sustainability

MDPI

Article

A Comparison of Linear and Non-Linear Machine Learning Techniques (PCA and SOM) for Characterizing Urban Nutrient Runoff

Angela Gorgoglione [1,*,†], Alberto Castro [2,3,†], Vito Iacobellis [4] and Andrea Gioia [4]

[1] Department of Fluid Mechanics and Environmental Engineering, School of Engineering, Universidad de la República, Montevideo 11300, Uruguay
[2] Department of Computer Science, School of Engineering, Universidad de la República, Montevideo 11300, Uruguay; acastro@fing.edu.uy
[3] Department of Electrical Engineering, School of Engineering, Universidad de la República, Montevideo 11300, Uruguay
[4] Department of Civil, Environmental, Land, Building Engineering and Chemistry, Politecnico di Bari, 70126 Bari, Italy; vito.iacobellis@poliba.it (V.I.); andrea.gioia@poliba.it (A.G.)
* Correspondence: agorgoglione@fing.edu.uy
† Equally contributed to the paper.

Abstract: Urban stormwater runoff represents a significant challenge for the practical assessment of diffuse pollution sources on receiving water bodies. Given the high dimensionality of the problem, the main goal of this study was the comparison of linear and non-linear machine learning (ML) methods to characterize urban nutrient runoff from impervious surfaces. In particular, the principal component analysis (PCA) for the linear technique and the self-organizing map (SOM) for the non-linear technique were chosen and compared considering the high number of successful applications in the water quality field. To strengthen this comparison, these techniques were supported by well-known linear and non-linear methods. Those techniques were applied to a complete dataset with precipitation, flow rate, and water quality (sediments and nutrients) records of 577 events gathered for a watershed located in Southern Italy. According to the results, both linear and non-linear techniques can represent build-up and wash-off, the two main processes that characterize urban nutrient runoff. In particular, non-linear methods are able to capture and represent better the rainfall-runoff process and the transport of dissolved nutrients in urban runoff (dilution process). However, their computational time is higher than the linear technique (0.0054 s vs. 15.24 s, for linear and non-linear, respectively, in our study). The outcomes of this study provide significant insights into the application of ML methods for the water quality field.

Keywords: nutrients; urban runoff; PCA; SOM; machine learning

Citation: Gorgoglione, A.; Castro, A.; Iacobellis, V.; Gioia, A. A Comparison of Linear and Non-Linear Machine Learning Techniques (PCA and SOM) for Characterizing Urban Nutrient Runoff. *Sustainability* **2021**, *13*, 2054. https://doi.org/10.3390/su13042054

Academic Editor: Juan Tomás García-Bermejo
Received: 31 December 2020
Accepted: 8 February 2021
Published: 14 February 2021

1. Introduction

Surface freshwater, among the aquatic ecosystems, is one of the fundamental components of the water cycle for human life worldwide. It is generated from surface water bodies, which provide drinking water, preserve biodiversity, control climate, and maintain phosphorus and nitrate cycling [1]. However, in recent decades, the quality of these water bodies has been threatened by anthropogenic activities (e.g., urbanization, agriculture, water extraction, sewage discharge) [2–5]. Therefore, in the management plan of pollution control at a watershed scale, a worthwhile initial step is the identification of the pollution sources. They can be point sources (PS) or non-point sources (NPS). PS control is more direct and quantifiable and, in several countries, its mitigation has been linked to water treatment, achieving lower pollutant concentrations before discharge. Instead, NPS pollution occurs when contaminants from diverse and widely spread sources are transported by runoff into water bodies. NPS pollution is more difficult to quantify due to its diffuse

nature and the fact that many small sources might contribute to the generation of NPS pollutants [6–8]. Stormwater runoff from urban areas has been considered one of the most critical types of NPS pollution [8]. Thus, in this context, efficient and sustainable strategies for urban catchment management play a significant role in the protection of surface water quality.

The dynamic and random nature of urban runoff quality may be related to different local factors [9]. Previous studies have categorized such factors into four main groups: (i) pollutant type (with its characteristics like composition, load, decay rate) [10]; (ii) water quality physical characteristics (temperature, pH, salinity) [11]; (iii) temporal factors (seasonality, antecedent dry days, first flush pattern) [12]; and (iv) spatial factors (land cover, land use, slope, soil type, and other catchment physical characteristics) [1,13]. In addition to the intrinsic random nature of urban runoff quality, most of the works mentioned above have also demonstrated that many multifaceted interactions among these factors have occurred in multi-dimensions. In such conditions, conventional univariate (e.g., ANOVA) and multivariate (e.g., linear regression) statistical techniques have been widely adopted to investigate the correlation among water quality characteristics and influential factors [3,14]. However, the outcomes were often affected by several sources of statistical bias, unless all the analysis requirements (e.g., model selection criteria, parametric assumptions, and interaction terms) were rigidly tackled or fixed [9,15]. Therefore, considering that these issues may prevent an in-depth understanding of water quality changes and their different effects on water bodies in response to various rainfall events, the challenge in evaluating the variability of urban stormwater quality is clear.

In this context, machine learning (ML) methods are used to explore the hidden information in a multidimensional water quality dataset [16,17]. During the last decade, linear and non-linear ML techniques have been well reviewed for surface water quality assessments due to their outstanding capability to overcome the above-mentioned issues by processing and analyzing large amounts of data in a relatively short time. Gorgoglione et al. [7] adopted principal component analysis (PCA) to assess the effect of rainfall, watershed, and drainage network characteristics on urban nutrient runoff in poorly gauged areas. Furthermore, they used hierarchical cluster analysis (HCA) to evaluate the performance of a hydrologic/hydraulic and water quality model implemented in two different study areas. Dutta et al. [18] also used PCA and HCA to investigate the geospatial differences in water quality monitoring locations and identify potential water pollution sources. Liu et al. [19] carried out a comprehensive investigation into the relationship between land-use type and mineral components in river sediments by exploiting the PCA technique.

However, several researchers have stated that the linear multivariate ML techniques are constrained by the assumption of linearity, which is an unverified hypothesis for the urban nutrient runoff process [17,20,21]. For this reason, lately, non-linear multivariate methods have received significant attention from environmental researchers. Among several techniques, the self-organizing map (SOM) is one of the most adopted in the water quality field [20,22,23]. It is a type of artificial neural network (ANN) composed of fully connected neuron arrays, able to describe an environmental phenomenon depending on different physical variables (represented by a high-dimensional space) through a new low-dimensional space (usually two dimensions) [24]. Nevertheless, prior to the network training, the user has to define the number and arrangement of neurons to outline the topology structure, which directly affects the classification results. On its side, the SOM has the advantages of providing a suitable representation of non-linear processes, as well as a large number of parallel distributed structures, and is capable of learning and induction. Ding et al. [4] used the SOM training to improve linear techniques to identify the temporal and spatial patterns of several water quality variables. The authors found that the sampling-site elevation affected the water quality throughout different seasons. In fact, the reservoir water quality was poorer in the rainy season and particularly for reservoirs located on plains than the ones located on the mountains. For a similar purpose, Jiang et al. [17] adopted SOM and the growing hierarchical self-organizing map (GHSOM) in the Songhua

river basin (China). These techniques were able to explore spatial and temporal features, the correlation between water quality parameters, and the major contaminants presented in the river (chemical oxygen demand, ammonia nitrogen, total phosphorus, and fecal coliform). Ki et al. [9] applied the SOM technique to the storm water monitored dataset to gain new insights about stream water quality profiles under different precipitation conditions. Among the several outcomes of this study, it was found that, for different monitoring sites and rainfall events, the SOM showed significant variability in trace metal concentrations, with a greater impact of runoff on river water quality at the upstream stations than at the downstream ones, except under low rainfall conditions ($\leq$4 mm).

Taking into account the successful applications of the PCA and SOM, this study aims to compare the results of these two approaches, which respectively belong to the linear and non-linear ML techniques, regarding the characterization of nutrient runoff from impervious surfaces in urban watersheds. In particular, the comparison is carried out following three main aspects: (i) the ability to represent the correlation among the selected variables to represent the system and, therefore, depict the build-up and wash-off processes (feature correlation); (ii) the capability to group the dataset, based on the variables that represent the build-up and wash-off processes (data point grouping); (iii) the ability to quantify the importance of each variable (feature importance). PCA and SOM are supported by other linear and non-linear methods to strengthen this comparison. These two techniques are both used for dimension-reduction but, as far as we know in the recent literature, they have never been computed for the same dataset and their results have never been compared.

A watershed located in Southern Italy was used as a case study, where: (i) precipitation, flow rate, and water quality (sediments and nutrients) were monitored; (ii) a hydrologic/hydraulic and water quality model was calibrated and validated; (iii) a model for synthetic rainfall generation was implemented. The findings of this study will provide essential insights into the application of ML techniques for water quality data exploration in urban areas.

2. Materials and Methods

2.1. Methodology Description

A flowchart that summarizes the methodology adopted in this study to accomplish the main and the specific objectives is presented in Figure 1. Four main steps can be identified. The first one is represented by the dataset creation, including a monitoring campaign, a precipitation-generation model, and a hydrologic/hydraulic/water quality model (see Section 2.3). The second and third steps include data analysis performed by PCA and SOM, respectively (see Section 3). The last step compares the outcomes obtained by the previous two phases (see Section 3). It is worth remarking that the comparison includes not only the physical processes that characterize urban nutrient runoff but also the computational cost (computational time and hardware resource requirements) of the two algorithms.

