Trends in Computational and Applied Mathematics

Trends in Computational and Applied Mathematics, **23**, N. 4 (2022), 749-767
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029
www.scielo.br/tcam
doi: 10.5540/tcam.2022.023.04.00749

A Novel In Silico Monte Carlo Approach to Optimize a PSD Estimation Problem. Generation of Data Fusion Experiment Rules

F. A. OTERO[1]* and G. FRONTINI[2]

Received on August 30, 2021 / Accepted on June 30, 2022

ABSTRACT. This article analyzes the performance of combining information from Scanning Electron Microscopy (SEM) micrographs with Static Light Scattering (SLS) measurements for retrieving the so-called Particle Size Distribution (PSD) in terms of experimental features. The corresponding data fusion is implemented using a novel Monte Carlo-based method consisting in a SMF (Sampling-Mapping-Filtering) approach. This approach provides an important reference to assess the strategy of the experiment for this specific problem by means of solving an inverse problem. Furthermore, low levels of volume fraction and a PSD represented by log-normal distributions are considered in order to reduce processing and model errors due to ill-posedness. The prior statistics corresponding to the SEM micrographs have been achieved by means of the Jackknife procedure used as a resampling technique. The likelihood term considers iid normal measurements generated from the Local Monodisperse Approximation (LMA) and also makes use of the same model as forward linear model, in an inversion case known as inverse crime. However, it has been proved that the LMA performs well in practice for low fraction volume systems as considered here. The PSD retrieval is measured in terms of improvement in precision with respect to one of the log-normal parameters in SEM micrographs, i.e., the desirability. Estimates are expressed as a function of a typical system parameter such as polydispersity, as well as experimental variables, i.e., number of particles per micrograph (PPM) and noise level ε in the SLS measurements. These estimations are then analyzed by means of the Box-Behnken (BB) design and the response surface methodology (RSM) in order to generate a surrogate model from which rules for the optimization of the experiment are made when desirability is maximized. Finally, a Rule-Based System (RBS) is proposed for future use.

Keywords: Inverse problem, particle size distribution, DOE, data fusion, Monte Carlo methods.

*Corresponding author: Fernando Otero – E-mail: foterovega@fi.mdp.edu.ar

[1] Department of Mathematics, College of Engineering, Universidad Nacional de Mar del Plata, Mar del Plata 7600, Argentina

Institute of Materials Science and Technology (INTEMA), Universidad Nacional de Mar del Plata and National Research Council (CONICET), Mar del Plata 7600, Argentina – E-mail: foterovega@fi.mdp.edu.ar https://orcid.org/0000-0002-6458-9665

[2] Department of Mathematics, College of Engineering, Universidad Nacional de Mar del Plata, Mar del Plata 7600, Argentina

Institute of Materials Science and Technology (INTEMA), Universidad Nacional de Mar del Plata and National Research Council (CONICET), Mar del Plata 7600, Argentina – E-mail: gfrontin@fi.mdp.edu.ar

1 INTRODUCTION

Inverse problems (IP) arise in a large number of processes in a diverse number of science and engineering fields. Among these IP, indirect estimation of a desired quantity is an often, almost unavoidable, aim at their solution. In particular, estimation of the Particle Size Distribution (PSD) is a common goal for several applications including areas such as biology, meteorology and nanomedicine. This estimation goal is, however, a very difficult one to perform due to limitations in experimental techniques. In fact, no single technique provides a complete description of the PSD [31]. As a consequence, a considerable number of signal processing methodologies has been developed, where the ones which perform the so-called multi-source information fusion (MSIF), i.e., methodologies which combine processing from two or more different measurement techniques stand out as the most promising. In this sense, there are several methodologies within the framework of MSIF techniques, with the only requisite of expressing information in a statistical manner, through a probability density function (pdf). For example, several articles have successfully applied Bayesian approaches to the PSD retrieval such as [7, 26]. Nevertheless, an effective employment of typical Bayesian methodologies such as Monte Carlo Markov Chain (MCMC) and Sequential Monte Carlo (SMC) methods to realistic cases often implies a large computational cost, mostly because in such Monte Carlo algorithms is the need to numerically evaluate the posterior distribution, up to a normalisation constant, commonly many thousands or millions of times [33]. This work, on the other hand, is focused on building an appropriate prior so a direct mapping from this prior on the solution field through an integral transform can be found using a Monte Carlo (MC) methodology. This is performed in a very similar fashion to previously mentioned Bayesian procedures but not attached to the Bayes theorem itself, as it will be seen later.

On the other side, design of experiments (DOE) is another cross-sectional field with many applications in science and engineering including chemometrics, one of the specific areas where this article can be addressed in. Nowadays, with the increasing power of computers, conventional experimental methods are sometimes replaced with computer code that serve as a proxy for the physical processes, especially when the physical processes are hard to study. In this sense, one can vary the inputs to the code and watch how the output is affected, as it happens with the physical experiment, Such experiments are called computer experiments [32]. Once again, MC methods have been successfully applied to computer experiments such as in [18]. Developments using computer experiments are sometimes called ‘in silico’ as appearing in the title of this article.

This article has also coupled DOE and knowledge engineering (KE) through rule-based systems (RBS). RBS are a tool omnipresent in science, technology and everyday life, although their encoding, analysis and design are seldom a matter of deeper theoretical investigation. The RBS are sets of rules imitating logical implication. Even after years of investigation of various other formalisms, rules proved to be generic core and very universal knowledge representation tool for the widest posible spectrum of applications [16]. In this paper, authors propose a RBS based on fuzzy logic for future work.

Finally, a fourth keystone applied in this article is data analysis. In this sense, as suggested by [19] authors have applied Exploratory Data Analysis (EDA) and confirmatory data analysis (CDA) in a complementary way. The process constructed by authors has followed Tukey [35] in his steps on both CDA: i) State the question(s) to be investigated ii) Design an experiment to address the questions iii) Collect data according to the designed experiment iv) Perform a statistical analysis of the data v) Produce an answer; and also on EDA: i) Start with some idea.and ii) Iterate between asking a question and creating a design. In fact, both EDA and CDA steps have been successfully applied here along the development process.

