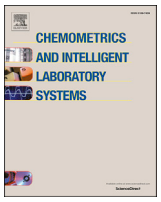
Chemometrics and Intelligent Laboratory Systems 225 (2022) 104563

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



Defining multivariate raw material specifications in industry 4.0

Joan Borràs-Ferrís [a,*,1], Daniel Palací-López [a,1], Carl Duchesne [b], Alberto Ferrer [a]

[a] Multivariate Statistical Engineering Group, Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València, València, Spain
[b] Chemical Engineering Department, Laval University, Quebec, Canada

ARTICLE INFO

Keywords:
Industry 4.0
Design space
Model inversion
Partial least squares
Prediction uncertainty
Raw material multivariate specifications

ABSTRACT

A novel methodology is proposed for defining multivariate raw material specifications providing assurance of quality with a certain confidence level for the critical to quality attributes (CQA) of the manufactured product. The capability of the raw material batches of producing final product with CQAs within specifications is estimated before producing a single unit of the product, and, therefore, can be used as a decision making tool to accept or reject any new supplier raw material batch. The method is based on Partial Least Squares (PLS) model inversion taking into account the prediction uncertainty and can be used with historical/happenstance data, typical in Industry 4.0. The methodology is illustrated using data from three real industrial processes.

1. Introduction

Raw materials properties are usually considered as Critical Input Parameters (CIPs) because their variability has an impact on Critical Quality Attributes (CQAs) of the final product. Thus, as commented by Duchesne and MacGregor [1], the development of specification regions for raw materials is crucial to ensure the desired quality of the product. In this paper, we propose a novel method to define a multivariate raw material specification region that is expected to provide assurance of quality with a certain confidence level for the CQAs. Our approach overcomes the drawbacks of the current industrial practice of setting univariate specifications for each property of raw material and allows the producer to make a decision on accepting or rejecting a raw material batch based on the confidence of producing good product quality prior to starting the manufacturing process.

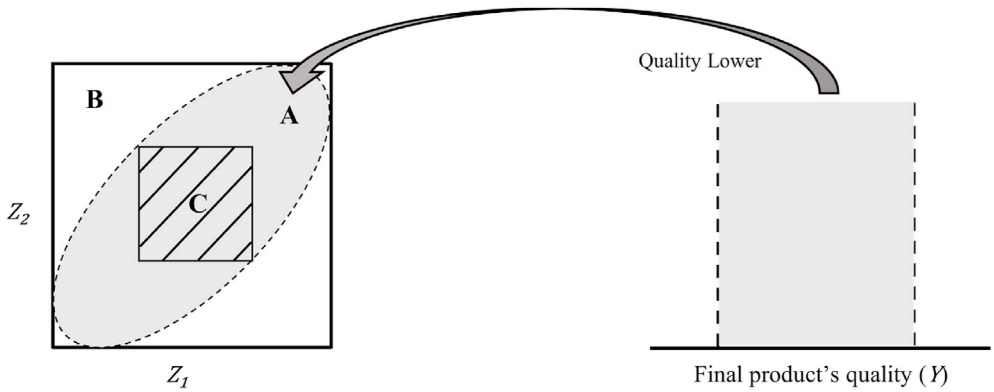
Despite their importance, specifications are usually defined in an arbitrary way based mostly on subjective past experience, instead of using a quantitative objective description of their impact on CQAs. Furthermore, in many cases, univariate specifications on each property are designated, with the implicit assumption that these properties are independent from one another. As a consequence, significant amounts of raw materials whose properties are correlated may be misclassified, as appropriate or otherwise, when univariate specifications are considered, as it is shown in Fig. 1.

Let us consider a raw material with two correlated properties, Z_1 and Z_2 (see Fig. 1) used in the manufacturing of a particular product with final product quality Y. The elliptical region "A" is the true multivariate region in Z_1 and Z_2 such that any batch of raw material used with Z_1 and Z_2 properties falling within it will provide good product quality (i.e., within the Y quality specification limits). On the contrary, raw material batches with properties outside this elliptical region correspond to unacceptable raw material batches, as they lead to poor product quality (i.e., outside the Y quality specification limits). The square region "B" corresponds to the univariate specification region when accepting the same variance on each individual property as the multivariate region. In this case, accepting raw material batches with properties outside region "A" and inside region "B" leads to manufacturing products with final product quality Y outside its specification limits. To avoid this, companies are forced to shrink the univariate specifications from region "B" to the region "C", at the cost of rejecting acceptable raw material batches (i.e., those outside region "C" but inside region "A"). Another consequence of setting these more restrictive univariate specifications is an increase in costs in the acquisition of raw material batches with tighter variations in their properties.

Multivariate specifications provide, therefore, much insight into what constitutes acceptable raw material batches when their properties are correlated (as usually happens). In order to cope with this correlation several authors suggest using multivariate approaches, such as Partial Least Squares (PLS) regression, to improve the definition of raw materials specifications.

* Corresponding author. Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València, Camino De Vera s/n, 7A, 46022, València, Spain.
E-mail address: joaborfe@eio.upv.es (J. Borràs-Ferrís).
[1] These authors have equal contributions.

https://doi.org/10.1016/j.chemolab.2022.104563
Received 2 November 2021; Received in revised form 19 February 2022; Accepted 11 April 2022
Available online 15 April 2022



A. True multivariate region.

B. Univariate specification region containing the true multivariate region.

C. Univariate specification region containing only acceptable raw material.

Fig. 1. *Problem of using univariate specifications on correlated raw material properties (Z_1 and Z_2).*

The first systematic study was reported by De Smet [2], where PLS regression is used first to build a model between raw materials properties and CQAs by using historical data. Then, a boundary in the model subspace is defined within which most of the values for the raw materials properties associated with good CQAs can be found. This multivariate region (in the latent space) can then be used to accept or reject new batches of raw materials. The key assumption of this method is that variability in the CQAs results exclusively from variations in the raw materials properties of a single material. Duchesne and MacGregor [1] generalized this method by assuming that both variation in raw materials properties and in process operating conditions are responsible for CQAs variations. Uncontrolled variability in the operating conditions will increase the variability of the CQAs and require tightening specifications on the raw material properties to make up for it. On the other hand, properly tuned feedback and feedforward controllers may compensate for CQAs variations allowing for wider raw material properties specifications [3]. Later on, García-Muñoz [4] extended the Duchesne-MacGregor method to combine data from multiple scales (e.g. lab or pilot scale and commercial scale) with different processing conditions and control strategies.

These approaches, however, focused on defining multivariate specification regions on the multiple properties of a single raw material. To overcome this limitation, MacGregor et al. [5] extended them to determine the acceptability of new raw materials from multiple suppliers and with multiple measured properties, as well as to assess the suitability of combining specific batches of raw materials currently in inventory to minimize the risk of manufacturing a poor quality product. Finally, Azari et al. [6] proposed a sequential multiblock PLS algorithm to better sort the contribution of raw materials and process operating conditions on CQA variations, considering two types of raw materials in this study.

In the aforementioned references, the aim was to determine the boundary in the latent space of the historical data that best separates acceptable from unacceptable raw materials by direct mapping (i.e., those leading to good and poor CQAs, respectively). Nonetheless, the general shape (e.g., an ellipsoid or a straight line) and locus of such boundary was decided based on subjective criteria, trying to best balance out the type I and type II risks.[2] In contrast to this, García-Muñoz, Dolph, and Ward [3] emphasized the use of mathematical and statistical models as an objective way to define such specifications by linking them with specification limits for CQAs. Thus, given a desired set of CQAs, and in order to predict an appropriate set of raw materials properties, it is necessary to carry out the inversion of the model relating inputs (raw materials properties) with outputs (CQAs). Recently, Paris, Duchesne and Poulin [7] carried out a comparison between direct mapping and model inversion stating their advantages and drawbacks.

However, when inverting PLS models, their prediction uncertainty is also back-propagated [8,9]. This issue has not been addressed in the past when defining multivariate raw materials specifications and, thereby, all the methods commented above are considered as descriptive approaches focused on historical data, lacking a probabilistic interpretation. For that, uncertainty is accounted in the form of prediction intervals, with a certain confidence level, finding a window within which any batch with raw material properties is expected to produce product with CQAs within specification limits with at least the predefined confidence level. In this regard, this window refers to the estimation of the so-called Design Space (DS), which is defined as the multidimensional combination and interaction of inputs variables (e.g., raw material properties) and process conditions that have been demonstrated to provide assurance of quality [10]. A preliminary approach to frame the DS by prediction intervals was used by Whitcomb and Anderson [11], but in the original space of inputs variables and using data from a Design of Experiments (DOE).

Although not explicitly applied to the definition of multivariate specifications, Bayesian approaches [12–14] can be used to include the model-parameter uncertainty and estimate the probability map of meeting the specifications imposed on the CQAs being used to identify the DS [15]. However, these methodologies define the DS by means of a predictive approach instead of carrying out the model inversion. Therefore, the representation of the DS a priori requires the discretization of the multidimensional input domain by sampling algorithms. Then, simulation methods, such as Markov-Chain Monte Carlo techniques, are required for each discretization point to determine if it is within the DS. Hence, these approaches do not represent analytically the DS in the input domain, with the additional drawback of being computationally costly.

The novelty of this paper is the implementation of the frequentist probabilistic interpretation in the definition of the multivariate specification region for raw materials in the latent space. For that, we propose a method to define analytically a window in the latent space of the raw material properties that is expected to provide assurance of quality for the CQAs with at least a certain confidence level.

[2] Type I risk is defined as the proportion of truly acceptable batches of raw materials that is rejected by the customer under a given specification region; type II risk consists of the proportion of truly unacceptable batches of raw materials that is accepted by the customer under a given specification region [1].

The paper is organized as follows. Data requirements for defining multivariate specification are first discussed, followed by a description of PLS model regression and the analytical definition of its inversion. How PLS inversion addresses the definition of multivariate specifications by considering a probabilistic approach is then presented. Finally, the methodology is illustrated by means of three industrial case studies.

2. Data requirements

The data required for developing raw materials multivariate specifications following the methodology proposed in this paper involves two blocks, **Z** and **Y**. **Z** ($N \times M$) is a matrix of inputs which includes a total of M measurements characterizing the properties of each of the N batches of a particular raw material, and the **Y** ($N \times L$) output matrix consists of L measurements of the CQAs of the final product obtained for each one of the N corresponding batches.

Furthermore, process variations may be under tight control to attenuate some raw material variations, whenever the eventual effect of such variability on the CQAs can be compensated by control systems. Specifications for incoming raw materials are nonetheless required, however, to account for variations in raw materials whose effect on the CQAs cannot be compensated by control systems. Therefore, if this situation prevails in the future there is no need to consider process data to establish the specification regions associated to the latter source of variation.

3. Latent variable regression model inversion

3.1. PLS regression model

PLS regression [16,17] is a latent variable-based approach used not only to model the inner relationships between the matrix of inputs **Z** and the matrix of output variables **Y**, but also to provide a model for both. This fact gives them a very nice property: uniqueness and causality in the reduced latent space no matter if the data come either from a DOE or daily production process (historical/happenstance data) typical in Industry 4.0 [18,19]. The PLS regression model structure can be expressed as follows:

$$\mathbf{T} = \mathbf{Z} \cdot \mathbf{W}^{*} \tag{1}$$

$$\mathbf{Z} = \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{2}$$

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{3}$$

where the columns of the matrix **T** ($N \times A$) are the PLS scores vectors, consisting of the first A latent variables (LVs) from PLS. These score vectors explain most of the covariance between **Z** and **Y**, and each one of them ($\boldsymbol{t}_a$, $a = 1, 2, \ldots, A$) is estimated as a linear combination of the original variables with the corresponding "weight" vector ($\boldsymbol{w}_a$, a = 1, 2, …, A) (Eq. (1)). These weights vectors are the columns of the weighting matrix $\mathbf{W}^{*}$ ($M \times A$).

The PLS scores vectors are also good "summaries" of **Z** according to the **Z**-loadings ($\mathbf{P}(M \times A)$) (Eq. (2)) and good predictors of **Y** according to **Y**-loadings ($\mathbf{Q}(L \times A)$) (Eq. (3)), where $\mathbf{E}(N \times M)$ and $\mathbf{F}(N \times L)$ are residual matrices. The sum of squares of $\boldsymbol{F}$ is an indicator of how good the model is in predicting the **Y**-space, and the sum of squares of $\boldsymbol{E}$ is an indicator of how well the model explains the **Z**-space.

In order to evaluate the model performance when projecting the n-th observation z_n onto it, the Hotelling T^2 in the latent space T_n^2 and the Squared Prediction Error SPE_{z_n} are calculated [20]:

$$\boldsymbol{\tau}_n = \mathbf{W}^{*\mathrm{T}} \cdot \boldsymbol{z}_n \tag{4}$$

$$T_n^2 = \boldsymbol{\tau}_n^{\mathrm{T}} \cdot \boldsymbol{\Lambda}^{-1} \cdot \boldsymbol{\tau}_n \tag{5}$$

$$SPE_{z_n} = (\boldsymbol{z}_n - \mathbf{P} \cdot \boldsymbol{\tau}_n)^T \cdot (\boldsymbol{z}_n - \mathbf{P} \cdot \boldsymbol{\tau}_n) = \boldsymbol{e}_n^{\mathrm{T}} \cdot \boldsymbol{e}_n \tag{6}$$

where z_n refers to the n-th row extracted from **Z** being defined as a column vector, $\boldsymbol{e}_n$ is the residual column vector associated to the n-th observation, $\boldsymbol{\Lambda}^{-1}$ is defined as the ($A \times A$) diagonal matrix containing the inverse of the A variances of the scores associated with the LVs, and $\boldsymbol{\tau}_n$ is the column vector of scores corresponding to the projection of the n-th observation z_n onto the latent subspace of the PLS model.

