Applied Mathematics, 2015, 6, 1947-1956
Published Online October 2015 in SciRes. http://www.scirp.org/journal/am
http://dx.doi.org/10.4236/am.2015.611172

Scientific Research Publishing

The CARSO (Computer Aided Response Surface Optimization) Procedure in Optimization Studies

Mauro Fernandi[1], Massimo Baroni[2], Matteo Bazzurri[2], Paolo Benedetti[2], Danilo Chiocchini[3], Diego Decastri[1], Cynthia Ebert[4], Lucia Gardossi[4], Giuseppe Marco Randazzo[2], Sergio Clementi[3*]

[1]Flint Group, Cinisello Balsamo, Milano, Italy
[2]Biology, Biotechnology and Chemistry Department, Section Chemistry, University of Perugia, Perugia, Italy
[3]M.I.A. Multivariate Infometric Analysis Srl, Perugia, Italy
[4]Department of Chemistry and Pharmaceutical Sciences, University of Trieste, Trieste, Italy
Email: [*]mia@miasrl.com

Received 15 September 2015; accepted 27 October 2015; published 30 October 2015

Abstract

The paper illustrates innovative ways of using the CARSO (Computer Aided Response Surface Optimization) procedure for response surfaces analyses derived by DCM4 experimental designs in multivariate spaces. Within this method, we show a new feature for optimization studies: the results of comparing their quadratic and linear models for discussing the best way to compute the most reliable predictions of future compounds.

Keywords

CARSO Procedure, Design Strategies, Double Circulating Matrices (DCMs), Optimization

1. Introduction

Following our recent papers [1]-[3], where we showed a new strategy for collecting the data needed for defining a response surface on the basis of an innovative method that requires only a very low number of experimental data, based on Double Circulant Matrices (DCMs). They are similar to Central Composite Designs (CCDs), which represent the best way to generate response surfaces. The final surface model is obtained by the formerly

[*]Corresponding author.

How to cite this paper: Fernandi, M., Baroni, M., Bazzurri, M., Benedetti, P., Chiocchini, D., Decastri, D., Ebert, C., Gardossi, L., Randazzo, G.M. and Clementi, S. (2015) The CARSO (Computer Aided Response Surface Optimization) Procedure in Optimization Studies. *Applied Mathematics*, **6**, 1947-1956. http://dx.doi.org/10.4236/am.2015.611172

developed CARSO method [4], where the surface equation is derived by a PLS model on the expanded matrix, containing linear, squared and bifactorial terms.

This paper is meant to discuss if this statistical form of the X block is the most appropriate within chemical studies of experimental properties either in terms of mixtures, or in terms of literature constants, mainly electronic and steric, reported for each substituent. For a long time, organic chemists collected data for a class of similar compounds (the field was called correlation analysis) and computed the slope of the resulting straight line, thus comparing the behaviour of different chemicals in the same reaction.

After a full immersion of one of us (SC, 1981) in the group of Svante Wold at Umea (Sweden) to learn chemometrics and to use their software SIMCA [5], a few of us decided to enter into this new fascinating world. In order to drive the interest of our colleagues towards a more rewarding result, e.g. so that the study should find the best conditions for optimizing the property or the process, we published CARSO in 1989 [4], and GOLPE in 1993 [6]. Since then we recognized that our procedures were children of SIMCA.

Now we discuss the best way to compute reliable predictions. It is clear that, for an optimization study, the first must is a plan of the experiments, but this is outside the objectives of this paper. See, for example, reference [1].

2. Optimization Studies: The Models

This paper is meant to make clear how an optimization study should be carried out. The first choice should be an experimental plan as round as possible: we selected the double circulant matrices (DCMs) that represent the maximal roundness [1]. The next thing to be done is to find how the y response(s) can be described by the best possible relationship in terms of the x values. In principle it is possible to use linear or non linear models. The linear models are simpler to be interpreted, but it is unusual that the response depends only by a single descriptor. Usually the response is affected by a complex mixture of different effects that contribute simultaneously: therefore a bell form should appear more appropriate to give a better model. Therefore each model will be evaluated by different approaches, namely the variance explained, the standard deviation and the intervals covered by the recalculations or predictions. Indeed we wish to show numerically the relative performances of linear and quadratic models.

