water

MDPI

Article

Wastewater Quality Screening Using Affinity Propagation Clustering and Entropic Methods for Small Saturated Nonlinear Orthogonal Datasets

George Besseris

Department of Mechanical Engineering, The University of West Attica, 12241 Egaleo, Attica, Greece; besseris@uniwa.gr

Abstract: Wastewater recycling efficiency improvement is vital to arid regions, where crop irrigation is imperative. Analyzing small, unreplicated–saturated, multiresponse, multifactorial datasets from novel wastewater electrodialysis (ED) applications requires specialized screening/optimization techniques. A new approach is proposed to glean information from structured Taguchi-type sampling schemes (nonlinear fractional factorial designs) in the case that direct uncertainty quantification is not computable. It uses a double information analysis–affinity propagation clustering and entropy to simultaneously discern strong effects and curvature type while profiling multiple water-quality characteristics. Three water quality indices, which are calculated from real ED process experiments, are analyzed by examining the hierarchical behavior of four controlling factors: (1) the dilute flow, (2) the cathode flow, (3) the anode flow, and (4) the voltage rate. The three water quality indices are: the removed sodium content, the sodium adsorption ratio, and the soluble sodium percentage. The factor that influences the overall wastewater separation ED performance is the dilute flow, according to both analyses' versions. It caused the maximum contrast difference in the heatmap visualization, and it minimized the relative information entropy at the two operating end points. The results are confirmed with a second published independent dataset. Furthermore, the final outcome is scrutinized and found to agree with other published classification and nonparametric screening solutions. A combination of modern classification and simple entropic methods which are offered through freeware R-packages might be effective for testing high-complexity 'small-and-dense' nonlinear OA datasets, highlighting an obfuscated experimental uncertainty.

Keywords: nonlinear factorial screening; wastewater recycling; water quality index; electrodialysis; affinity propagation clustering; surprise; entropy; heatmaps

Citation: Besseris, G. Wastewater Quality Screening Using Affinity Propagation Clustering and Entropic Methods for Small Saturated Nonlinear Orthogonal Datasets. *Water* **2022**, *14*, 1238. https://doi.org/10.3390/w14081238

Academic Editor: Maria Gavrilescu

Received: 25 March 2022
Accepted: 11 April 2022
Published: 12 April 2022

1. Introduction

Water is a substance that attracts continual attention because, while it is abundant in nature, only a small portion is suitable to cover human needs [1,2]. Nowadays, the demand for clean water is unprecedented. This is mainly due to an increasing world population and improving living standards—as evinced from broadly evolving consumerism—but also to daunting environmental deterioration on a global scale. Therefore, there is an international drive toward ensuring water availability through sustainable management of resources [3,4]. Polluted water resources pose various chemical challenges to scientists and engineers who strive to comprehend the complexities behind the indigenous separations phenomena. Thus, modern wastewater treatment methods and techniques are obliged to adapt to the new paradigms. Among the many routes of handling such difficult problems, "the improved water quality through effluent treatment" and the "improved water efficiency through application of the 5R principles: reduce, reuse, recover, recycle, replenish" are highly recommended pathways for workable action [3]. There are several methods and techniques to confront the filtration problem in wastewater supplies [5]. The electrochemical approaches have been an effective way to control water processing and

separation activities [6,7]. Particularly important is the efficient use and reuse of wastewater reserves in irrigating crops as it has been well addressed in several current reviews [8–10]. However, there are circumstances—agricultural irrigation in semi-arid and arid zones—where unfiltered, partially treated and even filtered wastewater bodies could pose several potential risks to food production sustainability and eventually to public health [11]. As the looming water scarcity causes the unavoidable exploitation of extant wastewater stocks, it becomes vital that optimized nutrient recovery be accomplished to balance the water reuse cycles [12,13]. The screening/optimization of water quality may inevitably include numerous resource origins that correspond to a wide range of water uses, from hydroponic neutralization to heavy metal removal applications [14–16]. This is due to the critical effect that wastewater irrigation has on regulating carbon and water neutral crops [17,18]. The screening/optimization of the water quality status requires extensive modeling to ensure quantification of the experimental uncertainty in the performed trials [19,20]. Recent studies take advantage of machine-learning engines to predict groundwater quality, presuming there are sufficient datasets to infer the underlying tendencies in the intricate chemical systems [21–24]. Nevertheless, environmentally oriented screening/optimization investigations may not always generate those ample datasets—for several reasons that machine learning algorithms need to thrive on [25]. This is because the resulting multivariate datasets are perhaps programmed to be intentionally brief—often using fractional factorial designs (FFDs) for planning [26]—if they are to eventually be frugal and hence, permissible.

A comprehensive wastewater electrodialysis (ED) study has been published that intended to discover how three key water quality indices were optimally adjusted for irrigating farms in arid zones [27]. Since the purpose of the study was to upscale the designed experimental units to larger operations, the researcher implemented structured Design of Experiments (DOE) to speed up data generation. Thus, a small data trial-planning scheme was selected from a family of nonlinear Taguchi-type orthogonal arrays (OAs) [28,29]. The motive was to substantially reduce the total number of trials and rapidly obtain useful information that otherwise would be extremely difficult to theorize for several reasons: (1) the complexities arising from the chemical/physical characteristics of the cation/anion exchange membrane ED cell, (2) the widespread wastewater sample nonuniformity due to the three main drainage sources, (3) the stochastic nature of the studied ED process, (4) the unknown interrelationships, if any, among the examined water quality indices, and (5) the necessity for pragmatic trial economy. With only nine prescribed recipes to be executed, the intention was to screen as many as four controlling factors that are widely known to potentially influence ED-based processes [27]. A pre-planned minimal effort was accounted for to synchronously diagnose any curvature effects for each of the examined controlling factors. The typical reduction in the overall trial volume was generous compared to alternative full-factorial schemes. As a matter of fact, the implemented $L_9(3^4)$ OA selection was suppressed by almost 90% of the data requirements of a respective full-factorial scheme. In addition, the decision in the study was not to replicate the specific recipe combinations but to further curtail the time and costs of experimentation.

What has been narrated above is in full agreement with the contemporary trends of data-driven lean and green manufacturing philosophies [30–32]. Data-centric engineering fast-tracks practical and cheap solution options to minimize production waste [33,34]; a strong motivation toward a cleaner environment forms the main tenets behind the lean and green initiatives. Generally speaking, the DOE approach encourages the overall utilization minimization of materials and resources, which is also in congruence with the concepts of sustainability [35,36]. DOE promotes simplicity and frugality in an attempt to gain specialized operational knowledge; it is meant to be immediately applicable to the point of use. The industrial philosophy behind the lean green initiatives is clearly also in harmony with the principles of sustainable chemical systems [37,38]. Specifically, DOE's OA-based screening sampler schemes automatically adhere to the '*green sampling*' principle #5 which favors sample/materials minimization [39]. Furthermore, DOE simultaneously satisfies the requirements of a key metric in green chemistry, through principle #4, the

minimization of chemical waste [40,41]. One may go on and claim that even principle #6 is fulfilled through the use of FFD and OA samplers. The high throughputs of examined parameters the FFD and OA samplers are equipped to handle at the same time make this especially true at the saturated/unreplicated array condition and for larger experimental plans [26,42]. By reducing the trials while increasing the number of candidate influences, a high-throughput trade-off is attained using FFD/OA plans. By exchanging the expected large-sample processing demands (principle #6) to simultaneously screen a larger number of effects, fractional-multifactorial-OA recipe formulations perform less experimental work, succeeding in profiling an equivalent group of parameters. Thus, they offer additional savings in materials, labor, and time. Since optimized OA plans, in terms of minimizing total trial volume, are saturated plans, and further economized if implemented in the 'unreplicated' condition, a great number of techniques have been developed to service FFD/OA under such conditions [43]. Among the more popular methods and techniques to approach the saturated–unreplicated fractional-multifactorial problems are the classic half-normal plot [44], the Lenth test [45], and the Box–Meyer method [46]. Ordinary tools such as the analysis of variance (ANOVA) and the general linear model (GLM) cannot be of immediate use because a combination of saturation and unreplication permits no estimation of the experimental error. This problem becomes even more pronounced when there is a multivariate output to be modelled. For screening designs, the method of Derringer and Suich [47], using a combination of regression and desirability analysis, has found great appeal. and it has been implemented by commercial statistical software packages.