This work considers computer experiments starting from an initial PSD resulting from a Scanning Electron Microscopy (SEM) micrograph which is resampled by means of the Jacknife procedure. This statistical information is then combined with data obtained from Static Light Scattering (SLS) measurements. Each computer simulation is solving an IP using the new proposed methodology. Finally, the performed computer simulations, which are designed under a Box-Behnken (BB) scheme, are employed to build a surrogate model (SM) by means of the classical response Surface methodology (RSM). From the numerical optimization of this SM a set of rules is drawn and a final RBS is generated in order to properly build the experimental set up for future measurements.

2 CONTRIBUTIONS

This article has three main contributions: first, the proposed inversion methodology and the analysis of its application to the specific problem of estimating the PSD; second, a study on the influence of the SLS noise, the number of particles per micrograph (PPM) and the polydispersity of the particle system on estimations achieved with this methodology; and finally, the generation of rules for combining these experimental variables in order to maximize the so-called desirability using a surrogate model obtained by means of the BB design, the RSM optimization and the fuzzy logic.

The rest of the work is organized as follows. In section 3, a brief preliminary for both employed SLS and SEM models is given. A second section of background concepts related to the methodology appears in section 4. Section 5 includes the complete description of the problem formulation. Section 6 describes the proposed methodology employed for combining information, as well as the other techniques employed in the article. Section 7 presents the computational implementation and the selected examples. In section 8, results are presented and discussed. Finally, conclusions are shown in section 9 and the URL for downloading the corresponding codes is provided in the Appendix.

3 PHYSICAL PROBLEM: SEM AND SLS MODELLING

The SEM micrographs are the result of a direct microscopy measurement technique and allow capturing details from the structure on the surface of the particles. However, they are experimentally expensive and the electron beam may distort the results [25]. Electron microscopy methods have been widely simulated using the MC method. It can be mentioned the pioneer job of Joy [13] and more recently works related to the field of biophotonics such as [11]. For reasons of simplicity, the complete process of SEM microscopic imaging has not been simulated here. This work uses a simplified random sampling MC routine as described in [26] and is applied to a sample of a low number of particles, that is far below the several hundreds needed in order to make a reliable statistics per se [25].

The SLS, on the other side, is an indirect measurement technique for the PSD estimation. It briefly consists on illuminating the sample with laser light and measures the scattered light intensity average over time at different angles and relating to the PSD according to some model. In this case, the used model is the Local Monodisperse Approximation (LMA) developed by Pedersen in [27]. According to the LMA model SLS intensities, denoted as $I_s(q)$, can be computed by solving a first kind Fredholm integral equation as in Eq. (3.1)

$$I_s(q) = K \int_0^\infty f(R) S(p,q,R) P(q,R) \ \mathrm{d}R \tag{3.1}$$

where q represents the magnitude of the scattering vector, $f(R)$ is the PSD with R as the radius as integration variable; $S(p,q,R)$ is the so-called structure factor where p as an effective model parameter; $P(q,R)$ is the shape factor and K is a global constant that 'absorb' all the proportionality terms.

The PSD is parameterized using a log-normal distribution as a function of parameters and g as in Eq. (3.2). Relations between these parameters and the mean and variance of the PSD can be seen in [25]. It is worthy to point out that this article has followed the parameterization of the PSD applied in [25], that is, representing the PSD by the two parameters R_0 and $g^* = \ln(g)$

$$f(R) = \frac{\sqrt{\frac{g}{\pi}}}{R} \exp\left\{ -g \left[\log\left(\frac{R}{R_0} \right) \right]^2 \right\} \tag{3.2}$$

The LMA is a reduced model but it can rigorously represent the complete model when particles are grouped according to their size, and the system is highly diluted so the structure factor equals to 1. Narrow and low-concentrated particle systems are best suited for the use of the LMA, however practice demonstrated that the estimation of the PSD remains quite unaffected as the other parameters 'absorb' the model approximations, at least for moderate particle concentrations and breadth of the PSD [24]. Authors also recommend [8,24] as extended references concerning the LMA model and its use in data analysis.

The corresponding compound experiment to be designed is the combination of SEM and SLS experiments, each of them performed on their own and when the proposed methodology is applied. It seems important to remark the dependence of the compound experiment on the methodology because the proper methodology is an active part of the experiment as in [28].

4 THEORETICAL BACKGROUND FOR THE METHODOLOGY

4.1 Inverse problem methodology

The methodology for combining information is following a 'SMF' (Sampling-Mapping-Filtering) strategy. It starts with the achieved micrograph of a given number n particles as a first sampling object to be processed by means of a resampling technique as the Jackknife procedure, which is briefly described next.

4.1.1 The Jackknife procedure

The Jackknife procedure, formerly formulated by Quenouille [30], applied to this specific case can be considered the preprocessing stage and it basically consists in generating n resamples of $n-1$ particles each one, obtained from the micrograph, where every resample is achieved by deleting a different particle at each time. From these resamples, approximate normal distributions are drawn for R_0 and g^* with means $\mu(R_0)$ and $\mu(g^*)$ (the original sample averages using the total number of PPM) and standard deviations $\sigma(R_0) = \sqrt{PPM}\sigma_{R_0}$ and $\sigma(g^*) = \sqrt{PPM}\sigma_{g^*}$ where $\sigma(R_0)^2$ and $\sigma(g^*)^2$ are the variances of each parameter obtained by resampling the micrograph with Jackknife [25].

After this procedure and corresponding drawing of normal distributions, the actual three stages in the SMF methodology are performed. These are shortly presented from sections 4.1.2 to 4.1.4.

4.1.2 Monte Carlo sampling

Monte Carlo (MC) sampling methods refers to a class of methods for randomly sampling from a probability distribution. Its motivation relies in that for most probabilistic models of practical interest, exact inference is intractable, and so we have to resort to some form of approximation [5]. In the particular case of this work normal distributions are managed through MC sampling instead of using an analytical approach.

4.1.3 Integral Transforms

An integral transform like Eq. (3.1) is a linear operation that converts the PSD to an intensity function in the scattering vector domain q. Integral transforms are used to map one domain into another in which the problem is simpler to analyze in the measurements space.

