water

MDPI

Article

Databionic Swarm Intelligence to Screen Wastewater Recycling Quality with Factorial and Hyper-Parameter Non-Linear Orthogonal Mini-Datasets

George Besseris

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

Abstract: Electrodialysis (ED) may be designed to enhance wastewater recycling efficiency for crop irrigation in areas where water distribution is otherwise inaccessible. ED process controls are difficult to manage because the ED cells need to be custom-built to meet local requirements, and the wastewater influx often has heterogeneous ionic properties. Besides the underlying complex chemical phenomena, recycling screening is a challenge to engineering because the number of experimental trials must be maintained low in order to be timely and cost-effective. A new data-centric approach is presented that screens three water quality indices against four ED-process-controlling factors for a wastewater recycling application in agricultural development. The implemented unsupervised solver must: (1) be fine-tuned for optimal deployment and (2) screen the ED trials for effect potency. The databionic swarm intelligence classifier is employed to cluster the $L_9(3^4)$ OA mini-dataset of: (1) the removed Na^+ content, (2) the sodium adsorption ratio (SAR) and (3) the soluble Na^+ percentage. From an information viewpoint, the proviso for the factor profiler is that it should be apt to detect strength and curvature effects against not-computable uncertainty. The strength hierarchy was analyzed for the four ED-process-controlling factors: (1) the dilute flow, (2) the cathode flow, (3) the anode flow and (4) the voltage rate. The new approach matches two sequences for similarities, according to: (1) the classified cluster identification string and (2) the pre-defined OA factorial setting string. Internal cluster validity is checked by the Dunn and Davies–Bouldin Indices, after completing a hyper-parameter $L_8(4^1 2^2)$ OA screening. The three selected hyper-parameters (distance measure, structure type and position type) created negligible variability. The dilute flow was found to regulate the overall ED-based separation performance. The results agree with other recent statistical/algorithmic studies through external validation. In conclusion, statistical/algorithmic freeware (R-packages) may be effective in resolving quality multi-indexed screening tasks of intricate non-linear mini-OA-datasets.

Keywords: non-linear factorial screening; wastewater recycling; water quality index; electrodialysis; databionic swarm intelligence; unsupervised cluster matching; dissimilarity validation

Citation: Besseris, G. Databionic Swarm Intelligence to Screen Wastewater Recycling Quality with Factorial and Hyper-Parameter Non-Linear Orthogonal Mini-Datasets. *Water* **2022**, *14*, 1990. https://doi.org/10.3390/w14131990

Academic Editor: Laura Bulgariu

Received: 27 May 2022
Accepted: 20 June 2022
Published: 21 June 2022

1. Introduction

Wastewater constitutes a precious resource, and its exigent exploitation has become an international issue [1]. Therefore, worldwide sustainable development initiatives include actionable goals that align with "the improved water quality through effluent treatment" and the "improved water efficiency through the application of the 5R principles: reduce, reuse, recover, recycle, replenish" [2]. There is a great demand for a variety of wastewater treatment technologies that may not be limited to producing cleaner water only to serve agricultural and industrial needs, but highly purified potable water as well [3–5]. Great emphasis has been placed on making the most of wastewater stocks, which are intended for agricultural irrigation purposes, especially in arid areas [6–8]. There is a universal water scarcity that gradually places restrictions on the amount of available water that may be distributed to irrigate farms; wastewater reuse techniques aspire to offer sustainable solutions [9]. However, wastewater separation processes are dictated by complex

chemical phenomena. There are several engineering concerns that emerge from regulating the removal of heavy metals from the influx stocks, while at the same time optimizing nutrient recovery. A delicate balance is sought, such that irrigation water does not perturb the groundwater quality status [10–12]. To solve the water crisis, numerous promising membrane technologies have been developed [13]. Specifically, electrodialysis (ED) techniques have become indispensable in industrial-level wastewater treatment projects around the world [14,15]. Recycling polluted wastewater and using ED separation cells to produce irrigation water for crop growth could entail multifarious prospects [16]. In general, to improve water treatment processes, it is crucial to conduct successful screening and optimization studies on the operating controls by heeding to pragmatic conditions. A water-quality-screening/optimization task, for a custom-made recycling ED process, is bound to encounter the well-known drawbacks that arise from dealing with environmental stochastic uncertainty in modelling water systems [17,18]. Modern machine learning techniques have been recommended to assist in circumventing difficult theoretical aspects in deciphering uncertainty through statistical data analysis [19–22]. Contemporary improvement problem-solving tactics must embody design of experiments (DOE) techniques that appeal to the fourth industrial revolution premise. A data-centric optimization approach must account for elements of lean manufacturing, quality 4.0 and artificial intelligence [23]. It is profitable that the same mentality that is dominated by expediting innovation is applied in water optimization projects [24]. A recommended pathway to achieve this is by adopting robust engineering techniques that rely on the structured DOE framework [25–27]. Robust DOE techniques and methods have embedded the 'lean-and-green' initiative, upon which sustainable operations are established [28–34]; 'lean-and-green' operations minimize waste and increase production uptime. Furthermore, green sampling is a principled innovative tactic that makes the most of the gleaned information by encouraging: (1) the minimization of the experimental trials, (2) the maximization of the investigated effects at a given research effort, (3) the minimization of chemical material usage and costs and (4) the minimization of environmental product impact [35,36]. Wastewater-recycling quality-improvement research is susceptible to the green and sustainable chemistry stipulation that establishes the necessity for green chemistry metrics to guide data-centric methods in order to reduce waste and increase yield [37–41].

The work by Abou-Shady [16] has demonstrated that it is feasible to conduct lean-and-green experiments on wastewater recycling by using Taguchi's DOE methodology, for robust product design, in a custom-made electrodialysis apparatus [27]. Taguchi methods align to lean-and-green goals in several aspects. The main tactic is the drastic reduction in trials by resorting to fractional factorial designs (FFDs) [25,27]. A second tactic consolidates factorial screening and parameter optimization in a single phase even though they are theoretically of different and distinct scope. A third tactic is that it minimizes the number of factorial settings that are needed to profile and optimize the investigated characteristic responses. A fourth conditional tactic is the designing of experiments without provisions for trial run replications. Abou-Shady [16] showed that only nine pre-determined formulations may be adequate to adjust four ED process factors to decontaminate wastewater to become suitable for crop development in arid areas. The Taguchi-type optimization scheme serves to collect concurrent information on factor strength along with the associated curvature effects. The characteristic guides are three regular water quality indices, which are meaningful in agricultural engineering. From a scientific perspective, the Abou-Shady [16] trials benefited from all mentioned tactics above. The trials were designed to maximize factorial engagement for the allowable runs; the selected nine-run four-factor non-linear orthogonal array (OA) scheme was saturated. Trial recipes were executed once (unreplicated condition). Factorial screening and parameter optimization were conducted in a single phase. The multi-response screening/optimization task was not carried out because it is not available in the ordinary Taguchi methods. Finally, factorial strength and curvature effects were pre-designed by selecting the proper experimental plan. A second complication that had remained to be resolved was the quantification of the hidden

experimental uncertainty. It is a difficult subject and has become a theme of intense research for a long time [42]. The problem of working with unreplicated schemes is not foreign to agricultural research [43–45]. There have been some attempts to exploit the benefit that unreplication offers in cost and time savings, but the data conversion process requires more specialized approaches. Most commercial software packages provide a statistical analysis for unreplicated–saturated FFD-collected datasets through one or more of the three options: (1) the half-normal plot [46], (2) the Box–Meyer method [47] and (3) the Lenth method [48]. Nevertheless, there seem to be disagreements in predictions among methods and software packages [49]. Some prediction disparities may be attributed to deeper theoretical issues, as in the case regarding the use of the half-normal plot [50]. Others may be identified in unavoidably counting on auxiliary 'pseudo-error' estimations, or even in tentative prior probabilities for the active factor group; speculative variance inflation factor estimations may be part of the solver input. Irrespective of the hindrances and bad practices in executing DOE projects [51], DOE is a worthwhile empirical vehicle that drives to better understanding large industrial-level processes because it is couched on simplicity [52,53]. The analysis process is a bit more complicated if a concurrent optimization is expected that implicates several response variables [54]. Chemical qualimetrics have recognized the contribution of multi-factorial screening for several characteristics as well as the influence of the Taguchi methods to systematize laboratory trials in chemometrics [55,56]. Classical analysis of variance (ANOVA) or general linear modeling (GLM) may not be readily deployable techniques in unreplicated–saturated FFD schemes [25,57]. A prevailing notion is to certainly assess predictions not only from the viewpoint of the statistical inferential methods, but also to compare them to solutions that have been obtained from algorithmic routines [58].

The unreplicated–saturated non-linear multi-response ED dataset of Abu-Shady [16] was analyzed using descriptive and graphical Taguchi methods; a 'one-response-at-a-time' analysis is a rudimentary data reduction phase in aquametrics. There were further attempts to add significance to the original experimental outcomes using: (1) a non-parametric multi-factorial technique [59], (2) a combination method of a micro-clustered dimension-reduction with rank learning [60] and (3) a combination method of unsupervised clustering with entropic methods [61]. The proposed work implements a recently published algorithmic classification method, the bionic swarm intelligence (DBSc) method [62], to structured multi-dimensional mini datasets. DBSc has been developed to classify large multi-dimensional datasets. However, the application of DBSc in this work is tested in the other end of the spectrum, i.e., for screening a wastewater recycling ED process by classifying unreplicated–saturated non-linear multifactorial multi-response FFD-sampled datasets. The new feature is that no other post-processing is required besides the usual solution check through internal validation. The advantage of adopting DBSc for clustering mini-data is multifaceted, and it stems from the inner-workings of the algorithm itself: (1) the swarm intelligence searching procedure, (2) optimized targeting in the absence of a definite global objective function, (3) the upgraded non-parametric annealing driver and (4) the Nash equilibrium process objective reachable by game theory and circumstantial symmetry. The specific innovative intervention is that it is posited that the dominant factorial setting sequence(s) mirror the clustered mini-dataset to a degree that it is internally measurable. The matching of the two partitioned sequences, designed runs and clusters is monitored by internal validity measures (Dunn Index [63], Davies–Bouldin Index [64] and Rand Index [65]). Hyper-parameter screening for the fine-tuning of the DBSc routine is performed before completing the wastewater recycling process screening. Hyper-parameter screening [66,67] is a prophylactic step to ensure the optimal operation of the algorithm on the particular ED process mini-dataset. To highlight the innovative aspects of this work, the concept's main features are summarized as follows:

1. The approach attempts to nest the validated cluster solution to the individual factorial recipe patterns.

2. The unsupervised dimension reduction in the examined multiple aquametric indices is independent of individual index performance goals.
3. Self-organizing and emergent influences appear first from intermingling and labeling multiple characteristic responses.
4. Matching validated clusters to factorial recipe patterns circumvents the need to gauge against a statistical reference law.
5. The technique is operable on saturated–unreplicated multi-response multi-factorial mini-datasets, in spite of standard aqua-qualimetrical treatments failing to return a significance estimate.
6. Easy-to-follow and trusted analysis steps are used because openly-sourced visual tools and state-of-the-art artificial intelligence R-packages are deployed.

Extensive internal cluster validity comparisons are provided. Openly available DOE and clustering tools are implemented on the freeware platform R for faster and more reliable data-centric decision making [68–70]. The outcomes are discussed against those results from previous published research.

2. Materials and Methods

2.1. Double Orthogonal Screening for Vital Factors and Solver Hyper-Parameters

A regular OA trial planner systematizes the progress of an accelerated product/process screening/optimization study by compressing the total volume of experiments through the balanced 'tight-fitting' of the scheduled factor setting combinations [25,27]. An OA sampler is a prescribed matrix that standardizes the minimum number (n) of the formulated trial recipes for a group of as many as m investigated effects. The sampling tactics in this work require two types of related screenings: (1) a saturated non-linear OA trial planner to profile the hierarchical influence of the examined factors and (2) a mixed-structure OA sampler to identify favorable hyper-parameter adjustments for the databionic algorithmic solver [62]. The saturated non-linear OA trial planner is aimed to organize the experimental runs in such a manner as to deliver simultaneous information on the potency status effect and on the curvature tendencies for each of the investigated effects separately. Therefore, an OA-based experimentation tactic speeds up the data-centric decision-making process while demanding minimal research effort. During such a brief endeavor, the high-density dataset configuration may also include a zero-replication regimen; it is the condition of unreplication.

The non-linear minimal-data OA scheme for exemplifying the Taguchi $L_9(3^4)$ matrix test case is studied in its 'full-utilization' pattern, i.e., by exploiting the maximum input assignment of four controlling factors [26,27]. This plan produces a total of nine data entries per recorded characteristic response. Thus, the data requirements are maintained to a minimum, because of the concurrent conditions of trial unreplication and factor saturation. In its general representation with coded factor settings, the original Taguchi-type $L_9(3^4)$ OA, which is highlighted in this work, is shown in Table 1.

Table 1. The coded three-level four-factor Taguchi-type $L_9(3^4)$ OA.

Run #	Factor A	Factor B	Factor C	Factor D
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

On the other hand, effectively implementing an unsupervised solver on cluster-partitioning tasks inherently requires fine-tuning feedback from its principal hyper-parameter gamut [66,67]. Consequently, an $L_8(4^1 \times 2^2)$ OA was re-arranged to probe the stability of the implemented databionic swarm intelligence analyzer on explaining the particular $L_9(3^4)$ OA mini-dataset [26,27]. The proposed $L_8(4^1 \times 2^2)$ OA scheme, which mixes three categorical hyper-parameter variables, each available in nominal scaling, is shown in Table 2; the hyper-parameter settings are listed in coded form.

Table 2. The coded mixed-level Taguchi-type $L_8(4^1 \times 2^2)$ OA for algorithmic hyper-parameter screening.