Screening analysis of nonlinear Taguchi-type datasets poses the extra impediment of dealing with additional settings during the profiling process. Moreover, modelling multiple water quality indices that may or may not correlate between them creates a more complex situation. The resulting data structure presents an even more clouded outlook, relying on the fact that indices intended for interpreting crop cultivation water quality performance may be available in ratio or percentage form; the percentage scale is a bounded scale that ranges between 0% and 100%. Additionally, robustness issues should be raised as the recorded datasets from ordinary separation processes, such as the ED, are collected from stochastic phenomena with transient trajectories [48].

Undoubtedly, the greatest practical incentive of using DOE in the industry is to manage to carry out the dual tasks of multifactorial screening and parameter optimization in a single experimental endeavor, as it was initially conceived through the classical Taguchi methods several decades ago [28,29]. It is still very desirable to apply DOE methods in modern manufacturing, while accomplishing both aims—screening and optimization. However, it is an arduous project for most companies to gain practical knowledge from DOE without substantial expertise in structured experimentation [49,50]. Part of this discrepancy stems from the fact that even the basic methods appear to engineers rather mystifying to implement. Empirical data are best examined by embracing diversity. If both stochastic and algorithmic methods are used cooperatively, any prediction disagreements between them are further scrutinized to discover the hidden causes [51]. DOE cannot be alienated from such circumstances, particularly in the phase of screening where the concepts of classification and optimization are combined to separate the influential factors from the weak [48,52]. A primary challenge for this work is mooted on the operability of screening methods. It is particularly focused on probing those multifactorial profilers that are intended to induce maximum (small) data exploitation. Maximum data utilization occurs under the unreplicated and saturated conditions [43]. An additional motivation is raised from past research findings regarding the necessity for improved reliability of predictions among different methods and across different software packages [53]. Unreplicated–saturated OA screening methods present an idiosyncrasy that diverges from the classical treatments of analysis of variance [54]. ANOVA estimates statistical significance by taking into account the magnitude of a variance due to a specific factor against the (remaining) experimental error variance. In contrast, unreplicated–saturated screening pits the effects directly against each other; those that do not conform with a statistically inert group are declared active.

Grouping effects is based on some affinity behavior, which becomes more pronounced for the weak effects. The sparsity assumption favors the occurrence of trivial influences in OA sampling schemes [43]. Then, it becomes obvious that reliable screening could involve robust comparisons based on both the magnitude of uncertainty in the trials as well as the relative factor strength as communicated among the examined factors [55]. The former feature is solely encountered in the classical ANOVA methods, and the latter are only found in the prevalent unreplicated/saturated OA/FFD methods. Uncertainty, then, is the sieve that separates strong from weak effects in ANOVA. In contrast, factorial irregularity is the sieve that tags strong effects in familiar unreplicated/saturated OA/FFD-profilers. Consequently, this dual group statistical detachment (inert/active) may appear relatively binding within each group or be perceived as perhaps promoting factor dependence [56]. On the contrary, algorithmic methods could provide 'mechanism-free' information to confirm results that have been obtained by various stochastic data models [51]. The purpose of this study is to use classification-based algorithms to perform a nonlinear multiresponse multifactorial screening. Multivariate screening techniques are generally in demand to chemometrics because of the overall assistance empirical methods offer in lowering production costs [57,58]. This is particularly welcomed today because the circular economy is deeply connected to operational sustainability through smart, lean, and green frameworks [59,60]. It is water quality indices that will be synchronously screened for a novel ED process using affinity propagation clustering and entropic methods [61,62]. This work offers freeware solver approaches by resorting to proven R-packages [63,64]. It also adds an interesting case study contribution in exhibiting the use of the R-based DOE toolbox in analyzing OA-inspired datasets in chemometrical problems [65,66]. The case study datasets include two independent experimental OA trials to allow for confirmation of the predictions [27]. The experimental ED setup has been well described, and its novel application in improving wastewater filtration for crop irrigation in arid areas has been published recently. The rest of the paper is organized by introducing a methodology to prepare and analyze nonlinear OA datasets using DOE, affinity propagation clustering, and entropic methods in an R-based statistical software environment. Next, in the Results section, the outcomes of the double information analysis are presented and contrasted. A Discussion section provides a confirmation part that verifies the outcomes with a second independent dataset as well as a critical review of the outcomes of the new method against past predictions from statistical and other empirical modeling methods. A Conclusion section summarizes the key points of this work.

2. Materials and Methods

2.1. The OA Sampler Structure

A typical OA plan organizes an expedient product/process screening/optimization study by properly prescribing the settings of a group of as many as m examined controlling factors [26]. An OA sampler is a table that specifies the execution of a minimum of n number of recipes to collect enough information to determine the strength of the investigated effects. For linear dependencies, the maximum throughput for an OA sampler is attained at the saturation condition, i.e., in the case where the number of runs and the number of controlling factors are related by the equality: n = m + 1. We assume that the conditions of saturation as well as unreplication are present in the developments that follow. Unreplication is construed to be the condition whenever there is only one opportunity to collect an observation—a sole experimental run is conducted. In case the saturated OA design should encompass potential curvature effects, to a minimal extent, then, the condition between the number of runs and the number of factors becomes n = 2 × m + 1. In this allotment, a third factor setting is added, and it is located within the studied range [28,29]. The overall consequence of implementing a saturated–unreplicated OA sampling plan is to impose the obvious constraint of 'dataset smallness' in the information generation cycle. An abstract depiction of an OA matrix is shown in Table 1. The examined controlling factors are coded

as: $\boldsymbol{X_j}$ for $1 \leq j \leq m$, and their respective factor settings are correspondingly coded as x_{ij} for $1 \leq i \leq n$, and $1 \leq j \leq m$.

Table 1. A general arrangement of an OA matrix.

