, which lead to two different causality regions. The clustering of data sets can be revealed from projected scores in the reduced dimensional space of principal components.

A scores plot is a scatter plot and it is normally used to visualize the scores of observations on the first two principal components (or any other pair of components) in order to show the distances between observations to discover the similarities (Varmuza and Filzmoser, 2009). On the other hand, the loading plot is widely used to show the correlation between variables in two dimensions, where each variable is represented by a vector whose coordinates are determined by the loading matrix.

In order to understand the relation between observations and variables, scores and loading plots are often merged together and plotted for the first two or three principal components. These are referred to as biplots and can be used to discover features of data and similarities between variables and scores (Gabriel and Odoroff, 1990). On the biplots, the scores are represented by points and the loadings are represented by vectors. An example of biplot for the nickel based alloys case study discussed in this paper is depicted in Fig. 6. In most cases a two dimensional biplot is able to give an indication about the contribution of the scores on variables (Jollife, 2002).

With the view of the data centring transformations proposed in Table 2 followed by the mean centred analysis of the PCA, it is noted that, for each variable, the origin in the principal component space corresponds to its central (or mean) value and the distance of the loading vector becomes proportional to the standard deviation (Gabriel, 1971; Jollife, 2002). A longer length for the response variables denotes higher standard deviation in penalty values or larger deviations from the expected results. This is a direct consequence of the data transformation for response variables. Thus, it is inferred that the proposed p-dimensional principal component space becomes a reference frame that embodies risk based thinking and allows choosing observations that are characteristic to the position of variables in this space. The observations, with positions far from the origin, account for larger proportion of the variance and in particular, the contribution of an observation to the variable variance is determined by its projection onto the loading vector of the variable. In the proposed algorithm only variables that correlate either positively or negatively with response values are chosen.

In order to relate scores with corresponding variables in the p-dimensional principal component space, a new plot for each variable is derived from projecting p-dimensional scores respectively on the variable and response vectors to infer which scores contribute in the variable direction and the response direction. It is noted that the response direction is considered as the datum direction and co-linearity indices, or cosines of angles, which are considered as measures of the correlation, are calculated with reference to the response direction. For instance, in the Nickel based alloy example, the variable Carbon is correlated negatively with the response. For risk based thinking approach, the response variable (e.g. %shrinkage defect occurred in a batch of components) measures the deviation from desired values with a penalty function. This means that higher values of Carbon are inversely correlated with higher penalty values for the response variable. Thus, the contribution area is bounded by a positively projected Carbon direction and negatively projected response (shrinkage) direction. In other words, it infers an hypothesis that within the range of the %Carbon in the observation data, the increase of the Carbon content is correlated with the reduction of the %shrinkage defects. It is also observed in Fig. 7 that the majority of corresponding scores in the bounded area have low penalty values and hence lower bubble diameters. The projection of score t_i on loading L_j in p dimension is given by the following equation.

$$t_i^* = \frac{\sum_{k=1}^{p} L_j(k) * t_i(k)}{\sqrt{\sum_{k=1}^{p} (L_j(k))^2}} = \frac{L_j . t_i}{\|L_j\|}$$

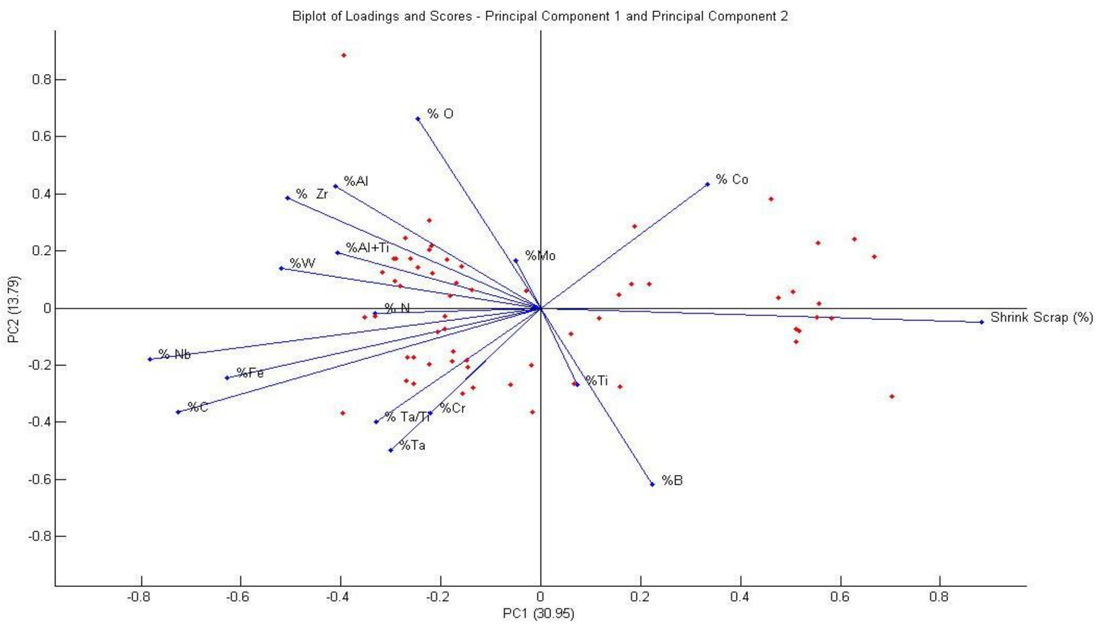


Fig. 6: Biplot for first two principal component of Nickel based alloy.