The Hotelling T^2 statistic of an observation (T_n^2) is the estimated squared Mahalanobis distance from the center of the latent subspace to the projection of such observation onto this subspace. The SPE statistic gives a measure of how close (in an Euclidean way) the n-th observation (z_n) is from the A-dimensional latent space. Upper confidence limits (with a specified confidence level) for both statistics, $SPE_{\lim}$ and $T_{\lim}^2$, can be calculated for Phase I (model building) and Phase II (model exploiting) based on theoretical distributions [21,22]. The normality assumption on which these calculations are based is usually quite reasonable in practice. Alternatively, these confidence limits can be obtained from distribution free methods by repeated sampling [23]. The only requirement is to have a large reference dataset. Besides, if this large dataset is available (as with historical/happenstance data), confidence limits for Phase II can also be used in Phase I. In the following sections, SPE and T^2 99% confidence limits are calculated from theoretical distributions.

Once the PLS regression model has been fitted, it can be used directly in order to obtain the prediction vector corresponding to a particular observation, z^{obs}, fulfilling that $T_{obs}^2 \leq T_{lim}^2$ and $SPE_{z^{obs}} \leq SPE_{\lim}$ for Phase II, as

$$\hat{\mathbf{y}}^{obs} = \mathbf{Q} \cdot \boldsymbol{\tau}^{obs} = \mathbf{Q} \cdot \mathbf{W}^{*\mathrm{T}} \cdot z^{obs} \tag{7}$$

However, predictions are not free from uncertainty, yielding prediction errors. Three different sources of uncertainties can affect the prediction error e_l^{obs} of the l-th response variable $\hat{y}_l^{obs}$ given a new observation z^{obs} [24]: (i) measurement uncertainty in both the regressor matrix (**Z**) and the response matrix (**Y**) used to calibrate the PLS model, (ii) uncertainty in the estimated model regression parameters, (iii) and uncertainty due to the unmodeled part of the response variable (structural model uncertainty).

Estimation of prediction uncertainty is done by using Ordinary Least Squares (OLS) as Faber and Kowalski [25] suggested. Although this approach is an approximation, it was observed to yield good results in practice [26]. First, it is assumed that the prediction error e_l^{obs} follows a normal distribution with zero mean and variance $\sigma_{e_l^{obs}}^2$ (Eq. (8)).

$$e_l^{obs} = y_l^{obs} - \hat{y}_l^{obs} \sim N\left(0,\ \sigma_{e_l^{obs}}^2\right) \tag{8}$$

Therefore $e_l^{obs}/s_{e_l^{obs}}$ follows a t-statistic with $N - df$ degrees of freedom and, consequently, the $(1-\alpha)$ prediction interval ($PI_{y_l^{obs}}$) on y_l^{obs} is calculated as:

$$PI_{y_l^{obs}} = \hat{y}_l^{obs} \pm t_{N-df,\alpha/2} \cdot s_{e_l^{obs}} \tag{9}$$

where N is the number of the PLS model calibration samples, df the degrees of freedom consumed by the model (it is set equal to the number of LVs of the model[3]), α the false alarm rate for the prediction interval (i.e. (1- α) × 100 confidence level) and $s_{e_l^{obs}}$ the estimated standard deviation of the prediction error. The latter is calculated using Eq. (10) when taking into account the second and third sources of uncertainty

[3] Although the derivation of the DF for PLS is not straightforward, they are expected to be low in comparison with the number of observations when dealing with historical data, $N - df$ tends to N, thus having a negligible effect on estimating the prediction uncertainty.

mentioned above. Note that to estimate the first source of uncertainty requires explicit knowledge about error variance in $\mathbf{Z}$ and $\mathbf{y}$, that is estimated from replications and thus this limits its use in practice. However, it seems to be more practical to assume that second and third sources of uncertainties dominate and to ignore the first one [26].

$$s_{e_l^{obs}} = SE_l \cdot \sqrt{1 + h^{obs} + \frac{1}{N}} \tag{10}$$

In the above expression, h^{obs} is the leverage of the observation (Eq. (11)) and SE_l the standard error of calibration (Eq. (12)).

$$h^{obs} = \boldsymbol{\tau}^{obs^T} \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}^{obs} \tag{11}$$

$$SE_l = \sqrt{\frac{\sum_{n=1}^{N} \left(y_{n,l} - \hat{y}_{n,l}\right)^2}{N - df}} \tag{12}$$

where $y_{n,l}$ and $\hat{y}_{n,l}$ are, respectively, the measured and estimated values of the l-th response variable for the n-th observation in the calibration dataset.

3.2. PLS model regression inversion and null space

The objective of model inversion is to find (predict) a window of inputs (raw materials properties, process conditions, etc.) for a desired product quality. Jaeckle and MacGregor [27] proposed a framework for the inversion of PLS models using historical data available on the process operating conditions and on the corresponding product quality. Using standard regression or machine learning models, the inversion is inadequate because those models do not contain any information about the covariance structure and, consequently, the inversion solution of the model almost certainly does not respect previous structural relationships, leading to unfeasible solutions. By contrast, when inverting a PLS model the inversion solution belongs to the latent space (defined by the latent variables) and, therefore, such solution is constrained to be physically feasible and consistent with the sets of process conditions and correlation structure from the past. In this respect, the PLS model inversion has been demonstrated to be a valid tool to support the development of new products and their manufacturing conditions using historical data in several case studies [28–33].

When considering the inversion of a PLS model (Eq. (1) and Eq. (3)), the set of raw materials properties (column vector $\boldsymbol{z}^{new}$) that will yield the desired set of CQAs (column vector $\boldsymbol{y}^{des}$) are obtained by solving the following system of linear equations:

$$\boldsymbol{y}^{des} = \mathbf{Q} \cdot \boldsymbol{\tau}^{new} \tag{13}$$

where $\boldsymbol{\tau}^{new}$ is the vector of scores corresponding to the projection of the observation $\boldsymbol{z}^{new}$, which is estimated by the inversion of the PLS model:

$$\boldsymbol{\tau}^{new} = f^{-1}\left(\boldsymbol{y}^{des}\right) \tag{14}$$

Then, $\boldsymbol{z}^{new}$ is estimated going back from the latent space to the raw materials properties space as follows:

$$\boldsymbol{z}^{new} = \mathbf{P} \cdot \boldsymbol{\tau}^{new} \tag{15}$$

Eq. (15) clearly shows that the solution $\boldsymbol{z}^{new}$, obtained by the PLS model inversion, is a linear combination of the loading vectors $\boldsymbol{p}_a$ (columns of $\mathbf{P}$) and thus belongs to the latent space. Besides, notice that the PLS model inversion involves solving a system of linear equations represented in a matrix form (Eq. (13)), where there are as many linear independent equations as the rank of $\mathbf{Y}$ (r_Y), and the number of unknown variables corresponds to the dimensionality of the latent space (A). Thus, three possible cases are considered based on dimensions r_Y and A:

- $r_Y > A$: the most likely case is that no solution provides the desired set of CQAs, but the least squares solution can be obtained as follows:

$$\boldsymbol{\tau}^{new} = (\mathbf{Q}^{\mathrm{T}} \cdot \mathbf{Q})^{-1} \cdot \mathbf{Q}^{\mathrm{T}} \cdot \boldsymbol{y}^{des}$$

- $r_Y = A$: a single solution exists that provides the desired set of CQAs.

$$\boldsymbol{\tau}^{new} = \mathbf{Q}^{-1} \cdot \boldsymbol{y}^{des}$$

- $r_Y < A$: it corresponds to an underdetermined system of linear equations, and has multiple solutions forming a vector space whose dimension is the difference between A and r_Y. Hence, multiple solutions $\boldsymbol{\tau}^{new}$ fall into a $(A - r_Y)$- dimensional subspace of the A-dimensional space, that theoretically yields the same desired set of CQAs. This subspace is so-called Null Space (NS) and, in such a case, the model inversion requires defining such a space.

The latter situation ($r_Y < A$) corresponds to the most common case and, for that reason, it has been widely studied. Jaeckle and MacGregor [28] defined the hyper-plane related to the NS by both the solution given by the pseudo-inverse with minimal Euclidean norm as a point which belongs to the NS (Eq. (16)), and the orthogonal directions referring to null variations in CQAs ($A - 1$ linearly independent vectors parallel to the NS).

$$\boldsymbol{\tau}^{new} = \mathbf{Q}^{\mathrm{T}} \cdot (\mathbf{Q} \cdot \mathbf{Q}^{\mathrm{T}})^{-1} \cdot \boldsymbol{y}^{des} \tag{16}$$

García-Muñoz et al. (2006) extended this approach proposing a linear equation system where each equation defines the NS for each CQA as proposed by Jaeckle and MacGregor [28] (i.e., by both a point and orthogonal directions of null variations).

On the other hand, Palací-López et al. [9] defined the NS for each l-th CQA by the analytical equation of a $(A - 1)$-dimensional hyper-plane, which spans the multiple inversion solutions for such l-th CQA. The general form of a hyperplane only requires a constant (v_{0l}) and a single orthogonal vector to the NS ($\mathbf{v}_l$). This vector corresponds to the direction of maximum variation of the l-th CQA. The intersection of all these NS (if they exist) gives the same solution as the one proposed by Jaeckle and MacGregor [28].

In this work, it is assumed that all variables are centered and scaled to unit variance as a pre-treatment. Thus, the l-th NS is defined as follows:

$$\begin{aligned} &v_{0l} + \mathbf{v}_l^{\mathrm{T}} \cdot \boldsymbol{\tau}^{NS,l} = 0 \\ &v_{0l} = -y_l^{des} \\ &\mathbf{v}_l = \boldsymbol{q}_l \end{aligned} \tag{17}$$

where $\boldsymbol{q}_l$ is the l-th row of $\boldsymbol{Q}$. When applied to all L CQAs:

$$\begin{aligned} &\boldsymbol{v}_0 + \mathbf{V} \cdot \boldsymbol{\tau}^{NS} = 0 \\ &\boldsymbol{v}_0 = \begin{bmatrix} v_{01} \\ v_{02} \\ \vdots \\ v_{0L} \end{bmatrix} = -\boldsymbol{y}^{des} \; ; \; \mathbf{V} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_L^{\mathrm{T}} \end{bmatrix} = \mathbf{Q} \end{aligned} \tag{18}$$

Indeed, Eq. (18) is equivalent to Eq. (13) but expressed as the intersection of the L NSs (if it exists). To put it briefly, Fig. 2 shows the PLS model inversion by means of a simple example. In this example, there are three raw material properties ($M = 3$) and the focus is on the l-th CQA, and a PLS model has been previously fitted using two components ($A = 2$). Then, given a desired l-th CQA, multiple solutions are predicted, which will theoretically result in such l-th CQA. These solutions belong to the one-dimensional NS.

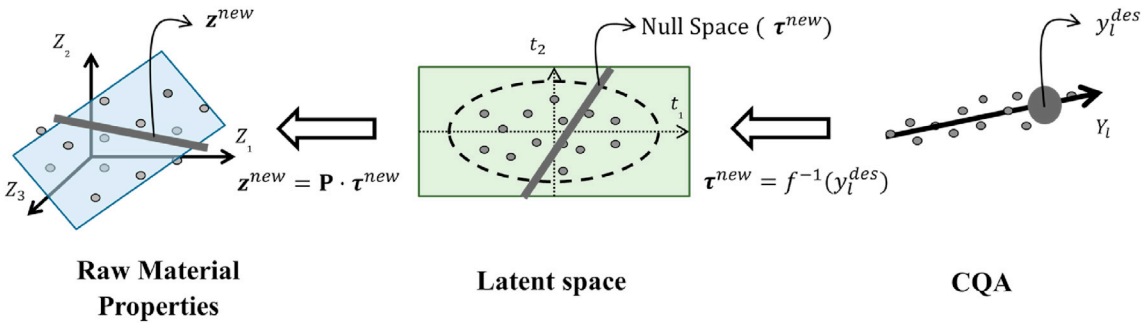


Fig. 2. Simple example of the model inversion where there are three raw materials properties and the focus is on the l-th CQA, and a PLS model has been fitted by two components.

4. Defining multivariate specifications in the latent space

As commented above, establishing the multivariate raw material specification regions in the latent space is equivalent to defining the multidimensional combination and interaction of raw materials properties that have been demonstrated to provide assurance of quality (i.e., the DS of raw materials). Hence, both terms (multivariate specifications region and DS) are used interchangeably in the remainder of the paper.

4.1. Design space with no uncertainty

If there is not prediction uncertainty, the DS must be defined as a region in the latent space associated with raw materials properties such that these properties yield an expected value of CQAs, according to Eq. (7), within their specification limits.

Besides, since PLS is an empirical model based on historical data, any new set of raw materials properties must respect the *correlation structure* and *range* of this historical data [28]. Regarding the *correlation structure*: since the DS is defined in the latent space, it ensures new observations to behave in the same way as the ones used to create the model, in the sense that the correlation structure of the model is respected. Regarding the historical *range*: when considering the Hotelling T^2 confidence limit as a raw material specification limit, the new set of raw material properties are constrained to be within historical *ranges* by a multivariate approach. Additionally, historical univariate *ranges* for each property (and other constraints) might be included.

