water

MDPI

Article

Internet of Things-Based Arduino Intelligent Monitoring and Cluster Analysis of Seasonal Variation in Physicochemical Parameters of Jungnangcheon, an Urban Stream

Byungwan Jo [1] and Zafar Baloch [1,2,*]

[1] Jae Sung Civil Engineering Building, Department of Civil and Environmental Engineering, Hanyang University, 222 Wasgsimini-ro, Seongdong-gu, Seoul 04763, Korea; joycon@hanmail.net

[2] Department of Civil Engineering, Faculty of Engineering and Architecture, BUITEMS, Quetta 87650, Balochistan, Pakistan

* Correspondence: engr.zafarbaloch@gmail.com; Tel.: +82-2-2220-3576 or +82-10-5927-0333

Academic Editor: Maria Filomena Camões
Received: 29 January 2017; Accepted: 13 March 2017; Published: 16 March 2017

Abstract: In the present case study, the use of an advanced, efficient and low-cost technique for monitoring an urban stream was reported. Physicochemical parameters (PcPs) of Jungnangcheon stream (Seoul, South Korea) were assessed using an Internet of Things (IoT) platform. Temperature, dissolved oxygen (DO), and pH parameters were monitored for the three summer months and the first fall month at a fixed location. Analysis was performed using clustering techniques (CTs), such as K-means clustering, agglomerative hierarchical clustering (AHC), and density-based spatial clustering of applications with noise (DBSCAN). An IoT-based Arduino sensor module (ASM) network with a 99.99% efficient communication platform was developed to allow collection of stream data with user-friendly software and hardware and facilitated data analysis by interested individuals using their smartphones. Clustering was used to formulate relationships among physicochemical parameters. K-means clustering was used to identify natural clusters using the silhouette coefficient based on cluster compactness and looseness. AHC grouped all data into two clusters as well as temperature, DO and pH into four, eight, and four clusters, respectively. DBSCAN analysis was also performed to evaluate yearly variations in physicochemical parameters. Noise points (NOISE) of temperature in 2016 were border points (ƥ), whereas in 2014 and 2015 they remained core points (ɋ), indicating a trend toward increasing stream temperature. We found the stream parameters were within the permissible limits set by the Water Quality Standards for River Water, South Korea.

Keywords: Arduino monitoring; urban stream; data analysis; clustering technique; internet of things

1. Introduction

Physicochemical parameters (PcPs) of a stream are critical to the health of aquatic ecosystems. In particular, knowledge of the temperature, pH, and dissolved oxygen content of a stream allows the determination of its ability to sustain aquatic life [1]. Aquatic ecosystem health is often impaired by anthropogenic activities, such as intake of urban sewage and urban storm water [2]. Therefore, continuous monitoring of urban streams is necessary to evaluate the impact of urban pollution on aquatic life [3]. Numerous methods exist to monitor and analyze these parameters, including discrete value analysis, water analysis, and use of monitoring networks [4]. However, continuous efficient monitoring of a stream with zero error is difficult. Recent advancements in technology have improved data acquisition and analysis techniques, as well as modes of information sharing with concerned individuals (end user). In this study, a novel technique for stream monitoring was described and data analyzed. Stream monitoring was based on an Internet of Things (IoT) approach linked to the Information Communication Technique (ICT) [5] using Arduino sensor modules (ASM). Collected data were analyzed using arbitrary dense shape clustering techniques (CTs) [6].

The data analysis techniques for stream data used by previous researchers, such as multivariate statistical methods [4], timely flow [7], and number of peak parameters [1], were collected using autonomous sensors [8]. In addition, long-term surveys [9] and large-scale monitoring programs [2] have relied on these autonomous sensors. Monitoring techniques for water bodies include remotely-wired control telemetry [7], wired online continuous monitoring [10], sensor placement network techniques [11], wired and wireless sensor network (WSN) approaches [8], and satellite sensing [12]. However, these conventional methods for monitoring stream parameters are costly, time-consuming, and involve risk of contamination [13]. Researchers have analyzed water quality using data mining techniques, such as CT [14,15] and density-based spatial clustering of applications with noise (DBSCAN) [16,17]. Moreover, CTs were also used to group similar data and describe relationships among large number of objects [18]. CTs are widely used in machine learning [19], disease classification [20], statistical analyses [21], and data mining [14]. Shape and density-based spatial clustering of application with noise (DBSCAN) can be used to classify water quality data into natural clusters along with pairwise arbitrary shapes [16]. Both CTs and DBSCAN can be used to assess seasonal behavior. Shape-based and time-series data analysis are novel approaches for analyzing freshwater parameters. Currently, there is no easy, low-cost, and efficient method to analyze data to determine the effects of anthropogenic activity on water quality in a comprehensive manner. Therefore, techniques to improve stream monitoring, such as ASM based on an IoT platform, are needed. In this study, we designed and evaluated an intelligent stream platform that consisted of open source user-friendly software, hardware, and smartphone-based data analysis capability for end-users.

In this study stream behavior was monitored by evaluating PcPs during the period from June to September, and cluster analysis of the collected data performed. Data were collected at a site on the Jungnangcheon stream, which is one of 14 tributaries that empties into the Han River. The Han River is in the middle of the Korean peninsula and drains into the Yellow Sea [22]. The Jungnangcheon stream originates in the Bulguksan mountain region south of Yang-ju city and passes through downtown Seoul. It is approximately 35 km in length with an average width of 8.6 m, depth of 0.4 to 0.9 m, and has bushes, weeds, and vegetation along the edges of the stream [15]. Stream data were collected during the three summer months and the first month of fall for three consecutive years (2014 to 2016) on weekdays. Anthropogenic activities and other sources of stream variations were also identified. We developed an IoT to share stream information with interested individuals (nearby residents, pedestrians walking by the stream, and end-users) on their smartphones. Stream seasonal data was simplified using groups defined by various CTs. The procedure and platform we established can be used by environmental statutory bodies, such as the Ministry of Environment, Republic of Korea, to efficiently monitor streams and other water bodies of interest [23,24].

2. Experimental Methods

2.1. Experimental Setup

2.1.1. Sampling Site

The sample site platform was located in the Jungnangcheon stream of the Han River in Haengdang-dong, Seongdong-gu, as shown in Figure 1. This urban stream passes through the center of Seoul, a major metropolitan city in South Korea. The IoT platform was built onsite at 37°33′12.2″ N 127°02′37.5″ E, at an altitude of 13.45 m, under a concrete girder bridge, the Seongdong Bridge of the Gangbyeon Expressway, which was constructed in 1987. The ASM was used to measure discreet parameters at a static location every second during three summer months and the first fall month for three consecutive years (2014, 2015, and 2016).

