Multiple Linear Regression: a powerful statistical tool to understand and improve APIs manufacturing processes

1. INTRODUCTION

It is known that, over time, all production processes tend to deviate from their initial conditions. This happens for the most diverse reasons:

- changes in materials, personnel, environment,
- technological improvements,
- acquisition of production experience, *etc.*

Among other things, it is precisely in these changes that the foundations for an improvement of the process itself lie.

This variability in the processes, which often goes unnoticed, is instead well intercepted by the data that Quality Control systematically collects for batch release purposes. Furthermore, these data also capture very well the interactions between the different analytical parameters that normally escape. Now if these data are analyzed with the right tools, they can reveal a great deal of the manufacturing processes that generated them.

This *product knowledge* is of great practical use to the Company as it allows to:

- understand which are the parameters that most affect the product quality and how they interact with each other,
- establish whether the parameters that are controlled are really the ones we need or, instead, which ones would be better to consider,
- define / improve the *product control strategy* (as *per* FDA Guidances on Process Validation and Quality Metrics, ICH Q8-Q10-Q12, Eudralex Annex 15) based on experimental data and quantitative models rather than speculation,
- define and graphically represent the *design space* (ICH Q8) inherent to the production process considered,
- identify possible ways to improve process performance and scientifically pilot this improvement,
- mitigate the Regulatory impact in case of changes.

To extract this knowledge from the data, a powerful tool provided by the Statistical Sciences helps, namely the Multiple Regression in its usually most used model, that is the *linear* one.

Multiple Linear Regression (or MLR) is, in fact, the extension of the conventional "simple regression" to the case in which the dependent variable (or *response*, *y*) is related to several independent variables (also called: *predictors, regressors* or *features* as in Machine Learning, x_i) instead of with only one[1].

Here below a simplified approach to MLR is used to analyze the release data of an Active Pharmaceutical Ingredient (API) of which thirty-one batches have been produced and which was initially studied in the first post of this series (*i.e.*, May 9, 2018). On that occasion, the focus was on identifying and displaying the correlations between the various variables and the organized data structures (or *clusters*) that may be present. The analysis revealed that at least three lots had very different characteristics from the remaining twenty-eight listed in the database.

A different analysis of that same data is described below. Starting from the same database, a simple mathematical model is built which allows to highlight which of the variables considered independent in this analysis have the most significant influence (alone or jointly) on the variable chosen as dependent.

Thanks to the use of effective graphic representations typical of the DoE (*Design of Experiments*) methodology such as:

- *MAIN EFFECTS PLOTS*,
- *INTERACTION PLOTS* and
- *CONTOUR PLOTS*

the interactions between the variables involved are easy to understand and therefore to use.

In practice, here we want to use the "backward" DoE methodology, *i.e.*, instead of designing *orthogonal experiments* and generating data to be analyzed, apply it to existing data to extract the information contained therein.

2. EXPERIMENTAL SECTION

Since only numerical data are suitable for calculations, only quantitative variables (*i.e.*, those analytical tests whose outcome is a number) were considered and in cases where the result was expressed as lower than a numerical threshold value (*e.g.*, <LOQ) this value was used directly.

For each lot of the dataset considered, were used the analytical parameters (or variables) listed in Table 1 together with their units of measurement and the specifications they must satisfy.

Table 1 is completed by the abbreviations used below to identify the different analytical parameters in the graphs.

Table 1

Analytical parameter (or variable)	Units	Allowed Range of Variability	Analytical Technique	Abbreviation
pH	pH units	5.0 – 8.0	pH-metry	ph
Residual water content	%	1.0 – 5.0	Karl-Fisher titration	h2o
Assay	%	80 - 92	HPLC	assay
Starting material residual content	%	≤ 0.20	HPLC	sm
Largest known impurity	%	≤ 0.20	HPLC	known
Largest unknown impurity	%	≤ 0.20	HPLC	unk
Total impurities content	%	≤ 1.0	HPLC	total
Residual solvent 1 content	%	≤ 5.0%	Gas-chromatography	solv1
Residual solvent 2 content	%	≤ 5.0%	Gas-chromatography	solv2
Residual solvent 3 content	%	≤ 1.0%	Gas-chromatography	solv3

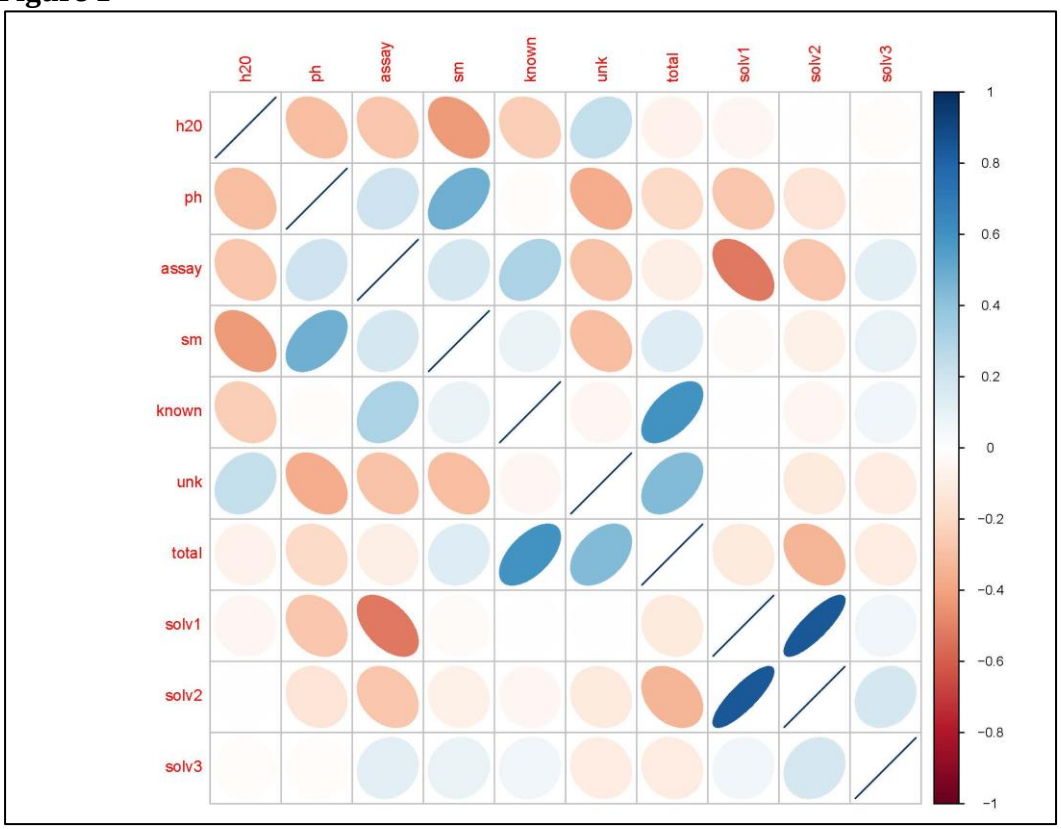
The dataset pertaining to the thirty-one lots considered consists of a table (called *double entry table*), whose rows each contain the data relating to a given lot while each column refers to a specific analytical parameter measured, or *variable*.
This data table, in statistical jargon, is usually referred to as a *data matrix*.

Data analysis and visualization were performed using RStudio (The R Foundation for Statistical Computing) and, above all, Minitab 19 (GMSL S.r.l. - Via Giovanni XXIII, 21 - 20014 Nerviano (Milan), Italy).
For the *correlogram* in Figure 1 it has been used the R package *corrplot* (T. Wei, Fujian Agriculture and Forestry University, China) [2, 3]

3. RESULTS AND DISCUSSION

Figure 1 shows the so-called *correlogram*, *i.e.*, the graph that displays the degree of linear correlation between the pairs of variables considered.

Figure 1



Each element of this diagram is a geometric figure that becomes more and more elliptical and intensely colored the more the two variables are *linearly correlated*. On the main diagonal, where the correlation is maximum (in fact the correlation of a variable with itself is equal to 1) the ellipses become a segment.

The ellipses are oriented to the right and blue colored if the two variables are positively correlated, while they are oriented to the left and red / brown colored if they are negatively correlated

The lack of numerous elongated and intensely colored ellipses indicates, already at a glance, that, in general, the variables are not very linearly correlated with each other. However, some exceptions are represented by couples:

- largest known impurity (*known*) and total impurities (*total*): positive correlation
- residual quantities of solvents 2 and 1: positive correlation
- residual quantity of solvent 1 (*solv1*) and assay (*assay*): negative correlation

Figure 1 also indicates the presence of weaker correlations (*i.e.*, non-elongated and faintly colored ellipses) such as those between the pairs:

- residual quantity starting material (*sm*) and pH value (*ph*): positive correlation
- largest unknown impurity (*unk*) and total impurities (*total*): positive correlation
- residual amount of water (*h2o*) and largest unknown impurity (*unk*): positive correlation
- residual quantity of starting material (*sm*) and residual amount of water (*h2o*): negative correlation.

These correlations, and particularly the stronger ones, reveal some characteristics worthy of further investigation. To learn more about them, the quantitative estimation of the degree of linear correlation between the different variables helps. This estimate is provided by the so-called *correlation matrix*, *i.e.*, the numerical basis on which the graph in Figure 1 was made. The correlation matrix is here below in Table 1.