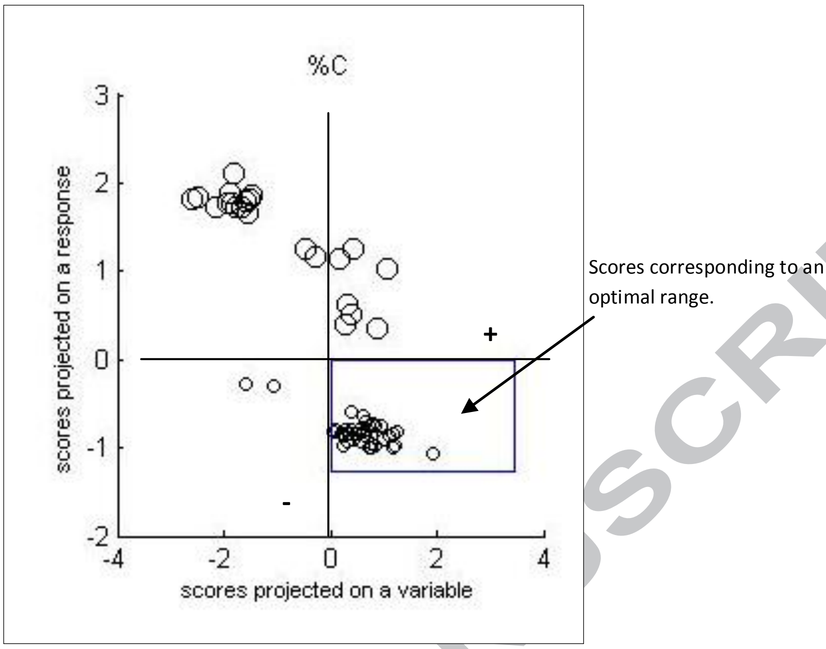


Fig. 7. Scores projection on %C and response to show the optimal scores area.

3.1 The quality correlation algorithm steps

The main steps for calculating the optimal system settings for quantitative and categorical variables are described in Table 3. Firstly the co-linearity index method, as described in Section 2.2.1, is used to find the correlated variables n_c. The scores are projected on all variables and responses. The corresponding scores for a j^{th} correlated variable are collected based on direction of variable and response as explained in the previous section. These scores relate to either optimal or avoid settings with reference to the correlated variable. The observations corresponding to the collected scores stored in new variables $x^j_{optimal}$ or x^j_{avoid} depending upon whether the correlation is positive or negative. The number of observations stored are counted and stored in a variable n^j_x. The minimum and maximum values for variable j are chosen from observations used in vectors $x^j_{optimal}$ or x^j_{avoid}. These values determine the range which is further explored for its optimality. The obtained range is considered as an optimal range if (i) the variable is correlated positively with low penalty response vector and, (ii) the majority of observations in $x^j_{optimal}$ or x^j_{avoid} have low penalty values. The range is considered as avoid if the variable is correlated positively with high penalty response vector and majority of observations in $x^j_{optimal}$ or x^j_{avoid} have high penalty values. In case of categorical variables, the corresponding original variables vector will have binary values. A new variable, percentage of occurrences Po^j, is introduced as an indication for the true positive occurrences for a specific variable j in x^j:

$$Po^j = \frac{\sum_{i=1}^{n_x^j} x(optimal)_i^j}{n_x^j} \times 100\% \text{ or } Po^j = \frac{\sum_{i=1}^{n_x^j} x(avoid)_i^j}{n_x^j} \times 100\%$$

Where n^j_x= number of elements in x^j.

The categorical variable j is chosen for recommendation as optimal classification if $Po^j \geq 60$ % and is negatively correlated with penalty values high penalty values. The recommendation will be 'avoid' for a categorical variable i if $Po^i \geq 60$ % and the variable is positively correlated with high penalty values.

3.2. Interactions

Interactions are used to discover strong combined effects of correlated variables when the combined effect is stronger than the individual ones. Traditional regression analysis and statistical based approaches calculate interactions by taking a product of mean centred variables. During the course of the research, it was realised that the process engineers preferred visualisation of interactions using bubble diagrams and penalty matrices rather than the traditional approach to predict interactions (Ransing et al. 2013). The proposed formulation allows robust calculation of optimal regions and the subsequent discovery of interactions so that they can be visualised using bubble diagrams and penalty matrices. The interactions and optimal regions discovered in this way may not always concur with traditional approaches and the authors do not expect the proposed method to give similar results obtained by the usual interaction formulation. Instead authors recommend process engineers to visualise interactions using penalty matrices, apply domain knowledge, follow the 7 steps of 7Epsilon and verify results via confirmation trials. The specific references to the statistical methods have also been dropped from the ISO9001:2015 revision. Hence, discovery of interaction using the proposed approach and its subsequent visualisation using penalty matrices has been considered as an appropriate step for satisfying the requirements of clause 6.1 of the ISO9001:2015 quality standard (Ransing and Ransing, 2015).