Run #	Hyper-Parameter A	Hyper-Parameter B	Hyper-Parameter C
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1
5	3	2	1
6	3	1	2
7	4	2	2
8	4	1	1

Nevertheless, there is an intrinsic relationship between the two distinct OA-based screenings. The controlling fractional factorial $L_9(3^4)$ OA generated the multi-response mini-dataset. The unsupervised clustering was firstly applied on the OA mini-dataset, and then the cluster-labelling results from the hyper-parameter (mixed OA) screening were obtained. Finally, the algorithmic stability of the clustered data predictions was internally checked on the structured and pre-labeled factorial (columns) vectors of the generically coded $L_9(3^4)$ OA scheme. This stability needs to be individually checked across all m controlling factors. Thus, the overall concept is naive and easy to comprehend. The examined controlling factors are labeled as: $\mathbf{X_j}$ for $1 \leq j \leq m$ ($m \in N$), and their respective factor settings are correspondingly coded as x_{ij} ($x_{ij} \in N$) for $1 \leq i \leq n$ ($n \in N$) and $1 \leq j \leq m$. Particularly, a non-linear OA scheme is pre-formed with k_j levels for the jth factor ($1 \leq j \leq m$) and $3 \leq k_j \leq K_j$ ($K_j \in N$), as shown in Table 3. Hence, the non-linear OA matrix is an input ordered array that produces an output matrix $\{r_{ic}\}$ with $\mathbf{R_c}$ multiple characteristic responses, with $1 \leq i \leq n$ and $1 \leq c \leq L$ ($L \in N$); each c^{th} matrix column is a single response vector.

Table 3. A general arrangement of an OA matrix, and its cluster-labelled output.

Controlling Factors (Input)						Multiple Characteristics (Output)						Labelled Clusters	
run #	X_1	X_2	.	.	X_m	***run #***	R_1	R_2	.	.	R_L	***run #***	$\mathbf{I_d}$
1	x_{11}	x_{12}	.	.	x_{1m}	**1**	r_{11}	r_{12}	.	.	r_{1L}	**1**	l_1
2	x_{21}	x_{22}	.	.	x_{2m}	**2**	r_{21}	r_{22}	.	.	r_{2L}	**2**	l_2
.	.	.			.	.	.	.			.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.			.	.	.	.			.	.	.
n	x_{n1}	x_{n2}	.	.	x_{nm}	***n***	r_{n1}	r_{n2}	.	.	r_{nL}	***n***	l_n

The mini-dataset is converted to a cluster vector, $\mathbf{I_d}$, which is labelled into a total number of Z clusters with cluster members: $l_i \mid 1 \leq l_i \leq Z$ ($Z \in N$) and $1 \leq i \leq n$. Since the $\mathbf{X_j}$ vectors in an OA provide probable 'pre-clustering' information through the locations of their pre-defined factor settings, they may be directly tested individually for accuracy in comparison to the final cluster identification vector $\mathbf{I_d}$. The hyper-parameters that favor accuracy amelioration are retained to permit the prediction of the final solution.

2.2. The Databionic Swarm Intelligence Classifier

The orthogonal multi-dimensional $L_9(3^4)$ OA mini-dataset is morphed into a single cluster identification vector by employing the databionic swarm intelligence classifier (DBSc) [62]. The DBSc data conversion engine is benefited from several attractive built-in

processor features, which may also promote the effective cluster member fingerprinting of high-density mini-datasets. Strictly speaking, the DBSc is geared toward cluster analysis for large high-dimensional datasets. Notably, the DBSc-extracted information is assembled from three conspicuous nature-inspired agents such as: (1) emergence, (2) self-organization and (3) swarm intelligence. Theoretically, DBSc may also be applicable, in a broader sense, to non-linear OA-apportioned multi-response mini-datasets. Two particularly intriguing traits of the DBSc data-conversion engine render its adoption highly desirable to OA-generated multi-response multifactorial mini-datasets: (1) the swarm intelligence searching procedure, which seeks optimized target states in the absence of an established global objective function and (2) the upgraded non-parametric annealing driver, which is astutely impelled toward a Nash equilibrium, and hence is guided by solid game theory and symmetry considerations. DBSc has been designated for this study, instead of other more familiar clustering methods, because it is the only unsupervised classifier that manages to fingerprint previously uncharacterized data. Briefly, the DBSc classifier comprises three sequential co-processors: (1) a parameter-free projection engine, (2) a parameter-free high-dimensional data visualization facility and (3) a zero-sensitivity cluster membership identifier. However, it is reiterated that the important caveat in this intended application is attributed to the inherent 'data smallness'. Unequivocally, the classifier visualization phase may not be responsive to the generated 'small-and-dense' OA dataset. The topographic map, which is an essential component of the DBSc solver, may be restricted in providing enough databot details that visually ascertain the aptness standing of the final cluster formations. Irrespective of the extent of the usability of the topographic map in clustering OA data, the DBSc engine advances to recommend a dendrogram and a final identification prediction of the cluster memberships. Probing the stability of the DBSc routine becomes advisable under such previously unexplored conditions. Therefore, three hyper-parameters reasonably emerged in congruence to modulating the cluster prediction accuracy. The three DBSc hyper-parameters were: (1) the type of the distance measure, (2) the structure type of clusters and (3) the position matrix preference. Since this is a first examination on the subject, the types of distance measures were limited to a group of ordinary candidate models. They were mainly chosen because the computing of their non-metric multi-dimensional scaling values was available by popular freeware modules in data science. For demonstrating purposes, the well-known Shepard–Kruskal approach was employed to furnish information on dissimilarity trends with the Shepard diagram [71] and to assist in differentiating between two nominated cluster numbers; their Kruskal stresses [72] were also computed and compared. With a lack of prior studies, it is worthwhile to start exploring how five of the most widely implemented continuous distance measure models impact the OA mini-data dissimilarity estimations: (1) the Euclidean distance, (2) the Maximum/Chebyshev distance, (3) the Manhattan distance, (4) the Canberra distance and (5) the Minkowski distance. After a pre-screening, an attempt was made to shorten the initial list of distance measure models. The four proposed distance measure models for the hyper-parameter screening were (Table 4): (1) the Euclidean distance, (2) the Maximum/Chebyshev distance, (3) the Manhattan distance and (4) the Canberra distance. Moreover, 'compact' and 'connected' structure types of clusters were altered in running the DBSc module [73]. Finally, the position options were set as either 'Projected Points' (automatic clustering projection from the Pswarm algorithm) or as 'Best Matches' (involving the 'GeneralizedUmatrix' module).

Table 4. The Taguchi-type $L_8(4^1 \times 2^2)$ OA arrangement for algorithmic hyper-parameter screening.

Run #	Distance Measure	Structure Type	Position Type
1	Euclidean	Connected	Best Matches
2	Euclidean	Compact	Projected Points
3	Maximum	Connected	Projected Points
4	Maximum	Compact	Best Matches
5	Manhattan	Compact	Best Matches
6	Manhattan	Connected	Projected Points
7	Canberra	Compact	Projected Points
8	Canberra	Connected	Best Matches

2.3. The Water Quality Case Study

To elucidate databionic double-screening on a crucial data-centric application, a published wastewater treatment optimization study was revisited [16]. The great importance of the selected case study rests on several key aspects: (1) a newly designed electrodialysis apparatus was launched, (2) the intended purpose of the research was to improve the recycling process of wastewater stocks in order to supply arid/semi-arid areas with agricultural irrigation water, (3) the ED cell dialyzed a great variety of polluted water resources from different locations and origins, (4) the wastewater properties could not be assumed homogeneous, owing to the underlying complex chemical systems, (5) a structured non-linear orthogonal mini-dataset was planned and gathered, (6) a triplet of water quality indices were to be concurrently screened/optimized, (7) the Taguchi-type OA sampler generated a one-time dataset to satisfy the synchronous screening and optimization tasks, (8) the optimized water quality indices were to achieve agricultural conformity levels comparable to those necessitated to crop nourishment, yield and safety, (9) the constraints of trial unreplication and factor saturation were present in the experimental plan, (10) the recommended Taguchi-type SNR data conversion was not applicable due to the unreplication constraint, (11) the multi-response dataset was in percentage or ratio form, i.e., dimensionless and bounded on both ends of the scale, (12) the remaining unexplainable experimental error was not feasible to be recovered from standard analysis methods and (13) the overall unique nature of the application.

The trial-planning scheme was a four-factor three-level nine-run $L_9(3^4)$ OA, and it was executed once. The implemented OA sampler was fully exploited and was thus saturated with the following four controlling factors: (1) A: the dilute flow; (2) B: the cathode flow; (3) C: the anode flow; and (4) D: the voltage rate [16]. The dataset consisted of the responses of the three water quality indices: (1) RS: the removed sodium content (%); (2) SAR: the sodium adsorption ratio; and (3) SSP: the soluble sodium percentage (%). There are previous comments on the investigated water quality indices and their pivotal value in the cultivation management of productive crops [60]. The dataset is described in ref. [16]; the entire input–output arrangement is in conjunction to Tables 3 and 4, respectively.

2.4. The Methodological Outline

The proposed methodology may be recapitulated as follows:

(1) Determine the relevant water quality characteristics that represent the wastewater recycling efficiency performance with respect to the specific agricultural application.
(2) Select a group of ED process controlling factors.
(3) Determine the minimum group of factor settings that covers the operational range and allows for the detection of potential characteristic non-linearities.
(4) Program lean trials by implementing the proper OA-sampling scheme that accommodates non-linear trend detection.
(5) Execute the prescribed OA runs (step 4), and gather the multi-characteristic mini-dataset.
(6) Pre-screen each characteristic response using visual information from the boxplot (original/notched) [74], the adjusted boxplot [75], the violin plot [76] and the bean plot [77].

(7) Determine the number of candidate (mini-dataset) clusters by pre-screening the available distance measures using the Shepard plot trends and the Kruskal stress estimations.
(8) Shortlist the distance measures from step 7 if there is evidence that their dissimilarity estimations are not as competitive as the alternative candidate models.
(9) Apply OA screening on the algorithmic hyper-parameters of the databionic swarm intelligence analyzer [62,73].
(10) Assess the best partitioning performance of the clustered mini-dataset against the original controlling factorial OA pre-labeled settings (individual testing of the $\mathbf{X_j}$ vectors and the final cluster identification vector $\mathbf{I_d}$).
(11) Obtain the cluster dendrogram and the solver-labelled clusters of the mini-OA-dataset.
(12) Evaluate cluster similarity using the Rand Index [65] across different hyper-parameter screenings; the proximity of the clustering results to the controlling factorial OA pre-settings is facilitated by the estimation of, respectively, the Dunn Index [63] and the Davies–Bouldin Index [64].
(13) Provide a hyper-parametrized hierarchy for the profiled factorial landscape.

2.5. The Computational Aids

All required computational work was conducted on the free statistical analysis software platform R (v. 4.1.3) [69]. The non-linear $L_9(3^4)$ OA array was constructed using the module 'param.design()' from the R-package 'DoE.base' (v. 1.2). Similarly, the L_8 ($4^2 \times 2^2$) OA was prepared to setup and carry out the hyper-parameter screening process; it is an assessment of the DBSc-based cluster membership accuracy against a pre-selected group of distance measures, the DBSc projection procedures and the structure type of clusters. The module 'boxplot()' was used from the R-package 'graphics' (v. 4.1.3) to obtain regular and notched visuals for the summary statistics of the three water quality characteristics. The module 'adjbox()' in the R-package 'robustbase' (v. 0.93-9) provided the corresponding adjusted boxplot depictions of the three responses. To prepare the customized violin and bean plots, the corresponding R-packages 'vioplot()' (v. 0.3.7) and 'beanplot()' (v. 1.2) were employed. Moreover, to assemble the recommended assortment of Shepard graphs in order to pre-assess the available options for the distance measures ('euclidean', 'maximum', 'manhattan', 'canberra' or 'Minkowski'), the modules 'isoMDS()' and 'Shepard()' were utilized from the R-package 'MASS' (v. 7.3-55); two k-dimensional cluster configurations were obtained from their generated (non-metric multi-dimensionally scaled) distance matrices, along with their respective computed Kruskal stresses. The self-organized clustering of the three water-quality-index mini-datasets was accomplished through the implementation of the swarm intelligence machinery, the R-package 'DatabionicSwarm' (v. 1.1.5) [73]. Furthermore, it furnished the 'ClusteringAccuracy()' and the 'DelaunayClassificatioError()' modules in order to estimate the normalized aggregation of all fingerprinted true-positive data points in addition to finding the first k-nearest Delaunay neighbors in the input space, and subsequently in order to evaluate them in the output space. The dendrogram visualization and the labelled cluster list was generated from the module 'DBSclustering()'. To coordinate the swarm-based databot projections, which were also required in half of the hyper-parameter screening runs, the module 'Pswarm()' was executed to complete the 2D reduction (polar) mapping. The parallel distance matrix computation using multiple threads was facilitated by the R-package 'parallelDist' (v. 0.2.6). The R-package 'GeneralizedUmatrix' (v. 1.2.2) supplied the 'Bestmatches' positions as the alternative to the 'ProjectedPoints', which were estimated by intermediate distance matrix processing using the 'GeneratePswarmVisualization()' module. The validation status of the cluster partitioning performance of the water quality index mini-dataset against the structured factorial OA formulations was conducted by estimating the Rand Index, the Dunn Index and the Davies–Bouldin Index, applying the modules accordingly: (1) 'rand.index()' (R-package 'fossil' (v. 0.4.0)), (2) 'dunn()' (R-package 'clValid' (v. 0.7)) and 'index.DB()' (R-package 'clusterSim' (v. 0.49-2)).

3. Results

3.1. Graphical Pre-Screening of the Three Water Quality Indices

The data pre-screening starts with the portrayals of the responses of the three water quality indices with respect to the three adaptations of boxplots: (1) the original, (2) the notched and (3) the adjusted. It can be immediately noticed from Figure 1 that the original and the adjusted versions of the boxplot do not agree with the removed sodium content (RS) dataset. The lone outlier point appears to be situated on opposite sides in the two graphs. Inspecting the adjusted boxplot suggests that the RS dataset is rather asymmetrical, and it may be worth exploring further with other techniques. This is focal because impending data non-normality may render outlier detection arbitrary. The original and adjusted boxplots for the sodium adsorption ratio (SAR) and the soluble sodium percentage (SSP), when contrasted, exhibit matching behavior in both cases. Moreover, from Figure 1, it is also evident that the (original boxplot) interquartile ranges for all three water quality indices coincide with the 95% confidence interval for the medians (notched boxplots). This is perceived as a substantial variation in locating a robust central tendency for all three examined water quality indices, and it is usually attributed to the small sample limitation. In Figure 2, the original boxplot, the violin plot and the bean plot provide additional comparative information for the response dispositions. The violin and the bean plots for the RS dataset seem to affirm the hint for the presence of an asymmetrical spread in the original/adjusted boxplots; the lower values of the data distribution may perhaps uphold a wider tail formation. Individual outliers are not expected to be detectible in the violin plots. The adjusted boxplot and the bean plot agree that the outlier point is on the lower end of the distribution. Multimodality for this RS dataset may not be ruled out according to its profile in the corresponding bean plot. In comparison, the estimated kernel density trends in the violin plot of the SAR dataset reveal a skewness that is barely visible in the original boxplot. It is, rather, credited to a higher concentration of datapoints below the mean value in the bean plot. Finally, the respective violin and bean plots for the SSP dataset depict a symmetric spread around the location point, in agreement to the two boxplots; the nature of the distribution tails [78] still remains not transparent.