2.1. Collection of Results by DCM4

Our experimental data set contains 13 different mixtures (objects) generated by different relative amounts of 4 variables according to a strategy in keeping to the experimental design illustrated by a DCM4 [2]. At the time we wanted to show which could be the minimum number of measurements needed to build a response surface (and we suggested to use only two of the six possible quartets of mixture, namely sm1 and sm3, plus two replicates at the centre point, *i.e.* 10 points). Indeed, in this case, with the objective to compare the soundness of our models, we prefer using a model with all three first quartets of the DCM4 and one single central point only [2].

Table 1 reports the results of our experimental data set. For each of the 13 different mixtures (objects) are listed the real value of the y (the inverse of the contact angle) and their coded values between 0 and 100, that helps to understand immediately if the following predictions are within or outside the range of the collected data.

Table 2 lists the recalculation by the three statistical models Q1, L1 and L2, indicating the variance explained. The ranges of each of the models are somewhat different. They span the ranges reported in the last rows of the Table, even showing the maximum and minimum values and their mean averages. On comparing the three models we can see that the Q1 model have a range of 103.8 (between −8 and 95.8) with an average of 43.9, while the L1 model have a range of 92.3 (between −16.9 and 98.2) with an average of 29.3, and the L2 model have a range of 114.7 (between −16.1 and 98.2) with an average of 41.3.

The data of the statistical model Q1 is a quadratic model with only one latent variable, which contains all linear, quadratic and bifactorial terms (and the PLS uses all the terms, as if all of these were “new” pseudolinear data), whereas L1 has only linear terms and one latent variable, and L2 has also only linear terms but computed by two latent variables, even if this type of modelling is well known only by researchers with confidence in chemometrics.

Table 2 reports the predictions, for the three models, on the 12 objects of this collection and excluding the central point, because it gives three different predictions and there is no way to choose the best of them.

Table 1. Selected experimental plan (training set).

N	ogg	1	2	3	4	1/CA	15 = 1/CA
		x1	x2	x3	x4	y sper	y cod
1	E11	−1	−0.58	0.58	1	8.13	0
2	E12	−0.58	0.58	1	−1	13.12	93.3
3	E13	0.58	1	−1	−0.58	8.46	6.2
4	E14	1	−1	−0.58	0.58	8.98	15.9
5	E21	−1	−0.58	1	0.58	8.45	6
6	E22	−0.58	1	0.58	−1	13.48	100
7	E23	1	0.58	−1	−0.58	8.96	15.5
8	E24	0.58	−1	−0.58	1	8.24	2.1
9	E31	−1	0.58	1	−0.58	12.85	88.2
10	E32	0.58	1	−0.58	−1	10.52	44.7
11	E33	1	−0.58	−1	0.58	8.15	0.4
12	E34	−0.58	−1	0.6	1	8.35	4.1
13	CC	0	0	0	0	8.33	3.7
Max						13.48	100
Min						8.13	0
Ave						10.81	50
Ran						5.35	100

Table 2. Modelling of the selected experimental plan (training set).

			Q1	d	d2	L1	d	d^2	L2	d	d^2
N	ogg	y cod	93%			67%			80%		
1	E11	0	−1.6	1.6	2.6	18.6	18.6	346	−5.9	5.9	34.8
2	E12	93.3	95.8	2.5	6.3	74	19.3	372.5	98.2	4.9	24
3	E13	6.2	17	10.8	116.6	37.5	31.3	979.7	12.9	6.7	44.9
4	E14	15.9	6.4	9.5	90.3	−13.2	29.1	846.8	11.8	4.1	16.8
5	E21	6	10.9	4.9	24	31.4	25.4	645.2	31.3	25.3	640.1
6	E22	100	94.4	5.6	31.4	75.4	24.6	605.2	74.9	25.1	630
7	E23	15.5	6.9	8.6	74	27.1	11.6	134.6	27.1	11.6	134.6
8	E24	2.1	6.4	4.3	18.5	−16.9	19.1	364.8	−16.5	18.6	346
9	E31	88.2	86.3	1.9	3.6	70.3	17.9	320.4	69.9	18.3	334.9
10	E32	44.7	32.2	12.5	156.3	50.3	5.6	31.4	50.1	5.4	29.2
11	E33	0.4	7.9	7.5	56.3	−11.8	12.2	148.8	−11.5	11.9	141.6
12	E34	4.1	−8	12.1	146.4	8.2	4.1	16.8	8.4	4.3	18.5
Sum					726.0			4812.2			2394.4
Ave					60.5			401.0			199.5
STD					7.8			20.0			14.1
Max			95.8			75.4			98.2		
Min			−8			−16.9			−16.5		
Range			103.8			92.3			114.7		

