Received: 29 November 2019 | Revised: 14 May 2020 | Accepted: 19 May 2020

DOI: 10.1002/qre.2697

SPECIAL ISSUE ARTICLE

WILEY

Communicating statistical conclusions of experiments to scientists

Martin Otava[1] | **Kalliopi Mylona**[2]

[1]Manufacturing and Applied Statistics, Janssen-Cilag s.r.o., Janssen Pharmaceutical Companies of Johnson & Johnson, Prague, Czech Republic

[2]Department of Mathematics, King's College London, London, UK

Correspondence
Kalliopi Mylona, Department of Mathematics, King's College London, London WC2R 2LS, UK.
Email: kalliopi.mylona@kcl.ac.uk

Abstract

The manuscript introduces a framework for presenting the results of the statistical analysis of experiments with multiple responses and multiple factors. We propose a utilisation of factors scaling to enable a transformation that combines main effects, quadratic effects and interactions into a meaningful summary that allows the scientist/experimenter to immediately recognise the most influential factors for a given response. The framework does not replace the thorough evaluation of the results but provides a clear high-level summary of the relative importance of findings. The visualisation of such factor importance, using intensity heatmaps, allows the immediate understanding of the results across multiple responses that efficiently guides a following detailed analysis of certain responses and factors and contributes in designing subsequent experiments. The methodology is applied to a real industrial experiment and to a simulated data set with a larger number of responses and factors.

KEYWORDS

data visualisation, experimental design, screening experiments

1 | INTRODUCTION

Experimentation underpins the most important scientific and technological advances. An efficient statistical design and a correct analysis of experiments have the potential to generate important new domain and application knowledge in numerous disciplines, such as in agroindustry, biopharmaceuticals, food science, material science, medicine, biology and genetics, and savings in time and money. The main aim of a statistically designed experiment is the identification and estimation of an appropriate quantitative relationship between the controllable input variables (factors) and the observed outputs (responses). Such a relationship, or model, can be used to gain scientific understanding on the causal links between systems inputs and outputs, predict the response at unobserved inputs and optimise the system by finding values of the factors that match a target output.

In industry, the life sciences and the physical sciences, the 'big data' revolution, enabled through modern instrumentation for carrying out experiments and automatically collecting data, brings the need for efficient visualisation techniques for the clear communication of the statistical results to the applied scientists, for example, engineers, medical doctors, psychologists and so forth. Results summarisation and consequent visualisation is crucial for (i) spotting patterns, (ii) external memorisation, (iii) stimulation/evaluation of hypothesis, (iv) saving of time and (v) ability to give people the whole picture.[1] In this work, we develop an exploratory summarisation and visualisation technique that gives the overall picture of the data analysis in experiments with multiple responses and multiple factors.

The method can be broadly applied to any kind of experiment that uses a multifactorial design, with multiple responses. For example, in response surface methodology, a sequential strategy of experimentation, statistical modelling and optimisation, the experiment aims to study and understand the relationship between the controllable factors and the response. Typically, such experiment only considers a few factors. If there are many factors that may be of importance, a fractional factorial, a definitive screening or a supersaturated design should be used first in a screening experiment to identify the important factors.[2-4] The experiments are usually designed assuming that the treatments (combinations of factor levels) are completely randomised to the experimental units. However, just as with any designed experiment, there may be some structure in the experimental units that should account for known nuisance factors that affect the response and should be used for blocking. We refer for more details on the different types of experiments to the *Handbook of Design and Analysis of Experiments*.[5]

The large number of factors in combination with the large number of responses and the further extensions of the traditional design framework, as suggested in the previous paragraph, make the comparison and presentation of the analysis results across many responses cumbersome, especially when a model selection needs to be applied and the final models differ across responses. The methodology proposed in this manuscript is a summarisation framework and graphical representation of the results of the statistical analysis of data from experiments with multiple responses. The approach, based on transforming parameters and summarising them in an interpretable way, is accompanied by intensity heatmaps that allow the practitioner just with a simple look at the graph to get an overall information on which factors are the most important ones for a given response, taking into account the main, the quadratic and the interaction effects. Additionally, the practitioner will be able to see easily how often certain factors appear to be important across multiple responses. This immediate understanding contributes to a correct combination of the statistical output with the scientific knowledge leading into the design of subsequent experiments and the modelling of the responses under study. However, this framework definitely does not replace the deeper evaluation of the individual models, responses and factors. It allows for smooth communication of the results and enables a statistician to focus on a detailed analysis of the responses and factors with the strongest effects. Finally, the results would need to be evaluated against scientific context and knowledge, taking into account the magnitudes of the effects within the context of each response as well as the relative importance of the responses themselves.

In Section 2, we present the motivating examples, in which our method was successfully applied. The first study comes from the literature and has a rather small number of factors involved. Its main role is illustrative, to provide a full understanding of the methodology, but even then, we can see a benefit when combining quadratic effects and interactions in a concise framework. The second example represents a screening study based on a definitive screening design with a much higher number of factors that would benefit largely from our visualisation framework. This motivating example was inspired by a real pharmaceutical case study performed by the first author in which the proposed technique was successfully applied. Unfortunately, the data set is proprietary and cannot be shown publicly in any form; hence, a simulated data set with the matching structure is used instead. Note that both case studies use as input the effect estimates, computed after a model selection step has been applied. This is typical for screening designs, when fitting the full model is rarely possible. However, if estimating the full model is feasible, the framework can be analogously applied with considerable benefit, when considering all the quadratic effects and the interactions. In Section 3, we describe the proposed methodology and possible extensions. In Section 4, we present the results from the application of our method to the real-life experiment and to the simulated data set. We provide the discussion in Section 5. In the supporting information, we provide the computer code in R[6] for the implementation of our technique to both case studies and a compilation of the R code into the HTML report.

