Functionalization of Microarray Devices: Process Optimization Using a Multiobjective PSO and Multiresponse MARS Modeling

Laura Villanova, Paolo Falcaro, Davide Carta, Irene Poli, Rob Hyndman, Kate Smith-Miles, *Senior Member, IEEE*

***Abstract*— An evolutionary approach for the optimization of microarray coatings produced via sol-gel chemistry is presented. The aim of the methodology is to face the challenging aspects of the problem: unknown objective function, high dimensional variable space, constraints on the independent variables, multiple responses, expensive or time-consuming experimental trials, expected complexity of the functional relationships between independent and response variables. The proposed approach iteratively selects a set of experiments by combining a multiobjective Particle Swarm Optimization (PSO) and a multiresponse Multivariate Adaptive Regression Splines (MARS) model. At each iteration of the algorithm the selected experiments are implemented and evaluated, and the system response is used as a feedback for the selection of the new trials. The performance of the approach is measured in terms of improvements with respect to the best coating obtained changing one variable at a time (the method typically used by scientists). Relevant enhancements have been detected, and the proposed evolutionary approach is shown to be a useful methodology for process optimization with great promise for industrial applications.**

I. Introduction

Process optimization plays an important role in many fields of application ranging from engineering and science to nonmanufacturing settings, such as marketing and technology commercialization. Any process can be considered a combination of resources (i.e. operations, machines, methods, and people) that transform some input (often a material) into an output [1]. The characteristics of the output are measured by one or more observable response variables and are influenced by the inputs (independent variables). The purpose of process optimization is to optimize the observed responses and identify the corresponding process conditions (values of the inputs). The current study is devoted to the optimization of a chemical process for the production of coatings for DNA microarray applications. DNA microarrays are high capacity systems [2] suitable for gene expression profiling of tens of thousands of genes in a single experiment [3]. This technology requires a functionalization of the substrate (glass slide or membrane), to efficiently bind biological products. Substrate functionalization is typically achieved via sol-gel or plasma deposition [4], [5] of a thin coating of poly-lysine, amino silanes or other reactive silanes [6]. In the present work, sol-gel amino-functionalized glass slides are investigated. The features of the obtained coatings have been measured by four quality characteristics (response variables). Six chemical components (independent variables or factors) have been used to produce the coatings, and six different levels for the number of moles (factor levels) have been defined for each component. To evaluate the contribution of each silica chemical precursor, two constraints have been set on the number of moles for three out of six chemicals. The high dimensionality of the problem (six independent variables or factors each with six levels, and four response variables), the presence of constraints, the expected complexity of the search space, time-consuming experiments and the unknown objective function are challenging aspects also for well-established statistical approaches such as the design of experiments (DOE) and the response surface methodology (RSM) [1]. A pioneering approach based on evolutionary algorithms [7] and statistical modeling has been adopted to select experiments, optimize multiple responses and obtain some insight into the process. The remainder of this paper is organized as follows. Section II summarizes some of the existing methodologies useful for process optimization purposes. Section III introduces the proposed evolutionary approach. Section IV illustrates the case study providing a detailed description of microarray technology, sol-gel process and developed algorithm. The experimental results are presented in section V and conclusions are drawn in Section VI.

Laura Villanova is with the Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy, and with the School of Mathematical Sciences, Faculty of Science, Monash University, Building 28, Wellington Road, Clayton, Australia.

Paolo Falcaro is with CSIRO, Materials Science and Engineering, Gate 5 Normanby Rd, Clayton VIC 3168, Australia.

Paolo Falcaro and Davide Carta were with Associazione CIVEN - Nanofabrication Facility, Via delle Industrie 9, 30175 Marghera, Venezia, Italy.

Irene Poli is with the Department of Statistics, University Ca' Foscari of Venice, San Giobbe, Cannareggio 873, 30121 Venezia, Italy, and with European Center for Living Technology, Ca' Minich, San Marco 2940, 30124 Venezia, Italy.

Rob Hyndman is with the Department of Econometrics and Business Statistics, Faculty of Business and Economics, Monash University, Building 11, Wellington Road, Clayton, Australia.

Kate Smith-Miles is with the School of Mathematical Sciences, Faculty of Science, Monash University, Building 28, Wellington Road, Clayton, Australia (phone: +61 3 9905 3170; fax: +61 3 9905 5020; e-mail: Kate.Smith-Miles@sci.monash.edu.au).

II. Existing Methodologies for Process Optimization

A. One factor at a time approach

The one-factor-at-a-time approach (OFAT) consists of varying only one factor at a time while keeping others fixed

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

[8].
Many engineers and scientists perform OFAT experiments because they rapidly see and react to their data, in the hope to learn something from each run [8], [9]. However, this practice is justified only when the effects are expected to be of magnitude 4σ or more, that is at least 4 times the random error per trial [9]. OFAT experiments are typically outperformed by statistically designed experiments that vary several factors simultaneously. Indeed, OFAT experiments require more resources for the amount of information obtained, produce less precise estimates of the effect of each factor, do not permit the estimate of interaction between factors, and generate experimental information in a smaller region of the search space [8].