On comparing the three models by their explained variance and STD (standard deviation), it is clear that Q1 is by far the best, showing a 93% of explained variance and a STD of 7.8, followed by L2, showing a 80% of explained variance and a STD of 14.1, and L1 showing a 67% of explained variance and a STD of 20.0.

2.2. Predictions on Hypervertices

In order to evaluate the relative sizes of variations of the three models shown in **Table 2**, we report the predictions, for each of the three models, on the 16 points located on the hypervertices of the 4D space, to study comparatively the span generated by each model. Therefore the values reported in **Table 3** could allow to find out which of the models span a greater interval between the minimum and maximum. The list of the vertices is ordered according to the Yates proposal.

2.3. Comparison of External Predictions (Hypervertices)

The comparison of the predictions listed in **Table 3** allows to evaluate the span for each of the three models used on the same set of experimental data. We remind that examining the real data we already reported the standard deviations: the Q1 model gives the best result (7.8), followed by L2 (14.1) and L1 (20).

On inspecting the external predictions we observe that Q1 shows an interval of 130.0 (from −12.7 to 117.3), while L1 shows an interval of 110.3 (from −25.9 to 84.4) and L2 shows an interval of 245.2 (from −93.4 to 151.9). This result clearly shows that this model (linear model with two latent variables), independently of other good parameters, can gives very risky external predictions outside the explored range. Because of that we stopped to continue to explore the characteristics of the L2 model.

2.4. Revision of Recalculations/Predictions

The data discussed after **Table 2** for the training set and after **Table 3** for the testing set clearly indicate that the quadratic models might be defined to be the most reliable because the intervals of predictions are closer to 100 (similar to the L1 one, but different from the L2 models).

Table 3. Predictions on the 16 hypervertices (testing set).

N	name	x1	x2	x3	x4	Q1	L1	L2
14	H1	−1	−1	−1	−1	27.5	29.2	29.2
15	H2	1	−1	−1	−1	6.6	10.4	69.2
16	H3	−1	1	−1	−1	57.8	60.0	1.2
17	H4	1	1	−1	−1	10.2	41.2	41.1
18	H5	−1	−1	1	−1	54.6	53.6	111.9
19	H6	1	−1	1	−1	32.8	34.8	151.9
20	H7	−1	1	1	−1	117.3	84.4	83.9
21	H8	1	1	1	−1	68.8	65.6	123.8
22	H9	−1	−1	−1	1	−0.6	−7.1	−65.4
23	H10	1	−1	−1	1	12.0	−25.9	−25.4
24	H11	−1	1	−1	1	22.1	23.7	−93.4
25	H12	1	1	−1	1	8.1	4.9	−53.5
26	H13	−1	−1	1	1	−12.7	17.2	17.4
27	H14	1	−1	1	1	−1.0	−1.6	57.3
28	H15	−1	1	1	1	42.5	48.0	−10.7
29	H16	1	1	1	1	27.5	29.2	29.2
MAX						117.3	84.4	151.9
MIN						−12.7	−25.9	−93.4
Range						130.0	110.3	245.2

These results show two significant characteristics to be interpreted: the ranges and the averages of recalculations/predictions of each model. The Q1 models cover roughly the range between 0 and 100. The L1 models have the smaller intervals of variation (92.3 for recalculations and 110.3 for predictions), but they are all shifted towards the lower part of the collected data. The L2 models have by far the larger intervals covering predictions data much larger than the highest ones and lower than the smallest figure, and therefore we did not investigate it any more.

However all these arguments (besides the standard deviations and the ranges of recalculations/predictions) might appear to be somewhat too poor to claim that the quadratic model is better than the linear ones.

Therefore we decided to approach our problem by constructing a unique list of predicted y values obtained by "inner" models, *i.e.* using the values predicted for each object by modelling using the two submatrices not including the object, and computing a sort of self-predictions of the left out objects for each model based on a couple of the three submatrices, as we always applied for validating each model.