2 | DATA

2.1 | The pastry dough experiment

The first case study is about an experiment performed to improve the quality of a pastry dough. A non-orthogonally blocked response surface design, with seven blocks of four runs, was used by Trinca and Gilmour.[7] The data set represents an experiment aimed to improve the quality of a pastry dough by adjusting three factors: the flow rate x_1, the moisture content x_2 and the screw speed x_3. The four evaluated responses[8,9] are a longitudinal expansion index y_1, a cross-sectional expansion index y_2 and two measures of light transmission in two different bands of the spectrum, y_3 and y_4.

The estimates of the fixed effects for the purpose of this manuscript, which is the graphical representation of the results of a statistical analysis, were retrieved from the literature.[10] These estimates were obtained using generalised least squares

	y_1	y_2	y_3	y_4
Intercept	11.8464	4.4655	13.1249	77.0781
x_1	0.9944	0	−0.1894	0
x_2	−1.4556	−0.6233	0.8783	−0.2844
x_3	0.7556	0.3256	−0.7094	0.2294
x_1x_2	0	0	0	0
x_1x_3	0	0	0	0
x_2x_3	0	0	0	−0.3706
x_1^2	0	0	0	0
x_2^2	2.0222	0.6870	−0.5904	0
x_3^2	0	0	0	0.4185

TABLE 1 Pastry dough data: The estimated coefficients of the variables based on generalised least squares and backward elimination

(GLS) and backward model selection. The usual way to analyse data from this type of experiments is to fit the linear mixed model. This is usually done by using GLS for the fixed effects combined with restricted maximum likelihood (REML) estimation of the variance components.[9,11] The estimated effects used can be seen in Table 1.

2.2 | Simulated data

The main benefit of the data set above is to simply explain the proposed methodology. However, studies with a large number of factors and responses will benefit the most from fast exploratory visual tools. The first author has applied the method on a definitive screening experiment evaluating 20 responses on 11 factors (that represented various settings of the process equipment). The simulated modelling output mimicking the structure of this real study will be used as a second case study in the manuscript. Because the visualisation and not a practical interpretation of the specific effects is the subject of the demonstration, there is no practical difference in presenting simulated numerical values for effect sizes instead of the real data.

The definitive screening designs,[3] unlike most of the screening designs proposed in the literature, include factors in three levels instead of two. In this way, it is possible to detect quadratic effects even at the initial stage of the experimentation with just a few runs. Specifically, the minimal definitive screening design with m factors at three levels have $2m - 1$ runs, while at least four extra runs are recommended for better estimate of the residual error.[12] The final design applied to this case study had 30 runs due to the presence of an additional categorical factor.

The simulated output was obtained as follows. For simplicity, we do not initially simulate the data and then the whole model selection process, but we simulate directly the hypothetical result of the model fitting and the model selection procedure. Hence, we directly simulate the set of factors (and their interactions and quadratic effects) that would be selected for each response by the hypothetical model selection procedure, that is, the final linear model for each response.

Therefore, the simulation process has initially randomly selected the number of selected main effects for a given response, then simulated the effect sizes from a normal distribution with zero mean and variance equal to 25. Quadratic and interaction effects were then simulated under the strong heredity principle, that is, a quadratic effect could only occur, if the corresponding main effect was non-zero and an interaction could only occur, if both respective main effects were non-zero. The quadratic effect and interactions were considered present with probability of 0.2 and 0.3, respectively. Finally, all the main and quadratic effects for factors 2, 6 and 11 were multiplied by 5 to create factors with higher relative importance. The approach was repeated for each response variable.

The simulation has resulted, in total (summed over all responses), in 118 main effects (out of 220 possible), 29 quadratic effects and 105 interaction effects. The simulation result comprises 20 distinct linear models, one per response. The models are not shown here due to their size, and the computer code to generate this data is submitted as supporting information for the readers' convenience and reproducibility. Note that we have created only single simulated data set to demonstrate the value of our visualisation method.

3 | THE METHODOLOGY

3.1 | The framework

Let us assume that we have implemented a design aiming at describing the relationship of N factors with each of K responses of interest. The starting point of the visualisation approach is the result of modelling. The final model can be the full model including all the effects of interest or some preliminary selection can be applied to reduce the number of

model terms. Further, we assume that the quadratic effects and the second-order interactions are the highest order effects of interest. The framework, however, can be extended to higher orders in a straightforward manner. Finally, we assume that the factors are centred and transformed in such a way that factors' levels with values of –1 and 1 are of the main interest. In practice, that would often mean that the factors would be transformed to have range between –1 and 1, but there may be exceptions, for example, in central composite designs (CCDs), the axial observations lie beyond the –1 and 1 boundaries. Even then, the shift of –1 and 1 is interpretable and relevant from a practical perspective.

Denote the factors as $X_1, \ldots, X_N$ and the responses as $Y_1, \ldots, Y_K$. Then, the full model considered for any given response Y_k with $k = 1, \ldots, K$ would be

$$E(Y_k) = \alpha + \sum_{n=1}^{N} \beta_{nk} X_n + \sum_{n=1}^{N} \gamma_{nk} X_n^2 + \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} \delta_{nmk} X_n X_m. \tag{1}$$

In Equation (1),

- α represents the intercept, that is, the mean response value if all the factors are set to zero,
- β_{nk} represents the main effect of X_n ($n = 1, \ldots, N$) in the model for Y_k,
- γ_{nk} represents the quadratic effect of X_n ($n = 1, \ldots, N$) in the model for Y_k, and
- δ_{nmk} represents the interaction between X_n and X_m ($n = 1, \ldots, N-1, m = n+1, \ldots, N$) in the model for Y_k.