B. RSM and Computer-Generated Designs

The RSM was introduced by Box and Wilson in 1951 [10]. It is a sequential procedure that aims at exploring the typically unknown form of the relationship between the response and the independent variables [1]. As stated by Montgomery [1], almost all RSM problems use low-order polynomial models (i.e. first- and second-order models) to suitably approximate the true functional relationship in a relatively small region of the independent variables. Because the interest is on a restricted area of the search space, polynomial models usually provide a reasonable approximation of the true functional relationship in that subarea [1]. The objective of RSM is to lead the experimenter rapidly and efficiently along a path of improvement toward the general vicinity of the optimum [1]. Despite their flexibility, standard response surface designs (such as the central composite design, the Box-Behnken design, the face-centered cube) may not be the best choice if non standard situations occur. Two non standard situations of interest for the microarray device optimization problem are: irregular experimental region and unusual sample size requirements [1]. An irregular experimental region is generated because of the presence of constraints on the independent variables, such as the constraints on the number of moles for some of the chemical components used to produce the coating. The unusual sample size requirements derive from the necessity of studying six independent variables because all the six chemical components are required to produce the coating. For the same reason, the objective of the experiment is not on learning which variables are important (factor screening), but rather on learning how the response variables change in correspondence to different values (levels) of the independent variables and their combinations as well. The number of levels for each input variable has been chosen to be equal to six in order to obtain a sufficiently detailed knowledge of the responses, while considering some practical limitations (such as experimental costs and precision of the laboratory instruments). In such a context, standard response surface designs might present severe drawbacks mainly due to the high number of experimental runs required and to alias structures that might arise (see [1] for a detailed description of statistical designs).
In the presence of non standard situations, computer-generated designs constitute a valuable alternative. Computer-generated designs are based on the concept of optimal design [11], [12], that is experimental designs that are optimal with respect to some criterion (i.e. D-optimality, A-optimality, G-optimality and V-optimality criterion) [12]. The usual approach is to specify a model, determine the region of interest, select the number of runs to make, specify the optimality criterion, and then chose the design points from a set of candidate points [12].

C. Two-stage approach

In 1998, Smith-Miles and Gupta introduced a two-stage neural network approach for optimization based on sample points [13]. In the first stage, the optimization function is approximated using a multilayer feedforward neural network (MFNN) based on a sample of evaluated data points. In the second stage, the approximated function is optimized using a feedback neural network to perform gradient descent [14]. The main advantage of this methodology consists in its applicability to a broad variety of continuous optimization problems since it makes no assumption about the nature of the objective function. Indeed objective functions that are non-quadratic, non-smooth, discontinuous, non-differentiable, or even difficult to represent mathematically, are approximated with a MFNN which provides a smooth, continuous and differentiable representation of the objective function. Furthermore, different approximation and optimization methods can be used in place of neural networks [13].
Smith-Miles and Gupta showed that the two-step approach identifies a local minima in the vicinity of the true local minima, when the objective function has been well approximated by the MFFN in stage one [14]. The necessity to ensure a good approximation of the true function over the entire space and the high number of evaluated data points required in stage one (from 1000 to 5000 points for the De Jong's Test Suite), constitute limitations for the microarray coating optimization purpose.

D. Design and Analysis of Computer Experiments

Design and Analysis of Computer Experiments (DACE) [15] is a widely studied approach in the statistics literature. It is a sequential procedure that combines Gaussian Process (GP) models (also known as kriging) and strategies for the selection of additional search points [16]. Similarly to the two-stage approach previously described, an approximation model is first obtained. Subsequently, the model is either optimised or used to select the next point(s) to be evaluated. Several strategies have been developed to select the new search points including minimizing a statistical lower bound, maximizing the probability of improvement and maximizing the expected improvement [17], [18]. All these methods pursue a common strategy. They rely on the standard error of the kriging predictor to force the algorithm to go back and explore regions where the sampled points are sparse [17]. So they rely on the approximation function obtained during the first stage. However, if the initial sample is

sparse a highly misleading view of the true function can be obtained. Because the misleading view is taken as 'correct' in stage two, the algorithm may stop prematurely or become excessively local in its selection of search points [17]. One-stage approaches overcome this difficulty and can handle extremely deceptive problems [17].