In this study, we initially focus on the *l*-th CQA and, hence, vector $\boldsymbol{y}^{des}$ degenerates to scalar y^{des}, and matrix $\mathbf{Q}$ degenerates to vector $\boldsymbol{q}_l^T$ (l – th row of matrix $\mathbf{Q}$). Besides, one might face three scenarios depending on the kind of specifications for it:

(1) $y_l = y_l^{des}$. In this first scenario, a specific value of the *l*-th CQA is required.
(2) $y_l^{LSL} \leq y_l \leq y_l^{USL}$. In the second scenario, it is desired that the *l*-th CQA is between a lower specification limit (y_l^{LSL}) and an upper specification limit (y_l^{USL}).
(3) In the third scenario, only one specification limit is considered, which might be lower ($y_l^{LSL} \leq y_l$) (scenario 3i) or upper ($y_l \leq y_l^{USL}$) (scenario 3ii).

Following the same framework as in Figs. 2 and 3 shows the DS in the latent space for the latter three scenarios assuming a PLS model with no uncertainty.

In the first scenario, the desired specific value for the *l*-th CQA yields a one-dimensional NS and, the DS is defined by the intersection of this NS and the Hotelling's T^2 confidence region. In the same way, in the second and third scenarios, each specification limit is defined in the latent space by its associated NS. Thus, the DS in the latent space is defined by the intersection of the scores fulfilling the specifications' NSs and the Hotelling T^2 confidence region.

Until now, the DS has been defined without taking into account the prediction uncertainty. However, since empirical models are subject to uncertainty, when a PLS model is inverted, the uncertainty is back-propagated to the calculated inputs (i.e., the DS calculation is probabilistic) [8,9].

4.2. Design space with uncertainty

4.2.1. Bracketing the design space

When prediction uncertainties are present, the DS without uncertainty shown in Fig. 3 does not correspond to the true DS. Therefore, it

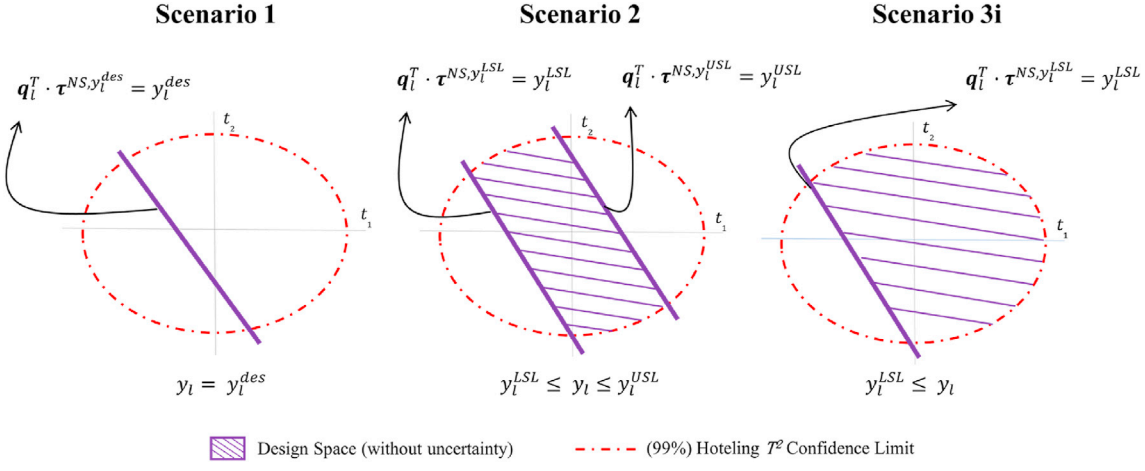


Fig. 3. The Design Space in the latent space, for the three scenarios, assuming a PLS model with no uncertainty. NS: null space.

might be possible to improve the estimation of the DS by running a set of experiments designed within the input domain that have already been used in the past (i.e., the so-called Knowledge Space (KS)). However, exploring the entire KS may be impractical due to the high number of experiments that may be needed to account for the variability in all accessible inputs [8]. For that reason, several approaches have already been proposed in order to define a subspace of the historical KS where the true DS is likely to lie with a predefined confidence level. This subspace is called the Experimental Space (ES).

In particular, Facco et al. [8] present a methodology to account for the backpropagation of the prediction uncertainty in model inversion to bracket the DS. This methodology resorts to the calculation of the prediction interval considering only the inversion solution by means of the pseudo-inverse (Eq. (16)). However, this approach does not consider the difference in the amplitude of the confidence region due to the leverage of different sets of scores along the NS. A proposed solution was given by Palací-López et al. [9] leading to non-linear confidence limits.

A graphical interpretation of the methodology proposed by Palací-López et al. [9] is shown in Fig. 4 assuming the first scenario ($y_l = y_l^{des}$).

Fig. 4 shows that, as expected, moving along the NS one would obtain the same prediction of the l-th CQA. Nevertheless, due to model uncertainty, it does not guarantee to obtain exactly such a prediction. When considering the prediction uncertainty, a prediction interval which is expected to contain the true value of an individual value with a predefined confidence level can be calculated. Note that, since the prediction interval depends on the leverage of the observation (Eq. (9), Eq. (10) and Eq. (11)), its amplitude is expected to be lower for observations close to the centre of projection (small leverage) than for those far away from it (high leverage) [9]. Then, the prediction intervals for the multiple solutions are backpropagated when the model is inverted. Thus, the KS is restricted in such a way as to identify an experimental space in the latent space, which has a high probability of containing the true DS. However, this does not mean high probability of providing assurance of quality, which is what we are interested in when defining multivariate specifications.

4.2.2. Proposed definition of the High-Confidence Design Space

The proposed methodology for defining multivariate raw material specifications is motivated by Facco et al. [8] and Palací-López et al. [9] ideas when back-propagating the uncertainty, but framing the knowledge space with a different purpose. The ES has a high probability of containing the true DS at the expense of including unacceptable raw material batches. By contrast, in this paper we propose considering the prediction uncertainty in a different way, when the model is inverted, in order to define a subspace of the KS where there is assurance of quality with a certain confidence level. For ease of understanding of the proposed methodology, we illustrate the second scenario (Fig. 5) where it is desired that l-th CQA is between y_l^{LSL} and y_l^{USL}.

As discussed above, even though working in the NS associated with the specification limit leads to a predicted value between specifications, it might yield out of specifications values for the l-th CQA due to prediction uncertainties. For that reason, focusing on the y_l^{LSL}, one should accept raw materials properties such that its projection in the latent space leads to a lower endpoint, which is equal or higher than the y_l^{LSL}, thus delimiting a lower confidence region (Eq. (19)).

$$y_l^{LSL} \leq \boldsymbol{q}_l^T \cdot \boldsymbol{\tau}^{new} - t_{\alpha/2,\,N-df} \cdot s_{e_l^{new}} \tag{19}$$

When calculating this confidence limit for the multiple solutions along the NS of y_l^{LSL}, a non-linear boundary is obtained for the y_l^{LSL} as is shown in Fig. 5a. Such boundary in the latent space refers to the Lower Specification Confidence Limit (LSCL). If working in the LSCL there will be a high probability to obtain the l-th CQA higher than the y_l^{LSL}.

In the same way, considering the y_l^{USL}, one should accept raw materials properties such that its projection in the latent space leads to an upper endpoint which is equal or lower than the y_l^{USL}, thus delimiting an upper confidence region (Eq. (20)).

$$y_l^{USL} \geq \boldsymbol{q}_l^T \cdot \boldsymbol{\tau}^{new} + t_{\alpha/2,\,N-df} \cdot s_{e_l^{new}} \tag{20}$$

Following an analogous reasoning as before, another non-linear boundary, called Upper Specification Confidence Limit (USCL), is obtained for the y_l^{USL} (see Fig. 5b). If working in the USCL there will be a high probability to obtain the l-th CQA lower than the y_l^{USL}.

Appendix A shows the analytical expression, which allows calculating the score belonging to both the lower and upper specification confidence limits given its respective score in the NS for the l-th CQA. Although Eq. (19) and Eq. (20) refer to one-sided prediction intervals, the t-statistic is calculated at the $\alpha/2$ significance level because two specifications limits are considered. In the case of having one specification limit (i.e., third scenario), Eq. (19) or Eq. (20), as appropriate, would be used at α significance level.

The intersection regions delimited by the LSCL, USCL and the Hotelling T^2 confidence ellipsoid, delimits the so-called High-Confidence Design Space, i.e., the Multivariate Raw Material Specification Region, where any batch of raw material properties results in a prediction interval for the CQA within specifications. Therefore, from a frequentist probabilistic interpretation, these batches are expected to produce product with CQAs within specification limits with a confidence level equal or higher than $1-\alpha$. In other words, this definition of the High-Confidence DS has been demonstrated to provide assurance of quality with at least a certain confidence level (Fig. 5c). The High-Confidence DS is a potential opportunity to establish real-time release (RTR), which is defined as the ability to evaluate and ensure the acceptable quality of

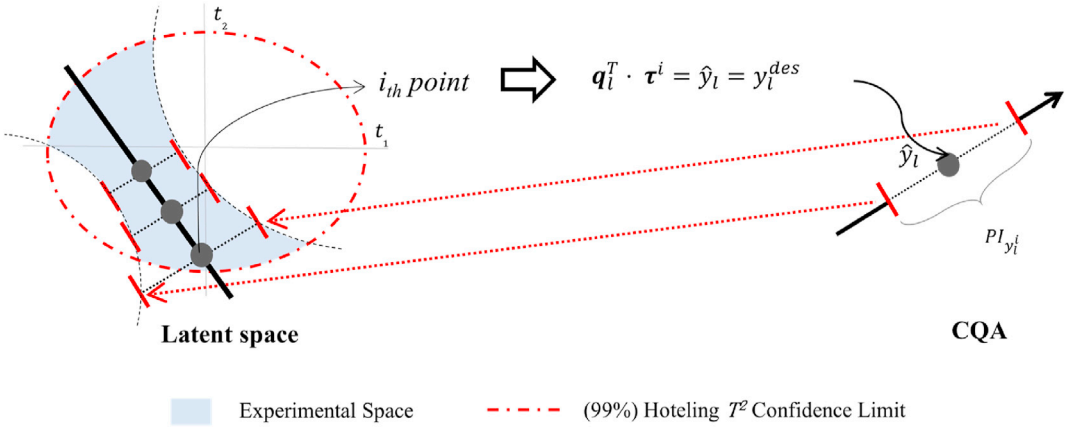


Fig. 4. *FIRST SCENARIO: The methodology proposed by Palací-López* et al. [9].

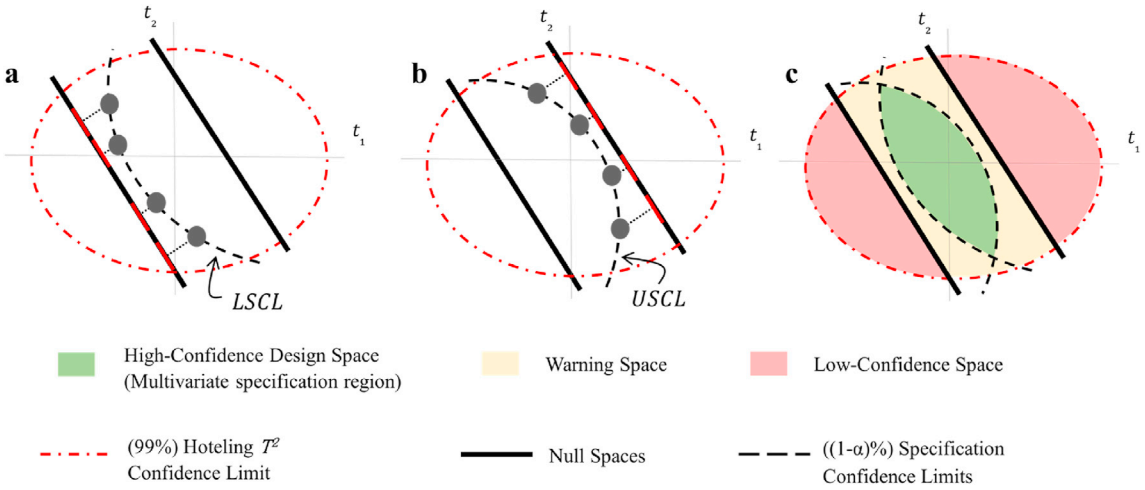


Fig. 5. SECOND SCENARIO: Graphical interpretation of the proposed definition of the High-Confidence Design Space. (a) lower specification confidence limit (LSCL). (b) upper specification confidence limit (USCL). (c) Splitting the KS into High-Confidence Design Space, Warning Space and Low-Confidence Space.

final product based on inputs variables (e.g., raw material properties) without using end-product testing [10].

Additionally, the intersection between the region bounded by the two NSs corresponding to the y_l^{LSL} and y_l^{USL}, and the Hotelling's T^2 confidence region, but outside the High-Confidence DS, defines the so-called Warning Space (Fig. 5c). Note that, although this space does not belong to the Multivariate Raw Material Specification Region as defined, it does not necessarily imply the rejection of batches. In fact, batches lying within the Warning Space lead to predicted values between specifications, but they result in prediction intervals for the CQA partially outside of specifications given the predefined confidence level $1-\alpha$. Namely, there is no assurance of quality due to the prediction uncertainty and, hence, RTR testing is not feasible. Instead of that, end-product testing may be employed, which usually involves undertaking specific lab-testing procedures on samples of the final product. This could be interesting when rejecting all batches in the Warning Space is not affordable. Finally, the Low-Confidence Space (Fig. 5c) leads to predicted values outside specifications. Although batches lying within this subspace may lead to response values between specifications, most of the time such values are expected to be outside.

Therefore, following the proposed approach, the KS is split into three regions: High-Confidence DS, Warning Space and Low-Confidence Space, providing a strategy where RTR or end-product testing, can be used as needed.

Note that, the High-Confidence DS is more restrictive that the unknown true DS, and the less uncertainty there is, the more similar the High-Confidence DS and the true DS are, as it is graphically shown in Fig. 6.