A binary interaction variable is created with a unit value when both individual variables are within the ranges as described in Section 3.1. A zero value is assigned otherwise. For example the interaction between variable Carbon and variable Iron (see section 4.1) is accomplished by creating a categorical interaction variable that takes value of one when Carbon is in the optimal range (0.093-0.112) and when the value of Iron is within (0.095-0.2). For all other combinations a zero value is assigned. The interaction algorithm is described in Table 4. This interaction variable is chosen if its co-linearity index is greater than the co-linearity index of its individuals. Using the method described in Section 2.1.2, the newly developed interaction variable is combined with all original variables and a combined co-linearity index plot is generated for discovering correlations and studying scores.

Table 3: The algorithm for discovering optimal/avoid range process settings

	Optimal range process settings for quantitative variables
Step1	Collect in-process data as shown in Table 1. Use the batch data collected in sequence for all good and bad batches. Apply the data pre-treatment method as described in Table 2. Choose one response and treat all other responses as factors. Determine co-linearity indices to discover correlated variables for mixed datasets for each response using the formulation proposed by Giannetti et al (2014) and as described in Sections 2.1 and 2.2. Specify the correlated variables (n_c) for each response.
Step2	Project scores on all variables and responses using the following equation: for score i and variable or response j $t_i^* = \lVert t_i \rVert \frac{L_j . t_i}{\lVert L_j \rVert \lVert t_i \rVert}$
Step3	**For quantitative variables** **If** variable *j* is correlated negatively **Then** Find the projected scores that lay in the plane of positive direction of variable *j* and negative direction of response. Find the original values corresponding to the scores in the last step $x^j_{optimal}$. Find the minimum and maximum values in $x^j_{optimal}$ and the resulting range [Min($x^j_{optimal}$), Max($x^j_{optimal}$)] that leads to low penalty values. **Else** Find the projected scores lay in the plane of positive direction of variable *j* and positive direction of response. Find the original values corresponding to scores in the last step x^j_{avoid}. Find the minimum and maximum values in x^j_{avoid} and the resulted range [Min(x^j_{avoid}) Max(x^j_{avoid})] that leads to high penalty values. **End**
Step 4	**For categorical variables** **If** variable j is negatively correlated with response **Then** Find the projected scores that lay in the plane of positive direction of variable j and negative direction of response. Find the original values corresponding to scores in the last step $x^j_{optimal}$. Find the percentage of occurrence of variable *j* on the resulted $x^j_{optimal}$ using: $Po^j = \frac{\sum_{i=1}^{n_x^j} x(optimal)_i^j}{n_x^j} \times 100\%$ If $Po^j > 60\%$ the occurrence of variable *j* lead to low penalty values.

Else

Find the projected scores that lay in the plane of positive direction of variable j and positive direction of response.

Find the original values corresponding to scores in the last step x^j_{avoid}.

Find the percentage of occurrence of variable *j* on the resulted x^j_{avoid} using:

$$Po^j = \frac{\sum_{i=1}^{n_x^j} x(avoid)_i^j}{n_x^j} \times 100\%$$

If Po^j > 60% the occurrence of variable *j* lead to high penalty values.

End

Table 4: The algorithm for discovering interactions among variables

	Interaction between two variables i,j
Step1	**For Quantitative variables**
	For any $q^i \in QVar^i$ **and any** $q^j \in QVar^j$
	If $Min(x^i_{optimal or avoid}) < q^i < Max(x^i_{optimal or avoid})$ and
	$Min(x^j_{optimal or avoid}) < q^j < Max(x^j_{optimal or avoid})$
	Then
	$Ivar^{i,j}=1$
	Else
	$Ivar^{i,j}=0$
	End
Step 2	**For categorical variables**
	If $CVar^i$ =1 and $CVar^j$ =1
	$Ivar^{i,j}=1$
	Else
	$Ivar^{i,j}=0$
	End
Step 3	Merge new variable $Ivar^{i,j}$ within the global matrix with data transformations as given in Table 2 and perform the covariance PCA using steps described in Section 2.2 (Giannetti et al 2014)
	Determine co-linearity indices
	If the correlation of $Ivar^{i,j}$ with response variable is stronger than the correlation of the original variables *i*,*j*.
	Keep the new interactions variable $Ivar^{i,j}$
	Else
	Ignore
	End

4. Results of a foundry case study for a nickel based alloy as described in Section 2.1

4.1. Co-linearity index and scores based bubble plots

The main aim of the analysis is to reduce the shrinkage defects in Nickel based alloy by analysing in-process data as described in Section 2.1. The co-linearity index method along with the corresponding data pre-treatment, as described in Section 2.2, is applied to understand the correlation of variables with shrinkage penalty vector. The co-linearity plot is depicted in Fig. 8. One variable, %Cobalt (%Co), showed positive correlation with high penalty vector and seven other variables showed negative correlation with high penalty vector.