III. THE EVOLUTIONARY APPROACH FOR PROCESS OPTIMIZATION

The evolutionary approach is an iterative procedure that employs computational intelligence methods (i.e. Genetic Algorithms, Particle Swarm Optimization, Ant Colony Optimization) and statistical modeling. It takes advantage of the ability of computational intelligence algorithms to cope with local optima by evolving several candidate solutions simultaneously [19]. Computational intelligence methods are used to iteratively define, implement and evaluate sets of experiments [7], whereas system response and statistical modeling are used to search for the best performing experiment. In the current study, a Particle Swarm Optimization (PSO) algorithm has been used. PSO has been chosen according to the results of an empirical study comparing the performances of a GA and a PSO. The purpose of the study was to generate a simulated problem with characteristics similar to those of the real problem. Data collected during the first stage of the experimentation have been used for that purpose. In particular, 88 sol-gel coatings have been obtained and their quality has been measured in terms of spot circularity, intensity, background and homogeneity. 40 out of 88 coatings have been obtained during a preliminary study aimed at defining suitable ranges for the input variables (chemical components). The remaining 48 coatings are random experiments and coincide to the first generation of a GA-based approach or to the first swarm of a PSO-based methodology. The empirical work, that will be published somewhere, resulted in a better performance of the PSO-based approach. PSO is a population based heuristic inspired by the flocking behavior of birds [20]. To simulate the behavior of a swarm, each bird (or particle) is allowed to fly towards the optimum solution. Within the evolutionary approach, the swarm corresponds to a set of experiments and each particle defines an experiment. At each iteration of the PSO algorithm the swarm moves and a new set of experiments is selected. The movement of the swarm is guided by the basic rules of a PSO algorithm. Because of the presence of multiple responses, a method based on distance functions has been adopted [21] and a weighted metric has been developed to take into account the different relevance of the responses. Furthermore, a statistical model has been included in the procedure to support the search of the optimum solution. The Multivariate Adaptive Regression Splines (MARS) approach [22] has been chosen for this purpose because of the many attractive features: flexibility, ability to cope with high-dimensional data and multiple responses, capacity of modelling complex non-linear relationships, and clear interpretability [23]. The PSO-MARS algorithm is described in section IV C.

The evolutionary approach discussed in the current paper, can be considered a one-stage approach [17] in that a target point is assumed and the selection of the additional points is not based on the properties of the approximation function previously obtained. At each iteration of the algorithm four approximating functions, one for each response, are obtained on the basis of a sample of evaluated data points. The approximating function is a multiresponse MARS model and the sample of evaluated data points are the experiments implemented and evaluated in the laboratory. The approximated functions are then used to find a target point, also called demand-level vector. As will be further explained, the objective point is not the optimum solution, like in the second step of the two-stage approach. Instead, it is an intermediate solution that helps the search of the optimum point. Even if a comparison has not been carried out, the evolutionary approach is expected to offer two main advantages with respect to the two-stage approach. They are the reduced number of evaluated data points necessary in the function approximation step, and the improved accuracy obtained in the vicinity of the local and global optima. Indeed, thanks to the iterative procedure, a relatively small number of evaluated data points is used to initially investigate the whole space and subsequently exploit the most promising regions. The cost for the increased accuracy is a lack of knowledge on the remaining part of the objective functions: the functional relationships between the input (chemical components) and the output variables are less reliable as the vicinity to the promising regions decreases. However, because the focus is on process optimization, this is an acceptable drawback.

IV. CASE STUDY

A. The microarray technology

DNA microarrays are high capacity systems to monitor the expression of tens of thousands of genes in parallel [2]. They are produced by spotting biological material onto a substrate (glass slides or membrane) in arrayed positions. The spotted biological material (target) is typically composed by oligonucleotides or PCR products [2], [3]. Oligonucleotides are short chain of single-stranded DNA or RNA [2]. PCR stands for Polymerase Chain Reaction, an amplification technique used to generate thousands to millions of copies of a DNA, known as PCR products [24]. Target molecules enconding for a specific gene, are associated to a given position in the substrate. To efficiently bind target molecules, the substrate need to be functionalized, that is covered with a thin coating of poly-lysine, amino silanes or other reactive silanes [6]. Coated slides are typically fabricated via sol-gel and plasma depositions [4], [5]. Subsequently, fluorescent-labeled (or radioactively-labeled) cDNAs are deposited over the spotted array and zipper up (hybridize) to the complementary strands of the target sequences. Unhybridized cDNA molecules are then washed off and the glass slide is scanned with a laser scanner. Hybridized dye molecules are therefore detected and digitalized images are recorded. The scanned images are composed by an arrayed grid of colored circles, named spots. The position of a spot in the grid identifies the

corresponding gene, whereas the intensity of the spot (color intensity) identifies the level of expression of the associated gene. Finally, the digitalized images undergo an analysis process for the statistical selection of significantly expressed genes. Because the main source of inherent variability in the expression data derives from the spot quality [25], ideal spots should be obtained. Ideal spots are perfectly circular and identical, have a uniform signal both inside (intensity signal) and outside (background or noise signal) the spot, and are characterized by a high signal-to-noise ratio [25], [5], [26], [27]. Crucial in obtaining good quality spots is the microarray surface, and therefore the glass slide coating [2]. In the current study, amino-functionalized sol-gel coatings are investigated. With this platform, target molecules bind the substrate by means of the attractive forces between the positive charges on the amino groups and the negative charges on the biomolecules [2]. Due to the preliminary nature of the study and to reduce the experimental cost, fluorescent-labeled cDNAs are replaced by fluorescein isothiocyanate (FITC). This is a dye molecule capable of simulating the behavior of cDNA molecules because of their chemical affinity with the amino group.