High uncertainty in the data is reflected in a low goodness of prediction model. But this does not limit the proposed methodology, indeed, the lower goodness of prediction, the more crucial it is to take uncertainties into account if product quality is to be guaranteed. In that point, the authors would like to challenge the widely held view that a low goodness of prediction model is useless and point out that low goodness of prediction model, typical from the industry 4.0 environment, can be useful if being cautious. In this sense, García-Muñoz and Mercado [34] already worked in a real process under control where a LV regression model, that had the ability to systematically predict 21% of the variability in the quality attribute, was used with a great potential for improvement. However, in certain situations a low goodness of prediction model may be a warning of non-linearities in the original dataset that is not captured adequately by the linear PLS model [35].

To summarize, Fig. 7 shows the DS (if there is not uncertainty in the model), the experimental space and our proposed High-Confidence DS for all scenarios.

The first scenario is a particular case of the second scenario where $y_l^{LSL} = y_l^{USL}$. In this case, there is not intersection between the LSCL and USCL and, therefore, the High-Confidence DS does not exist. Up to this

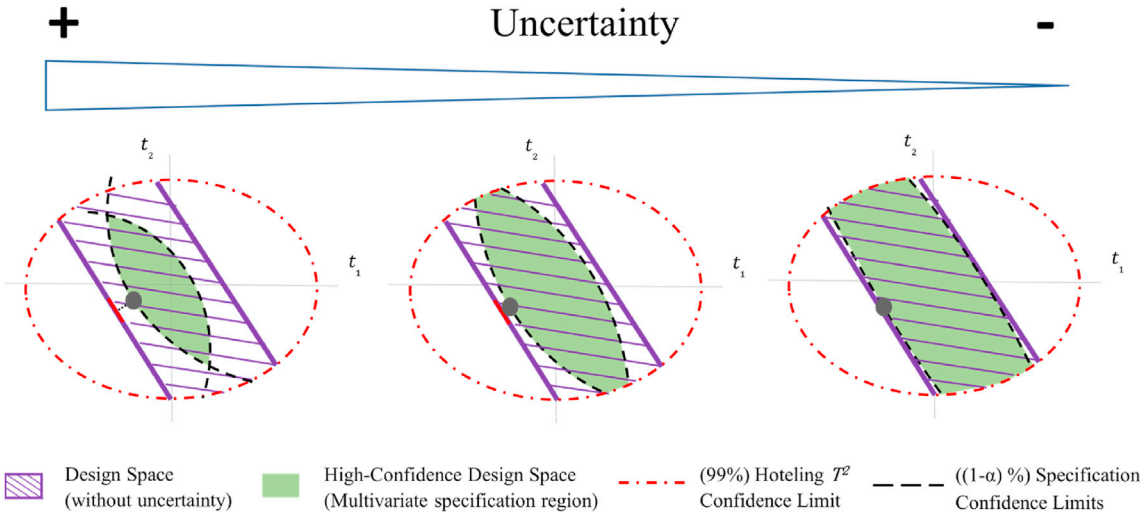


Fig. 6. SECOND SCENARIO: Effect of the uncertainty on the High-Confidence DS related to the DS without uncertainty.

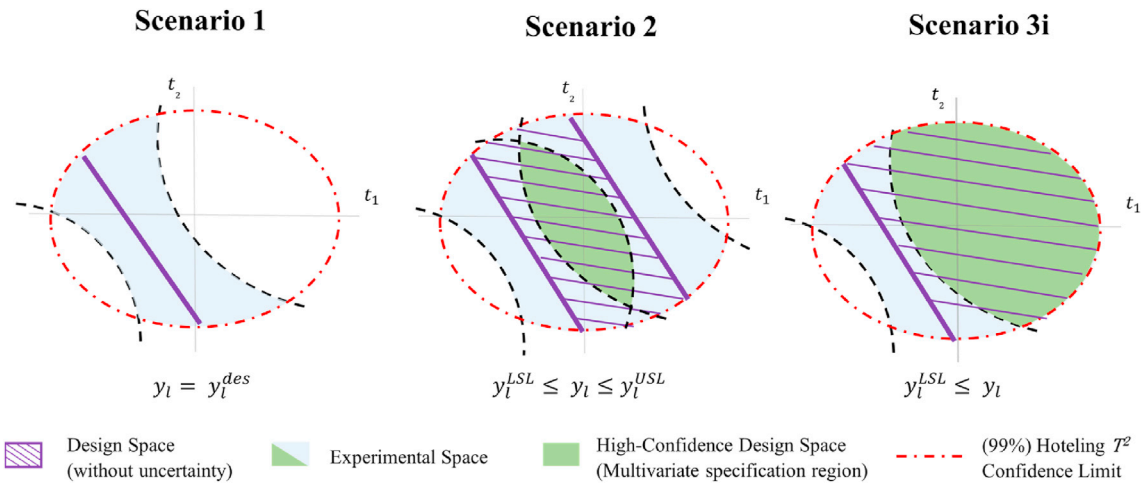


Fig. 7. Comparison of the DS (without uncertainty), ES and High-Confidence DS for the three scenarios.

point, we have defined the High-Confidence DS for the l-th CQA. The joint High-Confidence DS for the L CQAs will be obtained as the intersection of the L High-Confidence DSs for each CQA.

5. Exploiting the model

Once multivariate raw material specifications have been defined as discussed above, the model can be used to inspect every new batch of raw material, z^{obs} (Phase II). This allows the user to predict if the CQAs of the product, that would be manufactured using any new raw material batch, would be within specifications, and consequently accept or reject the raw material batch prior to introducing it into the production process. The procedure for that is as follows:

(4) Mean-center and scale z^{obs} using the same mean and scaling factor used on the calibration data when the PLS model was developed in Phase I.
(5) The scores $\boldsymbol{\tau}^{obs}$ are obtained from the linear combinations of mean-centered and scaled raw materials properties according to Eq. (4), and the $SPE_{z^{obs}}$ is obtained according to Eq. (6).
(6) The final decision on whether to accept or reject a new raw material batch is up to the user based on the values of $SPE_{z^{obs}}$ and $\boldsymbol{\tau}^{obs}$. When the $SPE_{z^{obs}}$ is higher than $SPE_{\lim}$, this suggest that their properties reflect a different correlation structure than that of the raw material batches from the historical dataset used to build the PLS model. It is then impossible to predict with the fitted PLS model the impact of this raw material batch on CQAs of the final product. Besides, in such a case, one could use the SPE contribution plot in order to examine which raw material properties contribute the most to this high SPE value, providing the supplier with useful information about deviations in the batch raw material properties. Regarding the projection in the latent space $\boldsymbol{\tau}^{obs}$, if these scores fall within the High-Confidence DS, this batch will be expected to produce product with CQAs within specification limits with at least a certain confidence level. Note that, instead of rejecting all the high $SPE_{z^{obs}}$ and high T^2_{obs} raw material batches, one may also process some of them (when deviations are not too important), incorporate them as new design points to augment the historical data matrices $\mathbf{Z}$ and $\mathbf{Y}$, and fit a new PLS model in order to better define sequentially the multivariate specification region.

6. Industrial case studies

6.1. First industrial case study: cereal extraction process

Historical data collected from a maize cereal extraction process is used to illustrate the proposed methodology. A schematic representation of such process is shown in Fig. 8.

The maize is fed to the production process where, initially, it is cleaned to free the maize of all kinds of impurities and then it is steeped. Subsequently, a grinding process takes place to grind the harder parts of the maize, followed by a degerminating process so that germ is separated from the fiber, gluten, and starch. Finally, after a sieving process carried out to separate the fiber, a primary separator splits by centrifugal force the stream in two fractions: gluten and slurry starch. The latter has a great interest as it has become a major industrial raw material.

The data available in this case are a compilation of eight raw material properties ($\mathbf{Z}$) of maize: promatest value, protein, acid value, specific weight, burnt grain, broken grain, starch and extractable lipids, and one response variable $\boldsymbol{y}$ (extraction yield of starch slurry). These variables are easily registered in order to assess the feasibility of a raw material batch. In total, 989 historical batches/observations were measured: $\mathbf{Z}$ (989×8) and $\boldsymbol{y}$ (989×1), and they were divided randomly in two sets: calibration set (70%) for Phase I and exploiting set (30%) for Phase II. Besides, a lower specification limit of 69% is considered for the response variable (i.e., this case refers to the scenario 3i).

Leave-one-out cross-validation (CV) was used for selecting the number of PLS components. Thus, two LVs were chosen to fit a PLS model ($R^2_{Zcum} = 36.8\%$, $R^2_{Ycum} = 26.8\%$ and $Q^2_{Ycum} = 25.1\%$) using the 693 calibration observations (Phase I). The R^2 values (goodness of fit) give the percentage of the total sum of squares of $\boldsymbol{y}$ and $\mathbf{Z}$, respectively, that are explained by the fitted PLS model, while the Q^2_{Ycum} (goodness of prediction) gives the percentage of the total sum of squares of $\boldsymbol{y}$ that can be predicted with the PLS model by CV. In Phase I it is also crucial to validate the model by monitoring charts for SPE and T^2 (shown in Fig. 9), in order to determine whether historical/happenstance data are consistent with normal process conditions (i.e., common cause process variations).

Fig. 9 shows that none of the historical batches exhibit any unusual behaviour caused by special cause process variations. Notice that, although some of them slightly exceed the upper confidence limit, they correspond approximately to 1% false alarm rate (expected when using 99% confidence limits). Fig. 10 illustrates the 99% Hotelling T^2 confidence limit, the NS associated with the LSL, and its 90% confidence limit when considering the prediction uncertainty (i.e., the Low Specification Confidence Limit. LSCL). The intersection of all confidence regions, defined by their limits, yields the High-Confidence DS (i.e., the proposed multivariate raw material specifications in the latent space) within which there is assurance of obtaining superior or equal yields to 69% with at least 90% confidence level.

To evaluate the performance of the definition of the multivariate raw material specification region, a diagnostic test is carried out using the

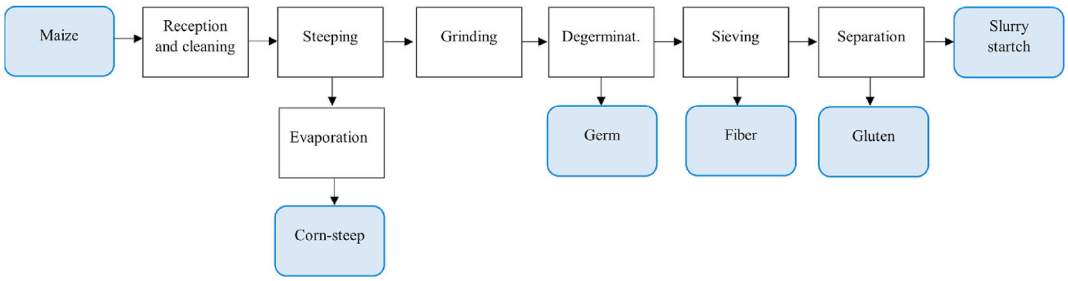


Fig. 8. Schematic representation of the maize cereal extraction process.

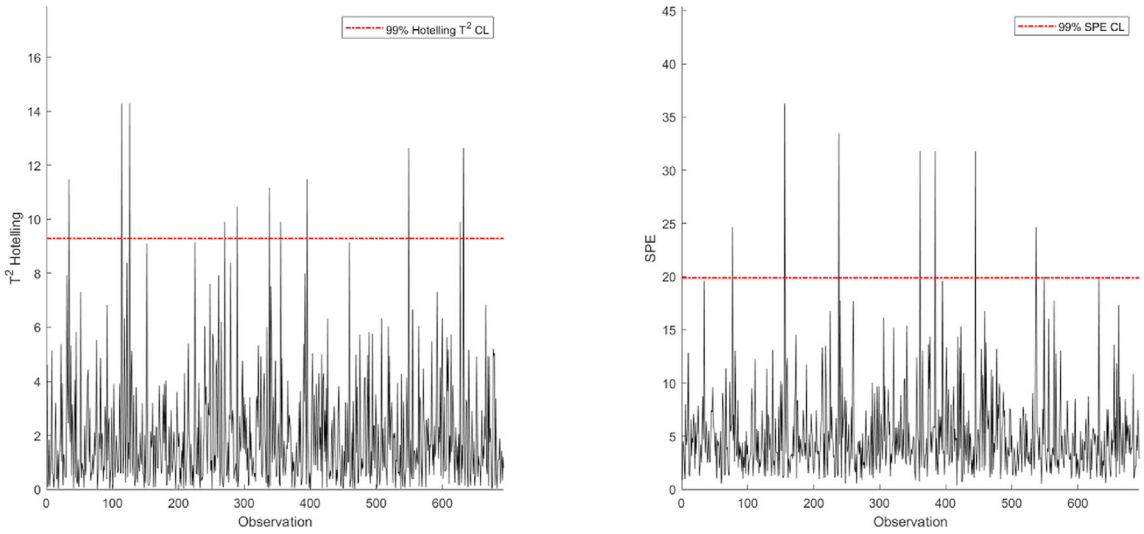


Fig. 9. *First industrial case study: Monitoring charts for SPE and T^2 in Phase I.*