Note that the effects listed above are defined on transformed factors, with a range of interest between –1 and 1 and the centre at zero. This means that the effects reflect changes in the response when a factor is changed by one at the transformed scale, that is, by half of the range at its original scale.

3.2 | The maximal possible influence

The aim of this manuscript is to provide a visual tool that allows the importance of the factors to be assessed at global perspective, that is, simultaneously for all $Y_1, \ldots, Y_K$. An intensity heatmap is the plot type fitting such aim. Let us assume the matrix with the responses $Y_1, \ldots, Y_K$ in its columns and the factors $X_1, \ldots, X_N$ in its rows and each cell value being equal to some measurable quantity representing the strength of the relationship between a specific factor and a response. The heatmap simply assigns a colour to the observed quantity in each cell of the matrix, with a colour intensity proportional to the value of the quantity. For example, if our focus would be on main effects only, we could represent the relationship between all responses and factors with the following matrix:

$$\begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_N \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \ldots & Y_K \end{array} \\ \begin{pmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1K} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2K} \\ \vdots & \vdots & \ldots & \vdots \\ \beta_{N1} & \beta_{N2} & \ldots & \beta_{NK} \end{pmatrix} \end{array}.$$

The visualisation of such a matrix is straightforward via the heatmap. However, considering only the main effects may be misleading in the presence of strong quadratic effects and interactions. Another metric to compare the relative contribution of factors is needed. Hence, we suggest a framework that summarises both types of second-order effects by one metric value per factor and response combination.

Let us define the 'maximal possible influence', MPI_{nk}, of the factor X_n on the response Y_k as *the maximal absolute change in Y_k that can be induced by increasing the X_n by one, within the range* $[-1, 1]$*, for a certain starting value of X_n and some favourable setting of the other factors $X_1, \ldots, X_{n-1}, X_{n+1}, \ldots, X_N$* . As we will see later, *a certain starting value* is necessary for adding the quadratic effects and *a favourable setting of other factors* for adding the interactions.

Let us consider a quadratic effect first. We define the numbers $b > 0$ and $c > 0$ such that for a factor X_n and a response Y_k, the absolute value of the main effect is $|\beta_{nk}| = b$ and the absolute value of the quadratic effect is $|\gamma_{nk}| = c$. We can see all possible combinations of positive and negative main and quadratic effects in Table 2. Given that X_n only takes values in the interval $[-1, 1]$, we can calculate the level change by one, between the zero and the boundaries. As we can see in the table, the higher of the two effects, regardless of the sign of the main and quadratic effect, can be always obtained using the following equation:

$$sign(\beta_{nk}) \cdot [|\beta_{nk}| + |\gamma_{nk}|]\,. \tag{2}$$

β_{nk}	γ_{nk}	$X_n = -1$	$X_n = 0$	$X_n = 1$	–1 to 0	0 to 1	max effect
b	c	$-b+c$	0	$b+c$	$b-c$	$b+c$	$b+c$
b	$-c$	$-b-c$	0	$b-c$	$b+c$	$b-c$	$b+c$
$-b$	c	$b+c$	0	$-b+c$	$-b-c$	$-b+c$	$-(b+c)$
$-b$	$-c$	$b-c$	0	$-b-c$	$-b+c$	$-b-c$	$-(b+c)$

TABLE 2 Quadratic effects summarisation example

Note: We use the equation $\beta_{nk}X_n + \gamma_{nk}X_n^2$.

We proceed analogously with the interaction effects. The interaction term δ_{nmk} represents the change in influence of factor X_n on response Y_k conditional on the value of factor X_m. When X_n is increased by one, the response Y_k will change by $\beta_{nk} - \delta_{nmk}$ if $X_m = -1$ and by $\beta_{nk} + \delta_{nmk}$ if $X_m = 1$ (assuming no quadratic effect). Hence, a value of X_m can be always selected such that the sign of β_{nk} matches either δ_{nmk} or $-\delta_{nmk}$. The choice of X_m to match the sign of interaction effect with the sign of the main effect of X_n reflects the most favourable setting towards the potential effect of X_n. Naturally, the concept above can be extended to all the interactions between X_n and the rest of the factors $X_1, \ldots, X_{n-1}, X_{n+1}, \ldots, X_N$. In summary, we select the most favourable setting for the remaining factors such as to maximise the effect of the factor X_n. Therefore, regardless of the signs of the main effect and interaction effects, we can obtain the maximal effect in absolute sense as

$$sign(\beta_{nk}) \cdot \left[|\beta_{nk}| + \sum_{m=n+1}^{N} |\delta_{nmk}| + \sum_{m=1}^{n-1} |\delta_{mnk}| \right]. \tag{3}$$

Note that Equations (2) and (3) have the same structure. Therefore, the maximal possible influence of X_n at Y_k, as defined above, can be calculated as

$$MPI_{nk} = sign(\beta_{nk}) \cdot \left[|\beta_{nk}| + |\gamma_{nk}| + \sum_{m=n+1}^{N} |\delta_{nmk}| + \sum_{m=1}^{n-1} |\delta_{mnk}| \right]. \tag{4}$$

Therefore, the final matrix looks as follows:

$$\begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_N \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & \ldots & Y_K \end{array} \\ \begin{pmatrix} MPI_{11} & MPI_{12} & \ldots & MPI_{1K} \\ MPI_{21} & MPI_{22} & \ldots & MPI_{2K} \\ \vdots & \vdots & \ldots & \vdots \\ MPI_{N1} & MPI_{N2} & \ldots & MPI_{NK} \end{pmatrix} \end{array},$$

which can be plotted directly within the intensity heatmap framework.