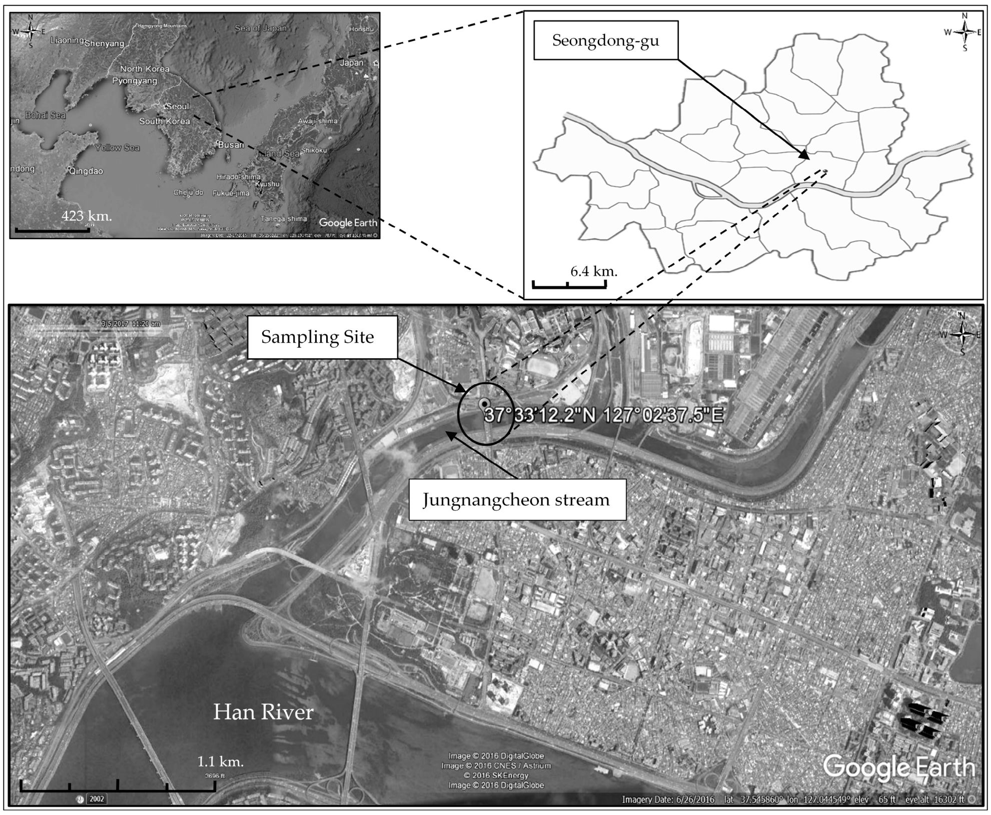

Figure 1. Geographical location of sampling site (Jungnangcheon stream).

2.1.2. Testbed Design

The experiment was designed to monitor the 24-h behavior of the stream on weekdays during summer and fall for three years (2014 to 2016). Instruments were placed downstream and anchored to the first pier for safety. A detailed schematic diagram of the ASM frame is provided in Figure 2a. The ASM instrument was placed in a porous galvanized steel strainer (70 cm length, 8.9 cm diameter, 3.35 kg weight) with a galvanized zinc cap (10 cm length). The holes of the strainer were 3 mm in diameter. Figure 2b,c show the strainer (bracket for sensor) fixed vertically to the pile with two bolts for protection against wave action. The strainer was horizontally flanged on the pile cap to keep the instrument safe from external impacts.

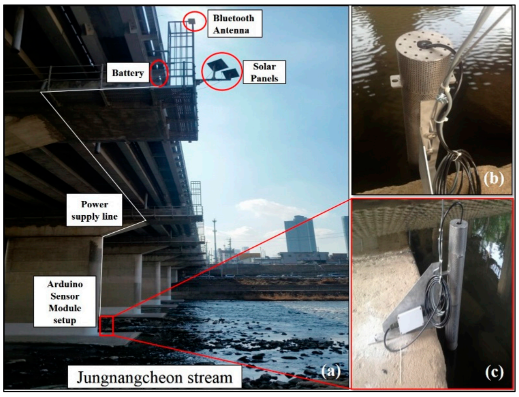

Figure 2. (**a**) Schematic illustration of Arduino sensor modules (ASM) monitoring installation setup; (**b**) top view of the stainless steel strainer casing for the ASM; and (**c**) side view of ASM strainer horizontally flanged on the pile cap anchored on the concrete girder for instrument safety from external impacts.

2.1.3. Instrument Calibration

All sensor modules were cleaned with a brush before calibration and the accuracy verified. The sensor guard was removed and attached to a calibration cup half-filled with deionized water while being shaken to dislodge any particles attached to the sensor surroundings. This practice was repeated 3–5 times on weekends. The dissolved oxygen (DO) sensors were calibrated using the saturated air method with tap water that had a specific conductance less than 0.5 mS/cm using Equation (1) below:

$$B\rho^{\circ} = B\rho - 2.5\left(\frac{A_m}{30.5}\right) \tag{1}$$

where $B\rho^{\circ}$ is the barometric pressure at altitude, $B\rho$ is the barometric pressure at sea level and A_m is the altitude in meters. A buffer solution (pH = 4) was poured into the cap of the pH sensor module; after 24 h, the sensor was calibrated using the Henderson–Hasselbalch equation to estimate the pH value of the buffer solution, as shown below:

$$\mathrm{pH} = pKa + \log\left(\frac{A^{-}}{A^{+}}\right) \tag{2}$$

where pKa is –logKa, Ka is the acid dissociation constant, A^{-} is the molar concentration of a conjugate base, and A^{+} is the molar concentration of an undissociated weak acid.