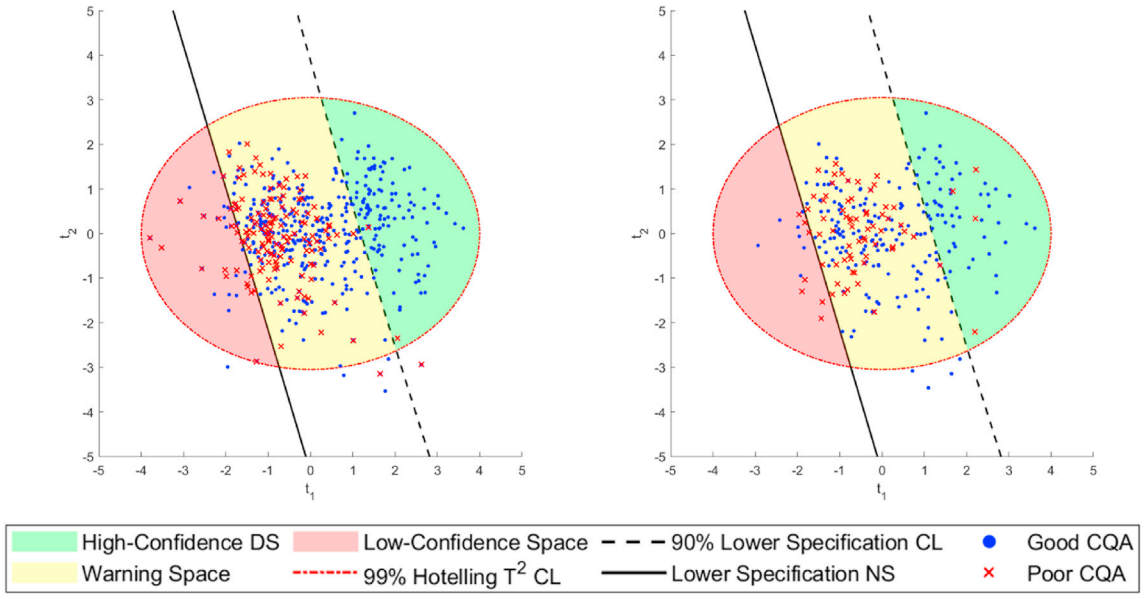


Fig. 10. First industrial case study (scenario 3i): Graphical definition of the High-Confidence DS (Multivariate Raw Material Specification Region), Warning Space and Low-Confidence Space by showing (a) calibration data and (b) exploiting data.

validation set. In particular, type I risk, type II risk and the Negative Predictive Value (NPV) are calculated for the High-Confidence DS. The NPV is the proportion of batches that actually result in a good product out of all those within the High Confidence Design Space, and, hence, this metric is directly connected to the definition of the High-Confidence DS itself.

De Smet [2] and Duchesne and MacGregor [1] approaches would have ended up defining a straight line or an ellipsoid in a subjective way, that would best balance type I and type II risks in Fig. 10a. Besides, if the PLS model was of a higher dimension (A≥3), it would be difficult to

decide the general shape and locus that best defines the separation between good and poor quality, unlike the proposed approach, which does not suffer from such handicap regardless of the dimensionality of the latent space. On the other hand, García-Muñoz, Dolph, and Ward [3] would have obtained a wider region, akin to the DS without considering the uncertainty (the joint of High-Confidence DS and Warning Space). However, because of the uncertainty, this approach would result in accepting almost every batch of raw materials (no matter if they are acceptable or unacceptable), leading to 9% type I risk, 90% type II risk and 75% NPV with exploiting data (see Fig. 10b). None of these approaches are probabilistic, and therefore they do not allow knowing the confidence level in meeting the final product quality specifications.

By contrast, our High-Confidence DS is defined with at least a 90% confidence level of obtaining superior or equal yields to 69%. Thus, one would expect that, of the batches lying within the High-Confidence DS, 90% or more would be acceptable batches (the NPV for the High-Confidence DS is 93.3%). On the other hand, the High-Confidence DS leads to 75.0% type I risk and 5.6% type II risk. This means that if only batches lying within the High-Confidence DS are accepted, 5.6% of unacceptable batches of raw materials will be accepted at the expense of rejecting 75.0% of acceptable batches. These results are the consequence of the low PLS goodness of prediction ($Q^2_{Ycum} = 25.1\%$) in this case study, due to the fact that historical data presents a low signal to noise ratio. Alternatively, one could accept batches lying within the Warning Space knowingly that the NPV in such space would be 70.8% and, hence, likely end-product test should be required. Another option would be to balance the type I and type II risks by modifying the confidence level of the High-Confidence DS. Fig. 11 shows the High-Confidence DS for different confidence levels (50, 70, 90 and 99%) with exploiting data. The corresponding type I risk, type II risk and NPV for the High-Confidence DS, and NPV for the Warning Space are shown in Fig. 12. Note that, the 50% confidence level case corresponds to the DS without considering the uncertainty.

Fig. 11 shows that as confidence level increases, a tighter High-Confidence DS is spanned, thereby, the type II risk is reduced at the expense of increasing the type I risk, as is shown in Fig. 12. Therefore, the

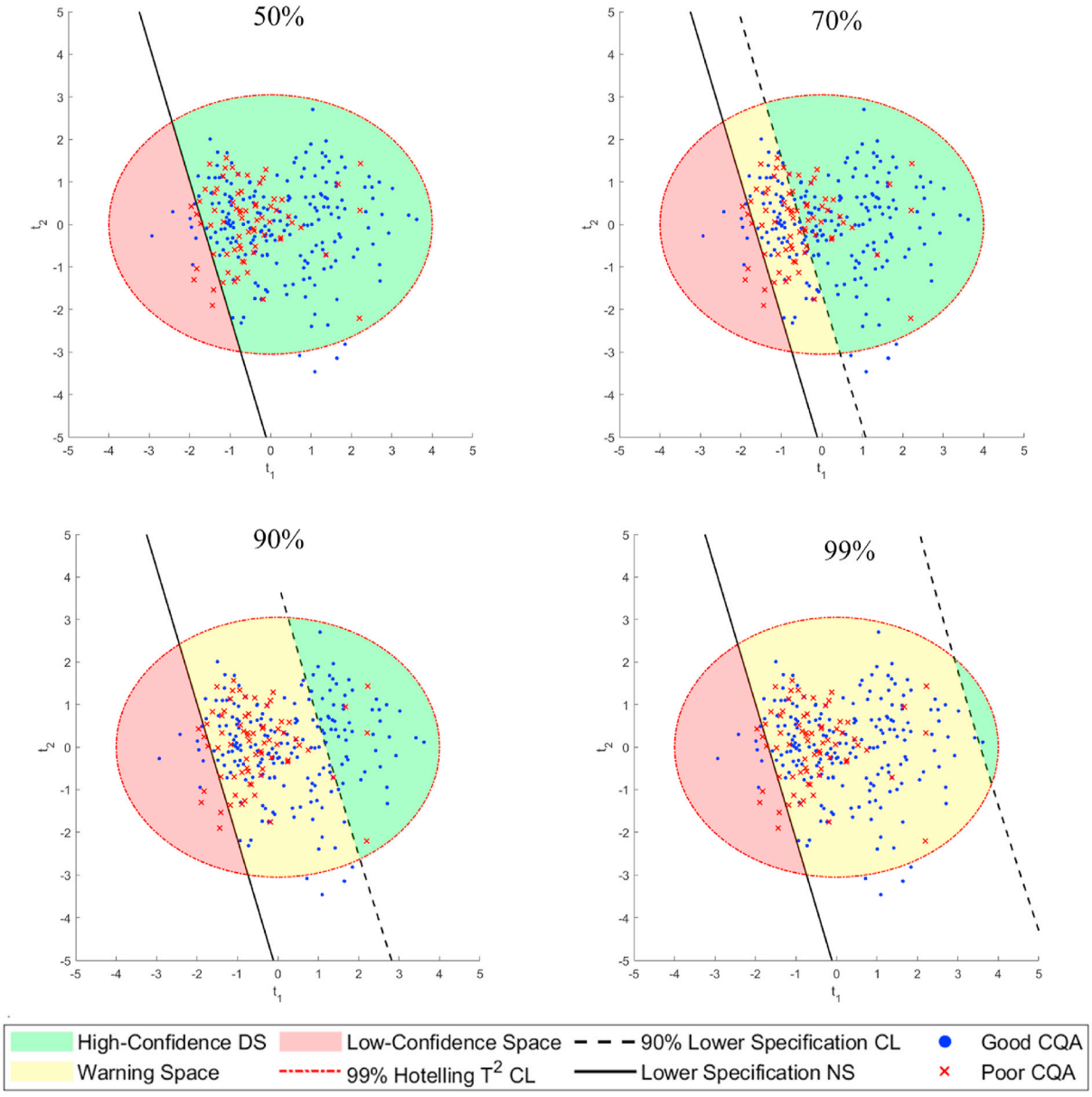


Fig. 11. First industrial case study (scenario 3i): High-Confidence DS (Multivariate Raw Material Specification Region), Warning Space and Low-Confidence Space for several confidence levels with exploiting data.

confidence level of the High-Confidence DS must be chosen according to the users by balancing the consequences of having type I and type II errors in their processes, and the total amount of such errors. Besides, for all cases, the NPV is equal or higher than its corresponding confidence level as expected.

In order to investigate how PLS goodness of prediction Q^2_{Ycum} affects the performance of the High-Confidence DS a simulation study is carried out. In these simulations we assume that the true model relating $\mathbf{Z}$ and $\boldsymbol{y}$ is, indeed, the one calculated by the calibration set. Hence, individual values of $\boldsymbol{y}$, y^{obs}, are obtained using Eq. (21) given a batch of raw material z^{obs} and the weighting matrices $\boldsymbol{q}$ $L = 1\mathbf{Q} = \boldsymbol{q}^T$ and $\mathbf{W}^*$:

$$y^{obs} = \boldsymbol{q}^T \cdot \mathbf{W}^{*\mathrm{T}} \cdot z^{obs} + e^{obs} \tag{21}$$

where e^{obs} is an independent random noise value from a normal distribution with zero mean and standard deviation σ. By modifying the value of such standard deviation, one can create simulated datasets yielding PLS models with different goodness of prediction. Fig. 13 shows the High-Confidence DS with 90% confidence level of obtaining superior or equal yields to 69% for different datasets simulated from the exploiting dataset by using a standard deviation of 0.025, 0.1, 0.5 and 1 yielding Q^2_{Ycum} of 90.8%, 72.5%, 38.7% and 20.8%, respectively. 2000 batches have been simulated for each dataset to obtain more accurate results with respect to the original data.

Fig. 13 shows that the lower the noise standard deviation, the higher the goodness of prediction and, consequently, the clearer the discrimination between acceptable and unacceptable raw materials. Besides, regardless the goodness of prediction, the proposed method defines the multivariate specification region given the same confidence level (90%). As can be see, lower values for the goodness of prediction result in narrower multivariate specification region where more acceptable material

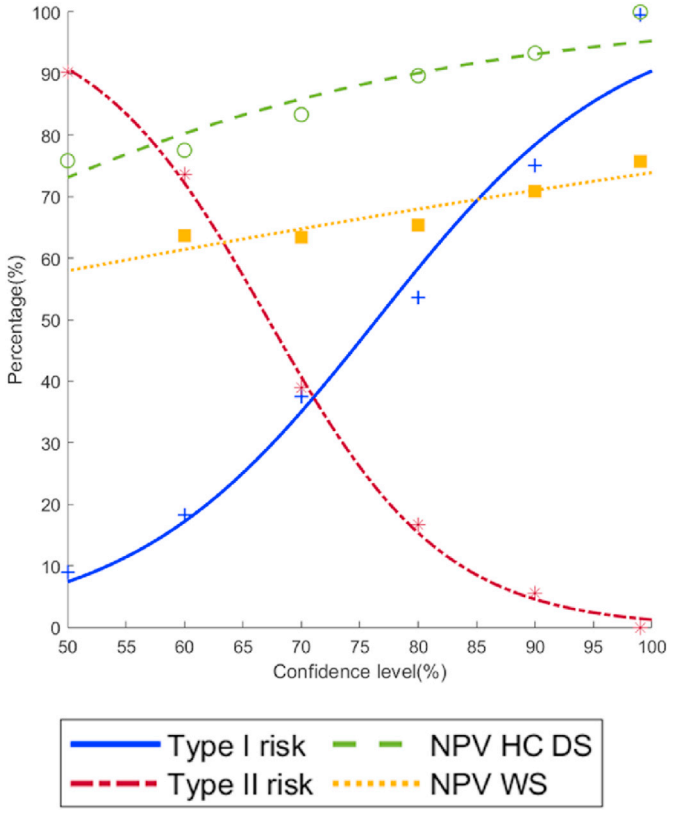


Fig. 12. First industrial case study (scenario 3i): type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) vs confidence level.

is rejected to guarantee such confidence level. This will affect the type I and type II risks as shown in Fig. 14.

Fig. 14 shows that with moderate/high values of Q^2_{Ycum} it is feasible to obtain High-Confidence DS with high confidence level and low type I and II risks, and high NPV. For example, given desired yields equal or superior to 69%, the High-Confidence DS with 90% confidence level and $\sigma = 0.025$ (Q^2_{Ycum} =90.8%) leads to 9.6% type I risk, 7.0% type II risk and 99.0% NPV. However, with low values of Q^2_{Ycum} it is more critical to consider the prediction uncertainty for guarantying quality (i.e., high NPV in the High-Confidence DS) at the expense of increasing the type I risk.

Note that the apparently bad performance for low values of Q^2_{Ycum} is solely due to the nature of the data and not the methodology, as noise refers to random variation with no pattern, and therefore usually unavoidable and unpredictable.

In case of desiring to increase the signal-to-noise ratio of the data sets, some process excitation is needed. Multivariate design of experiments can be used such that it provides the greatest amount of additional information with respect to the information available in the existing dataset [36]. Considering these new observations from experimentation in addition to the historical/happenstance data will improve the estimation of the DS (i.e., wide multivariate specification region with high confidence level and low type I and type II risks will be obtained).