2.2. Study Area

The urban area that we took into account for this work is located in Southern Italy (Puglia region), in Sannicandro di Bari (SB). From the climatic point of view, this region belongs to the southeastern Mediterranean area [25]. In the Koppen classification, the climate is designated as *Cs* to indicate a sub-tropical climate with dry summers [26]. Mainly, the *Cs* climate is characterized by rainy winters and dry summers, with peaks of precipitation in the shoulder seasons [27].

SB watershed has a surface equal to 31.24 ha. The average slope is equal to 1.56% and the average elevation is 169 m above sea level. The mean annual temperature is equal to 15.0 °C and the mean annual rainfall is equal to 586 mm. The impervious area represents the dominant land cover (70% of the entire catchment), while only 3.80% of the watershed is covered by green area (source: SIT Puglia) [28]. The stormwater drainage network is 1.96 km long and collects water into a concrete rectangular channel that is 1.20 m $\times$ 1.70 m.

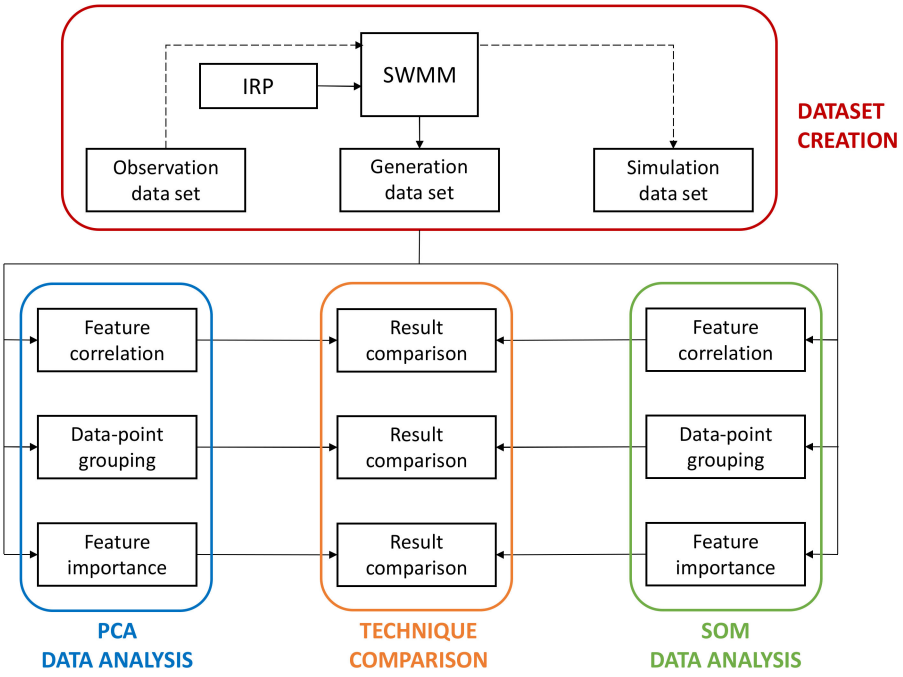


Figure 1. Flowchart of the methodology adopted in this study.

In Figure 2, the area of the watershed, the drainage network, and the outfall are shown.

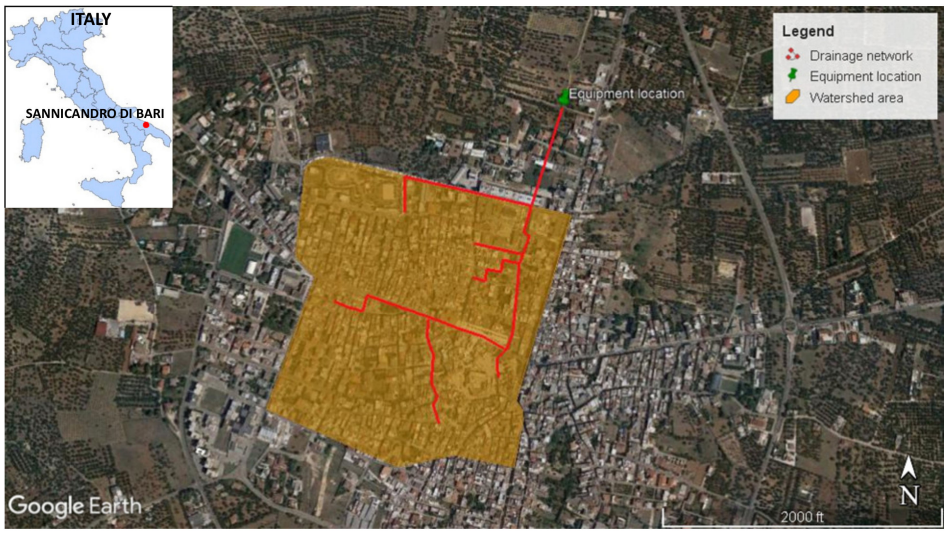


Figure 2. Sannicandro di Bari (SB) study area: basin surface, drainage network, and outfall (equipment location) (source: Google Earth).

2.3. Dataset

2.3.1. Observations

A monitoring campaign was carried out in SB to collect precipitation, flow rate, and water quality records. A rain gauge (ISCO 674 model), installed close to the basin's outlet, was used for recording the precipitation. A bubble flowmeter (ISCO 730 model) was adopted to measure discharge. Water quality data were monitored by collecting samples through an autosampler with 24 bottles of 0.5 L each and evaluated by adopting the standardized methods reported in Eaton et al. [29]. In Di Modugno et al. [30], we provided more details about the equipment used for the monitoring campaign. The monitored water quality variables were total suspended solids (TSS), total nitrogen (TN), and total

phosphorus (TP). Data of five rainfall events were collected (11/10/2006, 11/22/2006, 12/17/2006, 01/24/2007, and 02/10/2007).

Summaries of the observed rainfall-runoff (antecedent dry period (ADP), total rainfall, event duration, maximum rainfall intensity, runoff volume, and runoff peak) and water quality (minimum, maximum, and event mean concentration (EMC)) data are presented in Tables 1 and 2. Hereafter, this dataset is called "observations."

Table 1. Summary of the rainfall-runoff data for the monitored events at the SB basin.

Event	ADP (days)	Total Rainfall (mm)	Event Duration (min)	Max. Rainfall Intensity (mm/h)	Runoff Volume (m^3)	Runoff Peak (m^3/s)
10 November 2006	6	2.4	50	24	113.49	0.04
22 November 2006	11	4.3	112	6	148.86	0.04
1 December 2006	18	5.9	251	12	286.88	0.05
24 January 2007	19	1.6	37	12	111.62	0.05
10 February 2007	6	12.9	398	36	460.11	0.05

Table 2. Summary of the water quality data for the monitored events at the SB basin.

Event	TSS (mg/L)			TN (mg/L)			TP (mg/L)		
	min	max	EMC	min	max	EMC	min	max	EMC
11/10/2006	224.0	420.0	19.54	7.0	8.3	0.47	0.70	1.00	0.05
11/22/2006	124.0	2160.0	86.40	3.6	14.0	0.45	0.24	2.96	0.11
12/17/2006	6.0	217.0	6.040	-	-	-	-	-	-
01/24/2007	177.0	807.0	47.96	5.4	10.0	0.48	0.65	0.99	0.03
02/10/2007	541.0	2090.0	40.00	6.3	13.0	0.25	2.08	3.63	0.08

2.3.2. Simulations

From now on, we will call "simulations" those events characterized by observed precipitation (rainfall data described in Section 2.3.1) and simulated flow rate and water quality variables (TSS, TP, and TN). The simulations were obtained using the Storm Water Management Model (SWMM), a well-known, worldwide hydrologic/hydraulic and water quality model used to simulate hydrographs and pollutographs in urban areas [31]. It is worth mentioning that recent bibliographic works have used this model for characterizing urban runoff and for estimating pollutant loadings [32–37]. For our work, the observed precipitation of each monitored event was the input for the SWMM model, along with the physical characteristics of the watershed and the drainage network. Therefore, five "simulation events" were added to the global dataset adopted in this study.

SWMM operates in blocks or units. For the current study, we adopted the runoff and the transport block. In particular, the kinematic wave was implemented to simulate the runoff from impervious surfaces. The water losses considered in the system were represented by the infiltration process (Horton's equation) and the depression storage of the impervious portion of the watershed. Pollutant build-up, pollutant wash-off, and first-order decay were the water quality processes simulated by the model. The first two processes occur at a watershed scale, while the third occurs in the drainage network. Build-up and wash-off were both simulated with an exponential function (Equation (1) and Equation (2), respectively) [31]:

$$b = B_{max}\left(1 - e^{-k_B t}\right) \quad (1)$$

$$w(t) = bC_w q^{k_w} \quad (2)$$

where b is the pollutant build-up during the dry period [kg/ha], B_{max} is the maximum asymptotical limit of the build-up curve [kg/ha], k_B represents the build-up rate constant [1/d], t is the interval dry time [d], $w(t)$ is the cumulative mass of constituent washed off at

time t, C_w is the wash-off coefficient [1/mm], k_w represents the wash-off exponent, and q is the runoff rate over the subcatchment [mm/hr].

A comprehensive description of the above-mentioned physical processes can be found in the scientific literature [30,38,39].

The model was implemented in the study area, and the calibration and validation processes for the quantity and quality components were already successfully tackled in our previous work [30]. The calibrated parameters of Eq. 1 and Eq. 2 are summarized in Table 3.

Table 3. Water quality parameters (build-up and wash-off) calibrated at SB.