2.2. Establishment of an Arduino Sensor Modules Network

We devised an ASM network to monitor the stream. This intelligent stream platform was a prototype model based on the open source Arduino platform in an integrated development environment (IDE) and easy-to-use hardware (Arduino Bluetooth shield V2.1, Arduino AG, Ivrea, Italy). The Arduino UNO is a microcontroller chip (ATmega328P, Atmel, San Jose, CA, USA) capable of reading and writing data. We installed an Arduino temperature sensor (DHT 11, D-Robotics, London, UK), DO sensor (SEN-11194, Spark Fun electronics, Boulder, CO, USA), and pH meter (SKU: SEN0169, DFRobot, Beijing, China). This framework was managed through a Microsoft Surface device as the main server (MS) with Operating System 8 (Microsoft, Redmond, WA, USA). The temperature measurement of sensor module ranged between -5 and 50 °C ± 0.10 °C, the DO between 0 and 50 ± 0.2 mg/L, and pH from 0 to 14 ± 0.2 units. The sensor modules used in this study conforms to the water quality (WQ) guidelines laid by Ministry of Environment, Republic of Korea [24]. According to these guidelines, the permissible values for temperature, DO and pH range of stream water are (22.1 °C to 28.3 ± 1.3 °C) ($\geq$7.5 mg/L) and (6.5 to 8.5), respectively. The WSN efficiency was 99.99% and the error of the sensor drop-down transmission was ±20. This communication link was developed using a Bluetooth flat panel antenna at 2.4 GHz with a performance of 15.5 dBi (L-com, North Andover, MA, USA).

2.3. Internet of Things Site Monitoring Procedure

The IoT platform for stream monitoring was based on the following six procedural steps to allow sharing of simplified PcP values with end-users on their smartphones:

i. The sensor modules of temperature, DO, and pH were fabricated with Arduino shield V2.1 as shown in Figure 3a.

ii. This ASM frame was installed in stainless steel and fixed at the site, as shown in Figure 1b.

iii. Solar power energy of 3 V to 5 V was supplied to sensor nodes, while solar power energy of 12 V to 24 V was supplied to the Bluetooth antenna using a leakage, corrosion, and water-resistant sealed battery, as shown in Figure 3b.

iv. This prototype model was pre-programed with the permissible limits of water quality standards for river water in Korea in Arduino IDE 1.0.X (Arduino AG) [25].

v. ASM data acquisition and subsequent data were transferred to the main server. Communication was developed using the Bluetooth antenna, as shown in Figure 3c.

vi. Development of the ASM as part of the IoT and information sharing with residents on their smartphones are shown in Figure 3d.

vii. Stream data was transferred to the MS every second on weekdays for three years. Summary statistics of seasonal variations in stream data obtained via the ASM framework are given in Table 1.

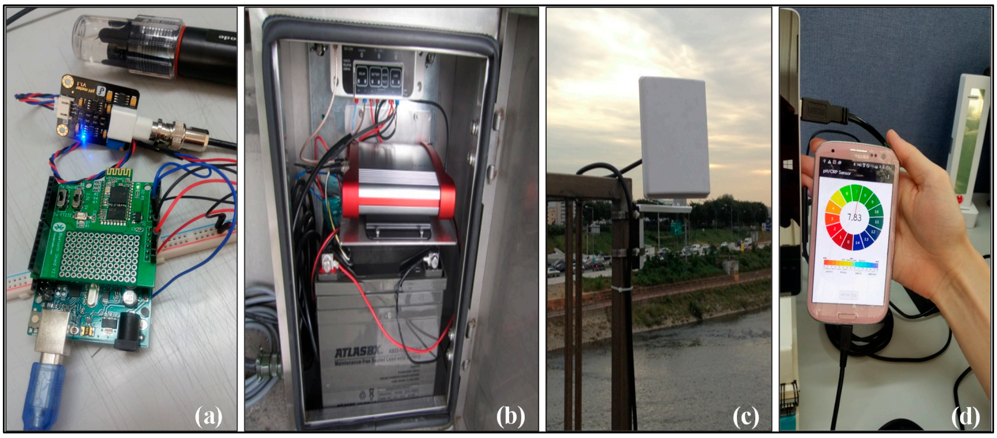

Figure 3. (**a**) Integrated Arduino-based Bluetooth shield pH sensor; (**b**) solar power-controlled battery (3 V to 5 V with 12 V to 24 V); (**c**) Bluetooth flat panel antenna (2.4 GHz with 15.5 dBi); and (**d**) Arduino-based mobile application for smartphones.

Table 1. Summary statistics of seasonal variations in stream data obtained via the ASM framework.

Seasonal Parameter	Temperature Data (MBps) Mean Values	DO Data (MBps) Mean Value	pH Data (MBps) Mean Value	Temperature Data (MBps^2) Variance	DO Data (MBps^2) Variance	pH Data (MBps^2) Variance
2014	0.01079911	0.010799088	0.01079922	281.31667	263.86667	287.59583
2015	0.01079916	0.01079918	0.01079919	161.18333	165.39583	181.58333
2016	0.01079922	0.01079919	0.01079927	342.9625	177.5625	105.45

Note: DO: Dissolved oxygen.

2.4. Clustering Techniques

2.4.1. K-Means Clustering Technique

We used a clustering technique to classify the raw dataset generated by the ASM installed at Jungnangcheon stream. Three-year data were grouped into similar clusters, while non-similar data formed individual groups. Clustering can either be supervised (cluster structure evaluated based on external information) or unsupervised (cluster structure evaluated without external information). We analyzed the data using K-means clustering, agglomerative hierarchical clustering (AHC), and DBSCAN. K-means clustering analyzes time series based on the centroid point of clusters using a partitioning technique such that $k \leq n$, where k is the number of clusters, and n is the maximum number of objects in the dataset. Generally, there are two partitioning techniques for data objects, K-means and K-medoid [26]. In brief, we applied a time series equation of K-means to determine the behavior of the stream. Data collected over 240 days was partitioned into 24-h segments using Equation (3) below:

$$S = \sqrt[t]{\sum_{\mathrm{i}=1}^{k} \sum_{\rho \in \mathrm{C}_{\mathrm{i}}} dist\left(\rho, \mathrm{C}_{\mathrm{i}}\right)^2} \tag{3}$$

where S is the integral of the squared error of total observations, k is the number of clusters, i is the centroid of the object group, ρ is the time series point of the object, $\mathrm{C_i}$ is the centroid of a cluster, time factor t and *dist* are the pairwise distance of observations of the time intervals of points a and b. To avoid duplicate tracks in $dist\ (a, b)$, the Euclidean distance (euclidean), an application of the Malinowski space, was used. The Malinowski space can be solved using the matrix value of $p = \infty$ in Equation (4). The pdist shows that object $\rho \in \mathrm{C_i}$ is the cluster of the object. Here, x the matrix of order $x \times y$, where m is the matrix, y is the axis, and n is the maximum large number of objects obtained x $(1 \times n)$. The row vector was calculated using the Euclidean Equation (5) below:

$$\mathrm{dist}^2_{\mathrm{ab}} = (\mathrm{x_a} - \mathrm{x_b})(\mathrm{x_a} - \mathrm{x_b})' \tag{4}$$

$$\mathrm{pdist_{ab}} = \sqrt[\mathrm{p}]{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\mathrm{x_{ai}} - \mathrm{x_{bi}}\right|^{\mathrm{p}}} \tag{5}$$

The initial centroid point can be used as the Euclidean, while the number of clusters can be estimated using the silhouette coefficient (ζ_i) [27], as shown in Equation (6) below:

$$\zeta_{\mathrm{i}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{(\beta_i - \alpha_i)}{\max(\alpha_i, \beta_i)} / \mathrm{N} \tag{6}$$

where ζ_i is the (i^{th}) object in the average distance (α_i) of an individual group of the same cluster; if not, then the minimum distance with respect to the second cluster value of β_i is selected. The value of ζ_i is between ± 1; a negative value corresponds to $(\alpha_i > \beta_i)$, a positive value indicates $(\alpha_i < \beta_i)$, assuming the coefficient of α_i is zero. This partitioning technique helps organize observed data into hierarchical clusters.