A sensitivity analysis was undertaken to assess the robustness of the High-Confidence DS with respect to i) the observations used to build the PLS model, and ii) the number of PLS components.

i) Happenstance data

To evaluate the sensitivity analysis of the High-Confidence DS with respect to happenstance data, the initial data set (989 historical batches/observations) are divided 100 times randomly in two sets: calibration set (70%) and exploiting set (30%). Then, the High-Confidence DS with a 90% confidence level is calculated each time using the corresponding calibration set. Fig. 15 shows the boxplots of the distribution of type I risk, type II risk and NPV for the High-Confidence DS, and NPV for the Warning Space from all PLS models using their corresponding exploiting set. Besides, as example, Fig. 16 shows the High-Confidence DS for two of the 100 models generated.

Fig. 15 shows that Type I and II risks, and NPV hardly vary for the different PLS models, and Fig. 16 shows that both subsets lead to a similar region of the High-Confidence DS.

ii) Number of PLS components

The number of components to be used is a very important property of a PLS model and their choice must be done according to the purpose of such model. In our case study, we have evaluated how changes in the number of components may affect the type I and type II risks of the High-Confidence DS with 90% confidence limit. Table 1 shows that no relevant differences in the performance of the diagnostic test are observed when adding PLS components. The reason for this is the fact that the goodness of prediction (Q^2_{Ycum}) is quite similar among the models.

6.2. Second industrial case study: petrochemical process

This industrial case study refers to a catalytic afterburner used as control device for oxidation of undesirable combustible gases in a petrochemical process. The properties of the catalyst have an impact on the afterburn quality process and, hence, it is not only crucial to determine the raw material properties of the catalyst, but also define its multivariate specifications for ensuring such quality.

The historical/happenstance data available are a compilation of nine properties of the afterburn catalyst ($\mathbf{Z}$) related to regenerated catalyst percentage, catalyst density, particle size distribution and chemical

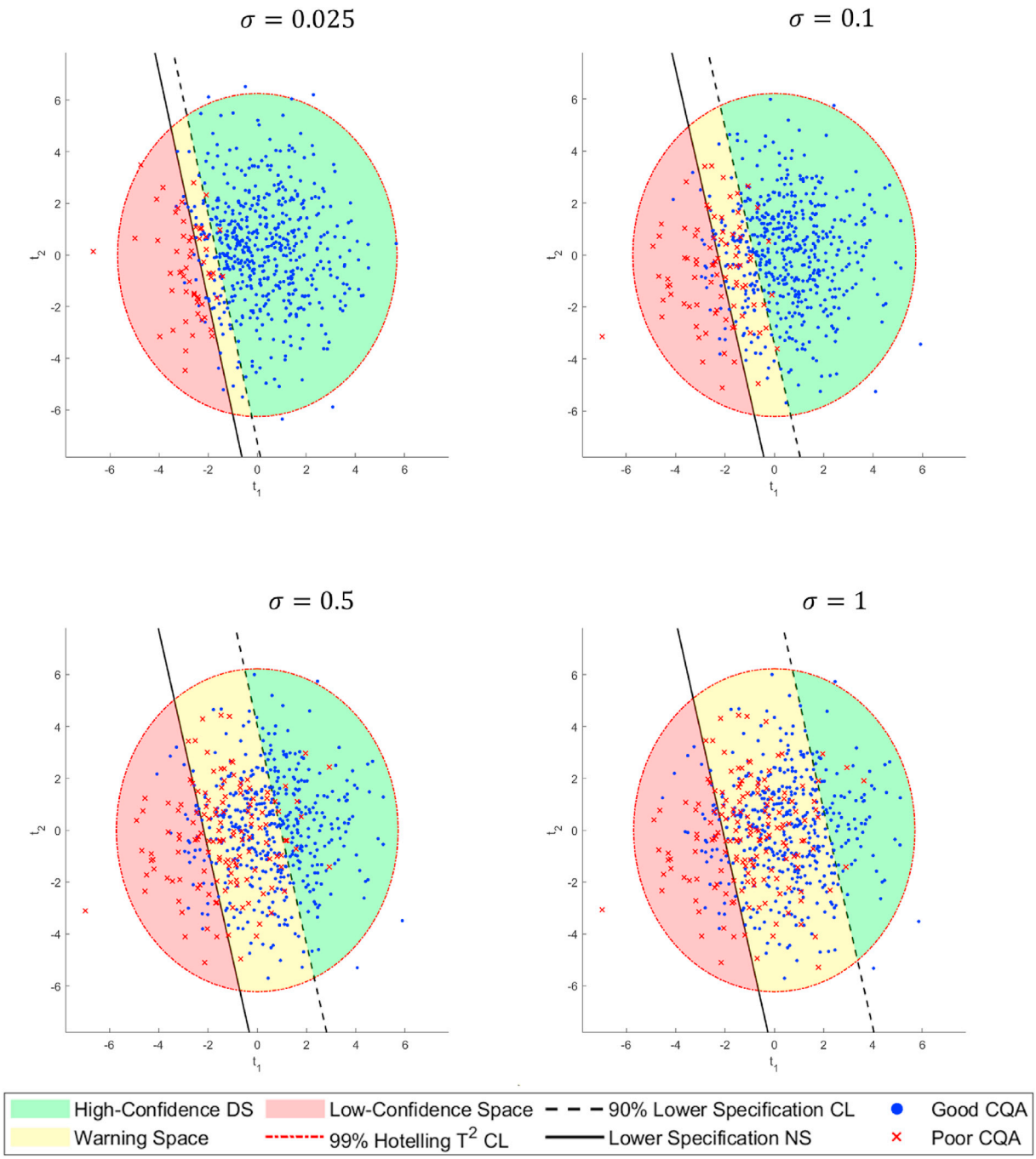


Fig. 13. High-Confidence DS (Multivariate Raw Material Specification Region), Warning Space and Low-Confidence Space for simulated data with different noise variability σ.

composition, and one response variable ***y*** (afterburn yield). In total, 9971 historical batches/observations were measured, and they were divided randomly in two sets: calibration set (70%) for Phase I and exploiting set (30%) for Phase II. Besides, a lower and upper specification limits are considered for the response variable, hence this case refers to the second scenario.

Leave-one-out CV was used for selecting the number of PLS components. Thus, two LVs were chosen to fit a PLS model (R^2_{Zcum} = 56.3%, R^2_{Ycum} = 73.6% and Q^2_{Ycum} = 73.5%) using calibration observations (Phase I). This is a case study with a moderate goodness of prediction (Q^2_{Ycum}). None of the historical observations exhibit any unusual behaviour caused by special cause process variations based on *SPE* and T^2 charts in Phase I (charts not shown).

Fig. 17a illustrates the High-Confidence DS with a 90% confidence level resulting in 41.6% type I risk, 9.6% type II risk and 98.0% NPV with exploiting data. However, if uncertainty had not been considered, 17.2% type I risk, 34.7% type II risk and 95.0% NPV would have been obtained. As expected, Fig. 17b shows that as confidence level increases, the type II risk is reduced at the expense of increasing the type I risk. It should be noticed that the type I and II risks and NPV not only depend on the goodness of prediction but also on other factors such as the scenario, the value of the specification limits or the validation data. For that reason, different case studies with the same Q^2_{Ycum} could result in slightly different type I and II risks for the same confidence level, as it happens if comparing the simulated case for the first case study using a standard deviation of 0.1 (Fig. 14b) and the second case study (Fig. 17b).

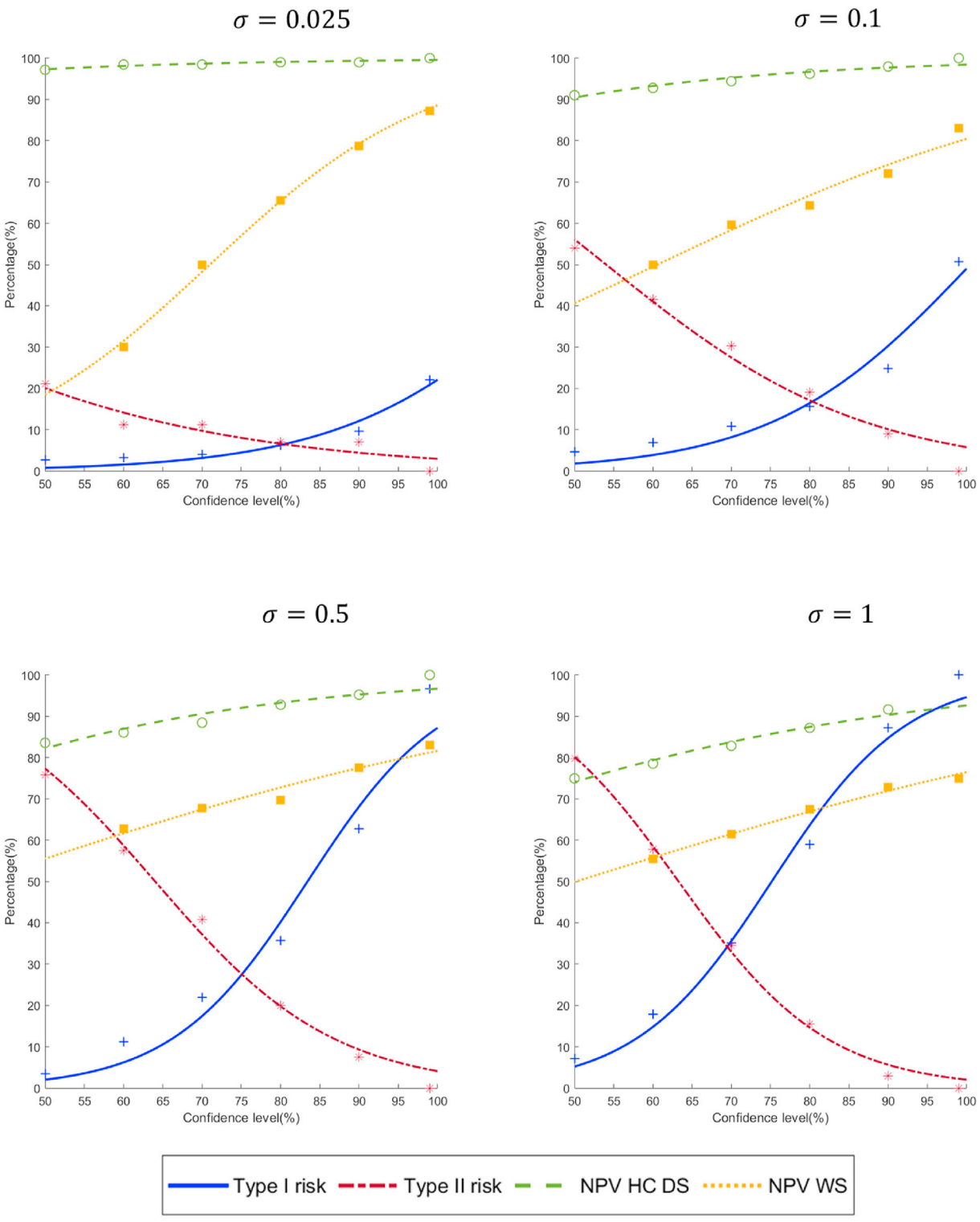


Fig. 14. Type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) vs confidence level for simulated data with different noise variability σ.

On the other hand, since in the second case study there is a substantial variation in the goodness of prediction when adding the second PLS component, the sensibility analysis of the number of the PLS components is also undertaken (Table 2).

Unlike the first case study (Table 1), Table 2 shows relevant improvements in the reduction of type I and II risks when adding the second PLS component, but not after adding more components. For that reason, it is concluded that the CV criterion for the selection of two PLS component results in good performances indices.

6.3. *Third industrial case study: blown film process*

In order to illustrate the proposed methodology in a case study with a high goodness of prediction, we consider data collected from an industrial blown film process presented by Duchesne and MacGregor [1]. In this example, 10 resin properties (**Z**) and 25 film characteristics (***Y***) were

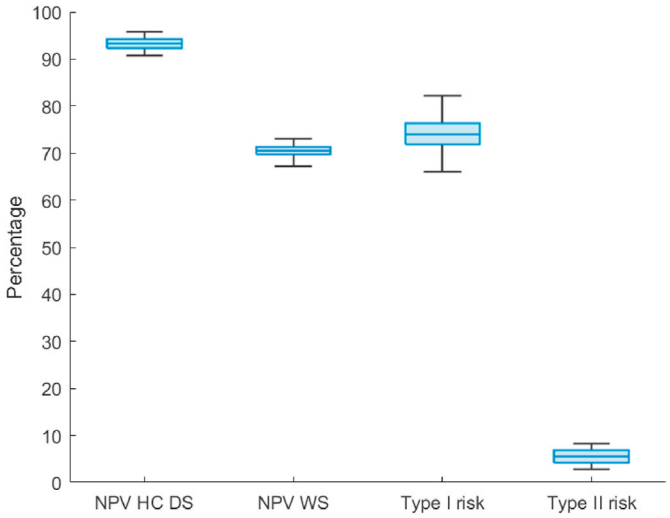


Fig. 15. Boxplots of type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) (from the 100 PLS models using their corresponding exploiting set).

measured for a series of 55 films/observations, but one of these observations was dismissed due to unusual behaviour. In addition, an overall judgment of film quality for the specific market into low, medium and high suitability is available for each film. This quality variable would be used for mapping the region of high film quality according to the De Smet [2] and Duchesne and MacGregor [1] approaches. But, since our approach defines the High-Confidence DS from specification limits in film characteristics, such quality variable was used to define the product acceptance limit that best discriminate high from medium and low quality. In fact, due to the high correlation of film characteristics, only a lower specification limit for one characteristic was required, referring to scenario 3i. Thus, in total 54 samples were analyzed: $\mathbf{Z}$ (54 $\times$ 10) and $\boldsymbol{y}$ (54 $\times$ 1) and they were divided randomly in calibration set (70%) and exploiting set (30%).