Controlling Factors				
run #	*X*1	*X*2	. . .	*Xm*
1	*x*11	*x*12	. . .	*x*1*m*
2	*x*21	*x*22	. . .	*x*2*m*
.	.	.		.
.	.	.	. . .	.
.	.	.		.
n	*xn*1	*xn*2	. . .	*xnm*

It is considered that the outcome of a given ith run may be synchronously quantified by the measured responses of as many as c product/process characteristics, $\boldsymbol{R_k}$ for $1 \leq k \leq c$. Then, their corresponding coded entries (output) could be denoted as: r_{ik}, for $1 \leq i \leq n$ and $1 \leq k \leq c$. From an informational point of view, the *i*th run introduces an ith sequence as input which identifies with the strict particulars of the *i*th recipe, and it is written as follows:

*i*th run (factorial recipe)-input sequence:

x_{i1}	x_{i2}	...	x_{im}

Then, the multiresponse output after the execution of the *i*th factorial recipe will be:

r_{i1}	r_{i2}	...	r_{ik}

*i*th run multiresponse entries:

2.2. The Unsupervised Analyzer

The input–output relationship resulting from conducting the nonlinear OA sampling plan is analyzed by the affinity propagation clustering approach [27] that furnishes exemplar information in the partitioned clusters. The structured multiresponse dataset is converted to a two-point similarity matrix, $s(i, j)$, where each data point i relays its affinity to the exemplar point j. To initiate cluster memberships, preference values are usually selected from the computed similarities that possess larger magnitudes. In the required two-way messaging process, a data point solicits candidate exemplars for potential membership through a responsibility matrix, $r(i, j)$ and affirms membership suitability through the availability matrix, $a(i, j)$. Since screening is the primary goal of this exercise, the end-point correlations across runs and their relationships to the formed 'most-distant' clusters are of much value. The practical and popular informational portrayal of their tendencies are easily visualized through an assorted heatmap. As a second means of testing, a classic comparison ought to involve the relative entropy evaluation on the resulting clusters from the affinity propagation method. In brief, if the *i*th cluster has n_i members ($n = \sum_{i=1}^{c} n_i$), x_{jik}, where the number of clusters is $1 \leq i \leq c$, with factor setting label $j_k \mid j_k \in \{1, 2, \ldots, s\}$ for a total number of s settings in the kth controlling factor, $1 \leq k \leq F$, for the total number of F factors in the selected OA plan. Then, the surprise for the kth factor, the *i*th cluster and the n_{ijk} members is:

$$u_{ik} = -\log_2(p_{ik}) \text{ with } p_{ik} = \sum_{jk=1}^{s} n_{ijk}/n_i$$

The expected relative surprise (relative entropy) will be [62]:

$$H = \sum_{i=1}^{c} p_{ik} u_{ik} / \log_2(n)$$

2.3. The Water Quality Case Study

The case study reflects a real published attempt to manage wastewater treatment using two electrodialysis setups [27]. The quality of the filtrated water was anticipated to be suitable for irrigation in arid areas. The unique screening/optimization effort was modelled by implementing a classical nonlinear Taguchi-type OA plan. The selected four-factor three-level nine-run $L_9(3^4)$ OA was conducted in two modes for practical purposes. Since the OA-specified runs were executed once, the collected dataset was available in the unreplicated conditions. In addition, in the first electrodialysis setup, the OA plan was arranged to carry the maximum load of investigated factors. The trial planner was saturated with the four controlling factors: (1) A–the dilute flow, (2) B–the cathode flow, (3) C–the anode flow, and (4) D–the voltage rate. In the second electrodialysis setup, the anode flow was irrelative. Therefore, the dataset from the first setup was intended to generate a prognosis [27], and the second (independent) dataset was to be utilized for confirmation on the prognostic outcome. It is very important to mention that what made this case study so unique was that despite the inherent data smallness, the collected output was characterized by three water quality indices, which are meaningful in cultivating crops. The three water quality indices are: (1) RS–the removed sodium content (%), (2) SAR–the sodium adsorption ratio, and (3) SSP–the soluble sodium percentage (%). All three water quality indices have been described in detail in earlier publications, and they were found statistically uncorrelated among them [27,48,52]. Thus, the experimental design was programmed to deliver a multiresponse output. Subsequently, the screening/optimization modelling was formulated according to a nonlinear multiresponse multifactorial unreplicated–saturated OA dataset. This means that screening prognostics are to be generated against unknown uncertainty. No degrees of freedom are available for quantifying the statistical error.

2.4. The Methodological Steps

The proposed methodology is accordingly summarized:

(1) Define the wastewater quality characteristics that monitor the direction of the ED progress and quantify the recycling efficiency improvement.
(2) Select the proper group of the ED process controlling factors that are deemed relevant to regulating the influent wastewater condition, and directly screen the multivariate effluent tendencies.
(3) Determine a practical operating range for each of the controlling factors, such as to induce adequate variability, that could potentially detect a presence of curvature effects.
(4) Select a suitable OA sampling plan [26,28,29] that economically accommodates the nominated group of controlling factors (from step 2).
(5) Execute the prescribed OA runs (step 4) and compile the multiresponse dataset.
(6) Apply affinity propagation clustering [61,63] to corral the cluster members from the multiresponse dataset.
(7) Ensure convergence of the estimations of the predicted exemplar preferences and fitness (maximizing overall net similarity) to proceed in determining the cluster hierarchy.
(8) Prepare the cluster dendrogram and the visualized clustering result, including the designated exemplar points. Pinpoint on a similarity–matrix heatmap the contoured clustering performance to assess the correlation between potential operational limits.
(9) Provide a double verification of the strong effect predictions (from step 8) by reassessing the clustering outcomes by their estimated relative surprise measure, leading to a relative entropy measure for each labelled cluster [62].
(10) Confirm the results with additional independent datasets.

2.5. The Computational Aids

The computational work was carried out on the free statistical software R (v. 4.1.2) [64]. The module 'param.design()' from the R-package 'DoE.base' (v. 1.2) provided the nonlinear OA sampler. The R-package 'cluster' (v. 2.1.2) was utilized in grouping memberships. The affinity propagation approach, which offered exemplar-based agglomerative clustering

capabilities, was implemented through the R-package 'apcluster' (v. 1.4.9) [61,63]. The transformation of distances to similarities was mediated by maintaining the exponent in the default mode (Laplace kernel) in the *apcluster* function (*negDistMat* similarity matrix). The generated graphs delivered convergence performance for the fitness score, the sum of exemplars, and the sum of similarities. Primary graphical information was supplied by a standard dendrogram, displaying the cluster hierarchy, the exemplar-based groupings and the assorted heatmaps to quickly recognize contrasting factor settings. The R-package 'entropy' (v. 1.3.1) was used to estimate the cluster entropies.

3. Results

The convergence performance of the affinity propagation clustering algorithm (the R-package 'APCluster') [63] for the three water quality index datasets (Table 4 in ref. [27]) was successful, and it is shown in Figure 1. For the nine tested samples, the input preference, the sum of similarities, the sum of preferences, and the net similarity were found to be −7.59, −26.59, −22.78, and −49.37, respectively. This resulted in three clusters (dendrogram in Figure 2) with exemplar OA runs: 1, 4, and 9. More specifically, cluster #1 (exemplar 1) included OA runs # 1, 2, and 3; cluster #2 (exemplar 4) included OA runs 4, 5, 6, and 8; and cluster #3 (exemplar 9) included OA runs 7 and 9. Cluster #3 was observed to be the most dissimilar with respect to the other two. There is an association to this behavior that mainly relates to the influence of dilute flow (Table 3 in ref. [27]) on the most affected water quality index—the removed Na^+ response (Table 4 in ref. [27]). This becomes more pronounced when observing the full clustering two-response plot matrix in Figure 3. The first water quality index (removed Na^+) generates the required clustering impetus when compared to the other two indices. From Figure 4, the mini-data groups that are formed from OA runs 3, 5, and 7, 9, respectively, appear to create the most antithetical contrast across all runs in the cluster heatmap visualization; it implies that the stronger clustering drive emanates from the removed Na^+ response and, hence, it identifies the dilute flow as the primary controlling factor. The third setting (fixed at 10 L/h) of the dilute flow (Table 3 in ref. [27]), provides a great response separation with respect to the other two adjustments. Since this behavior remains indistinguishable for the first two settings, the dilute flow might be optimally regulated in the prescribed factorial landscape at the setting of 2 L/h. A post-verification is accomplished by executing the R-package 'APCluster' for a two cluster scheme (Figure 5).