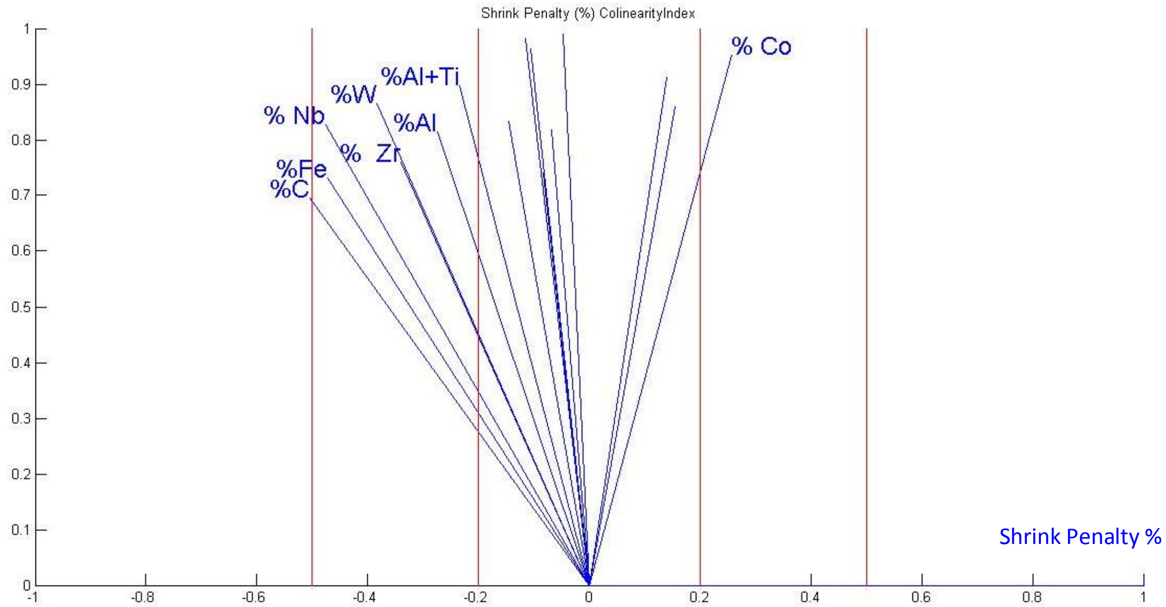


Fig. 8. Co-linearity index plot for the in-process data used in Nickel based alloy.

Projection of scores on loadings, calculated with the algorithm proposed in Section 3, is shown in Fig. 9. The optimal and avoid ranges are obtained by using Steps 3 and 4 (Table 3) for continuous and categorical data respectively. The rectangles in Fig. 9 show the data points chosen for deciding optimal or avoid limits. The data points are plotted in a lower dimensional space with the associated scores. As discussed in Steps 3 and 4 (Table 3), the minimum and maximum process variable value associated with these data points constitute optimal or avoid ranges. The optimal settings for all variables are shown in Fig. 10. The optimum range is plotted beside the whole range for each variable. The interaction co-linearity index plot and the interaction between variables are shown respectively in Fig. 11, Fig. 12 and Table 5. Even though the results are shown for a single response, the approach is applicable for multiple responses. When co-linearity indices are analysed for a chosen response, all other response variables are treated as factors and analysed together with remaining factors. The results can then be compiled to study the influence of each factor setting on all responses in order to make an informed decision on choosing the factor and its setting for a confirmation trial.