Two LVs were chosen by Leave-one-out CV to fit a PLS model (R^2_{Zcum} = 58.0%, R^2_{Ycum} = 89.5% and Q^2_{Ycum} = 85.0%). None of the historical observations exhibit any unusual behaviour caused by special cause process variations based on *SPE* and T^2 charts in Phase I (charts not shown).

Table 1
First industrial case study (scenario 3i): Goodness of prediction ($Q^2_{Y,\,cum}$), type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) as a function of the number of PLS components (High-Confidence DS for 90% confidence level).

A	$Q^2_{Y,\,cum}$ (%)	Type I (%)	Type II (%)	NPV HC DS (%)	NPV WS (%)
1	24.3	76.3	6.9	91.4	70.8
2	25.1	75.0	5.6	93.3	70.8
3	25.2	74.1	4.2	95.1	70.3
4	25.3	74.1	5.6	93.5	69.6
5	25.3	74.1	5.6	93.5	69.8
6	25.3	74.1	4.2	95.1	69.8
7	25.3	74.1	4.2	95.1	70.0
8	25.3	74.1	4.2	95.1	70.0

Fig. 18 illustrates the High-Confidence DS with a 90% confidence level resulting in 0% type I risk and 0% type II risk with exploiting data. The same results would be obtained by De Smet [2] and Duchesne and MacGregor [1] approaches, and without considering the prediction uncertainty. This result was expected and suggests that when having high goodness of prediction, considering prediction uncertainty is less critical. However, this situation is not frequent when working with historical/happenstance data, typical from Industry 4.0.

7. Conclusions

In this paper, we propose a novel approach to define an analytical expression for defining the multivariate raw material specification region in the latent space where there is assurance of quality with a certain confidence level for the CQAs of the final product (i.e., the so-called High-Confidence design space). Thus, it would allow evaluating the capability of the raw material batches of producing product with CQAs within specification limits, before producing a single unit of the product, and based on that information, making a decision about accepting or not the supplier raw material batch. This is totally different to existing

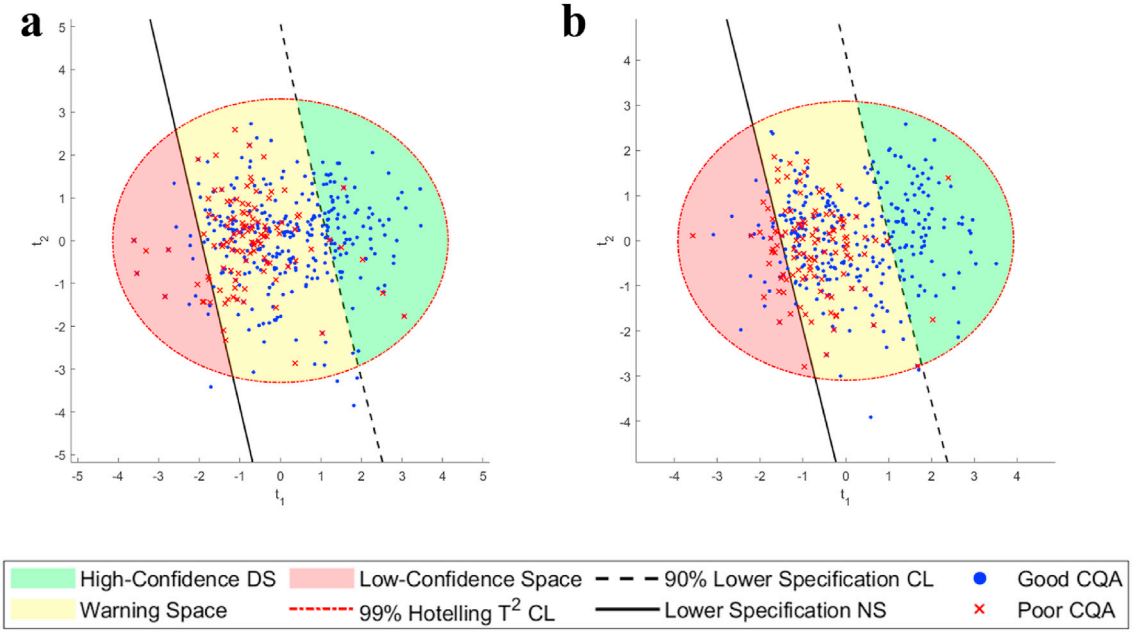


Fig. 16. Graphical definition of the High-Confidence DS (Multivariate Raw Material Specification Region), Warning Space and Low-Confidence for two of the 100 models generated: (a) and (b).

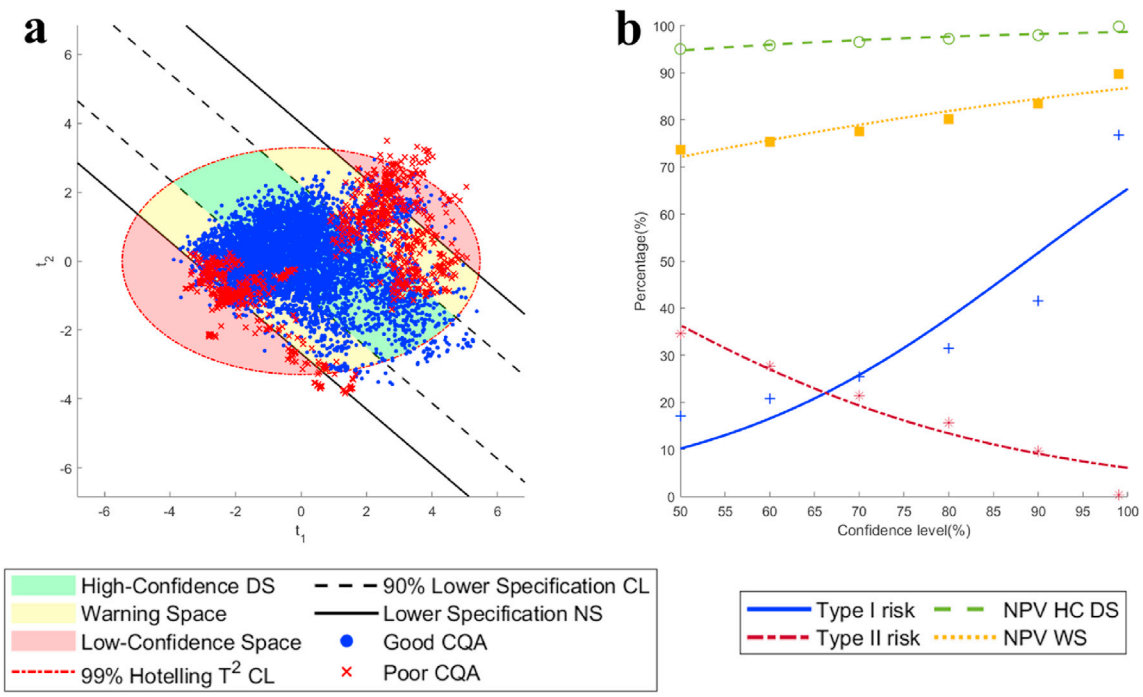


Fig. 17. Second industrial case study (scenario 2): (a) Graphical definition of the High-Confidence (HC) DS (Multivariate Raw Material Specification Region), Warning Space (WS) and Low-Confidence by showing exploiting data. (b) Type I risk, type II risk and NPV for the High-Confidence DS, and NPV for the Warning Space vs confidence level.

Table 2
Second industrial case study (scenario 2): Goodness of prediction ($Q^2_{Y,\,cum}$), type I risk, type II risk and NPV for the High-Confidence (HC) DS, and NPV for the Warning Space (WS) as a function of the number of PLS components (High-Confidence DS for 90% confidence level).

A	$Q^2_{Y,\,cum}$ (%)	Type I (%)	Type II (%)	NPV HC DS (%)	NPV WS (%)
1	47.3	66.7	16.9	94.0	86.4
2	73.5	41.6	9.6	98.0	83.5
3	74.2	40.5	8.1	98.3	83.7
4	74.5	41.4	7.5	98.4	83.0
5	74.5	41.5	7.2	98,5	82.8
6	74.6	40.6	6.6	98.6	82.9
7	74.7	41.3	7.5	98.4	83.5
8	74.9	40.3	7.8	98.4	82.8

approaches that evaluate (and also accept or reject) raw material batches based on their raw material properties but not on the desired final product properties.

This methodology is based on the inversion of the PLS model, and the most remarkable advantages are:

- It can be used with historical data (i.e., daily production data not coming from any experimental design but with varying raw material properties, typical from Industry 4.0 environment) since, when fitting PLS models, causality can be inferred in the latent space, which allows the meaningful inversion of the model.
- It considers a multivariate approach providing much insight into what constitutes acceptable raw material batches when their properties are correlated.

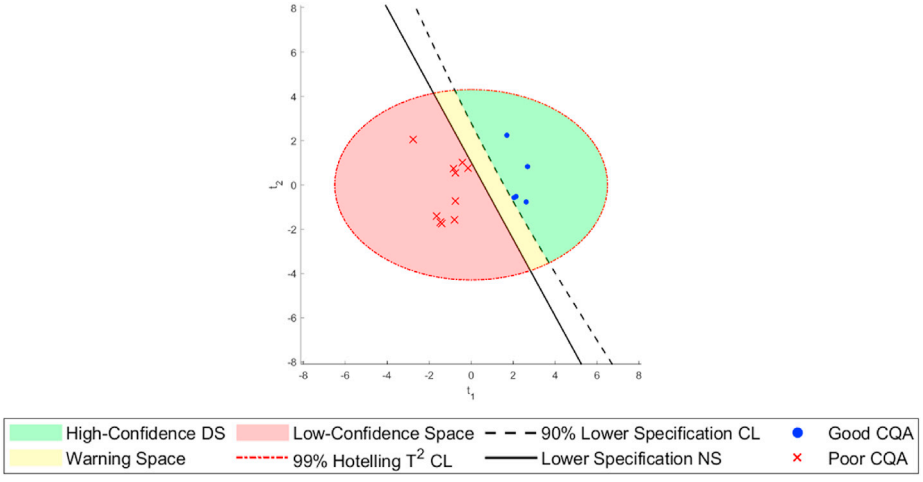


Fig. 18. Third industrial case study (scenario 3i): Graphical definition of the High-Confidence DS (Multivariate Raw Material Specification Region), Warning Space and Low-Confidence by showing exploiting data.

- The use of mathematical and statistical models as a way to define such raw material specifications by linking them with specification limits for CQAs of the final product.
- It allows a frequentist probabilistic interpretation. The multivariate raw material region is expected to produce product with CQAs within specification limits with a confidence level equal or higher than $(1-\alpha)\times 100$.
- It provides the analytical definition of the limits of the multivariate raw material specifications.
- It provides a strategy where RTR (for batches in the multivariate raw material specification region or High-Confidence Design Space), or end-product testing (for batches in the Warning Space) can be used as needed.

The methodology presented here is illustrated using three industrial case studies.

Our approach assumes that if process variations are correlated with raw material properties due to control actions through manipulated variables, they will remain in place in the future. However, several works have already emphasized that such control actions could be improved in order to compensate for some of the raw materials variability [3,5,6]. Hence, wider raw materials specifications can be used if an effective process control system attenuating most raw material variations is implemented. In this sense, future research is needed to model the relationships between not only raw materials properties and CQAs, but also process conditions. The proposed approach provides a good starting point when raw material multivariate specifications, defined analytically by considering a probabilistic approach, are precisely linked with CQAs of the final product. Finally, the logical extension of defining raw material specifications is to measure how far suppliers can consistently operate inside such specifications in order to select them by means of multivariate capability indices. This will deserve future work.

Author statement

Alberto and Carl: **Conceptualization,** Ideas; formulation or evolution of overarching research goals and aims. Alberto, Carl, Daniel and Joan: **Methodology,** Development or design of methodology; creation of models. Daniel and Joan: **Software,** Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components. Carl, Daniel and Joan: **Validation,** Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs. Daniel and Joan: **Formal analysis,** Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data. Alberto, Carl, Daniel and Joan: **Investigation,** Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection. Alberto, Carl, Daniel and Joan: **Resources,** Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools. Daniel and Joan: **Data Curation,** Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse. Joan: **Writing – original draft preparation,** Creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation). Alberto, Carl and Daniel: **Writing – review and editing,** Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages. Daniel and Joan: **Visualization,** Preparation, creation and/or presentation of the published work, specifically visualization/data presentation. Alberto and Carl: **Supervision,** Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team. Alberto: **Project administration,** Management and coordination responsibility for the research activity planning and execution. Alberto: Funding **acquisition,** Acquisition of the financial support for the project leading to this publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially funded by the Valencian Regional Government: Dirección General de Ciencia e Investigación (AICO/2021/111), the Spanish Ministry of Economy, Industry and Competitiveness (DPI2017-82896-C2-1-R), and the European Social Fund (ACIF/2018/165).