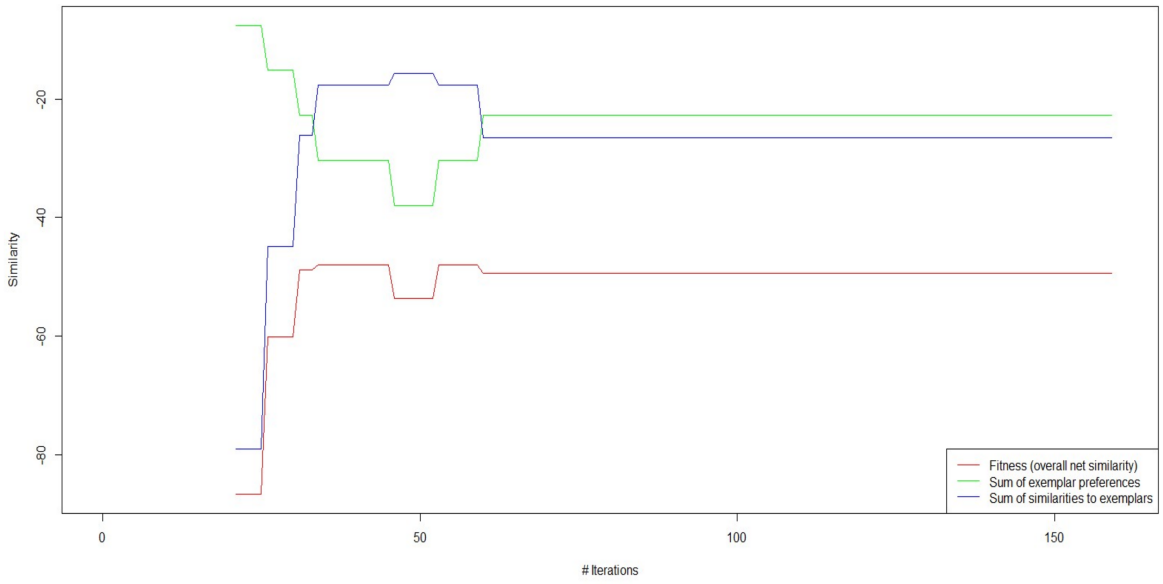

Figure 1. Similarity convergence for: (green) sum of exemplar preferences; (purple) sum of similarities to exemplars; and (red) fitness.

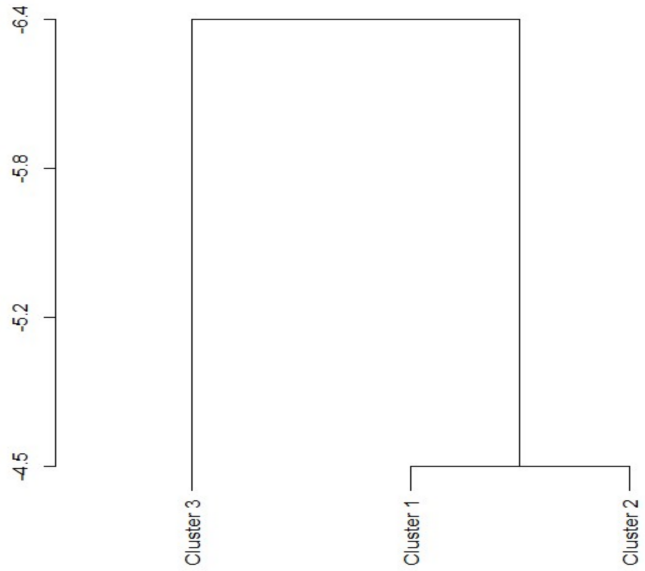

Figure 2. The cluster dendrogram for the ED-process (the three water quality indices dataset).

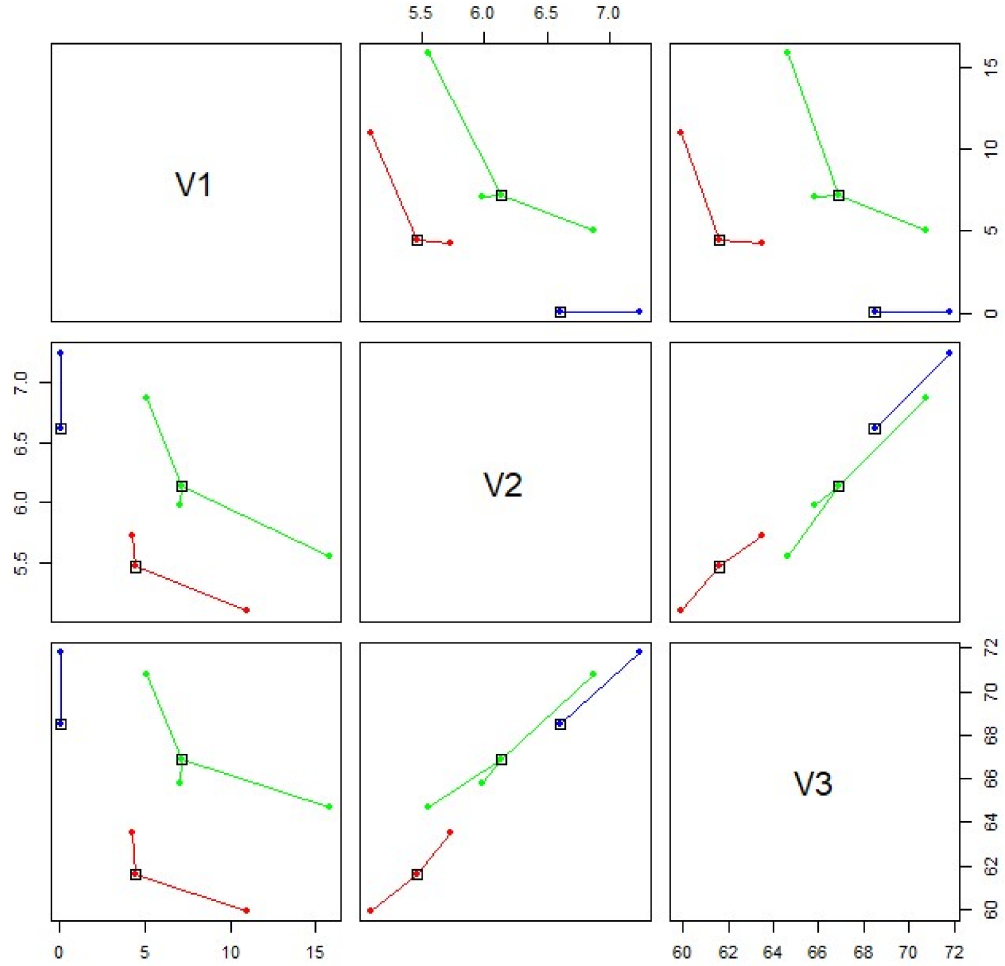

Figure 3. Clustering result for two-response combinations (three-cluster scheme) of the three water quality indices (V1 = removed Na^+, V2 = SAR, V3 = Na^+ ratio).

