How to extend the shelf life of an API ?
Look at its Stability Data from a Multivariate standpoint !

The real voyage of discovery consists not
in seeking new landscapes, but in having new eyes.

M. Proust

1. INTRODUCTION

According to ICH guideline Q1A (R2), the purpose of the stability studies is to "*provide evidence on how the quality of a drug substance or drug product varies with time under the influence of a variety of environmental factors such as temperature, humidity, and light, and to establish a re-test period for the drug substance or a shelf life for the drug product and recommended storage conditions.*" [1]

To this end, the guideline provides indications on how to carry out these studies (temperature, humidity, frequency of measurements, *etc.*) and on the evaluation of the experimental results. In general, the trend of a "quantitative attribute" (the *assay*, usually) is followed, which is expected to vary over time because of the degradation process, and its behavior is compared with respect to some batches. Under normal conditions, the decrease of this "quantitative attribute" occurs linearly and the slopes of the regression lines relating to the lots under study are similar to each other. Usually only the intercepts differ. The product shelf life is established by looking at what time the "95 one-sided confidence limit for the curve intersects the acceptance criterion". [1,2]

Among the various operating conditions for conducting stability studies described in the ICH Q1A (R2) guideline, there are also the so-called "accelerated" ones whose purpose is to speed up the chemical degradation or change in the physical state of a *drug substance* or a *drug product*. These studies are particularly important, especially when conducted on validation batches, as they are used to estimate a possible shelf life of the product under consideration. Furthermore, they are completed long before the "long-term" studies and therefore the data under "accelerated conditions" are available for analysis within a few months (usually six) from the start of the stability studies.

The approach to the analysis of stability data described so far, which is the one used in common practice, only records the occurrence of the degradation process. The attention is in fact focused on the variation over time of a single quantitative attribute (*e.g.*, assay) and therefore, precisely by construction, this type of *univariate* data analysis can only return a limited amount of information.
At each stability time point, however, other quality attributes are also determined in addition to the assay value such as, for example: pH, water content, total impurities, *etc.* However, their information content is usually ignored and therefore lost.

In this post I want to propose a different, *multivariate approach* to the analysis of stability data and, in particular, to those pertinent to APIs aged under accelerated conditions.
From this new perspective, using all the data available at each stability time point, it is in fact possible to identify those parameters, among those that are detected, that most influence the degradation process. This allows us to hypothesize improvement actions on the process aimed at reducing, if not even minimizing, degradation and therefore, ultimately, extending the shelf life of the product itself.
As a case study, stability data obtained under "accelerated conditions" were chosen precisely because, being available before the others, they allow the degradation process to be investigated immediately, identifying any weak points and therefore also the precautions to avoid them.

2. EXPERIMENTAL SECTION

Table 1 shows the data relating to a hypothetical stability study under accelerated conditions (*e.g.*, 40°C ± 2°C / 75% RH ± 5% RH) conducted on three lots of a given advanced intermediate.

Table 1

Lot No.	Time (months)	Assay (%)	Water content (%)	Total impurities (%)	solv1 (%)	solv2 (%)	solv3 (%)	solv4 (%)	Color (AU)
1	0	99,3	0,1	0,90	0,0500	0,0170	0,0200	0,0110	0,018
	1	99,0	0,1	1,00	-	-	-	-	0,024
	2	98,9	0,1	1,10	-	-	-	-	0,039
	3	98,8	0,1	1,20	-	-	-	-	0,068
	6	98,5	0,1	1,30	0,0400	0,0120	0,0200	0,0060	0,126
2	0	99,8	0,2	1,00	0,1000	0,0160	0,0200	0,0110	0,016
	1	99,6	0,2	1,00	-	-	-	-	0,023
	2	99,5	0,2	1,10	-	-	-	-	0,038
	3	99,4	0,2	1,20	-	-	-	-	0,070
	6	99,2	0,2	1,20	0,0600	0,0100	0,0200	0,0050	0,132
3	0	99,8	0,3	1,00	0,0700	0,0150	0,0200	0,0120	0,023
	1	99,7	0,3	1,00	-	-	-	-	0,033
	2	99,6	0,3	1,10	-	-	-	-	0,048
	3	99,5	0,0	1,20	-	-	-	-	0,082
	6	99,2	0,3	1,30	0,0500	0,0100	0,0200	0,0100	0,134

As can be seen from the data in Table 1, for each of the three validation batches the determination of residual solvents was carried out only at the beginning and at the end of the study. Since continuous variables are needed to build a model, this lack of information (*missing data*) has been filled by assuming a linear trend between the two known values (*linear interpolation*).

Since from the data in Table 1 it can be observed that the residual content of solvent 3 is practically constant over time for all three batches, it will not be taken into consideration in the following investigation.