B. The sol-gel chemistry and the microarray fabrication process

Sol-gel is defined as the synthesis of ceramic materials by preparation of a sol, gelation of a sol, and removal of the solvent [28]. This technique has already been applied for the productions of microarray coatings because of many attractive features including fine control of chemical composition, low post thermal treatment, low cost of equipment, reduced steric hindrance problems associated with covalent immobilization methods, higher density of reactive groups on the surface, preserved activity of immobilized biomolecules, and improved hybridization kinetics [28], [26]. Furthermore, sol-gel microarray coatings with tunable properties can be produced using different patterning technologies [29], [30], [31]. Sol-gel coatings have been produced by dipping a pre-cleaned glass slide into a chemical solution (dip-coating deposition). The chemical solution, named sol, is a mixture of six components: ethanol (EtOH), deionised water (H_2O), hydrochloric acid (HCl 1M), aminopropyltriethoxysilane (APTES), methyltriethoxysilane (MTES) and tetraethoxysilane (TEOS). Each component can enter the mixture in different quantities. On the basis of 40 preliminary experiments, six number of moles (factor levels) have been defined for each component, as reported in Table 1. Ranges and levels of the factors have been determined with the purpose of considering a wide range of compositions. Simultaneously, practical limitations, such as costs and precision of the laboratory instruments, have been taken into consideration. Relevance has been assigned also to experience-based aspects: only levels of the independent variables likely to generate detectable changes in the sol have been used. Furthermore, two constraints have been defined: the total number of moles for APTES, MTES and TEOS must be equal to $5 \cdot 10^{-3}$

TABLE I

FACTOR LEVELS (NUMBER OF MOLES (M) AND MICROMOLES (MM) USED BY THE PSO-MARS ALGORITHM TO GENERATE THE SOL-GEL RECIPES (EXPERIMENTS))

EtOH (M)	H_2O (M)	HCl (M)	TEOS (mM)	MTES (mM)	APTES (mM)
0.170	0.014	0.010	2.19	0.25	0.21
0.306	0.039	0.028	2.63	0.70	0.58
0.442	0.064	0.046	3.07	1.15	0.97
0.578	0.089	0.064	3.51	1.60	1.34
0.714	0.114	0.082	3.95	2.05	1.72
0.850	0.139	0.100	4.39	2.50	2.10

$$\text{APTES} + \text{MTES} + \text{TEOS} = 5 \cdot 10^{-3}\text{moles} \quad (1)$$

and the number of TEOS moles must be more than the 50% of $5 \cdot 10^{-3}$

$$\text{TEOS} > 2.5 \cdot 10^{-3}\text{moles}. \quad (2)$$

Once the sol has been obtained, dip-coating deposition is performed using 2.5 mm s^{-1} withdrawal speed at 20% relative humidity (RH). Subsequently, the sol evolves towards the formation of a gel thus generating a film (coating). The film is then dried for 30 minutes at 100°C in order to remove the solvent.

Functionalized slides are spotted using a FTA 1000 contact angle tester in order to control a 50 μL drop deposition. The spotted solution is obtained by adding Triton X to a sodium phosphate buffer solution. FITC dye (7 μg/ml) is then deposited over the spotted slide. The coated slide is aged for one hour (30°C and 68% relative humidity) to simulate the hybridization conditions. The slide is finally rinsed with methanol and water to remove the unreacted dye molecules, and then dried. Digital images of the spotted slides are obtained using a laser scanner GenePix 4000B (Axon Instruments). Measurements are performed with the 15 μm resolution, under excitation with 532 and 633 wavelengths. Genepix software has been employed to analyze the scanned images and derive measures of the coatings quality by considering Cir, Int, Back, and Hom.

C. The evolutionary PSO-MARS algorithm

In order to optimize the sol-gel process using the evolutionary approach, a customized algorithm has been developed. To implement the methodology the statistical software R [32] has been employed. The multi-objective PSO, inspired by the work of Tripathi [33], has been combined with a multiresponse MARS model obtained using the R package earth [34].

Let's first recast the case study in a process optimization framework. Here, the resources are the material scientists, the laboratory instruments, the sol-gel chemistry, and so on. The inputs are compositional variables (EtOH, H_2O, HCl, APTES, MTES, and TEOS) and process variables

(withdrawal speed, relative humidity, heating temperature, heating time, and so on). To avoid an exponential growth of the problem dimensionality, process variables are kept constant and only compositional variables are investigated. The output is the sol-gel coating and its characteristics are measured by four response variables: spot circularity (Cir), spot intensity (Int), background intensity (Back) and homogeneity of the intensity signal inside the spot (Hom). The four responses are not considered equally relevant, therefore a set of normalized weights has been assigned to them. Higher weights are associated to the more important response variables. In particular, a weight of 0.4 has been assigned to Cir and Int, a weight of 0.12 has been assigned to Back, and a weight of 0.08 has been assigned to Hom. The aim of the sol-gel process optimization problem is to identify the recipe (process conditions) that generate the best coating (optimize the four responses simultaneously).