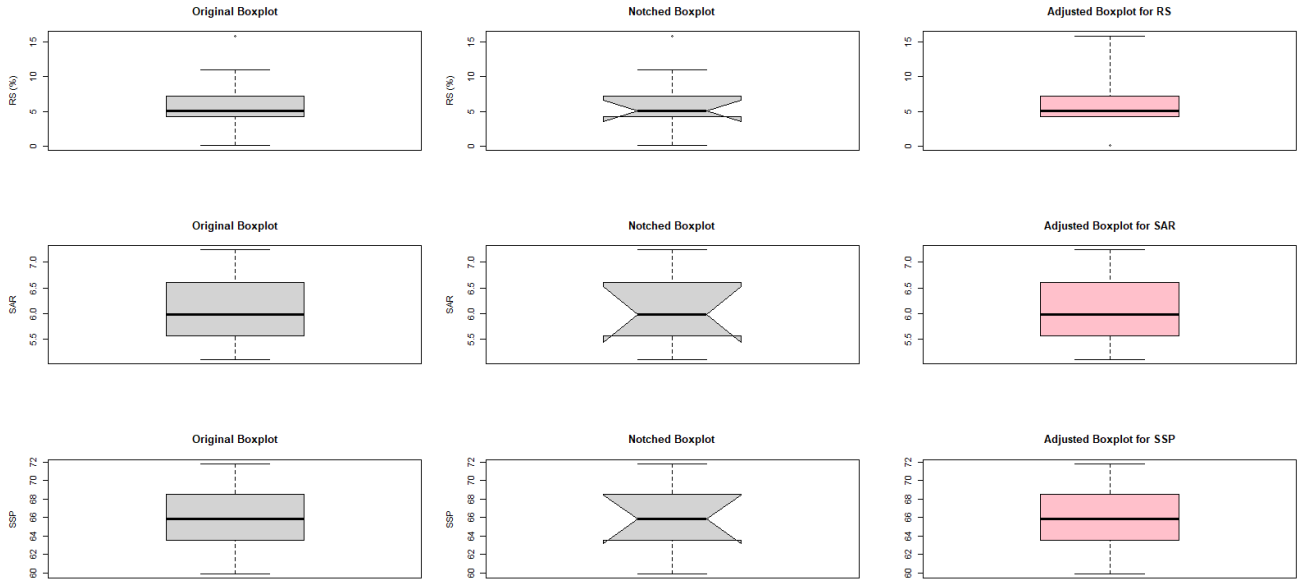


Figure 1. Original, notched and adjusted boxplots for the three water quality indices: (1) RS (**first row**), (2) SAR (**second row**) and (3) SSP (**third row**).

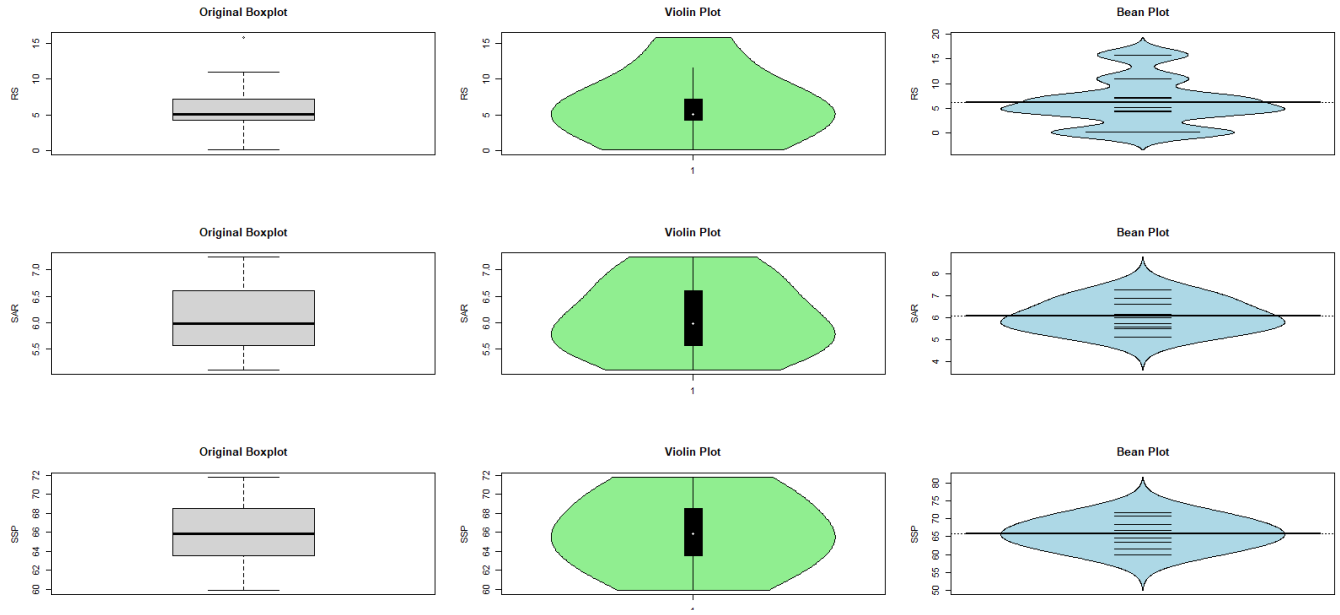


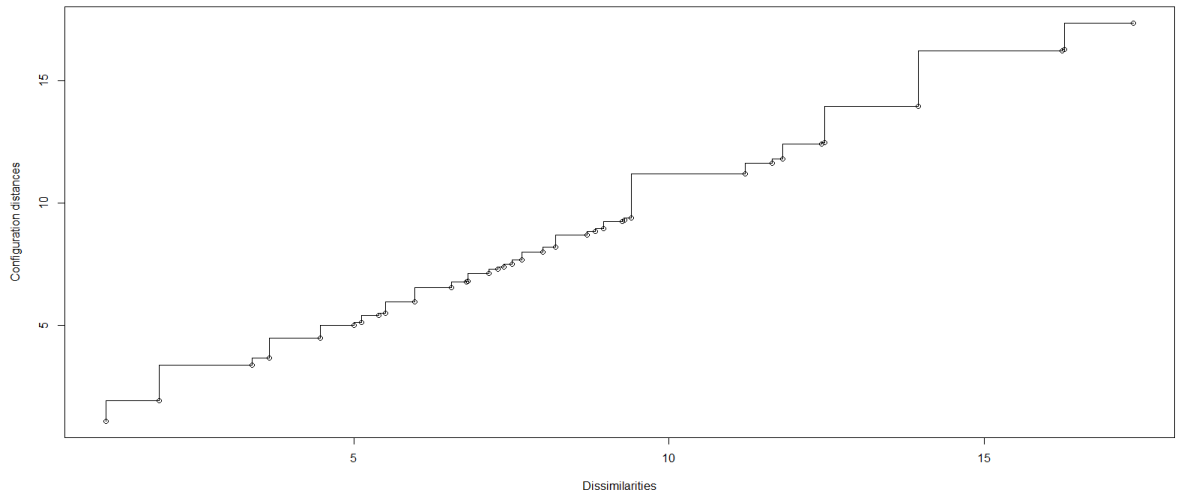
Figure 2. Boxplots, violin plots and bean plots for the three water quality indices: (1) RS (**first row**), (2) SAR (**second row**) and (3) SSP (**third row**).

3.2. Graphical Pre-Screening of the Candidate Distance Measures

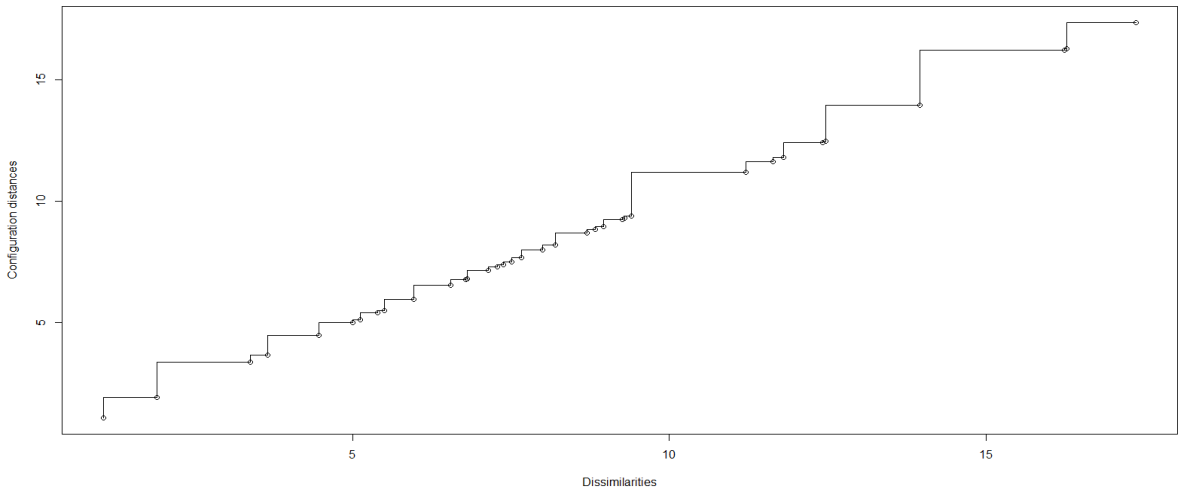
Choosing a suitable distance measure for the Databionic Swarm classifier is simplified by a naïve exploration of the group of the nominated distance measures in two cluster number settings of relevance. Since the multi-response dataset was programmed to deliver information at three preset adjustments, for each probed factor, the dominant factor(s) are netted in three-cluster configurations if they are indeed non-linear, or otherwise in two-cluster reduced form if they behave linearly. Therefore, cluster number settings of 2 and 3 were decided to be tested. Moreover, the candidate distance measures were indicatively selected: (1) the Euclidean distance, (2) the Maximum/Chebyshev distance, (3) the Manhattan distance, (4) the Canberra distance and (5) the Minkowski distance. For the Minkowski distance measures, the model parameter p was tested at values 2 and 4, respectively. The measure parameter p set at a value of two coincides with the Euclidean distance measure definition. Figure 3 provides visual differences for the tendencies of the configuration distances against their similarities for distance measures and both cluster numbers using the illuminating Shepard graphs. It is immediately obvious that the Euclidean distance performs more optimally than the rest of the distance models, and it does not discriminate between cluster numbers. As expected, the same behavior is observed for the Minkowski model set at $p = 2$. However, for $p = 4$, the similarity-tracking tendencies dramatically deteriorate. A general observation is that the measure-based similarity efficiency favors the three-cluster option. The same conclusion is reached by computing the corresponding Kruskal stresses for the 12 plots in Figure 3. In Table 5, it becomes evident that the Euclidean distance measure (Minkowski model at $p = 2$) drastically minimizes the Kruskal stresses. Stepping down from a three-cluster to two-cluster model, it is accompanied by a 35% increase in the magnitude of its Kruskal stress prediction. The three-cluster partitioning consistently minimizes the Kruskal stress estimations in all examined distance measure models. There is a significant difference in the Kruskal stress estimations using the Minkowski model set at $p = 2$ and $p = 4$.

Table 5. Kruskal stress pre-screening for the distance measures at two cluster number settings.

Distance Measure	Cluster Number	Kruskal Stress Value
Euclidean	2	7.35×10^{-14}
Euclidean	3	5.44×10^{-14}
Maximum	2	2.28
Maximum	3	0.72
Manhattan	2	1.96
Manhattan	3	5.73×10^{-3}
Canberra	2	8.67×10^{-3}
Canberra	3	7.31×10^{-3}
Minkowski ($p = 2$)	2	7.35×10^{-14}
Minkowski ($p = 2$)	3	5.44×10^{-14}
Minkowski ($p = 4$)	2	6.75×10^{-3}
Minkowski ($p = 4$)	3	2.18×10^{-3}

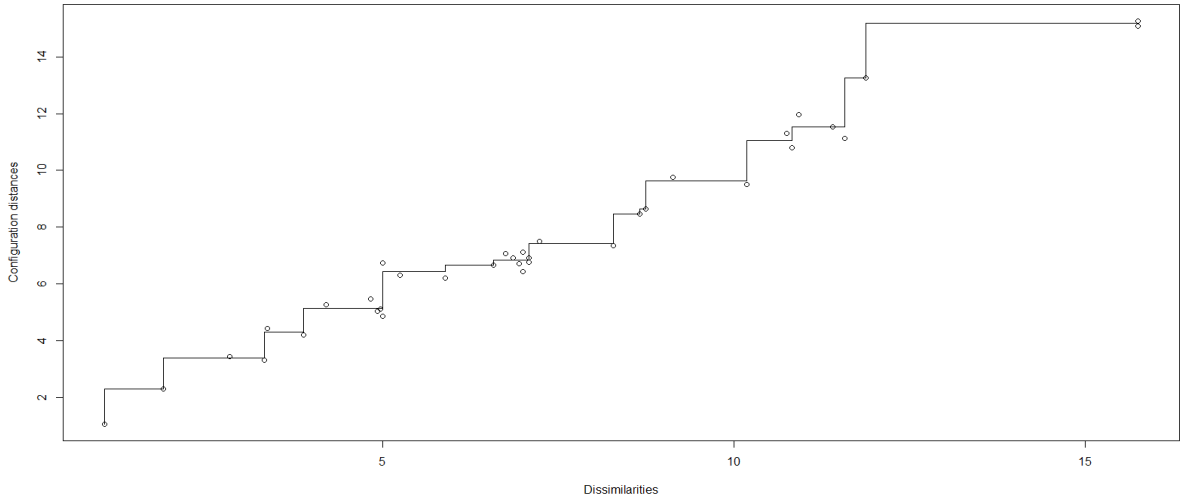


(**A**)