However, Equation (4) will be practically useful only if the scales of the responses are comparable. Otherwise, the colour intensity will be entirely dominated by the responses with the higher absolute effects. Besides invalid comparison across responses, the visual reading of heatmap will be compromised. Therefore, we recommend scaling the results using one of the following two options. The first would be normalising the responses themselves to obtain comparable means and variances. After such transformation, Equation (4) can be applied as is.

The framework may also be used on the model output without access to raw data or in cases where refitting the models is impossible or when the experimenter prefers to fit the model using the original responses. In such case, an alternative approach could be applied by standardisation within each response Y_k towards the highest (in sense of absolute value) observed $MPI_{nk}, n = 1, \ldots, N$. In other words, for each response (column), we visualise the MPI_{nk} relative to the maximum $|MPI_{nk}|$ achieved for that response, that is, we visualise the 'relative importance' of the factors within each Y_k:

$$MPI_{nk}^{std} = MPI_{nk} / \left(max_{n=1,\ldots,N} |MPI_{nk}| \right). \tag{5}$$

Additionally, in the screening context, the sign of the effect often is not of the utmost importance, but rather the selection of factors with strong effects in either direction that will be further evaluated in following stages of the process characterisation. Therefore, instead of the MPI_{nk} or the MPI_{nk}^{std}, the absolute value of the respective metrics, $|MPI_{nk}|$ or the $|MPI_{nk}^{std}|$, can be computed and visualised using a heatmap. Reducing to only a one-directional intensity scale will help in a visual evaluation of the intensity ranges.

3.3 | Considerations for follow-up experiment

Screening designs often represent the first part of a two-stage approach, intended to select the factors that will be carefully studied in a second stage experiment. In practice, the factors included in the screening stage might not be selected solely based on their inferential properties (e.g., statistical significance), but also on their actual practical relevance. Hence, it might be useful to see how the MPI_{nk} of a given factor changes when other factors (not planned to be investigated further) would be fixed to the centre value (note that concept can be extended to any other value than centre). For example, it may happen that X_n has a relatively small interaction with multiple other factors, leading to practically relevant MPI_{nk}, whereas when all the other parameters are set to centre values (leading to zero contribution of the respective interactions), the actual potential of X_n is reduced considerably. From this perspective, it may be useful to explore how the heatmap of the potential effects changes if certain factors are excluded from the final set. In terms of the MPI_{nk}, the interactions with excluded factors will not be added into the sum.

It is often the case that the ranges of the factors are reduced (or extended) during the follow-up studies. The reasons for a range reduction would typically include higher cost or run time at extremes leading to a demanding implementation of the experiments. Hence, if we are interested in the potential of the factors to influence the response in a future study, we should take into account the range adjustments. Then, the impact on the interaction terms should be addressed, because the MPI_{nk} as defined in Equation (4) assumes the whole range to be between –1 and 1 for all the factors, so the interacting factors can be set to the most favourable values of –1 or 1. Hence, if we assume a narrower range of the factors, the interaction terms δ_{nmk} have to be scaled down appropriately (i.e., if a range between –0.7 and 0.7 is considered for factor X_n, we should multiply all interactions with this factor by 0.7). Generalisation towards non-symmetrical ranges is straightforward.

Note that the considerations above apply as well for the case of extending the range of the factors' levels in the follow-up studies. An example of such a case may be the augmentation of a screening factorial design towards a central composite design, where levels of factors beyond the range of –1 to 1 are explored. However, scaling up the MPI_{nk} for extended ranges