Table 1

	h20	**ph**	**assay**	**sm**	**known**	**unk**	**total**	**solv1**	**solv2**
ph	-0,307								
assay	-0,275	0,210							
sm	-0,424	0,486	0,171						
known	-0,243	-0,020	0,317	0,088					
unk	0,239	-0,361	-0,283	-0,308	-0,041				
total	-0,061	-0,190	-0,084	0,131	**0,590**	0,432			
solv1	-0,050	-0,271	**-0,526**	-0,020	-0,007	0,010	-0,110		
solv2	-0,009	-0,141	-0,277	-0,077	-0,041	-0,117	-0,339	**0,843**	
solv3	-0,015	-0,019	0,115	0,092	0,061	-0,100	-0,108	0,062	0,180

The correlation matrix, in practice, is nothing more than a table whose elements are the *linear correlation coefficients of Bravais - Pearson* each calculated for a given row-column pair.

$$\rho_{ij} = \rho(X_i, X_j) = \frac{Cov\,(X_i, X_j)}{\sqrt{Var(X_i)\,Var\,(X_j)}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \qquad \forall i \neq \text{j}$$

where:

X_i = *i*-row of the data matrix

X_j = *j*-column of the data matrix

$\rho_{ij} \in [-1, +1] \quad \forall i \neq \text{j}$

Pearson's correlation coefficients shown in Table 1 highlight some important aspects for the purpose of creating a Multiple Linear Regression model and precisely:

- some independent variables (*e.g.*, residual quantities of solvents 1 and 2, *assay*, *etc.*) are highly correlated with each other (*i.e.*, ρ_{ij} > | 0.5 |) and therefore, reasonably, will have to be excluded from the model to prevent problems of *multicollinearity*. In fact, the ideal would be that all independent variables were significantly correlated with the dependent variable, but not with each other;
- considering *assay* as a dependent variable (*y*), it is observed that, except for the residual content of solvent1 (*solv1*), it is not strongly correlated with the other available regressors (*i.e.*, ρ_{ij} < | 0.5 |), in fact:

	h20	**ph**	**sm**	**known**	**unk**	**total**	**solv1**	**solv2**	**solv3**
assay	-0,275	0,210	0.171	0.317	-0.217	-0.084	**-0,526**	-0.277	0.115

- from the values reported above it is also to be expected that regressors such as the total content of impurities (*total*) do not appear in the final model due to their poor correlation ($\rho_{assay,total}$ = -0.084) with the dependent variable (*assay*).

For the purposes of building a model it is therefore necessary, first of all, to investigate the relationship of each independent variable with the dependent variable and any relationships existing between the independent variables. This analysis is carried out using *scatterplots* and *simple linear regression* models as here below documented only for some independent variables, namely those that show the highest linear correlation coefficients.

- **Regression Analysis: assay *vs.* solv1 (*R* = -0.526)**

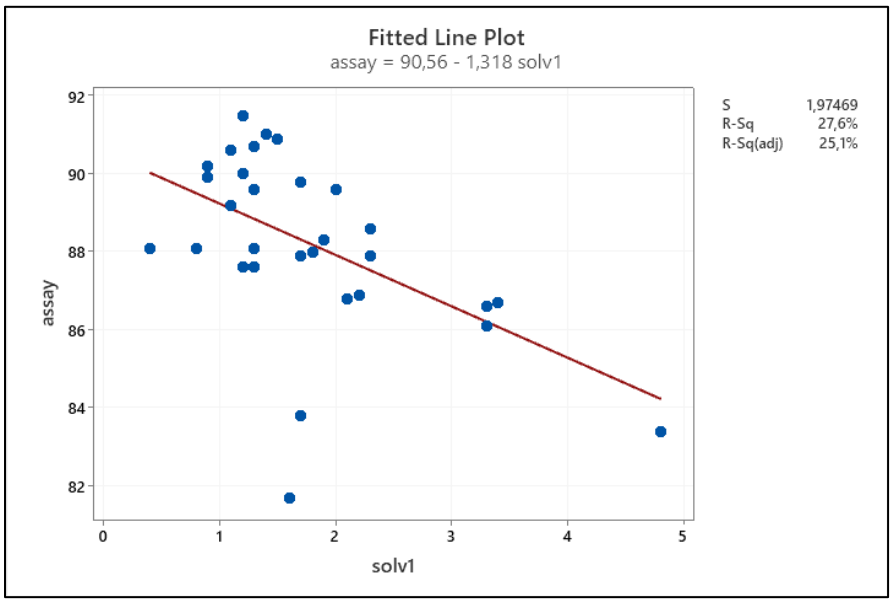
The regression equation is **assay = 90,56 - 1,318 solv1**

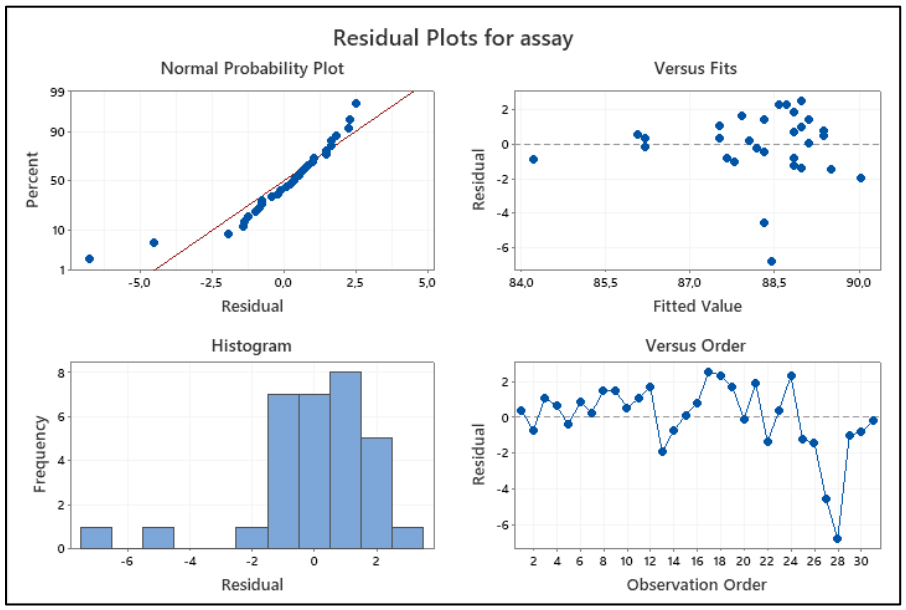
Model Summary

S	R-sq	R-sq(adj)
1,97469	27,63%	25,13%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	43,166	43,1655	11,07	0,002
Error	29	113,083	3,8994		
Total	30	156,248			





- **Regression Analysis: assay *vs.* known (*R* = 0.317)**

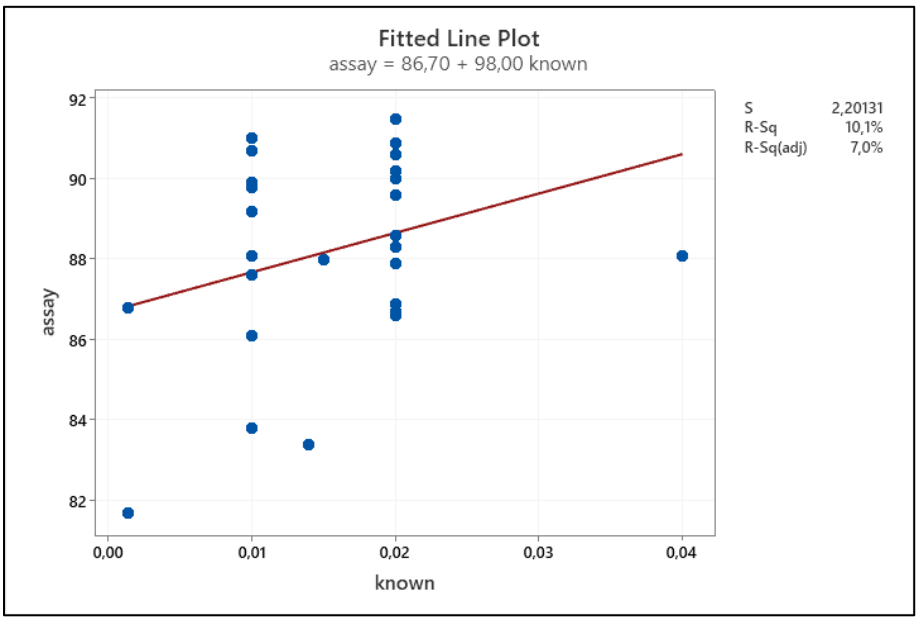
The regression equation is **assay = 86,70 + 98,00 known**

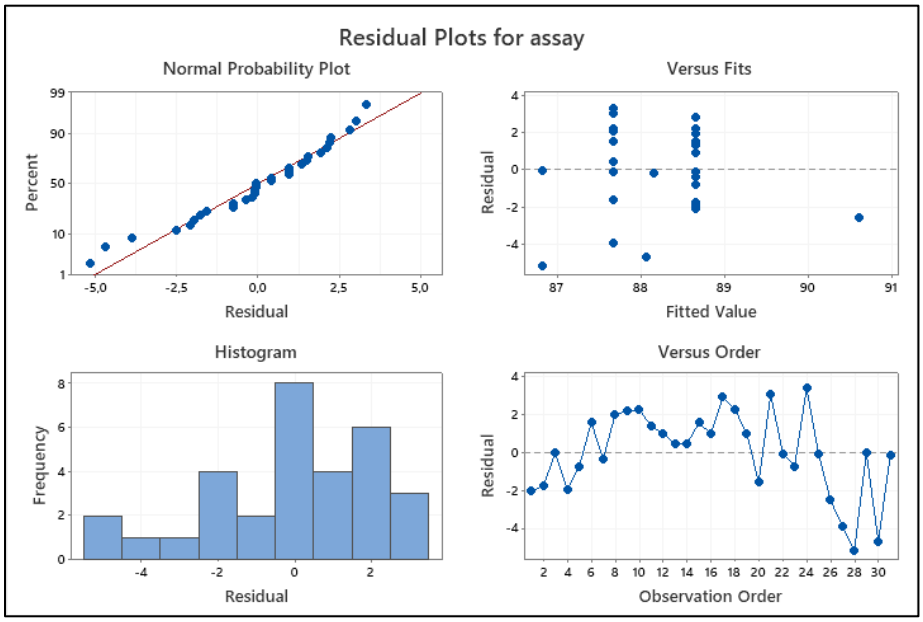
Model Summary

S	R-sq	R-sq(adj)
2,20131	10,06%	6,96%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15,721	15,7213	3,24	0,082
Error	29	140,527	4,8458		
Total	30	156,248			





