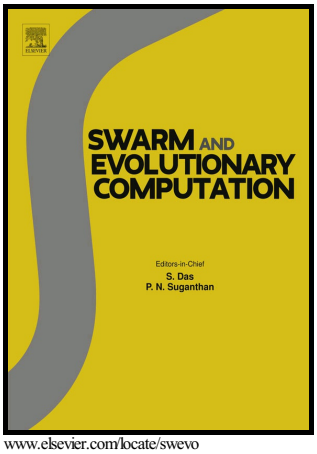
Screening dense and noisy DOX-datasets with NN-blending and "dizzy" swarm intelligence: Profiling a water quality process

George J Besseris



www.elsevier.com/locate/swevo

PII: S2210-6502(16)30196-1
DOI: http://dx.doi.org/10.1016/j.swevo.2016.08.003
Reference: SWEVO230

To appear in: *Swarm and Evolutionary Computation*

Received date: 25 February 2015
Revised date: 21 November 2015
Accepted date: 16 August 2016

Cite this article as: George J Besseris, Screening dense and noisy DOX-dataset with NN-blending and "dizzy" swarm intelligence: Profiling a water quality process, *Swarm and Evolutionary Computation* http://dx.doi.org/10.1016/j.swevo.2016.08.003

Screening dense and noisy DOX-datasets with NN-blending and "dizzy" swarm intelligence: Profiling a water quality process

George J Besseris*

Advanced Industrial & Manufacturing Systems program, Mechanical Engineering Department, TEI of Piraeus (Greece) and Kingston University (UK).

E-Mail: besseris@teipir.gr

Abstract

A novel nature-inspired method is presented in this work for resolving product/process development or improvement with design of experiments (DOX). The technique is suitable for difficult Taguchi-type multifactorial screening and optimization studies that need to simultaneously contain the double hassle of controllable and uncontrollable noise intrusions. The three-part sequential processing routine requires: 1) a regressive data-compression preprocessing, 2) a smart-sample generation using general-regression neural networks (GRNN), and 3) a screening power prediction using 'reverse' swarm intelligence (SI). The approach is primed to confront potential non-linearity and data messiness in the examined effects. The Taguchi-type orthogonal-array (OA) sampler is tuned for retrieving information in controllable (outer OA) and uncontrollable (inner OA) noises. The OA-saturation condition is elicited for maximum data exploitation. GRNN-fuzzification consolidates into a single contribution the uncertainty from all possible sources. The resulting 'smart' sample is defuzzified by a robust-and-agile data reduction. Screening-solution meta-power is controlled with a new SI-variant. The independent swarm groups, as many as the studied effects, are tracked toward preassigned targets, i.e. their ability to return to their host beehives. The technique is illustrated on a complex purification process where published multifactorial data had been collected for a critical wastewater paradigm and thus may be used to test the benchmark solution. However, environmental water-qualimetrics are profoundly dominated by messy data as justified in this work. We elucidate on several issues that regular Taguchi methods may be benefited by the proposed GRNN/SI processing while emphasizing the consequence of overlooking the underlying assumptions that govern standard comparison models. The new swarm itelligence method offered a practical way to estimate a first-time "soft" power measure for the inner/outer OA optimization case that was impossible with ordinary statistical multi-factorrial treatments. Key performance advantages in efficiency, robustness and convenience are highlighted against alternative approaches.

Keywords: Stochastic optimization, Taguchi methods, swarm intelligence, GRNN, smart sampling, water quality.

1. Introduction

1.1 *Product/process development with nature-inspired methods*

Modern process/product development poses a major challenge and nature-inspired techniques have been promising in conquering it [1-4]. Real-world products and processes are intricate systems that do not surrender trivially to ordinary data mining methods. It is non-linearity, non-normality and incompleteness that mainly fuels data messiness in pragmatic operational environments [5, 6]. Moreover, high complexity stems from the implication of a large number of interweaved effects which confound the knowledge discovery process with their veiled ipseities. To resolve convincingly the potency of vital effects in product/process development studies requires structured experimentation that confronts data messiness. 'Many-assumptions' data mining deters robust screening which is essential for reliable product/process diagnostics [7-9]. On the other hand, nature-inspired mining methods have been proved advantageous because of their capability to search for a solution often stripped of any overt assumptions [10-13]. Robust and intelligent multivariate inference when is founded on well-planned structured data will provide screening results with high-fidelity effect-signatures.

1.2 *Product/process data generation with Taguchi-type samplers*

The adoption of Taguchi methods has been proved fruitful in ameliorating industrial operations [14]. This has been accomplished by placing great emphasis on snappy, yet structured, data generation. Taguchi methods promote easy-to-use stochastic screening and optimization tools for the quick and economic delivery of enhanced product/process performance. Successful implementations span a multitude of applications in diverse environments [15]. The Taguchi-mindset focuses on brief yet meticulous design of experiments (DOX) to create concise datasets. It was also akin to simple multi-factorial data-processing in order to gather cheap and convenient information. The common-core steps for executing a Taguchi improvement study is: 1) plan constrained experiments with orthogonal arrays (OAs) – a type of fractional factorial designs (FFDs) [16, 17], 2) compress the replicated trial-runs using the signal-to-noise ratio (SNR) estimator, 3) analyze the ensuing multifactorial problem with analysis of variance (ANOVA), and 4) confirm the predictions. In a nutshell, setting up a screening/optimization study relies on field data to be acquired through the implementation of a proper Taguchi-OA-sampler plan.

1.3 *Taguchi-type screening/optimization for water quality improvement*

Water quality improvement studies are extremely popular today due to the worldwide increasing scarcity in clean water resources [18, 19]. The dramatic necessity for this commodity has directed efforts to recover consumable water primarily by purifying wastewater pools as well as desalinating seawater [20]. For water to be drinkable, as well as usable for other household and industrial needs, high-purity standards and specifications have been established by local, national and international agencies [21]. Therefore, in high demand are water-quality screening/optimization studies in connection with pollution remediation due to heavy elements and other industrial organics [22]. Instructive contributions of Taguchi-inspired water-quality studies may be identified with respect to increasing toxic-substance removal efficiencies [23, 24]. Structured experimentation which is assisted by Taguchi-type solvers has demonstrated promising outcomes for eco-designing operations for innovative wastewater treatment [25-28].

1.4 *A need for a new nature-inspired for translating inner/outer Taguchi-type samplers*

In this work, we will propose a new nature-inspired soft-computing method to confront the stochastic screening/optimization of product/process characteristics under a candidate set of non-linear controlling factors with a declared (controlled) noise factor in addition to other residual (uncontrolled) uncertainty. Thus, the input-output dataset configuration is conveniently constructed by an inner-outer Taguchi-type non-linear OA-sampler. The novelty of the proposed methodology hinges upon its three-pronged design: 1) compression of the outer-OA dataset to saturation with regression analysis, 2) smart-

sampling to discover the effect hierarchy, and 3) "dizzy"-swarm intelligence (DSI) to verify the power of the effect-potency predictions.

The application that motivated this research attempt was a water-quality improvement effort. The implementation of a saturated Taguchi-type OA-sampler anticipated severe ambient interference. Saturated OA schemes known for delivering maximum exploitability continue to intrigue modern intelligent chemometrics [29, 30]. The investigated controlling factors are bound to be screened synchronously for strength and curvature. The new proposition will entail the dual arrangement – inner/outer OA layout – as it has been recommended by the original Taguchi-DOX for probing simultaneously controlling factors against perturbing noise variables. To reinforce the credibility of our approach we will borrow for testing the experimental datasets of Barrado et al. [31]. We select the particular published datasets on grounds of being a pioneering research in a field of escalating challenge; the instructing of concurrent screening and optimization for water purification from toxic metals in decontaminating wastewater. Managing the information generation in two otherwise sequential data-processing phases - screening and optimization - in parallel, as in the study by Barrado et al. [31], also adds to the perplexity of the new developments. The essential reason for revisiting that specific case study becomes the transparency of the replicated data and the deployment of a non-linear OA at the saturation limit. In a nutshell, the aforementioned article attempted an 'all-inclusive' optimization task with a broad spectrum outlook that even today is rather rare to access in current publications. Nevertheless, we will ponder over several assumptions that may not hold true in executing a typical multifactorial treatment using analysis of variance (ANOVA) as exemplified in the paper of Barrado et al. [31]. Thus, this necessity commands a nature-inspired soft-computing alternative.

It will be demonstrated in the Results section the natural messiness undermining the explored dataset [7, 32]. This is by no means foreign to describing intricate phenomena in an aqueous environment. It simply may urge us to modify respectively our mode of attacking the combined screening/optimization problem in order to ensure the objectiveness of the data conversion [33]. We contemplate upon earlier research that recommended that it is generally invalid to make strict presumptions on the distribution laws governing water-quality datasets [34]. Surely, it is a major impediment in water qualimetrics the fact that normality does not necessarily dictate the stochastic fate of aqua-based chemical systems. Thereby, we adopt a non-parametric framework which is controlled by swarm intelligence for translating scarce but smartly-transformed datasets. The chosen case study is also of raised criticality because it appears to display evidence of fingerprints incurred from unknown and unknowable intrusions. Additionally, it is imperative to be astutely vigilant for the presence of dataset anomalies when a sub-optimal sampling strategy has been imposed for practical/economical purposes on the experimental design. New developments justifiably target the indispensable interpretation of messy water OA-datasets with non-linear micro-analytics. The offered new angle allows inner-workings to remain essentially distribution-free throughout all the stages of data processing. Thus, we suggest a new robust and agile methodology to circumvent several analysis breakdowns and traps in connection to the above published paradigm. The new intelligent and nonparametric multifactorial solver aspires to have a broader applicability and immediate deployability in the extremely complex phenomena which are related to 'aqua-limetrics'. The intelligent part of our methodology is novel because of the combo-treatment with general regression neural networks (GRNN) and "dizzy" swarm intelligence (DSI). Even though the synergistic benefits from forging NN and nature-inspired intelligence have been heralded earlier, the spotlighted tandem action of smart sampling by GRNN and power meta-verification with DSI is unique [35, 36].

The article is organized as follows: 1) we present the new methodology for the regressive/intelligent/nonparametric multifactorial solver, 2) we generate 'new-view' solutions for the Barrado et al. dataset [31], 3) we pinpoint and discuss assumption breakdowns and stochastic dilemmas in applying directly the Taguchi method for the studied inner/outer OA, and 4) we compare our results with the original outcomes and by resorting to predictions from other competing methods.

2. Methods

2.1 Taguchi OA arrangement for controlling factors and controllable noise

Taguchi methods accelerate the knowledge discovery cycle by curtailing dramatically the processing demands along with its associated expenses for data generation. This is accomplished by executing a mini-schedule of astute multifactorial recipes that ensnare, in stochastic terms, the strength profile and the accompanied optimal adjustments for a tested group of explanatory variables. The Taguchi toolbox is suitable for programming experiments when the endpoints of the examined effects have been previously established. Therefore, both, screening and optimization, tasks belong to the same constrained stochastic optimization category. We should clarify that in the traditional experimental design, effect screening precedes parameter optimization [16]. However, screening materializes by fulfilling an obvious optimization aim - to identify and filter-out any weak effects from the initially-nominated pool of explanatory factors. Thus, screening essentially narrows down (minimizes) the original profiling scope by removing those statistically inert effects. Actually, Taguchi methods are considered 'fast-track' experimentation since screening and optimization are conducted synchronously reducing the cycle time for gathering data and extracting information. A flexible yet exclusive rubric of multifactorial recipes accommodates the trial patterns for increasing hordes of investigated effects. Tabulating apt recipe guides that interweave the precise combinations of the investigated factor-settings have been standardized in the augmenting families of orthogonal arrays (OAs) [8, 37].

One of the most challenging OA layouts is the one that makes provisions for parallel investigation of controlling factors and controllable noise factors by allotting for their distinctly dichotomous inputting [14, 15]. Such specialized arrangement for planning trial runs is said to be cast in the '*inner-outer*' array formation. We outline the theoretical framework for the case when the noise factor makes a "messy" appearance which is manifested through the scheduled-recipe trials. This means that the solver is coerced to march on blind to any reference law or in lack of any concrete model that describes the controllable noise. In Figure 1, we display a general representation of an *inner-outer* configuration of an $L_9(3^4)$ OA, since it mimics the real layout as it was set up by Barrado et al. [31]. As a matter of convenience, we retain the acronyms of the investigated controlling factors without loss of generalization for the developments that follow. The studied effects are symbolized as: *T*, *P*, *F* and *H* (Table 1). The quality characteristic that is sought to be minimized is the total residual concentration of metal ions in solution (TRC) and was measured in mg/l. The controllable noise factor (*N*) is the added $KMnO_4$ concentration. For statistical purposes, the controlling-factor settings are initialized as $T = \{T_i,\ i=1,2,\ldots,9\}$, $P = \{P_i,\ i=1,2,\ldots,9\}$, $F = \{F_i,\ i=1,2,\ldots,9\}$, $H = \{H_i,\ i=1,2,\ldots,9\}$. Data are programmed by tweaking the controllable noise *N* at *m* settings, then $N = \{N_i,\ i=1,2,\ldots,m\}$. If the TRC response is replicated *r* times for each *N*-noise setting, then $TRC = \{TRC_{ijk},\ i=1,2,\ldots,m;\ j=1,2,\ldots,r;\ i=1,2,\ldots,9\}$. For the Barrado et al. [31] example, we assign $m=3$ and $r=2$.