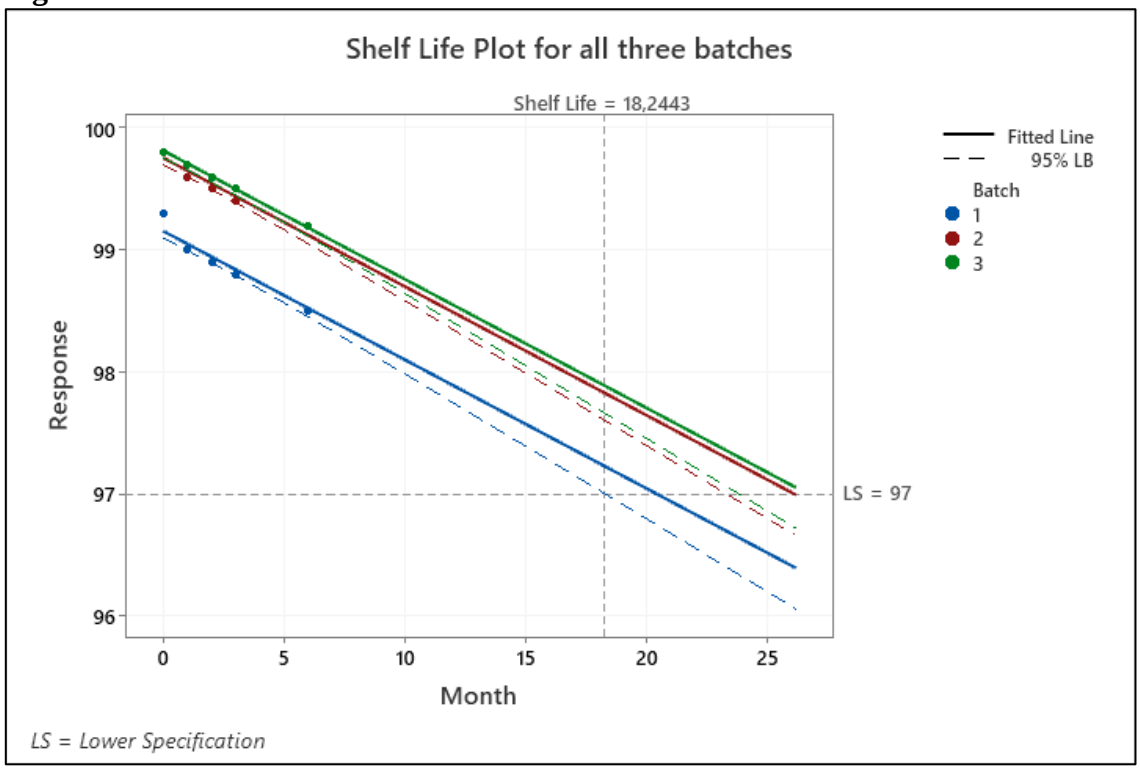
Data analysis and visualization were conducted using Minitab 20 (GMSL S.r.l. - Via Giovanni XXIII, 21 - 20014 Nerviano (Milan), Italy).

3. RESULTS AND DISCUSSION

The classic approach to stability studies, which involves the analysis over time of the assay trend, if applied to the case under study, leads to a graph like the one shown in Figure 1. From it we can estimate a shelf life of about 18 months for the advanced intermediate under investigation. In fact, the 95% confidence interval of the regression line calculated on lot 1 values meets the lower specification limit (*i.e.*, 97.0%) earlier (*i.e.*, 18.2 months) than what happens for the analogous relative confidence intervals to lots 2 and 3 (*i.e.*, 22.3 and 22.8 months).

For a quantitative attribute known to decrease with time, the term *shelf life* is generally intended as the "time period in which you can be 95% confident that at least 50% of response is above lower spec limit".

Figure 1



This graph also highlights other aspects such as, for example:

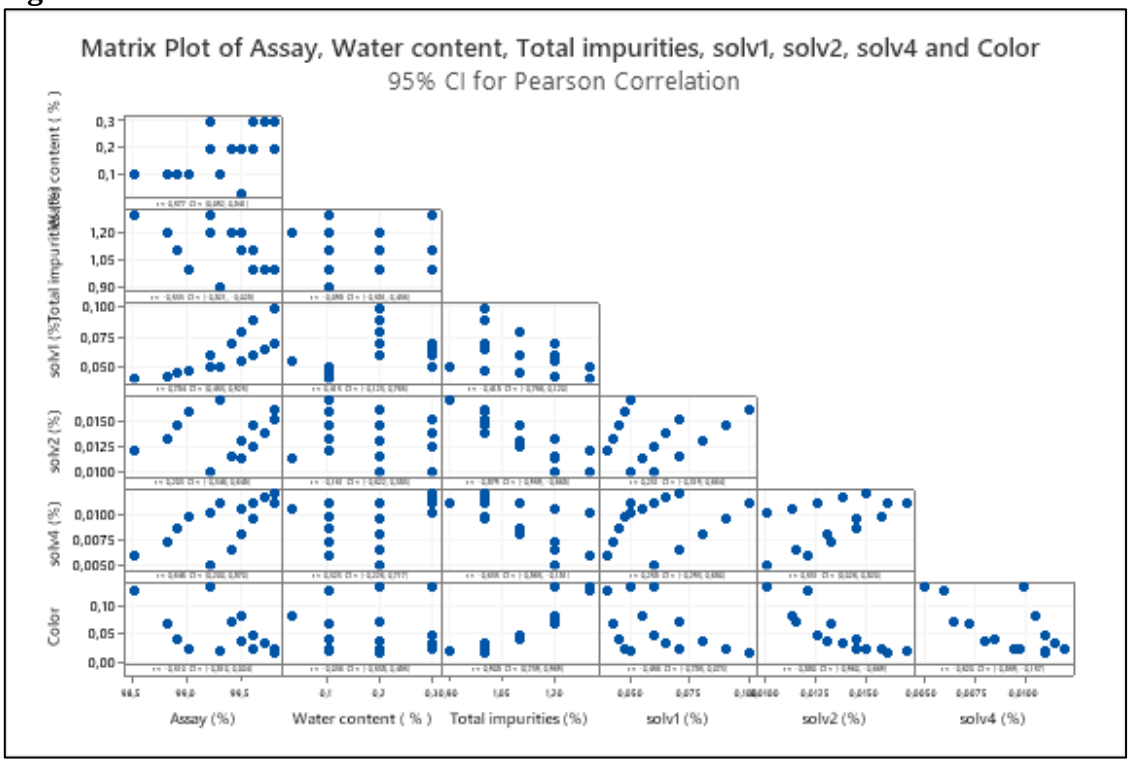
- the differences in the intercepts of the lots,
- the substantial parallelism between the slopes of the three lines, a guarantee of a common degrading behavior over time,
- the substantial similarity between lots 2 and 3 and the diversity of lot 1.