- **Regression Analysis: assay *vs.* solv2 (*R* = - 0.277)**

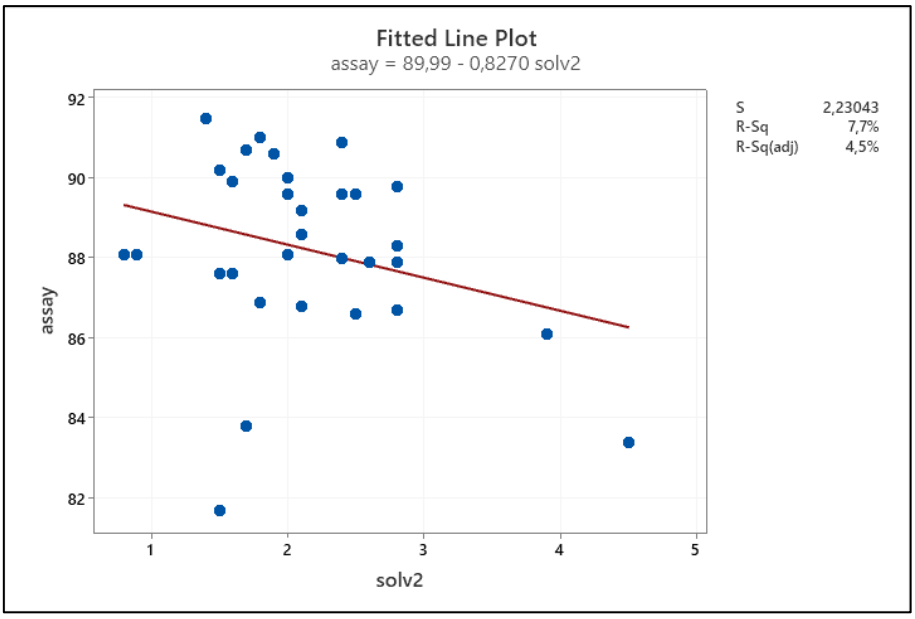
The regression equation is **assay = 89,99 - 0,8270 solv2**

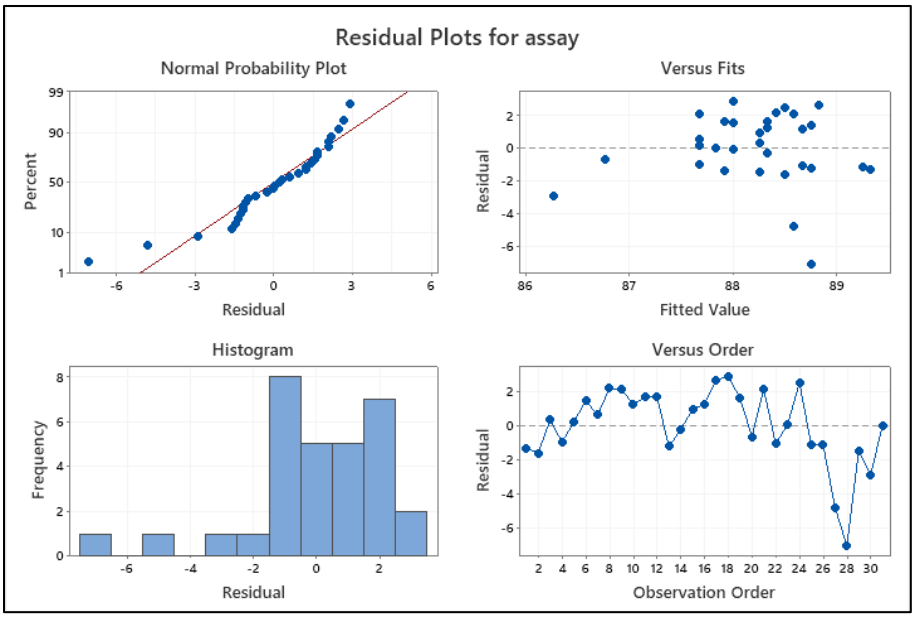
Model Summary

S	R-sq	R-sq(adj)
2,23043	7,67%	4,48%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	11,979	11,9791	2,41	0,132
Error	29	144,269	4,9748		
Total	30	156,248			





From the examination of the results relating to the relationships between the dependent variable (*y*, *assay*) and the independent variables, some observations common to all cases can be drawn:

- the low levels of linear correlation initially found (Figures 1 and 2) are related to the wide dispersion of experimental data around the regression lines. In all models shown here above, the *S* value (*i.e.*, standard error of the regression) is close to or greater than two. *S* represents the standard deviation of how far the data values fall from the fitted values and it is measured in units of the response variable. Since, approximately, 95% of the observations should fall within ± *2S* (which is a quick approximation of a 95% prediction interval), in the above regressions about 95% of the observations should fall within ± 4 – 4.4 % of the fitted lines;
- previous point suggests that the final model will also be characterized by a certain standard error value. A regression analysis, in fact, can only be as good as the data on which it is based;
- the residuals, which can be seen as the realization of the error associated with each model, except for some anomalous values (which anyway result from anomalous data), do not show specific patterns.

By extending the above analysis to the other dependent variables and ordering them on the basis of the absolute values of the linear correlation coefficient, they are identified as possible variables on which to build the model:

	S	R-sq	R-sq(adj.)	Correlation
solv1	1,97469	27,63	25,13	**-0,526**
known	2,20131	10,06	6,96	**0,317**
unk	2,22606	8,03	4,86	**-0,283**
solv2	2,23043	7,67	4,48	**-0,277**
h2O	2,23182	7,55	4,36	**-0,275**
pH	2,26966	4,39	1,09	**0,210**
sm	2,28681	2,94	0	**0.171**

The remaining independent variables (*solv3*, *total*) are not considered as they are even less linearly related to the dependent variable *assay* and therefore even less able to contribute significantly to the model.

Considering now only the correlations between independent variables, the values shown in Table 2 below show that, in addition to the pair (*solv1*, *solv2*) to which is associated an *R* value equal to 0.843, other pairs of variables also show significant correlations: (*sm*, ph) with *R* = 0.486 or (*sm*, *h2o*) with *R* = -0.424.

Table 2

	h20	ph	sm	known	unk	solv1
ph	-0,307					
sm	**-0,424**	**0,486**				
known	-0,243	-0,02	0,088			
unk	0,239	-0,361	-0,308	-0,041		
solv1	-0,05	-0,271	-0,02	-0,007	0,01	
solv2	-0,009	-0,141	-0,077	-0,041	-0,117	**0,843**

Given all this, a model based on the functional relationship is built:

$$assay = f(solv1, known, unk, solv2, h2o, ph, sm)$$

Taking into account these variables and all second-order interactions, the model described by the regression equation (1) is obtained:

Assay = 37 + 126 solv2 - 826 known - 37,1 h20 + 12,2 ph - 494 sm + 119 unk
- 108,0 solv1 - 3,59 h20*solv1 + 4,93 h20*ph - 452 h20*sm + 561 h20*known
+ 138 h20*unk - 0,88 h20*solv2 - 120 ph*sm - 70 ph*known - 64 ph*unk
+ 18,0 ph*solv1- 19,9 ph*solv2 - 11753 sm*known + 15202 sm*unk **(1)**
- 265 sm*solv1 + 874 sm*solv2 - 2775 known*unk + 433 known*solv1
- 266 known*solv2 + 27,6 unk*solv1- 41,2 unk*solv2 + 0,04 solv1*solv2

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1,56321	96,87%	53,08%	0,00%

It deals of a model in 28 variables (between single and double) and a constant.
In this model and in what follows it can be identified three different types of terms:

- a CONSTANT, that in this case is equal to 37
- PURELY LINEAR TERMS (monomial grade = 1) whose general structure is:

  *numerical parameter * independent variable*

- MIXED TERMS (monomial grade = 2) whose general structure is:

  *numerical parameter * independent variable$_1$ * independent variable$_2$*

This clarification will help later when two different types of graphs (*i.e.*, CONTOUR PLOTS) will be compared with each other.

The R-sq value, which measures the percentage of variation in the data explained by the model is, in this case, about 97% and it is calculated as:

$$R - sq = 1 - \frac{SS_e}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where:

$SS_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 =$ error sum of squares (variation not explained by model)

$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2 =$ total sum of squares (total variation in the model).