The customized evolutionary PSO-MARS algorithm is as follows. Let $X = (x_1, \ldots, x_p)$ be an $N \times p$ matrix of independent variables, where N is the number of feasible experiments and $p = 6$ is the number of factors. Let $Y = (y_1, \ldots, y_q)$ be an $N \times q$ matrix of response variables, where $q = 4$ is the number of performance metrics. Each independent variable is studied at 6 levels giving rise to 6^6 experiments. Because of the constraints (1) and (2), this number reduces to $N = 8424$ feasible experiments. The value of each independent variable x_j, $j = 1, \ldots, p$, indicates the number of moles of the components used to prepare the sol (x_1 = moles of EtOH, x_2 = moles of HCl 1M, x_3 = moles of H_2O, x_4 = moles of MTES, x_5 = moles of TEOS, x_6 = moles of APTES). The response values y_z, $z = 1, \ldots, q$, measure the goodness of the coating in terms of the spots characteristics (y_1 = Cir, y_2 = Int, y_3 = Back, y_4 = Hom). To control the influence of factors, such as humidity and temperature, during thermal curing and testing processes, reference samples are prepared. Reference samples are obtained by covering glass slides with the best coating found using the OFAT approach. Once a new set of coated slides is synthesized a reference sample is prepared. The response variables are therefore ratios and their values compare the performances of the coatings obtained with two different approaches (evolutionary vs OFAT). The ith row of the matrix X corresponds to the ith experiment, that is the ith sol-gel recipe. It is fully characterized by the $p \times 1$ vector $x_i^T = (x_{i1}, \ldots, x_{ip})$ and the $q \times 1$ vector $y_i^T = (y_{i1}, \ldots, y_{iq})$, where the value x_{ij}, $j = 1, \ldots, p$, is the number of moles of the jth chemical component X_j in the ith experiment, and y_{iz} is the value of the zth response Y_z measured in the ith experiment. The purpose is to find the sol-gel recipe that optimize all the responses simultaneously under the constraints (1) and (2). Cir, Int and Hom must be maximized, whereas Back must be minimized. A trade-off exists between the responses Int and Back, because an increased intensity signal generally corresponds to an increased background signal. Indeed, intensity and background signals both depend on the same elements, such as amount of bounded dye molecules and grafting conditions. Factor levels in Table 1 have been considered.

The PSO-MARS algorithm is divided into an initialization and an iteration step. The initialization step occurs at the iteration $t = 0$, whereas the iteration step occurs at $t > 0$. During the initialization step, a swarm of $N_0 = 48$ particles (set of N_0 experiments) is randomly generated: a random sample is obtained from the matrix X. The selected recipes are used to produce N_0 coatings and the quality of the obtained functionalized slides are measured. For each experiment, the four measured responses are used to calculate the fitness value f_i. The fitness values indicate the coatings quality and are used to rank the selected recipes on the basis of all the measured response values. They are calculated as follows. A multiresponse MARS model is fitted to the experimental data obtained at $t = 0$ (N_0 evaluated data points) and used to predict the remaining $N - N_0$ data points. The response variable Back is transformed into 1-Back, so that all the responses can be maximized. The response values of all the N experiments are normalized in the range $[0, 1]$. The weighted response values are calculated using the vector $w = (w_1, \ldots, w_q)$ of weights, where $w_1 = 0.4$, $w_2 = 0.4$, $w_3 = 0.12$, and $w_4 = 0.08$. An objective point, or demand-level vector [21], y_{obj}^T, is identified by combining the maximum value of each single response. That is, the demand-level vector is given by $y_{\text{obj}}^T = (\max(\text{Cir}), \max(\text{Int}), \max(\text{Back}), \max(\text{Hom}))$ and, because of the normalized weights, it is equal to $y_{\text{obj}}^T = (0.4, 0.4, 0.12, 0.08)$. The fitness value is then calculated using the weighted response values $t_{i,z} = w_z \cdot y_{i,z}$, and the weighted objective point $t_{\text{obj},z} = w_z \cdot y_{\text{obj},z}$, according to the function

$$f_{i(t)} = \sum_{z=1}^{q} |t_{i,z} - t_{\text{obj},z}| \tag{3}$$

where, $f_{i(t)}$ indicates the fitness value of the ith particle at iteration t. Due to time and cost limitations, only $N_s = 15$ experiments are investigated during the subsequent iteration step. The best N_s recipes are therefore selected from the N_0 investigated experiments and the vector of fitness values reduces to $(f_{1(t)}, \ldots, f_{Ns(t)})$. Once the fitness values $(f_{1(t)}, \ldots, f_{Ns(t)})$ are selected, t is increased and the algorithm enter the iteration step. The iteration step repeats the subsequent stages for a predetermined number of iterations, $C = 5$. The fitness values obtained so far are used to identify two kind of elements: the global best and the personal bests. The global best is the recipe that outperforms the compositions studied by the whole swarm. Let S be the set of investigated recipes, and let F_s be the associated set of fitness values. The global best at iteration t is the $1 \times p$ vector Gb_t, and it is identified according to

$$Gb_t = x_i^T, \text{ s.t. } f_i = \max(f_u \in F_s) \text{ and } x_i^T \in S \tag{4}$$