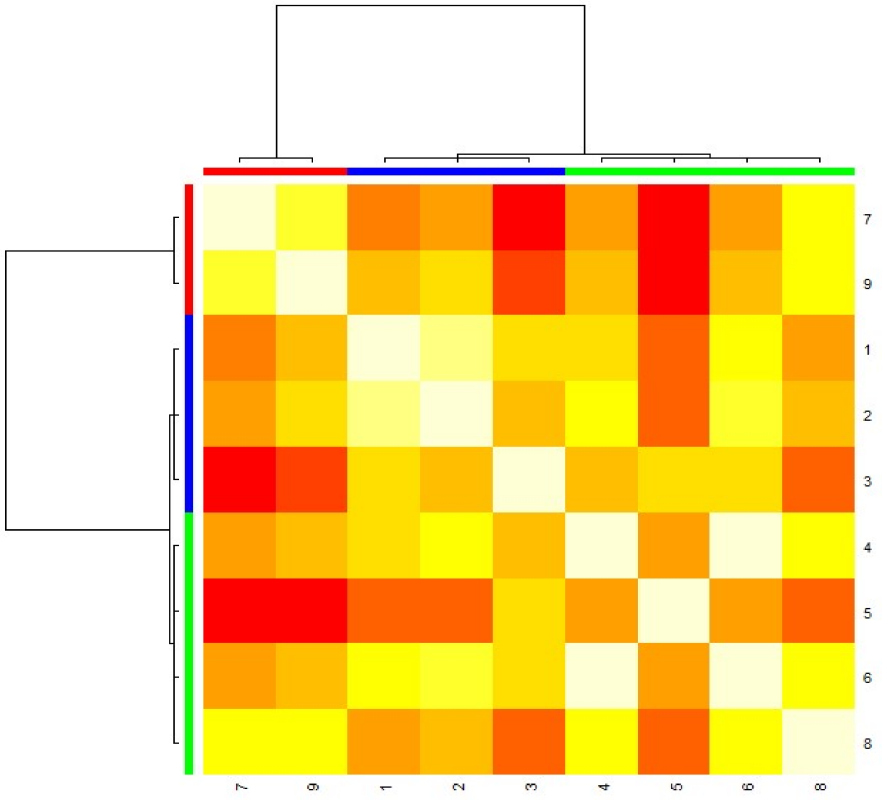

Figure 4. Cluster heatmap of the nine OA runs simultaneously taking in account the response of the three water quality indices.

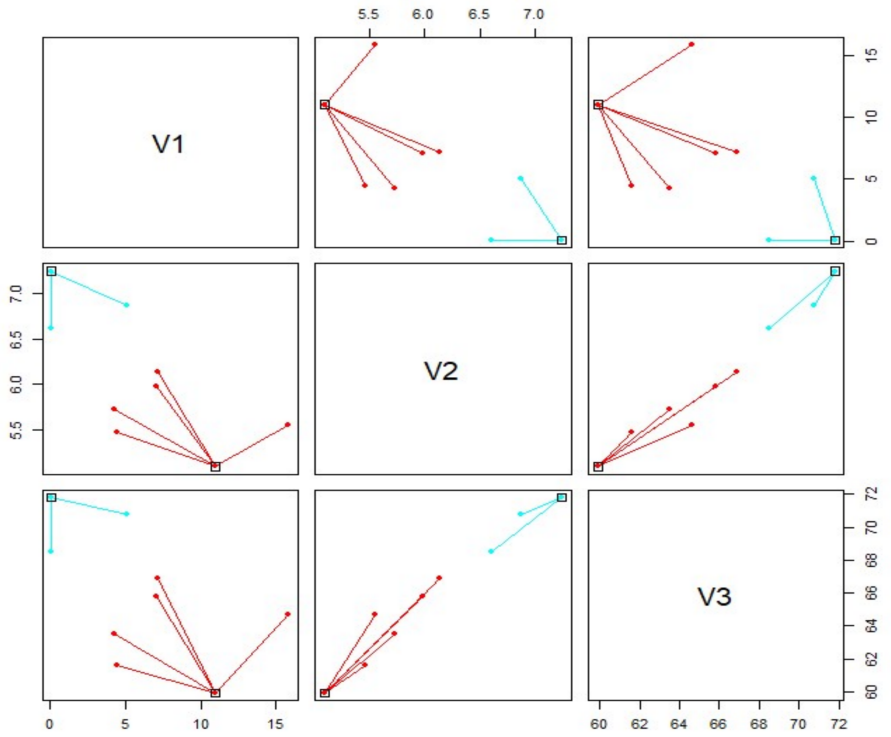

Figure 5. Clustering result for two-response combinations (two-cluster scheme) of the three water quality indices (V1 = removed Na^+, V2 = SAR, V3 = Na^+ ratio).

In this case, the two exemplars partner to cluster #1 (exemplar 3) with OA run cluster members: 1, 2, 3, 4, 5, 6, and cluster #2 (exemplar 7) with OA run cluster members: 7, 8, 9. For the two-cluster post-verification scheme, the input preference, the sum of similarities, the sums of preferences, and the net similarity were found to be −9.0, −44.90, −18.0, and −62.9, respectively. The same final screening outcome seems to also be supported by the verification attempts by a nonparametric solver and a recent microclustering-denominalization technique [48,52]. Finally, a pure informational approach using the relative surprise of the clustered OA runs (for the three-cluster scheme) and then the expected relative surprise (relative entropy) mirrors the same conclusions as listed in Table 2. We observe that the only settings with the minimum information are the end points of the dilute flow, furnishing additional design evidence about the strong influential status of this controlling factor.

Table 2. Relative entropy for clustered factor settings.

Factor/Setting	P(X)	Relative Entropy
A1	1	0.00
A2	0.75/0.25	0.41
A3	1	0.00
B1	0.33/0.33/0.33	1.00
B2	0.25/0.25/0.5	0.75
B3	0.5/0.5	1.00
C1	0.33/0.33/0.33	1.00
C2	0.25/0.25/0.5	0.75
C3	0.5/0.5	1.00
D1	0.33/0.33/0.33	1.00
D2	0.25/0.25/0.5	0.75
D3	0.5/0.5	1.00

4. Discussion

To confirm the results from the previous section, a second independent dataset was investigated on the same three water quality indices of the ED process (Table 10 in ref. [27]). The convergence performance of the affinity propagation clustering algorithm on the dataset was successful and is shown in Figure 6. For the nine tested samples, the input preference, the sum of similarities, the sum of preferences, and the net similarity were found to be −23.70, −39.91, −71.10, and −111.01, respectively. This resulted in three clusters (dendrogram in Figure 7) with exemplar OA runs: 3, 7, and 8. More specifically, cluster #1 (exemplar 3) included OA runs # 1, 2, and 3; cluster #2 (exemplar 7) included OA runs 4, 5, 6, 7, 9; and cluster #3 (exemplar 8) comprising the single OA run 8. Again, cluster #3 is observed to be the most dissimilar with respect to the other two. There is an association to this behavior that mainly relates to the influence of dilute flow (Table 9 in ref. [27]) on probably all three water quality indices (Table 10 in ref. [27]). Surprisingly, the full clustering two-response plot matrix in Figure 8 demonstrates the separation of all water quality indices. All responses affect the clustering process. From Figure 9, the mini-data groups that are formed from OA runs 8 and 1, 2, 3, respectively, appear to create the most opposing effect across all runs in the cluster heatmap visualization. It identifies the dilute flow as the primary controlling factor. The low setting (2 L/h) OA runs of the dilute flow (Table 9 in ref. [27]), ensure the great response separation with respect to the high endpoint adjustment.