Unfortunately, the R-sq (adj) value is much lower than R-sq (53% *ca. vs.* 97% *ca.*) and, above all, this model totally lacks any predictive capacity, in fact R-sq (pred) = 0.00%. This finding is typical of *overfitting* and in fact the model shows many terms that are not significant as indicated by *P-values* >> 0.05 (Table 3). An example for all is represented by the factor solv1 * solv2 to which corresponds a *P-value* = 0.983. In this respect it is always useful to bear in mind that:

It is frequently helpful to have a procedure that can guard against overfitting the model, that is, adding terms that are unnecessary. The adjusted R^2 penalizes us for adding terms that are not helpful, so it is very useful in evaluating and comparing candidate regression models [1]

Table 3 - Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	28	151,361	5,40575	2,21	0,359
h20	1	2,430	2,43001	0,99	0,424
ph	1	0,190	0,18952	0,08	**0,807**
sm	1	0,007	0,00667	0,00	**0,963**
known	1	0,014	0,01377	0,01	**0,947**
unk	1	0,032	0,03191	0,01	**0,919**
solv1	1	3,498	3,49756	1,43	0,354
solv2	1	3,718	3,71792	1,52	0,343
h20*ph	1	2,194	2,19376	0,90	0,443
h20*sm	1	1,290	1,29026	0,53	0,543
h20*known	1	2,071	2,07097	0,85	0,454
h20*unk	1	2,787	2,78709	1,14	0,397
h20*solv1	1	1,613	1,61329	0,66	0,502
h20*solv2	1	0,102	0,10228	0,04	**0,857**
ph*sm	1	0,179	0,17874	0,07	**0,812**
ph*known	1	0,003	0,00315	0,00	**0,975**
ph*unk	1	0,296	0,29631	0,12	0,761
ph*solv1	1	2,740	2,73975	1,12	0,401
ph*solv2	1	3,568	3,56756	1,46	0,350
sm*known	1	0,003	0,00319	0,00	**0,974**
sm*unk	1	1,957	1,95727	0,80	0,465
sm*solv1	1	4,142	4,14200	1,70	0,323
sm*solv2	1	5,957	5,95673	2,44	0,259
known*unk	1	1,598	1,59772	0,65	0,504
known*solv1	1	4,800	4,79984	1,96	0,296
known*solv2	1	1,313	1,31339	0,54	0,540
unk*solv1	1	0,276	0,27646	0,11	0,769
unk*solv2	1	0,730	0,72964	0,30	0,640
solv1*solv2	1	0,001	0,00150	0,00	**0,983**
Error	2	4,887	2,44362		
Total	30	156,248			

The Pareto diagram shown in Figure 2, which distinguishes *significant* effects from *insignificant* ones, in this case is of no use due to the presence of so many highly correlated terms that make the model inadequate.

Figure 2

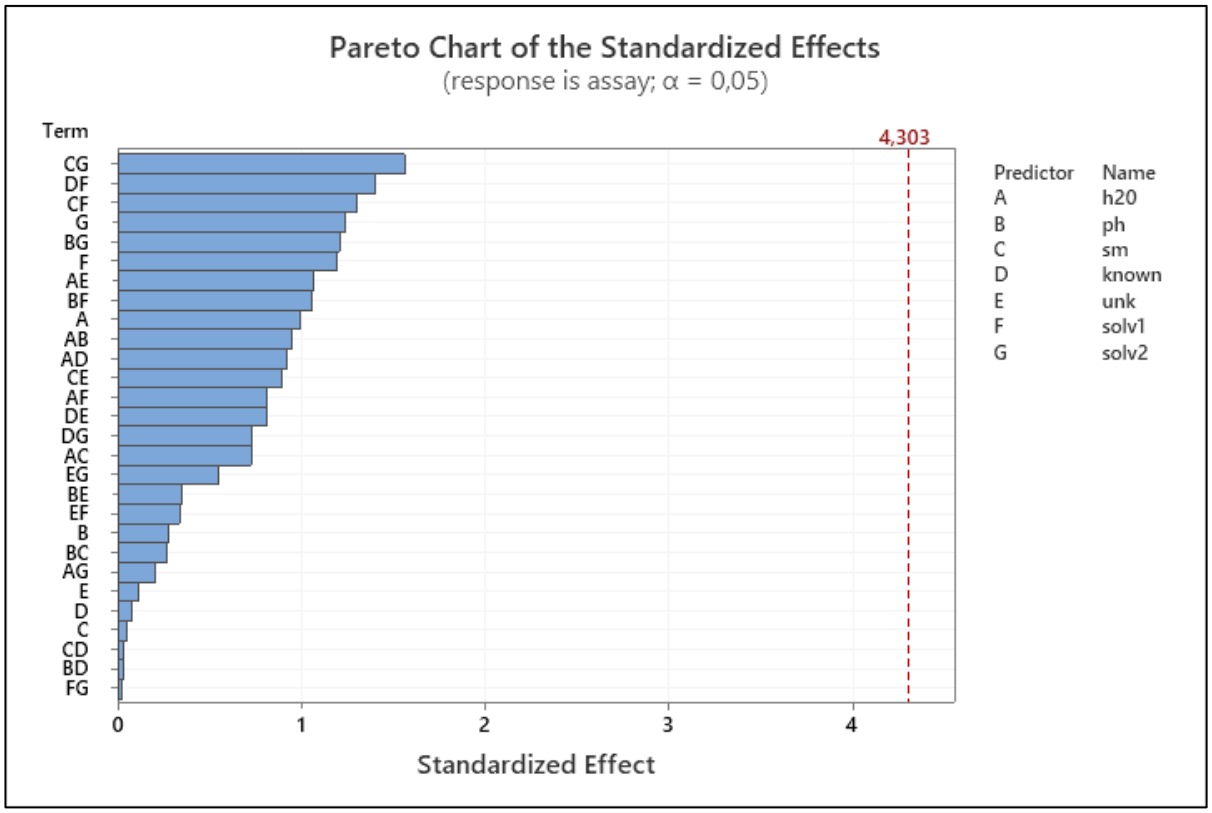
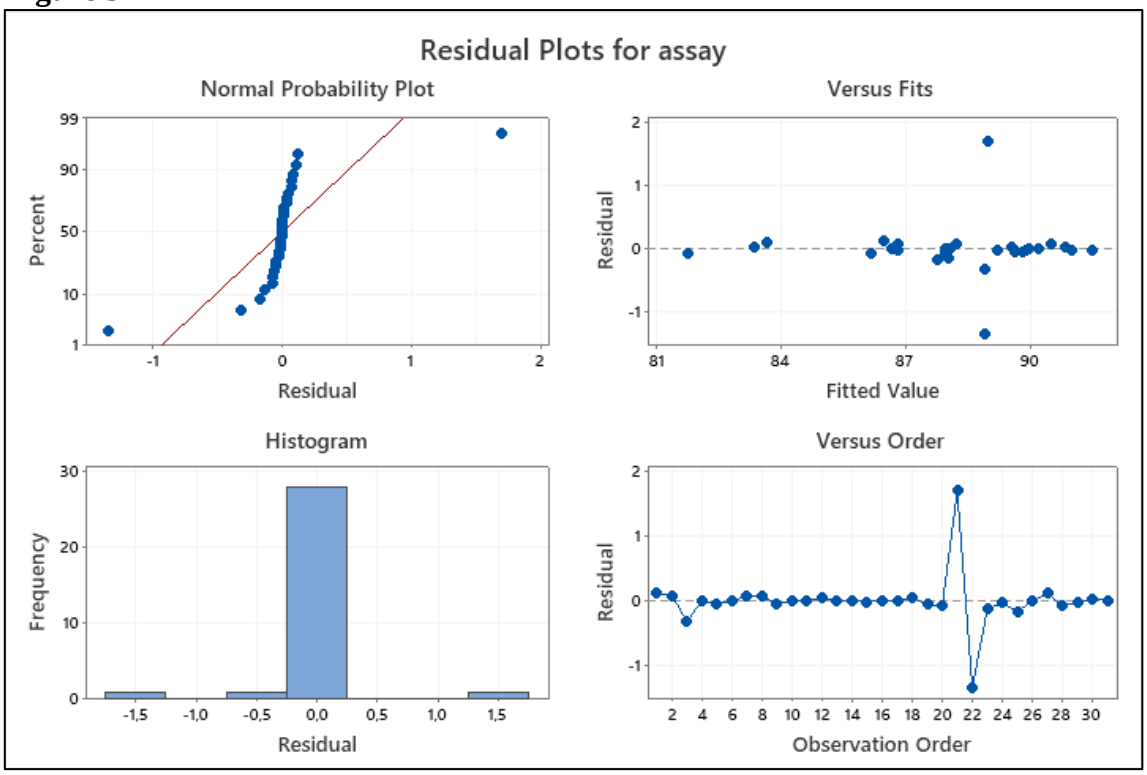


Figure 3, here below, summarizes the residual diagrams and they too are poorly informative given the inadequacy of this initial model which, although it explains about 97% of data variability, is practically unusable.