3. Inner Predictions by Each Couple of Submatrices

In order to evaluate the predictions by partial models, so that the objects left out are really "predicted" and not "recalculated", we computed the two models using only the two submatrices that does not contain the object to be evaluated for correctly obtaining "external" or "inner" predictions.

The quadratic predictions listed in **Table 4** are computed for the objects of the left out submatrix, reported in italics, and allow to compute the standard deviations of the predictions. **Table 5** lists the linear predictions with one latent variable The results shown in **Table 4** and **Table 5** allow to compare the models by their standard deviation: 11.8 for the Q1 model, 21.3 for the L1 model.

Moreover we can compare the intervals found for each model. In order to evaluate the predictions by partial models, so that the objects left out are really "predicted" and not "recalculated", we computed the two models using only two submatrices for correctly obtaining "inner" predictions for each object of the third one. The interval found for the Q1 model is 100.7, while for the L1 model is 91.2. Therefore the intervals of both models Q1 and L1 with inner predictions are very similar.

Table 4. Quadratic inner predictions (1LV) compared with coded experimental data.

Ogg	Sm 1 + 2	Sm 1 + 3	Sm 2 + 3	Inner	Coded	Delta	Delta^2
LV%	92	88	94	y	y		
E11	−6.8	2.8	0.8	0.8	0	0.8	0.64
E12	93.4	92.6	92.4	92.4	93.3	0.9	0.81
E13	11.5	13.8	25.8	25.8	6.2	19.6	384.16
E14	11.7	6.9	0.5	0.5	15.9	15.4	237.16
E21	2.2	18.9	11.9	18.9	6	12.9	166.41
E22	92.9	84.5	87.3	84.5	100	15.5	240.25
E23	0.4	8.7	12.5	8.7	15.5	−6.8	46.24
E24	15.1	4.9	−1.2	4.9	2.1	2.8	7.84
E31	79.2	86.4	85.6	79.2	88.2	−9	81.00
E32	28.8	24.6	42.0	28.8	44.7	15.9	252.81
E33	10.8	12.2	0.4	10.8	0.4	10.4	108.16
E34	−8.3	−6.1	−7.4	−8.3	4.1	12.4	153.76
			Max	92.4		Sum	1681.1
			Min	−8.3		Ave	140.1
			Range	100.7		STD	11.8
			ave	28.9			

Table 5. Linear inner predictions (1LV) compared with coded experimental data.

Ogg	Sm 1 + 2	Sm 1 + 3	Sm 2 + 3	Inner	Coded	Delta	Delta^2
LV%	57	65	73	y	y		
E11	14.9	20.9	17.1	17.1	0	17.1	292.41
E12	71.4	73.7	74.7	74.7	93.3	18.6	345.96
E13	35.5	32.5	41.7	41.7	6.2	35.5	1260.25
E14	−13.9	−13.1	−15.9	−15.9	15.9	31.8	1011.24
E21	28	34.3	29.1	34.3	6	28.3	800.89
E22	72.3	73.5	78	73.5	100	26.5	702.25
E23	25.9	22.7	29.7	22.7	15.5	7.2	51.84
E24	−18.4	−16.5	−19.2	−16.5	2.1	18.6	345.96
E31	66.9	70.4	71.7	66.9	88.2	21.3	453.69
E32	48.6	45.8	53.8	48.6	44.7	3.9	15.21
E33	−12.9	−13.4	−12.6	−12.9	0.4	13.3	176.89
E34	5.3	11.2	5	5.3	3.7	1.2	1.44
			Max	74.7		Sum	5458.03
			Min	-16.5		Ave	454.84
			Range	91.2		STD	21.3
			Ave	29.1			

4. Summary of the Data Collection

The data reported in the Tables give a clear answer to our question: which of the models can be defined the most reliable for new predictions, to be computed by the data of **Table 1**. The discussion of these results clearly shows that the Q1 model is by far the best, with respect to L1, while L2 is the worst, because its predictions cover intervals from very high to very low, outside from those of the experimental data.

Because of this the best way of describing the trends of a series of compounds appear to be a quadratic model, that finds out reliable results, usually within the explored space. On the contrary the linear model with one latent variable gives predictions within a much smaller interval, which is also shifted download, towards lower numbers, while the linear model with two latent variables (which is less used by researchers) spans a much larger space, with quite higher and lower results.