2.4.2. Agglomerative Hierarchical Clustering

In the second phase of data analysis, we performed hierarchical clustering (HC) using an unsupervised clustering method to classify identical pairs and merge them into a single cluster. HC performs two types of clustering, agglomerative (bottom-up merging) and divisive (top-down splitting). In this study, we used AHC of the Jungnangcheon stream data to observe the distance level created by this partitioning technique. This technique presents the data in the form of a tree, also known as a dendrogram [28]. Clusters in a dendrogram can be linked using a single link method or a complete link method. The single link $(\gimel_\ell)$ method joins two clusters that have maximum similarities and are located at a minimum distance from each other. The complete link $(\beth_\ell)$ joins two clusters that have minimal similarity that are located at a maximum distance from each other. D_{link} is the minimum distance with maximum similarities obtained using the equation $D_{sl} = \gimel_\ell$. To summarize, we evaluated stream data collected over 240 days arranged at a 24-h diurnal interval using AHC. The shape-based AHC consisted of a column matrix with order $x \times y$ and the pairwise distance (*pdist*) observed on a plot of the percentage of D_{link}/D_{max}.

2.4.3. Density-Based Spatial Clustering of Applications with Noise

We also analyzed data obtained from Jungnangcheon stream using DBSCAN. This technique estimates the point number specified by the radius point size of the object neighborhood. The neighborhood radius (eps) is the centroid point of the object and while the minimum number of points (MinPts) is defined by the domain, as given in Equation (7). The maximum large number of objects of DBSCAN points are classified into three points: core points (ɋ), border points (ƥ), and noise points (noise); ɋ refers to the interior points of the dense region, ƥ are edge points of the dense region, and noise are neither ɋ or ƥ [26]. The Ester Martin algorithm [16] was applied to the Jungnangcheon stream data using MATLAB R16a (MathWorks, Natick, MA, USA). Our work performance fulfilled the conditions of CLARANS [29] and revised DBSCAN [17,30,31]. Moreover, cluster points were joined together through maximum density reachability (MDR). The MDR distance was calculated using the object points of the directly density reachable (DDR) and density reachable (DR) distances. DDR is the point distance from the x-point to the z-point of the cluster and is known as the core point condition. The DDR in the present study was calculated using Equation (8). The neighborhood distance points of ɋ and ᵽ indicate the DDR and the DR reachability, respectively. In brief, these distances were used to calculate the noise for set points whose critical distance (D_{min}) was not part of a cluster. Equation (9) was used to formulate the maximum large number of Noise for the seasons as follows:

$$\mathrm{Eps}\left(\text{ᵽ}\right) = \left\{\text{ɋ} \in \mathrm{D}|\mathrm{distance}\left(\text{ᵽ},\text{ɋ}\right) \leq \mathrm{E}\right\} \tag{7}$$

where:

$$\text{ᵽ}_1 = \text{ɋ} \text{ and } \text{ᵽ}_\mathrm{n} = \text{ᵽ}.$$

Let:

$$\text{ᵽ}_1, \text{ᵽ}_2, \text{ᵽ}_3, \ldots . \text{ᵽ}_\mathrm{n} \text{ and } \text{ᵽ} \in \mathrm{Eps}(\text{ɋ}) \text{ and } |\mathrm{Eps}(\text{ɋ})| \geq \mathrm{MinPts}$$

$$\text{ᵽ} \in \mathrm{N_{Eps}}(\text{ɋ}) \text{ and } \left|\mathrm{N_{Eps}}(\text{ɋ})\right| \geq \mathrm{MinPts} \tag{8}$$

$$\mathrm{N_{Noise}} = \left\{\text{ᵽ}\epsilon\mathrm{d} \mid \forall\, \mathrm{i} : \text{ᵽ} \notin \mathrm{C_i}\right\} \tag{9}$$

3. Results and Discussion

3.1. Experimental Approach

PcPs of Jungnangcheon stream were monitored using an IoT platform. The computational overheads of the measurements had an efficiency of 99.99% during the study period; the data expected was 7.4157715 MBps, whereas 7.415203 MBps data was received. Figure 4 shows the response time of storing and processing the sensed data. The drop-off in the network delay in the temperature signal was 86,353 bps on days 10, 49, 75, and 160. The same trend was observed for DO and pH. In addition, the drop-off in temperature data to 86,378 bps during different day intervals are shown in Figure 4. The network delay of 596 Bps was measured; the reason for this is unclear and may have been caused by delays in the wireless logger or program functioning. It shows that the proposed IoT platform is reliable and highly efficient in data transmission, which leads to the high response time of the sensor, reliable processing, and operational data storage, without any significant loss.

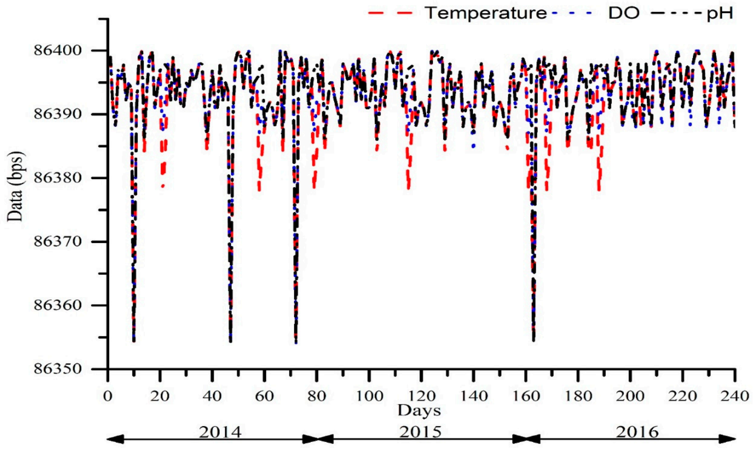

Figure 4. Data in bps received by the main server (MS) during the three summer months and the first fall month for three consecutive years (2014–2016). DO: Dissolved oxygen.