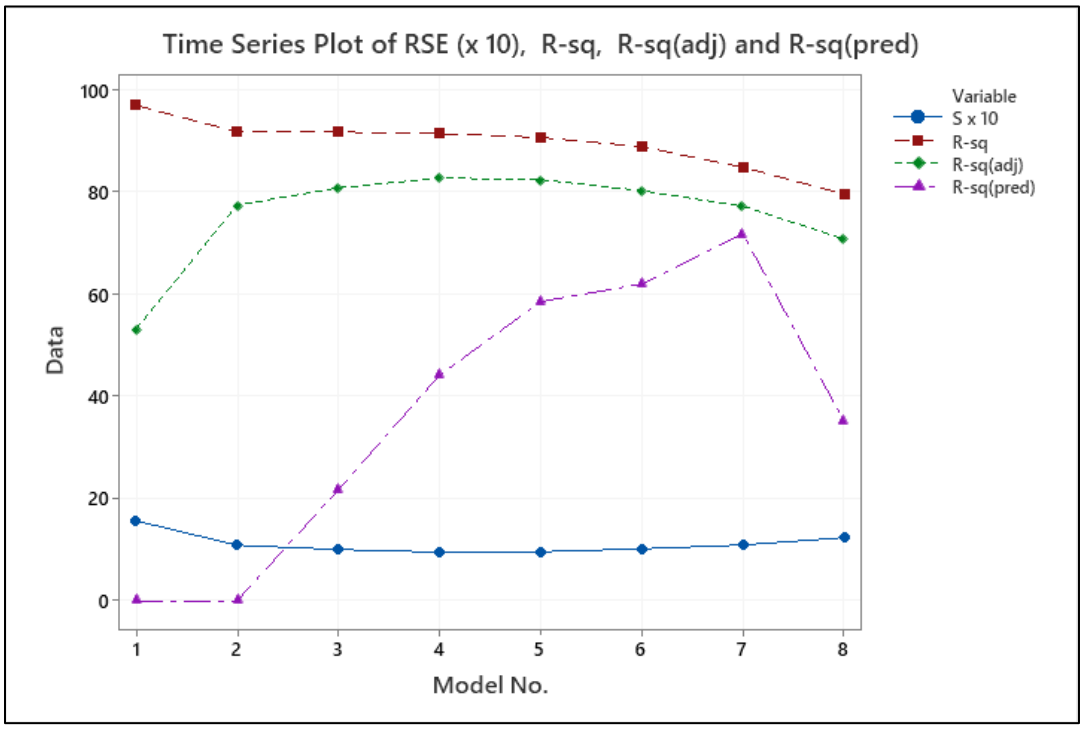
Figure 3



This initial model was then refined by progressively eliminating the insignificant terms. The refinement was carried out as summarized in Figure 4, namely trying to keep the R-sq value as high as possible and concurrently increase both the values of R-sq (adj) and, above all, that of R-sq (pred).

The refinement process has been considered completed when this continuous growth stopped, and further attempts of improvement were only leading to an increase in the standard error that began to increase.

Figure 4



The numerical values corresponding to the points in the graph of Figure 4 are shown below in Table 4 and indicate in model no. 7 the best compromise. In fact, it combines the highest values of R-sq, R-sq (adj) and R-sq (pred) with the lowest *S* value.

Comparing the initial model (1) with the final one described by the regression equation (2) here below, it is evident that the refinement process although at the cost of a loss of approximately 12% in predictive capacity (*i.e.*, 84.76% *vs.* 96.87%), has led to a model characterized by:

- a standard error 30% lower than the initial figure (*i.e.*, 1.09133 *vs.* 1.56321),
- a R-sq (adj) value which only differs from R-sq by 9% *ca.* (*i.e.*, 77.13 *vs.* 84.76%) against the 45% observed in the initial model (*i.e.*, 53.08% *vs.* 96.87%),

but most of all:

- a predictive capacity that increased up to 71.74% starting from an initial 0.00%.

Finally, while the initial model (1) was based on 28 variables, the refined one (2) uses only 10.

Figure 4, but even more Table 4, show that further adjustments immediately lead to significant losses in the characteristics of the model.

Table 4

Model No.	S	R-sq	R-sq(adj.)	R-sq(pred)
1 (initial)	**1,56321**	**96,87%**	**53,08%**	**0,00%**
2	1,08621	91,69%	77,35%	0,00%
3	1,00178	91,65%	80,73%	21,70%
4	0,950528	91,33%	82,65%	44,08%
5	0,958665	90,59%	82,35%	58,48%
6	1,01619	88,76%	80,17%	61,93%
7 (final)	**1,09133**	**84,76%**	**77,13%**	**71,74%**
8	**1,23492**	**79,50%**	**70,72%**	**35,01%**

The regression equation associated with the best model resulting from the optimization process is:

$$\begin{aligned} \text{Assay} = {} & 93{,}26 - 4{,}136\ \text{h20} - 43{,}34\ \text{solv1} + 31{,}37\ \text{solv2} + 136{,}9\ \text{h20*known} \\ & + 25{,}45\ \text{h20*unk} + 5{,}22\ \text{ph*solv1} - 3{,}776\ \text{ph*solv2} - 2742\ \text{known*unk} \\ & + 577\ \text{known*solv1} - 429{,}8\ \text{known*solv2} \end{aligned} \qquad \textbf{(2)}$$

Three of the ten variables in model (2) appear as single independent variables (or *factors*) while the remaining seven are *interactions between two variables*. The three factors that appear individually (*solv1*, *solv2* and *h2o*) are among the variables that, in the initial analysis, were linearly well correlated with *assay*, in fact:

- R (*assay*, *solv1*) = -0.526
- R (*assay*, *solv2*) = -0.277
- R (*assay*, *h20*) = -0.275

Other variables, also well correlated linearly with *assay*, such as:

- *known* : R (*assay*, *known*) = 0.317
- *unk* : R (*assay*, *unk*) = -0.283
- *ph* : R (*assay*, *ph*) = 0.210

and which were used to build the initial model, appear instead in the mixed terms (*interactions*).

The remaining less linearly correlated variables with *assay* such as:

- *sm* : *R* (*assay*, *sm*) = 0.171
- *solv3* : *R* (*assay*, *solv3*) = 0.115
- *total* : *R* (*assay*, *total*) = -0.084

do not appear in the model.

Figure 5 shows the Pareto diagram which highlights how all the effects considered in model (2) are significant.

Figure 5

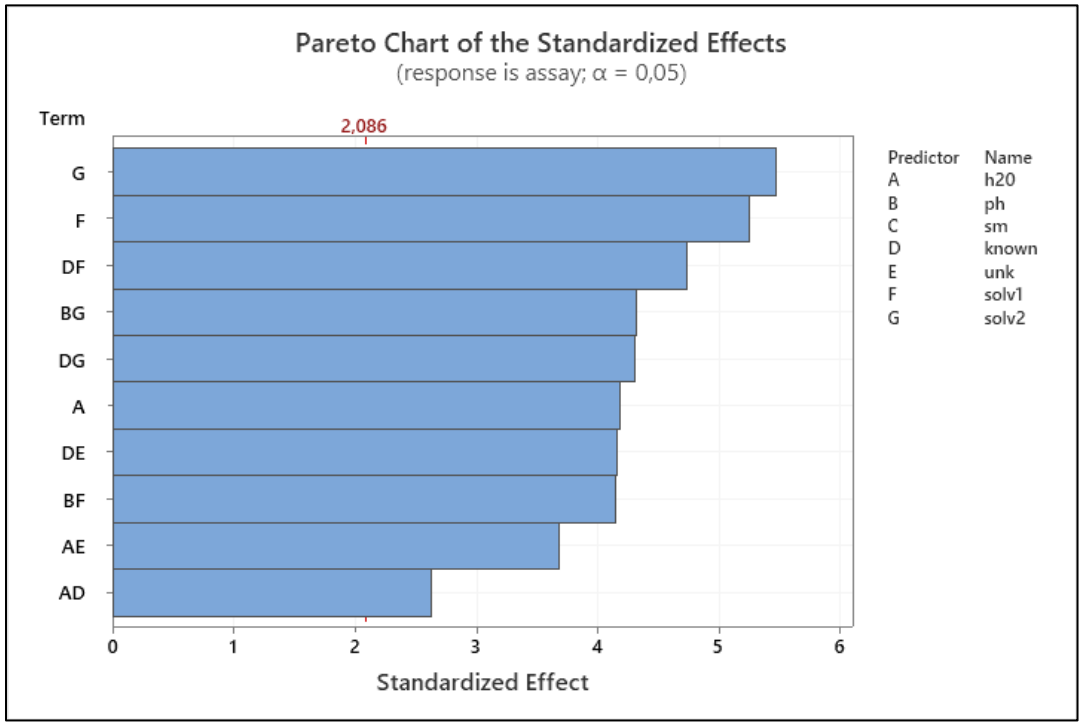


Figure 6 shows the graphs relating to the residuals. The *normal probability plot* and the histogram indicate a practically normal trend. The diagrams on the right show a scattering of points around zero practically free from patterns or trends.

Figure 6

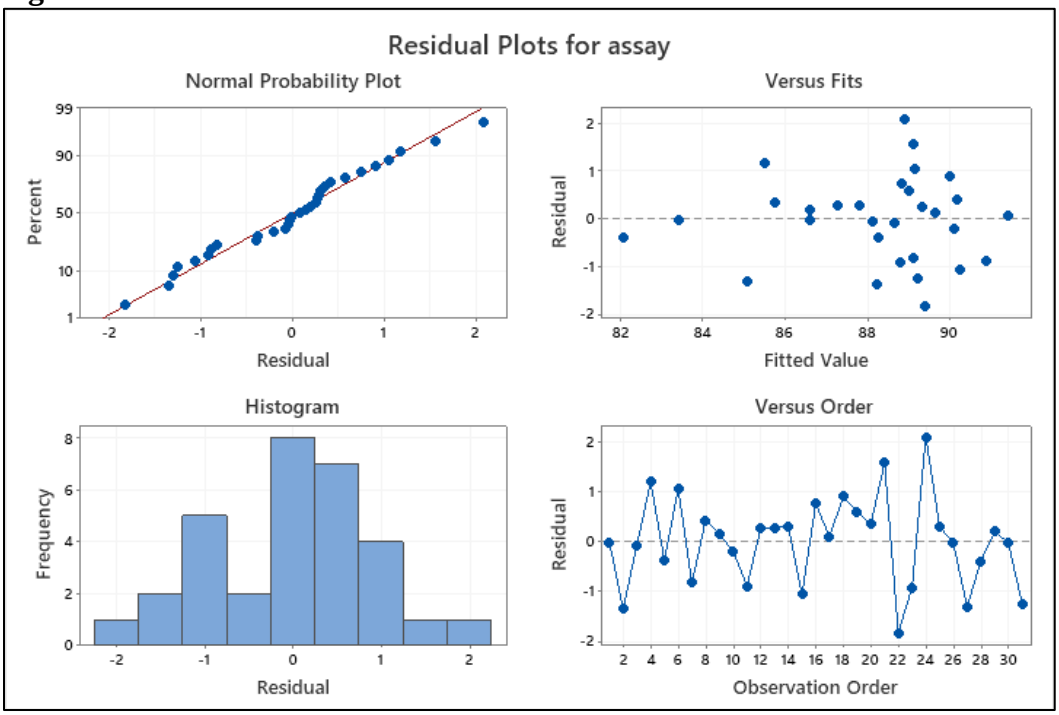


Figure 7 shows the good level of approximation of the experimental assay values provided by model (2). The initial experimental data (*Assay exp. values*) are in fact represented by a green line while the limits, lower (*Assay calc. - 2S*) and upper (*Assay calc. + 2S*), calculated using the model (2) are represented by two broken lines respectively in red and blue.