However, apart from these considerations, this approach says nothing more.

On the other hand, examining the data from a *multivariate point of view* [3], other types of findings can be obtained which are useful for understanding the degradation process just described.

As the first part of the survey, for example, the degree of linear correlation existing between the various variables measured (*i.e.*, the analytical parameters) can be examined. In this regard, Figure 2 shows the *matrix plot* resulting from the data reported in Table 1.

Figure 2



What shown in Figure 2 is a symmetric matrix (of which only the lower part is shown) whose elements are scatterplots, each relating to a given pair of variables. At the bottom of each scatterplot is reported, among others, the numerical value of the *linear correlation coefficient of Bravais - Pearson*, ***r***, pertinent to a given pair of variables.

The correlations shown here, and in particular the most significant ones, provide important information regarding the degradation process of the intermediate under study. To facilitate the investigation, it helps to have the direct quantitative estimate of the degree of linear correlation existing between the various variables, *i.e.*, the actual *correlation matrix* in which are collected the values of the linear correlation coefficients, ***r***, for each pair of variables. This matrix is represented in Table 2.

Table 2

	Assay (%)	Water content (%)	Total impurities (%)	solv1 (%)	solv2 (%)	solv4 (%)
Water content (%)	0,577					
Total impurities (%)	-0,533	-0,098				
solv1 (%)	0,784	0,415	-0,418			
solv2 (%)	0,203	-0,161	**-0,879**	0,231		
solv4 (%)	0,646	0,323	-0,635	0,258	0,531	
Color (AU)	-0,510	-0,036	**0,908**	-0,456	**-0,880**	-0,620

Despite:

- the variation ranges for each parameter are limited:

	Assay (%)	Water content (%)	Total impurities (%)	solv1 (%)	solv2 (%)	solv3 (%)	solv4 (%)	Color (AU)
Maximum	99,8	0,30	1,30	0,100	0,017	0,020	0,012	0,134
Minimum	98,5	0,03	1,00	0,040	0,010	0,020	0,005	0,016

- and the database is limited to only 15 series of values (*i.e.*, 5 time points for 3 lots)

the examination of the linear correlation coefficients in Table 2 highlights some interesting aspects from a chemical point of view as well as useful for the creation of a Multiple Linear Regression model.

In particular:

- some pairs of independent variables (*e.g.*, *total impurities* and *residual quantity of solvent 2*, *etc.*) are highly correlated with each other (*i.e.*, $\boldsymbol{r} >> |0.5|$). In the case of the *total impurities* and *color* pair, the degree of linear correlation approximates ideality ($\boldsymbol{r} = 0.908$ *vs.* $\boldsymbol{r} = |1|$).

  These correlations, which in the specific case it is reasonable to assume reflect interesting and useful chemical aspects, are also important for the construction of a model. Variables that are so highly correlated with each other should in fact be excluded to prevent problems of *multicollinearity*. In fact, the ideal is that all independent variables ($\boldsymbol{x_i}$) are significantly correlated with the dependent variable ($\boldsymbol{y}$), but not among them.

- considering *assay* as dependent variable (***y***), it is observed that, except for what concerns the residual content of solvent1 and 4 (*solv1*, *solv4*), it is not strongly correlated (*i.e.*: R >> | 0.5 |) with the available regressors, indeed:

	Water Content	**Total Impurities**	**solv1**	**solv2**	**solv4**	**Color**
assay	0,577	-0.533	**0.784**	0.203	**0.646**	-0.510

As already seen in a previous post, for the purposes of building a model it is therefore necessary, first of all, to deepen the relationship of each independent variable with the dependent variable and the relationships that may exist between the independent variables.

The analysis, carried out using scatterplots and simple linear regressions, is summarized in Table 3, below, where the independent variables are ordered on the basis of the absolute values of the linear correlation coefficient, from the largest to the smallest.

Table 3

	S	**R-sq**	**R-sq(adj)**	**Correlation**
solv1	0,248748	61,40%	58,43%	**0,784**
solv4	0,3057	41,70%	37,22%	**0,646**
water content	0,326947	33,32%	28,19%	**0,577**
total impurities	0,338788	28,40%	22,89%	**-0.533**
color	0,344484	25,97%	20,28%	**-0.510**
solv2	0,392084	4,10%	0,00%	**0.203**

Based on the above, a model was then built on the basis of the functional relationship:

$$assay = f(solv1, solv4, water\ content, total\ impurities, color, solv2)$$

Taking into account these variables and considering only the first order terms, we obtain the model described by the regression equation:

Assay (%) = 102,20 + 0,018 Water content (%) - 1,902 Total impurities (%) +
11,23 solv1 (%) - 158,1 solv2 (%) + 82,1 solv4 (%) - 1,95 Color (AU) **(1)**

Model (1) Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,104575	95,80%	92,65%	51,40%

The model is therefore based on six variables and one constant. The value of $\boldsymbol{R^2}$ (**R-sq**), which measures the percentage of variation in the data explained by the model, is, in this case, about 96%. The value of **R-sq (adj)** is close to that of **R-sq** (~ 93% *vs.* ~ 96%) precisely because of how the model was built. Only the predictive capacity of the model is significantly lower than the previous ones, in fact **R-sq(pred)** = 51.4%.