Process	Parameter	Range	Value
Build-up	B_{max}	87.000–446.000	115
	k_B	0.002–6.000	0.08
Wash-off	C_w	0.110–0.190	0.18
	k_w	0.000–3.000	2.35

2.3.3. Generations

Thereafter, we will call "generations" those events characterized by synthetic precipitation, produced by the Iterated Random Pulse (IRP) model, and simulated flow rate and water quality variables (TSS, TP, and TN) obtained using the SWMM model. In this case, the synthetic precipitation events were used as input of the SWMM model for generating hydrographs and pollutographs for each event at SB.

The IRP model was proposed by Veneziano and Iacobellis [40] and Veneziano et al. [41]. It adopts the classical depiction of the exterior process of the precipitation as an alternating sequence of dry and wet periods with independent lengths that describe the arrival, duration, and average intensity of rainfall events at the synoptic scale. The dry and wet periods are assumed to follow a Weibull and exponential distribution, respectively. The average precipitation intensities in various wet periods are independent and follow an exponential distribution. Precisely, the wet periods of the exterior model are scattered through the "interior" scheme, where the precipitation is represented as the overlapping of pulses with a hierarchically nested structure of temporal occurrences, with multifractal properties of intensity and location [42].

The IRP model was implemented at SB and provided a 15-year-long precipitation time series with 15 min of aggregation. Considering the regional regulation [43], single rainfall events were identified considering 48 h of the antecedent dry period. Consequently, 567 synthetic rainfall events were defined and introduced in SWMM for getting the simulated flow rate and water quality load and concentration.

2.4. Variable Selection

Suitable rainfall and water quality characteristics were chosen to better represent the cause–effect process of nutrient urban runoff.

For the rainfall characteristics, antecedent dry period (*ADP*), total rainfall (*Tot_Rainfall*), and runoff volume (*Runoff_Vol*) were chosen. Respectively, they represent the no-rainy days before the rainfall event (dry period), the input (*Tot_Rainfall*), and the output (*Runoff_Vol*) of the hydrologic component (wet period).

For the water quality characteristics, the event mean concentration (EMC) and the event mean load (EML) of TSS, TN, and TP were considered to not overshadow any process related to dissolved and particulate nutrients. In particular, the following equations were adopted [44]:

$$EMC = \frac{\sum_{i=1}^{n} C_i V_i}{V} \quad (3)$$

$$EML = \sum_{i=1}^{n} C_i V_i = EMC \cdot V \quad (4)$$

where V is the total runoff volume for each event [L], C_i is the average pollutant concentration at time step i [mg/L], V_i represents the runoff volume proportional to the flow rate at the time i [L], and n is the total number of samples collected during a rainfall event. EMC and EML were calculated for TSS (*EMC_TSS* and *EML_TSS*), TN (*EMC_TN* and *EML_TN*), and TP (*EMC_TP* and *EML_TP*).

2.5. Machine Learning Techniques

The two groups of ML techniques adopted for this study were linear and non-linear. For the linear methods, PCA was chosen and supported by Pearson's correlation coefficient (r). For the non-linear algorithms, SOM was selected and supported by Spearman's rank correlation coefficient (ρ).

The main objective of PCA and SOM is reducing the dimensionality of a dataset that contains a large number of interrelated variables while preserving most of the dataset variance [45]. In this study, Pearson's r and Spearman's ρ were adopted as supporting techniques (linear and non-linear, respectively) to investigate beforehand the correlation among data points to confirm or add further information to the outcomes obtained with both dimension-reduction methods.

Both groups of techniques belong to the family of unsupervised methods, where information about other response variables or group belonging is not used to obtain results. This makes these techniques suitable for exploratory analysis, where the main aim is hypothesis generation rather than hypothesis verification [5,46].

2.5.1. Linear Techniques: PCA and Pearson's *r*

The PCA decreases the dimensionality and, therefore, the complexity of a given dataset of independent variables. This method generates a new set of variables containing orthogonal-uncorrelated variables. The latter, known as principal components (PCs), are linear combinations of the original features and are arranged by decreasing variance [47,48]. The eigenvalues quantify the importance of the PCs. They are able to expose possible emerging characteristics of the system that may be hidden if we emphasize one original variable at a time [44].

Pearson's r is a measure of linear correlation between two variables. Its value lies between –1 and +1, –1 indicates a total negative linear correlation, 0 shows no linear correlation, and 1 indicates a total positive linear correlation.

2.5.2. Non-Linear Techniques: SOM and Spearman's ρ

The SOM is an ANN proposed by Kohonen [24]. It is a competitive self-organizing network, constituted by fully connected neuron arrays, which can create a two-dimensional space mapping starting from a multidimensional space. In SOM, neurons learn in an unsupervised way since no network is required to provide a specific target or objective. Neurons compete with each other to better describe the input data, with the activation of only one neuron (or one node of neurons) when a data pattern is defined (*competition phase*) [1,49]. In the training phase, the input values are progressively adjusted to maintain the neighborhood relationship in the given input dataset (*adaptation phase*). This phase generates a mapping between the multidimensional space input and the two-dimensional space output (*co-operation phase*). In this way, the SOM clusters alike data close to each other in the 2D space [50,51].

Spearman's ρ is a measure of the monotonic correlation between two variables, and, therefore, works better in catching non-linear monotonic correlations than Pearson's r. Its value ranges between –1 and +1, –1 indicates a total negative monotonic correlation, 0 shows no monotonic correlation and 1 indicates a total positive monotonic correlation.

3. Results and Discussion

3.1. Dataset Profiling

The resulting dataset was composed of 577 storm events: 5 observed, 5 simulated, and 567 generated. The data profiling process was programmed and run in Python 3.8, using the *pandas_profiling* library [52]. In Table 4, the quintile statistics (minimum, 5th percentile, median, 95th percentile, and maximum) and the descriptive statistics (standard deviation, coefficient of variation, kurtosis, mean, and variance) of each variable of the dataset (577 storm events) are reported. The histograms that represent the frequency of each variable are reported in the Supplementary Materials (SM-1).

Table 4. Quintile and descriptive statistics of the dataset.

	ADP (days)	Tot_Rainfall (mm)	Runoff_Vol (m^3)	EMC_TSS (mg/L)	EML_TSS (mg)	EMC_TN (mg/L)	EML_TN (mg)	EMC_TP (mg/L)	EML_TP (mg)
min	2.010	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5th percentile	2.177	0.617	5.580×10^{4}	4.104×10^{-6}	0.214	1.991×10^{-7}	1.444	1.050×10^{-7}	0.022
median	5.031	13.684	2.791×10^{6}	1.617×10^{-4}	735.796	1.562×10^{-6}	4.415	7.628×10^{-7}	2.225
95th percentile	18.140	100.346	2.105×10^{7}	6.576×10^{-4}	2209.532	3.635×10^{-5}	12.721	2.376×10^{-6}	5.323
max	59.135	422.099	1.489×10^{8}	1.428×10^{-3}	3655.108	2.185×10^{-4}	36.648	5.462×10^{-6}	10.780
sd	6.041	39.059	1.171×10^{7}	2.221×10^{-4}	743.537	1.951×10^{-5}	3.929	7.648×10^{-7}	1.649
coef. variat.	0.879	1.472	1.966	1.004	0.893	2.413	0.724	0.799	0.700
kurtosis	18.812	26.378	65.981	6.231	0.306	37.433	10.376	4.182	2.129
mean	6.874	26.529	5.960×10^{6}	2.211×10^{-4}	832.201	8.084	5.425	9.572×10^{-7}	2.355
variance	36.489	1525.611	1.373×10^{14}	4.934×10^{-8}	55,2847.603	3.805×10^{-10}	15.435	5.850×10^{-13}	2.720

The correlation among variables was investigated in advance by calculating the Pearson's r and the Spearman's ρ (Figure 3).

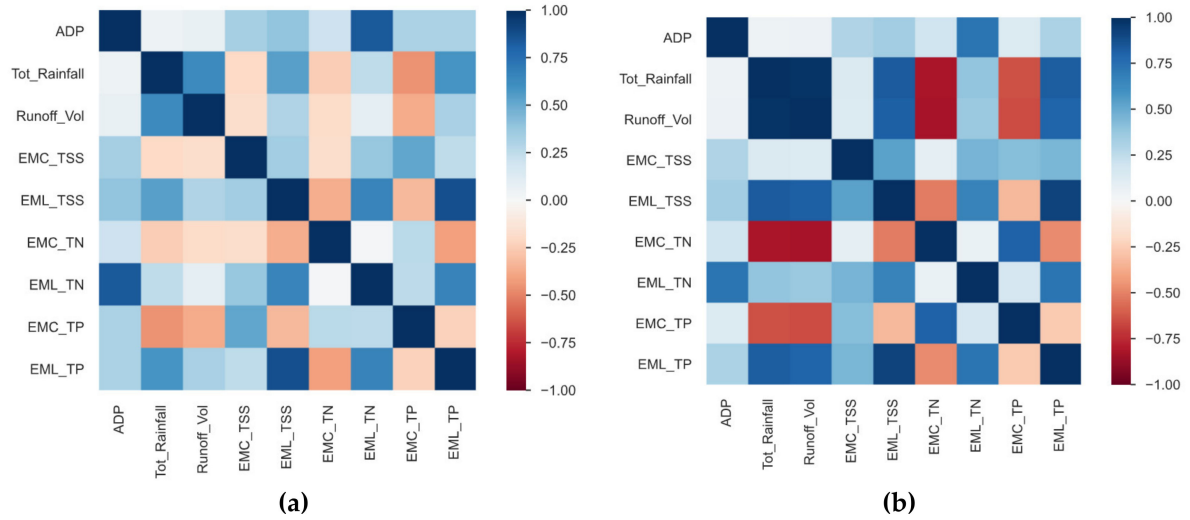


Figure 3. Matrix of **(a)** the Pearson's correlation coefficient (r) and **(b)** the Spearman's rank correlation coefficient (ρ).