Figure 7

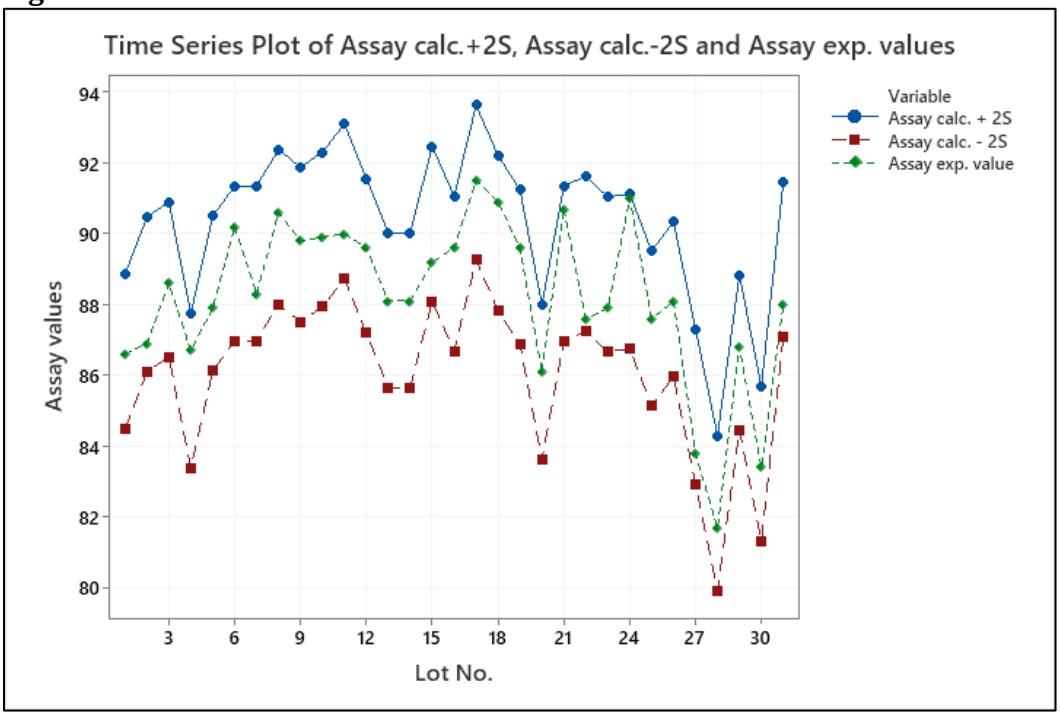
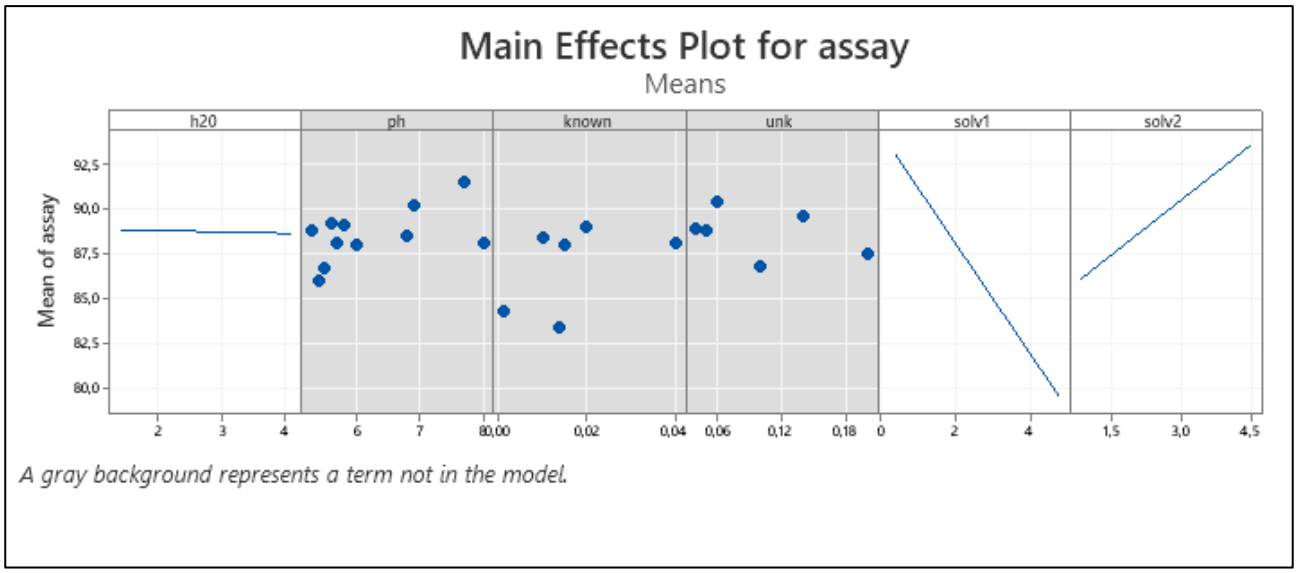


Figure 7 shows that experimental values (green line) are included among those calculated on the basis of model (2).

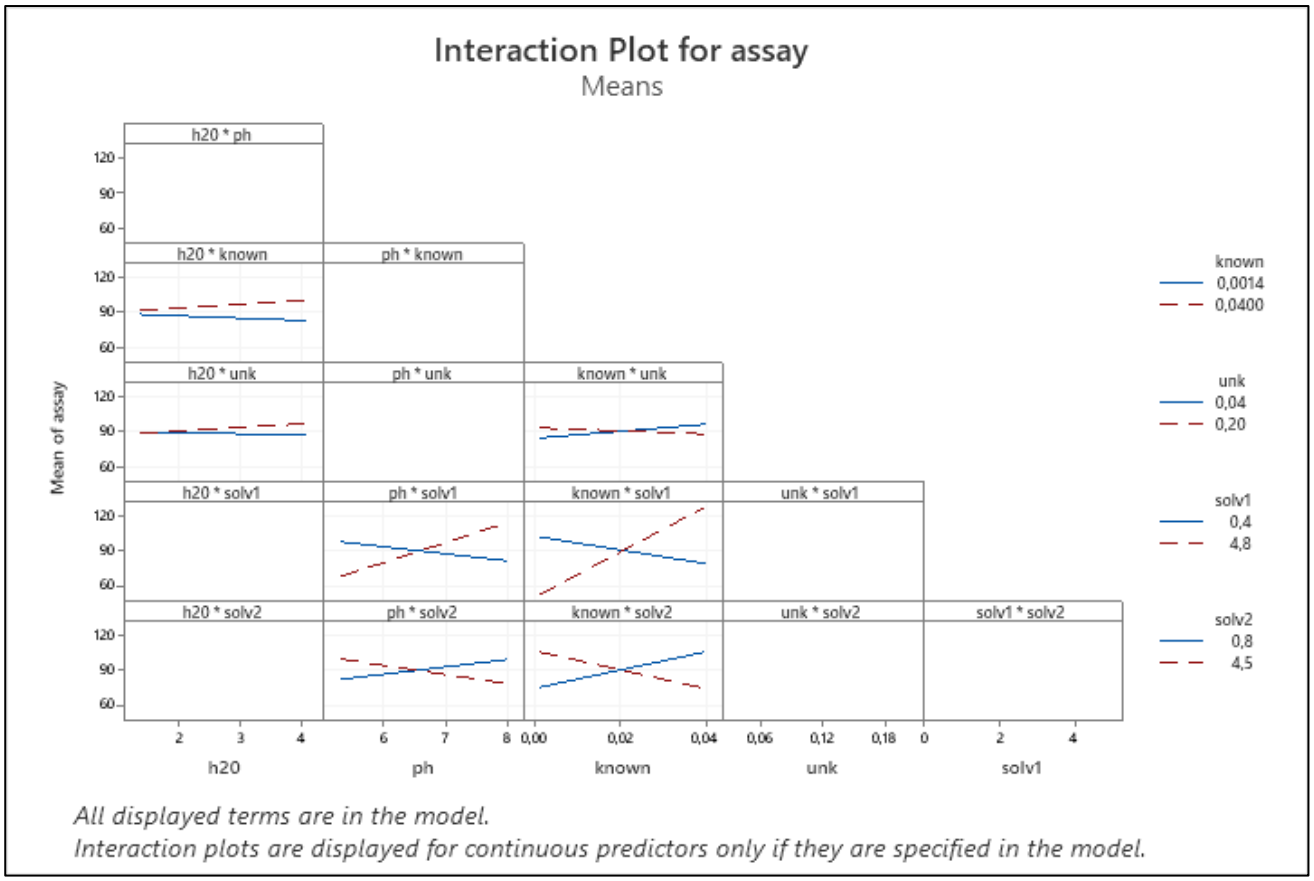
The two figures below show the so-called *FACTORIAL PLOTS* relating to the *MAIN EFFECTS* (Figure 7) and to the *INTERACTIONS BETWEEN FACTORS* (Figure 8).