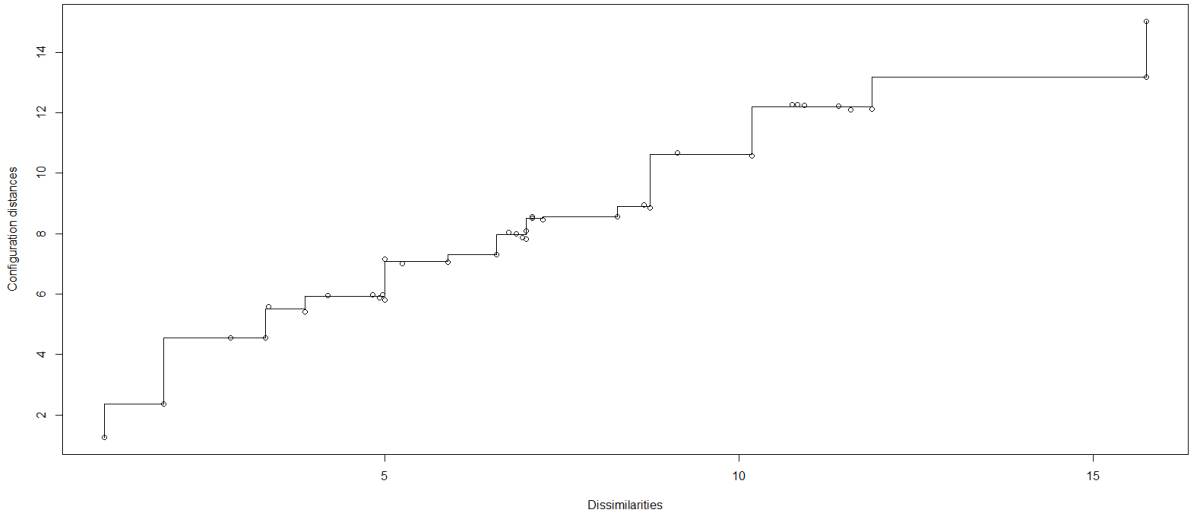


(**B**)

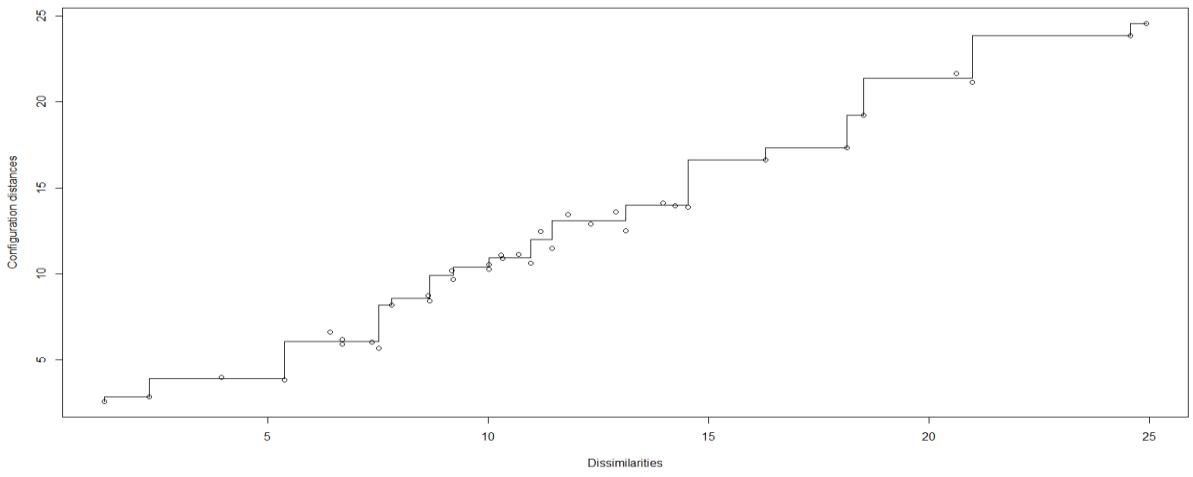
Figure 3. *Cont.*



(**C**)

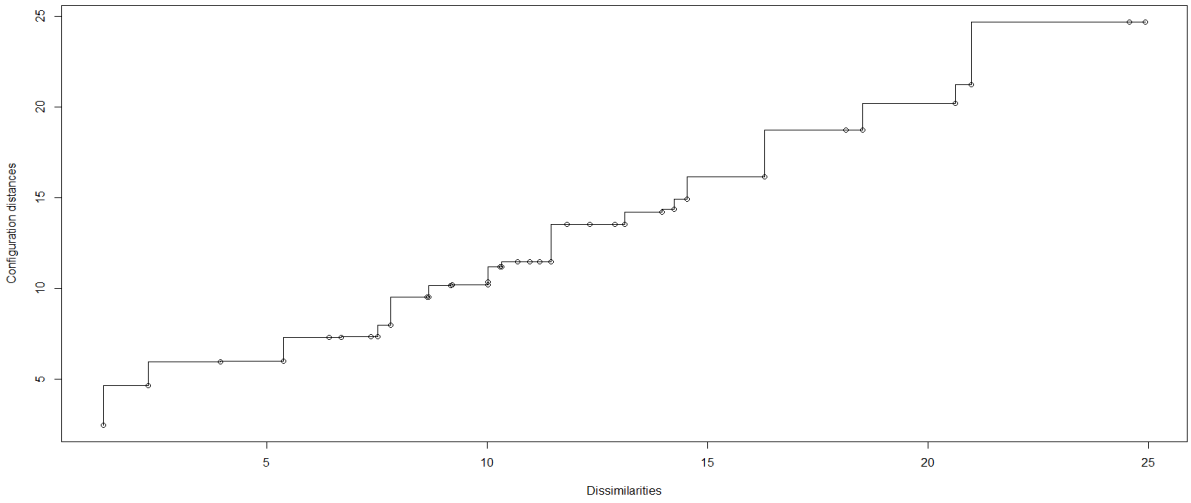


(**D**)

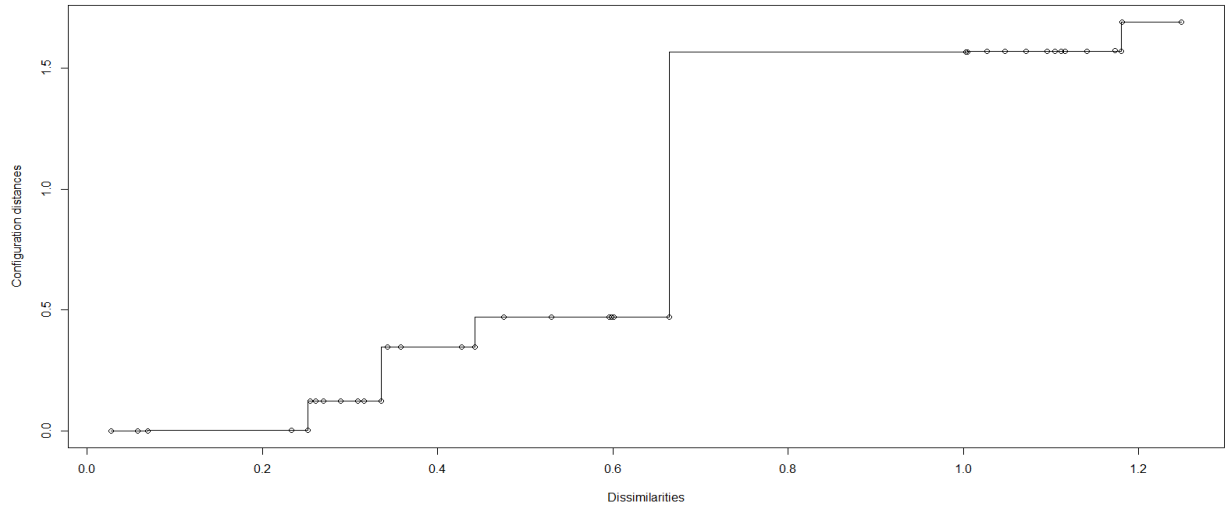


(**E**)

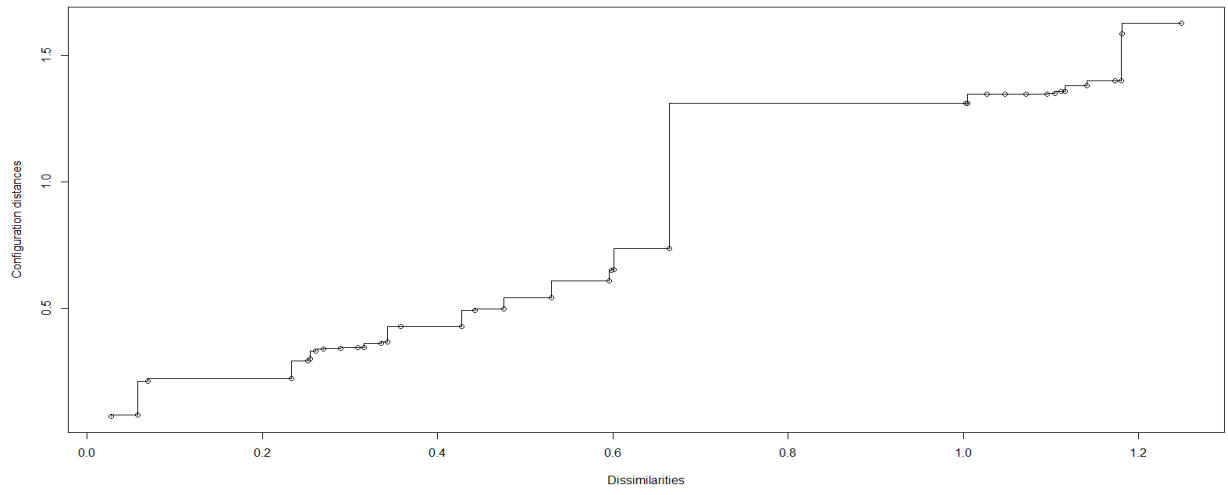
Figure 3. *Cont.*



(**F**)

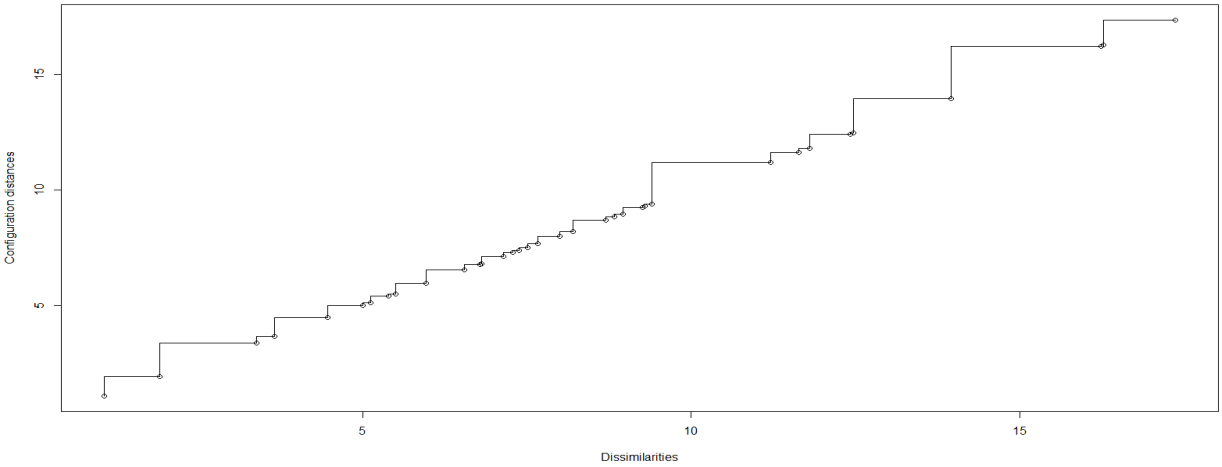


(**G**)

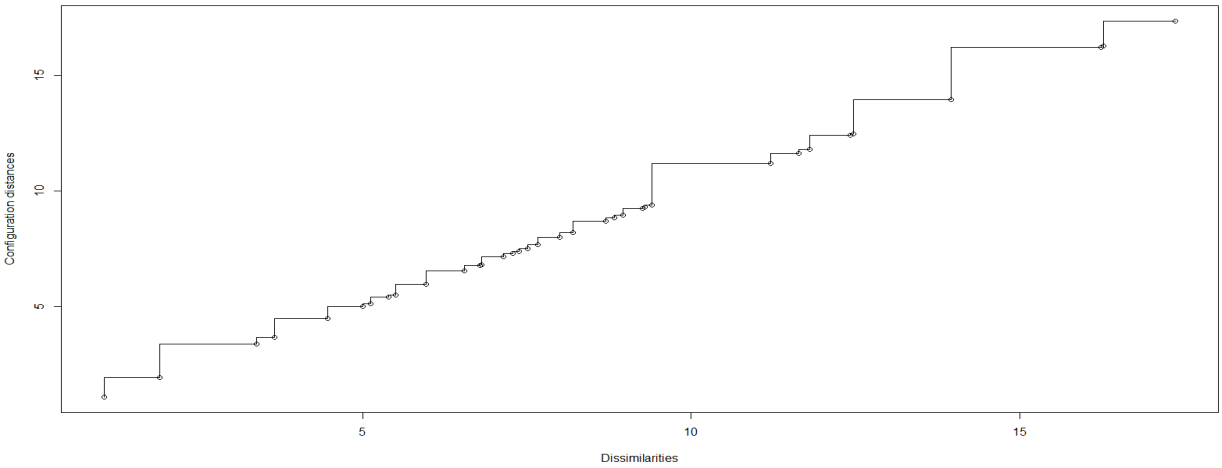


(**H**)