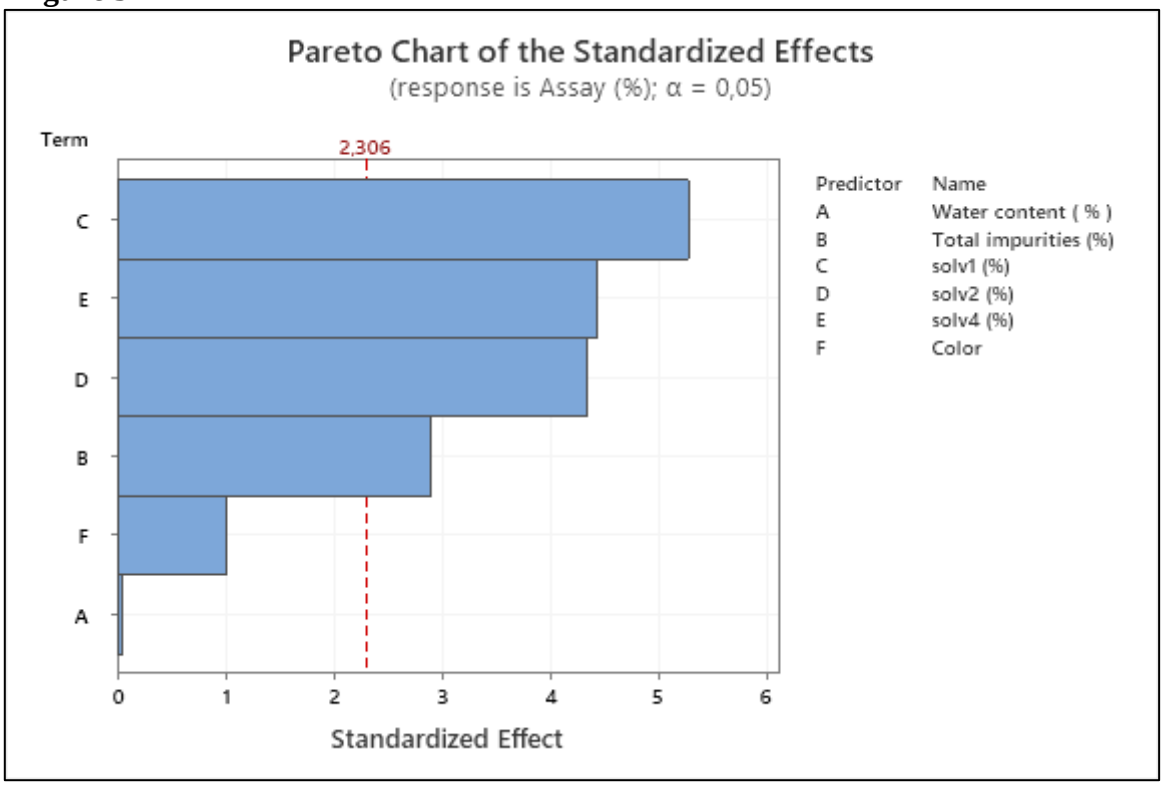
The Analysis of Variance relating to the model (1), summarized in Table 4, shows that two of the terms that appear in the model are not statistically significant, *i.e.*, they are characterized by *P-value* >> 0.05. An example for all is represented by the *water content* factor to which *P-value* = 0.966.

Table 4 - Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	1,99651	0,332752	30,43	0,000
Water content (%)	1	0,00002	0,000021	0,00	**0,966**
Total impurities (%)	1	0,09215	0,092145	8,43	0,020
solv1 (%)	1	0,30458	0,304578	27,85	0,001
solv2 (%)	1	0,20598	0,205980	18,84	0,002
solv4 (%)	1	0,21461	0,214615	19,62	0,002
Color (AU)	1	0,01129	0,011291	1,03	**0,339**
Error	8	0,08749	0,010936		
Total	14	2,08400			

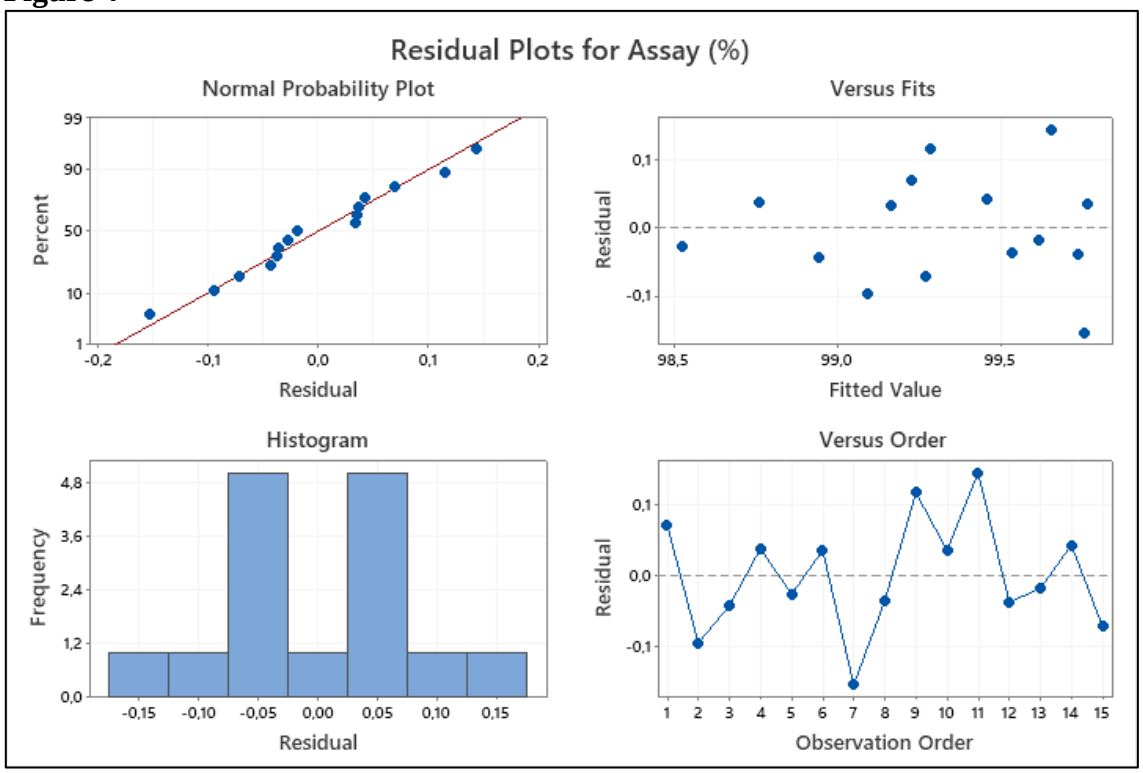
The Pareto diagram, shown in Figure 3, which distinguishes the significant effects on the process output from the insignificant ones, is in line with the results of Table 4.