Figure 7



The graphs in Figure 7 show the average effect on *assay* of those factors such as the residual levels of water (*h2o*) and the two solvents 1 and 2 (*solv1*, *solv2*) as their respective levels vary. In general, the steeper the segment, the more significant the effect of the factor. In this case it is observed how the residual content of the two solvents significantly influences *assay* while the residual water content (*h2o*) shows only an almost zero effect.
In the gray fields of Figure 7 are the main effects of those factors (*ph*, *known*, *unk*) which do not appear individually in the model.

Figure 8



The INTERACTION PLOT in Figure 8 shows if there are interactions between the levels of the factors considered. In general, the more the segments exhibit significant slopes or even intersect, the greater the significance of the levels of interaction between the factors involved. In this case it can therefore be observed how, for example, the interactions between the residual levels of water and unknown (*h2o* ↔ *unk*) or known impurity (*h2o* ↔ *known*) or between the residues of unknown and known impurities (*known* ↔ *unk*) are of little significance.

On the contrary, the interactions between:

- pH and residual solvent content 1 or 2 (*ph* ↔ *solv1*, *ph* ↔ *solv2*)
- known impurity level and residual solvent 1 or 2 content (*known* ↔ *solv1*, *known* ↔ *solv2*)

are all significant.

The Main Effects graphs in Figure 7 show how the residual content of solvents 1 and 2 significantly influences the *assay* value. To further investigate this type of specific relationship, you can use the so-called *level curves* (or CONTOUR PLOT), a graphical representation of the response variable (*assay*) as a function of two or more factors based on the linear model developed. Figure 9 shows the behavior of the response variable (*assay*) as a function of the two factors *solv1* and *solv2* which is obtained using the linear model developed and keeping all the other variables constant. For them the median value was chosen as the reference.

Figure 9

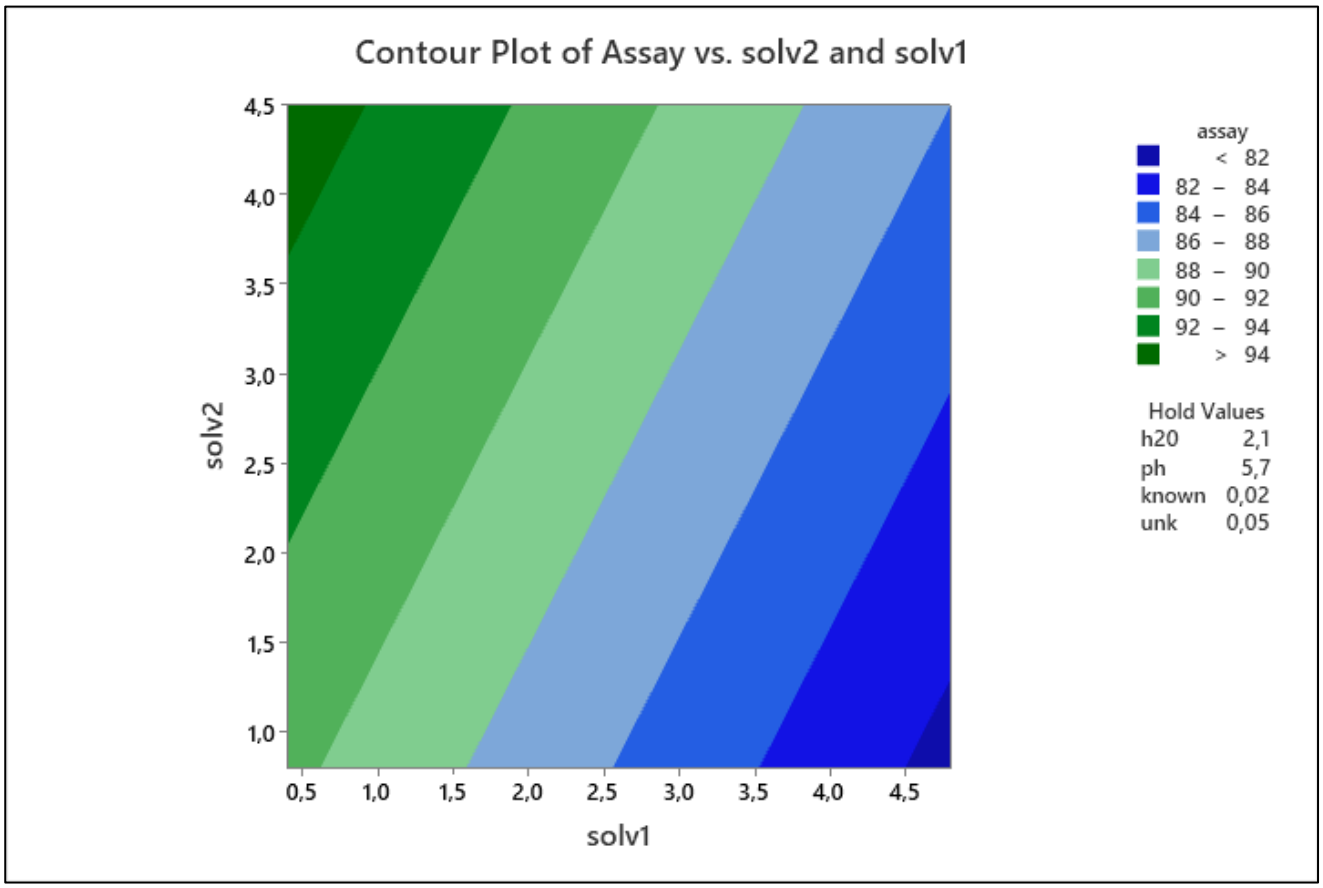
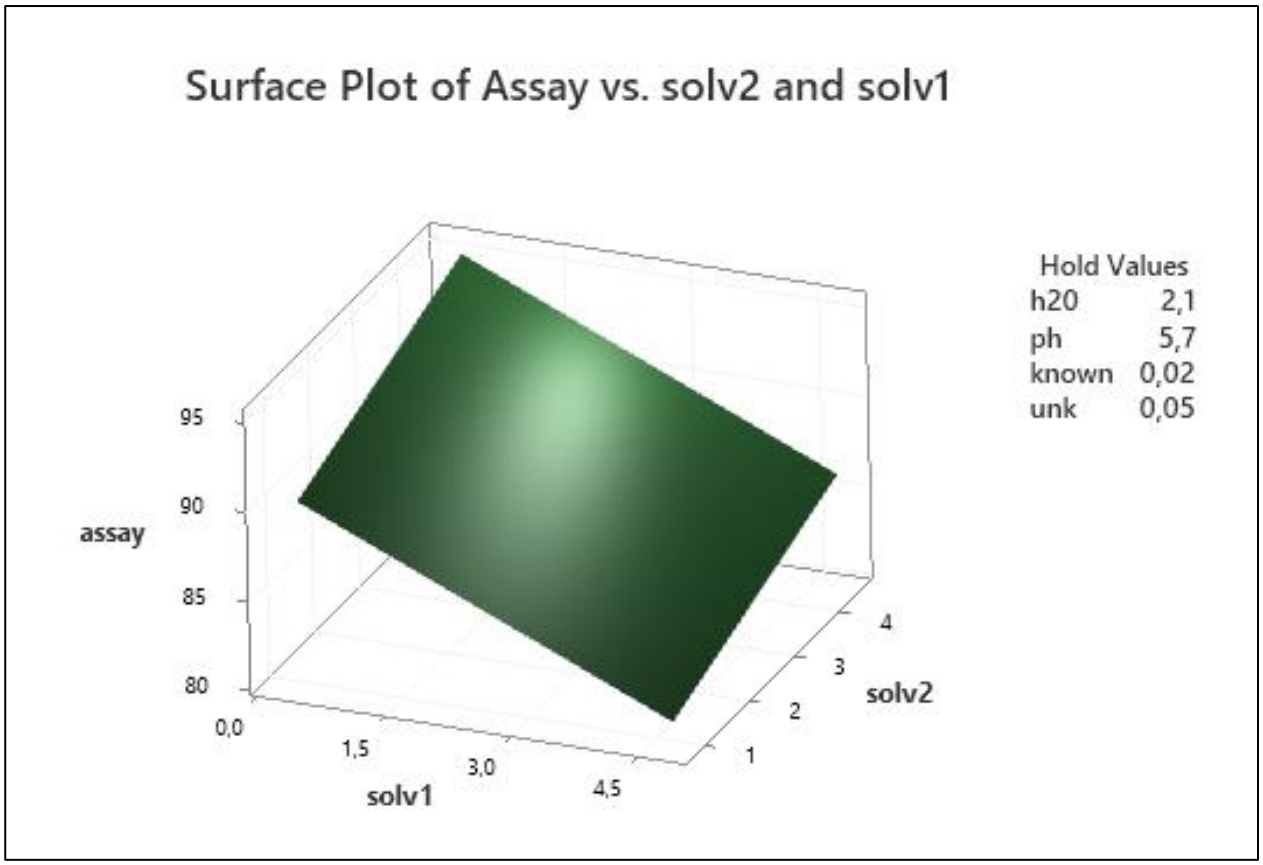


Figure 9 shows that, with the other variables being equal, the maximum assay value (dark green area) is obtained in the vicinity of *solv1* equal to approx. 0.5%. and *solv2* equal to 4.5% approx.