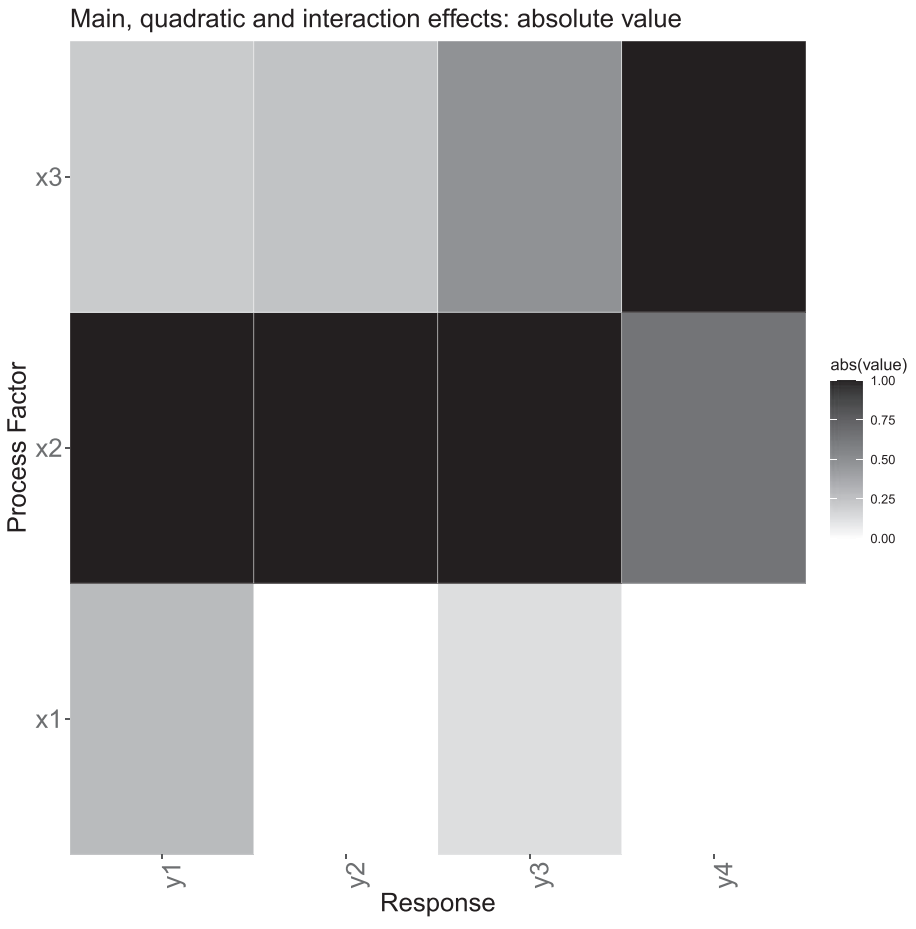

FIGURE 1 Pastry dough data: $|MPI^{std}_{nk}|$

relies on the assumption that the estimated effects and the relationship extrapolate outside the studied factor ranges. Such assumption may be often inappropriate.

4 | RESULTS

The ideas described in the previous section have been applied to both the literature case study and the simulated data set, described in Section 2.

4.1 | The pastry dough experiment

The MPI_{nk} for the pastry dough data as defined in Equation (4), that is, the basic version without any formatting scaling or further adjustments is not shown in this manuscript; however, for a coloured version of this plot, see the supporting information. Given that the responses have different scales, we proceed directly with the adjusted $|MPI^{std}_{nk}|$, by considering the absolute value in Equation (5), that allows us to compare the relative importance of factors across responses. To provide a concrete example, let us explicitly calculate MPI_{34}. We consider the factor x_3 and the response y_4 in the presence of the main effect, the quadratic effect and one interaction. See Table 1 for the estimates of the individual effects: the main effect $\beta_{34} = 0.2294$, the quadratic effect $\gamma_{34} = 0.3706$ and the interaction $\delta_{234} = 0.4185$. Following Equation (4), $MPI_{34} = sign(\beta_{34})\,[|\beta_{34}| + |\gamma_{34}| + |\delta_{234}|] = 1 \cdot [0.2294 + 0.3706 + 0.4185] = 1.0185 = |MPI_{34}|$. Analogously, we obtain $MPI_{14} = 0$ and $MPI_{24} = 0.655$. Hence, after scaling relatively to the maximal $|MPI_{n4}|$, the scaled effects become $MPI^{std}_{14} = 0$, $MPI^{std}_{24} = 0.655/1.10185 = 0.59445$ and $MPI^{std}_{34} = 1$.

The resulting heatmap for all $|MPI^{std}_{nk}|$, calculated analogously, is shown in Figure 1. The heatmap is interpreted as follows: relatively to other factors, the potential effect of x_2 seems to be the most important across y_1, y_2 and y_3. For y_4, x_2 is dominated by x_3. Factor x_1 does not seem to be (relatively to the effects of the other factors) very important, with exemption for y_1.

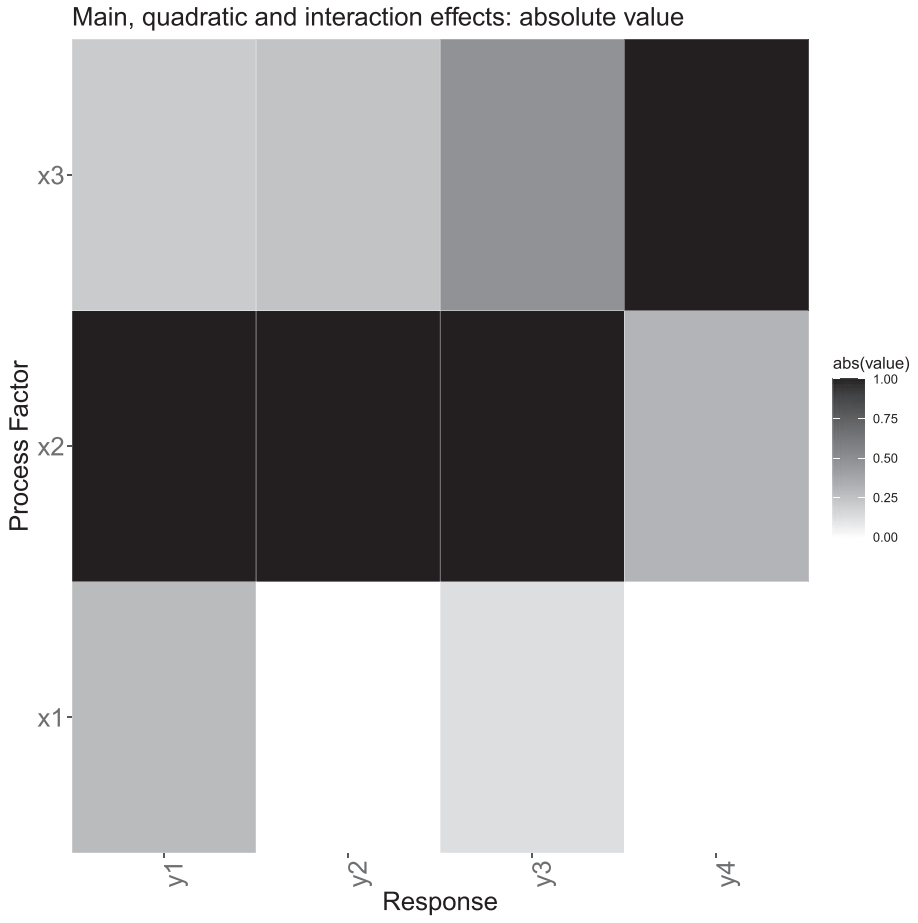

FIGURE 2 Pastry dough data: $|MPI^{std}_{nk}|$ with restriction of x_3 to $(-0.1, 0.1)$