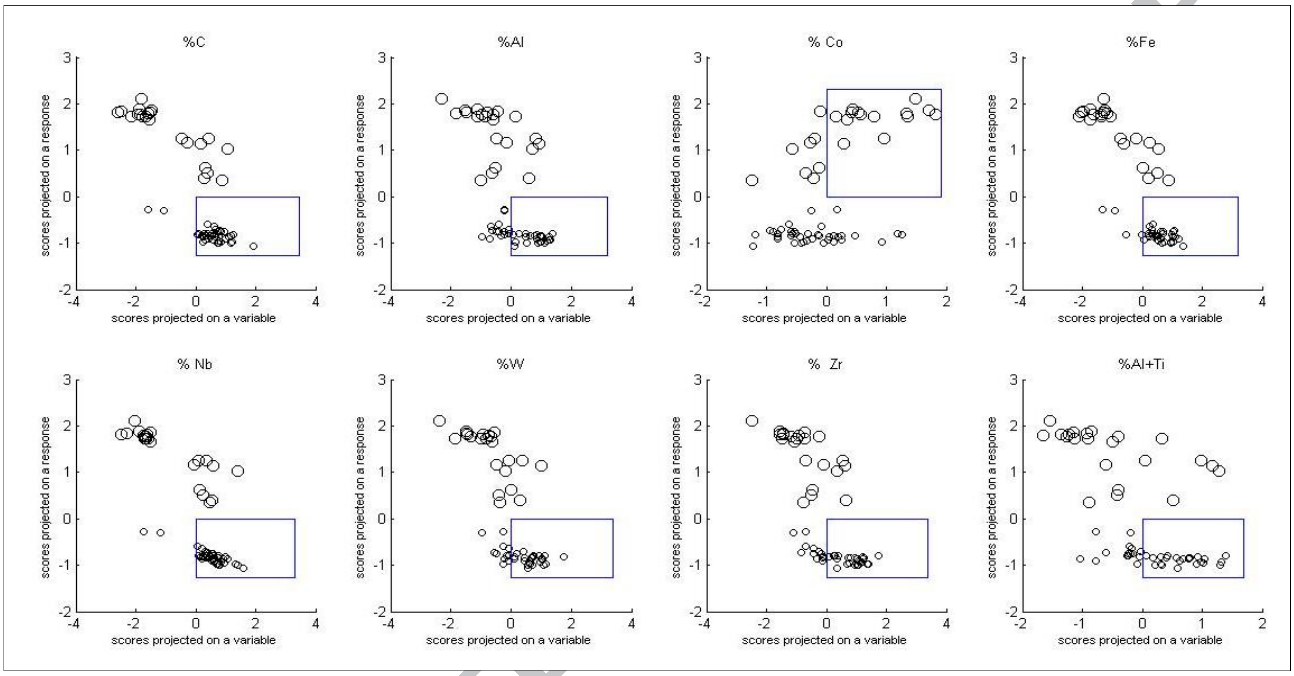


Fig. 9. Scores projection on variables and response of Nickel based alloy for 8 principal components.

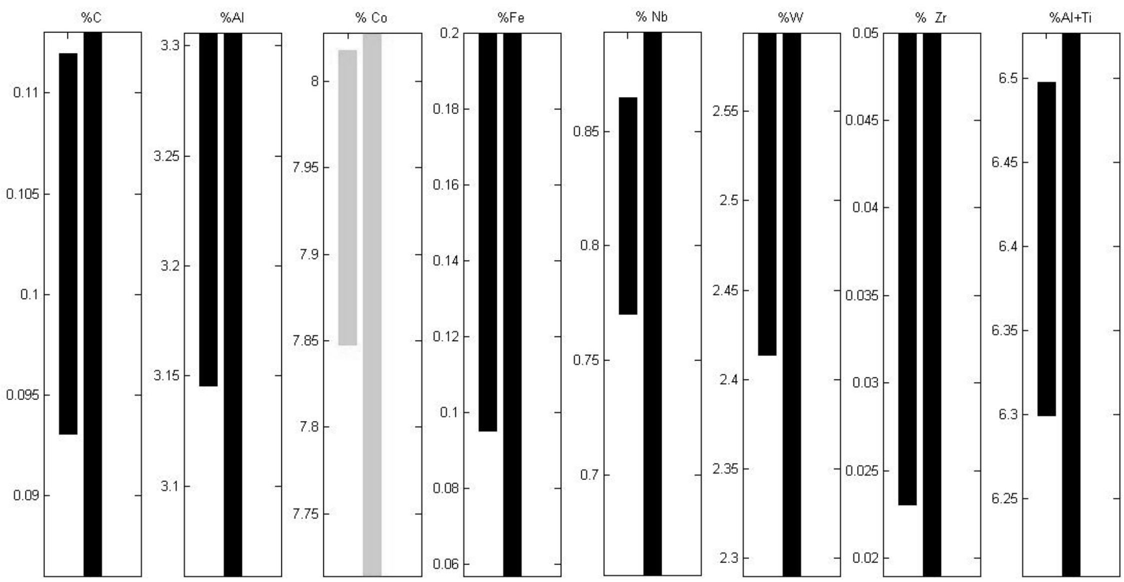


Fig. 10. The optimal range for variables, for each variable the left hand bar represent the optimal (black bars) or avoid (the light bar) range, obtained range for each variable corresponding to scores bounded by rectangle in **Fig.(9)** for the variable.