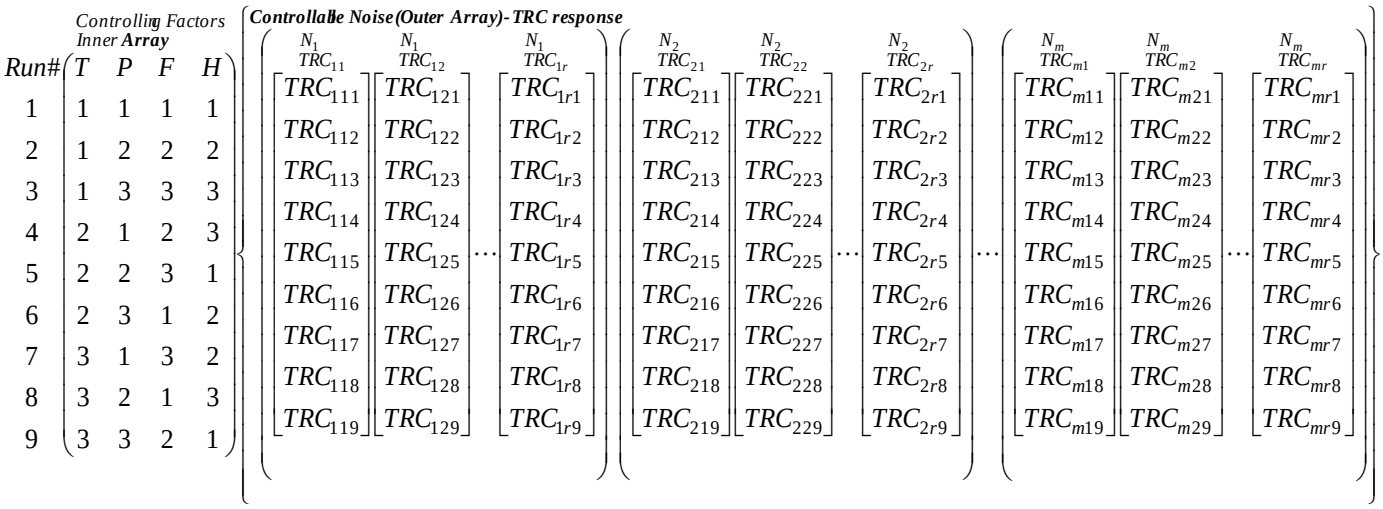


Figure 1: The generalized saturated $L_9(3^4)$ OA layout with replications that include one non-linear noise in the outer array.

Table 1: Initialization of the controlling factors and the controllable noise [31].

	Controlling Factors				Controllable Noise
	Temperature	pH	Ratio of Fe(II) in total metal content	Ageing time	$KMnO_4$ concentration
Symbol	T	P	F	H	N
Units	°C			h	mol/l
Level 1	25	8	2	1	0.00375
Level 2	50	10	7	2	0.0375
Level 3	75	12	15	3	0.075

2.2 Data reduction with effective slope variation

We reduce the data volume which is generated because of the combined action of replication and outer-array expansion. We suggest fitting the data for each trial recipe independently using the simple (second-order) curvilinear equation:

$$log(TRC) = \alpha + \beta \cdot log(N) + \gamma \cdot log^2(N) \quad (1)$$

The logarithmic transformation for both the dependent and independent variables is voluntarily inserted in the modeling proposition to allow for noise settings to transcend several orders of magnitude in the event of dealing with a strong intrusion. The actual noise settings are permitted to ascend at least one order of magnitude in the case study of Barrado et al. [31] while the response magnitudes differentiate by as much as two orders. Furthermore, by fitting data which are produced from an individual recipe, it means that the $m \cdot r$ entries are compressed each time such that to be represented uniformly by the three fitting constants - α, β and γ. Immediately, we recognize three pertinent cases that may be encountered after the completion of the regression analysis step:

1) Equation 1 returns $\beta = 0, \gamma = 0$: This means that the tested noise has no particular effect on the inner-array (recipe) combination.
2) Equation 1 returns $\beta \neq 0, \gamma = 0$: This means the tested noise has a linear effect on the inner-array (recipe) combination.
3) Equation 1 returns $\gamma \neq 0$: This means the tested noise has a non-linear effect on the inner-array (recipe) combination.

Intuitively, we may simplify the noise-effect quantification per each inner-array recipe by introducing the "effective slope" concept. This tactic narrows down to a single estimator the extent of describing the perturbed behavior. The effective slope, β'_ι (i=1,2,…,9), could then be defined as follows:

1) For case 1: $\beta'_\iota = \beta_\iota = 0$, since there is no superimposed component on the response by the induced disturbance.
2) For case 2: $\beta'_\iota = |\beta_\iota| \neq 0$, since there is a superimposed component on the response by the induced disturbance.
3) For case 3: This case may be divided in two separate sub-cases:

   a) Negative or positive curvilinear relationship (Figure 2A and 2B) - we define as effective slope, β'_ι, the absolute magnitude of the slope of the "end-to-end" line passing through the two preset noise boundaries (perforated lines).

   b) Convex or concave (U-shaped curvilinear) relationship (Figure 2C and 2D) - we define as effective slope, β'_ι, the largest of the two absolute magnitudes of the slopes formed by the line that connects either of the two noise boundaries with the respective minimum or maximum point (perforated lines).

Hence, the inner-outer array dataset arrangement of Figure 1 is converted to a saturated-unreplicated layout as shown in Figure 3.

2.3 Homogenizing OA effective-slope data with an intelligent solver

Once the condensed form for the input-output arrangement has been finalized as in Figure 3, it is necessary next to analyze the resulting unreplicated-saturated OA dataset. A promising way to accomplish this is to homogenize the data first using a proven intelligent processor [9]. To convert non-linear relationships for small datasets, the general-regression neural-network (GRNN) processor has become of undisputable worth [38]. However, the introduction of GRNN to defuzzify the OA-dataset is not intended to furnish the terminal profiling outcome of the examined effects. This is not feasible for unreplicated-saturated OA designs due to the additional messiness which is ushered inherently by the random recipe-partitioning requirement which is imposed by the intelligent solver [9, 37]. Instead, the GRNN is utilized to create "smart" samples. Smart samples retain the effect hierarchies which are garnered in the sensitivity analysis reports after multiple GRNN-code executions (Outline 1).

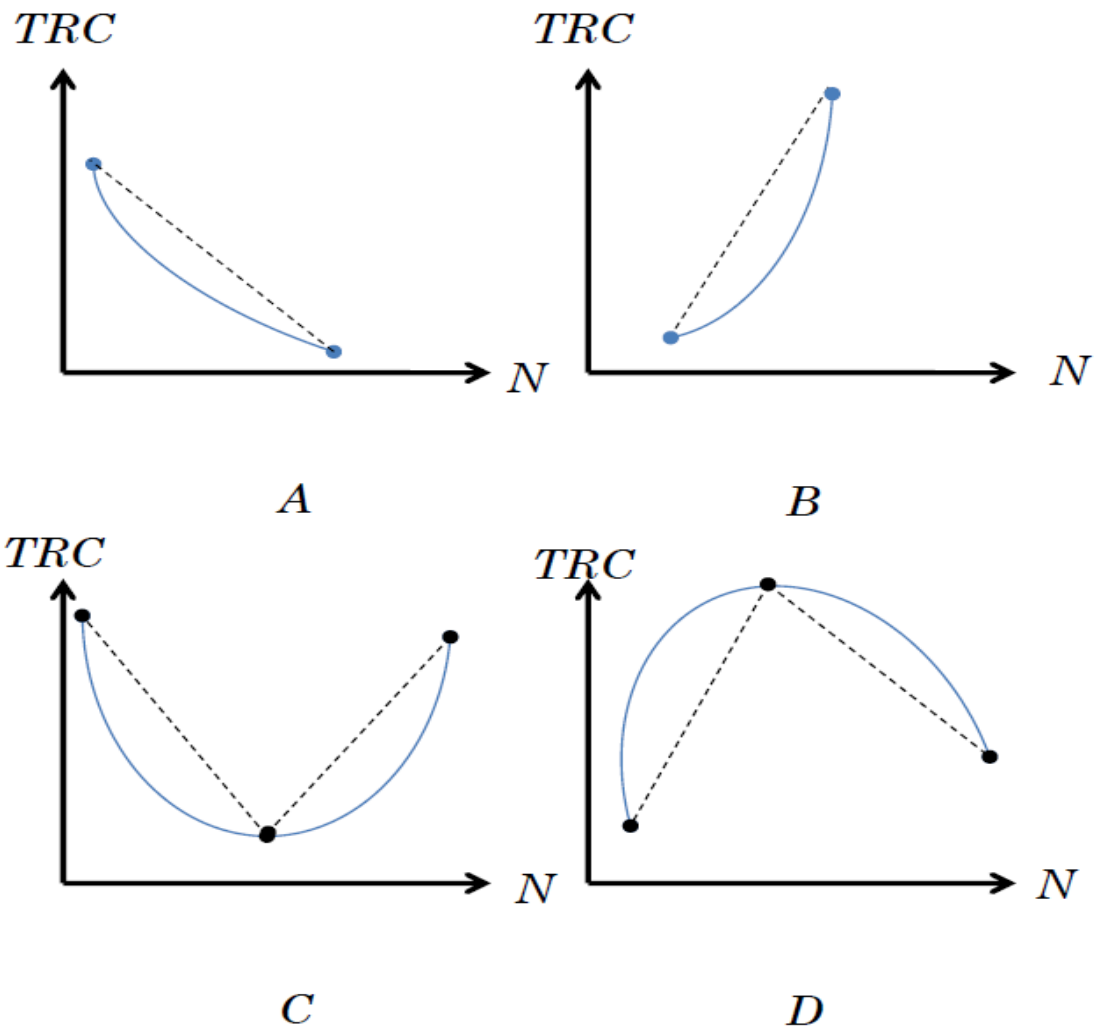


Figure 2: Types of possible curve shapes (A-D) for converting to an effective slope estimator.

$$
\begin{array}{c}
\\
Run\# \\
1 \\
2 \\
3 \\
4 \\
5 \\
6 \\
7 \\
8 \\
9
\end{array}
\begin{array}{c}
\overset{Controlling\ Factors}{\begin{array}{cccc} T & P & F & H \end{array}} \\
\begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 2 & 2 & 2 \\
1 & 3 & 3 & 3 \\
2 & 1 & 2 & 3 \\
2 & 2 & 3 & 1 \\
2 & 3 & 1 & 2 \\
3 & 1 & 3 & 2 \\
3 & 2 & 1 & 3 \\
3 & 3 & 2 & 1
\end{pmatrix}
\end{array}
\begin{array}{c}
\beta\ -slope \\
\begin{bmatrix}
\beta'_1 \\ \beta'_2 \\ \beta'_3 \\ \beta'_4 \\ \beta'_5 \\ \beta'_6 \\ \beta'_7 \\ \beta'_8 \\ \beta'_9
\end{bmatrix}
\end{array}
$$

Figure 3: The converted inner-outer $L_9(3^4)$ OA in terms of the effective slope (β'_ι) response.

In the GRNN architecture, information flows in four layered operations: 1) inputting, 2) patternization, 3) summation, and 4) outputting [39, 40]. The input and output layers store information about the k independent controlling variables, $X \in \{ x_{ij} \mid i=1,2,\ldots,k;\ j=1,2,\ldots,n \}$, and a response, $y \in \{y_j \mid j=1,2,\ldots,n\}$, respectively. The pattern layer distributes information on p nodes. A training sample of size p ($< n$) is randomly selected from the input-layer space x_{ij} to form the relative paired group (x'_{iq} , y'_q) (i=1, 2,…, k; q=1, 2,…, p). Thus, each x'_{iq} (i=1, 2,…, k) is associated with the corresponding q^{th} pattern layer node. A similarity measure advances learning in two contrasting patterns - x_{ij} and x'_{iq} - on the q^{th} pattern node. Then, a suitable distance function, $D_{iq}(x_{ij}, x'_{iq})$ (i=1, 2,…, k; q=1,2,…,p) is defined as:

$$D_{jq}(x_{ij}, x'_{iq}) = \sum_{i=1}^{k} \left(x_{ij} - x'_{iq}\right)^2 \tag{2}$$

On the summation layer, the two respective kernel operations (S_{1i}, S_{2i}) are performed correspondingly on the two available nodes:

$$S_{1j} = \sum_{q=1}^{p} y'_q \exp\left[-\frac{D_{jq}^2}{2\sigma^2}\right] \text{ for } j=1, 2,\ldots, n \tag{3}$$

$$S_{2j} = \sum_{q=1}^{p} \exp\left[-\frac{D_{jq}^2}{2\sigma^2}\right] \text{ for j=1, 2,..., } n \tag{4}$$

The training-sample variance, σ^2, regulates the machine learning pace. On the output layer, the prediction is normalized by balancing the information stored in the two kernels. The prediction of the response vector y as perceived by the GRNN output nodes, $\hat{y}_i$ (i=1,2,…,n), is computed as:

$$\hat{y}_j = \frac{S_{1j}}{S_{2j}} \tag{5}$$

For the screening/optimization problem (Figure 3), the appropriate GRNN topology that manipulates the $L_9(3^4)$ OA-dataset for our case study is depicted in Figure 4 with the proper initializations.