Let us assume that we would decide to perform a follow-up optimisation study, where we would like to reduce the range of x_3 to $(-0.1, 0.1)$. A practical reason might be that the experiment was taking too long to perform at low screw speed while leading to a risk of equipment failure at high screw speed. Such a change would only impact the contribution of the interaction x_2x_3 for y_4 towards the MPI_{24}. As a matter of fact, the main and quadratic effects do not need to be adjusted, because we define the MPI_{nk} as a change when the factors are increased by one (in a transformed scale). Hence, the range adjustment would only affect the contribution of the corresponding interactions (that requires the most favourable value setting for x_3, which is affected by the range adjustment). Hence, we would expect the colour intensity of MPI_{24} to decrease, due to the reduced interaction term value. We can see that this is indeed the case in Figure 2.

4.2 | Simulated large study

The scaled absolute value results, that is, the $|MPI^{std}_{nk}|$ per response, for the simulated study are shown in Figure 3. As expected, factor 6 seems to be the most relevant overall (recall that it was chosen to have high relative importance during the simulation), whereas factors 2 and 11 are slightly less pronounced. As noted in Section 1, any final decision taken based on these findings needs to be confronted with scientific knowledge, for example, relative practical importance of various responses. For example, if responses 5 and 9 would be extremely important, whereas other responses are comparatively less relevant, then factor 10 could be deemed more important than factors 2, 6 and 11.

Let us assume that only six factors will be selected for further evaluation in the optimisation stage: factors, 1, 2, 4, 5, 6 and 11. Figure 4 demonstrates why it is important to explore the reduced set of factors before finalising the set. Factor 1 could have been selected due to its relative importance with respect to the response 7 or 14. However, after removing the interactions of the removed factors, it seems that factor 1 has limited ability to influence these responses. Additionally, in the final set, there is no factor that has been selected as non-zero for the response 9 (uniformly white column). It may merit discussion if there is added value to measure response 9 in further experiments or not.