Every day a sudden drop of 1.27 °C was observed early in the morning, followed by an average gradual rise of 1.2 °C at noon and then another drop in temperature in the evening, as shown in Figure 5. The DO concentration increased smoothly with an average difference in concentration of 0.22 mg/L between midnight and dawn. The drop-off in DO concentration was high during the day and low at night. The pH of the stream (7.63) was slightly acidic. A decrease of 0.10 decrease in pH was observed at midnight and an increase of 0.17 at midday. Table 2 shows the strongest correlation between temperature and DO was in September, followed by June, July, and August, indicating DO changes were strongly affected by high temperatures in 2014, 2015, and 2016. A strong correlation coefficient (R^2) between temperature vs. pH was also observed in the month of July, followed by September, August, and June during study periods (Table 2). In general, it is observed that with the rise in temperature the pH also increases steadily. The temperature and DO were also found to be affected by the bed slope of the stream. The variation in stream parameters are essential for aquatic life [32].

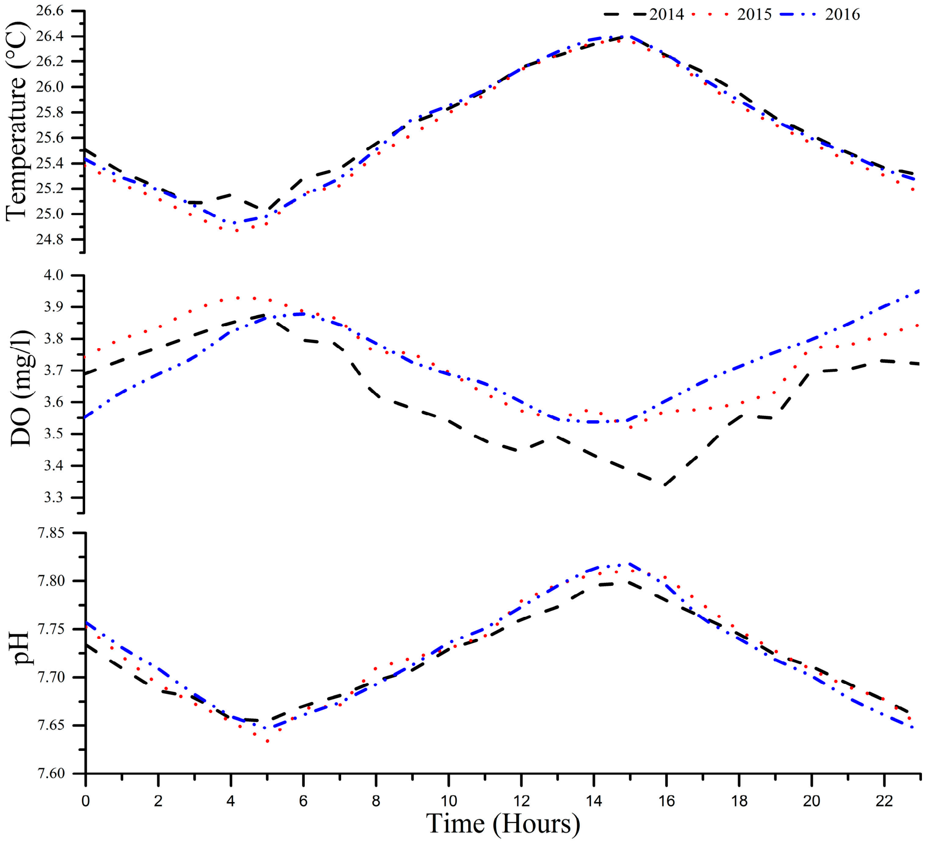

Figure 5. Stream behavior based on stream water temperature during the three summer months and the first fall month for three consecutive years.

Table 2. Results of correlation coefficient for linear regression (R^2) values of temperature and DO and temperature vs. pH from 2014 to 2016.

R^2 Correlation Values								
	Temperature vs. DO				**Temperature vs. pH**			
Year	**June**	**July**	**August**	**September**	**June**	**July**	**August**	**September**
2014	0.4021	0.3275	0.3226	0.4492	0.0875	0.6494	0.2909	0.4631
2015	0.4317	0.3318	0.363	0.5808	0.0910	0.6739	0.4578	0.5109
2016	0.2524	0.1512	0.4035	0.5115	0.0837	0.6232	0.4514	0.498

3.2. K-Means Clustering

The raw data generated by the sensors installed in the urban stream are partitioned into three sub-groups of physicochemical parameters (temperature, DO, and pH) using K-means clustering. We used the K-means algorithms to cluster the raw data obtained from the sensors. Separate PcP clusters are shown in Figure 6a. This figure shows the partition of the raw data into three clusters. Since we collected the data over the specified study period, we clustered each PcP value into a sub-cluster. Figure 6b shows the raw temperature datasets. In this figure, it is observed that cluster 2 is well classified as the values on the x-axis are higher, followed by clusters 3 and 1. The cluster 1 shows the higher values of ζ_i compared to the other clusters, which indicates the variability of data obtained over the period of study. Moreover, it is ascertained from Figure 6c that there are significant changes in the variance of the three clusters for raw DO data. Figure 6d shows that cluster 1 has larger variance with respect to ζ_i and higher width values, followed by clusters 2 and 3.

K-means clustering was applied to stream parameters to classify the number of clusters based on a comparison of tightness and separation. These clusters display cohesion on the x-axis, while the data width of the cluster is shown on the y-axis, calculated using ζ_i. The positive increase on the x-axis indicates that the clusters are strongly compacted, owing to small variations in the data obtained. A higher cohesion value is preferable for distance (α_i) calculations using the Euclidean distance, while a decreasing trend indicates the ζ_i value is reduced, which results in the creation of a second cluster (β_i). This dissimilarity splits clusters into subsectors of less similar cluster values. The K-mean ζ_i provides an indication of cluster compactness. Moreover, it is envisaged from the figures that the negative ζ_i values represent negative dissimilarity in the cluster values. Thus, a large amount of real-world data generated by distributed sensors was classified using K-means clustering. This clustering technique was also useful in predicting the quality of stream data.