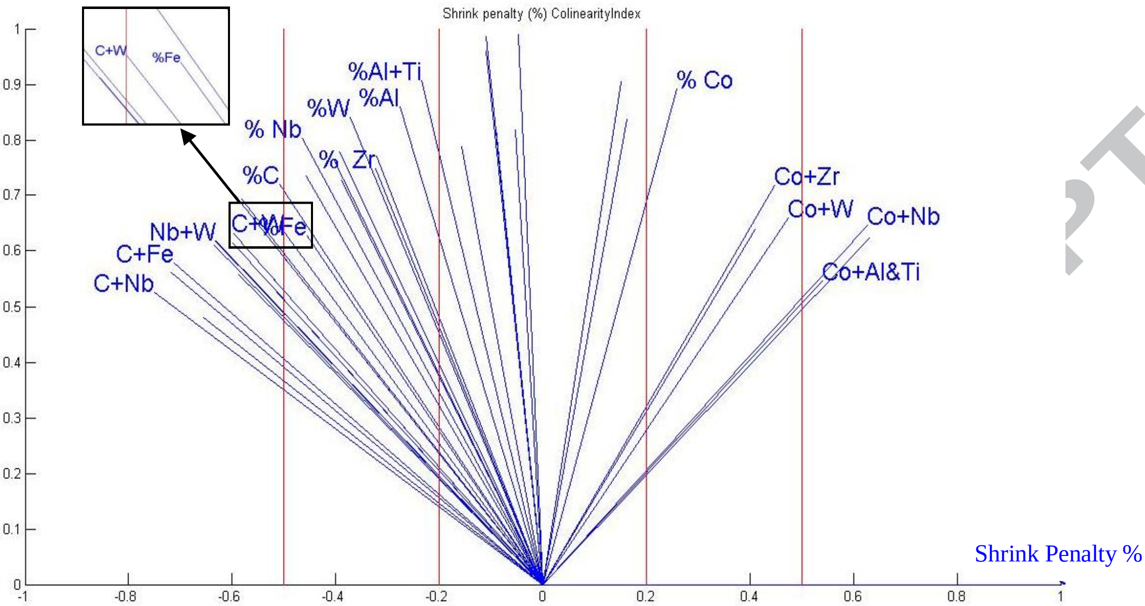


Fig. 11. The interaction co-linearity index for Nickel based alloy.

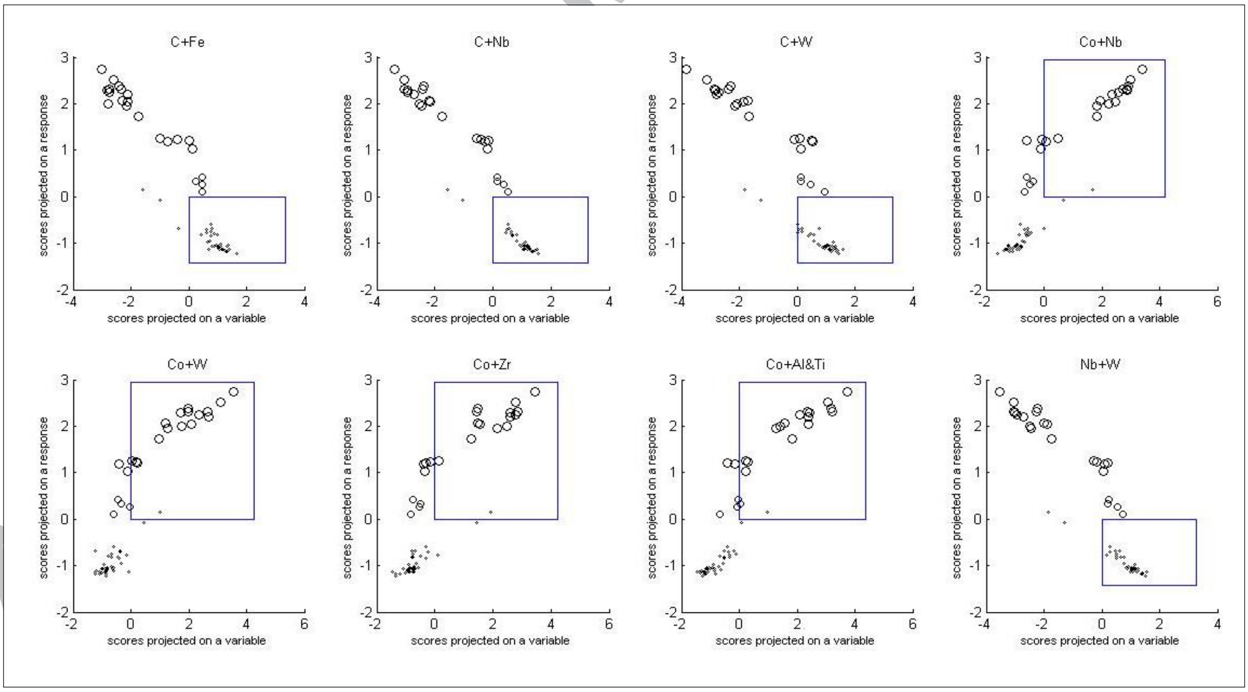


Fig.12. Scores projection on variables and response for 8 interactions resulted from using ranges in **Fig. (10)**.

Table (5) Interactions Table with percent of occurrence (PO%) for interacted variables

Variables	C	Co	Nb
Fe	100%	-	-
Nb	100%	75%	-
W	94.4%	35.3 %	94.4%
Zr	-	53.3%	-
Al&Ti	-	38.8%	-

4.2 Discussion of results and comparison with the quartile based classification by penalty matrices

The comparison of present results with results of the penalty matrix approach (Ransing et al. 2013) (see Table 6) shows that there is a slight difference. The most important reason for the deviation in results may be attributed to the range of variables used. In the penalty matrix method the optimal and avoid ranges are obtained from comparing the number of observations in each quartile, whereas in the present work the obtained range depends on the projection of scores on loadings to find out which scores contribute to each variable. The interactions table (Table 5) shows interactions, where most of the correlated variables tend to interact and produce stronger variables. The proposed algorithm's ability to account for mixed data types has allowed comparing the categorical interactions variable with the corresponding continuous variables on the same co-linearity index plots. In this case, the analysis of interactions gives additional information and leads to a better understating of the process. The resulted variable interactions provide additional hypotheses that need to be checked by process engineers whilst satisfying the requirements of clause 6.1 of the ISO9001:2015 standard. The standard further requires process engineers to develop strategies for managing the new knowledge discovered and continually upgrade its organisational knowledge library..