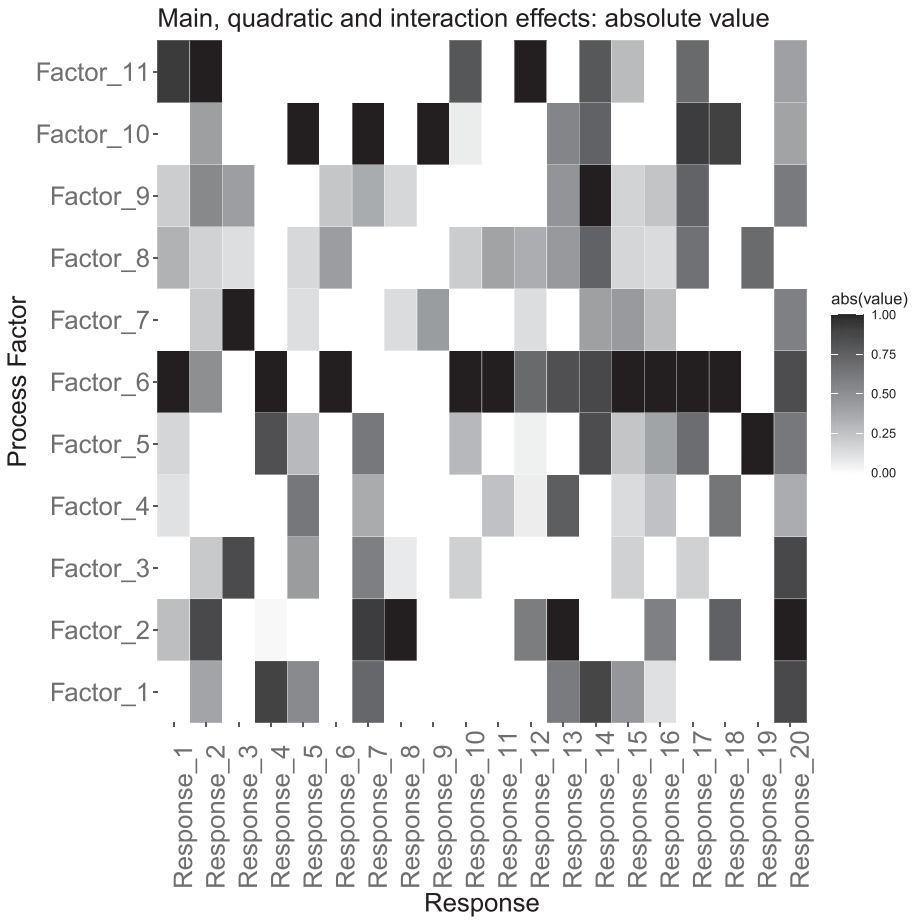

FIGURE 3 Simulated data: $|MPI^{std}_{nk}|$

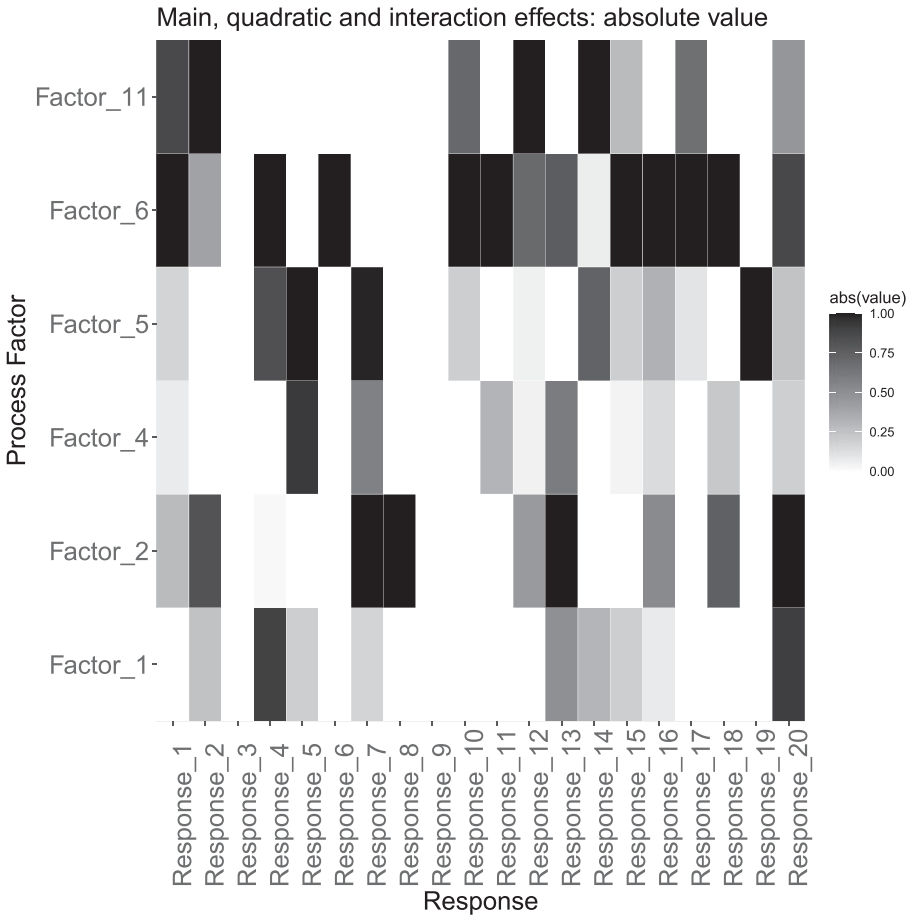

FIGURE 4 Simulated data: $|MPI_{nk}^{std}|$ after selection of factors $\{1, 2, 4, 5, 6, 11\}$ only

The most important feature of the presented framework is that it allows a fast exploration of the relative effects to efficiently focus an attention at deeper level towards the relevant choices of factor sets. It is extremely efficient for transmitting statistical results to applied scientists, such as engineers, pharmacists, medical doctors and so forth.

5 | DISCUSSION

The framework for summarisation and visualisation provided above offers fast and easy overall evaluation of the data set. The definition of *MPI* allows for pooling together all the effects related to certain factors and quantify their potential to alter the responses of interest. The framework is flexible to accommodate further modifications as scaling, absolute value, changes in ranges or exclusion of factors. It can be easily extended further, for example, by adding higher order interactions and/or higher order polynomial effects. We do not provide one particular metric and visualisation, but rather the platform/framework to visualise the results of modelling across multiple responses in general way. From this perspective, the framework may be generalised further, beyond linear models applied to multiple responses of experimental results to virtually any modelling-based multiresponse framework.

Throughout the manuscript, we have implicitly assumed that all factors considered in the experiment are quantitative. The framework can be readily applied to two-level factors by using analogous equations, a baseline level and no quadratic effects. Various options can be considered in the case of factors with multiple levels, for example, for each factor taking the maximal effect across all its levels as the *MPI* or treating each level as a separate dummy variable compared with the baseline or with the overall average effect across all factor levels. The most suitable solution depends on the context of each experiment.

Any modelling approach to reach the final linear model for each response is compatible with the proposed framework. Hence, both univariate and multivariate approaches may be used. When a multivariate model is being applied, the correlation among the responses is implicitly reflected in the estimated coefficients. Although the correlation structure among responses could be useful in interpreting the visualisation output (e.g., by downweighing the responses that are highly cor-

related and focusing on the independent ones), we do not recommend to attempt to incorporate the correlation structure into the heatmap visualisation, because it typically results into very complicated graphics and reduces its communication potential. Hence, we would recommend to produce the correlation matrix (or its heatmap-like visualisation) aside and use it in the interpretation step.

The role of interactions in the visualisation may be tuned as needed for the application. If the interest is in main and quadratic effects only, the interactions can be entirely omitted and the framework can be simplified (although we discourage such practice without first checking for potential important interactions). Implicitly, the role of the interactions should be addressed in the model selection step.

The proposed framework is conditional to the model selection step and starts by using the estimates of the chosen models. Model selection would be typically needed in cases such as screening experiments that include a large number of factors and rarely allow for the estimation of all the main, the quadratic and the interaction effects. Naturally, the framework can be applied on the full model as well. We note that if possible to fit larger models, it is especially important to consider the impact of the interactions and not to evaluate the main effects in isolation.