Figure 3



In Figure 4, here below, are summarized the diagrams for residuals analysis.

Figure 4



The initial model, described by equation (1), was refined by progressively eliminating the insignificant terms. Proceeding in line with the findings that emerged from the Variance Analysis of Table 4, and in just two steps, the new model described by the regression equation (2) was obtained:

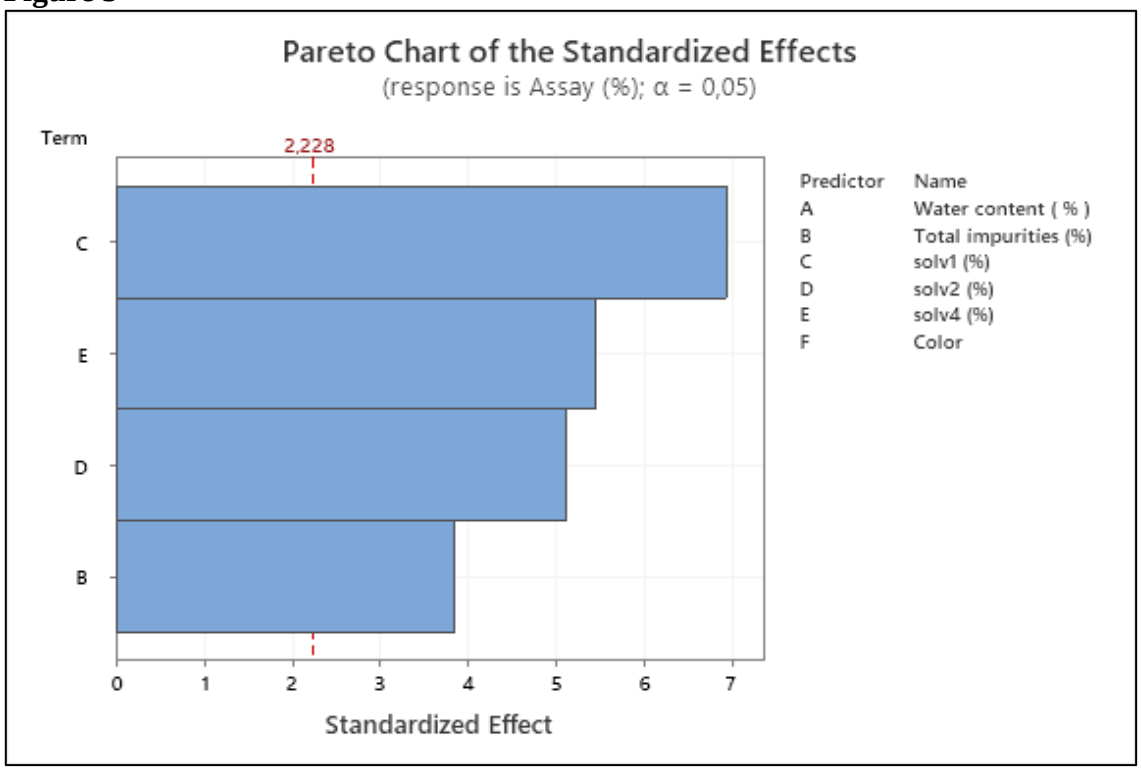
Assay (%) = 102,04 - 2,149 Total impurities (%) +12,07 solv1 (%) - 140,5 solv2 (%) + 86,0 solv4 (%) **(2)**

Model (2) Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0994065	95,26%	93,36%	90,41%