The interactions Carbon with Iron, Niobium and Tungsten and the interaction of Niobium and Tungsten (Fig. 11) are considered strong as they have high corresponding values of P_O% (Table 5). On the other hand, Fig. 11 also illustrates that Cobalt has a strong interaction with Tungsten, Zirconium and Al&Ti but the corresponding P_O% values are low as shown in Table 5. Hence, these interactions are not suggested by the proposed algorithm.

Table (6) Comparison of obtained ranges with penalty matrix approach (Ransing et al. 2013).

Variables	Penalty Matrix Range	Range predicted by the proposed algorithm
%C	0.095-0.113 (Optimal)	0.093-0.112 (Optimal)
%Al	3.24-3.306 (Optimal)	3.1453.306 (Optimal)
%Co	7.84-8.028 (Avoid)	7.847-8.018 (Avoid)
% Fe	0.114- 0.2 (Optimal)	0.095-0.2 (Optimal)
%Nb	0.77-0.827 (Optimal)	0.77-0.865 (Optimal)
%W	2.451-2.594 (Optimal)	2.413-2.594 (Optimal)
%Zr	0.026- 0.05 (Optimal)	0.023-0.05 (Optimal)
%Al+Ti	6.299–6.527 (Optimal)	6.299-6.498 (Optimal)

5. Conclusion

Many industries measure the quality of process outputs in terms of percentage of defective components in a production batch. A large number of process inputs across sub-processes influence the quality of this batch. A quality correlation algorithm is described that penalises undesired and unexplained quality variation among batches and discovers the optimal and avoid settings for various process inputs (factors). The approach takes into account the variance contribution of each process input and is applicable for mixed data sets. The median based data transformations and covariance PCA makes the approach applicable to situation where in-process data associated with factors may not always follow Guassian distributions. The co-linearity concept is constrained by the linearity assumptions of the PCA however, the discovery of optimal and avoid settings and the new definition of interactions allow discovery of quality correlations even if the factor response relationship is non-linear. The algorithm is generic and is applicable to discover complex correlations and interactions among continuous and categorical variables using ISO9001:2015's risk based thinking strategy. However, the proposed algorithm is unable to work missing data. Future research is necessary to address this technological gap.

In the proposed work, the co-linearity index procedure presented by Giannetti et al. (2014) is used to predict the correlations between variables and responses for in process data for casting process. A new algorithm has been proposed to predict the optimal process settings for correlated variables. The algorithm uses the analogy between loading and scores of principal component analysis in bi-plots but extends the concept for *p* number of principal components. The data pre-treatment of response variables has been shown to embed the risk based thinking as proposed by the ISO9001:2015 quality standard. The proposed approach found optimal or avoid ranges from the observation data. It separated the corresponding scores into different clusters according to its contribution on loading vectors of variables in the p-dimensional space. Another advantage of the proposed algorithm is that it has allowed the discovery of interactions between variables by creating new categorical variables and comparing its co-linearity index on the original co-linearity index plot. Additional constraints to further qualify interactions have also been proposed. The results are verified with previously published literature (Ransing et al. 2013).

References

Abdi H, Williams L. (2010) Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2(4):433-459.

Bécue-Bertaut M, Pagès J. (2008) Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Computational Statistics & Data Analysis 52(6):3255-3268.

Begam F, Kumar S. (2014) Visualization of Chemical Space Using Principal Component Analysis. World Applied Sciences Journal 29 (Data Mining and Soft Computing Techniques):53-59.

Cattell R.B. (1996) The scree test for the number of factors, Multivariate Behaviour reserach 1: 245–276.

Ceglarek, D., Huang, W., Zhou, S., Ding, Y., Kumar, R., & Zhou, Y. (2004). Time-Based Competition in Multistage Manufacturing: Stream-of-Variation Analysis (SOVA) Methodology—Review. International Journal Of Flexible Manufacturing Systems, 16(1), 11-44.

Chen L, Carpita N, Reiter W, Wilson R, Jeffries C, McCann M. (1998) A rapid method to screen for cell-wall mutants using discriminant analysis of Fourier transform infrared spectra, The Plant Journal16(3):385-392.

Colledani, M., Tolio, T., Fischer, A., Iung, B., Lanza, G., Schmitt, R., & Váncza, J. (2014). Design and management of manufacturing systems for production quality. CIRP

Annals - Manufacturing Technology, 63(2), 773-796.

Escofier, B. & Pagès , J. (1994) Multiple factor analysis (AFMULT package). Computational Statistics & Data Analysis 18(1): 121 – 140

Gabriel, K. R. (1971) The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. Biometrika 58(3): 453-467.

Gabriel, K.R. and Odoroff, C.L. (1990) Biplots in biomedical research. Statist. Med. 9(5): 469–485.

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, M., Ransing, R., Bould, D., Gethin, D., & Sienz, J. (2015) Organisational Knowledge Management for Defect Reduction and Sustainable Development in Foundries. International Journal Of Knowledge And System Science (IJKSS), 6(3): 20 pages.