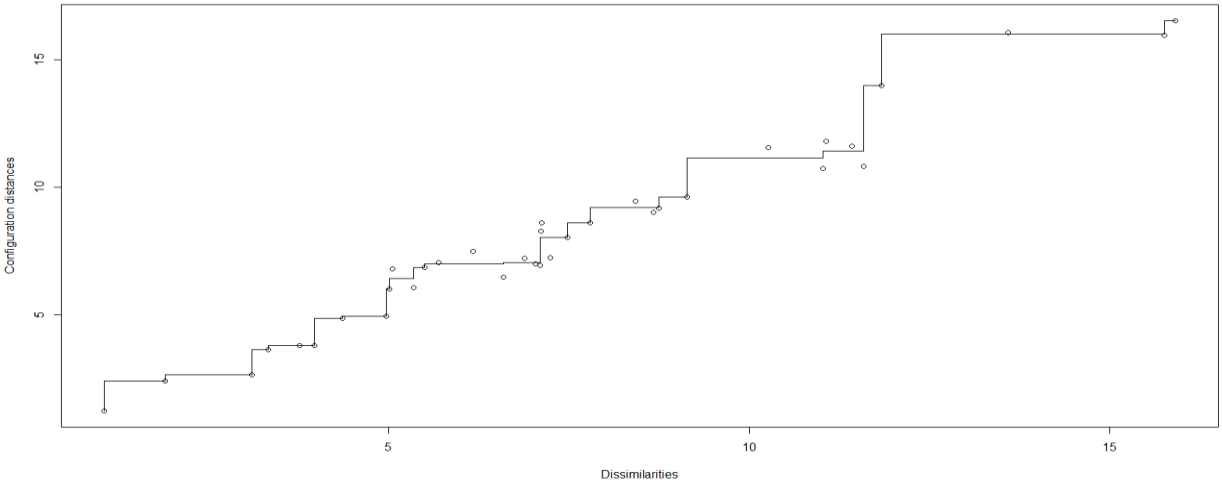
Figure 3. *Cont.*



(I)

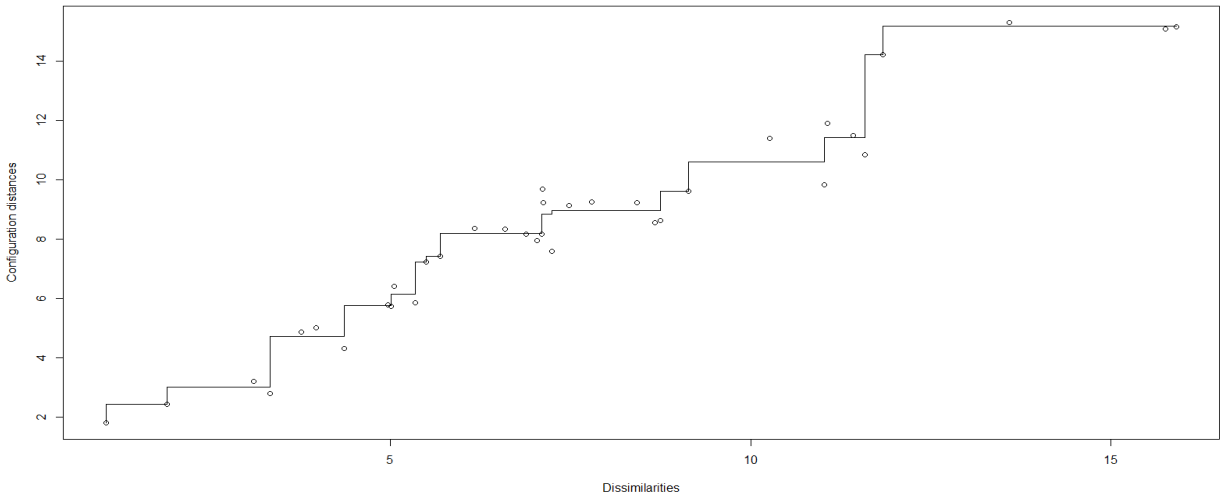


(J)



(K)

Figure 3. *Cont.*



(**L**)

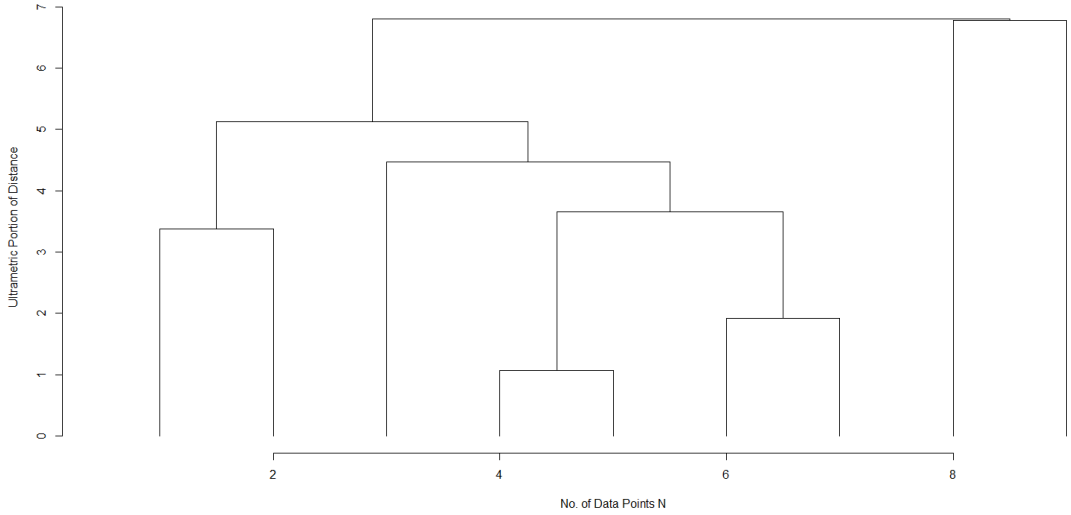
Figure 3. Shepard graphs of the three water quality index datasets with respect to the five measure types at the two cluster number settings, k: (1) Euclidean distance ((**A**): k = 2, (**B**): k = 3), (2) Maximum distance ((**C**): k = 2, (**D**): k = 3), (3) Manhattan distance ((**E**): k = 2, (**F**): k = 3), (4) Canberra distance ((**G**): k = 2, (**H**): k = 3), (5) Minkowski distance for $p = 2$ (**I**): k = 2, (**J**): k = 3) and (6) Minkowski distance for $p = 4$ ((**K**): k = 2, (**L**): k = 3).

3.3. Unsupervised Multi-Response Screening Using the Databionic Classifiier

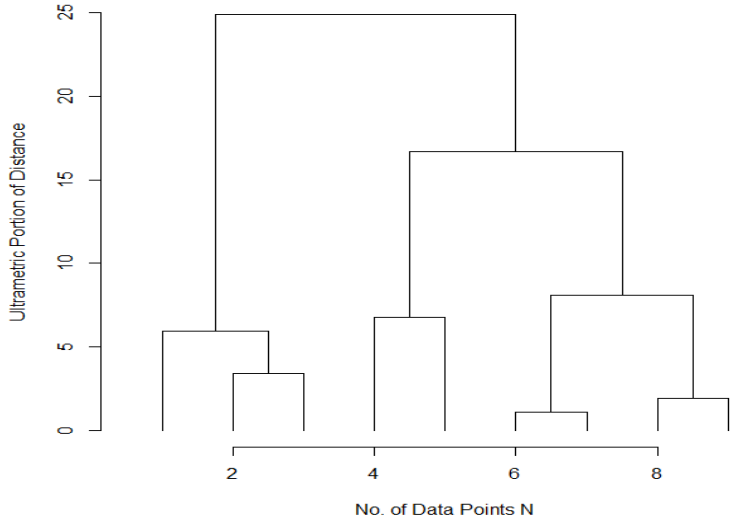
Based on the pre-screening information of Sections 3.1 and 3.2, it was decided to examine in more depth the behavior of the four distance measures: (1) the Euclidean distance, (2) the Maximum/Chebyshev distance, (3) the Manhattan distance and (4) the Canberra distance in the unsupervised multi-response screening output using the Databionic classifier. A three-cluster-partition setup was maintained constant in all trial runs. To investigate the influence of the three proposed hyper-parameters on the grouping accuracy of the DBSc, the $L_8(4^1 \times 2^2)$ OA arrangement (Table 4) was conducted. The eight resulting dendrograms are shown in Figure 4. The screening process generated grouping patterns that follow different pairing pathways for different hyper-parameter combinations. However, from Table 6, out of the eight executed recipes, only three outcomes are distinct. Six outcomes produced the same membership groupings (runs #3–#8). The outcomes from runs #1 and #2 differentiate between them in four membership identifications. The sequence outcome from run #2 does not agree in only one position with respect to the rest of the runs (runs #3–#8).

Table 6. Cluster membership identifications from screening the three DBSc hyper-parameters.

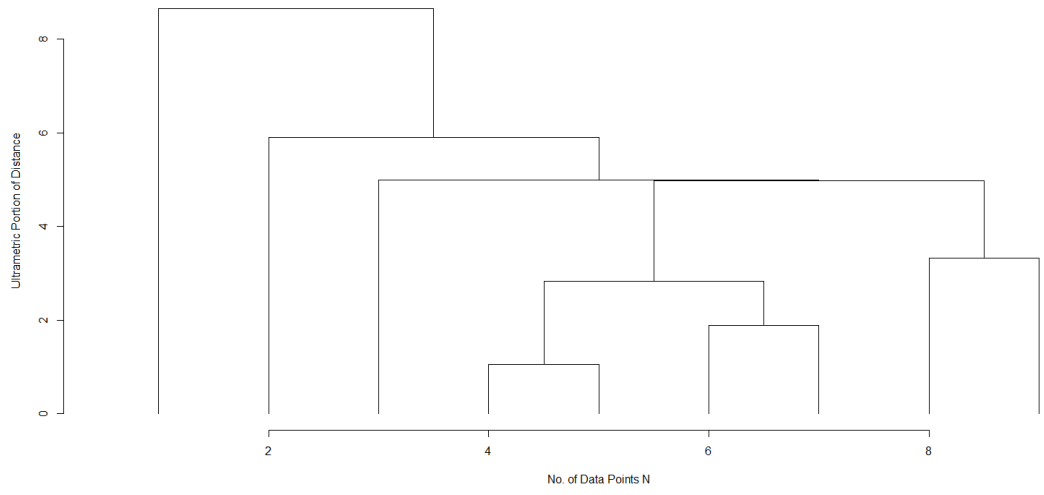
	Cluster Member Identification								
Hyper-Parameter Screening Run #	**Run #1**	**Run #2**	**Run #3**	**Run #4**	**Run #5**	**Run #6**	**Run #7**	**Run #8**	**Run #9**
1	1	1	2	1	3	1	1	1	1
2	1	1	2	1	2	1	3	3	3
3	1	1	2	1	2	1	3	1	3
4	1	1	2	1	2	1	3	1	3
5	1	1	2	1	2	1	3	1	3
6	1	1	2	1	2	1	3	1	3
7	1	1	2	1	2	1	3	1	3
8	1	1	2	1	2	1	3	1	3



(**A**)

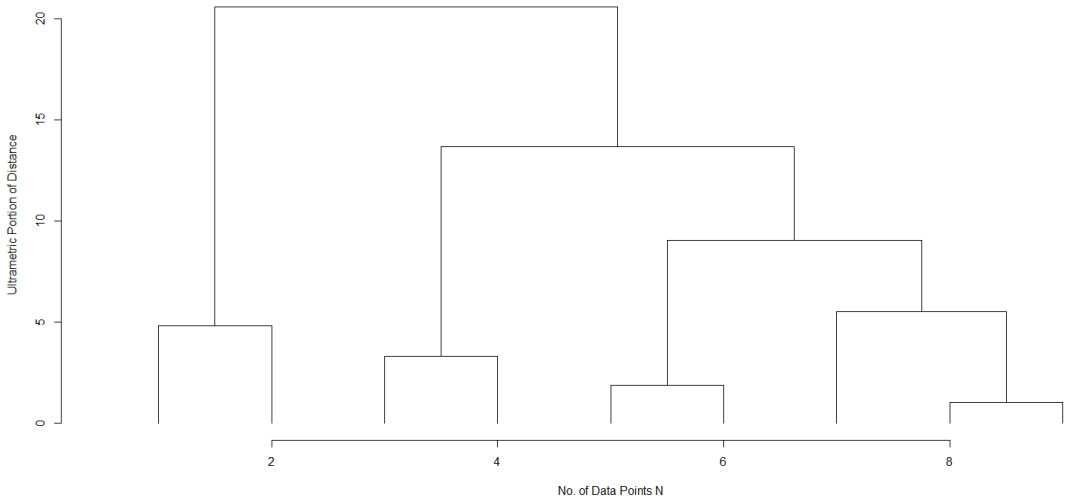


(**B**)

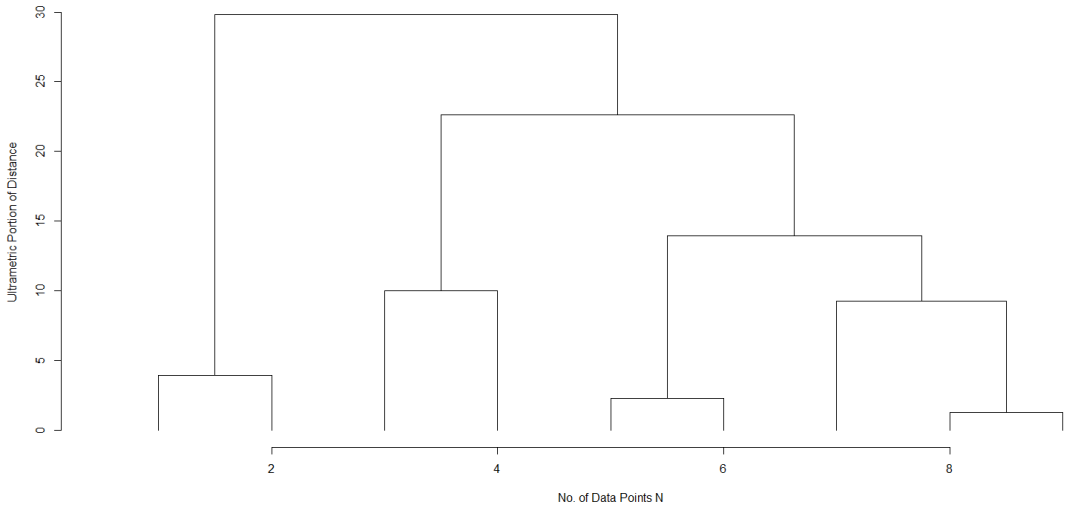


(**C**)