From Figure 3a, it is possible to identify a high linear correlation between *Runoff_Vol* and *Tot_Rainfall*, *EML_TN* and *ADP*, *EML_TP*, and *EML_TSS*. It is noteworthy to remark that Pearson's r is not able to comprehensively describe the non-linear rainfall-runoff process represented by the variables *Runoff_Vol* and *Tot_Rainfall*, which, instead, is well represented by Spearman's ρ (Figure 3b). Furthermore, it was able to catch the nutrient transport driven by sediments (*EML_TP* and *EML_TSS*), showing, in particular, the highest particulate nature of TP compared to TN. However, no information about the dissolved portion of nutrients is provided.

Further information about non-linear processes is given by Spearman's ρ (Figure 3b). In this matrix, not only the correlation between *Runoff_Vol* and *Tot_Rainfall* is higher, as mentioned before, but it is also able to represent the dilution process of the dissolved nutrient portion: the higher *Tot_Rainfall*, and therefore *Runoff_Vol*, the lower the concen-

tration of TP and TN (*EMC_TP* and *EMC_TN*). Furthermore, it is possible to see that the phosphorus off-site movement occurs mainly due to sediment transport. In contrast, the nitrogen mobilization occurs primarily due to water surface runoff (*EMC_TN* has a stronger correlation with *Tot_Rainfall* and *Runoff_Vol* compared to *EMC_TP*).

The plots that represent each of these correlations are reported in the Supplementary Materials (SM-2).

3.2. PCA and SOM Run

A (577 × 9) matrix was used as input for the PCA and SOM analysis, where 577 refers to the number of events that occurred at SB (observed + simulated + generated) and 9 are the selected variables. Prior to the analysis, the variables were standard normalized (i.e., mean = 0, standard deviation = 1) to give equal weight to each of them and deal with their various measurement units.

In this work, we coded all the algorithms in Python 3.8 and ran them on a 2.6 GHz Intel i7 PC with 32 GB of memory.

The first two PCs were selected for SB since they represent 66.61% of the total variance (39.19% is represented by PC1 and 27.42% by PC2). PCA was implemented using the *scikit-learn* library [53]. Regarding the computing time, solving PCA with our dataset on our PC took 0.0054 s.

For evaluating the SOM map size, we calculated the neuron number from the number of data points of the training dataset using the equation proposed by Vesanto et al. [54]:

$$M \approx 5\sqrt{N} \tag{5}$$

where M represents the number of neurons, rounded to the nearest integer, and N is the number of data points. In this work, N = 577, therefore $M \approx 121$; this means a map with 11 × 11 neurons.

In this study, SOM implementation was programmed using the *minisom* package [55]. Solving SOM with the same input of PCA on the same PC took 15.24 s.

It is noteworthy that SOM's computational load increases quadratically with the number of data points. In our study, there is a four orders of magnitude difference, in terms of computational time consumption, between PCA and SOM.

3.3. Feature Correlation

The first aspect that was compared between PCA and SOM was the ability to correlate rainfall characteristics to water quality variables, in other words, to correctly represent build-up and wash-off processes. With this aim, the PCA-loading plot and the SOM weight map are represented in Figure 4.

It is important to remark that in the PCA-graphical representation, vectors representing variables that form an acute angle are considered as correlated features, while those that are perpendicular are considered as uncorrelated. In the SOM weight map, variables are correlated if they activate the same neurons in the map (red neurons, also called positive neurons). The phases of the map-weights initialization and the map training were both realized by picking samples at random from the input dataset. After a training phase of 10,000 epochs (all the input data samples were used 10,000 times), a quantization error of 0.57 and a topographic error of 0.042 were obtained, assuring the resulting map's quality.

Regarding the rainfall-related features (*ADP*, *Tot_Rainfall*, and *Runoff_Vol*), in both plots, it is possible to detect the strong relationship between *Tot_Rainfall* and *Runoff_Vol*., while *ADP* is independent of these two variables. In particular, from Figure 4b, it is possible to observe that *ADP* activates exactly the symmetric neurons of *Runoff_Vol* (neurons (1, 11) and (1, 10)) and slightly activates neuron (1, 9), whose symmetric neuron is activated by *Tot_Rainfall.*

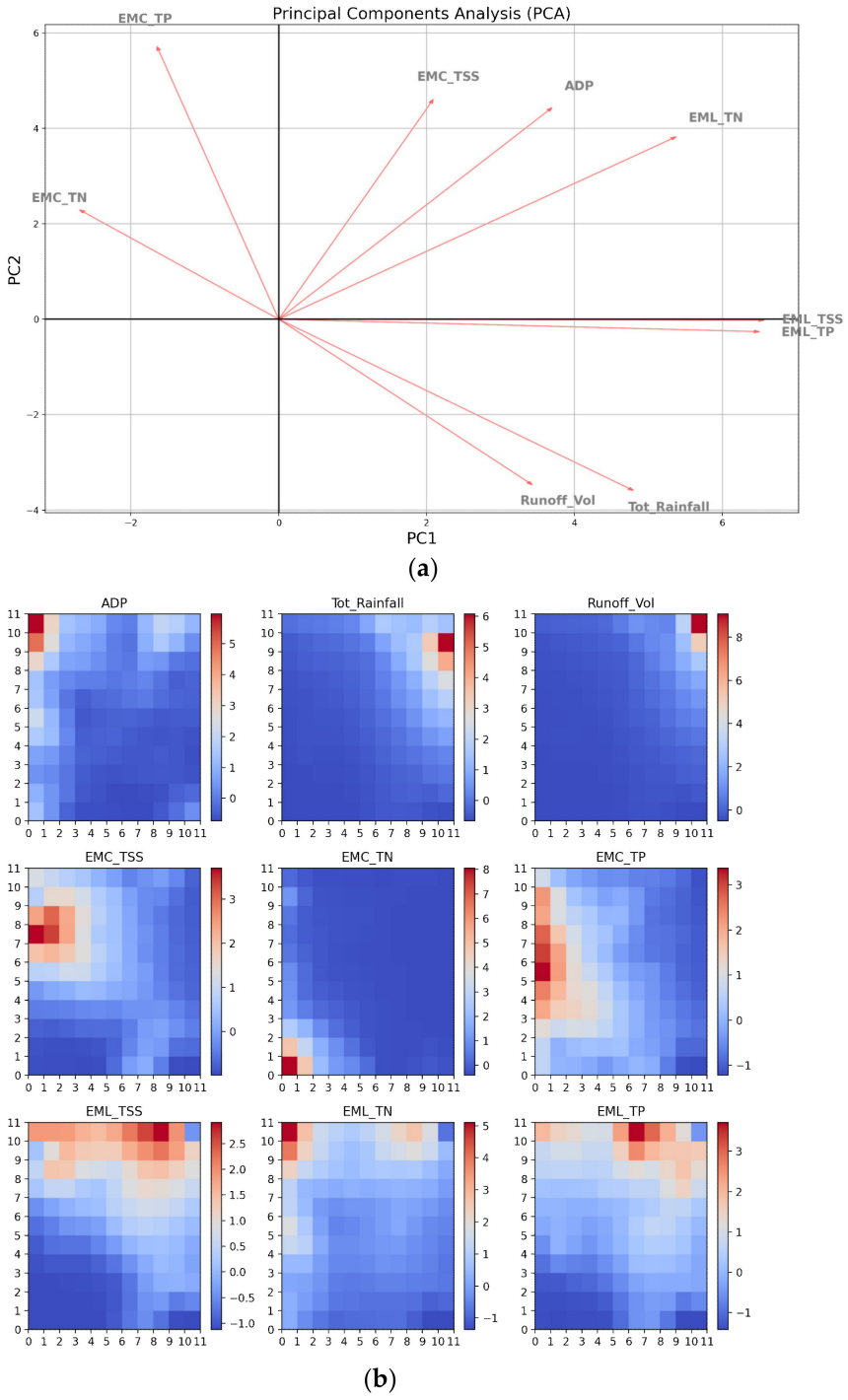


Figure 4. (**a**) PCA (principal component analysis)-loading plot and (**b**) SOM (self-organizing map) weight map for identifying feature correlations at SB.

Considering the water-quality-related variables, in the biplot (Figure 4a), we can identify two groups of variables: pollutant EMCs and EMLs. For both groups, one of the most significant results is represented by the reliable correlation between TSS and TP, and a weaker relationship between TSS and TN. These correlations suggest that sediment transport is critical in the process of nutrient mobilization from impervious surfaces. Notably, in our study watershed, phosphorus had a higher particle-bound component than nitrogen. It is possible to also identify these patterns in Figure 4b, where the stronger correlation between *EML_TSS* and *EML_TP* is represented by the same dark red neurons.

Furthermore, this graphical representation shows that the variable *EML_TSS* covers the highest variance of the system since it activates the highest number of neurons.