4.1.4 Discrepance principle and noise estimation

The discrepancy principle, formulated by Morozov [22], can be applied when available information about the noise level ε or a corresponding bound. The basic idea in this method, (originally employed to select a regularization parameter) is to choose the value of the model parameters to estimate in such manner that the norm of residues is equal to the noise level or a bound. In this article, since we are working with simulations, the SLS noise level is known, however in practice this situation can hardly happen. As an alternative, in order to estimate this noise, it is posible to apply the singular spectrum analysis, also called averaging Hankel or Cadzow's basic algorithm [9]. Summarizing the Cadzow iteration algorithm, It takes the data and forms a Hankel matrix, say H, then performs a singular value decomposition $H = USV^T$ and retains only the significant singular values. It then reconstructs the matrix with $H_a = U_1S_1V_1{}^T$. Then it averages over the main anti-diagonals to reconstruct a Hankel matrix. It does another SVD and repeats the procedure over and over again until the Hankel matrix has an exact rank. The level of noise can be estimated (i.e., noise floor). The selection of ε is also discussed in [4].

4.2 DOE methodology

The second part of this work is devoted to the generation of rules for the DOE, specifically in terms of the PPM, the noise level ε and the PSD breadth σ (degree of polydispersity) which are the important factors in the experimental set up. In this sense, there are few works combining DOE and RBS through fuzzy logic, such as the recent articles [1, 14]. This second part starts with the selected examples solved with the SMF methodology following certain design, i.e., the Box-Behnken design, that will be seen next.

4.2.1 Box-Behnken Design

Box–Behnken (BB) designs are experimental designs for Response Surface methodology (RSM) (next to be seen), devised by George E. P. Box and Donald Behnken in 1960 [6]. Each factor is placed at one of three equally spaced values, usually coded as -1, 0, +1. In fact, at least three levels are needed. The design should be sufficient to fit a quadratic model, that is, one containing squared terms, products of two factors, linear terms and an intercept. BB design is still considered to be more proficient and most powerful than other designs such as the three-level full factorial design, central composite design (CCD) and Doehlert design, despite its poor coverage of the corner of nonlinear design space. Authors recommend [20] for further information.

4.2.2 Response Surface Methodology

RSM hinges on a rather simple idea - that of obtaining an approximate form of the objective function by simulating the system at a finite number of points, which are carefully sampled from the function space [10]. RSM has been employed here to optimize and analyze the effects of several independent factors on a treatment process to obtain the maximum output, which is called here *desirability D*. In particular, polynomial response Surface methodology (PRSM) uses regression analysis and analysis of variance to determine the relationship between design variables and responses. In PRSM linear polynomial is used to approximate the implicit limit state equation. The coefficients of the linear polynomial are determined through the BB design described above [12]. The general form of a PRS model relation is shown in Eq. 4.1

$$D(x) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m} \sum_{j \geq i}^{m} \beta_{ij} x_i x_j + \ldots + E \tag{4.1}$$

where E is the statistical error, x_i is the i-th component of the m-dimensional predictor and β_0, β_i and β_{ij} are parameters to be estimated and can be arranged in a certain order to form a column vector $\vec{\beta}$. In this case, m=3 and x_1, x_2,x_3, are the considered factors, correspondingly, the PSD standard deviation σ, the number of PPM, and the noise level ε.

4.2.3 Rule-based systems

Rule-Based Systems (RBS) or to be more precisely a Rule-Based Fuzzy System (RBFS) as the one employed here- have flexible enough structure to represent adequately non-linearity and uncertainty of real processes and they are transparent enough to be easy for inspection, analysis, incorporation of existing knowledge and suppression of undesired one. They have been developed during the last four decades as a result of an interaction of the Fuzzy Set Theory and the Control Theory [3]. RBFS contains four components—rules, fuzzifier, inference, and output processor—that are interconnected as shown in Fig. 1. Once the rules have been established, the fuzzy system can be viewed as a mapping from the factors to the output desirability D and this mapping can be expressed quantitatively as $D = f(x_1, x_2, x_3)$. Rules are quantified using the mathematics of fuzzy sets, and that mathematics is different for type-1, interval type-1, and general type-2 fuzzy sets. [21]. To see further about this topic applied to this work readers should go to the second part of the methodology in section 6, where a RBFS is proposed.

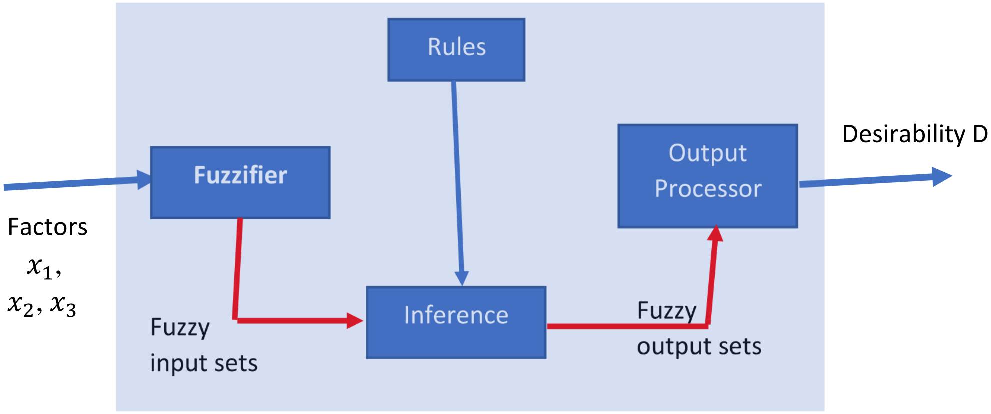

Figure 1: Scheme of a RBFS with input factors x_1, x_2, x_3 and output D.

5 FORMULATION OF THE PROBLEM

There are two main problems to be considered in this work:

The first one, defined as a part of an IP, can be formulated as the selection of the parameters R_0 and g^* such as

$$\|I_\varepsilon(q) - I_s(\check{R}_0, \check{g}^*)\| < \varepsilon \tag{5.1}$$

where $I_\varepsilon(q)$ are the noisy SLS measurements, I_s are simulated SLS measurements generated by Eq. (3.1) for the parameters $\check{g}^*$ and $\check{R}_0$ which are the sampled values from the approximate normal distributions developed after performing the Jackknife procedure; meanwhile ε corresponds to the SLS level noise.