where u ranges from 1 to the cardinality of F_s. The personal best is the recipe that outperforms the recipes studied by the same particle up to iteration t. Indeed, as t increases, each particle is allowed to move and study new recipes (points in the search space). The number of personal bests is equal to

the number of particles, that is $N_s = 15$. Let S_r be the set of recipes investigated by the rth particle, with $r = 1, \ldots, N_s$, and let F_{Sr} be the associated set of fitness values. The personal best of the rth particle at iteration t is the $1 \times p$ vector $Pb_{r(t)}$ and is identified according to

$$Pb_{r(t)} = x_i^T, \text{ s.t. } f_i = \max(f_v \in F_{Sr}) \text{ and } x_i^T \in S_r \quad (5)$$

where v ranges from 1 to the cardinality of F_{Sr}. The personal bests of all the particles are recorded in the matrix Pb_t. When $t = 1$, the personal bests coincide with the current positions of the particles and are recorded in the $N_s \times p$ matrix $Pb_1 = (x_1^T, \ldots, x_{Ns}^T)$. The new particle positions are then calculated by updating the current position of each single particle. The update consists in adding a velocity component that depends on the following time-varying parameters: inertia weight w_t, and acceleration coefficients $c_1 t$ and $c_2 t$. For a detailed description of these parameters and their updating formulas see Tripathi [33]. The position of each single particle $x_i^T = (x_{i1}, \ldots, x_{ip})$ is obtained by updating the elements x_{ij} ($i = 1, \ldots, N$ and $j = 1, \ldots, p$) according to:

$$v_{ij} = w_t v_{ij} + c_{1t} r_1 (Pb_{ij} - x_{ij}) + c_{2t} r_2 (Gb_j - x_{ij}) \quad (6)$$

$$x_{ij} = x_{ij} + v_{ij}. \quad (7)$$

Here, v_{ij} is the velocity of the element x_{ij}, Pb_{ij} is the element in the ith row and jth column of the matrix Pb_t, and Gb_j is the jth element of the vector Gb_t. Along with the identification of the new particle positions, a new set of N_s experiments is selected. To introduce a certain degree of randomness in the particles flight, a mutation operator is finally applied to each element x_{ij}. See Tripathi for more details on the mutation operator utilized [33]. The probability with which the elements x_{ij} are mutated decreases with t and has been varied from 0.1 at $t = 0$, to 0.07 at $t = 1$, 0.04 at $t = 2$, and 0.01 at $t > 2$. Once the new selected experiments are implemented and the coatings are obtained, the quality of the spots is measured and the new fitness values are calculated using the procedure previously described. The final output produced by the algorithm is the best recipe found during the whole procedure.

V. EXPERIMENTAL RESULTS

A total of 123 recipes have been investigated in six iterations of the algorithm ($t = 0, \ldots, 5$). The best recipe identified by the PSO-MARS algorithm is as follows: EtOH=0.17 moles, HCl 1M=0.00046 moles, H20=0.014 moles, TEOS=0.00263 moles, MTES=0.0016 moles, APTES= 0.000588 moles. Figure 1 shows the scanned images of two spots: one measured on the best coating found using the OFAT approach (left) and the other measured on the best coating found using the PSO-MARS algorithm (right). With respect to the best recipe found using the OFAT approach, significant improvements have been obtained in terms of the signal-to-noise ratio (Int/Back). The PSO-MARS algorithm successively faced the trade-off problem involving the response variables Int and Back. A good balance of

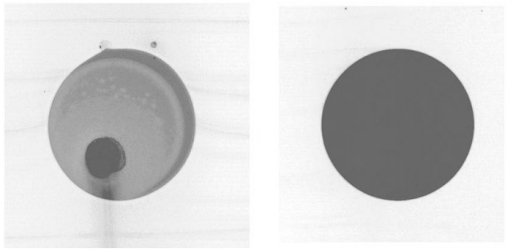

Fig. 1. Scanned images of two spots. Spots are the output of a microarray analysis. They should be circular, characterized by high signal (fluorescent intensity inside the spot), low background (noise; fluorescent intensity outside the spot) and high homogeneity of the signal. The spot measured on the best coating identified using the OFAT approach (left) is circular but has low homogeneity (the fluorescent dye is more concentrated along the circumference and in the bottom left of the spot). The spot on the right has been measured on the best coating identified using the PSO-MARS algorithm. With respect to the spot on the left, it presents improved features: perfect circularity, high homogeneity, and high signal-to-noise ratio (signal/background)

the chemical components has been identified and a coating with improved signal-to-noise value has been produced. An improvement higher than 240% has been measured. Furthermore, the enhancement achieved in terms of the spot homogeneity (Hom) has been around 90% (Figure 1). Perfectly circular spots have been found by both the investigated approaches (OFAT and evolutionary). With respect to the best OFAT coating, the PSO-MARS algorithm has therefore been able to identify a recipe resulting in perfectly circular spots, drastically increased homogeneity and signal-to-noise ratio.