The same final screening outcome agrees with the verification attempt by a recent microclustering-denominalization technique [52]. The classical (informational) check was repeated using the relative surprise concept for the clustered OA runs to obtain the expected relative surprise (relative entropy). The conclusions match the overall result as is easily observed from Table 3. Again, it is the end-point settings of the dilute flow that minimize information and, hence, reaffirm its only dominant status in the final ED process design of relevant controls. From Table 3, it becomes apparent that one setting from each of the other examined factors delivers minimum information (B3 and C3). Nevertheless, no contrast

is witnessed for those two factors as the performance for the other two settings is well over 50%.

Table 3. Relative entropy for clustered factor settings (confirmation dataset).

Factor/Setting	P(X)	Relative Entropy
A1	1	0.00
A2	0.6/0.4	0.42
A3	1	0.00
B1	0.33/0.33/0.33	1.00
B2	0.4/0.4/0.2	0.65
B3	1	0.00
C1	0.33/0.33/0.33	1.00
C2	0.4/0.4/0.2	0.65
C3	1	0.00

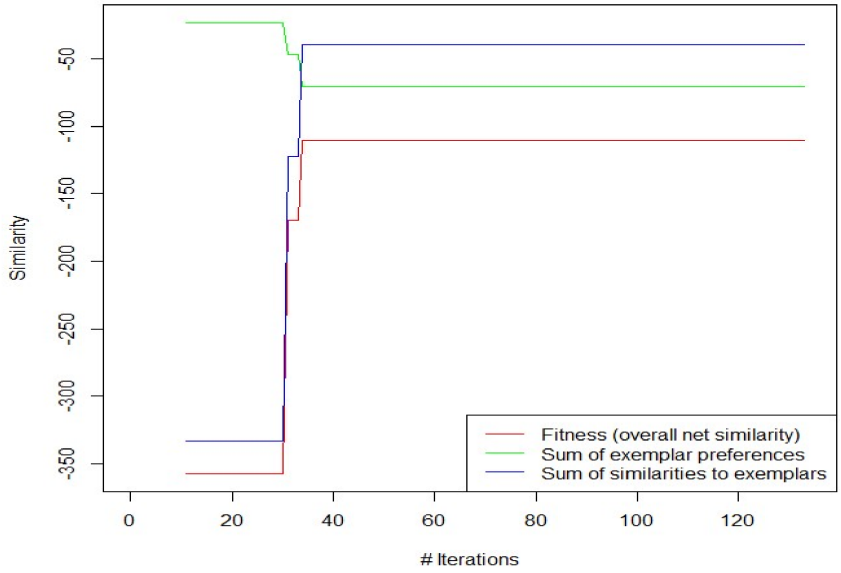

Figure 6. Similarity convergence for: (green) sum of exemplar preferences; (purple) sum of similarities to emplars; and (red) fitness (confirmation dataset).

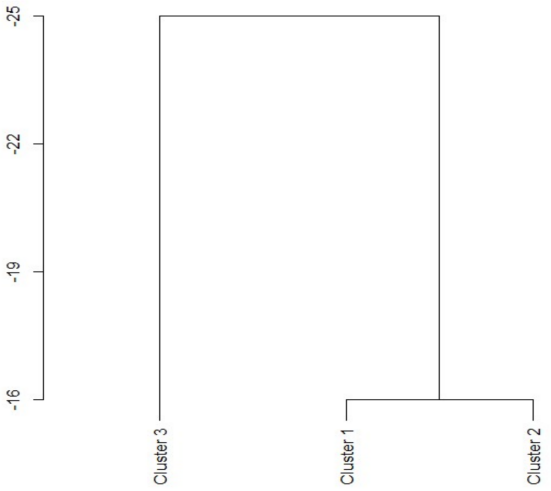

Figure 7. The cluster dendrogram for the confirmation ED-process dataset (the three water quality indices).

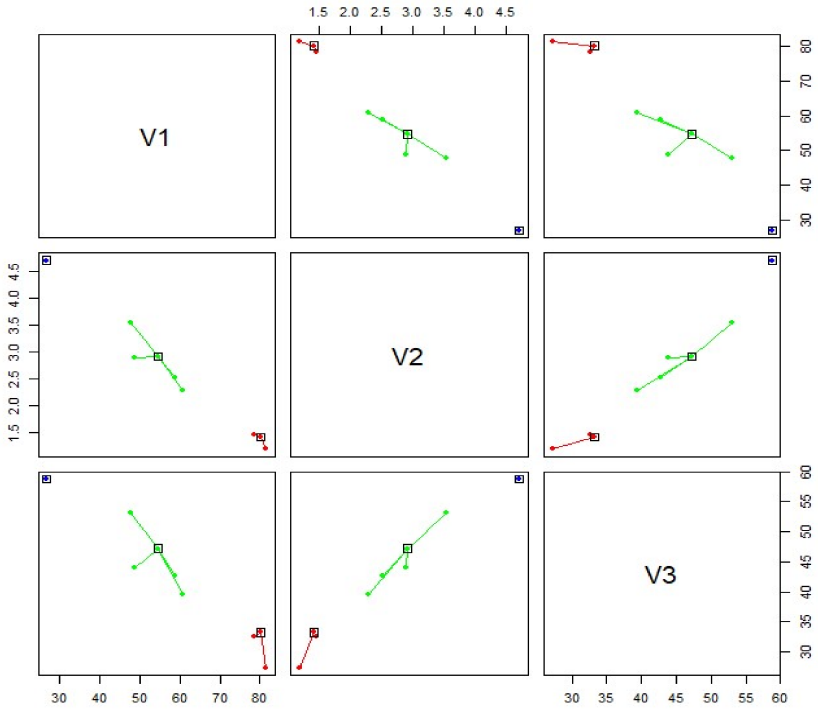

Figure 8. Clustering result (confirmation dataset) for two-response combinations (three-cluster scheme) of the three water quality indices (V1 = removed Na^+, V2 = SAR, V3 = Na^+ ratio).

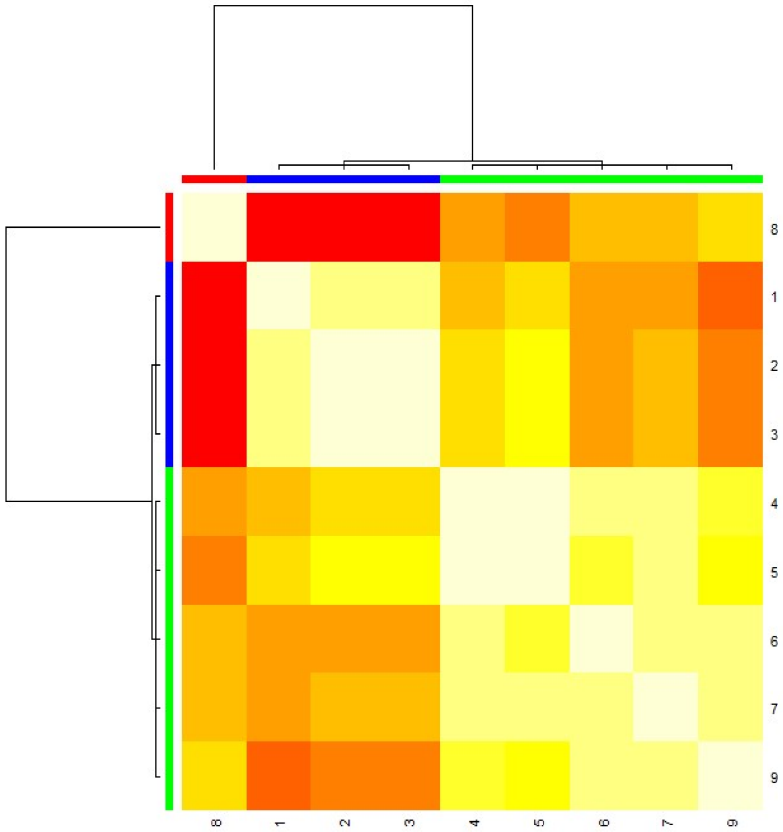

Figure 9. Cluster heatmap of the nine OA runs simultaneously taking into account the response of the three water quality indices (confirmation dataset).