However the data listed in **Table 1** can be considered self-referenced, since they are all experimental values, not evaluated by predictions. On the contrary, the results reported in **Table 4** and **Table 5** are real predictions obtained by the differences between the coded experimental data and their inner values.

5. Dissection of the Information

This approach can be used as an alternative way to compare the relative reliability of the quadratic and linear models based on the dissection of information according to the Pythagoras' theorem. Given a non expanded DCM4 it is not possible to execute a Multiple Linear Regression. Indeed the expanded DCM4 cannot be treated by MLR because of two reasons:

a) The number of objects (13) is smaller than the number of variables (14);
b) In the linear blocks each column of the DCM4 is a linear combination of the other three.

The only possibility of running MLR on a DCM4 is eliminating one column (say x4), and adding on the left a column of numbers "1" for determining the intercept. The coefficients of MLR are listed in **Table 6**, and it is possible to recalculate each object listed in **Table 7**.

Table 6. Coefficients of multiple linear regression.

y	x0	x1	x2	x3		
0.0	1	−1.00	−0.58	0.58		Coeff.
93.3	1	−0.58	0.58	1.00	x0	29.24
6.2	1	0.58	1.00	−1.00	x1	67.25
15.9	1	1.00	−1.00	−0.58	x2	33.28
6.0	1	−1.00	−0.58	1.00	x3	88.65
100.0	1	−0.58	1.00	0.58		
15.5	1	1.00	0.58	−1.00		
2.1	1	0.58	−1.00	−0.58		
88.2	1	−1.00	0.58	1.00		
44.7	1	0.58	1.00	−0.58		
0.4	1	1.00	−0.58	−1.00		
4.1	1	−0.58	−1.00	0.58		
3.7	1	0.00	0.00	0.00		

Table 7. Dissection of information.

x0	x1	x2	x3	Exp	Rec y	TSS	MSS	RSS
1	−1.00	−0.58	0.58	0.0	−5.90	−29.2	−35.14	5.90
1	−0.58	0.58	1.00	93.3	98.18	64.1	68.94	−4.88
1	0.58	1.00	−1.00	6.2	12.88	−23.0	−16.36	−6.68
1	1.00	−1.00	−0.58	15.9	11.80	−13.3	−17.44	4.10
1	−1.00	−0.58	1.00	6.0	31.33	−23.2	2.09	−25.33
1	−0.58	1.00	0.58	100.0	74.93	70.8	45.69	25.07
1	1.00	0.58	−1.00	15.5	27.15	−13.7	−2.09	−11.65
1	0.58	−1.00	−0.58	2.1	−16.45	−27.1	−45.69	18.55
1	−1.00	0.58	1.00	88.2	69.94	59.0	40.70	18.26
1	0.58	1.00	−0.58	44.7	50.11	15.5	20.87	−5.41
1	1.00	−0.58	−1.00	0.4	−11.46	−28.8	−40.70	11.86
1	−0.58	−1.00	0.58	4.1	8.37	−25.1	−20.87	−4.27
1	0.00	0.00	0.00	3.7	29.24	−25.5	0.00	−25.54
						17971	14927	3044

Table 7 shows the dissection of information according to the Pythagoras' theorem: TSS = MSS + RSS, where TSS = sum[yexp − (y ave exp)]^2 (total information = 17,971), MSS = sum[rec − (y ave exp)]^2 (information explained by the model = 14,927), RSS = sum[yexp-rec (y by MLR)]^2 (information not explained by the model = 3044). Formally the word "information" should be substituted by the stastistical term "deviance", but in chemometrics we prefer the term information which will be understood by a larger number of readers.

On applying the same computations with any other possible triplet of variables (x2, x3, x4 shown; x1, x3, x4; x1, x2, x4; x1, x2, x3) we obtained always different results for the MLR coefficients (excluding the intercept):

see **Table 8**.

Although we found that the MLR coefficients are different on using diverse triplets of variables it is noteworthy that the vectors of the y values computed (not shown) by the data of one of them (x1, x2, x3) is identical to the other one (x2, x3, x4).

This surprising result, observed also for the triplet (x1, x2, x4) even if the variables are not listed in sequence, may be attributed to the roundness of the DCM. As a consequence this happens even if the variables are not in an ordered sequence.