2.4 Nonparametric analysis of smart samples

To initialize a smart sample, the GRNN module is run thirty times [9]. Each time, we retain the rank assignment which is appended in the GRNN sensitivity analysis report. Thus, the smart sample is a list of rank scores where each data column is identified to its respective controlling factor. Based on the start-up smart-sample, sample count adequacy is checked for each effect individually. We are only concerned with the largest sample-count prediction among the examined effects. If this sample count prediction is larger than a value of 30, we complete the additional GRNN module runs and we supplement the initial smart sample with the extra collected data. We confirm the adequacy of the smart sample by iterating this process until the prediction converges to a final smart-sample count value.

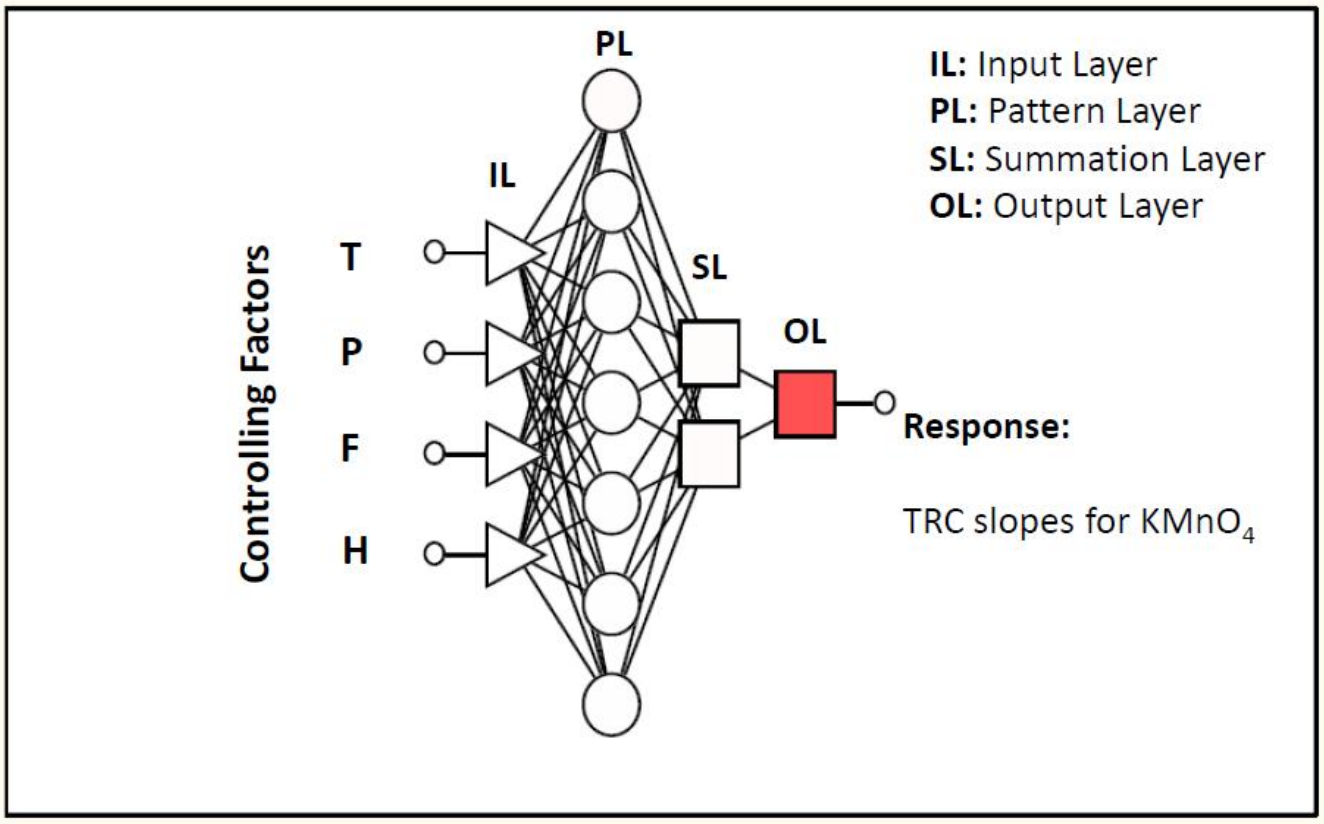
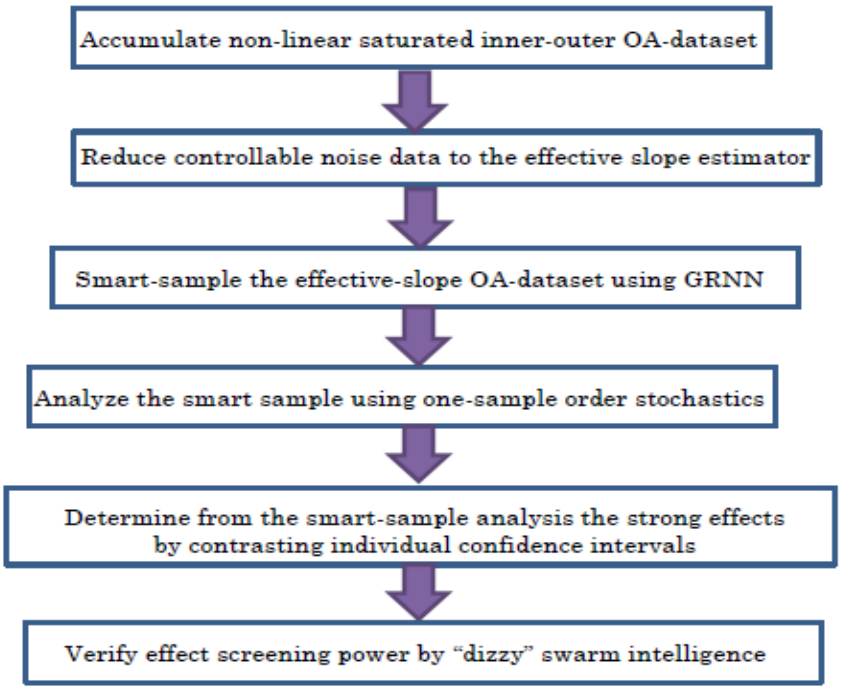


Figure 4: The GRNN topology for homogenizing the four non-linear effects (T,P,F, and H) to a messy condensed response (the effective slope for TRC).



Outline 1: Basic scheme for completing 'smart'-sample analysis.

We incorporate in the final smart sample all additional data which have been produced from the required rounds. On the other hand, if the maximum smart sample count is predicted to be less than a value of 30, the intelligent data-generation process terminates and the data from all thirty runs participate in forming the final smart sample. We note that for all sample count predictions the margin of error for the confidence interval - estimated at 95% - is the halfway distance between two consecutive rank values, i.e. (±) 0.5. This limit signifies the boundary for traversing - up or down the scale - to the adjacent full rank position.

The intelligent embodiment of the messy-data phenomena and other sampling anomalies into the effect rankings does not subscribe to any retrievable reference law that we know of [41]. Therefore, a simple distribution-free inspection of the stability of the effect hierarchy in the smart sample is deemed a reasonable overture. To achieve this, the one-sample Wilcoxon signed-rank test [42, 43] will be used to estimate non-parametrically the 95%-confidence intervals for each of the controlling factors separately. Thus, the proposed robust-intelligent diagnostics will allow for a more reliable resolution of the screened effects.

2.5 "Dizzy" swarm intelligence for testing effect screening power

A new algorithm is developed for verifying the test power of the effect screening outcomes. It is the "dizzy" swarm particle method. It mimics various stages of the colony collapse disorder (CCD) where bee navigational controls succumb to interference from unknown sources. The jamming backlash is expressed computationally by compelling a swarm to also espouse other routes besides their strong propensity to return to their host hive. Offering alternative directions gives opportunities to the swarm to be misguided and eventually to get stranded away from the hive. The new concept may be viewed as a "reverse-swarm intelligence" method because bee-swarms possess inbred capability for returning to their host hives, but now their navigational intelligence has been tampered with an affliction. This affliction is tantamount to an escalating uncertainty about where to land.

Theoretically, the number of nominated swarm groups is as many as the groups of the tested (screening) effects. Suppose that the examined screening effects are k, then, we need S_s (s=1,2,…,k) groups of swarms. Suppose also that each swarm has a total of N_s (s=1,2,…,k) members. Then, the test power, P_s (s=1,2,…,k), will indicate the fraction of a given swarm reaching their hive, i.e. the number of returning bees, n_s (s=1,2,…,k). P_s is also defined as the maximum ratio n_s/N_s of a dispersed swarm group with 'multiple landing' options. It is clear that the disparity N_s-n_s denotes the number of bees in the S_s swarm which have been stricken by CCD and thus missed their hive arrival. This means that each "dizzy" bee (swarm particle) is characterized by as many as k different directions that may opt upon return. Of course, according to the analysis of the smart sample in the previous sub-sections, each of the k directions has its own statistically generated impact. Obviously, the set of weights for each of the implicated swarm group maps the probabilistic effect hierarchy which is derived from the preceding intelligent OA-data conversion. Thus, the "dizziness" of each swarm is predefined by quantifying stochastically the extent of the path that each bee may take. We postulate that in each computational step the direction of the bee is selected randomly, but the size of the step is dictated by a statistical value and it is different for different directions. In our methodology, the statistical step size – expressed in fractions - for each swarm particle is identified by the frequency rating for each of the examined effects which is collected from smart sampling (Section 2.4). Therefore, the overall motion of each swarm group is asymmetric due to the

different statistical distances which are assigned for the k available directions. It makes sense to set the minimum roaming area, i.e. the computational range for all swarm groups, as $A_r = (\lambda \cdot \max(N_s))^2$ (s=1,2,…,k) where λ=1 is each expected iterative step. It is expected only when the bees are totally free from CCD. The value for an approximate start-up value of A_r has resulted after trying out increasing territory ranges. Practically, it was found that reasonable predictions may be obtained with the stated approximation for A_r. All particles from all swarm groups are randomly initialized on the search area A_r. However, the swarm motion is not dictated by λ but by the directional-step fraction, α_{si} (s=1,2,…,k; i=1,2,…,k). Thus, for all k swarms we have:

$$\lambda = 1 = \sum_{i=1}^{k} \alpha_{si} \tag{6}$$

We note that it is the accumulated length of α_{si} in a particular direction for any given swarm particle that determines if a particle has exited the search area, and not the λ. Exiting the search area means the swarm particle reached a target which is either right (its own bee hive) or wrong (to its eventual collapse). Usually, $pd_s = \max(\alpha_{si})$ for all α_{si} (i=1,2,…,k) is the prime direction – the direction to the bee hive - for each separate swarm group, s (s=1,2,…,k). It follows that the amount of 'total dizziness' (td_s) for each swarm group, s, will be:

$$td_s = 1 - pd_s \tag{7}$$

A swarm particle $sp_{si}(t)$, $\{sp_{si}(t) \mid i=1,2,\ldots,N_s;\ s=1,2,\ldots,k\}$, hovers at any iterative step, t, based on the rule evaluation of the directional probability, $dir_{si}(t) = rand()$ where $0 \le rand() \le 1$. The directional rule is simple:

Direction i with i= 1, 2, …, k-1: $(i-1)\cdot\lambda/k \le dir_{si}(t) < i\cdot\lambda/k$ (8)

Direction i=k: $(k-1)\cdot\lambda/k \le dir_{si}(t) \le \lambda$ (9)

The algorithm finishes when all swarm groups have cleared out the roaming area (Algorithm 1). Thus, the individual swarm-particle binning for each separate swarm group - in its predefined horizon - permits the quantification of their mission success by reaching to their respective host hives. It is this binning that completes the simple power estimation by measuring the fraction of success in a particular direction. For each swarm particle $sp_{si}(t)$, at any given time t, we track its distance $dsp_{sij}(t)$ $\{dsp_{sij}(t) \mid i=1,2,\ldots,N_s;\ j=1,2,\ldots,k;\ s=1,2,\ldots,k\}$ from exiting the search territory on the j^{th} direction (j=1,2,…,k). Hence, for each marching step in the j^{th} direction, $dsp_{sij}(t)$ reduces by a value of α_{sj} (not λ). It is easily seen now that the condition for an $sp_{si}(t)$ particle to terminate its computational tracking, and be binned in the j^{th} direction, is simply: $dsp_s(i,j) \le 0$.