If we look at all the features (rainfall and water quality characteristics), *Tot_Rainfall* and *Runoff_Vol* are inversely correlated to *EMC_TP* and *EMC_TN*; this means that the more it rains, the higher the runoff volume from impervious surfaces and the smaller the nutrient concentration due to a dilution process. In both plots, this effect is more evident for *EMC_TN*, confirming the hypothesis mentioned above about the dissolved TN and the particle-bound TP. In particular, in Figure 4b, *EMC_TN* activates the two opposite neurons to *Runoff_Vol* that is the variable that represents the wash-off process. Furthermore, only in the SOM weight map, the actual strong relationship between *EMC_TSS* and *EMC_TP* is clear that, instead, is not clearly visible in the PCA biplot. Again, the assumption of dominant particulate TP at SB is confirmed. Another aspect that can be better depicted in the SOM weight map is the high correlation between *ADP* and pollutant loads, particularly *EML_TN*: the longer the no-rainy period, the more significant the amount of pollutant load accumulated on the impervious surfaces. However, this process is usually characterized by an asymptotic superior limit and degradation processes (e.g., street cleaning, wind) that cannot be described by these techniques.

Another important aspect to highlight in this comparison is that, to make a human-readable-graphical representation, it is common to plot only the first two (or three) PCs that, in our case, cover 66.61% of the variance of the system under study. SOM, instead, considers 100% of the variance in two dimensions.

3.4. Data Point Grouping

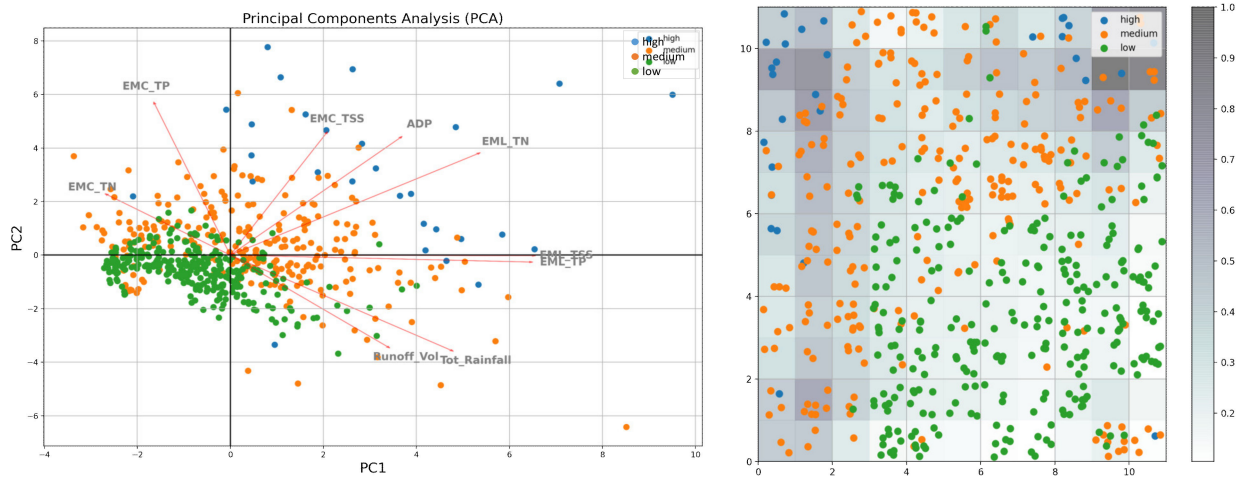
The second aspect that was considered in this comparison was the ability of the PCA and SOM techniques to group data points based on a specific feature. In particular, their capability to graphically represent data point similarities was discussed. Since we are representing nutrient build-up and wash-off (respectively depicted by dry and wet days), we tested the two methods by grouping the data points based on the *ADP* and the *Tot_Rainfall* values. We identified three categories for both variables:

$$\begin{cases} ADP \leq \text{5th percentile} \\ Tot\_Rainfall \leq \text{5th percentile} \end{cases} \text{low values}$$

$$\begin{cases} \text{5}th\ percentile < ADP \leq \text{95th percentile} \\ \text{5}th\ percentile < Tot\_Rainfall \leq \text{95th percentile} \end{cases} \text{medium values}$$

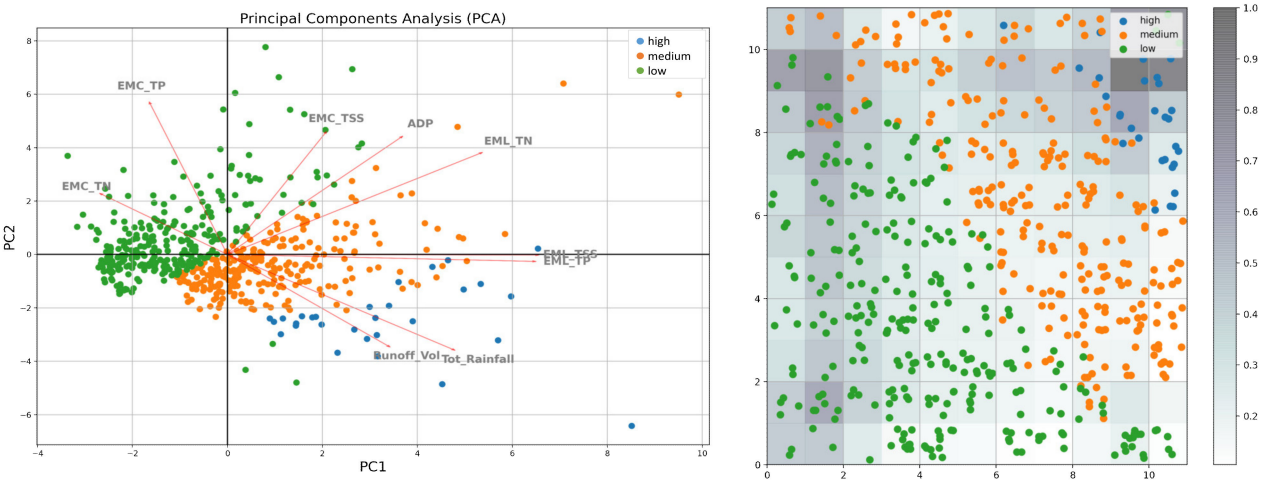
$$\begin{cases} ADP > \text{95th}h\ \text{percentile} \\ Tot\_Rainfall > \text{95th percentile} \end{cases} \text{high values}$$

In Figures 5 and 6, the PCA biplot and the SOM distance frequency map for both variables are shown. The PCA biplot is a combination of the PCA loading plot (shown in Figure 4a) and the score plot. The latter represents the original data points in the new rotated coordinated system. The grouping was applied to the scores. The SOM distance map is a tool that visualizes how much neurons differ from each other in a two-dimensional space. When two neurons correspond to different sets of data points, they would be separated by a larger distance, represented by a lighter color. On the contrary, neurons representing similar data points are separated by shorter distances, symbolized by a darker color. In the SOM frequency map, data points are plotted in the cell of the corresponding activated neuron, i.e., the winner of the competition phase.



(**a**) (**b**)

Figure 5. (**a**) PCA biplot and (**b**) SOM distance frequencies map for grouping data points based on *ADP*.



(**a**) (**b**)

Figure 6. (**a**) PCA biplot and (**b**) SOM distance frequencies map for grouping data points based on *Tot_Rainfall*.

Overall, it is possible to appreciate that both techniques are able to group data points well based on *ADP* and *Tot_Rainfall* features. It is interesting to see that for both PCA and SOM the groups are better defined with *Tot_Rainfall* (Figure 6) than with *ADP* (Figure 5).

The PCA groups data points following the same direction of the vector chosen for labeling them. This can be seen for both *ADP* and *Tot_Rainfall* vectors: data points characterized by a high value of *ADP* (or *Tot_Rainfall*) are closer to the head of the *ADP* vector (or *Tot_Rainfall* vector), while the points with a low *ADP* (or *Tot_Rainfall*) value are located in the opposite quadrant of the *ADP* vector (or *Tot_Rainfall* vector).

From the visual point of view, SOM provides the distance map that helps in better identifying the grouping borders just by checking the shade of the neurons: the darker the neuron, the more different the data points that belong to that neuron are from their neighbors. Furthermore, it is possible to combine different results to get further information. For instance, by overlapping the distance map with the frequency map, we can identify the storm events that belong to each neuron. In this case, knowing that data points that belong to the same neuron are very similar, it is possible to identify more substantial similarities (or dissimilarities) within groups that cannot be detected in the PCA. Moreover, it is easier to detect events that belong to different groups but that are similar to each other since they belong to the same neuron. Additionally, we can understand which is the feature that

characterizes each neuron (the most important one) by combining the feature-importance map (see Section 3.5) with the previous maps. Consequently, only with the SOM can we identify the most significant variable for each data point.

It is clear that one of the advantages of the SOM technique is the possibility to overlap maps that contain different information and combine their insights to get new results.

3.5. Feature Importance

The third aspect considered for this comparison study is the ability of both methods to represent feature importance. To correctly compare this capability, it is worth noting that the PCA is able to calculate the meaningful features for each PC with reference to the entire system (or, at least, for the percentage of the system represented by the PCs chosen). While, the SOM, even though it represents 100% of the system variance, identifies the most critical features only for each neuron (local feature importance) and not for the entire system.

The PCA-feature importance can be visualized by looking at the eigenvalues (loadings) showed in Table 5. For the SOM, it is possible to map the n most important features in each neuron (Figure 7). For this example, we use the first four PCs, whose variance is respectively equal to 39.19%, 27.42%, 13.53%, and 7.65%.