The second one, defined as a part of the RBS generation problem, can be formulated as the optmization problem of the so-called desirability function D as in:

$$\max_{A,B,C}\{D(comb(A,B,C))\}$$

where the desirability function D is specifically the relative improvement in the precision for parameter R_0 from Eq. (3.2) in terms of several combinations of factors appearing in Eq. (4.1), and $x_1 = A$, $x_2 = B$ and $x_3 = C$, where A is the PSD standard deviation σ, B is the number of PPM, and C is the noise level ε.

6 METHODOLOGY

The selection problem formulated in the last section as in Eq. (5.1) is solved using the proposed SMF methodology, however this problem is just the final step (filtering) involved in it. The complete methodology for combining SEM and SLS data (including the preprocessing stage described in 4.1., i.e., the Jackknife procedure and the successive approximate normal distributions) is described in the flowchart in Fig. 2

It can be asked a reasonable question after studying this inversion procedure and the LMA: What happened with the rest of the parameters in the model besides those related to the PSD? How should we take them into account? The answer is not unique and authors give two possibilities: 1) these nuisance parameters, i.e., p and K in Eq. (3.1) can be estimated using an standard inversion method, such as the Levenberg-Marquardt algorithm [15, 17] and then to be used this estimate point or 2) their corresponding confidence intervals (CIs) can be drawn after performing several MC simulations. In this work authors have considered the first option.

When analyzing the second part of this work, once again, the numerical optimization in section 5 is just a step in the RBS generation problem, which has three phases: in the first one, the BB design is drawn. In the second one, the RSM is applied. In the last one, the final RBS is built from the results of solving such optimization. The first two phases are standard and can be seen in more detail in good and amenable books such as [2, 20]. However, it seems important to remark that for RSM models with orders greater than linear or with interactions, univariate optimization will not reach to the optima. In this sense, the algorithm used in the RSM numerical optimization

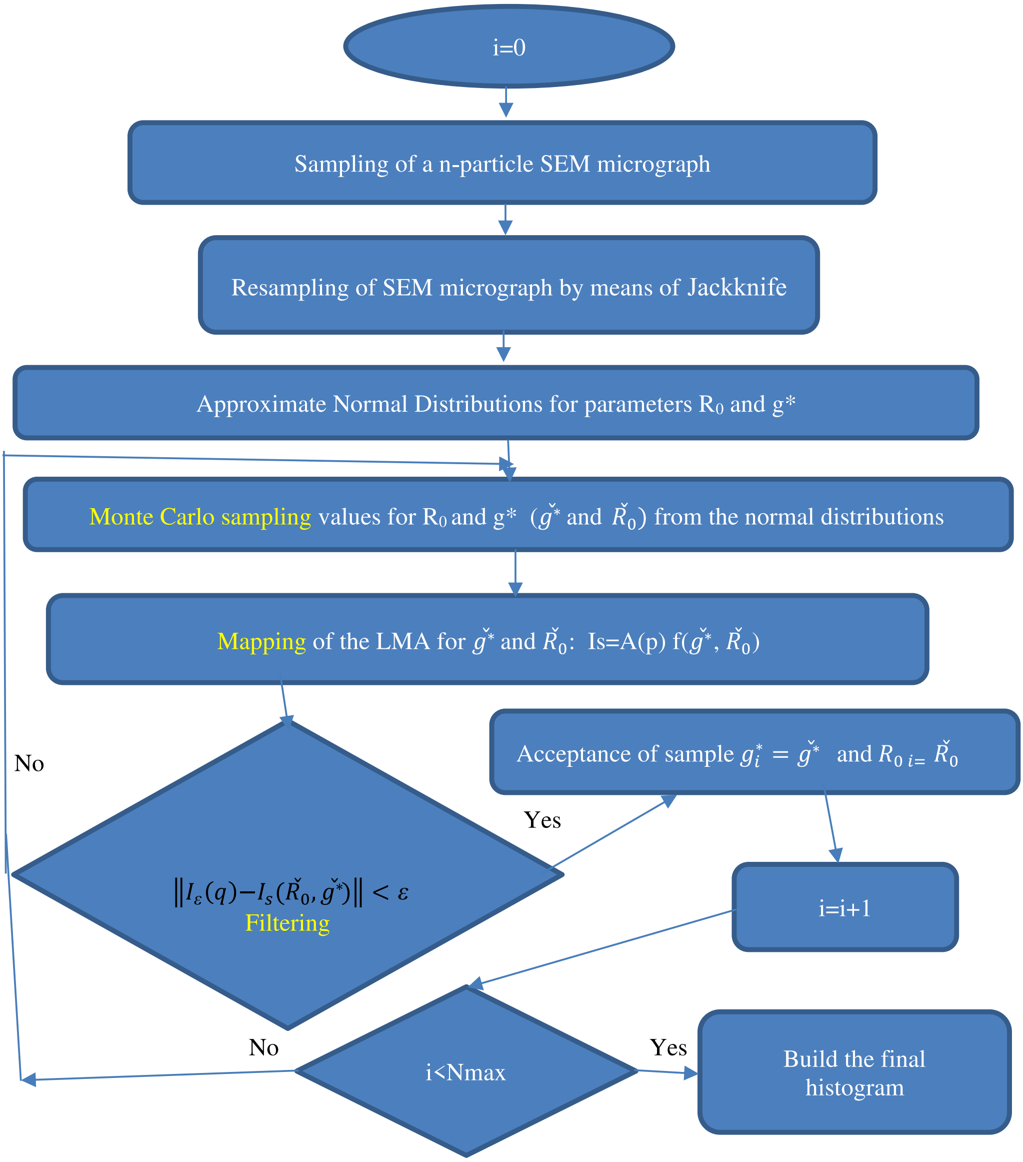
i=0
Sampling of a n-particle SEM micrograph
Resampling of SEM micrograph by means of Jackknife
Approximate Normal Distributions for parameters R_0 and g*
Monte Carlo sampling values for R_0 and g* ($\check{g}^*$ and $\check{R}_0$) from the normal distributions
Mapping of the LMA for $\check{g}^*$ and $\check{R}_0$: Is=A(p) f($\check{g}^*$, $\check{R}_0$)
No
$\|I_\varepsilon(q) - I_s(\check{R}_0, \check{g}^*)\| < \varepsilon$
Filtering
Yes
Acceptance of sample $g_i^* = \check{g}^*$ and $R_{0\,i} = \check{R}_0$
i=i+1
No
i<Nmax
Yes
Build the final histogram

Figure 2: Flowchart of the proposed SMF methodology for the resolution of the IP.

is the downhill simplex (Nelder-Mead) multi-dimensional pattern search whose code is extracted from [29] and it has a good performance in multivariate optimization, however convergence to global optima is not guaranteed.