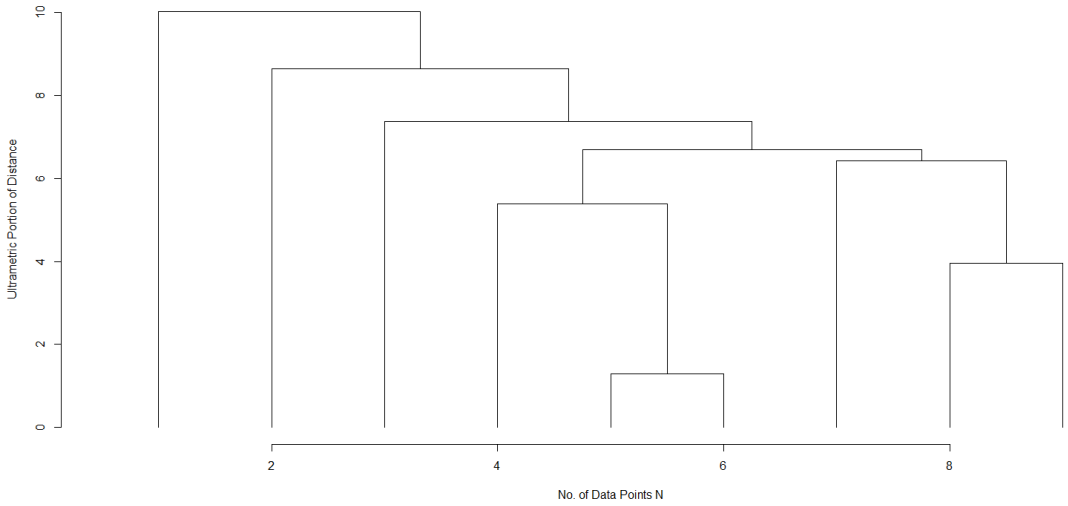
Figure 4. *Cont.*



(**D**)

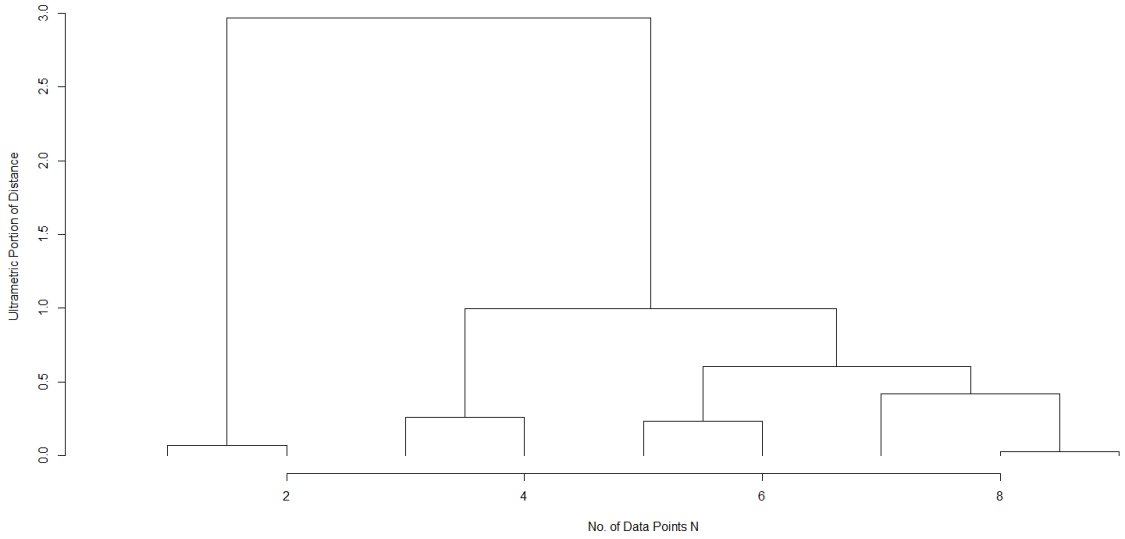


(**E**)

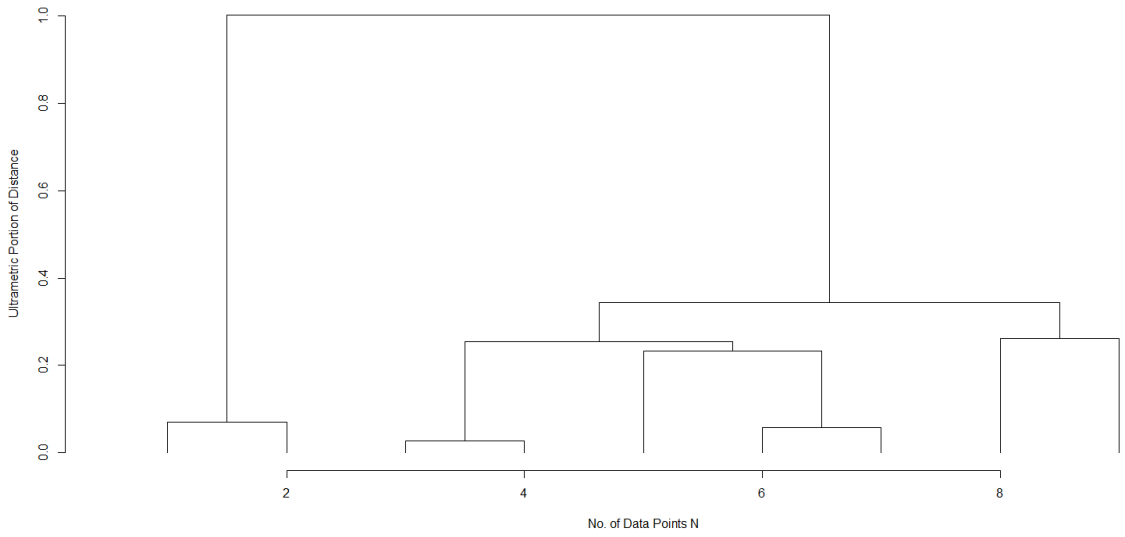


(**F**)

Figure 4. *Cont.*



(**G**)



(**H**)

Figure 4. The eight dendrogram outputs from the hyper-parameter OA screening (Table 4). (**A**) Run #1, (**B**) Run #2, (**C**) Run #3, (**D**) Run #4, (**E**) Run #5, (**F**) Run #6, (**G**) Run #7 and (**H**) Run #8.

Using the DBSc's 'ClusteringAccuracy()' function, the clustering result sequence was contrasted against each of the four factorial setting sequences as prescribed in $L_9(3^4)$ OA. A radar graph (Figure 5) summarizes all this information, from which there is no clear indication that some factor prevails over the others in affecting cluster membership predictions for a particular hyper-parameter combination. It is noted that the radar graph encompasses synchronous information from both experimental designs, i.e., from $L_9(3^4)$ OA and $L_8(4^1 \times 2^2)$ OA schemes. Similarly, a radar graph (Figure 6) was prepared where the Rand Index evaluation of the eight hyper-parameter screening clustering solutions are laid out for contrasting against the $L_9(3^4)$ OA factorial setting sequences. Again, it is not conclusive that one particular factor dominates the rest of the group. However, a bar chart (Figure 7) of the Dunn Index performance across the eight hyper-parameter combinations reveals that run #2 maximized internal consistency (compact clusters using Euclidean distance measure with projected points). Cluster validity for all runs varied from 0.659 (run #2) to 0.553 (runs #3–8). The only factorial (pre-set) partition pattern that may standout in comparison to the three unsupervised cluster predictions is identified to factor A (dilute

flow), as the performance of the remaining three factors is close to a Dunn Index value of zero for all hyper-parameter screening runs (Figure 7). It may be inferred that factor A may simultaneously influence all three water quality indices. Irrespective of adopting a Euclidean or non-Euclidean distance measure (Figure 8), the Delaunay classification error is more comparable to the cluster predictions, and it is generally minimized when the factorial partitioning for factor A is imposed on the mini-OA-dataset. Hence, it is implied that a factorial classification may be feasible when the setting sequence of factor A in the OA-planner nearly mimics the cluster solution, which is derived from the unsupervised outcome; factor A adequately reflects back a visualization of the most accurate two-dimensional projection of the three-response data system. Along these lines, what are equally convincing are the two radar graphs for the Davies–Bouldin Index performances across all eight run outcomes (Figure 9). The Davies–Bouldin Index has been prepared for both centrotype cases: (1) medoids (Figure 9A) and (2) centroids (Figure 9B). The result is similar for both versions: factor A appears to provide an approximate membership distribution that tightly encircles the unsupervised cluster identification solution. Factor A is capable of noticeably minimizing the Davies–Bouldin Index, and its OA setting groupings may pose as sequence imitations. The remaining factorial performances are much more distant to describing the cluster solution. The Dunn Index and the two versions of the Davies–Bouldin Index (centrotype cases: medoids and centroids) estimations of the cluster membership identification results are accumulated in Table 7. The Lenth test (function 'LenthPlot' in R-package BsMD (30 April 2020)) was used to complete the hyper-parameter screening using the three internal validity responses (BDIc, BDIm and DI). In the Lenth plots, the coded hyper-parameters are A: distance measure type; B: structure type; and C: position type. It becomes apparent that none of the internal checks of cluster validity promote a specific hyper-parameter setting in any of the three Lenth plots of Figure 10. This means that hyper-parameter settings carry no critical statistical significance, and they may be adjusted with respect to practicality. Therefore, the Euclidean distance measure is selected due to its efficiency, as is demonstrated in Table 5. The simpler structure and position types are also preferred. Hence, the two other hyper-setting selections that are assigned in the final solution are: the 'Compact' structure and the 'Projected Points' position (Table 4). The final clustering solution is identified for the sequence generated from run #2 (Table 6). Consequently, the overall factorial screening prediction suggests the predominance of factor A (dilute flow), as it is now clearly construed from Figures 7 and 9. The final outcome agrees with: (1) a non-parametric statistical method prediction [59], (2) a combination of semi-unsupervised (silhouette method)/statistical method with confirmation data [60] and (3) a combination of unsupervised (affinity propagation clustering)/information-entropic methods with confirmation data [61].

Table 7. Validity checks of the cluster identification results (Table 6) with Dunn Index (DI) and Davies–Bouldin Index (BDIc: centroids, and BDIm: medoids).

Run #	BDIc	BDIm	DI
1	0.416	0.721	0.605
2	0.732	0.839	0.659
3	0.416	0.721	0.474
4	0.766	0.912	0.544
5	0.766	0.912	0.574
6	0.416	0.653	0.531
7	0.766	0.912	1.022
8	0.766	0.912	1.022

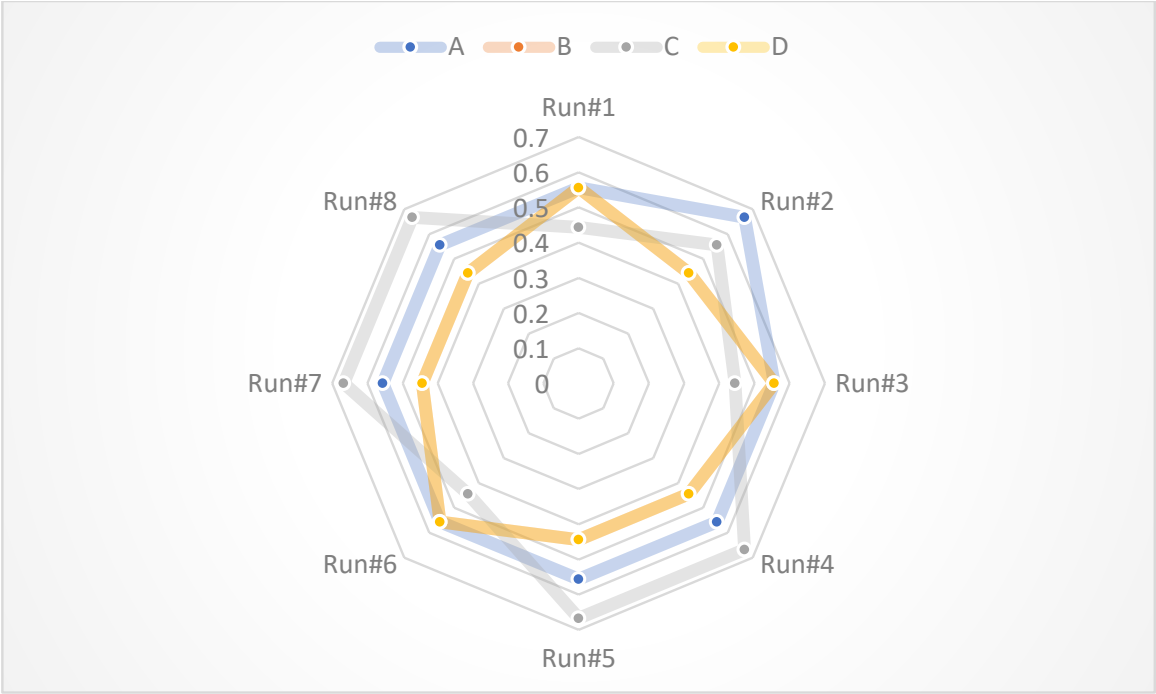


Figure 5. Radar graph of the clustering accuracy, estimated from the self-organized databionic swarm intelligence algorithm.