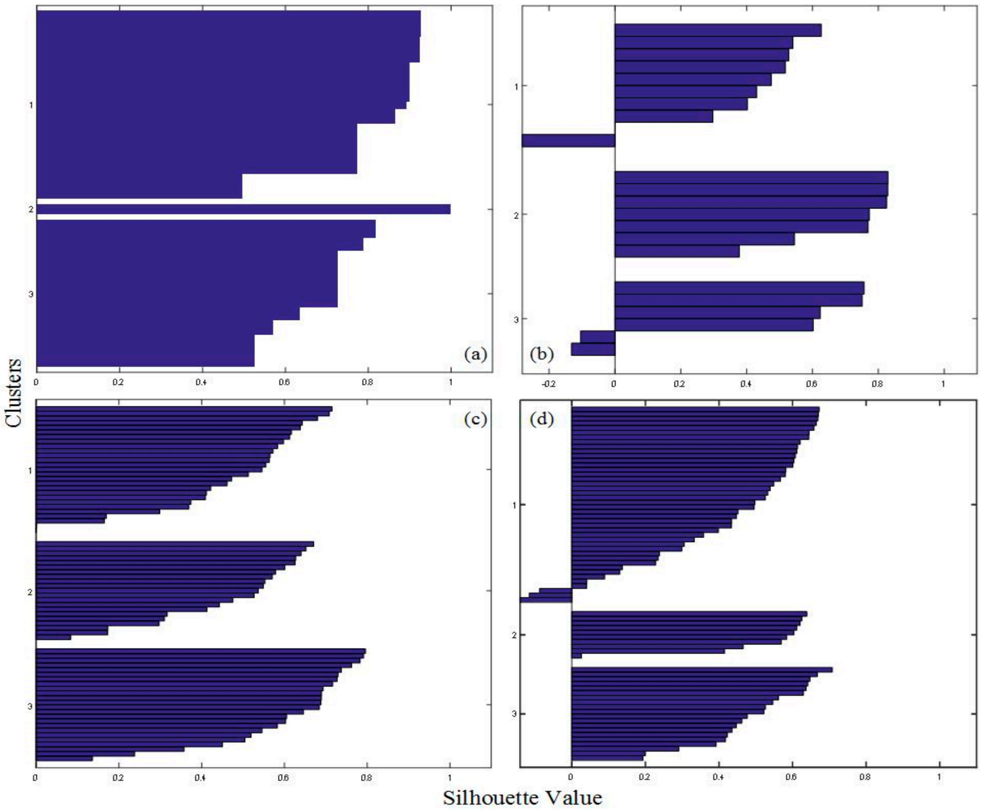

Figure 6. K-means cluster analysis results of Jungnangcheon stream data obtained during the three summer months and first fall month from 2014 to 2016. (**a**) Obtained data clusters from the summer and fall months for three consecutive years; (**b**) temperature clusters; (**c**) DO clusters; and (**d**) pH clusters.

3.3. Agglomerative Hierarchical Clustering

AHC classifies the urban stream behavior for 24-h diurnal cycles, by bottom-up merging into identical pairs of clusters. We also employed unsupervised AHC using the average linkage method to detect clusters in the data. The results are presented in the form of dendrograms. Two 24-h sampling data clusters, four temperature data clusters, eight DO clusters, and four pH clusters were detected, as shown in Figure 6.

The dendrogram shows a well-separated $\gimel_\ell$ cluster as shown in Figure 7a. A second cluster of data was subdivided into 23 h subsets of agglomerative clusters based on bits received to the MS, as shown in Figure 4. This demonstrates the greater overhead efficiency of our IoT platform for large data transmission. Thus, our Arduino setup successfully sensed, processed, and stored the raw data obtained by the ASM.

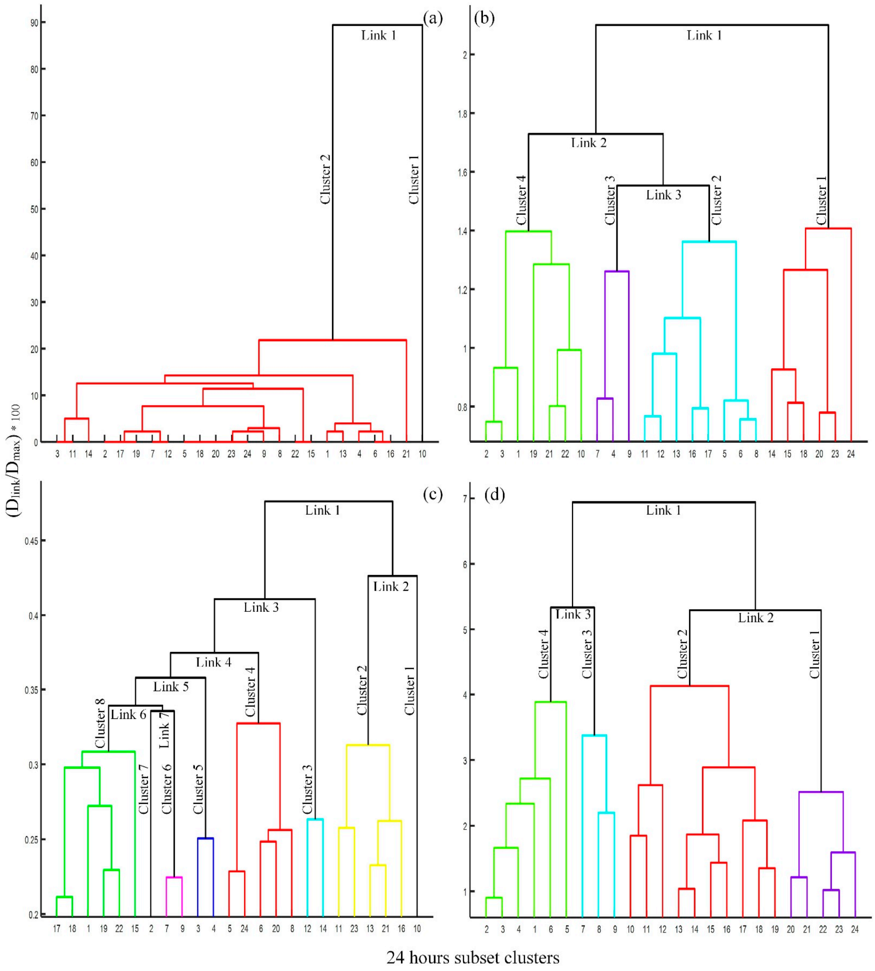

Figure 7. Dendrogram showing the hierarchical binary clusters of pdist time intervals of 24-h unsupervised static sampling site data collected via the ASM framework from Jungnangcheon stream during the three summer months and first fall month from 2014, 2015, and 2016. (**a**) Analysis of Physicochemical parameters (PcPs) obtained via the ASM during the experimental period; (**b**) nested stream temperature, in which red, blue, purple, and green lines represent the time of minimum, increased, moderate, and decline in stream water temperature, respectively; (**c**) dendrogram of DO, in which the black, yellow, blue, red and indigo, pink and black, and green line represent the time for the decreased, increased, minimum, maximum, initial increase in, and increased concentration level of stream respectively; (**d**) pH dendrogram, in which purple, blue, red, and green lines represent the drop, increase, peak and decrease in pH level.