Now with all of the above, authors may focus on the proposed RBS generation. First, as a first approach, rules are drawn according to the following setting given in the corresponding software (section 7) (noted as SETTING1):

i) One factor is set with goal: “equal to ->” a value in the range $[-1,1]$
ii) The other two factors are set with goal: “in range” with lower limit -1 and upper limit 1
iii) The desirability is set with goal: “To maximize”

In this case, factors in i) and ii) are varied between values $\{-1,0,1\}$, i.e., the three levels, resulting in 9 possible configurations. Some important conclusions can be derived from these results and probably more detalied configurations should be set as SETTING2:

i) Two factors are set with goal: “equal to ->” a value in the range $[-1,1]$
ii) The other factor is set with goal: “in range” with lower limit -1 and upper limit 1
iii) The desirability is set with goal: “To maximize”

In this case, once again factors are varied between values $\{-1,0,1\}$ resulting in 27 possible configurations. These results should probably be enough to build an acceptable set of rules.

Finally, if it is still found to be an uncomplete set of rules, a third option is to consider all the 81 possibilities varying the values for factors between $-1,0,1$ given in SETTING3:

i) The three factors are set with goal: “equal to ->” a value in the range $[-1,1]$
ii) The desirability is set with goal: “To maximize”

It is also possible (but not considered in this work) to use intermediate not integer values between -1 and 1. In this case, it should be used once again SETTING1, SETTING2 or SETTING3.

Second, it is worth to mention that results from SETTING1, SETTING2 and SETTING3 may probably give intermediate values. These values should be fuzzified according to the set of memberships, which are given next.

Third, authors have selected to work with type-1 fuzzy sets. Also, a set of membership functions (mf’s) and corresponding attributes should be proposed for the three (now normalized) factors and desirability D. This is needed for the fuzzification, the inference and the output processor blocks in Fig. 1. Typical mf’s include triangular, trapezoidal, sigmoidal and Gaussian. In this case, a generic trapezoidal mf has been chosen for a generic factor between -1 and 1 with a profile such as the one in fig. 3, with three attributes, “low”, “medium” and “high” and variations of these with the use of logical modifiers. For the case of the desirability D, authors have fed back the RBFS with the RSM results and the induced rules. As a consequence the attributes with the corresponding mf’s appear in fig. 4.

Fourth and finally, there is need of finishing the setting for the inference and the output processor blocks by defining the type of logic. As in [23] authors have proposed compensatory fuzzy logic for defining the corresponding connectors and modifiers based on its performance as a sensitive and idempotent inference system. See also [23] for the structure of these connectors and modifiers. In this work, authors use the modifiers 'hyper' and 'quite'. Authors have also used "hardly high" with a specific-defined mf.

All of these are very important issues to be solved, since rules are defined on the base of the different mf's and the type of sets and logic employed in the inference, fuzzification and output processor. In fact, it is a common procedure to build mf's in function of type of used logic and the set of rules.

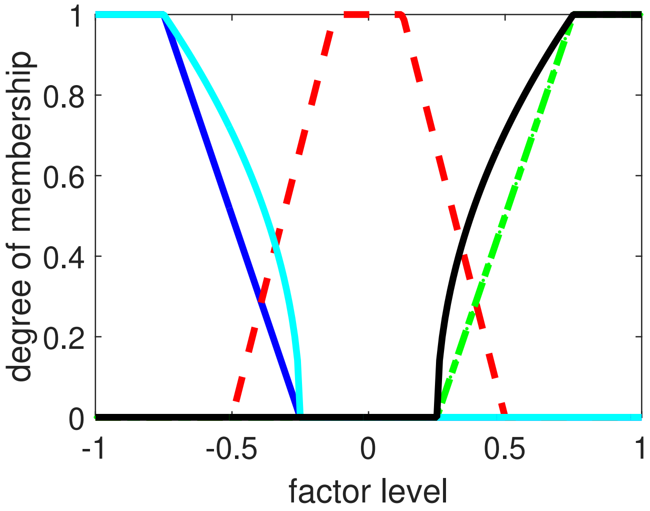

Figure 3: Corresponding mf's to attributes "low" (straight blue line), "quite low" (straight light blue line), "medium" (red dashed line), "quite high" (straight black line) and "high" (green dashed-dotted line) for a generic normalized factor.

7 COMPUTATIONAL IMPLEMENTATION AND EXAMPLES

The first part of the work, i.e., the computer simulations performing the SMF methodology have been implemented using the MATLAB© 2015b package. Results from this part are considered the input of the analysis phase from the second part of the work, i.e. the design of the experiment. This phase, as well as the analysis, the RSM and the numerical optimization of the DOE, have been implemented by means of the Design-Expert© 7.0 software. Meanwhile, the final phase of this second part, i.e., the generation of rules, is of human-crafted nature.