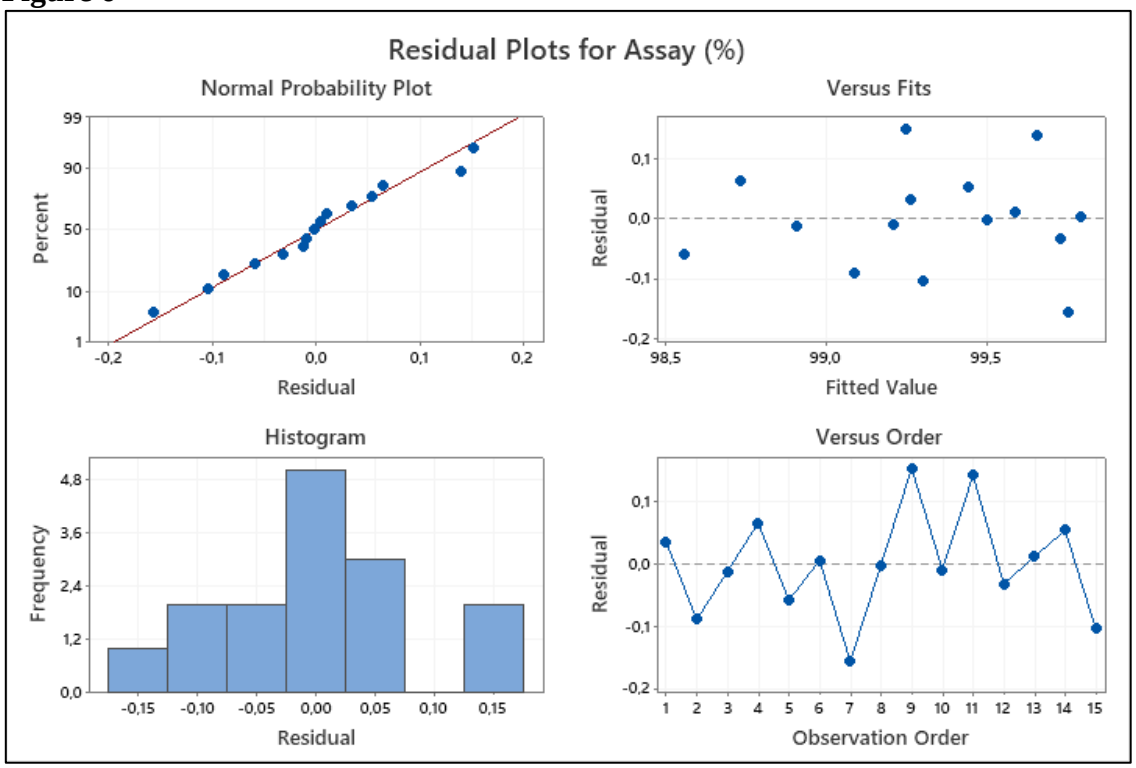
Figure 5 shows the Pareto diagram which highlights how all the factors considered are now significant.

Figure 5



The residuals analysis, summarized in Figure 6, shows a normal probability plot and a histogram with a substantially normal trend. Scatterplot and line plot show a scattering of points around zero practically free from patterns or trends.

Figure 6



In Table 5 is summarized the refinement process which has been carried out with the aim of:

- reduce the standard error value ***S***,
- keep the value of ***R-sq*** as high as possible and
- concurrently increase the values of ***R-sq(adj)*** and ***R-sq(pred)***.

Table 5

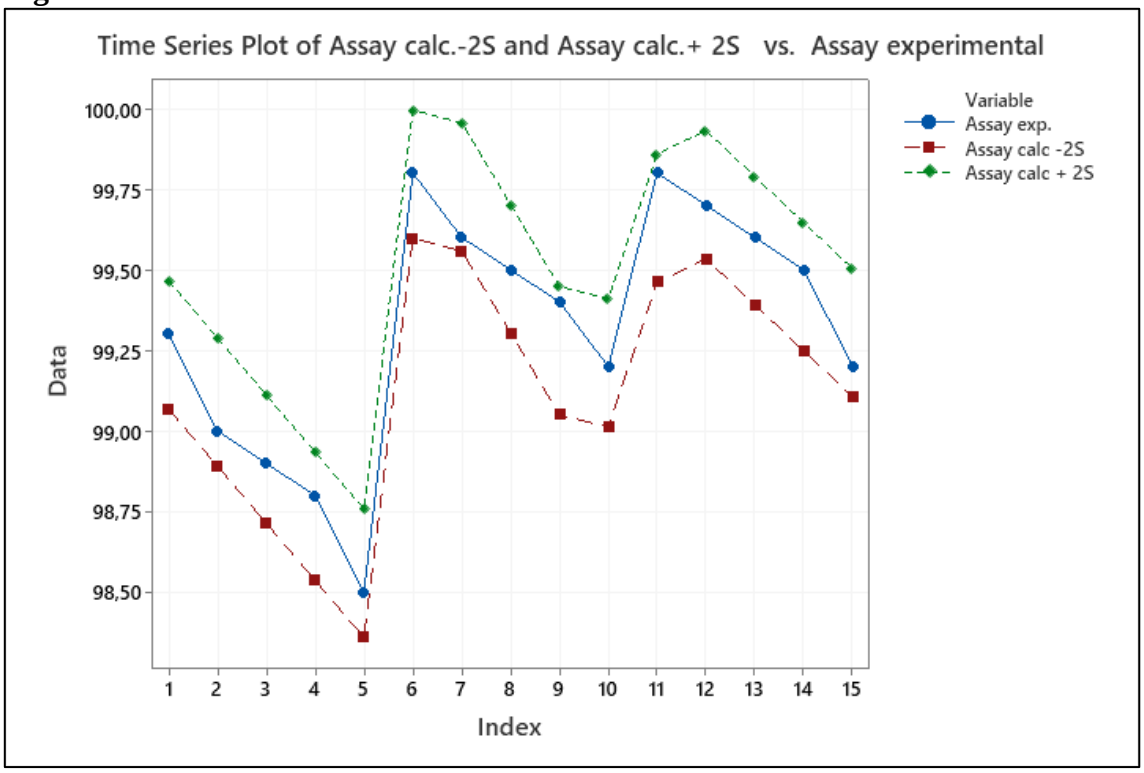
Model No.	S	R-sq	R-sq(adj.)	R-sq(pred)
1 (iniziale)	**0,104575**	**95,80%**	**92,65%**	**51,40%**
2	0,098606	95,80%	93,47%	88,34%
3	**0,099407**	**95,26%**	**93,36%**	**90,41%**

From the comparison between the two models (1) and (2) it appears that the refinement process has led to a final model (2) which has:

- a standard error ***S*** 5% lower than the initial model,
- a value of R-sq (adj) which differs from R-sq by 2% approx. compared to 3.4% that was observed in the initial model, but above all
- a final predictive capacity of 90% approx. compared to an initial ~ 51%.