Maximization of the test (screening) power is the main objective in our optimization routine but is not sufficient. We adopt the generally-accepted strong value for statistical power of 0.8. Thus, it holds that $P_s \ge 0.8$ for all s = 1,2,…,k. This strong threshold has been set based on the reasoning that Type I errors (alpha risk) are at least four times more serious than Type II errors (beta risk). In turn, setting alpha risk to a regular error rate of 0.05, then the beta risk should comply at a maximum value of 0.2 [44]. This limit signifies at least 80% probability of detecting an effect when it is actually present. It constitutes a new type of maximization objective while self-imposing a minimum constraint cut for solution aptness [45].

Algorithm 1: Pseudo-algorithm for "Dizzy" swarm particle intelligence

Initialize N_s, Ar for k swarm groups

Initialize the k groups of $sp_{si}(t=0)$ swarm particles and their random locations on the search space.

Initialize fractional step α_{si} for each swarm group.

Initialize the possible directions $dir_{si}(t=0)$ for each swarm group based on the rand() outcome (eq. 8, 9).

Determine initial distances of all swarm group particles from all search area borders, $dsp_{sij}(t=0)$.

Loop

While $dsp_{sij}(t) > 0$ do

t=t+1

Update $dir_{si}(t)$ for all particles in all k swarm groups

Update $dsp_{sij}(t) = dsp_{sij}(t) - \alpha_{sj}$

Update $sp_{si}(t)$ position on search space

End Loop

Count $sp_{si}(t)$ on last t for each s swarm separately

Compute P_s for all k swarm groups

If $P_s \geq 0.8$: Declare powerful outcome for the z ($0 \leq z \leq k$) swarm groups – z swarm groups returned to their host bee hives.

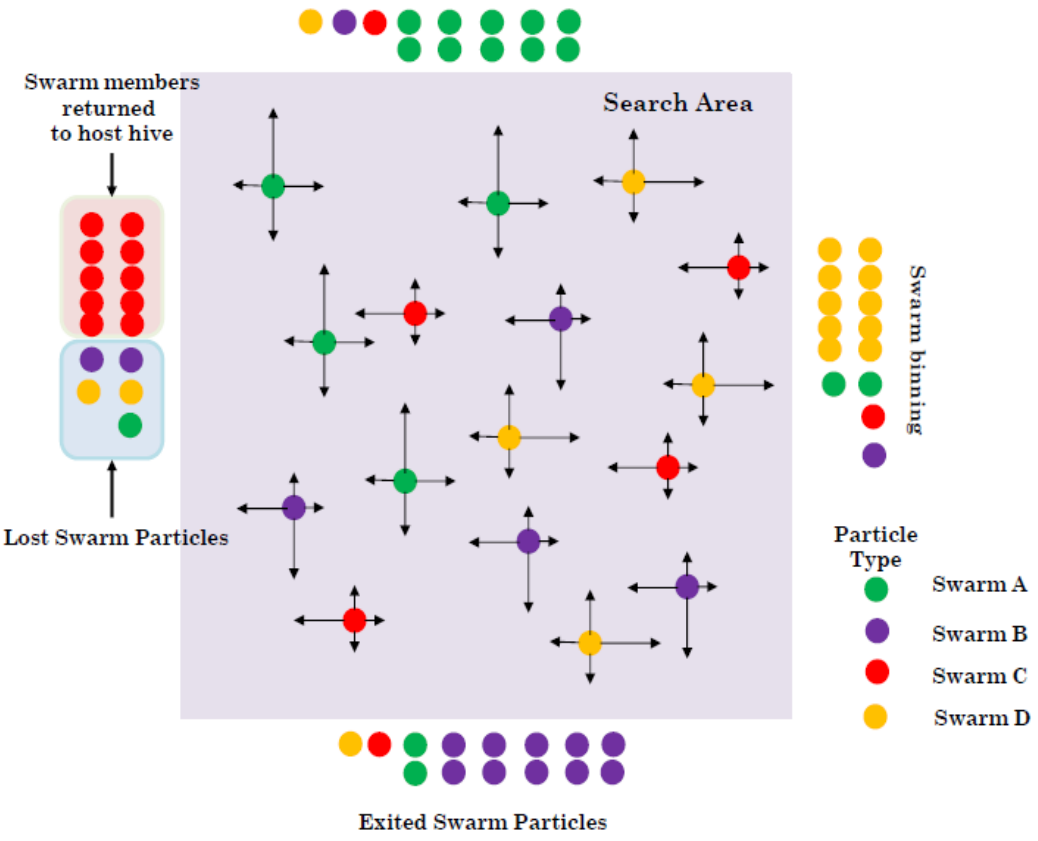


Figure 5: Indicative swarm particles roam and bin for four different swarms on a search area - signifying four different screening effects (A-D).

2.6 Computational platform

The nine emulated recipes of the $L_9(3^4)$ OA along with their respective effective-slope estimations (Appendix A) are loaded on a worksheet for data-processing on the '*Intelligent Problem Solver*' (IPS) module in accord to the topology which is mapped in Figure 4. The IPS module is accessed from the submenu '*Neural Networks*' of the professional software Statistica 7.0 (StatSoft). We select the '*GRNN*' network-type option. The maximum limit of tested networks is set at 10,000. The preferred criterion for retaining networks is: 'Balance error against diversity'. Due to the smallness and messiness of the OA-dataset, we reinforce the training subset size by dedicating to it as many as seven output observations. Thus, we allocate each one of the remaining two possible output entries for subsetting selection and testing, respectively. The randomization process for assigning subset elements is performed once at the onset of each learning cycle. The IPS output delivers the sensitivities of the four non-linear controlling factors for each individual IPS run. The extracted information is recorded in terms of input network-error ratio values which are also complemented by their rank-ordered representation. However, the coherent post-processing of the sensitivity information in the smart sample is expedited by acquiring only the rank performances of the effects.

Smart sample adequacy is checked using the module '*Power and Sample Size for 1-sample t-test*' which is selected through the *STAT* menu of the statistical software package MINITAB 17.0.

The effective-slope OA-dataset is subjected to a preliminary profiling with the ordinary main-effects graph, the half-normal plot and the box-plot. An auxiliary normality screening is also conducted.

For illustrational purposes, ANOVA and GLM results are appended to exemplify the conversion disability which is experienced by implicating standard multi-factorial treatments in conjunction with saturated-unreplicated OA-schemes. Furthermore, the test of equal variances (Levene's test) is performed to confirm if the homoscedasticity assumption holds for the ANOVA comparisons. All graphical portrayals for the above screenings and their respective data analysis including the linear and curvilinear regression fitting of the original TRC dataset have been greatly facilitated by the processing capabilities of the software package MINITAB 17.0.

Nonparametric analysis of the smart sample returns the estimation of the median and its associated confidence interval for each of the profiled inputs individually. Columned data representing the performance log for each effect are loaded in a MINITAB 17.0 worksheet to undergo the one-sample Wilcoxon signed-rank test ('*1-sample Wilcoxon*' module, MINITAB 17.0).

The algorithm 1 was developed in Microsoft Visual Basic 2010 Express.

3. *Results*

3.1 Data pre-screening

A normality pre-screening for each of the nine trials in the original data [31] is shown in Figure 5. It is evident that the groups of the trial runs associated with plots #1-6 do not conform to a normal-distribution fit at least at a confidence level of 0.05. This means two-thirds of the total dataset comply with obscure reference laws which depart significantly from the standard Gaussian model. The messiness of the datasets intensifies the search as three out of the nine trials exhibit outliers (trials # 1, 4, 5), i.e. points digressing on or outside of their group's confidence interval. Furthermore, in Figure 6, we portray an evaluation of replicate reproducibility which is discerned for the three noise settings. It becomes apparent through a simple linear regression that the slope of the fitted duplicates is deviant from unity in all three plots. The observed departures are deemed substantial. The slope magnitude fluctuates bilaterally of the fitted lines thus accumulating evidence that fuels even more the data messiness aspect. For example, the slope estimate elevates higher than 15% above the unity baseline on one end (Fig

ure 6C) while it drops lower than 25% on another instance (Figure 6A). In all three plots, we witness the presence of one or more outliers when contrasted against the estimated 95% confidence interval for each fitted line. Moreover, the coefficient of determination dips remarkably low in two of the three fittings – at 69 % - implying a mysterious tampering with the replicate stability.

3.2 Compressing replicated data with regression fitting

In Appendix B, we list the corresponding (nine) regression fittings for the original (replicated) OA trials (Appendix A). Each fitting relays information of the duplicated TRC observations across the three $KMnO_4$ concentration settings. We immediately notice that the OA-dataset obeys to diversified trends while modelling the various recipes, thus escalating its messy mien. The responses from the nine recipes are best-fitted to a blend of constant, linear and curvilinear curves. The TRC observations as well as the $KMnO_4$ setting-values have been previously transformed in logarithmic units. The least-squares method has been applied to deliver the statistical relationship between them. Trials # 3, 4, 5 and 7 are not influenced by the presence of the nuisance variable ($KMnO_4$) as their slopes are statistically horizontal.

As a matter of simplicity, we select the median estimator to provide the central tendency measure for summarizing the response of those four specific trials (Appendix B). Next, we notice that trials # 2, 8, and 9 are linearly dependent on $KMnO_4$ concentration. The remaining two trials, i.e. # 1 and 6, are closely approximated by retaining the quadratic term. We conclude that the majority of the trials (5/9) are influenced by the noise factor. Thus, the incumbent combo (screening-and-optimization) scheme is called upon to synchronously detect and fine-tune those strong controlling factors against an imperviously "messy" ambience.

We employ the predicted effective slope as the statistical 'compressor-translator' device that embeds disturbance information from the nuisance factor in the studied response. Easily then, we construct the effective slope vector (Appendix A). Trials # 3, 4, 5 and 7 are represented by zero values. Trials # 2, 8 and 9 contribute directly their fitted slope values from the listed equations in Appendix B. However, describing trials # 1 and 6 requires an intermediate step in order to convert their two-term varying behavior into a single-slope indicator. To compute such an "effective" slope, first we search for the min/max values of the fitted quadratic equation as long as such extremes exist in the investigated range for the $KMnO_4$ concentration. For example, for trial #1, we compute that the minimum value of log(TRC) is:

$$\min\{\log(\mathrm{TRC})\} \Rightarrow \frac{\partial \log(\mathrm{TRC})}{\partial \log(\mathrm{KMnO_4})} = 0 = 16.11 + 8.156 \cdot \log(\mathrm{KMnO_4}) \Rightarrow \log(\mathrm{KMnO_4}) = -1.975 \text{ and } \min\{\log(\mathrm{TRC})\} = -0.731$$

All it is needed now is to select the steeper of the dual two-point slopes. They are simply formed between the minimum value we just found in connection to either of the two boundary points. It turns out that the sharper slope is computed to a magnitude of 3.47 with respect to the noise upper boundary ($\log(KMnO_4)$ = - 1.125 and log(TRC) = 2.218). Similarly, for trial #6, the calculated magnitude of the slope is 1.310 (Appendix A).