The examples considered in this paper follows the previous article [25] and they belong to a typical range of polymeric systems such as the ones studied for a solid polymer matrix in [34]. Three values of PSD standard deviation of 0.02, 0.05 and 0.10 μm for a mean radius of 0.25 μm as well as three values for SLS noise levels of additive normal distribution of 0.1%, 1% and 10% from the measurement peak, and three values of PPM (50, 80 and 100 particles). All

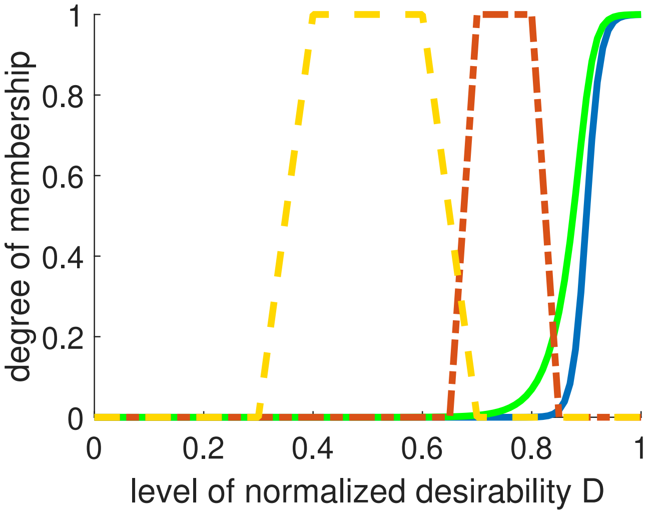

Figure 4: Corresponding mf's to attributes "medium" (dashed orange line), "hardly high" (brown dashed-dotted line), "high" (green straight line) and "hyper high" (straight blue line) for a normalized desirability D.

of these cases correspond to volume fractions of 1%. It has been performed a MC sampling of 100 samples for parameters R_0 and g^*. The BB design places three equally spaced values, coded as -1, 0 and $+1$ as stated in section, which are in correspondence to the three values for every factor in an increasing order. This design includes 15 runs with 3 center points per block. Since authors have included the study of the lack of fit (LOF), repetitions are needed. In this case, this implies three repetitions at the center $(0,0,0)$ as can be seen in Table 1.

8 RESULTS AND DISCUSSION

As stated in the introduction, this work is involved with both qualitative and quantitative data analysis, i.e. EDA and CDA. EDA was concerned about the question: how can we design the experiment including the inversion methodology? And more specifically, how can we fuse data from SLS and SEM? How can we measure improvement in such a data fusion? What type of model can we make for optimizing the design of the experiment? CDA was concerned about the process of responding these questions in practice following those steps dictated by Tukey.

It was observed from the results for the different IP's that improvements in the CI for parameter R_0 between 30% and 48% were found for the lesser number of PPM (50) correspondingly to a decreasing value of SLS noise level. However, just for a few cases in general, improvement in the CI for parameter g^* were found. It was also observed that improving the number of PPM to 80, for the narrower PSD and for the lower SLS noise level, decrease the initial CI for R_0 around a $50\ \%$. Once again, results on improvement for the CI's have shown a direct dependency on the SLS noise level, i.e., the lower the noise is, the greater the CI shrinking is. Nevertheless, the observed behavior seems to have a non-trivial dependency on the other two variables (factors A and B), which has to be evaluated in a further analysis in the second part. As a sort of example for the statistical improvements in the retrieved PSD's, results corresponding to the run 5 in table 1 are shown in Fig. 5

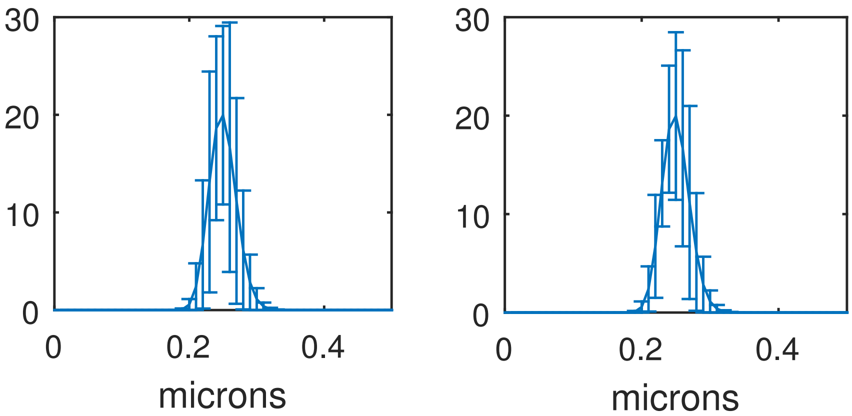

Figure 5: Corresponding error bars for the initial PSD achieved from Jackknife on the left and for the final PSD achieved with the SMF methodology on the right.

Table 1 shows the results for desirability D and their values normalized to the maximum after solving the different inverse problems for the 15 runs of the BB design. Authors have selected the relative improvement in the precision for the parameter R_0 as the desirability D due to the almost null improvement in the initial CI for the parameter g^* in a relevant number of cases.

The "Sequential Model Sum of Squares [Type I]" suggested the linear model is adequate when analyzing the selection of the highest order polynomial where the additional terms are significant and the model is not aliased. However, in a more exhaustive analysis using the Analysis of Variance (ANOVA), as seen in Table 2, authors have chosen a reduced 2FI (two factor interaction) model. In this case, it was included the interaction "AB" when its corresponding p-value is 0.083. Values greater than 0.10 indicate the model terms are not significant, but values lesser than 0.05 are significant. However, this range between 0.05 and 0.10 is a sort of fuzzy and after studying the behavior of this model and in particular for this interaction, authors decided to include it. The model F-value of 25.34 including this interaction (also seen in Table 2) implies the model is significant and the F-Value for the LOF of 2.95 implies the LOF is not significant relative to the pure error. There is a 27.74% chance that a "Lack of Fit F-value" this large could occur due to noise. It should be noticed that authors also included factor B in the model even when its p-value is so high. This inclusion is due to the need for a hierarchical model. Furthermore, the "Pred R-Squared" of 0.8169 is in reasonable agreement with the "Adj R-Squared" of 0.8743. The analysis also has shown an adequate precision of 15.411, which measures an appropriate signal to noise ratio (SNR)

Table 1: Results from the BB design for the DOE.