6. Using PLS Instead of MLR

In this section we show what happens on using PLS instead of MLR.

The first appearance of the PLS algorithm in the chemical literature was a merit by Herman Wold in 1966, and followed by many others, among which his son Svante. Half century later PLS showed to be much more reliable for finding quantitative relationships between chemical structure and properties.

Obviously the old version of MLR cannot work both on the expanded and the non expanded matrices, because their rank is not the same of the number (k) of independent variables. In other words the squared matrix (XtX), of order k, cannot be inverted.

Applying PLS on the expanded matrix (model called Q1) and on the non expanded matrix (model called L1, because it keeps only the first latent variable) we obtained the results reported in **Table 9**.

Table 8. MLR coefficients for two triplets of variables.

	Coeff.	x0, x1, x2, x3 MLR coefficients		x0, x2, x3, x4 MLR coefficients	N Obj	Rec y	TSS	MSS	RSS
x0	29.24	29.24	x0	29.24	1	−5.9	−29.2	−35.1	5.90
x1	67.25	67.25	x2	−33.97	2	98.2	64.1	68.9	−4.88
x2	33.28	33.28	x3	21.40	3	12.9	−23.0	−16.4	−6.68
x3	88.65	88.65	x4	−67.25	4	11.8	−13.3	−17.4	4.11
					5	31.3	−23.2	2.1	−25.33
					6	74.9	70.8	45.7	25.07
					7	27.1	−13.7	−2.1	−11.65
					8	−16.5	−27.1	−45.7	18.55
					9	69.9	59.0	40.7	18.26
					10	50.1	15.5	20.9	−5.41
					11	−11.5	−28.8	−40.7	11.86
					12	8.4	−25.1	−20.9	−4.27
					13	29.2	−25.5	0.0	−25.54
							17971.3	14927.4	3044.0

Table 9. Dissection of the information for models Q1 and L1.

	Q1	L1
RSS	1194.7	5461.9
MSS	16682.0	13250.9
RSS + MSS	17876.7	18712.8
TSS	17971.3	17971.3
Abs (TSS − MSS − RSS)	94.6	741.6

The results reported in **Table 9** show that only Q1 mimics the relationship

$$TSS = MSS + RSS$$

which has an interesting geometrical interpretation in the 2D space. In other words this relationship simulates the geometrics of the Pythagoras' theorem for a right triangle having a hypotenuse of lenghth^2 equal to TSS, the longer side having a lenghth^2 equal to MSS, and the shorter one having a lenghth^2 equal to RSS.

Furthermore the data of **Table 9** also allows to demonstrate the superiority of the Q1 model over L1 because:

a) Q1 shows that the weight of the MSS component is greater than that of RSS;
b) Q1 is the closest to the ideal null value of Abs(TSS-MSS-RSS), the indicator of the geometric idealistic, while L1 is much more distant;
c) In conditions of almost idealistic geometrics (typical of MLR) Q1 shows a RSS value of 1195, whereas L1 shows a value of 5461: this means that the Q1 model preserves only the 22% of the data involved in the L1 model, but gives a better picture of the situation.
d) This means that, under these conditions, we can eliminate 78% of information that appears to be non systematic.

7. Revaluation of the CARSO Software

Besides the comments given so far, we wish to remember that the core of this paper is the CARSO procedure illustrated in ref. 4, published in spring 1989. A few months later a similar paper was published in ref. 8 by our Swedish friends. At the time we suggested to apply this new approach to any data set to be used in optimization studies. Therefore we considered CARSO a possible new module to be inserted into the SIMCA software.

The CARSO module makes a simple, but significant, change of the matrix, that is expanded, on adding the squares and the cross products terms to the linear ones. This approach, some 25 years ago, allowed to model by PLS the expanded matrix, that is still used, in the mode now called Q1, as we showed in this paper.

Indeed, at the time, the main interest of the quadratic model (the linear one cannot give this information) was focused on the search of the operative intervals of each independent variable for the optimization of the y variable(s). Because of this we could use the canonical analysis, searching the coordinates for a maximum, if it exists, or for the stationary points, within the explored space, or even the extreme points on the frontier of the experimental domain. In other words the CARSO method is a full software tool for optimization studies.