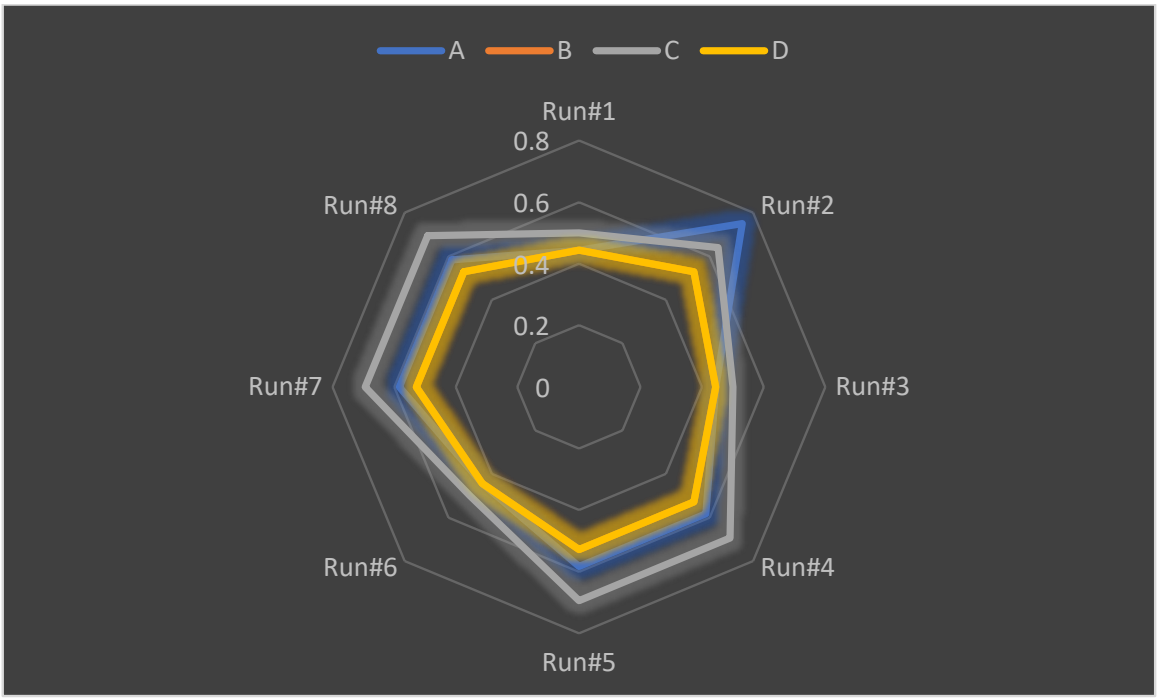


Figure 6. Radar graph of the Rand Index for the hyper-parameter screening clustering solutions.

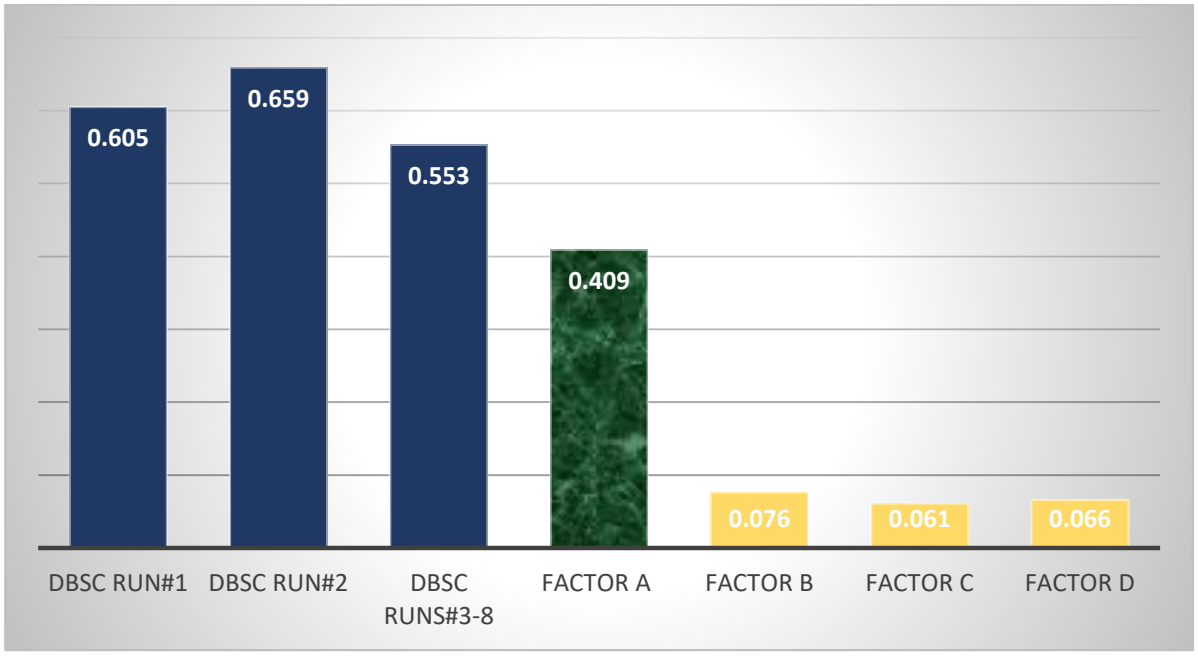


Figure 7. Bar chart for the Dunn Index performance (factorial presetting vs. cluster partitioning).

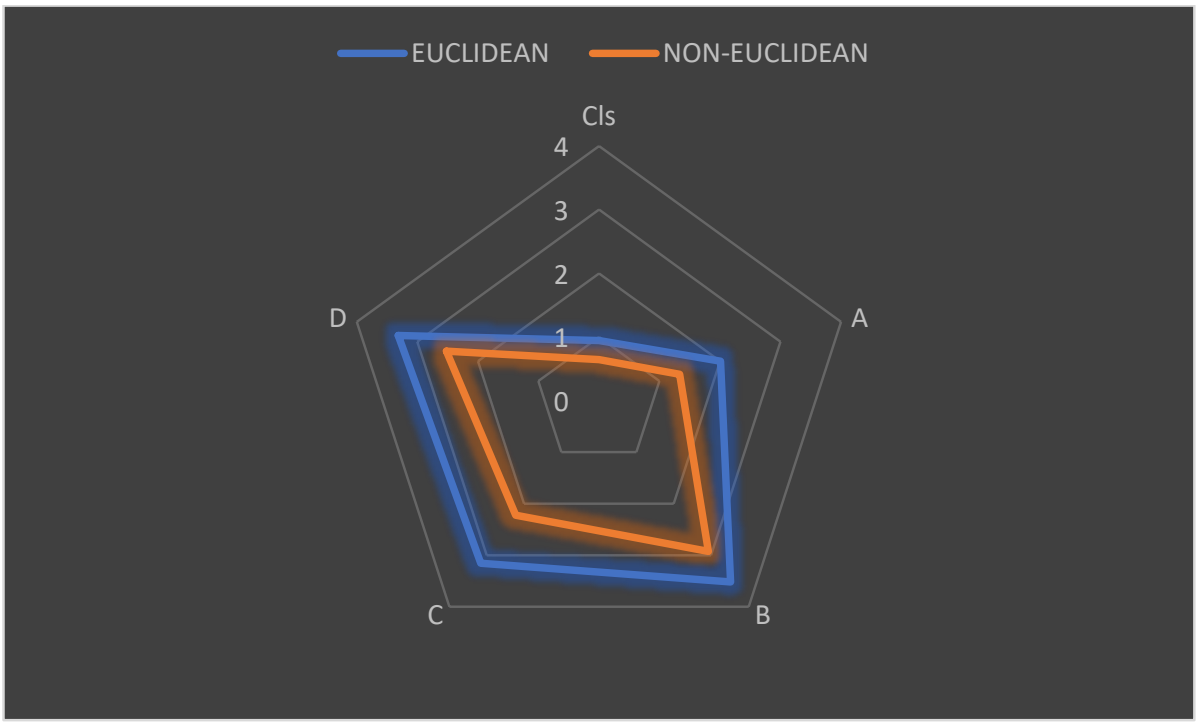
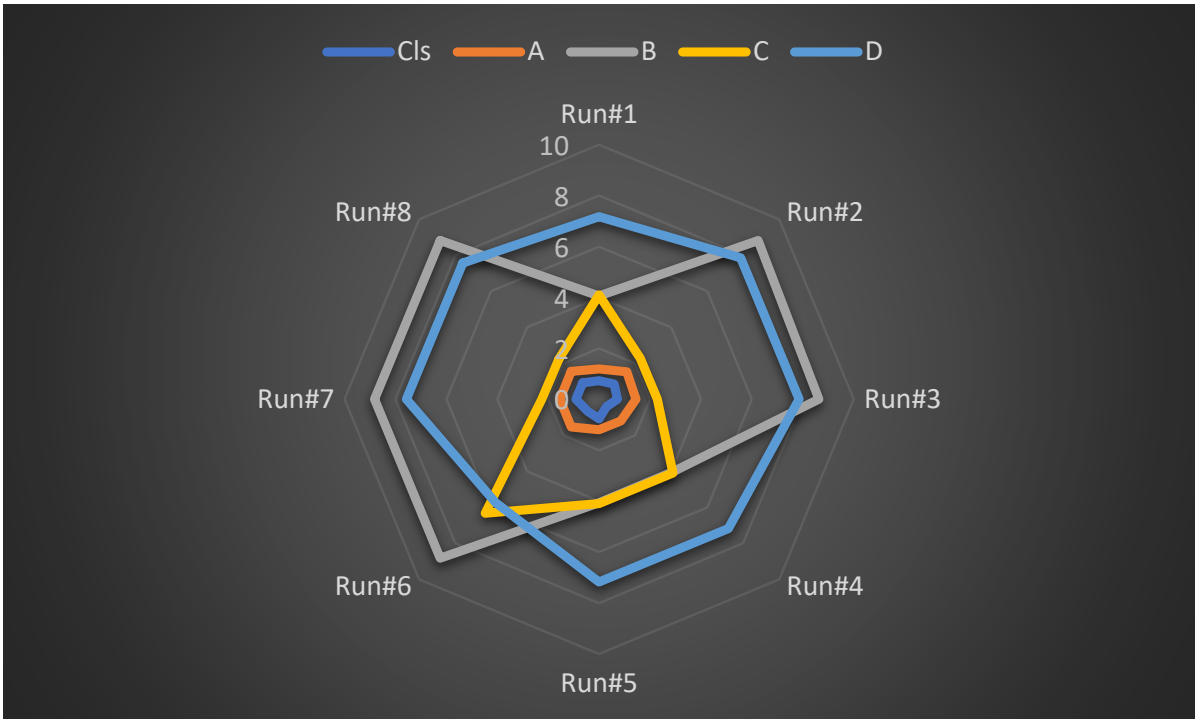
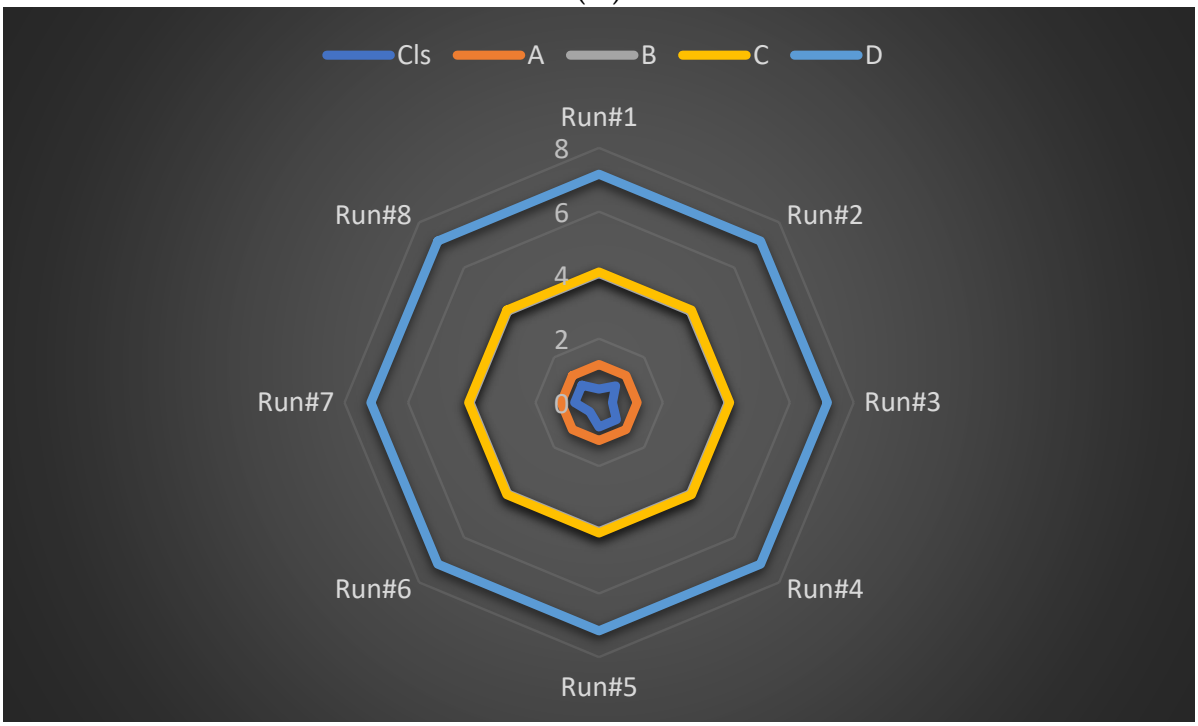


Figure 8. Radar graph for DCE comparison of Euclidean and non-Euclidean distance measure performance.



(**A**)



(**B**)

Figure 9. Radar graphs of the Davies–Bouldin Index for the centrotype cases: (**A**) medoids and (**B**) centroids.

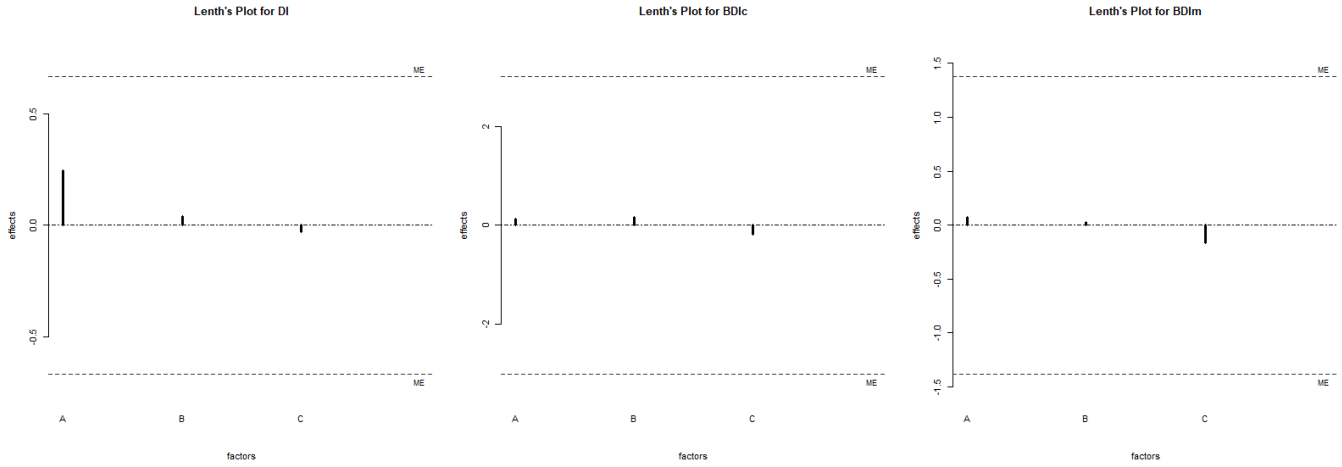


Figure 10. Lenth method results for the direct factorial screening of the BDIc, BDIm and DI estimations of Table 7 (function 'LenthPlot' in R-package 'BsMD' (30 April 2020)).