Table 5. PCA loadings.

	ADP	Tot_Rainfall	Runoff_Vol	EMC_TSS	EMC_TN	EMC_TP	EML_TSS	EML_TN	EML_TP
PC1	0.278	0.361	0.258	0.157	−0.203	−0.124	0.493	0.404	0.489
PC2	0.407	−0.330	−0.319	0.423	0.211	0.525	−0.002	0.352	−0.024
PC3	0.374	0.203	0.283	−0.456	0.675	−0.035	−0.126	0.201	−0.134
PC4	−0.021	0.193	0.716	0.331	−0.140	0.449	−0.260	−0.145	−0.176

Note: Loadings with absolute value higher or equal to 0.5 are highlighted in blue.

Figure 7. SOM feature-importance map.

PCA loadings with an absolute value higher or equal to 0.5 are usually considered to have a significant influence on the related factor [18,56] (highlighted in blue in Table 5). It is worth noting that PC1 is described by *EML_TSS* and *EML_TP*, and partially by *EML_TN*. This component can be interpreted as representative of the particle-bound nutrient wash-off. PC2 and PC3 can be mainly explained by *EMC_TP* and *EML_TSS* − *EMC_TN*, respectively. These two components represent the nutrient dilution process occurring during wash-off. It is noteworthy that for the first time, the linear technique is able to detect this process well. PC4 symbolizes the runoff process since it is represented by *Runoff_Vol*. In conclusion, it is possible to state that nutrient runoff in our study area is well represented by the first three PCs since EMCs and EMLs are the most important features for the entire system. However, this is not always true. In fact, in some cases, more than three PCs may be needed to represent this process. In such a case, even though the PCA loadings table is readable, the corresponding biplot would not be a human-readable-graphical representation anymore, and a non-linear technique has to be adopted.

In the feature-importance map computed by SOM (Figure 7), it is possible to arbitrarily decide the number of the most important variables to plot per neuron (three in our case). This representation allows identifying areas characterized by particular features. For instance, the bottom part, from left to right, is represented by *EMC_TN*, *EMC_TP*, and *EMC_TSS*. *ADP*, *Tot_Rainfall*, and *Runoff_Vol*, from left to right, characterize the upper part of the map. EMLs contribute to the middle area. These results confirm the previous findings, particularly those obtained from the weight map (Figure 4b). As previously mentioned, the SOM technique's advantage is represented by the possibility of coupling these different maps and detecting further hidden information.

3.6. Further Discussion

Wash-off from impervious polluted surfaces generates transport phenomenon from a range of pollutants (i.e., nutrients) such as TN and TP. Therefore, a comprehensive understanding of urban nutrient runoff is essential for water managers and environmental engineers for an efficient stormwater-treatment design in the context of sustainable urban watershed management and surface water quality protection. In this work, we aimed to assess the capability of the adopted ML methods to characterize the main processes of urban nutrient runoff. Particularly, three main aspects were analyzed: (i) the ability to represent the correlation among water quality and water quantity variables that describe the build-up and wash-off processes aiming at finding interesting insights; (ii) the ability to group the dataset to detect similarities or dissimilarities among data points; (iii) the capability to quantify the importance of each variable. In this section, the main findings are compared to the previous scientific literature related to urban runoff.

In the recent literature, some authors confirmed strong relationships between TSS and different common pollutants that can be found in urban runoff (i.e., nutrients, metals, pesticides) [57–60]. This is in accordance with the correlations highlighted in Figure 3 between TSS and TP (obtained by using Pearson's and Spearman's coefficients). This is also confirmed by Borda et al. [61]. They evaluated agronomic management's effect on the potential risk of P losses from soil to water bodies, where P losses were estimated using a simple dispersion test and the amount of suspended solids. Viviano et al. [62] found different relationships between turbidity and TP concentration in the investigated urban watershed, distinguishing between the TP from point (domestic wastewaters) and diffuse (surface runoff) sources.

Considering the results reported in Figure 4a,b, a reliable correlation between TSS and TP is evident. This confirms that phosphorus off-site movement occurs mainly due to sediment transport that is able to trigger the nutrient mobilization from impervious surfaces. In contrast, a weaker relationship between TSS and TN suggests that the nitrogen mobilization occurs primarily due to water-surface runoff (EMC_TN has a stronger correlation with Total_Rainfall than EMC_TP). In this context, Ciaponi et al. [60] corroborated

a stronger correlation between TP and TSS rather than TN and TSS. Moreover, different recent studies confirmed the N transport carried out by surface runoff [63–66].

Looking at Figure 3a,b, it is possible to recognize a strong dependence between nutrient load (EML_TN and EML_TP) and runoff volume. This is confirmed by the detailed study conducted by De Girolamo et al. on the Celone river, located in the same region of our study area (Puglia Region—Southern Italy) [67]. This work quantified the nutrient loads delivered to the downstream reservoir (Capaccio dam) on a seasonal and annual time scale. In particular, De Girolamo et al. [67] demonstrated the importance of flood event contribution to the annual nutrient load, stating that "nitrogen and phosphorus loads tend to be substantially higher during years of high precipitation, because of increased erosion and transport of the nutrients to stream channels." They also found that in the winter season, the high level of nutrient load is primarily due to surface runoff.

Another aspect that can be better depicted in the SOM weight map (Figure 4b) is the high correlation between ADP and pollutant loads, particularly EML_TN: the longer the no-rainy period, the more significant the amount of pollutant load accumulated on the impervious surfaces. This is confirmed by Gorgoglione et al. [7,20]. Moreover, in this context, Li et al. [68] found that the antecedent dry weather period and runoff volume were the determining factors in the generation of urban pollution runoff. Bian et al. [69] presented a significant positive correlation between water quality parameters and the ADP.

Considering the relationship between ADP and EMC_TSS, well represented in Figures 3 and 4, Lee et al. [70], analyzing four events in South Korea, found that the ADP and rainfall intensity were the main factors affecting TSS and COD concentrations and the loading mass of highway runoff in urban areas.

Further information about non-linear processes is given by Spearman's ρ (Figure 3b), where the dilution process of the dissolved nutrient portion is represented: the higher Tot_Rainfall, and therefore Runoff_Vol, the lower the concentration of TP and TN, meaning that the more it rains, the higher the runoff volume from impervious surface and the smaller the nutrient concentration due to dilution process. This is confirmed by Gorgoglione et al. [7,20].

4. Conclusions

The primary purpose of this work was the comparison of linear and non-linear ML techniques, PCA and SOM, respectively, to characterize urban nutrient runoff. In particular, this comparison was based on three main aspects: (i) the ability to represent the correlation among the variables chosen to represent the system and, therefore, depict the build-up and wash-off processes (cause–effect process) (feature correlation); (ii) the ability to group the dataset based on the two variables that symbolize build-up and wash-off processes (*ADP* and *Tot_Rainfall*) (data point grouping); (iii) the ability to identify and quantify the importance of each variable (feature importance). To strengthen this comparison, these techniques were supported by other linear (Pearson's r) and non-linear (Spearman's ρ) methods. The main results can be summarized as follows:

- Pearson's r was able to represent the main urban nutrient runoff processes detected in the study area: rainfall-runoff and phosphorus transport driven by sediments. Spearman's ρ, by strengthening the rainfall-runoff process, was also able to depict the transport of dissolved nutrients in urban runoff.
- Regarding feature correlation, both PCA and SOM methodologies captured the primary process that symbolizes nutrient build-up and wash-off. Notably, both were able to represent the critical role played by TSS in the nutrient mobilization from impervious surfaces. This was proved particularly for phosphorus, which dominantly was particle-bound, while nitrogen transport mainly occurred through water (dissolved). The latter was better depicted by the SOM analysis.
- Regarding datapoint grouping, both techniques were able to group data points well. The PCA groups data points following the same direction of the vector chosen for labeling them. The SOM better delineates the groups by assigning different shades

to the neurons: the lighter, the more similar to its neighbors (distance map). Furthermore, by overlapping distance and frequency map, we can identify similarities (or dissimilarities) among data points that belong to the same group.

- Concerning feature importance, the main difference between the two techniques is that the PCA can compute the meaningful variables for the system, while the SOM can only provide the feature importance for each neuron. PCA loadings are able to detect the dilution process that was never well detected by previous linear techniques. The SOM outcomes can detect the main processes under study by confirming the previous results. Furthermore, SOM maps can be coupled to extract further information.

To conclude, according to the outcomes of this work, we suggest that the SOM technique can provide a useful complementary tool to other methods, such as PCA, and can be successfully adopted for water quality research. Although both techniques can be run with the same hardware resources, it must be considered that the benefits of SOM, regarding data insights, come at a high computational cost, particularly when compared to PCA. In fact, in our study, there was a four orders of magnitude difference in terms of computational time.

The results presented in this work are expected to assist researchers and water managers in improving their water quality assessment ability. Furthermore, they support decision-making in the design of management strategies to reduce pollution impacts on receiving waters and, consequently, protect the surrounding ecological environment. An interesting aspect that will support the findings of this study is a more extensive monitoring campaign at the study area to enlarge the observation dataset. However, it is important to highlight that by exploiting accurate models that were properly calibrated and validated (IRP and SWMM), we were able to successfully characterize the nutrient urban runoff with both PCA and SOM. Based on these considerations, further steps to be considered in future works include an integrating field campaign, planned for considering more rainfall events and various pollutants.