The framework as presented above uses point estimates of the effects. Hence, it does not address uncertainty in the effects' estimates, which would result into unreliable results if the variability is generally large or severely imbalanced among the effetcs (e.g., in split-plot designs). In such cases, several straightforward extensions of the framework are possible. Standard errors of the effects' estimates can be visualised in analogous fashion to create a second heatmap to aid the interpretation of the point estimates. The relative standard error can be considered instead. Replacing the point estimates with upper or lower confidence bounds can be considered as well. Ideally, multiple of these visualisations may be combined to provide a sensitivity analysis to the simple point estimate-based analysis.

It needs to be clearly understood what the platform does not offer: the *MPI* and heatmaps cannot be used for direct decision-making on which factors are important and which are not in absolute sense. The final decision needs to combine the modelling output with scientific knowledge. However, this is not a limitation unique to this framework, but an essential property of any statistical tool: interpretation needs to be accompanied with domain knowledge in order to reach scientifically valid decisions. Visualisation accompanying the statistical analysis output can convey strong support for the evidence evaluation and an informed decision-making. The proposed technique certainly offers an easy way of presenting statistical conclusions on complicated large-scale experiments with a large number of factors and responses. This can be extremely handy in multidisciplinary projects where scientists from different fields need to find a common language to communicate the results.

Finally, the main motivation for the publication of this framework is the lack of such visualisation techniques in scientific communication as well as ready-to-make solutions in standard statistical software. We have not managed to find any alternative to be used for direct comparison. Visualisations are often focused on the main effects only or visualising all effects while losing the link between main and quadratic effect of the same factor. The closest tool to the purpose of the presented framework is probably the 'prediction profiler' graph from JMP,[13] which shows the predictions simultaneously for all the factors and responses. The main and quadratic effects are directly incorporated in the visualisation, but interactions are addressed only by the option to interactively adjust the levels of the corresponding factors. Hence, although the prediction profiler graph directly addresses the expected value of the response at certain combination of factors, it does not allow for immediate comparison across multiple factors. Additionally, it does not display the effect sizes using the *MPI*, but rather the resulting predicted response value. Hence, although both visualisation techniques are addressing similar goals in demonstrating modelling results, the prediction profiler is more focused on response surface experiments with a small number of factors rather than screening studies with a large number of factors and responses.

ACKNOWLEDGEMENTS

We would like to thank both referees for the thorough review and very useful comments.

ORCID

Martin Otava https://orcid.org/0000-0002-6150-913X
Kalliopi Mylona https://orcid.org/0000-0002-1460-0715

REFERENCES

1. Chen M., Golan A. What may visualization processes optimize? *IEEE Trans Vis Comput Graph*. 2016;34:187-200.
2. Wu C., Hamada M. *Experiments: Planning, Analysis, and Optimization*. 2nd. New Jersey: Wiley; 2001.

3. Jones B., Nachtsheim C. A class of three-level designs for definitive screening in the presence of second-order effects. *J Qual Tech*. 2011;22:2619-2634.
4. Georgiou S. Supersaturated designs: a review of their construction and analysis. *J Stat Plan Infer*. 2014;144:92-109.
5. Dean A., Morris M., Stufken J., Bingham D. *Handbook of Design and Analysis of Experiments*. Boca Raton: Chapman & Hall/CRC Handbooks of Modern Statistical Methods; 2015.
6. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019. URL http://www.R-project.org
7. Trinca L. A., Gilmour S. G. An algorithm for arranging response surface designs in small blocks. *Comput Stat Data Anal*. 2000;33:25-43.
8. Gilmour S. G., Ringrose T. J. Controlling processes in food technology by simplifying the canonical form of fitted response surfaces. *Appl Stat*. 1999;48:91-101.
9. Gilmour S. G., Trinca L. A. Some practical advice on polynomial regression analysis from blocked response surface designs. *Commun Stat Theor Meth*. 2000;29:2157-2180.
10. Mylona K., Goos P. Penalized generalized least squares for model selection under restricted randomization; 2011.
11. Letsinger J. D., Myers R. H., Lentner M. Response surface methods for bi- randomization structures. *J Qual Technol*. 1996;28:381-397.
12. Jones B., Nachtsheim C. Effective design-based model selection for definitive screening designs. *Technometr*. 2017;59:319-329.
13. JMP®. Version 13.2.1. SAS Institute Inc., Cary, NC. URL https://www.jmp.com/; 1989.

AUTHOR BIOGRAPHIES

Martin Otava is a principal statistician in Quantitative Sciences of Janssen Pharmaceutical Companies of Johnson & Johnson. He provides statistical support to research and development activities in pharmaceutical manufacturing, mainly for small molecules products, from early process characterization to process validation and early commercial production. Major areas of interest are statistical challenges in implementation of continuous manufacturing. From methodology perspective, he is interested in design of experiments, Bayesian approach towards hierarchical modelling and problems of equivalence. Martin received a doctorate in statistics and Master of Science in biostatistics from Hasselt University, Belgium. He also has a master's degree in mathematical statistics from Charles University, Czech Republic.

Kalliopi Mylona is a lecturer in statistics at King's College London, United Kingdom. She holds a Master of Science degree in applied mathematical and physical sciences from the National Technical University of Athens in Greece and obtained a PhD in statistics from the same university in 2009. Her main research areas are the design of factorial experiments and the analysis of experimental data, both the development of new statistical methodology and its application to real scientific problems.

How to cite this article: Otava M, Mylona K. Communicating statistical conclusions of experiments to scientists. *Qual Reliab Engng Int*. 2020;36:2688-2698. https://doi.org/10.1002/qre.2697

SUPPORTING INFORMATION

The supporting information file contains the R software codes using the RMarkdown framework and the respective HTML report. They do not provide any additional information beyond the manuscript content, but the reader can use it as a starting point for the implementation of the framework in R.