Iso.org (2010) ISO 9001 certifications top one million mark, food safety and information security continue meteoric increase. http://www.iso.org/iso/news.htm?refid=Ref1363. Accessed 27 February 2015.

Iso.org (2014) ISO 9001 revision moves on to final stage. http://www.iso.org/iso/home/news_index/news_archive/news.htm?refid=Ref1905. Accessed 26 February 2015.

Jolliffe I. (2002) Principal Component Analysis. New York, Springer.

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

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

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

Lewis, R. W., Ransing, R. S., Pao, W. K. S., Kulasegaram, K., & Bonet, J. (2004). Alternative techniques for casting process simulation. International Journal of Numerical

Methods for Heat and Fluid Flow, 14(2), 145–166.

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

Parente A, Sutherland J. (2013) Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity. Combustion and Flame 160(2):340-350.

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

Prakash, Eduardo Izquierdo, L., & Ceglarek, D. (2009). Functional process adjustments to reduce No-Fault-Found product failures in service caused by in-tolerance faults. CIRP Annals - Manufacturing Technology, 58(1), 37-40.

Ransing, R. S., & Lewis, R. W. (1997). A semantically constrained neural network for manufacturing diagnosis. International Journal of Production Research, 35(9), 2639–2660.

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).

Ransing, RS and Ransing M.R. (2015) http://www.slideshare.net/7Epsilon/key-changes-to-iso-9001-2015-and-7-steps-of-7-epsilon. Accessed 18 November 2015.

Stricker, N., & Lanza, G. (2014). The Concept of Robustness in Production Systems and its Correlation to Disturbances. Procedia CIRP, 19, 87-92.

Van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K. & Van der Werf1, M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7(142)

Varmuza K and Filzmoser P. (2009) Introduction To Multivariate Statistical Analysis In Chemometrics. Boca Raton, CRC Press.

% Shrinkage Defect Per Melt (or Heat)	Carbon	Aluminium	Boron	Cobalt	Chromium	Iron	Molybdenum
0.12	0.101	3.23	0.009	7.857	15.2	0.086	1.663
0	0.093	3.145	0.009	7.971	15.295	0.086	1.644
0.15	0.107	3.249	0.009	7.781	15.248	0.152	1.691
0	0.103	3.249	0.008	8.028	15.096	0.105	1.653
0	0.105	3.183	0.008	7.781	15.001	0.124	1.682
0	0.107	3.107	0.008	7.8	15.295	0.19	1.663
0	0.109	3.145	0.01	7.866	15.267	0.095	1.691
0	0.112	3.287	0.009	7.743	15.305	0.19	1.663
0.02	0.106	3.145	0.009	7.838	15.352	0.095	1.644
0	0.106	3.249	0.008	7.809	15.276	0.095	1.634
0	0.108	3.097	0.008	7.781	15.286	0.095	1.653
0	0.108	3.183	0.008	7.828	15.286	0.095	1.634
0	0.106	3.24	0.008	7.857	15.02	0.143	1.663
0	0.108	3.268	0.009	7.895	15.267	0.171	1.672
0	0.102	3.306	0.008	7.885	15.248	0.114	1.634
0.07	0.102	3.306	0.009	7.942	15.229	0.067	1.663

Niobium	Tungsten	Tantalum	Titanium	Zirconium	Aluminium	Nitrogen	Oxygen	Ta/Ti
0.846	2.556	1.587	3.23	0.037	6.46	33.25	6.65	0.492
0.798	2.594	1.558	3.211	0.05	6.365	11.4	6.65	0.486
0.893	2.423	1.653	3.278	0.031	6.527	38	10.45	0.505
0.865	2.489	1.568	3.211	0.035	6.46	22.8	7.6	0.489
0.808	2.423	1.52	3.107	0.032	6.289	21.85	5.7	0.49
0.808	2.442	1.615	3.145	0.022	6.251	20.9	8.55	0.514
0.77	2.451	1.653	3.192	0.024	6.337	20.9	9.5	0.518
0.817	2.461	1.672	3.211	0.023	6.498	26.6	5.7	0.521
0.808	2.461	1.596	3.164	0.023	6.308	38.95	11.4	0.505
0.817	2.48	1.558	3.173	0.024	6.422	30.4	3.8	0.492
0.836	2.432	1.625	3.202	0.023	6.299	29.45	7.6	0.508
0.798	2.48	1.577	3.145	0.026	6.327	20.9	7.6	0.502
0.865	2.518	1.615	3.173	0.03	6.413	32.3	5.7	0.509
0.865	2.47	1.568	3.211	0.03	6.489	32.3	7.6	0.489
0.817	2.489	1.492	3.145	0.027	6.451	36.1	11.4	0.475
0.865	2.47	1.549	3.183	0.031	6.489	20.9	7.6	0.487

Highlights

- An approach to embed ISO 9001:2015's risk based thinking for in-process quality improvement is proposed.
- The algorithm determines optimal and avoid ranges within the process variation including process interactions.
- It is shown how the technique can be used to satisfy the requirements of Clause 6.1 of the ISO9001:2015 standard.