VI. CONCLUSIONS

The proposed method can cope with problems simultaneously characterized by challenging aspects such as unknown fitness function, high dimensional variable space, constraints on the independent variables, multiple responses, expensive or time-consuming experimental trials, expected complexity of the functional relationships between independent and response variables. With respect to the OFAT approach, the proposed methodology has some advantages such as estimating the interaction between factors, identifying the relationship between factors and responses, recognizing the important factors, exploring a bigger area of the search space. With respect to RSM, the proposed approach can face nonstandard situations deriving from the presence of constraints on the factors and high dimensionality; furthermore, increased flexibility is achieved using MARS in place of polynomial models. If applied for experimental design purposes, the two stage approach would require an excessively high number of trials to obtain a good approximation of the unknown objective function over the entire space. The proposed approach has the advantage of requiring an acceptable number of executed experiments to achieve significant improvements of multiple responses. The PSO-MARS algorithm focuses on the most promising areas of the search space and employs strategies to overcome the problem of premature convergence. Differently from

computer-generated design, the new experimental points are selected according to the system performance.

From the EC perspective, the proposed approach is a surrogate methodology to integrate models into an evolutionary optimization algorithm [35]. The novelties are the distribution-free one-stage approach and the use of the MARS approach for fitness function modeling. Similarly to the Gaussian process models, the MARS approach does not require a predefined structure because no strong model assumptions are made. Furthermore, complex non-linear relationships can be approximated, and the resulting model is clearly interpretable. The advantage of MARS over Gaussian process models is the reduced computational cost. Indeed, for N data points, Gaussian process models require $O(N^3)$ steps to construct the GP, $O(N)$ to predict the mean function value at a new point, and $O(N^2)$ to predict the standard deviation [35]. Instead, the computational cost of the MARS procedure can be made proportional to $nNM_{\max}$, where n is the number of factors, N is the number of data points, and $M_{\max}$ is the maximum number of basis functions [22].

Because the actual objective function is unknown, the convergence properties of the evolutionary algorithm are unclear [36]. This is an area we need to address. Benchmark functions could be used for that purpose [14], [37], [36]. The performance of the proposed methodology has been evaluated with respect to the best result identified using OFAT, the approach traditionally used by scientists. Good quality coatings for microarray applications have been obtained and significant improvements have been achieved. The customized algorithm successfully faced the challenging aspects of the investigated problem and showed to be a potentially efficient methodology for process optimization purposes.

ACKNOWLEDGMENT

This work has been supported by the Fondazione di Venezia within the DICE project. CIVEN is acknowledged for supporting characterizations.

REFERENCES

[1] D. C. Montgomery, *Design and Analysis of Experiments*, Hoboken NJ: Wiley, 2009.

[2] M. Schena, *Microarray analysis*, New York: Wiley-Liss, 2003.

[3] D. J. Duggan and M. Bittner and Y. Chen and P. Meltzer and J. M. Trent, "Expression profiling using cDNA microarrays", *Nature Genetics*, vol. 21, pp. 10–14, Jan. 1999.

[4] S. D. Conzone and C. G. Pantano, "Glass Slides to DNA microarrays", *Nature Genetics*, vol. 7, pp. 20–26, Mar. 2004.

[5] J. Donggeun and Y. Sanghak and K. Jinmo and K. Bongjun and J. Bohwan and R. Doug-Young, "Formation of amine groups by plasma enhanced chemical vapor deposition and its application to DNA array technology", *Surf. Coat. Technol.*, vol. 200, pp. 2886–2891, Feb. 2006.

[6] M. Schena and D. Shalon and R. Heller and A. Chai and P. O. Brown and R. W. Davis, "Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes", *Proc. Natl. Acad. Sci USA*, vol. 93, pp. 10614–10619, Oct. 1996.

[7] M. Forlin and I. Poli and D. De March and N. Packard and G. Gazzola and R. Serra, "Evolutionary experiments for self-assembling amphiphilic systems", *Chemom. Intell. Lab. Syst.*, vol. 90, pp. 153–160, Oct. 2008.

[8] V. Czitrom, "One-Factor-at-a-time Versus Designed Experiments", *The American Statistician*, vol. 53, pp. 126–131, May 1999.

[9] C. Daniel, "One-at-a-Time Plans", *Journal of the American Statistical Association*, vol. 68, pp. 353–360, Jun. 1973.

[10] G. E. P. Box and K. B. Wilson, "On the Experimental Attainment of Optimum Conditions", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 13, pp. 1–45, 1951.

[11] J. Kiefer, "Optimum Experimental Designs", *Journal of the Royal Statistical Society. Series B (methodological)*, vol. 21, pp. 272–319, 1959.

[12] J. Kiefer and J. Wolfowitz, "Optimum Design in Regression Problems", *The Annals of Mathematical Statistics*, vol. 30, pp. 271–294, Jun. 1959.

[13] K. Smith-Miles and J. Gupta, "Integrating Feedforward and Feedback neural networks for optimization" in *Handbook of Intelligent Control: Neural*, *Fuzzy*, *and Adaptive Approaches* (Intelligent Engineering Systems Through Artificial Neural Networks: Neural Networks, Fuzzy Logic, Evoultionary Programming, Data Mining, and Rough Sets), Edited by C. Dagli et al., ASME Press, 1992.