Figure 7b shows three separate links joined with four clusters. Cluster 1 (red lines) represents the time of the minimum temperature in early evening and at night. Cluster 2 (blue lines) represents the time of increased temperature in early morning and afternoon until the evening. Cluster 3 (purple line) represents the time of moderate temperatures in early morning. Cluster 4 (green line) represents the time of minimum decline in temperature in midnight and late evening. This classification behavior of urban stream temperature into four clusters is based on diurnal time.

Figure 7c shows eight clusters connected through seven links. Clusters 1 (black lines) and 2 (yellow lines) show when the DO concentration decreased and increased, respectively. Cluster 3 (blue lines) shows when the DO concentration is at a minimum. Clusters 4 (red lines) and 5 (indigo lines) show when the DO concentrations is maximum. Clusters 6 (pink lines) and 7 (black lines) show the time of initial increase in DO concentration. Cluster 8 (green lines) shows when the DO concentration increased.

Figure 7d shows four clusters attached with three links. Cluster 1 (purple lines) shows a drop in pH level in the late evening until night. Cluster 2 (red lines) shows the increase in pH level in the morning to reach maximum level in early afternoon and then the decrease in late afternoon to early evening. Cluster 3 (blue lines) shows an increase in pH level in the early morning. Cluster 4 (green lines) shows the decrease in pH level reaches minimum level in midnight. The average pH value of stream water obtained was greater than 7, which shows the slightly basic nature of stream water with a small degree of diurnal variation in pH.

These results indicate that for assessment of stream PcPs, all four clusters are needed to perform an accurate analysis of the water quality for the urban stream. This technique requires fewer sampling sites and samples. It is also ascertained that the clustering technique is useful in analyzing the stream behavior and can be employed in water quality assessment.

3.4. Density-Based Spatial Clustering of Applications with Noise

DBSCAN was used to evaluate seasonal yearly variations in stream PcPs data. The arbitrary density value of noise for temperature, DO, and pH were calculated at different trial values. For these parameters, the trial values (eps in cm, MinPts) were set to $(1.9073^{-6}, 3)$, $\left(2.3842^{-7}, 12\right)$, and $(-50, 0.01)$, respectively. Noise was identified using the following three different color outlines: First the red outline identifies MDR and DDR. Greater variation in temperature was observed in 2016 than in 2014 and 2015. Second, the black outlines show the outlier points, mostly DR, marked in the figures. pH levels and DO concentration values were scattered in 2014, 2015, and 2016. Third, blue outlines display the dispersed noise that were not reachable. Noise were values within the same year; 2015 had the most outlier temperatures, DO values, and pH values.

Figure 8a shows the majority of noise in 2016 were ƥ, while noise in 2014 and 2015 were q. Noise reflects the reachability by MDR and DDR. Noise of temperature values was lower in 2016 on the x-axis (24.2) and y-axis (23.6) as compared with 2014 and 2015. Figure 8b shows noise in similar years were grouped together. Moreover, DO had outlier noise scattered within the purple outlines. DO concentration changed randomly with year, as shown from the dissimilar axis values. The outlines in Figure 8c are mostly black and purple. Moreover, only single DDR and DR points were present, indicating no significant change in pH level over the seasons. In short, the temperature points in 2016 were increased, while the increase in temperature points in 2014 and 2015 was not pronounced. This shows a trend of a steady increase in urban stream temperature. Based on the results obtained, it is demonstrated that a targeted sampling collection procedure can be designed to monitor stream PcPs with greater accuracy.