Appendix A. Specification confidence limits when considering the prediction uncertainty for the *l*-th CQA

Let $\boldsymbol{\tau}^{NS}$ be a vector of scores belonging to the NS associated to either the upper or lower specification limit for the l-th CQA (y_l^{SL}), and $\boldsymbol{\tau}^{SCL}$ the vector of scores belonging to such specification confidence limit. Thus, the vector defined by $(\boldsymbol{\tau}^{NS}-\boldsymbol{\tau}^{SCL})$ is orthogonal to the NS (i.e., as vector $\mathbf{v}_l$ defining the hyperplane of the NS (Eq. (17)), and the direction depends on whether it refers to y_l^{LSL} ($\boldsymbol{\tau}^{LSCL}$) or y_l^{USL} ($\boldsymbol{\tau}^{USCL}$):

$$\left(\boldsymbol{\tau}^{NS}-\boldsymbol{\tau}^{SCL}\right)=\mathbf{v}_l\cdot\lambda \tag{A.1}$$

where λ is a scalar that can be negative or positive depending on it referring to the $\boldsymbol{\tau}^{LSCL}$ or $\boldsymbol{\tau}^{USCL}$, respectively. Besides, the lower (if y_l^{LSL} is considered) or upper (if y_l^{USL} is considered) endpoint of its prediction interval must match the specification limit.

$$y_l^{SL}=\boldsymbol{q}_l^T\cdot\boldsymbol{\tau}^{NS} \tag{A.2}$$

$$y_l^{SL}=\boldsymbol{q}_l^T\cdot\boldsymbol{\tau}^{LSCL}-t_{\alpha/2\ ,N-df}\cdot s_{e_l^{LSCL}} \tag{A.3}$$

$$y_l^{SL}=\boldsymbol{q}_l^T\cdot\boldsymbol{\tau}^{USCL}+t_{\alpha/2\ ,N-df}\cdot s_{e_l^{USCL}} \tag{A.4}$$

By substitution and reorganization of either Eq. (A.1), Eq. (A.2) and Eq. (A.3), or Eq. (A.1), Eq. (A.2) and Eq. (A.4) the same quadratic equation is defined (Eq. (A.5)).

$$s^2_{e_l^{SCL}} \cdot t^2_{\alpha/2,N-df} = \left(\boldsymbol{q}_l^T \cdot \mathbf{v}_l\right)^2 \cdot \lambda^2 \tag{A.5}$$

Notice that there will be a negative solution attributed to the y_l^{LSL} and a positive solution attributed to the y_l^{USL}. Furthermore, since $s^2_{e_l^{SCL}}$ depends on the leverage of the unknown $\boldsymbol{\tau}^{SCL}$ (either $\boldsymbol{\tau}^{LSCL}$ or $\boldsymbol{\tau}^{USCL}$) according to Eq. (10) and Eq. (11), it must can be expressed as a function of $\boldsymbol{\tau}^{NS}$ by taking into account Eq. (A.1) as follows:

$$s^2_{e_l^{SCL}} = SE_l^2 \cdot \left(\mathbf{v}_l^T \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{v}_l \cdot \lambda^2 - 2 \cdot \mathbf{v}_l^T \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}^{NS} \cdot \lambda + 1 + {}^1/_N + \boldsymbol{\tau}^{NS^T} \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}^{NS}\right) \tag{A.6}$$

Substituting Eq. (A.6) in Eq. (A.5):

$$SE_l^2 \cdot \left(\mathbf{v}_l^T \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{v}_l \cdot \lambda^2 - 2 \cdot \mathbf{v}_l^T \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}^{NS} \cdot \lambda + 1 + {}^1/_N + \boldsymbol{\tau}^{NS^T} \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}^{NS}\right) \cdot t^2_{\alpha/2,N-df} = \left(\boldsymbol{q}_l^T \cdot \mathbf{v}_l\right)^2 \cdot \lambda^2 \tag{A.7}$$

and reorganizing terms:

$$a \cdot \lambda^2 + b \cdot \lambda + c = 0 \tag{A.8}$$

where:

$$a = SE_l^2 \cdot \mathbf{v}_l^T \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \mathbf{v}_l \cdot t^2_{\alpha/2,N-df} - \left(\boldsymbol{q}_l^T \cdot \mathbf{v}_l\right)^2$$

$$b = -SE_l^2 \cdot 2 \cdot \mathbf{v}_l^T \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}^{NS} \cdot t^2_{\alpha/2,N-df} \tag{A.9}$$

$$c = SE_l^2 \cdot \left(1 + {}^1/_N + \boldsymbol{\tau}^{NS^T} \cdot (\mathbf{T}^{\mathrm{T}} \cdot \mathbf{T})^{-1} \cdot \boldsymbol{\tau}^{NS}\right) \cdot t^2_{\alpha/2,N-df}$$

The values of λ that satisfy Eq. (A.8) are the solutions of a quadratic equation and, as commented above, there will be a positive and a negative one. Besides, it is known that c is positive given terms that define it. For all this, it can be deduced that the quadratic function is concave down (i.e., the second derivative is negative) and, consequently, a must be negative. Because a is negative and c is positive, it is determined that the discriminant $(b^2 - 4 \cdot a \cdot c)$ is positive and, therefore, there are two distinct roots as follows:

$$\lambda_1 = \frac{-b + \sqrt{b^2 - 4 \cdot a \cdot c}}{2 \cdot a} \lambda_2 = \frac{-b - \sqrt{b^2 - 4 \cdot a \cdot c}}{2 \cdot a} \tag{A.10}$$

where both of them are, by definition, real numbers. Since the root of the discriminant is higher than b and a is negative, it is deduced that λ_1 is negative (it refers to y_l^{LSL}) and λ_2 is positive (it refers to y_l^{USL}). Thus, Eq. (A.11) shows the analytical expression of the specification confidence limits when considering the prediction uncertainty.

$$\boldsymbol{\tau}^{LSCL} = \boldsymbol{\tau}^{NS} - \mathbf{v}_l \cdot \lambda_1 \quad \boldsymbol{\tau}^{USCL} = \boldsymbol{\tau}^{NS} - \mathbf{v}_l \cdot \lambda_2 \tag{A.11}$$

References

[1] C. Duchesne, J.F. MacGregor, Establishing multivariate specification regions for incoming materials, J. Qual. Technol. 36 (2004) 78–94, https://doi.org/10.1080/00224065.2004.11980253.

[2] J.A. De Smet, Development of Multivariate Specification Limits Using Partial Least Squares Regression, McMaster University, 1993.

[3] S. García-Muñoz, S. Dolph, H.W. Ward, Handling uncertainty in the establishment of a design space for the manufacture of a pharmaceutical product, Comput. Chem. Eng. 34 (2010) 1098–1107, https://doi.org/10.1016/j.compchemeng.2010.02.027.

[4] S. García-Muñoz, Establishing multivariate specifications for incoming materials using data from multiple scales, Chemometr. Intell. Lab. Syst. 98 (2009) 51–57, https://doi.org/10.1016/j.chemolab.2009.04.008.

[5] J.F. MacGregor, Z. Liu, M.J. Bruwer, B. Polsky, G. Visscher, Setting simultaneous specifications on multiple raw materials to ensure product quality and minimize risk, Chemometr. Intell. Lab. Syst. 157 (2016) 96–103, https://doi.org/10.1016/j.chemolab.2016.06.021.

[6] K. Azari, J. Lauzon-Gauthier, J. Tessier, C. Duchesne, Establishing multivariate specification regions for raw materials using SMB-PLS, IFAC-PapersOnLine 48 (2015) 1132–1137, https://doi.org/10.1016/j.ifacol.2015.09.120.

[7] A. Paris, C. Duchesne, É. Poulin, Establishing multivariate specification regions for incoming raw materials using projection to latent structure models: comparison between direct mapping and model inversion, Front. Anal. Sci. 1 (2021) 1–15, https://doi.org/10.3389/frans.2021.729732.

[8] P. Facco, F. Dal Pastro, N. Meneghetti, F. Bezzo, M. Barolo, Bracketing the design space within the knowledge space in pharmaceutical product development, Ind. Eng. Chem. Res. 54 (2015) 5128–5138, https://doi.org/10.1021/acs.iecr.5b00863.

[9] D. Palací-López, P. Facco, M. Barolo, A. Ferrer, New tools for the design and manufacturing of new products based on Latent Variable Model Inversion, Chemometr. Intell. Lab. Syst. 194 (2019), https://doi.org/10.1016/j.chemolab.2019.103848.

[10] ICH Harmonised Tripartite, Guidance for Industry Q8(R2), Pharmaceutical Development, 2009.

[11] P.J. Whitcomb, M.J. Anderson, Using DOE with tolerance intervals to verify specifications, in: 11th Annu. ENBIS Conf., 2011.

[12] J.J. Peterson, M. Yahyah, A bayesian design space approach to robustness and system suitability for pharmaceutical assays and other processes, Stat. Biopharm. Res. 1 (2009) 441–449, https://doi.org/10.1198/sbr.2009.0037.

[13] G. Bano, P. Facco, F. Bezzo, M. Barolo, Probabilistic Design space determination in pharmaceutical product development: a Bayesian/latent variable approach, AIChE J. 64 (2018) 2438–2449, https://doi.org/10.1002/aic.16133.

[14] E. del Castillo, M.S. Reis, Bayesian predictive optimization of multiple and profile response systems in the process industry: a review and extensions, Chemometr. Intell. Lab. Syst. 206 (2020), 104121, https://doi.org/10.1016/j.chemolab.2020.104121.

[15] E. Rozet, P. Lebrun, B. Debrus, B. Boulanger, P. Hubert, Design Spaces for analytical methods, Trends Anal. Chem. 42 (2013) 157–167, https://doi.org/10.1016/j.trac.2012.09.007.

[16] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (1988) 211–228, https://doi.org/10.1002/cem.1180020306.

[17] S. Wold, M. Sjostrom, L. Eriksson, PLS-Regression - a basic tool of chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1.

[18] A. Ferrer, Discussion of " A review of data science in business and industry and a future view " by Grazia Vicario and Shirley Coleman, Appl. Stoch Model Bus. Ind. 36 (2020) 23–29, https://doi.org/10.1002/asmb.2516.

[19] J.F. MacGregor, M.J. Bruwer, I. Miletic, M. Cardin, Z. Liu, Latent variable models and big data in the process industries, IFAC-PapersOnLine 48 (2015) 520–524, https://doi.org/10.1016/J.IFACOL.2015.09.020.
[20] T. Kourti, J.F. MacGregor, Multivariate SPC methods for process and product monitoring, J. Qual. Technol. 28 (1996) 409–428, https://doi.org/10.1080/00224065.1996.11979699.
[21] J.F. MacGregor, P. Nomikos, Multivariate SPC charts for batch monitoring processes, Technometrics 37 (1995) 41–59, https://doi.org/10.2307/1269152.
[22] N.D. Tracy, J.C. Young, R.L. Mason, Multivariate control charts for individual observations, J. Qual. Technol. 24 (1992) 88–95, https://doi.org/10.1080/00224065.1992.12015232.
[23] A. Ferrer, Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process, Qual. Eng. 19 (2007) 311–325, https://doi.org/10.1080/08982110701621304.
[24] G. Bano, P. Facco, N. Meneghetti, F. Bezzo, M. Barolo, Uncertainty back-propagation in PLS model inversion for design space determination in pharmaceutical product development, Comput. Chem. Eng. 101 (2017) 110–124, https://doi.org/10.1016/j.compchemeng.2017.02.038.
[25] K. Faber, B.R. Kowalski, Prediction error in least squares regression: further critique on the deviation used in the Unscrambler, Chemometr. Intell. Lab. Syst. 34 (1996) 283–292, https://doi.org/10.1016/0169-7439(96)00022-6.
[26] L. Zhang, S. Garcia-Munoz, A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective, Chemometr. Intell. Lab. Syst. 97 (2009) 152–158, https://doi.org/10.1016/j.chemolab.2009.03.007.
[27] C.M. Jaeckle, J.F. MacGregor, Product design through multivariate statistical analysis of process data, AIChE J. 44 (1998) 1105–1118, https://doi.org/10.1016/0098-1354(96)00182-2.
[28] C.M. Jaeckle, J.F. MacGregor, Industrial applications of product design through the inversion of latent variable models, Chemometr. Intell. Lab. Syst. 50 (2000) 199–210, https://doi.org/10.1016/S0169-7439(99)00058-1.
[29] F. Yacoub, J.F. MacGregor, Product optimization and control in the latent variable space of nonlinear PLS models, Chemometr. Intell. Lab. Syst. 70 (2004) 63–74, https://doi.org/10.1016/J.CHEMOLAB.2003.10.004.
[30] S. García-Muñoz, T. Kourti, J.F. MacGregor, F. Apruzzese, M. Champagne, Optimization of batch operating policies. Part I. Handling multiple solutions, Ind. Eng. Chem. Res. 45 (2006) 7856–7866, https://doi.org/10.1021/ie060314g.
[31] E. Tomba, M. Barolo, S. García-Muñoz, General framework for latent variable model inversion for the design and manufacturing of new products, Ind. Eng. Chem. Res. 51 (2012) 12886–12900, https://doi.org/10.1021/ie301214c.
[32] E. Tomba, P. Facco, F. Bezzo, S. García-Muñoz, Exploiting historical databases to design the target quality profile for a new product, Ind. Eng. Chem. Res. 52 (2013) 8260–8271, https://doi.org/10.1021/ie3032839.
[33] D. Palací-López, J. Borràs-Ferrís, L. Thaise da Silva de Oliveria, Multivariate six sigma: a case study in industry 4.0, in: Processes vol. 8, 2020, pp. 1–20, https://doi.org/10.3390/pr8091119.
[34] S. García-Muñoz, J. Mercado, Optimal selection of raw materials for pharmaceutical drug product design and manufacture using mixed integer nonlinear programming and multivariate latent variable regression models, Ind. Eng. Chem. Res. 52 (2013) 5934–5942, https://doi.org/10.1021/ie3031828.
[35] G. Bano, P. Facco, F. Bezzo, M. Barolo, Probabilistic Design space determination in pharmaceutical product development: a Bayesian/latent variable approach, AIChE J. 64 (2018) 2438, https://doi.org/10.1002/aic.16133.
[36] S. Wold, M. Josefson, J. Gottfries, A. Linusson, The utility of multivariate design in PLS modeling, J. Chemom. 18 (2004) 156–165, https://doi.org/10.1002/cem.861.