Today this practice is not widely used [7] [8], because a complex system cannot be described by linear terms only, but it requires to involve deviations from linearity (squared terms) and the synergic and antagonist effects (cross product terms). In general, on increasing the polynomial degree used for interpolating the data, grows the fitting and decreases the predictability. The second degree function needed for describing the response surface is the best compromise between these two properties, without involving higher terms.

The equality [TSS = RSS + MSS], that can be represented graphically by the relationships of the areas of the squares built on the sides of a rectangular triangle can be applied (high degree of approximation) only to the Q1 model. Furthermore the Q1 model (CARSO) lowers the unexplained information (RSS) with respect to the linear model. To sum up we can state that the CARSO power in this dataset is the result of both the Q1 power multiplied by the DCM power, so that also objects sum up to zero.

Four years later we published a further paper [6] on almost the same problems (GOLPE: Generating Optimal Linear PLS Estimations) where the declared main field of application was PCA and PLS models. The software, that includes the CARSO modules, adds partial information not shown in SIMCA, but the most important data are limited to the results of the module COEFFER that transforms the PLS loadings (called also gamma) into the coefficients of the second degree polynomial.

8. Conclusions

This paper was done for showing that in optimization studies it is needed to use a quadratic model. In other words, it means that only this model can be used for deriving reliable predictions of further compounds. This has been shown numerically here, but this is also implied into this problem the need of requiring a hyperbell for finding out the operative intervals. Because of this, the best way of describing the trends of a series of compounds is a quadratic model that finds out reliable results, usually within the explored space. On the contrary, the linear model with one latent variable gives much lower data, which seems unreliable.

This choice is in keeping with the position referred by Rosipal [7], who discussed the nonlinear PLS models on the basis introduced by Svante Wold *et al*. [8]. Because of this, the best way of describing the trends of a series of compounds is a quadratic model that can be found by using a linear model on the expanded matrix, as we did since the beginning.

The main goal of this paper was finding out which statistical method was more reliable for computing the predictions of new objects outside the training set used: we took into account a quadratic model and a linear model and we could demonstrate that the quadratic model is by far the best.

Acknowledgements

The authors wish to thank Dr. Matthias Henker (Flint Group Germany) for financing a partnership contract with MIA srl and Prof. Svante Wold and his former coworkers in Umeå (Sweden) for introducing SC to chemometrics.

References

[1] Clementi, S., Fernandi, M., Decastri D. and Bazzurri, M. (2014) Mixture Optimization by the CARSO Procedure and DCM Strategies. *Applied Mathematics*, **5**, 3026-3039. http://dx.doi.org/10.4236/am.2014.519290

[2] Clementi, S., Fernandi, M., Baroni, M., Decastri, D., Randazzo, G.M. and Scialpi, F. (2012) MAURO: A Novel Strategy for Optimizing Mixture Properties. *Applied Mathematics*, **3**, 1260-1264. http://dx.doi.org/10.4236/am.2012.330182

[3] Fernandi, M., Bazzurri, M., Decastri D. and Clementi S. (2015) Experimental Design for Optimizing a Mixture of Materials plus an Evaporating Solvent. *Applied Mathematics*, **6**, 1740-1746. http://dx.doi.org/10.4236/am.2015.610154

[4] Clementi, S., Cruciani, G., Curti, G. and Skagerberg, B. (1989) PLS Response Surface Optimization: The CARSO Procedure. *Journal of Chemometrics*, **3**, 499-509. http://dx.doi.org/10.1002/cem.1180030307

[5] SIMCA 4.01. www.umetrics.com

[6] Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S. (1993) Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quantitative Structure-Activity Relationships*, **12**, 9-20. http://dx.doi.org/10.1002/qsar.19930120103

[7] Rosipal, R. (2011) Non Linear Least Squares: An Overview. In: Lodhi, H. and Yamanishi, Y., Eds., *Chemoinformatics and Advanced Machine Learning Perspectives*: *Complex Computational Methods and Collaborative Techniques*, ACCM, IGI Global, 169-189.

[8] Wold, S., Ketteneh-Wold, N. and Skagerberg, B. (1989) Nonlinear PLS Modeling. *Chemometrics and Intelligent Laboratory Systems*, **7**, 53-65. http://dx.doi.org/10.1016/0169-7439(89)80111-X