5. Conclusions

Wastewater filtration is extremely important for arid regions. Recently, electrodialysis has been suggested as a potential option to resolve the intricate phenomena in a manner competitive to other separation methods. ED process trials on polluted wastewater stocks of various origins may be lengthy to carry out and arduous to stochastically interpret. Using modern affinity propagation clustering and regular entropic methods 'small-and-dense' multiresponse, OA datasets were profiled against several controlling factors, while statistically examining the potential effects of curvature in each of them. A re-examined real case study, which was published recently, involved three water quality characteristics pertinent to crop irrigation performance and four controlling factors that were assumed might regulate the filtration performance of an innovative electrodialysis apparatus. The dense dataset was generated from nine specific recipes as prescribed by the selected Taguchi-type $L_9(3^4)$ OA sampling scheme. The proposed approach managed to detect the strong influence of the diluted flow in the ED cell trials. It additionally furnished visual representation of the pertinent effect of polarity across factor settings. A combination of expert-level freeware R-packages was utilized to instill credibility in the probing style as well as to facilitate reproducibility checks and convenience in completing the analysis steps. The approach may supplement other techniques in predicting 'hard-to-handle' FFD/OA datasets in the case that the overall experimental error is not assessable. Future work could involve denser FFD/OA matrices with mixed-level variables that produce mixed-type datasets by blending nominal and continuous variables.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Finney, J. *Water: A Very Short Introduction*; Oxford University Press: Oxford, UK, 2015.
2. Ball, P. *H_2O: The Biography of Water*; Orion Publishing Co.: London, UK, 2000.
3. SDG Compass. *Ensure Availability and Sustainable Management of Water and Sanitation for All*; United Nations: New York, NY, USA, 2015. Available online: https://sdgcompass.org/sdgs/sdg-6/ (accessed on 7 August 2021).
4. WWAP (United Nations World Water Assessment Programme). *The United Nations World Water Development Report 2017: Wastewater: The Untapped Resource*; UNESCO: Paris, France, 2017.
5. Younas, F.; Mustafa, A.; Rahman Farooqi, Z.U.; Wang, X.; Younas, S.; Mohy-Ud-Din, W.; Hameed, M.A.; Abrar, M.M.; Maitlo, A.A.; Noreen, S.; et al. Current and emerging adsorbent technologies for wastewater treatment: Trends, limitations, and environmental implications. *Water* **2021**, *13*, 215. [CrossRef]
6. Zito, R. *Electrochemical Water Processing*; Wiley-Scrivener: Hoboken, NJ, USA, 2011.
7. Tanaka, Y. *Ion Exchange Membrane Electrodialysis: Fundamentals, Desalination, Separation*; Nova Science: New York, NY, USA, 2013.
8. Zhang, Y.; Shen, Y. Wastewater irrigation: Past, present, and future. *WIRE's Water* **2019**, *6*, 1234. [CrossRef]
9. Jaramillo, M.F.; Restrepo, I. Wastewater reuse in agriculture: A review about its limitations and benefits. *Sustainability* **2017**, *9*, 1734. [CrossRef]
10. Lopez-Serrano, M.J.; Velasco-Munoz, J.F.; Arnar-Sanchez, J.A.; Roman-Sanchez, I.M. Sustainable use of wastewater in agriculture: A bibliometric analysis of worldwide research. *Sustainability* **2020**, *12*, 8948. [CrossRef]
11. Elgallal, M.; Fletcher, L.; Evans, B. Assessment of potential risks associated with chemicals in wastewater used for irrigation in arid and semiarid zones: A review. *Agric. Water Manag.* **2016**, *177*, 419–431. [CrossRef]
12. Ungureanu, N.; Vladut, V.; Voicu, G. Water scarcity and wastewater reuse in crop irrigation. *Sustainability* **2020**, *12*, 9055. [CrossRef]
13. Saliu, T.D.; Oladoja, N.A. Nutrient recovery from wastewater and reuse in agriculture: A review. *Environ. Chem. Lett.* **2021**, *19*, 2299–2316. [CrossRef]
14. Richa, A.; Touil, S.; Fizir, M.; Martinez, V. Recent advances and perspectives in the treatment of hydroponic wastewater: A review. *Rev. Environ. Sci. Biotechnol.* **2020**, *19*, 945–966. [CrossRef]
15. El Batouti, M.; Al-Harby, N.E.; Elewa, M.M. A review on promising membrane technology approaches for heavy metal removal from water and wastewater to solve water crisis. *Water* **2021**, *13*, 3241. [CrossRef]
16. Saleh, T.A.; Mustaqeem, M.; Khaled, M. Water treatment technologies in removing heavy metal ions from wastewater: A review. *Environ. Nanotechnol. Monit. Manag.* **2022**, *17*, 100617. [CrossRef]