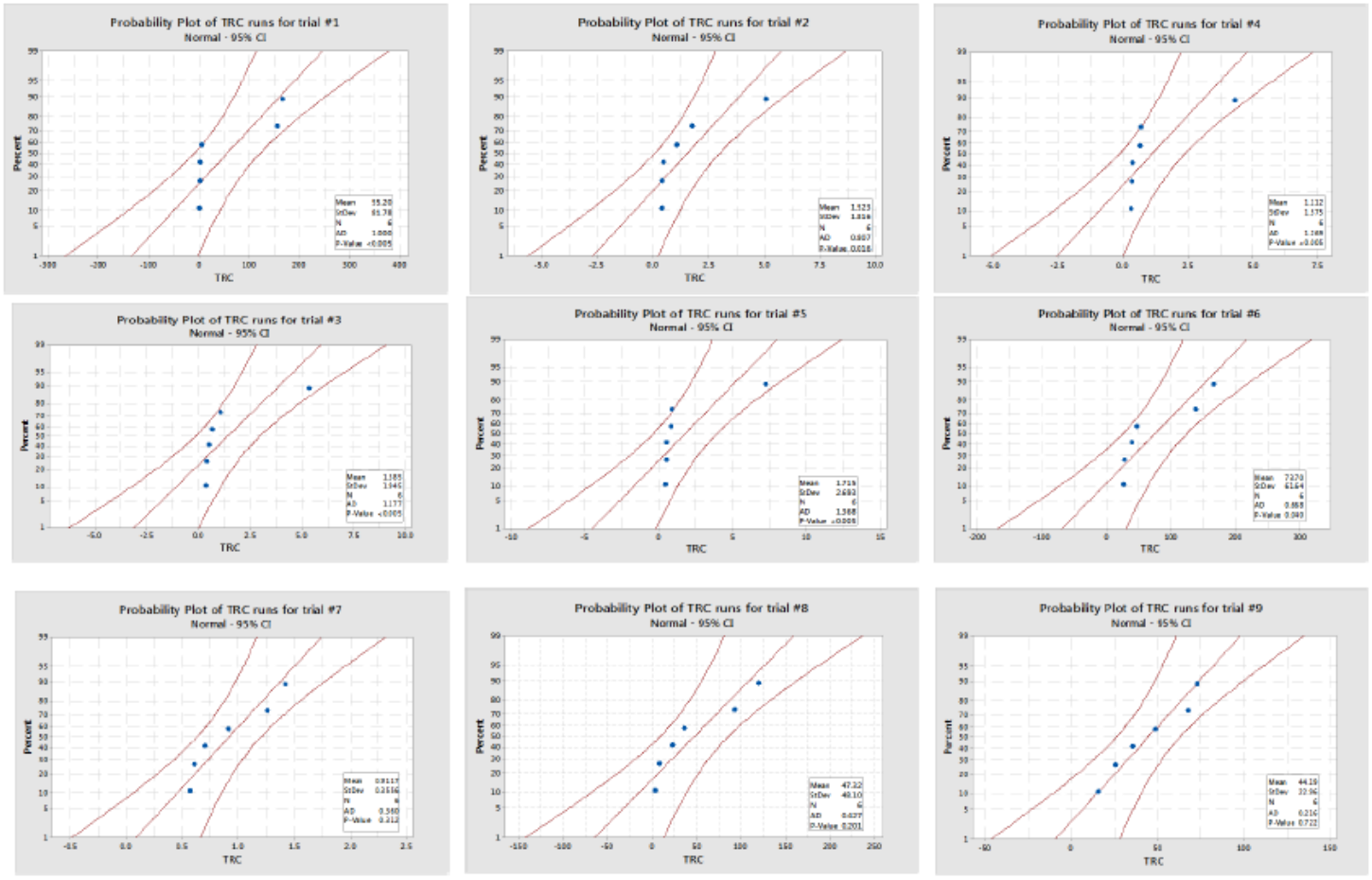
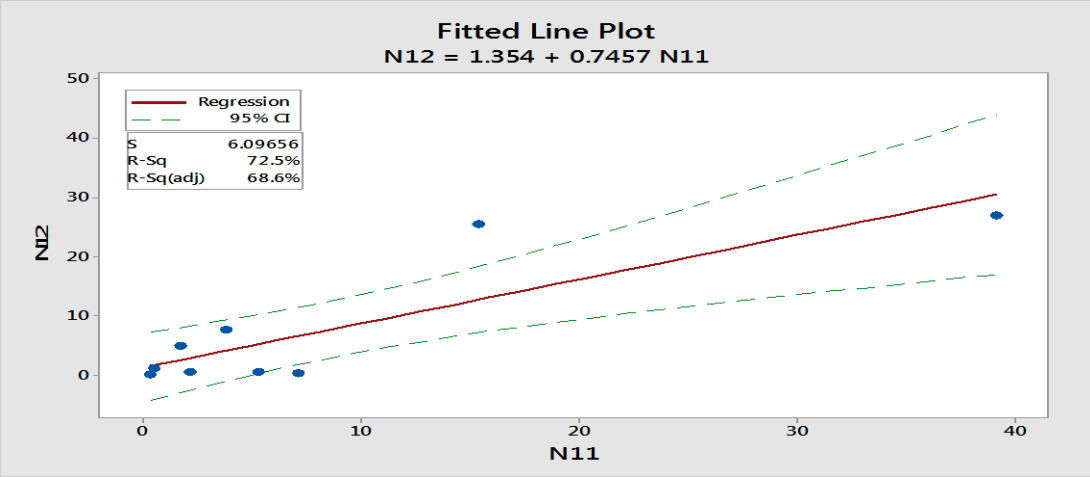
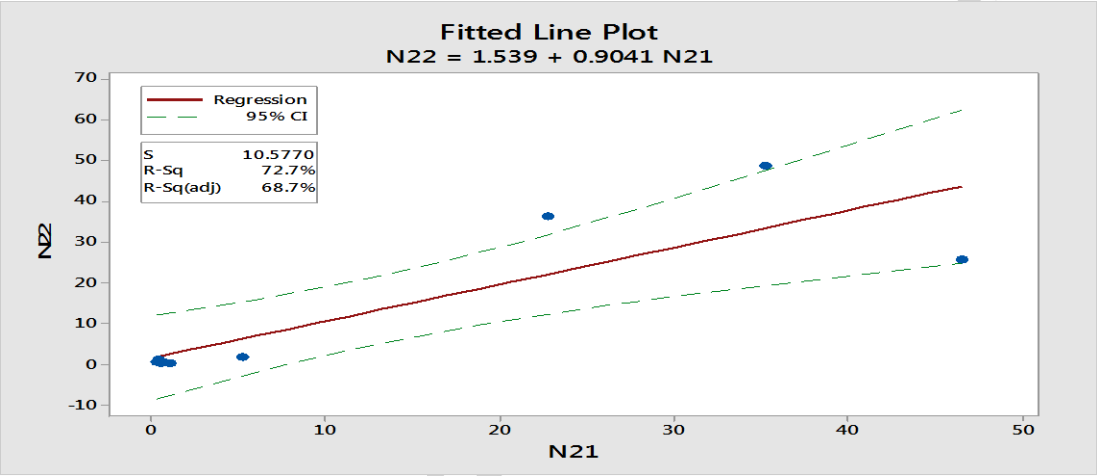


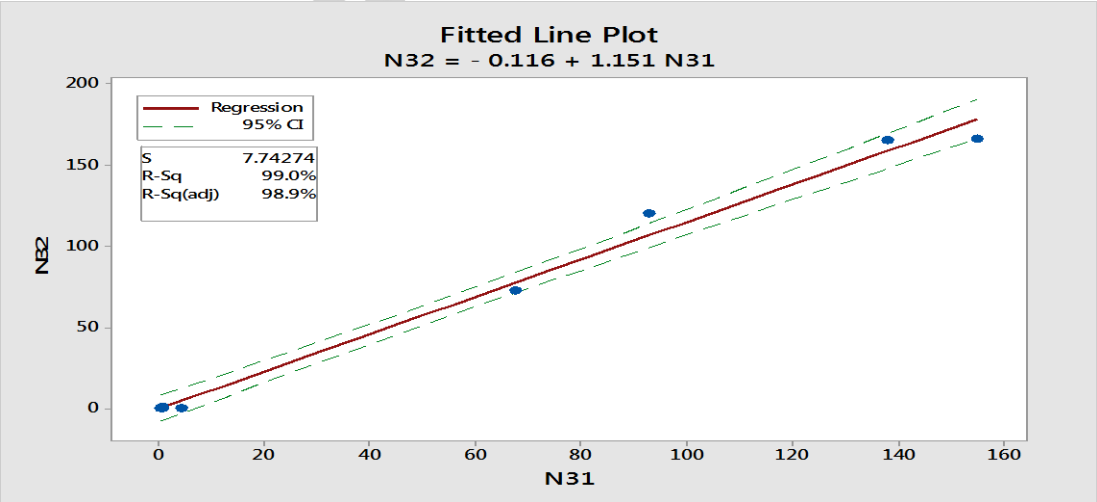
Figure 5: Data-normality prescreening of the original trials (Appendix A).



A.



B.



C.

Figure 6: Consistency checks for the original duplicated data (Appendix A).

3.3 Concurrent screening and optimization

The duplicated data for each of the three $KMnO_4$ concentration settings have been reduced to merely a single vector that relays now solely the effective-slope information. The disturbance which is interjected simultaneously by the four controlling factors and the controllable noise factor has also been engrained in the effective-slope expression. The initial $L_9(3^4)$ OA-dataset (Appendix A) has been cast into a non-linear saturated-unreplicated form where the output role is replaced by the 'slopes' vector. Variance homogeneity cannot be verified at this stage due to the saturated-unreplicated condition imposed on the OA [8, 37]. Thus, the input-output OA-data arrangement is transferred to the GRNN converter for intelligent processing.

The rankings of the controlling effects as collected from the GRNN sensitivity-analysis report have been tabulated for thirty independent runs in Appendix C. Before assessing the effect potency from the information in the rank data-log, we verify first the adequacy of the collected smart sample. Out of the four controlling factors, the maximum (rank) standard deviation is identified to the T-factor since it possesses a value of 0.973. Hence, the predicted smart-sample size for a 95%-confidence-interval is computed to be 18 when the margin of error is set at 0.5. Thereby, the start-up smart-sample size of 30 is judged as sufficient. It is remarked that even when sensitizing the method's detectability to the more stringent margin of error of 0.4, the predicted smart-sample size does not exceed a value of 24 (95% confidence interval).

The estimated confidence intervals for the rank performance of the four controlling factors (Appendix C) have been listed in Table 2. It stands out that the F-factor spawns a disturbance that is statistically superior with respect to the rest of the group. In spite of the complex interference of the nuisance effect on TRC readings, the nonparametric diagnostics declare F-factor as the predominant controlling effect since it occupies solely the top-rank position. Indeed, the upper bound of the 95% confidence interval for the F-effect is confined to a rank prediction that never exceeds a value of 1.5. On the other hand, the 95% confidence intervals for factors T, P and H overlap in extended ranges that interlock lower and upper bounds of ranks 2 and 4, respectively. As a result, the contribution of those three effects may not be resolved with regards to modulating the TRC behavior. Thus, all three effects are filtered out and eliminated from further consideration.

The main effects plot for the effective slopes (Figure 7A) demonstrates the monotonous behavior of the slope behavior owing to the F-factor. The optimal setting is easily identified at a Fe(II) ratio value of 15. Nevertheless, it may be argued that the effect of the Fe(II) ratio set at either levels of 7 or 15 may be statistically indistinguishable, thus urging for a robust contrast among the three settings instead. In Figure 8, we display comprehensively the respective box plots of log(TRC) at the three Fe(II)-ratio settings. We observe that at the endpoint settings - 2 and 15 - the response is significantly disparate. However, the response which is elicited by the middle Fe(II)-ratio setting (7) does not differentiate itself from the responses due to either endpoint settings. This suggests that the Fe(II) ratio may be adjusted equitably to either response-minimizing values (7 or 15). The final selection between the two settings is then decided on grounds of practicality and cost.

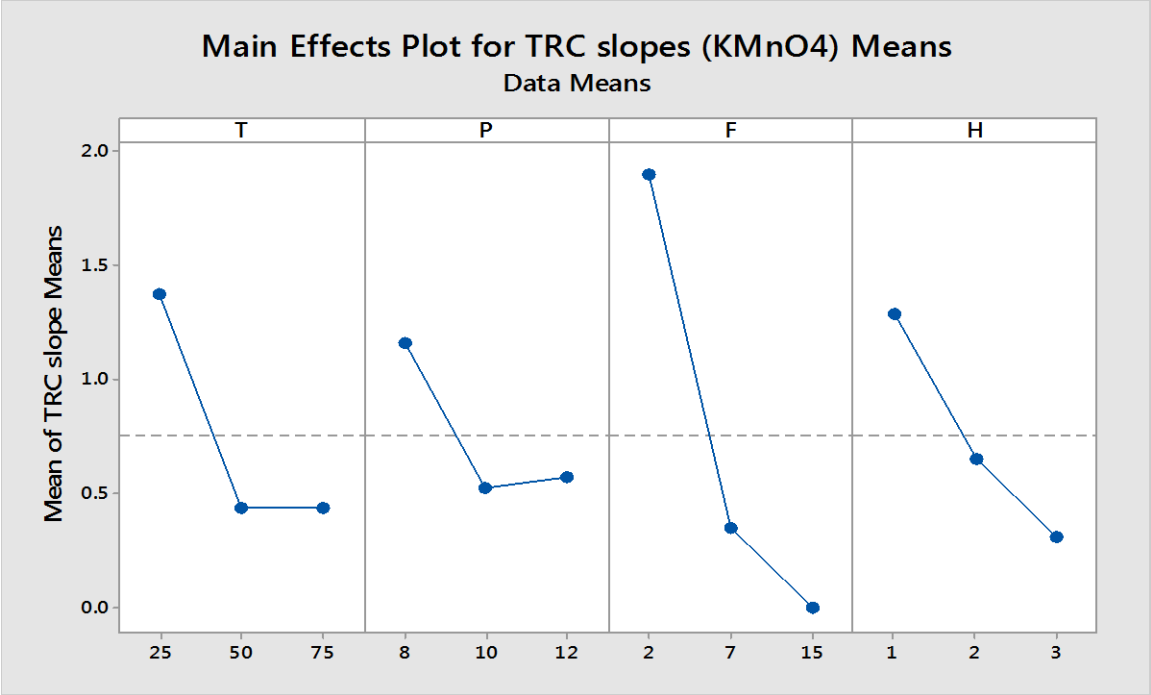
It is interesting to note that the boxplot depiction of the log(TRC) behavior versus the $KMnO_4$ concentration reveals no overall median differences (Figure 9). This finding is also confirmed by the Kruskal-Wallis test which detects no significant difference among the three settings ($p=0.377$). Nevertheless, it is observed that the variability of the log(TRC) increases as the $KMnO_4$ concentration

increases (Figure 9). In Figure 10, we refine the grouped boxplots by arranging for partial depiction of the behavior of the log(TRC) against the Fe(II) ratio and the $KMnO_4$ concentration. Now, we may draw more solid conclusions by noting that the central tendency and the dispersion of the log(TRC) is minimized simultaneously when the Fe(II) ratio is locked at the setting value of 15. The best response result is favored when the injected noise - $KMnO_4$ concentration - is sustained to values of 0.0375 mol/l or higher.