A more spatial, three-dimensional view of the interaction between *solv1* and *solv2* on the assay value is offered by the *response surface* (or SURFACE PLOT) shown in Figure 10 which shows the *planarity* (*i.e., linearity in space*) of this relationship.

Figure 10

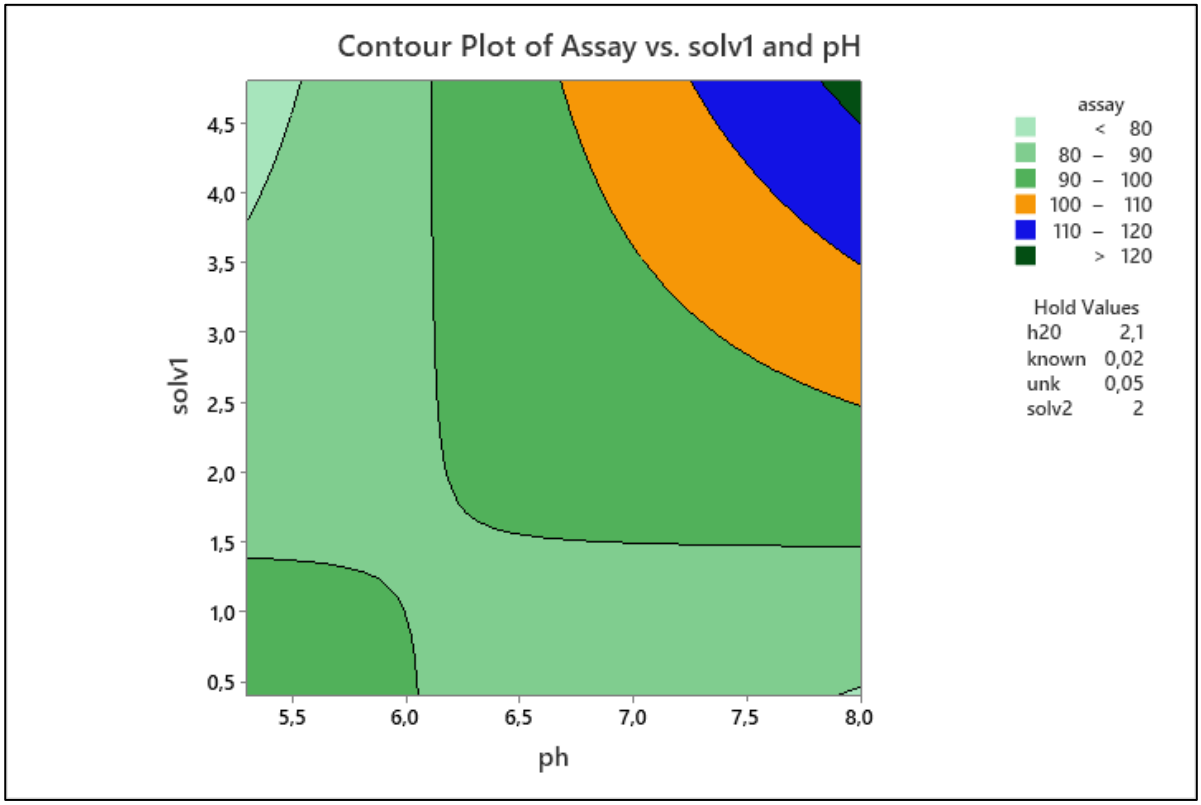


The use of these survey tools, which allow us to define and represent the *design space* (ICH Q8 (R2)), is even more useful when we want to study, in order to exploit them practically, the interactions between different variables.

Figure 8, for example, shows a significant interaction between pH value (*ph*) and the residual content of solvent 1 (*solv1*).

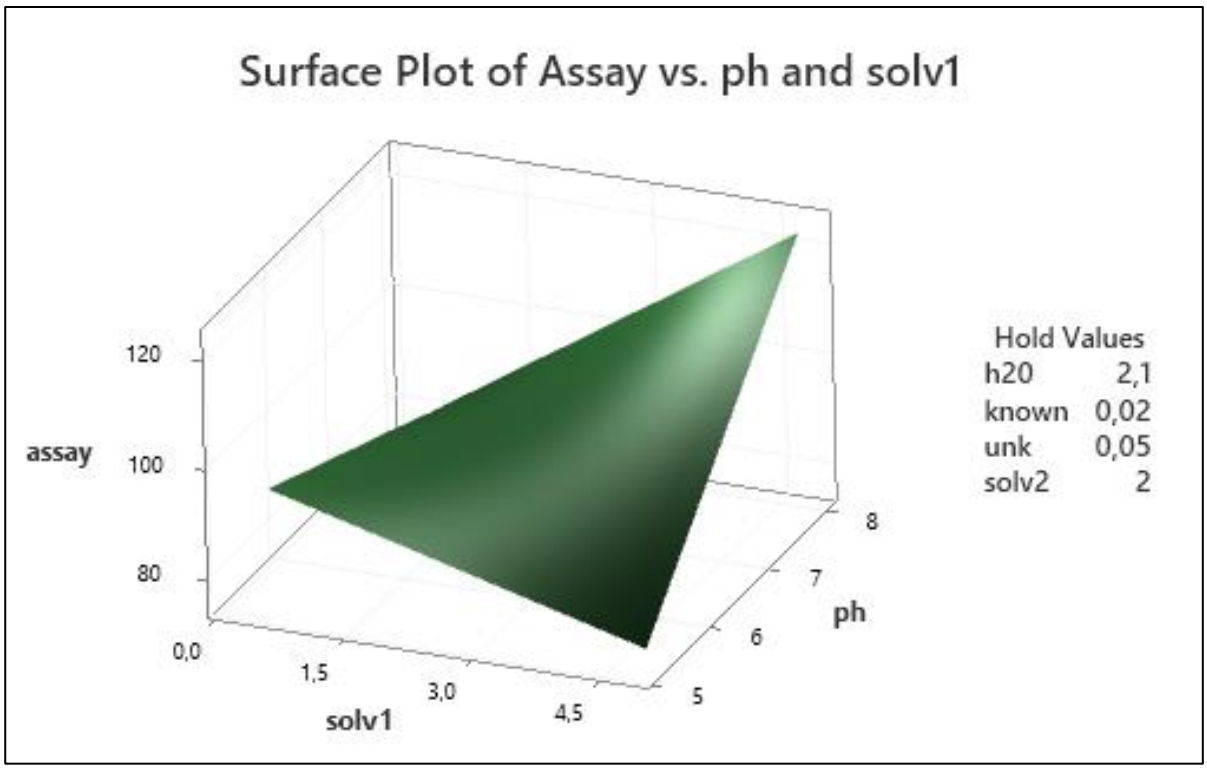
Figure 11 shows the behavior of the response variable (*assay*) as a function of the two factors *ph* and *solv1* which is obtained using the linear model developed and keeping all the other variables constant. As before, the median value was chosen as a reference for the other variables. In this case, since the relationship between dependent variable (*assay*) and independent variables (*solv1*, *ph*) is of "mixed nature" and not "purely linear" as it was for (*assay* ↔ *solv1*, *solv2*), the regions of the useful operating ranges instead of appearing as *parallel bands* (Figure 9) show up as *curved bands* as shown in Figure 11. The operating range associated with assay values between 90% and 100% is indicated by the dark green band adjacent to the orange one.

Figure 11



A spatial view of the interaction between *solv1* and *ph* on the *assay* value is given by the *response surface* in Figure 11.

Figure 11



4. CONCLUSIONS

It was shown how, applying MLR to the data that Quality Control systematically collects for release purposes on different samples of a given API, it is possible to extract information about the manufacturing process behind such data.

In particular, by choosing the *assay* values as dependent variable and all others that define the API's purity profile as independent variables, were identified the parameters that most affect the assay and how they interact with each other. On this basis it is therefore possible to establish whether the parameters that are controlled are really those that are needed and therefore to define a *product control strategy* based on experimental data.

It was then shown how, thanks to the *factorial graphics*, it is possible to identify and graphically represent the *design space* present in the production process used. The operating ranges highlighted in the Contour Plots suggested, for example, areas where single variables or their combinations allowed to maximize the assay value and thus improve the process.

Moreover, the availability of data pertinent to several years of production of a given active ingredient would allow to establish whether a given model maintains its predictive character, and therefore its validity, over time.

It is however clear, in all cases, that the quality of the models that can be obtained and therefore of the deductions / predictions that can be made strongly depend on the data available.

The approach detailed here can be extended to other situations such as, for instance:

- stability studies, where: y = assay value - x_i = stability indicating parameters

or

- any manufacturing process, where: y = process yield - x_i = parameters measured in-process

provided, of course, that data show some degree of variability.

All the considerations made so far have an undoubted value for the purposes of a process knowledge based on quantitative data, however, given their chemical nature, they need to be supplemented by specific knowledge of R&D and Production.

5. BIBLIOGRAPHY

1. D.C. Montgomery, E.A. Peck, G.G. Vining, *Introduction to Linear Regression Analysis*, 5th Ed., (2012) Wiley
2. M. Friendly, *Corrgrams: Exploratory displays for correlation matrices*, The American Statistician, 56 (2002) 316-324
3. D.J. Murdoch, E.D. Chow, *A graphical display of large correlation matrices*, The American Statistician, 50 (1996) 178-180