[14] K. Smith-Miles and J. Gupta, "Continuous function optimization via gradient descent on a neural network approximation function" in *Handbook of Intelligent Control: Neural*, *Fuzzy*, *and Adaptive Approaches* (Connectionist Models of Neurons, Learning Process, and Artificial Intelligence), Lecture Notes in Computer Science, vol. 2084, Springer-Verlag, Berlin, pp. 741-748, 2001.

[15] J. Sacks and W. J. Welch and T. J. Mitchell and H. P. Wynn, "Design and Analysis of Computer Experiments", *Statistical Science*, pp. 409-423, 1989.

[16] F. Hutter and H. H. Hoos and K. Leyton-Brown and K. P. Murphy, "An Experimental Investigation of Model-Based Parameter Optimisation: SPO and Beyond", *GECCO 09*, July 8-12, 2009.

[17] D. R. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces", *Journal of Global Optimization*, vol. 21, 345383, 2001.

[18] A. Sòbester and S. J. Leavy and A. J. Keane, "A Taxonomy of Global Optimization Methods Based on Response Surfaces", *Journal of Global Optimization*, vol. 21, 345383, 2001.

[19] S. Paterlini and T. Krink, "Differential evolution and particle swarm optimization in partitional clustering", *Computational Statistics and Data Analysis*, vol. 50, pp. 1220–1247, Jan. 2006.

[20] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," *Proc. IEEE Int. Conf. on Neural Networks*, Piscataway, NJ, 1995, pp. 1942–1948.

[21] N. Srinivas and K. Deb, "Multiobjective Optimization using Nondominated Sorting in Genetic Algorithms", *Evol Comput*, vol. 2, pp. 221–248, 1994.

[22] J. H. Friedman, "Multivariate Adaptive Regression Splines", *Ann Stat*, vol. 19, pp. 1–141, 1991.

[23] T. S. Lee and I. F. Chen, "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines", *Expert Syst Appl*, vol. 28, pp. 743–752, 2005.

[24] G. Schochetman and C. Y. Ou and W. K. Jones, "Polymerase Chain Reaction", *The Journal of Infectious Diseases*, vol. 158, pp. 1154–1157, Dec. 1988.

[25] X. Wang and S. Ghosh and S. W. Guo, "Quantitative quality control in microarray image processing and data acquisition", *Nucleic Acids Res*, vol. 29, e75, 2001.

[26] S. Venkatasubbarao, "Microarrays - status and prospects", *Trends Biotechnol.*, vol. 22, pp. 630–637, 2004.

[27] N. Giannakeas and D. I. Fotiadis, "An automated method for gridding and clustering-based segmentation of cDNA microarray images", *Comput. Med.Imaging Graph.*, vol. 33, pp. 40–49, 2000.

[28] C. J. Brinker, *Sol-gel science, the physiscs and chemistry of sol-gel processing*, Academic Press: Boston, 1990.

[29] P. Falcaro and M. Takahashi, "Patterning Techniques for Mesostructured Films", *Chem Mater*, vol. 20, pp. 607–614, 2008.

[30] P. Falcaro and S. Costacurta and L. Malfatti and M. Takahashi and T. Kidchob and M. Casula and M. Piccini and A. Marcelli and B. Marmiroli and H. Amenitsch and P. Schiavuta and P. Innocenzi, "Fabrication of Mesoporous Functionalized Arrays by Integrating Deep X-Ray Lithography with Dip-Pen Writing", *Adv. Mater.*, vol. 20, 2008.

[31] P. Falcaro and P. Innocenzi, "X-rays to study, induce, and pattern structures in sol-gel materials", *J. Sol-Gel Sci. Technol.*, published online, DOI: 10.1007/s10971-009-2127-7.

[32] R Development Core Team (2009). “R: A language and environment for statistical computing”. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

[33] P. K. Tripathi and S. Bandyopadhyay and S. K. Pal, “Multi-objective Particle Swarm Optimization with time variant inertia and acceleration coefficients”, *Inf Sci*, vol. 177, pp. 5033–5049, 2007.

[34] S. Milborrow derived from mda:mars by T. Hastie and R. Tibshirani. (2009). “earth: Multivariate Adaptive Regression Spline Models”. R package version 2.3-5. http://CRAN.R-project.org/package=earth

[35] D. Bürche and Nicol N. Schraudolph and P. Koumoutsakos, “Accelerating Evolutionary Algorithms with Gaussian Process Fitness Function Models”, *IEEE T Syst Man Cyb*, vol. XX, No. Y, pp. 1–12, 2004.

[36] Y. Jin and M. Olhofer and B. Sendhoff, “A Framework for Evolutionary Optimization with Approximate Fitness Functions”, *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 481-494, 2002.

[37] Y. Sun and D. Wierstra and T. Schaul and J. Schmidhuber, “Efficient Natural Evolution Strategies”, *Genetic and Evolutionary Computation Conference*, pp. 481-494, 2002.