Supplementary Materials: The following are available online at https://www.mdpi.com/2071-1050/13/4/2054/s1, SM-1: Rain gauge and bubble flowmeter specifications, SM-2: Frequency histograms of all the variables involved in the system, SM-3: Correlation plots among variables.

Author Contributions: Conceptualization, A.G. (Angela Gorgoglione), A.C., A.G. (Andrea Gioia), and V.I.; methodology, A.G. (Angela Gorgoglione); formal analysis, A.G. (Angela Gorgoglione) and A.C.; data curation, A.C.; writing—original draft preparation, A.G. (Angela Gorgoglione); writing—review and editing, A.C., A.G. (Andrea Gioia), and V.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article Di Modugno et al., 2015.

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

References

1. Keeler, B.L.; Polasky, S.; Brauman, K.A.; Johnson, K.A.; Finlay, J.C.; O'Neill, A.; Kovacs, K.; Dalzell, B. Linking water quality and well-being for improved assessment and valuation of ecosystem services. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 18619–18624. [CrossRef] [PubMed]
2. Ranieri, E.; Gorgoglione, A.; Petrella, A.; Petruzzelli, V.; Gikas, P. Benzene removal in horizontal subsurface flow constructed wetlands treatment. *Int. J. Appl. Eng. Res.* **2015**, *10*, 14603–14614.
3. Namugize, J.N.; Jewitt, G.; Graham, M. Effects of land use and land cover changes on water quality in the uMngeni river catchment, South Africa. *Phys. Chem. Earth* **2018**, *105*, 247–264. [CrossRef]
4. Ding, L.; Li, Q.; Tang, J.; Wang, J.; Chen, X. Linking land use metrics measured in aquatic–terrestrial interfaces to water quality of reservoir-based water sources in Eastern China. *Sustainability* **2019**, *11*, 4860. [CrossRef]

5. Gorgoglione, A.; Gregorio, J.; Ríos, A.; Alonso, J.; Chreties, C.; Fossati, M. Influence of land use/land cover on surface-water quality of Santa Lucía river, Uruguay. *Sustainability* **2020**, *12*, 4692. [CrossRef]
6. Khatri, N.; Tyagi, S. Influences of natural and anthropogenic factors on surface and groundwater quality in rural and urban areas. *Front. Life Sci.* **2015**, *8*, 23–29. [CrossRef]
7. Gorgoglione, A.; Gioia, A.; Iacobellis, V. A Framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds. *Sustainability* **2019**, *11*, 4933. [CrossRef]
8. Todeschini, S. Hydrologic and environmental impacts of imperviousness in an industrial catchment of northern Italy. *J. Hydrol. Eng.* **2016**, *21*, 05016013. [CrossRef]
9. Ki, S.J.; Kang, J.-H.; Lee, S.W.; Lee, Y.S.; Cho, K.H.; An, K.-G.; Kim, J.H. Advancing assessment and design of stormwater monitoring programs using a self-organizing map: Characterization of trace metal concentration profiles in stormwater runoff. *Water Res.* **2011**, *45*, 4183–4197. [CrossRef]
10. Surbeck, C.Q.; Jiang, S.C.; Ahn, J.H.; Grant, S.B. Flow fingerprinting fecal pollution and suspended solids in stormwater runoff from an urban coastal watershed. *Environ. Sci. Technol.* **2006**, *40*, 4435–4441. [CrossRef]
11. Nguyen, H.L.; Leermakers, M.; Elskens, M.; De Ridder, F.; Doan, T.H.; Baeyens, W. Correlations, partitioning and bioaccumulation of heavy metals between different compartments of Lake Balaton. *Sci. Total Environ.* **2005**, *341*, 211–226. [CrossRef]
12. Lee, H.; Lau, S.L.; Kayhanian, M.; Stenstrom, M.K. Seasonal first flush phenomenon of urban stormwater discharges. *Water Res.* **2004**, *38*, 4153–4163. [CrossRef]
13. Gobel, P.; Dierkes, C.; Coldewey, W.C. Storm water runoff concentration matrix for urban areas. *J. Contam. Hydrol.* **2007**, *91*, 26–42. [CrossRef] [PubMed]
14. Staponites, L.R.; Barták, V.; Bíly, M.; Simon, O.P. Performance of landscape composition metrics for predicting water quality in headwater catchments. *Sci. Rep.* **2019**, *9*, 14405. [CrossRef]
15. Cho, K.H.; Kang, J.H.; Ki, S.J.; Park, Y.; Cha, S.M.; Kim, J.H. Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: A case study of the Yeongsan reservoir, Korea. *Sci. Total Environ.* **2009**, *407*, 2536–2545. [CrossRef] [PubMed]
16. Almeida, S.F.; Elias, C.; Ferreira, J.; Tornés, E.; Puccinelli, C.; Delmas, F.; Dörflinger, G.; Urbanič, G.; Marcheggiani, S.; Rosebery, J.; et al. Water quality assessment of rivers using diatom metrics across Mediterranean Europe: A methods inter calibration exercise. *Sci. Total Environ.* **2014**, *476*, 768–776. [CrossRef] [PubMed]
17. Jiang, M.; Wang, Y.; Yang, Q.; Meng, F.; Yao, Z.; Cheng, P. Assessment of surface water quality using a growing hierarchical self-organizing map: A case study of the Songhua River Basin, northeastern China, from 2011 to 2015. *Environ. Monit. Assess.* **2018**, *190*, 260. [CrossRef]
18. Dutta, S.; Dwivedi, A.; Kumar, M.S. Use of water quality index and multivariate statistical techniques for the assessment of spatial variations in water quality of a small river. *Environ. Monit. Assess.* **2018**, *190*, 718. [CrossRef]
19. Liu, A.; Duodu, G.O.; Goonetilleke, A.; Ayoko, G.A. Influence of land use on river sediment pollution. *Env. Pollut.* **2017**, *229*, 639–646. [CrossRef]
20. Gorgoglione, A.; Castro, A.; Gioia, A.; Iacobellis, V. Application of the Self-Organizing Map (SOM) to Characterize Nutrient Urban Runoff. In *Computational Science and Its Applications—ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science*; Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B.O., Rocha, A.M.A.C., Tarantino, E., Torre, C.M., Karaka, Y., Eds.; Springer: Cham, Switzerland, 2020; Volume 12252.
21. Gamble, A.; Babbar-Sebens, M. On the use of multi-variate statistical methods for combining in-stream monitoring data and spatial analysis to characterize water quality conditions in the White River Basin, Indiana, USA. *Environ. Monit. Assess.* **2012**, *184*, 845–875. [CrossRef] [PubMed]
22. Sengorur, B.; Koklu, R.; Ates, A. Water quality assessment using artificial intelligence techniques: SOM and ANN—A case study of Melen River Turkey. *Water Qual. Expo. Health* **2015**, *7*, 469–490. [CrossRef]
23. Park, Y.-S.; Kwon, Y.-S.; Hwang, S.-J.; Park, S. Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map. *Environ. Model. Softw.* **2014**, *55*, 214–221. [CrossRef]
24. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480. [CrossRef]
25. Marisco, A.; Caldara, M.; Capolongo, D.; Pennetta, L. Climatic characteristics of middle-southern Apulia (southern Italy). *J. Maps* **2007**, *3*, 342–348. [CrossRef]
26. Köppen, W. *Das geographische System der Klimate, In Handbuch der Klimatologie*; Borntraeger: Berlin, Germany, 1936; Volume 1.
27. Zito, G.; Cacciapaglia, G. Precipitazioni in Puglia: Mappe stagionali. In Proceedings of the 5th Workshop Progetto Strategico Clima, Ambiente e Territorio nel Mezzogiorno, Amalfi, Italy, 28–30 April 1993; pp. 223–253.
28. SIT Puglia. Available online: http://www.sit.puglia.it/ (accessed on 16 December 2020).
29. Eaton, A.D.; Clesceri, L.S.; Greenberg, A.E. *Standard Methods for the Examination of Water and Wastwater*, 19th ed.; American Public Health Association (APHA) Association: Baltimore, MD, USA, 1995.
30. Di Modugno, M.; Gioia, A.; Gorgoglione, A.; Iacobellis, V.; La Forgia, G.; Piccinni, A.F.; Ranieri, E. Build-up/wash-off monitoring and assessment for sustainable management of first flush in an urban area. *Sustainability* **2015**, *7*, 5050–5070. [CrossRef]
31. Rossman, L.A. *Storm Water Management Model User's Manual Version 5.1*; EPA- 600/R-14/413b; National Risk Management Research Laboratory Office of Research and Development U.S. Environmental Protection Agency: Cincinnati, OH, USA, 2009.