Run	Factor A	Factor B	Factor C	*D* (desirability)	*D* normalized
1	1	-1	0	29	0.5577
2	1	0	-1	44.2	0.85
3	-1	1	0	38	0.7308
4	0	0	0	33.8	0.65
5	0	-1	1	17.8	0.3423
6	1	1	0	18.7	0.3596
7	0	0	0	32	0.6154
8	1	0	1	0	0
9	0	-1	-1	37.6	0.7231
10	0	1	-1	42	0.8077
11	-1	-1	0	30	0.5769
12	0	0	0	28	0.5385
13	-1	0	1	15.6	0.3
14	0	1	1	17	0.3269
15	-1	0	-1	52	1

The final model is:

$$D = 0.56 - 0.11A + 0.003125B - 0.30C - 0.088AB$$

Results show the significance of the SLS noise level in the surrogate model from the RSM. As it has been seen in section 3, the range of PPM should be between 50 and several hundreds in order to be comparable to the corresponding SLS noise. That is the reason for the preponderancy of factor C. In order to make these two factors comparable between each other, several micrographs should be taken into account

Furthermore, the analysis of data for the interaction between factors A and B shows that a combination of the two has a dramatic effect on maximization of D. This can be seen through Table 3 where no matter the value of C, the best combination for factors A and B is correspondingly -1 and 1. It is important to stand out that quantitative analysis is as subjective as qualitative analysis: significant isn't the same as meaningful. The more interactions are being compared, the more likely it is that "significant" interactions that aren't really significant are going to be found. This is the reason for choosing the parsimony law as a guide in the model construction and to consider just this interaction, which is justified by the ANOVA results and the rest of model tests. Subjectivity is also present in the author's choice of mf's. In this sense, definition and parameter identification of membership functions constitute the trickiest issue in the practical use of fuzzy sets theory. There are a lot of discussions and no common fundamental recipe for doing this. The mf's for two users could be quite different depending upon their many factors, or, they can

Table 2: ANOVA for the RSM (extracted from the Design-Expert© 7.0).

Source	Sum of Squares	Mean Square	F-Value	P-Value Prob > F-Value	
Model	0.85	0.21	25.34	<0.0001	Significant
A-A	0.088	0.088	10.57	0.0087	
B-B	7.80E-05	7.80E-05	9.36E-03	0.9249	
C-C	0.73	0.73	87.07	<0.0001	
AB	0.031	0.031	3.71	0.083	
Residual	0.083	8.35E-03			
LOF	0.077	9.62E-03	2.95	0.2774	Not significant
Pure Error	6.51E-03	3.26E-03			

be designed using optimization procedures [3,21]. Here, authors have used expert knowledge to define mf's.

Table 3: Optimal values for factors A and B for different levels of factor C.

	$C=-1$	$C=-0.5$	$C=0$	$C=0.5$	$C=1$
A	-1	-1	-1	-1	-1
B	1	1	1	1	1
D	1*	0.905	0.755	0.604	0.453**

* One of the 30 solutions found.

** One of 2 solutions found.

In this article, authors have just considered results from SETTING1 as a way of simplifying the analysis. The construction of basic rules drawn using the SETTING1 (including suboptimal solutions) are the following:

1- IF *A* IS LOW OR MEDIUM, AND *B* IS LOW AND *C* IS LOW THEN *D* IS HIGH
2- IF *A* IS LOW OR QUITE LOW AND *B* IS MEDIUM OR HIGH OR QUITE HIGH AND *C* IS LOW THEN *D* IS HYPER HIGH
3- IF *A* IS LOW AND *B* IS HIGH AND *C* IS MEDIUM THEN *D* IS HARDLY HIGH
4- IF *A* IS LOW AND *B* IS HIGH AND *C* IS HIGH THEN *D* IS MEDIUM
5- IF *A* IS HIGH AND *B* IS LOW OR QUITE LOW AND *C* IS LOW THEN *D* IS HIGH
6- IF *A* IS MEDIUM AND *B* IS MEDIUM OR HIGH AND *C* IS LOW THEN *D* IS HIGH

9 CONCLUSIONS

This article has successfully applied a novel inversion methodology to the specific problem of estimating the PSD (represented parametrically by a log-normal distribution) by means of fusing data from SLS and SEM measurements. This methodology based on Monte Carlo simulations can be described as a three-step strategy: Sampling-Mapping-Filtering (SMF); specifically, the inversions are used to generate data fusion rules in terms of three factors: the SEM particles per micrograph, the SLS noise and the PSD breadth. Results of inversion have shown that shrinking of the CI's with respect to initial from SEM were only significant for parameter R_0. This is the reason for choosing the improvement on this parameter as a measure for the objective function, called desirability, in the design of the generation of the experiment rules. It is important to remark the dependence of the computer experiments on the proposed inversion methodology. Experiment rules, which are human crafted, have presented two main features in order to optimize the desirability (even when suboptimal solutions include other features): i) a low SLS noise level, and ii) a combination of low polydispersity and high number of SEM particles per micrograph. It has also been proposed a rule-based fuzzy system based on Type-I fuzzy sets and compensatory fuzzy logic in order to derive solutions for other possible configurations.

Acknowledgments

Authors would like to acknowledge the National Research Council (CONICET) and the Faculty of Engineering at Universidad Nacional de Mar del Plata (UNMdP) for their respective financial support. This work is part of the project 15/G610, named "Inverse problems, modeling and computer simulation for Non-destructive Evaluation essays", which is supported by UNMdP. Authors also would like to acknowledge Mr. André Pontis for his invaluable help with LaTeX editing.

REFERENCES

[1] F. Aginab, R. Khosravaniana, M. Karimifardc & A. Jahanshahid. Application of adaptive neuro-fuzzy inference system and data mining approach to predict lost circulation using DOE technique (case study: Maroon oilfield). *Petroleum*, **6** (2020), 423–437.

[2] M.J. Anderson & P.J. Whitcomb. "RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments". CRC Press, Boca Raton, Florida (2017).

[3] P.P. Angelov. "Evolving Rule-Based Models: A Tool for Design of Flexible Adaptive Systems". Springer-Verlag, Berlin, Heidelberg (2002), chapter 3: Flexible models, p. 25.

[4] R.C. Aster, B. Borchers & C.H. Thurber. "Parameter Estimation and Inverse Problems". Elsevier Academic Press, New York (2005), chapter 4: Rank deficiency and ill-conditioning, p. 67.

[5] C.M. Bishop. "Pattern Recognition and Machine Learning". Springer, New York (2006), chapter 11: Sampling methods, p. 523.

[6] G. Box & D. Behnken. Some new three level designs for the study of quantitative variables. *Technometrics*, **2** (1960), 455–475.