Table 2: Confidence intervals for smart-sample effect rankings (Appendix C) using the Wilcoxon signed-rank test.

				Confidence Interval	
Factor	n	Estimated Median	Achieved Confidence	Lower	Upper
T	30	2.5	94.9	2	3
P	30	3.5	94.9	3	4
F	30	1.5	94.9	1	1.5
H	30	3.0	94.9	2.5	3

A. Main effects graph for the TRC effective slopes



B. Analysis of Variance for effective slope means

Source	DF	Seq SS	Adj SS	Adj MS	F	P
T	2	1.7522	1.75219	0.87610	*	*
P	2	0.7498	0.74982	0.37491	*	*
F	2	6.1336	6.13360	3.06680	*	*
H	2	1.4899	1.48992	0.74496	*	*
Residual Error	0	*	*	*		
Total	8	10.1255				

Figure 7: Multifactorial comparison of the effective slope data (Appendix A): A) Main effects graph and B) ANOVA output (MINITAB 17).

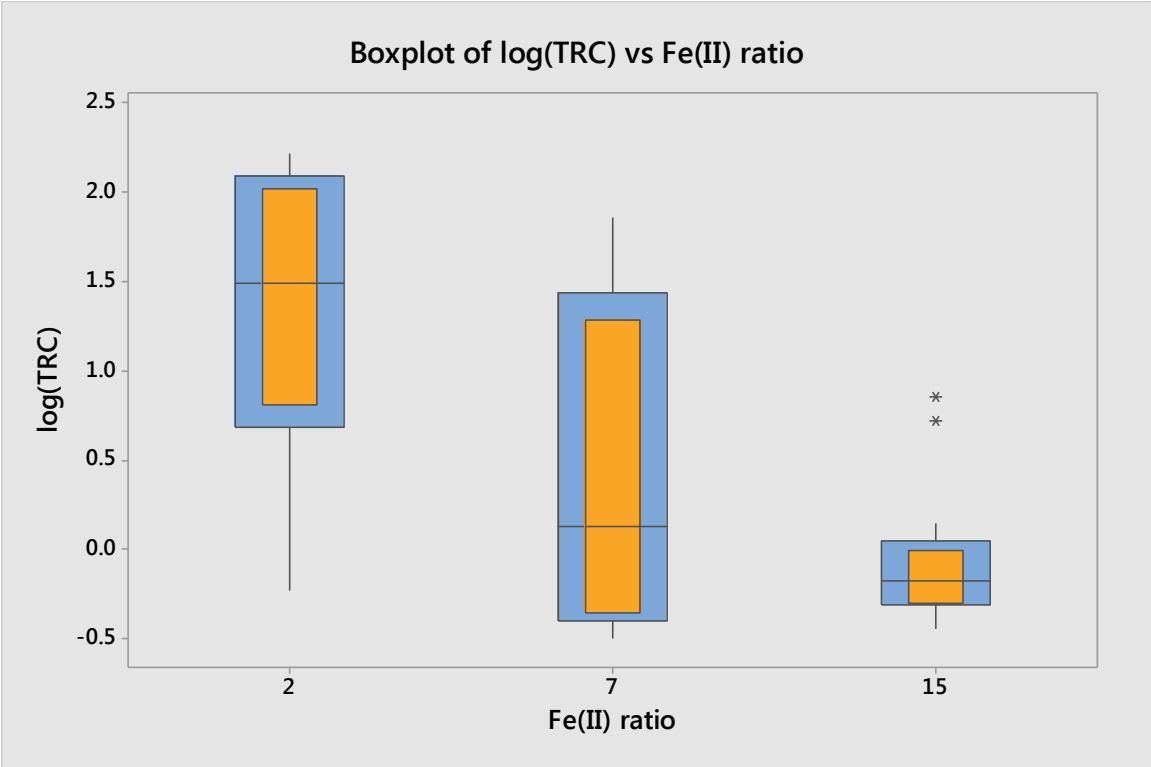


Figure 8: Box plot with median confidence intervals for log(TRC) versus Fe(II) ratio (MINITAB 17.0).

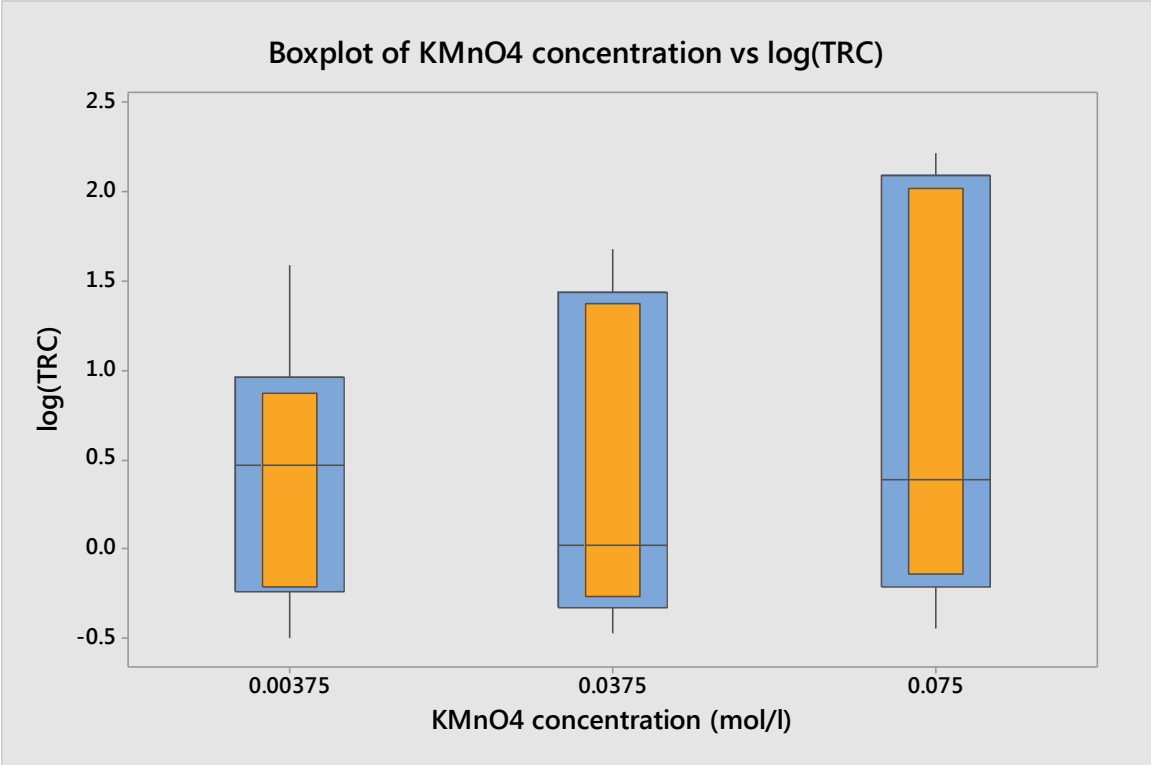


Figure 9: Box plot with median confidence intervals for log(TRC) versus $KMnO_4$ concentration (MINITAB 17.0).

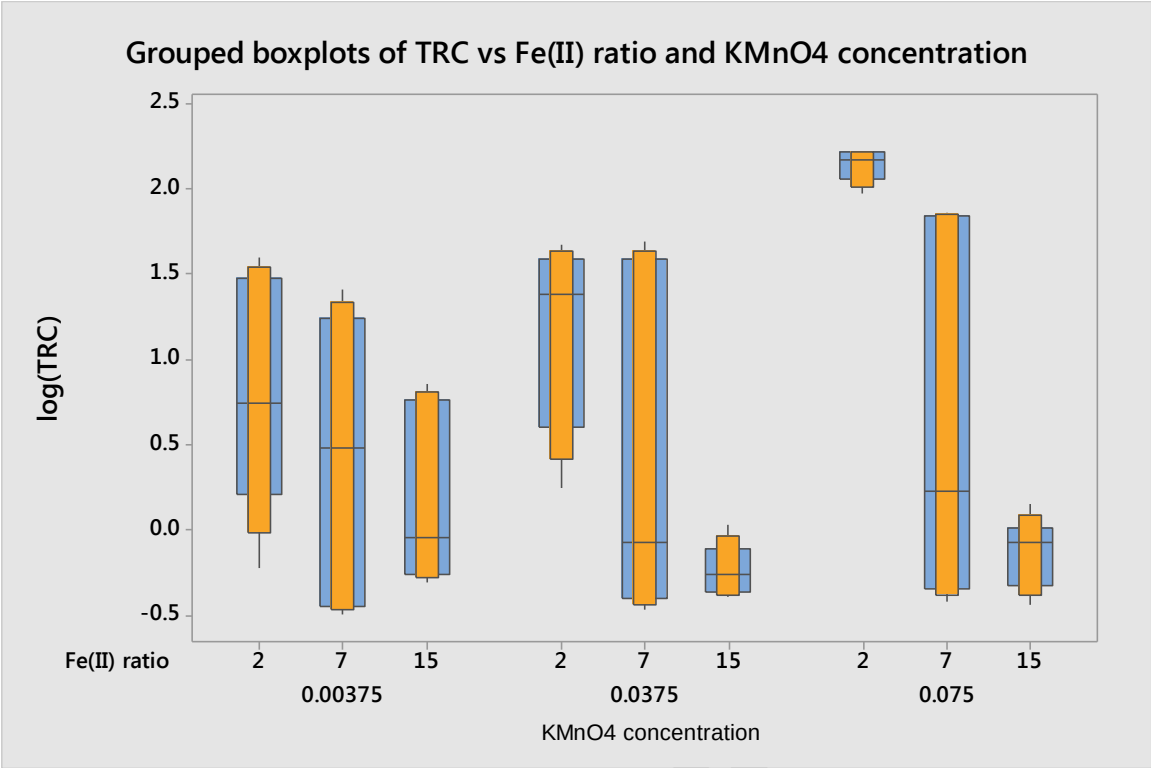


Figure 10: Grouped box plots with median confidence intervals for log(TRC) versus $KMnO_4$ concentration and Fe(II) ratio (MINITAB 17.0).

3.4 Screening-effect test power using "dizzy" swarm intelligence

In Table 3, we initialize the α_{sj} values for the four swarms, T, P, F and H. These values are simply the frequencies of their four rank values which are turned into fractions as computed from the smart sample of Appendix C. In Table 4, we tabulate the test power has been computed for a swarm size of 1000 (A_r = 1,000,000) where all four groups where maintained equal (N_T=N_P=N_F=N_H). The selected tested swarm group sizes were: 100, 1000 and 10000. Trying increasing sizes aided in understanding power convergence tendencies and computational performance. At a swarm group size of 10000, power prediction variance is tiny but code performance runs up to several seconds. On the other hand, for a swarm group size of 100, the code executes under a second but there is substantial prediction variance. For reporting, we found that 1000 particles offered acceptable prediction variance while maintaining execution speed under a second. Notably, the median power tendency for all three tested sizes is comparable.

The median estimation of power for all four effects has been established at an achievable confidence of 95% (Table 4). We conclude that that only for the F-effect the median power (0.73) is near but under the threshold for strong evidence. The rest of the three effects are confined in a range of values of 0.53 to 0.60 declaring a remote chance for be deemed serious influencers. Thus, the power evaluation has been materialized because of the creation of competing solution probabilities that have been assessed through a new pseudo-distribution algorithm [46].

Table 3: Inputs of the α_{si} fractions for the four swarm groups – screening effects: T, P, F and H.

	T		P		F		H	
Effect Rank (i)	Frequency	Fraction α_{Ti}	Frequency	Fraction α_{Pi}	Frequency	Fraction α_{Fi}	Frequency	Fraction α_{Hi}
1	3	0.10	2	0.07	21	0.70	4	0.13
2	15	0.50	3	0.10	6	0.20	6	0.20
3	5	0.17	8	0.27	3	0.10	14	0.47
4	7	0.23	17	0.57	0	0.00	6	0.20

Table 4: Test power for the four swarm groups (N_T=N_P=N_F=N_H= 1000).