The goodness-of-fit of experimental data provided by model (2) is well shown in Figure 7 where experimental data (*i.e.*, Assay exp.) are represented by a blue line while the lower limit (*i.e.*, Assay calc. - 2S) and higher limit (*i.e.*, Assay calc. + 2S), calculated using model (2) and the standard error on the regression ***S***, are represented by two red and green broken lines respectively. Examining Figure 7 it is observed that all experimental values are included among those calculated based on the model (2).

Figure 7



In light of the above evidence, the model described by equation (2) therefore represents a good approximation of the experimental assay data and can therefore be used to study the degradation process that occurred during the accelerated stability study.

To this end, the examination of the regression equation (2) provides important information:

- the factors that appear in it, and which are therefore linked in some way to the degradation process, are: total impurities and the residual content of solvents 1, 2 and 4. Solvent 3 does not appear as it was not included since the beginning.
- the three solvents all have numerical coefficients higher than that of total impurities content (*i.e.*, 12.07, 86.0, 140.5 *vs.* 2.149)
- the coefficient of variable *solv2* in the regression equation has a much higher value than that of variables *solv1* and *solv4* (*i.e.*, 140.5 *vs.* 12.07 and 86.0).

The examination of the regression coefficients is very important as they represent the average change in the dependent variable (*i.e.*, *assay*) resulting from a unit change in a given dependent variable (*e.g.*, *solv2*), keeping all the other variables constant.

Since the low *P-values* (*i.e.*, *P-value* <0.05) in the ANOVA table associated with model (2), and shown in Table 6, indicate that all dependent variables are statistically significant, it is clear that the change in the residual solvent 2 content is the one that most of all influences the *assay* value. Furthermore, having this coefficient negative sign, it follows that as the residual solvent 2 content increases, the *assay* value decreases.

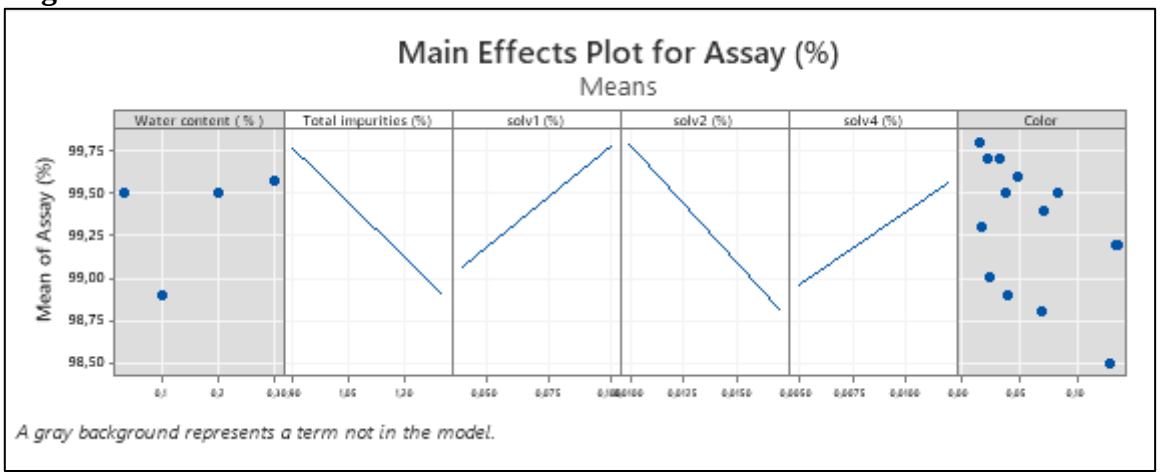
Such a result should not be surprising if we imagine, for example, that solvent 2 could be an oxygenated solvent and that the degradation process under study is characterized by the formation of oxidized species which are then responsible for the yellow color taken from the powder.

Table 6 - Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	1,98518	0,496296	50,22	0,000
Total impurities (%)	1	0,14626	0,146256	14,80	0,003
solv1 (%)	1	0,47521	0,475215	48,09	0,000
solv2 (%)	1	0,25870	0,258703	26,18	0,000
solv4 (%)	1	0,29204	0,292044	29,55	0,000
Error	10	0,09882	0,009882		
Total	14	2,08400			