32. Yazdi, M.N.; Ketabchy, M.; Sample, D.J.; Durelle, S.; Hehuan, L. An evaluation of HSPF and SWMM for simulating streamflow regimes in an urban watershed. *Environ. Model. Softw.* **2019**, *118*, 211–225. [CrossRef]
33. Lee, S.B.; Yoon, C.G.; Jung, K.W.; Hwang, H.S. Comparative evaluation of runoff and water quality using HSPF and SWMM. *Water Sci. Technol.* **2010**, *62*, 6. [CrossRef] [PubMed]
34. Jeon, J.H.; Yoon, C.G. Pollutant loading estimates from watershed by rating curve method and SWMM. *Korean J. Environ. Agric.* **2000**, *19*, 419–425.
35. Kim, J.H.; Paik, D.H. A study on runoff characteristics of combined sewer overflow (CSO) in urban area using GIS & SWMM. *Korean J. Environ. Health* **2005**, *31*, 467–474.
36. Baek, S.S.; Ligaray, M.; Pyo, J.; Park, J.P.; Kang, J.H.; Pachepsky, Y.; Chun, J.A.; Cho, K.H. A novel water quality module of the SWMM model for assessing low impact development (LID) in urban watersheds. *J. Hydrol.* **2020**, *586*, 124886. [CrossRef]
37. Bisht, D.S.; Chatterjee, C.; Kalakoti, S.; Upadhyay, P.; Sahoo, M.; Panda, A. Modeling urban floods and drainage using SWMM and MIKE URBAN: A case study. *Nat. Hazards* **2016**, *84*, 749–776. [CrossRef]
38. Gorgoglione, A.; Bombardelli, F.A.; Pitton, B.J.L.; Oki, L.R.; Haver, D.L.; Young, T.M. Uncertainty in the parameterization of sediment build-up and wash-off processes in the simulation of sediment transport in urban areas. *Environ. Model. Softw.* **2019**, *111*, 170–181. [CrossRef]
39. Tu, M.C.; Smith, P. Modeling pollutant buildup and washoff parameters for SWMM based on land use in a semiarid urban watershed. *Water Air Soil Pollut.* **2018**, *229*, 121. [CrossRef]
40. Veneziano, D.; Iacobellis, V. Multiscaling pulse representation of temporal rainfall. *Water Resour. Res.* **2002**, *38*, 131–1313. [CrossRef]
41. Veneziano, D.; Furcolo, P.; Iacobellis, V. Multifractality of iterated pulse processes with pulse amplitudes generated by a random cascade. *Fractals* **2002**, *10*, 209–222. [CrossRef]
42. Gorgoglione, A.; Gioia, A.; Iacobellis, V.; Piccinni, A.F.; Ranieri, E. A rationale for pollutograph evaluation in ungauged areas, using daily rainfall patterns: Case studies of the Apulian region in Southern Italy. *Appl. Environ. Soil Sci.* **2016**, *2016*, 9327614. [CrossRef]
43. Regional Regulation, 9 December 2013, n° 26, "Stormwater Runoff and First Flush Regulations" (Implementation of Article 13 of Legislative Decree n° 152/06 and Subsequent Amendments). Available online: https://www.indicenormativa.it/sites/default/files/R_26_09_12_2013.pdf (accessed on 18 December 2020).
44. Gorgoglione, A.; Bombardelli, F.A.; Pitton, B.J.L.; Oki, L.R.; Haver, D.L.; Young, T.M. Role of sediments in insecticide runoff from urban surfaces: Analysis and modeling. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1464. [CrossRef] [PubMed]
45. Adams, M.J. *Chemometrics in Analytical Spectroscopy*, 2nd ed.; Royal Society of Chemistry: Cambridge, UK, 2007; pp. 67–95.
46. Mishra, S.P.; Sarkar, U.; Taraphder, S.; Datta, S.; Swain, D.P.; Saikhom, R.; Laishram, M. Multivariate statistical data analysis/principal component analysis (PCA). *Int. J. Livest. Res.* **2017**, *7*, 60–75.
47. Massart, D.L.; Vandeginste, B.G.M.; Deming, S.M.; Michotte, Y.; Kaufman, L. *Chemometrics—A Text Book*; Elsevier: Amsterdam, The Netherlands, 1988; Chapters 1–4; pp. 14–21.
48. Arriola, A.; Pastorini, M.; Capdehourat, G.; Grampín, E.; Castro, A. Large-Scale Internet User Behavior Analysis of a Nationwide K-12 Education Network Based on DNS Queries. In *Computational Science and Its Applications—ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science*; Gervasi, O., Murgante, B., Misra, S., Garau, C., Blečić, I., Taniar, D., Apduhan, B.O., Rocha, A.M.A.C., Tarantino, E., Torre, C.M., Karaka, Y., Eds.; Springer: Cham, Switzerland, 2020; Volume 12249.
49. An, Y.; Zou, Z.; Li, R. Descriptive Characteristics of Surface Water Quality in Hong Kong by a self-organising map. *Int. J. Environ. Res. Public Health* **2016**, *13*, 115. [CrossRef] [PubMed]
50. Balamurali, M.; Silversides, K.L.; Melkumyan, A. A comparison of t-SNE, SOM and SPADE for identifying material type domains in geological data. *Comput. Geosci.* **2019**, *125*, 78–89. [CrossRef]
51. Balamurali, M.; Melkumyan, A. Detection of outliers in geochemical data using ensembles of subsets of variables. *Math. Geosci.* **2018**, *50*, 369–380. [CrossRef]
52. Pandas_Profiling Library. Available online: https://github.com/pandas-profiling (accessed on 29 December 2020).
53. Scikit-Learn Library. *Scikit-Learn: Machine Learning in Python, Pedregosa et al., JMLR 12*; MIT Press Microtome Publishing: Cambridge, MA, USA, 2011; pp. 2825–2830.
54. Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. *SOM Toolbox for Matlab 5*; Technical Report A57 2000; Neural Networks Research Centre, Helsinki University of Technology: Helsinki, Finland, 2000.
55. Vettigli, G. Minisom: Minimalistic and Numpy-Based Implementation of the Self Organizing Map. Available online: https://github.com/JustGlowing/minisom (accessed on 29 December 2020).
56. Gorgoglione, A.; Alonso, J.; Chreties, C.; Fossati, M. Assessing temporal and spatial patterns of surface-water quality with a multivariate approach: A case study in Uruguay. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *612*, 012002. [CrossRef]
57. Artina, S.; Maglionico, M.; Marinelli, A. *Le Misure di Qualità nel Bacino Urbano Fossolo, Modelli Quali-Quantitativi del Drenaggio Urbano*; CSDU: Milano, Italy, 1997; pp. 21–78.
58. Milano, V.; Pagliara, S.; Della Casa, F. Urban stormwater quantity and quality in the experimental urban catchment of Picchianti. In Proceedings of the 2nd International Conference: New Trends in Water and Environmental Engineering for safety and Life: Eco-compatible Solutions for Aquatic Environments, Capri, Italy, 24–28 June 2002.

59. Han, Y.H.; Lau, S.L.; Kayhanian, M.; Stensrtom, M.K. Correlation analysis among highway stormwater pollutants and characteristics. In Proceedings of the IWA 8th International Conference on Diffuse/Nonpoint Pollution, Kyoto, Japan, 24–29 October 2004.
60. Ciaponi, C.; Papiri, S.; Todeschini, S. *Analisi e Interpretazione Della Correlazione tra Alcuni Parametri Inquinanti Nella Rete Fognaria di Cascina Scala in Tempo di Pioggia*; XXX° Convegno di Idraulica e Costruzioni Idrauliche—IDRA: Ancona, Italy, 2006.
61. Borda, T.; Celi, L.; Zavattaro, L.; Sacco, D.; Barberis, E. Effect of agronomic management on risk of suspended solids and phosphorus losses from soil to waters. *J. Soils Sediments* **2011**, *11*, 440–451. [CrossRef]
62. Viviano, G.; Salerno, F.; Manfredi, E.C.; Polesello, S.; Valsecchi, S.; Tartari, G. Surrogate measures for providing high frequency estimates of total phosphorus concentrations in urban watersheds. *Water Res.* **2014**, *64*, 265–277. [CrossRef] [PubMed]
63. Ng Kee Kwong, K.F.; Bholah, A.; Volc, Y.L.; Pyne, E.K. Nitrogen and phosphorus transport by surface runoff from a silty clay loam soil under sugarcane in the humid tropical environment of Mauritius. *Agric. Ecosyst. Environ.* **2002**, *91*, 147–157. [CrossRef]
64. Chen, N.; Hong, H. Nitrogen export by surface runoff from a small agricultural watershed in southeast China: Seasonal pattern and primary mechanism. *Biogeochemistry* **2011**, *106*, 311–321. [CrossRef]
65. Inamdar, S.; Dhillon, G.; Singh, S.; Parr, T.; Qin, Z. Particulate nitrogen exports in stream runoff exceed dissolved nitrogen forms during large tropical storms in a temperate, headwater, forested watershed. *J. Geophys. Res. Biogeosci.* **2015**, *120*, 1548–1566. [CrossRef]
66. Chen, Y.H.; Wang, M.K.; Wang, G.; Chen, M.H.; Luo, D.; Li, R. Nitrogen runoff under simulated rainfall from a sewage-amended lateritic red soil in Fujian, China. *Soil Tillage Res.* **2012**, *123*, 35–42. [CrossRef]
67. De Girolamo, A.M.; Calabrese, A.; Pappagallo, G.; D'ambrosio, E.; Lo Porto, A. Impact of anthropogenic activities on a temporary river. *Fresenius Environ. Bull.* **2012**, *21*, 3278–3286.
68. Li, L.Q.; Yin, C.Q.; Kong, L.L.; He, Q.C. Effect of antecedent dry weather period on urban storm runoff pollution load. *Huan Jing Ke Xue* **2007**, *28*, 2287–2293. [PubMed]
69. Bian, B. Effect of antecedent dry period on water quality of urban storm runoff pollution. *Huan Jing Ke Xue* **2009**, *12*, 3522–3526.
70. Lee, J.Y.; Kim, H.; Kim, Y.; Han, M.Y. Characteristics of the event mean concentration (EMC) from rainfall runoff on an urban highway. *Environ. Pollut.* **2011**, *159*, 884–888. [CrossRef] [PubMed]