Swarm Group	Median P	Confidence Interval	Achieved Confidence
T	0.55	0.54-0.56	95%
P	0.60	0.59-0.61	95%
F	0.73	0.72-0.75	95%
H	0.53	0.52-0.54	95%

4. Discussion

The TRC-response optimization was non-linearly programmed and analyzed in the past using a four-factor (saturated) $L_9(3^4)$ OA experimental design with the aim to ameliorate water quality [31]. Experimental provisions demanded duplicated trial-runs. Simultaneously, it was investigated the possible non-linear influence of a highly-critical redox trigger ([$KMnO_4$]) posting as a controllable noise factor. The noise factor was arranged in the special outer-array configuration as prescribed by the advanced Taguchi methods. Overall, the published TRC-minimization study was a typical inner/outer array application of the Taguchi methods using the signal-to-noise ratio concept to compress multiple replicates. The SNR transformation reduced a series of duplicated runs which were conducted in increasing noise-setting triads down to a single-column dataset. The conversion compacted information from the initial multi-column data arrangement to a saturated-unreplicated OA form. However, the variance homogeneity assumption which is expected to hold for securing the reliable performance of the ANOVA solver cannot even be checked on SNR-transformed data for saturated designs. This is due to the SNR-compression process which caused the dramatic draining of the initial pool of degrees of freedom in association with the conducted trial-runs (Figure 11B). Therefore, any subsequent use of ANOVA should be accompanied with precautions for potentially leading to subjective decision-making as its use at this stage is considered only of qualitative worth. Before utilizing the main-effects graph and the ANOVA output to gauge effect potencies, the compression of the replicates was instructed to be carried-out for the apt handling of the induced as well as other remnant (ambient) noise [14, 15]. We recreate the corresponding main-effects graph (Figure 11A) which is further supplemented with an ANOVA-treatment listing (Figure 11B). From the main effects graph, we merely deduce that the F-factor may head the efficacy hierarchy. The F-factor exhibits an increasingly monotonous trend forged in a rather curvilinear shape. The statistical significance of those findings is not readily interpretable by making the standalone usage of the main-effects graph. Similarly, by focusing on the ANOVA output, the largest value of the means of sums of squares is associated with the F-factor. The F-factor packs up more than a six-fold

variation when compared to the next less potent effect – the H-factor. However, the experimental uncertainty remains elusive as all emerging variation has been channeled exclusively through the ANOVA engine to the tested effects. Hence, no inference is possible at this stage which awards no statistical significance to the contrasts. Ostensibly, the analysis is rendered inconclusive due to the known incompatibility of non-linear unreplicated-saturated OAs with ordinary multi-variable converters [9, 30, 37]. A practical (subjective) outlet to arrive to an approximate solution is to follow Taguchi's recommendation to pool and disaffect some of the weaker effects. By fostering the notion that the residual error possesses an indistinguishable stochastic fingerprint in relation to the inert effects, this trick enables ANOVA to switch on again. For this to happen, the "pooled-ANOVA" approach needs to warrant a priori that some of the examined effects must be stochastically imperceptible. In other words, there will always be resolvable to identify and separate stronger from weaker effects since both always coexist. This is the so-called sparsity assumption. On the contrary, our intelligence-based approach using smart sampling at its core ignores this chimerical limitation.

TRC is a characteristic that is sought to be minimized. By definition, Taguchi's SNR-expression for a 'smaller-is-better' response is [14, 15]:

$$SNR = -10 \cdot \log_{10}\left(\frac{\sum_{i=1}^{n} y_i^2}{n}\right) \tag{10}$$

For equation 10 to be objectively functional and the compression capability of the SNR to be valid, the sum of squares has to reflect a characteristic with its y_i-observations to follow a normal distribution. This is because:

$$E(y^2) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = Var(y) + E^2(y) \tag{11}$$

Hence, the expectation of the squared response, $E(y^2)$, is a robust estimator when the variance and the expectation of the response - $Var(y)$ and $E(y)$ - are both robustly estimated. This holds true strictly for data obeying the normal distribution. As we demonstrated in the '*Results*' section, the majority of the TRC data points does not follow any familiar reference law. This is a major weakness which undermines the accuracy of the SNR estimator. It is an alarming situation to attempt to surmise on the effect tendencies under enormous uncertainty, which is dissipated in so many different modes as witnessed in this work. It is this aspect that invites and thus justifies the double checking of the original TRC optimization results with an alternative method. Henceforth, we realize the imperative necessity for building a new hybrid technique that is relieved of the numerous burdening (functional) assumptions, we discussed earlier, by introducing soft computing mentality and distribution-free data manipulation at its nexus.

Another intriguing fact may be gleaned from Table 5 where we sequentially pool the arising weak effects as suggested by the Taguchi methods. The T-factor incites the least variability in the SNR-measure (Figure 11B). Suppose we lend the T-factor variance in order to animate a germinal yet superficial – "seeding" – estimate for the residual error. Then, the adjusted mean of the sum of squares for the residual error becomes 28.26 (dB^2) granting an F-test p-value for the F-factor of 0.032 which is deemed as significant at a level of 0.05. By eliminating the T-factor from the starting group, the

subsequent regression fitting returns a satisfactory coefficient of determination at 90.10%. However, it is instantly revealed that P- and H-factors are statistically insignificant even at the cruder level of 0.1. The Taguchi methodology usually recommends the elimination of as many as 50% of the effects by pooling them out to form the potent group. Hence, the next weaker is the P-effect. By pooling P-effect too, the adjusted mean of the sum of squares for the cumulative residual error climbs to 72.84. Now, the F-effect p-value becomes more significant at a value of 0.21, but the coefficient of determination drops to a questionable rate of 74.4%. At this point, there is no reason to retain the last of the weak effects - the H-factor. By completing an exhausting three-way pooling, the p-value for the F-factor descends further at 0.016 while the coefficient of determination dips to 66.60%. This means that 1/3 of the total variability is still unexplainable which would prompt for a broader discovery cycle. What is particularly puzzling is that while the F-effect sharpens its role in the profiling performance, two non-significant variables, P and H, cause a more that 35% drop in the fitting efficiency of the terminal model and this may be viewed perhaps as self-contradictory. This oddity is likely rooted deeply in the messiness of the data. It also presents another opportunity to stress still from another angle the usefulness of revisiting the original TCR experiments.

Table 5: Relationship of pooled effects on F-factor p-value and coefficient of determination (R^2).

	Residual Error	F-factor		
Pooled Effects	Adj MS	Adj MS	p-value	R^2(Adj)
T	28.26	852.68	0.032	90.10%
T,P	72.84	852.68	0.021	74.40%
T,P,H	95.14	852.68	0.016	66.60%

In spite of the novelty of our method and its promising capabilities, its trustworthiness might be established only upon its applicability to more diverse circumstances. Unfortunately, owing to the spotlighted idiosyncrasies of the revisited theoretical problem, no alternative methods are really available to analyze it by comparison. The only method that is known to handle a non-linear saturated-unreplicated OA-dataset scheme is the amended Lenth method [47]. Clearly, the corrected Lenth method might be only applicable as long as a valid multi-response data fusion has been antecedently fulfilled. The corresponding outcomes from the improved Lenth-method version are listed in Table 6 where the pseudo-standard-error (PSE) has been estimated at 2.040. We observe that the t-statistic of the linear component of the F-factor (F_1 = 1.975) may be significant if the individual error rate (IER) is maintained at 1.671 (p=0.1) and the experimentwise error rate (EER) is respectively set at 1.839 (p=0.4). In comparison, our technique is more competitive since it offers prediction at a significance level of 0.05. This is because the Lenth test when applied for profiling non-linear effects coerces the virtual doubling-up of the studied contrasts. This aspect causes stochastic dilution on the quantification accuracy of each effect's strength magnitude. Thus, the dichotomized effects are gauged against a broadened group that acknowledges the presence of their identities twice for each its contrasted members. Moreover, Lenth statistics introduce two arbitrary yet rigid constants for truncating a standard error while disregarding any information

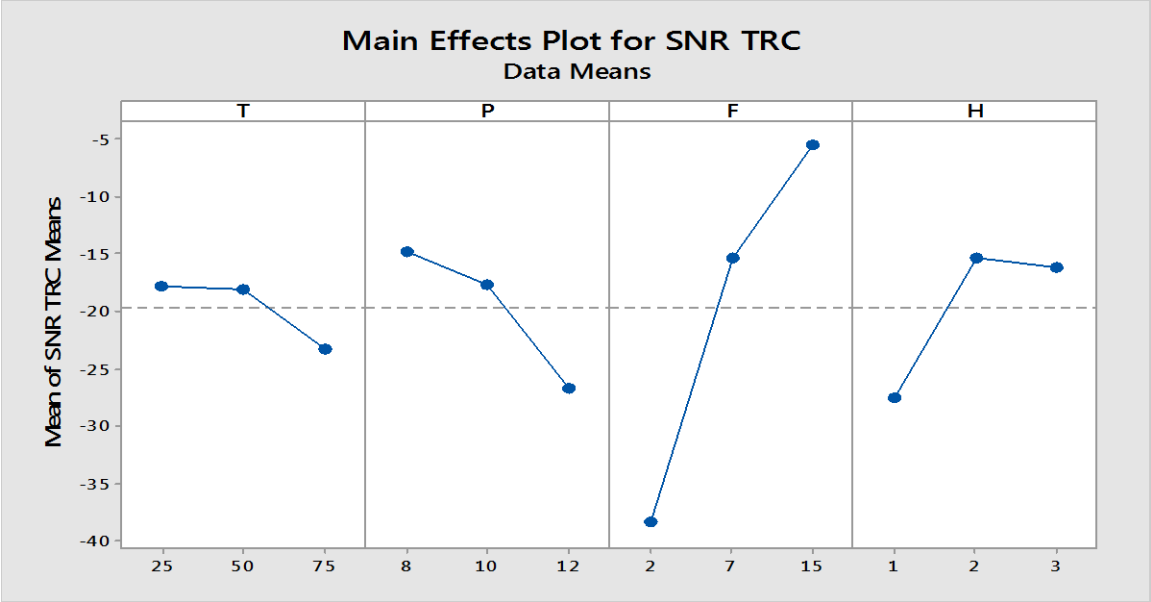
about the shape of the contrasts. Thus, the assumption that the PSE is invariant for all effects to permit the normalization of the contrasts might be situationally weak. Hopefully, this study has shed more light in handling modern a-qualimetrics. It may aid in robustly profiling water quality performance against great uncertainty and messy data. Finally, the notion of smart sampling that was delineated in this study offers a new option for ramping up the pace in cases where perplexing parameter identifiability is expected.

Table 6: Corrected Lenth test linear and quadratic terms.

Factor/term	$\hat{\theta}$	t-statistics
T_l	-1.984	0.973
P_l	-1.247	0.612
F_l	-4.029	1.975
H_l	-2.085	1.022
T_q	-1.149	0.563
P_q	-0.833	0.408
F_q	-1.472	0.722
H_q	-0.353	0.173

From the above developments, we may now encapsulate the benefits and the shortcomings of the new GRNN/swarm approach. Its main field of application is to decipher complex input-output relationships for nondeterministic datasets. Key features and restrictions for selecting the implementation of the GRNN/swarm approach are:

1) Structured (input-output) multi-factor datasets
2) Non-linear Taguchi-type OA sampling
3) Constrained stochastic multi-factor screening
4) Constrained stochastic multi-factor optimization
5) No available deterministic function to describe the investigated phenomena
6) Uncertainty creeps in the relationship between controlling factors and response
7) Uncertainty may be dealt with Taguchi's inner/outer array form
8) Effect saturation allowed (maximum effect exploitation)
9) Virtual response unreplication due to Taguchi's SNR transformation is dealt
10) Sparsity assumption waived
11) Homoscedasticity assumption waived
12) Normality assumption waived
13) Messy-data friendly
14) Distribution-free data reduction
15) Intelligent uncertainty wrap-up through GRNN runs
16) Predictions are based on quick smart samples
17) Multiple effect contribution is verified by swarms
18) Swarms outmaneuver dataset uncertainty through intelligence forwarded from GRNN



B. Analysis of Variance for SNR-TRC Means

Source	DF	Seq SS	Adj SS	Adj MS	F	P
T	2	56.52	56.52	28.261	*	*
P	2	234.86	234.86	117.428	*	*
F	2	1705.37	1705.37	852.685	*	*
H	2	279.46	279.46	139.732	*	*
Residual Error	0	*	*	*		
Total	8	2276.21				

Figure 11: A) Main effects graph for SNR transformed data (Appendix A), and B) ANOVA output for SNR transformed TRC-data (MINITAB 17).