4. Conclusions

Improving the efficiency of wastewater separation processes is a very crucial subject, especially when the applied technology is geared toward supplying large water volumes for irrigating crops in arid regions. Remote locations may benefit from certain advantages that the electrodialysis technique can offer. To regulate the efficiency of the wastewater ED treatment process to a level at which the filtered water product is suitable for agricultural usage, an empirical screening of the ED process, based on pertinent water quality characteristics, is pivotal. No single theory can describe the complex phenomena that govern a custom-made ED separation process. This is because the ED design variables may be unique to the particular application. In addition, the wastewater resources are anticipated to be of diverse origins with unpredictable non-homogeneous chemical properties. Consequently, there is an urgency from an engineering perspective to facilitate quick and economic ED process profiling. Diagnosing palpably strong factors eases the concurrent adjustment of several operational quality characteristics to an optimal level. A newly published artificial intelligence method, the databionic swarm intelligence classifier, was attempted to be adapted to deal with a constrained (small) aquametric dataset. The non-parametric self-organized search for an 'objective-free' Nash equilibrium, which is tracked by emerging databots, may be doable in classifying the multi-response output from high-density OA trials. Implementing fractional factorial designs downsizes the sampling requirements to a mini-dataset. In the case study, the resulting 'small-and-dense' ED-based dataset carries information for three key water quality indices that are associated with crop production performance. The four ED-controlling factors were screened for effectiveness as well as for curvature trends at the same time. The experimental sampling plan was organized according to an unreplicated–saturated Taguchi-type $L_9(3^4)$ orthogonal array. The databionic swarm intelligence classifier received hyper-parameter fine-tuning before the machine learning capabilities of the unsupervised profiler were assessed on predicting factorial strength and tendencies. The three controlling hyper-parameters were: (1) the distance measure time, (2) the structure type and (3) the position type. The algorithmic screening was conducted by programming an unreplicated $L_8(4^1 \times 2^2)$ OA sampler. In the proposed implementation, the databionic swarm intelligence analyzer was found to be impervious to any hyper-parameter tuning. Despite being destined to handle large high-dimensional data problems, the databionic swarm intelligence solver may be also useful for extracting information from structured mini-datasets. The novel introduction in this work is the validity assessment of cluster similarity, as evidenced from the sequence matching of the unsupervised cluster formation strings to each of the pre-defined factorial setting strings dictated by the implemented OA scheme. The internal validity measures, such as

the Dunn Index and the Davies–Bouldin Index, were evaluated. Estimations of the data cluster similarities were supplemented by also using the Rand Index. The algorithmically recommended Delauney classification error furnished estimations of cluster similarities while focusing on how well-visualized the reduction was in the three-dimensional input space to the designated two-dimensional projection. The databionic swarm intelligence solver identified the wastewater diluted flow to be the only dominant variable that regulated the water quality performance in the ED cell. Therefore, the new approach resolves the effect for each of the studied ED process controlling factors by using them as a magnifying glass, and the transparent information is directly gleaned from the tightness of the grip of the individual factorial trial recipe output as it encircles the validated cluster solution. It is stressed that this is achieved in a novel fashion by self-organizing the multiple aquametric indices into cluster members against no predefined characteristic-based objectives. Future work can seek unsupervised factorial screenings for larger numbers of investigated water quality indices as well as larger numbers and variable types of the nominated controlling factors.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. WWAP (United Nations World Water Assessment Programme). The United Nations World Water Development Report 2017. In *Wastewater: The Untapped Resource*; UNESCO: Paris, France, 2017.
2. SDG Compass. Ensure Availability and Sustainable Management of Water and Sanitation for All. United Nations. 2015. Available online: https://sdgcompass.org/sdgs/sdg-6/ (accessed on 7 August 2021).
3. 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]
4. Davis, M. *Water and Wastewater Engineering: Design Principles and Practice*; McGraw Hill: New York, NY, USA, 2019.
5. Edzwald, J. *Water Quality and Treatment: A Handbook on Drinking Water*; McGraw Hill: New York, NY, USA, 2010.
6. Zhang, Y.; Shen, Y. Wastewater irrigation: Past, present, and future. *WIRE's Water* **2019**, *6*, 1234. [CrossRef]
7. Jaramillo, M.F.; Restrepo, I. Wastewater reuse in agriculture: A review about its limitations and benefits. *Sustainability* **2017**, *9*, 1734. [CrossRef]
8. 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]
9. Ungureanu, N.; Vladut, V.; Voicu, G. Water scarcity and wastewater reuse in crop irrigation. *Sustainability* **2020**, *12*, 9055. [CrossRef]
10. 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. *Agr. Water Manag.* **2016**, *177*, 419–431. [CrossRef]
11. Saliu, T.D.; Oladoja, N.A. Nutrient recovery from wastewater and reuse in agriculture: A review. *Environ. Chem. Lett.* **2021**, *19*, 2299–2316. [CrossRef]
12. 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]
13. 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]
14. Martinez-Huitle, C.A.; Rodrigo, M.A.; Scialdone, O. *Electrochemical Water and Wastewater Treatment*; Butterworth-Heinemann: Oxford, UK, 2018.
15. Parker, P.M. *The 2023–2028 World Outlook for Electrodialysis Equipment*; ICON Group International, Inc.: Rockville, MD, USA, 2022.
16. 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]
17. Burn, D.H.; McBean, E.A. Optimization modelling of water quality in an uncertain environment. *Water Resour. Res.* **1985**, *21*, 934–940. [CrossRef]
18. 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]

19. Singha, S.; Pasupuleti, S.; Singha, S.S.; Singh, R. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* **2021**, *276*, 130265. [CrossRef]
20. Malviya, A.; Jaspal, D. Artificial intelligence as an upcoming technology in wastewater treatment: A comprehensive review. *Environ. Technol. Rev.* **2021**, *10*, 177–187. [CrossRef]
21. 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 Poll.* **2021**, *232*, 411. [CrossRef]
22. Pilar Callao, M. Multivariate experimental design in environmental analysis. *Trends Anal. Chem.* **2014**, *62*, 86–92. [CrossRef]
23. George, M.; Blackwell, D.; Rajan, D. *Lean Six Sigma in the Age of Artificial Intelligence: Harnessing the Power of the Fourth Industrial Revolution*; McGraw-Hill: New York, NY, USA, 2019.
24. George, M.; Works, J.; Watson-Hemphill, K. *Fast Innovation: Achieving Superior Differentiation, Speed to Market, and Increased Profitability*; McGraw-Hill: New York, NY, USA, 2005.
25. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for experimenters—Design, Innovation, and Discovery*; Wiley: New York, NY, USA, 2005.
26. 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.
27. Taguchi, G.; Chowdhury, S.; Wu, Y. *Quality Engineering Handbook*; Wiley-Interscience: Hoboken, NJ, USA, 2004.
28. Dhingra, R.; Kress, R.; Upreti, G. Does lean mean green? *J. Clean. Prod.* **2014**, *85*, 1–7. [CrossRef]
29. Johansson, G.; Sundin, E. Lean and green product development: Two sides of the same coin? *J. Clean. Prod.* **2014**, *85*, 104–121. [CrossRef]
30. 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]
31. Fercoq, A.; Lamouri, S.; Carbone, V. Lean/Green integration focused on waste reduction techniques. *J. Clean. Prod.* **2016**, *137*, 567–578. [CrossRef]
32. 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]
33. 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]
34. 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]
35. 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]
36. Anastas, P.T.; Zimmerman, J.B. The United Nations sustainability goals: How can sustainable chemistry contribute? *Curr. Opin. Green Sustain. Chem.* **2018**, *13*, 150–153. [CrossRef]
37. Cucciniello, R.; Anastas, P.T. Design for degradation or recycling for reuse? *Curr. Opin. Green Sustain. Chem.* **2021**, *31*, 100528. [CrossRef]
38. Zimmerman, J.B.; Anastas, P.T.; Erythropei, H.C.; Leitner, W. Designing for a green future. *Science* **2020**, *367*, 397–400. [CrossRef]
39. Sheldon, R.A.; Bode, M.L.; Akakios, S.G. Metrics of green chemistry: Waste minimization. *Curr. Opin. Green Sustain. Chem.* **2022**, *33*, 100569. [CrossRef]
40. Constable, D.J.C. Green and sustainable chemistry: The case for a systems-based, interdisciplinary approach. *iScience* **2021**, *24*, 103489. [CrossRef]
41. Sajid, M.; Plotka-Wasylka, J. Green analytical chemistry metrics: A review. *Talanta* **2022**, *228*, 123046. [CrossRef]
42. Hamada, M.; Balakrishnan, N. Analyzing unreplicated factorial experiments: A review with some new proposals. *Stat. Sin.* **1998**, *8*, 1–41.
43. Perrett, J.J.; Higgins, J.J. A Method for Analyzing Unreplicated Agricultural Experiments. *Crop. Sci.* **2006**, *46*, 2482–2485.
44. Stewart-Oaten, A.; Bence, J.R.; Osenberg, C.W. Assessing effects of unreplicated perturbations: No simple solutions. *Ecology* **1992**, *73*, 1396. [CrossRef]
45. Pagliari, P.H.; Ranaivoson, A.Z.; Strock, J.S. Options for statistical analysis of unreplicated paired design drainage experiments. *Agr. Water Manag.* **2021**, *244*, 106604. [CrossRef]
46. Daniel, C. Use of the half-normal plots in interpreting factorial two-level experiments. *Technometrics* **1959**, *1*, 311–341. [CrossRef]
47. Box, G.E.P.; Meyer, R.D. An analysis for unreplicated fractional factorials. *Technometrics* **1986**, *28*, 11–18. [CrossRef]
48. Lenth, R.V. Quick and easy analysis of unreplicated factorials. *Technometrics* **1989**, *31*, 469–473. [CrossRef]
49. Fontdecaba, S.; Grima, P.; Tort-Martorell, X. Analyzing DOE with Statistical Software Packages: Controversies and proposals. *Am. Stat.* **2014**, *68*, 205–211. [CrossRef]
50. Lenth, R.V. The case against normal plots of effects. *J. Qual. Technol.* **2015**, *47*, 91–97. [CrossRef]
51. Costa, N. Design of experiments-overcome hindrances and bad practices. *TQM J.* **2019**, *31*, 772–789. [CrossRef]
52. 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]
53. 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]

54. Derringer, G.; Suich, R. Simultaneous optimization of several response variables. *J. Qual. Technol.* **1980**, *12*, 214–219. [CrossRef]
55. 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]
56. Stone, R.A.; Veevers, A. The Taguchi influence on designed experiments. *J. Chemometr.* **1994**, *8*, 103–110. [CrossRef]
57. Fisher, R.A. *Statistical Methods, Experimental Design, and Scientific Inference*; Oxford University Press: Oxford, UK, 1990.
58. Breiman, L. Statistical modeling: The two cultures. *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
59. 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]
60. 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]
61. Besseris, G. Wastewater Quality Screening Using Affinity Propagation Clustering and Entropic Methods for Small Saturated Nonlinear Orthogonal Datasets. *Water* **2022**, *14*, 1238. [CrossRef]
62. Thrun, M.C.; Ultsch, A. Swarm intelligence for self-organized clustering. *Artif. Intell.* **2021**, *290*, 103237. [CrossRef]
63. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
64. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *1*, 224–227. [CrossRef]
65. Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850. [CrossRef]
66. Bergstra, J.; Bengio, Y. Random search for hyperparameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
67. Feurer, M.; Hutter, F. *Hyperparameter Optimization*; Springer: Cham, Switzerland, 2019.
68. Stewart Lowndes, J.S.; Best, B.D.; Scarborough, C.; Afflerbach, J.C.; Frazier, M.R.; O'Hara, C.C.; Jiang, N.; Halpern, B.S. Our path to better science in less time using open data science tools. *Nat. Ecol. Evol.* **2017**, *1*, 0160. [CrossRef]
69. R Core Team. *R (Version 4.1.3): A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022; Available online: https://www.R-project.org/ (accessed on 10 March 2022).
70. Lawson, J. *Design and Analysis of Experiments with R*; CRC Press: Boca Raton, FL, USA, 2014.
71. Shepard, R.N. The analysis of proximities: Multidimensional scaling with an unknown distance function-Part II. *Psychometrika* **1962**, *27*, 219–246. [CrossRef]
72. Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 1–27. [CrossRef]
73. Thrun, M.C. *Projection-Based Clustering through Self-Organization and Swarm Intelligence*; Springer: Berlin/Heidelberg, Germany, 2018; ISBN 978-3658205393.
74. McGill, R.; Tukey, J.W.; Larsen, W.A. Variations of box plots. *Am. Stat.* **1978**, *32*, 12–16.
75. Hubert, M.; Vandervieren, E. An adjusted boxplot for skewed distributions. *Comp. Stat. Data Anal.* **2008**, *52*, 5186–5201. [CrossRef]
76. Hintze, J.L.; Nelson, R.D. Violin plots: A box plot-Density trace synergism. *Am. Stat.* **1998**, *52*, 181–184.
77. Kampstra, P. Beanplot: A boxplot alternative for visual comparison of distributions. *J. Stat. Soft.* **2008**, *28*, 1–9. [CrossRef]
78. Taleb, N.N. *Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications*; STEM Academic Press: Cambridge, MA, USA, 2020.