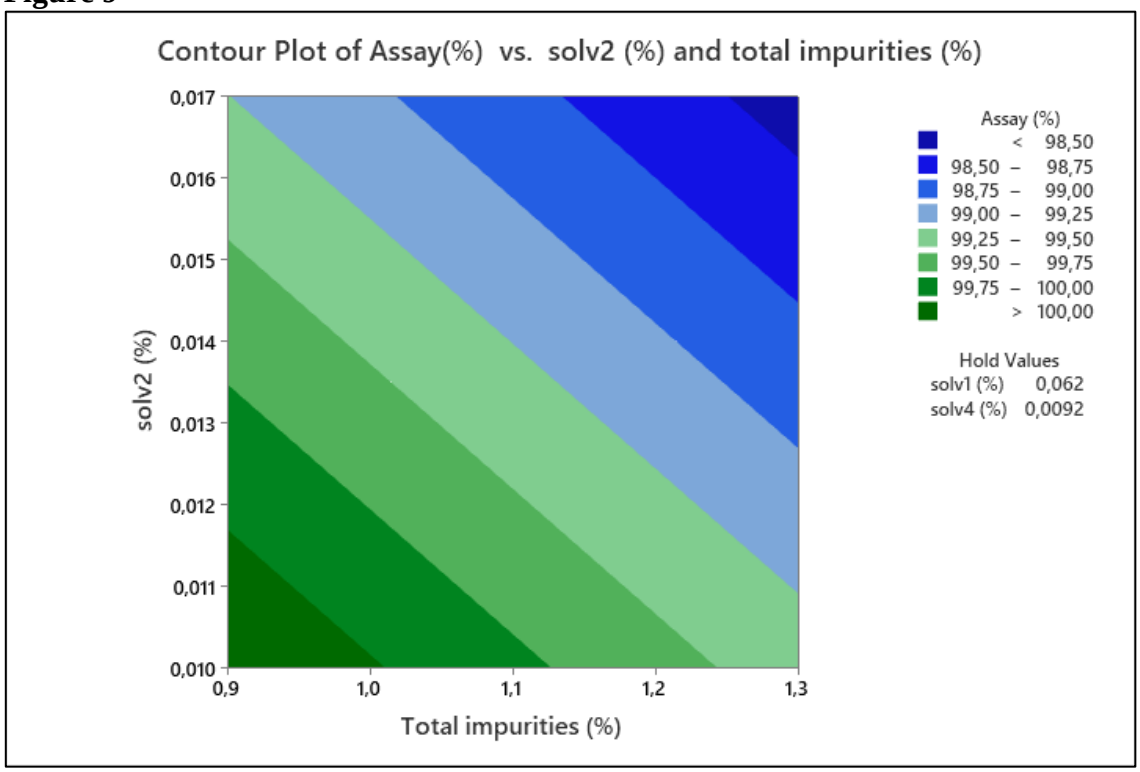
As expected from the considerations made so far, the graph in Figure 8 (MAIN EFFECT PLOT) shows the average effect on *assay* of the various independent variables. Since, in general, the steeper the segment and the more significant the effect of the factor, it is clear that all those considered in the model count. The gray fields of Figure 8 show the main effects for those factors which, despite being included in the initial model (1), do not appear in the final one (2).

Figure 8



The CONTOUR PLOT represented in Figure 9 is a graphical representation of the response variable (*assay*) as a function of *total impurities* and *residual content solvent 2* variables.

Figure 9



This graphical representation, obtained by fixing the values of the other two variables present in the model (*i.e.*, residual solvent 1 and solvent 4 content) around their average values, shows that *assay* reaches its highest values (dark green area) in correspondence to a low residue of *total impurities* and *solvent 2*. The latter variable, in fact, appears in model (2) with a minus sign and therefore its increase negatively affects the *assay* value.

In light of this, it is therefore reasonable to assume that a lower content of solvent 2 at time zero should increase the shelf life of the product by reducing the formation of impurities. This is in fact what can be observed by examining the raw data in Table 1 where it is found that batches 2 and 3, characterized by a shelf life longer than batch 1, have at time zero solvent 2 contents slightly lower than that of batch 1 (*i.e.*, 0.0160% and 0.0150% *vs.* 0.0170%).

It is interesting to note that, in the case under study, even using only the stability data obtained up to the third month under accelerated conditions, after refinement, a model quite similar to (2) is still obtained, and precisely:

$$\text{Assay (\%)} = 100{,}85 - 1{,}357\ \text{Total impurities (\%)} + 11{,}92\ \text{solv1 (\%)} - 125{,}1\ \text{solv2 (\%)} + 101{,}8\ \text{solv4 (\%)} \quad \textbf{(3)}$$

Comparing (3) with (2) we see that, apart from some differences in the numerical values of the coefficients, the algebraic signs are the same in both models.

Even the performance of model (3) is comparable to that found for model (2), in fact:

Model (3) Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0950028	95,10%	92,30%	86,47%

The combination of these evidences therefore suggests that, albeit with due precautions, concrete assessments can still be made by having only the data of the third month accelerated.

4. CONCLUSIONS

Stability studies are a mandatory activity that, in general, is routinely conducted and equally routinely monitored as *per* official guidelines.

However, this activity plays a particularly important role in the initial phase, *i.e.*, when studying validation batches.

The traditional approach to stability studies is limited exclusively to recording the occurrence of a degradation process, with the sole purpose of estimating a possible shelf life for the product.

This approach, due to its *univariate* nature, however, is not able to say anything about the possible causes of the degradation phenomenon and eventually suggest a way to improve things.

The *multivariate approach*, on the other hand, by fully grasping the relationships that exist between the different analytical parameters that are measured, and which define the evolution of the purity profile over time, reveals aspects that would otherwise go unnoticed.

In the case study chosen, for example, it was shown that in the presence of three validation batches, two of which are more similar, the multivariate approach identified the residual content of a solvent (*i.e.*, solvent 2) as a possible and significant cause of the process degradative. In this way, the conditions are created for an improvement of the process that can be controlled in a scientific and targeted way.

Experimentally it was also observed that even with only the data of the third month it was possible to obtain a model similar to that obtained with the data of the sixth month, thus allowing us to advance hypotheses regarding the degradation process underway already three months after the start of the studies.

Considering this, it is therefore reasonable to assume that the use of additional accelerated aging techniques (*e.g.*, $40°C \leq T \leq 80°C$ and $10\% \leq RH \leq 75\%$) will make the data available for analysis in an even shorter time frame.[4]

In any case, it is essential that, for a correct interpretation of the results, the data analyzes are continuously verified against an in-depth chemical knowledge of the process.

5. BIBLIOGRAPHY

1. ICH Q1A(R2), *Stability testing of new drug Substances and Products*, February 2003
2. ICH Q1E, *Evaluation for Stability Data*, February 2003
3. D.C. Montgomery, E.A. Peck, G.G. Vining, *Introduction to Linear Regression Analysis*, 5th Ed., (2012) Wiley
4. K.C. Waterman, R.C. Adami, *Accelerated aging: Prediction of chemical stability of pharmaceuticals*, Int. Journal of Pharmaceutics 293 (2005) 101–125