5. *Conclusions*

Water treatment processes are complex by nature and thus collected data on improving water quality are avidly messy. Attempting concurrent screening-and-optimization water-quality studies only exacerbates the task of reaching to a trustworthy solution. Water-quality indicators are bewildering to profile because qualimetrical predictions stumble upon intricate stochastics which are deterred by the presence of prevalent non-normal and non-linear phenomena. While accentuating the inherent messiness in dealing with Taguchi-type a-qualimetrics, by revisiting the pioneering work by Barrado et al. [31], a new intelligent approach was proposed to facilitate the robust interpretation of the published dataset. Evidence of data messiness stemmed from a strong intermixing of normal and non-normal phenomena that regulated the water quality status along with a baffling model structure which was varying unpredictably when fitting the imposed noise on different OA recipes. We assembled smart samples generated by the intelligent conversion of noise-tweaked effective-slope data using the GRNN engine. Smart samples simply accumulated all messiness to a new transformed dataset overcoming the need to account for its many sources. Analyzing nonparametrically the smart sample, it translated to a robust stochastic rating which was assigned to the investigated effects. Furthermore, the sychronous defuzzification and distribution-free data reduction tactic added convenience and agility in the data processing stage. Consequently, the pinpointing of the optimal settings for the dominant controlling

factors was greatly enlightened. The new idea of introducing the "dizzy" swarm intelligence based on the smart samples allowed power estimation for the multifactorial test. It is noteworthy that our version of predictions has been obtained by withdrawing several assumptions often imposed on other competing multifactorial solvers in conjuction with inner/outer OA applications – sparsity assumption, normality assumption and heteroscedasticity.

Our predictions are concordant to the predictions published by Barrado et al. [31] using classical Taguchi analysis of experiments. The predominant factor, the ratio of Fe(II) in total metal content is optimized at the setting value of 15. However, though the "dizzy" swarm intelligence method the power performance was not found to meet the threshold for a strong performance. Accordingly, the outer-array nuisance variable ([$KMnO_4$]) should be maintained at concentration levels higher than $3.75 \cdot 10^{-2}$ mol/l to boost ferrite removal. Furthermore, we showed that a lot of the assumptions which were required for the Taguchi methods to be applicable eventually could not be verifiable. Sometimes the assumptions do not seem to hold, i.e. data normality, and in many instances the assumptions could not even be possible to be tested, i.e. variance homogeneity. Swarm itelligence offered a practical way to estimate a "soft" power measure for the inner/outer OA optimization case. Such power prediction could not be inferred alone from the classical statistical treatment resulting from the SNR-ANOVA calculations because of the non-linear effect saturation.

The noise dependence of the fitted response model on particular trial recipes was statistically evaluated through the effective slope. Surfacing modeling oddities and data anomalies which innately lurk in deploying mainstream screening/optimization routines in complex water solutions were demonstrated to be more efficiently outmanouvered with NN-DSI-adapted and robustly-resolved processing.

Appendix A: Original data from Barrado et al. [31] and converted effective-slopes (Appendix B).

					TRC							
Trial#	T	P	F	H	Ml1	N12	X21	N22	N31	N32	5NRA1	slopes
1	25	8	2	1	2.24	0.59	5.29	1.75	155.34	166.27	-39.36	3.470
2	25	10	7	2	1.75	5.07	1.05	0.41	0.3S	0.4S	-7.05	0.650
3	25	12	15	3	5.32	0.65	0.4	1.07	0.51	0.36	-7.05	0.000
4	50	S	7	3	0.37	0.32	0.34	0.6S	4.31	0.65	-5.19	0.000
0	50	10	15	1	7.2	3.49	0.4©	0.44	O.S	o.ss	-9.54	0.000
6	50	12	2	2	39.17	27.05	46.54	25.77	138.0S	165.61	-39.34	1.310
'	75	8	15	2	0.57	1.26	0.61	3.7	0 91	1 42	0.28	0.000
3	75	10	2	3	3.88	7.85	22.74	36 33	92.S	120.33	-36.20	0.913
9	75	12	7	1	15.42	25.52	35.27	4S.61	67.56	72.73	-33.79	0.396

Appendix B: Regression-fitted replicated TRC log-data against log($KMnO_4$) (Appendix A).

Trial #1: $\log(TRC) = 15.18 + 16.11 \cdot \log(KMnO_4) + 4.078 \cdot \log^2(KMnO_4)$

Trial #2: $\log(TRC) = -1.105 - 0.65 \cdot \log(KMnO_4)$

Trial #3: $\log(TRC) = -0.892$; Median {TRC} = 0.58

Trial #4: $\log(TRC) = 0.531$; Median {TRC} = 0.51

Trial #5: $\log(TRC) = -0.636$; Median {TRC} = 0.65

Trial #6: $\log(TRC) = 7.161 + 6.243 \cdot \log(KMnO_4) + 1.614 \cdot \log^2(KMnO_4)$

Trial #7: $\log(TRC) = 0.008$; Median {TRC} = 0.81

Trial #8: $\log(TRC) = 2.931 + 0.918 \cdot \log(KMnO_4)$

Trial #9: $\log(TRC) = 2.243 + 0.396 \cdot \log(KMnO_4)$

Appendix C: GRNN output from multiple runs (smart sample): Ranked effects from sensitivity analysis.

Run#	T	P	F	H
1	2	4	1	3
2	2	4	1	3
3	2	4	1	3
4	2	3	1	4
5	3	4	1	2
6	1	4	2	3
7	3	2	1	4
8	4	3	2	1
9	4	3	2	1
10	4	2	3	1
11	1	3	2	4
12	3	4	1	2
13	2	3	1	4
14	2	4	1	3
15	2	4	1	3
16	2	4	1	3
17	2	4	1	3
18	2	3	1	4
19	4	3	1	2
20	2	4	1	3
21	2	4	1	3

22	2	3	1	4
23	3	4	1	2
24	4	1	2	3
25	3	4	1	2
26	2	4	1	3
27	4	2	3	1
28	4	1	2	3
29	1	4	3	2
30	2	4	1	3

References

[1] Y. Shi, Innovations and Developments of Swarm Intelligence Applications, IGI Global, PA, USA, 2012.

[2] B.K. Panigrahi, Y. Shi, M.-H. Lim, Handbook of Swarm Intelligence: Concepts, Principles and Applications, Adaptation, Learning and Optimization Series, Springer, UK, 2011.

[3] S. Dehuri, A.K. Jagadev, M. Panda, Multi-objective Swarm Intelligence: Theoretical Advances and Applications, Studies in Computational Intelligence, Springer, UK, 2015.

[4] X.-S. Yang, Z. Cui, R. Xiao, A.H. Gandomi, M. Karamanoglu, Swarm Intelligence and Bio-inspired Computation: Theory and Applications, Elsevier Insights, UK, 2013.

[5] R.K. Pearson, Mining Imperfect Data: Dealing with contamination and incomplete records, SIAM, New York, NY, 2005.

[6] G.A. Milliken, D.E. Johnson, Analysis of Messy Data Volume I: Designed Experiments, Chapman and Hall/CRC, Boca Raton, FL, 2004.

[7] G.J. Besseris, Profiling effects in industrial data mining by non-parametric methods, European Journal of Operational Research 220 (1) (2013) 147-161.

[8] G.J. Besseris, A distribution-free multi-factorial profiler for harvesting information from high-density screenings, PLOS ONE 8 (2013) e73275.

[9] G.J. Besseris, Multi-response non-parametric profiling using Taguchi's qualimetric engineering and neurocomputing methods: Screening a foaming process in a solar collector assembly, Applied Soft Computing 22 (1) (2014) 222-237.

[10] J. Kennedy, R. Eberhart, Particle Swarm optimization, in: Proceedings of the IEEE International Conference on Neural Networks, 4. IEEE (1995) 1942-1948.

[11] J. Kennedy, R.C. Eberhart, Swarm Intelligence, Morgan-Kaufmann, CA, USA, 2001.

[12] S. Ghosh, A. Konar, An overview of computational intelligence algorithms, Studies in Computational Intelligence 437 (2013) 63-94.

[13] X.-S. Yang, Metaheuristic optimization: Nature-inspired algorithms and applications, Studies in Computational Intelligence 427 (2013) 405-420.

[14] Taguchi, G., Chowdhury, S., Taguchi, S. (2000). *Robust Engineering: Learn How to Boost Quality While Reducing Costs and Time to Market*. McGraw-Hill, New York, NY.

[15] Taguchi, G., Chowdhury, S., Wu, Y. (2004). *Quality Engineering Handbook*. Wiley-Interscience, Hoboken, NJ.

[16] Box, G.E.P., Hunter, J.S., Hunter, W.G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley-Interscience, Hoboken, NJ.

[17] R. Mukerjee, C.F.J. Wu, A Modern Theory of Fractional Design, Springer, UK, 2006.

[18] Barlow, M., (2014). *Blue Future: Protecting Water for People and the Planet Forever*. New Press, New York, NY.

[19] Chartres, C., and Varma, S. (2010). *Out of Water: From Abundance to Scarcity and How to Solve the World's Water Problems*. FT Press, Upper Saddle River, NJ.

[20] Benjamin, M.M., Lawler, D.F. (2013). *Water Quality Engineering: Physical / Chemical Treatment Processes*. Wiley, Hoboken, NJ.

[21] Edzwald, J. (2010). *Water Quality & Treatment: A Handbook on Drinking Water* (Water Resources and Environmental Engineering Series). American Water Works Association. McGraw-Hill Professional, New York, NY.

[22]Eckenfelder, W., Englande, A. (2008). *Industrial Water Quality*. McGraw-Hill Professional, New York, NY.

[23] Madaeni, S.S. and Koocheki, S. (2006). Application of Taguchi method in the optimization of wastewater treatment using spiral-wound reverse osmosis element. Chemical Engineering Journal, 119, 37-44.

[24] Mousavi, S.M. , Kiani, S., Farmad, M.R., Hemati, A., and Abbasi, B. (2012). Extraction of Arsenic(V) from Water Using Emulsion Liquid Membrane. *Journal of Dispersion Science and Technology*, 33 (1), 123-129.

[25] Lochmatter, S., Holliger, C. (2014). Optimization of operation conditions for the startup of aerobic granular sludge reactors biologically removing carbon, nitrogen, and phosphorous. *Water Research*, 59(1), 58-70.

[26] Zirehpour, A., Rahimpour, A., Jahanshahi, M., and Peyravi, M. (2014). Mixed matrix membrane application for olive oil wastewater treatment: Process optimization based on Taguchi design method. *Journal of Environmental Management*, 132, 113-120.

[27] Mohammadi, T., Kazemi, P. (2014). Taguchi optimization approach for phenolic wastewater treatment by vacuum membrane distillation. *Desalination and Water Treatment,* 52(7-9), 1341-1349.

[28] Besseris, G.J. (2012). Eco-design in total environmental quality management: Design for environment in milk-products industry. *The TQM Journal,* 24 (1), 47-58.

[29] G.J. Besseris, Multi-response multi-factorial master ranking in non-linear replicated-saturated DOE for qualimetrics. Chemometrics and Intelligent Laboratory Systems 116 (1) (2012) 47-56.

[30] G.J. Besseris, A fast-and-robust profiler for improving polymerase chain reaction diagnostics. *PLOS ONE*, 9 (2014) e108973, 1-13.

[31] E. Barrado, M. Vega, R. Pardo, P. Grande, J.L. Del Valle, Optimization of a purification method for metal-containing wastewater by use of a Taguchi experimental design, Water Research, 30 (10) (1996) 2300-2314.

[32]G.A. Milliken, D.E. Johnson, Analysis of Messy Data, Volume II: Nonreplicated Experiments. Chapman and Hall/CRC, Boca Raton, FL, 1989.

[33] R. Mead, S.G. Gilmour, A. Mead, Statistical Principles for the Design of Experiments: Applications to Real Experiments, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK, 2012.

[34] R.W. Crabtree, I.D. Cluckie, C.F. Forster, Percentile estimation for water quality data, Water Research, 21(5) (1987) 583-590.

[35] S. Dehuri, S. Ghosh, S.-B. Cho, Integration of Swarm Intelligence and Artificial Neural Network, Machine Perception and Artificial Intelligence, World Scientific Publishing Company, Singapore, 2011.

[36] S. Ding, H. Li, C. Su, J. Yu, F. Jin, Evolutionary artificial neural networks: A review, Artificial Intelligence Review 39 (2013) 251-260.

[37] G.J. Besseris, Concurrent multi-response non-linear screening: Robust profiling of webpage performance, European Journal of Operational Research 241 (2015) 161-176.

[38] M.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, Cambridge, MA, 2012.

[39] D.F. Specht, Probabilistic neural networks, Neural Networks 3(1) (1990) 109-118.

[40] D.F. Specht, A general regression neural network, IEEE Transactions on Neural Networks 2(6) (1991) 568-576.

[41] J. Schmidhuber, Deep Learning in neural networks: An overview, Neural Networks 61 (2015) 85-117.

[42] J. Derrac, S. Garcia, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm and Evolutionary Computation 1 (1) (2011) 3-18.

[43] S. Garcia, A. Fernandez, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Information Sciences 180 (10) (2010) 2044-2064.

[44] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, NJ, 1988.

[45] E. Mezura-Montes, C.A. Coello, Constraint-handling in nature-inspired numerical optimization: Past, present and future, Swarm and Evolutionary Computation 1 (4) (2011) 173-194.

[46] M. Hauschild, M. Pelikan, An introduction and survey of estimation of distribution algorithms. Swarm and Evolutionary Computation 1(3) (2011) 111-128.

[47] K.Q. Ye, M. Hamada, Critical values of the Lenth method for unreplicated factorial designs, Journal of Quality Technology 32(1) (2000) 57-66.