17. Mora, A.; Torres-Martinez, J.A.; Capparelli, M.V.; Zabala, A.; Mahlknecht, J. Effects of wastewater irrigation on groundwater quality: An overview. *Curr. Opin. Environ. Sci. Health* **2022**, *25*, 100322. [CrossRef]
18. Lahlou, F.-Z.; Mackey, H.R.; Al-Ansari, T. Role of wastewater in achieving carbon and water neutral agricultural production. *J. Clean. Prod.* **2022**, *339*, 130706. [CrossRef]
19. Burn, D.H.; McBean, E.A. Optimization modelling of water quality in an uncertain environment. *Water Resour. Res.* **1985**, *21*, 934–940. [CrossRef]
20. Rehana, S.; Rajulapati, C.R.; Ghosh, S.; Karmakar, S.; Mujumdar, P. Uncertainty Quantification in Water Resource Systems Modeling: Case Studies from India. *Water* **2020**, *12*, 1793. [CrossRef]
21. Singha, S.; Pasupuleti, S.; Singha, S.S.; Singh, R. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* **2021**, *276*, 130265. [CrossRef]
22. Malviya, A.; Jaspal, D. Artificial intelligence as an upcoming technology in wastewater treatment: A comprehensive review. *Environ. Technol. Rev.* **2021**, *10*, 177–187. [CrossRef]
23. Hanoon, M.S.; Ahmed, A.N.; Fai, C.M.; Birima, A.H.; Razzaq, A.; Sherif, M.; Sefelnasr, A.; El-Shafie, A. Application of artificial intelligence models for modeling water quality in groundwater: Comprehensive review, evaluation and future trends. *Water Air Soil Pollut.* **2021**, *232*, 411. [CrossRef]
24. Sunayana; Kalawapudi, K.; Dube, O.; Sharma, R. Use of neural networks and spatial interpolation to predict groundwater quality. *Environ. Dev. Sustain.* **2020**, *22*, 2801–2816. [CrossRef]
25. Pilar Callao, M. Multivariate experimental design in environmental analysis. *Trends Anal. Chem.* **2014**, *62*, 86–92. [CrossRef]
26. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters—Design, Innovation, and Discovery*; Wiley: New York, NY, USA, 2005.
27. Abou-Shady, A. Recycling of polluted wastewater for agriculture purpose using electrodialysis: Perspective for large scale application. *Chem. Eng. J.* **2017**, *323*, 1–18. [CrossRef]
28. Taguchi, G.; Chowdhury, S.; Wu, Y. *Quality Engineering Handbook*; Wiley-Interscience: Hoboken, NJ, USA, 2004.
29. Taguchi, G.; Chowdhury, S.; Taguchi, S. *Robust Engineering: Learn How to Boost Quality while Reducing Costs and Time to Market*; McGraw-Hill: New York, NY, USA, 2000.
30. Dhingra, R.; Kress, R.; Upreti, G. Does lean mean green? *J. Clean. Prod.* **2014**, *85*, 1–7. [CrossRef]
31. Johansson, G.; Sundin, E. Lean and green product development: Two sides of the same coin? *J. Clean. Prod.* **2014**, *85*, 104–121. [CrossRef]
32. Garza-Reyes, J.A. Lean and green—A systematic review of the state of the art literature. *J. Clean. Prod.* **2015**, *102*, 18–29. [CrossRef]
33. Fercoq, A.; Lamouri, S.; Carbone, V. Lean/Green integration focused on waste reduction techniques. *J. Clean. Prod.* **2016**, *137*, 567–578. [CrossRef]
34. Dieste, M.; Panizzolo, R.; Garza-Reyes, J.A.; Anosike, A. The relationship between lean and environmental performance: Practices and measures. *J. Clean. Prod.* **2019**, *224*, 120–131. [CrossRef]
35. Bhattacharya, A.; Nand, A.; Castka, P. Lean-green integration and its impact on sustainability performance: A critical review. *J. Clean. Prod.* **2019**, *236*, 117697. [CrossRef]
36. Teixeira, P.; Sa, J.C.; Silva, F.J.G.; Ferreira, L.P.; Santos, G.; Fontoura, P. Connecting lean and green with sustainability towards a conceptual model. *J. Clean. Prod.* **2021**, *322*, 129047. [CrossRef]
37. Anastas, P.T.; Zimmerman, J.B. Design through the 12 principles of green engineering. *Environ. Sci. Technol.* **2003**, *37*, 94–101. [CrossRef]
38. Constable, D.J.C. Green and sustainable chemistry: The case for a systems-based, interdisciplinary approach. *iScience* **2021**, *24*, 103489. [CrossRef]
39. Lopez-Lorente, A.I.; Pena-Pereira, F.; Pedersen-Bjergaard, S.; Zuin, V.G.; Ozkan, S.A.; Psillakis, E. The ten principles of green sample preparation. *Trends Anal. Chem.* **2022**, *148*, 116530. [CrossRef]
40. Sajid, M.; Plotka-Wasylka, J. Green analytical chemistry metrics: A review. *Talanta* **2022**, *228*, 123046. [CrossRef] [PubMed]
41. Sheldon, R.A.; Bode, M.L.; Akakios, S.G. Metrics of green chemistry: Waste minimization. *Curr. Opin. Green Sustain. Chem.* **2022**, *33*, 100569. [CrossRef]
42. Besseris, G.J. A Distribution-Free Multi-Factorial Profiler for Harvesting Information from High-Density Screenings. *PLoS ONE* **2013**, *8*, e73275. [CrossRef] [PubMed]
43. Hamada, M.; Balakrishnan, N. Analyzing unreplicated factorial experiments: A review with some new proposals. *Stat. Sin.* **1998**, *8*, 1–41.
44. Daniel, C. Use of the half-normal plots in interpreting factorial two-level experiments. *Technometrics* **1959**, *1*, 311–341. [CrossRef]
45. Lenth, R.V. Quick and easy analysis of unreplicated factorials. *Technometrics* **1989**, *31*, 469–473. [CrossRef]
46. Box, G.E.P.; Meyer, R.D. An analysis for unreplicated fractional factorials. *Technometrics* **1986**, *28*, 11–18. [CrossRef]
47. Derringer, G.; Suich, R. Simultaneous optimization of several response variables. *J. Qual. Technol.* **1980**, *12*, 214–219. [CrossRef]
48. Besseris, G.J. Concurrent multiresponse multifactorial screening of an electrodialysis process of polluted wastewater using robust non-linear Taguchi profiling. *Chemom. Intell. Lab. Syst.* **2020**, *200*, 103997. [CrossRef]
49. Ilzarbe, L.; Alvarez, M.J.; Viles, E.; Tanco, M. Practical applications of design of experiments in the field of engineering: A bibliographical review. *Qual. Reliab. Eng. Int.* **2008**, *24*, 417–428. [CrossRef]

50. Tanco, M.; Viles, E.; Ilzarbe, L.; Alvarez, M.J. Implementation of Design of Experiments projects in industry. *Qual. Reliab. Eng. Int.* **2009**, *25*, 478–505. [CrossRef]
51. Breiman, L. Statistical modeling: The two cultures. *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
52. Besseris, G. Micro-Clustering and Rank-Learning Profiling of a Small Water-Quality Multi-Index Dataset to Improve a Recycling Process. *Water* **2021**, *13*, 2469. [CrossRef]
53. Fontdecaba, S.; Grima, P.; Tort-Martorell, X. Analyzing DOE with Statistical Software Packages: Controversies and proposals. *Am. Stat.* **2014**, *68*, 205–211. [CrossRef]
54. Fisher, R.A. *Statistical Methods, Experimental Design, and Scientific Inference*; Oxford University Press: Oxford, UK, 1990.
55. Besseris, G.J. Order Statistics for a Two-Level, Eight-Run Saturated-Unreplicated Fractional-Factorial Screening. *Qual. Eng.* **2009**, *21*, 416–431. [CrossRef]
56. Mee, R.W.; Lu, X. Don't use rank sum tests to analyze factorial designs. *Qual. Eng.* **2010**, *23*, 26–29. [CrossRef]
57. Carlson, R.; Nordahl, A.; Barth, T.; Myklebust, R. An approach to evaluating screening experiments when several responses are measured. *Chemom. Intell. Lab. Syst.* **1991**, *12*, 237–255. [CrossRef]
58. Lepeniotis, S.S.; Vigezzi, M.J. Lowering manufacturing cost of material by formulating it through statistical modeling and design. *Chemom. Intell. Lab. Syst.* **1995**, *29*, 133–139. [CrossRef]
59. Lim, M.K.; Lai, M.; Wang, C.; Lee, Y. Circular economy to ensure production operational sustainability: A green-lean approach. *Sustain. Prod. Consum.* **2022**, *30*, 130–144. [CrossRef]
60. Touriki, F.E.; Benkhati, I.; Kamble, S.S.; Belhadi, A.; El Fezazi, S. An integrated smart, green, resilient, and lean manufacturing framework: A literature review and future research directions. *J. Clean. Prod.* **2021**, *319*, 128691. [CrossRef]
61. Frey, B.J.; Dueck, D. Clustering by passing messages between data points. *Science* **2007**, *315*, 972–976. [CrossRef] [PubMed]
62. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
63. Bodenhofer, U.; Kothmeier, A.; Hochreiter, S. APCluster: An R package for affinity propagation clustering. *Bioinformatics* **2011**, *27*, 2463–2464. [CrossRef] [PubMed]
64. R Core Team. *R (Version 4.1.2): A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021; Available online: https://www.R-project.org/ (accessed on 16 February 2022).
65. Lawson, J. *Design and Analysis of Experiments with R*; CRC Press: Boca Raton, FL, USA, 2014.
66. Stone, R.A.; Veevers, A. The Taguchi influence on designed experiments. *J. Chemom.* **1994**, *8*, 103–110. [CrossRef]