[7] L.A. Clementi, J.R. Vega, L.M. Gugliotta & H.R.B. Orlande. A Bayesian inversion method for estimating the particle size distribution of latexes from multiangle dynamic light scattering measurements. *Chemometr. Intell. Lab. Syst.*, **107** (2011), 165–173.

[8] G.E. Eliçabe & F.A. Otero. Static Light Scattering of Concentrated Particle Systems in the Rayleigh-Debye-Gans Regime: Modeling and Data Analysis. *Particulate Science and Technology*, **28** (2010), 485–497.

[9] J. Gillard. Cadzow's basic algorithm, alternating projections and singular spectrum analysis. *Statistics and Its Interface*, **3** (2010), 335–343.

[10] A. Gosavi. "Simulation-based Optimization: Parametric Optimization Techniques and Reinforcement Learning". Springer Science+ Business Media, New York (2003), chapter 4: Parametric Optimization: Response Surfaces and Neural Networks, p. 57.

[11] M. Gu, X. Gan & X. Deng. "Microscopic Imaging through Turbid Media. Monte Carlo Modeling and Applications". Springer-Verlag, Berlin, Heidelberg (2015).

[12] P. Jiang, Q. Zhou & X. Shao. "Surrogate Model-Based Engineering Design and Optimization". Springer Nature, Singapore (2020), chapter 2: Classic Types of Surrogate Models, p. 7–9.

[13] D.C. Joy. "Monte Carlo Modeling for Electron Microscopy and Microanalysis". Oxford University Press, New York, Oxford (1995).

[14] H. Khoshdast, A. Soflaeian & V. Shojaei. Coupled fuzzy logic and experimental design application for simulation of a coal classifier in an industrial environment. *Physicochem. Probl. Miner. Process.*, **55** (2019), 504–515.

[15] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, **2** (1944), 164–168.

[16] A. Ligêza. "Logical Foundations for Rule-Based Systems". Springer-Verlag, Berlin (2006), chapter Preface, p. 5–6.

[17] D.W. Marquardt. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, **11** (1963), 431–44.

[18] H. Martens, G.B. Dijksterhuis & D.V. Byrne. Power of experimental designs, estimated by Monte Carlo simulation. *J. Chemometrics*, **14** (2010), 441–462.

[19] W.L. Martinez & A.R. Martinez. "Exploratory Data Analysis with MATLAB". Chapman & Hall/CRC Press, Boca Raton, Florida (2005), chapter 1: Introduction to Exploratory Data Analysis, p. 3–4.

[20] R.L. Mason, R.F. Gunst & J.L. Hess. "Statistical Design and Analysis of Experiments with Applications to Engineering and Science". John Wiley & Sons, Inc., Hoboken, New Jersey (2003), chapter 17: Designs and Analyses for Fitting Response Surfaces, p. 568–635.

[21] Mendel. "Uncertain Rule-based Fuzzy Systems. Introduction and New Directions". Springer Nature, Cham, Switzerland (2017), chapter 3: Type-1 Fuzzy Systems, p. 101–160.

[22] V.A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Math Dokl*, **7** (1966), 414–417.

[23] F. Otero, G. Eliçabe & G. Frontini. A modified processing method for light scattering measurements in particulate systems by means of splines and fuzzy logic. In "16th Workshop on Information Processing and Control (RPIC 2015)", volume 1. Córdoba, Argentina (2016). doi:10.1109/RPIC.2015.7497104.

[24] F.A. Otero, G.L. Frontini & G.E. Eliçabe. Evaluation of light scattering models to characterize concentrated polymer particles embedded in a solid polymer matrix. *Journal of Polymer Science Part B: Polymer Physics*, **48** (2010), 958–963.

[25] F.A. Otero, G.L. Frontini & G.E. Eliçabe. PSD Retrieval by Bayesian Data Fusion via Metropolis-Hastings and the Jackknife Procedure. In "17th Workshop of Information and Control (RPIC 2017)". Mar del Plata, Argentina (2017). doi:10.23919/RPIC.2017.8214311.

[26] F.A. Otero, H.R.B. Orlande, G.L. Frontini & G.E. Eliçabe. Bayesian approach to the inverse problem in a light scattering application. *J. Appl. Stat.*, **42** (2015), 994–1016.

[27] J.S. Pedersen. Determination of size distribution from small-angle scattering data for systems with effective hardsphere interactions. *J. Appl. Cryst.*, **27** (1994), 595–608.

[28] R. Prabhu, W.R. Whittington, S.S. Patnaik, Y. Mao, M.T. Begonia, L.N. Williams, J. Liao & M.F. Horstemeyer. A Coupled Experiment-Finite Element Modeling Methodology for Assessing High Strain Rate Mechanical Response of Soft Biomaterials. *Journal of Visualized Experiments*, **99** (2015), 1–17.

[29] W.H. Press, S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. "Numerical Recipes in Pascal". Cambridge University Press, New York (1989), chapter 10: Minimization or maximization of functions, p. 326.

[30] M.H. Quenouille. Notes on bias in estimation. *Biometrika*, **43** (1956), 353–360.

[31] R.A. Reynolds, D. Stramski, V.M. Wright & S.B. Woźniak. Measurements and characterization of particle size distributions in coastal waters. *J. Geophys. Res.*, **115** (2010), 1–19.

[32] T.J. Santner, B.J. Williams & W.I. Notz. "The Design and Analysis of computer Experiments". Springer-Verlag, New York (2003), chapter Preface, p. 7.

[33] S.A. Sisson, Y. Fan & M.A. Beaumont. "Handbook of Approximate Bayesian Computation". Chapman & Hall/CRC, Boca Raton, Florida (2018), chapter 1: Overview of ABC, p. 4.

[34] E.R. Soulé & G.E. Eliçabe. Determination of Size Distributions of Concentrated Polymer Particles Embedded in a Solid Polymer Matrix. *Particle & Particle Systems Characterization*, **125** (2008), 84–91.

[35] J.W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, **34** (1980), 23–25.

APPENDIX

MATLAB© codes from the first part and Design-Expert© file from the second part are made available at `http://www3.fi.mdp.edu.ar/virtuallab/MAMI/CODE_I_II.rar`.