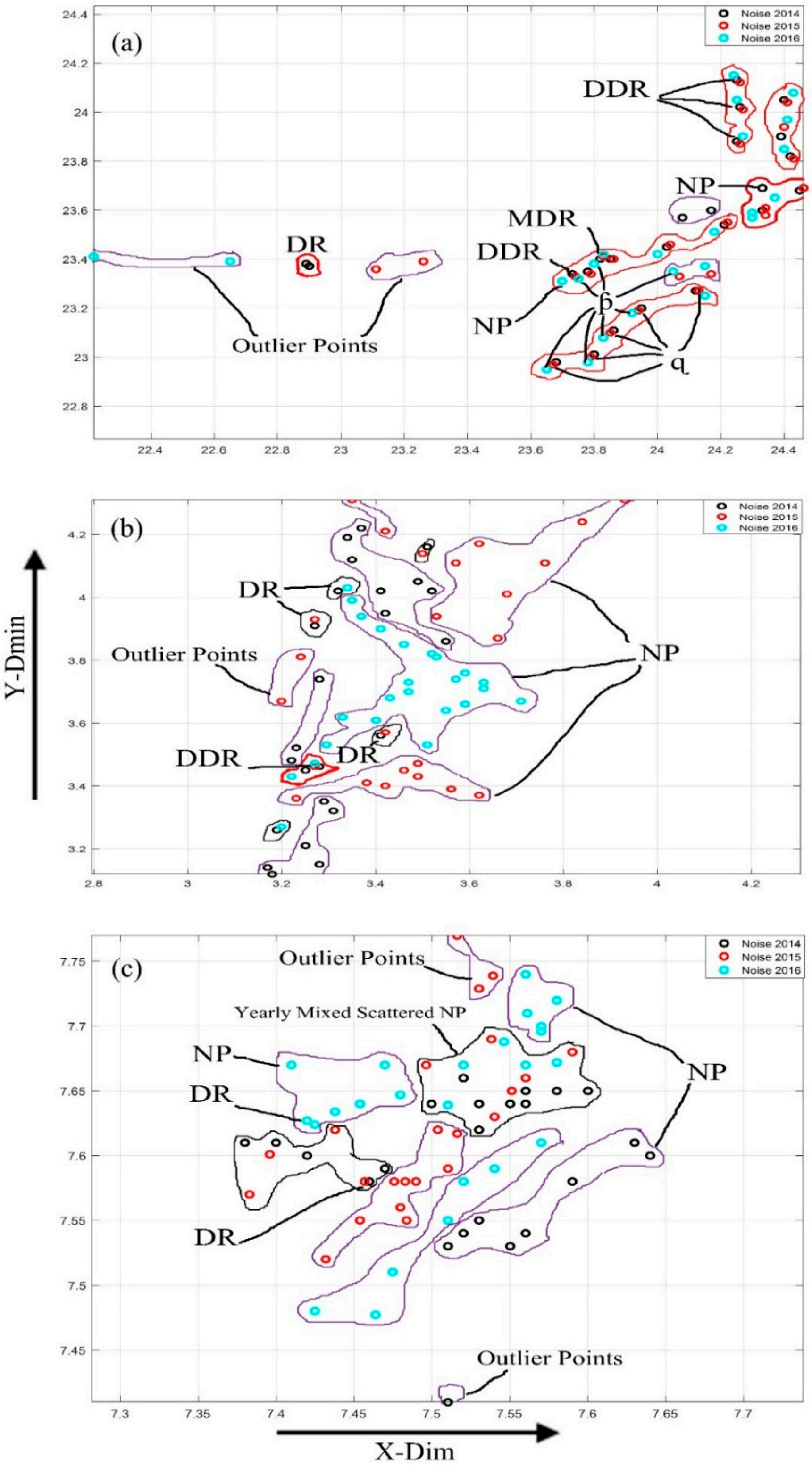

Figure 8. Density-based spatial clustering of applications with noise (DBSCAN) analysis of seasonal variation in stream parameters. eps is in cm, and MinPts is the number of points in the group N_noise. (**a**) Temperature DBSCAN clustering where eps = 1.9073×10^{-6} and MinPts = 3; (**b**) DO DBSCAN clustering where eps = 2.3842×10^{-7} and MinPts = 12; and (**c**) pH DBSCAN clustering where eps = -50 and MinPts = 0.01. DR: Density Reachability; DDR: Directly Density Reachability; MDR: Maximum Density Reachability; NP: NOISE/Noise Points; ɋ; Core Points; ƥ: Border Points.

4. Conclusions

In this study, we developed the IoT-based Arduino platform for continuous monitoring of Jungnangcheon stream. Monitoring was conducted for collected data every second every day for five days a week for four months (June, July, August, September) in 2014, 2015, and 2016.

(1) Three parameters, temperature, DO, and pH, were measured at a fixed location with 99.99% efficiency using an IoT Arduino platform. Simplified information was provided to residents (end users) on their smartphones. Hence, the proposed IoT platform is highly efficient and reliable in data transmission.

(2) AHC analysis segmented all data into two clusters, temperature into four clusters, DO into eight clusters, and pH into four clusters. AHC did not provide significant results; however, the optimal time for monitoring individual samples was identified allowing for a reduction in the number of sampling sites.

(3) DBSCAN results showed that temperature points in 2016 were þ (increased), while temperature points in 2014 and 2015 were ᑫ (no significant change). The measures showed a trend toward an increase in global temperature. Therefore, a targeted sampling collection can be designed to monitor stream PcPs.

(4) Our results indicated streams can be monitored and the collected data interpreted through data mining. This interpreted information can be shared on smart devices, such as smartphones, smart screens, and navigation devices.

(5) We performed monitoring using the IoT prototype based on the Arduino shield, only installed a handful of sensors, and only monitored conditions over a period of four months. However, the results indicated this application can help identify seasonal behavior and efficiently monitor PcPs in a low-cost manner.

(6) Replication of this work could establish a procedural framework for the Ministry of Environment, Republic of Korea, to allow monitoring of civil infrastructure through intelligent monitoring networks. Information can be shared with end users on their smartphones, which may also benefit researchers.

Author Contributions: Zafar Baloch designed and performed the experiment. Zafar Baloch also performed the data analysis and wrote the manuscript. Byungwan Jo conceived the idea and provided the materials. Byungwan Jo is a Professor in the Department of Civil and Environmental Engineering Lab of NT-IT (Nanotechnology and Information Technology) at Hanyang University, Seoul, South Korea. Zafar Baloch is the corresponding author and a Ph.D. candidate in the Department of Civil and Environmental Engineering of Hanyang University, Seoul, South Korea. He is also a faculty member at the Department of Civil Engineering, Faculty of Engineering and Architecture, BUITEMS, Quetta, Balochistan, Pakistan. At the time this paper was written, he was on sabbatical in South Korea.

Conflicts of Interest: The authors declare no conflicts of interest. Study sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Shrestha, S.; Kazama, F. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environ. Model. Softw.* **2007**, *22*, 464–475. [CrossRef]
2. Jackson, F.L.; Malcolm, I.A.; Hannah, D.M. A novel approach for designing large-scale river temperature monitoring networks. *Hydrol. Res.* **2016**, *47*, 569–590. [CrossRef]
3. Gasperi, J.; Sebastian, C.; Ruban, V.; Delamain, M.; Percot, S.; Wiest, L.; Mirande, C.; Caupos, E.; Demare, D.; Kessoo, M.D.; et al. Micropollutants in urban stormwater: Occurrence, concentrations, and atmospheric contributions for a wide range of contaminants in three french catchments. *Environ. Sci. Pollut. Res. Int.* **2014**, *21*, 5267–5281. [CrossRef] [PubMed]
4. Koklu, R.; Sengorur, B.; Topal, B. Water quality assessment using multivariate statistical methods—A case study: Melen river system (Turkey). *Water Resour. Manag.* **2010**, *24*, 959–978. [CrossRef]

5. Granell, C.; Havlik, D.; Schade, S.; Sabeur, Z.; Delaney, C.; Pielorz, J.; Uslander, T.; Mazzetti, P.; Schleidt, K.; Kobernus, M.; et al. Future internet technologies for environmental applications. *Environ. Model. Softw.* **2016**, *78*, 1–15. [CrossRef]
6. Iyigun, C.; Turkes, M.; Batmaz, I.; Yozgatligil, C.; Purutcuoglu, V.; Koc, E.K.; Ozturk, M.Z. Clustering current climate regions of turkey by using a multivariate statistical method. *Theor. Appl. Climatol.* **2013**, *114*, 95–106. [CrossRef]
7. Lee, H.S.; Lee, J.H.W. Continuous monitoring of short term dissolved oxygen and algal dynamics. *Water Res.* **1995**, *29*, 2789–2796. [CrossRef]
8. Huang, X.; Yi, J.; Chen, S.; Zhu, X. A wireless sensor network-based approach with decision support for monitoring lake water quality. *Sensors* **2015**, *15*, 29273–29296. [CrossRef] [PubMed]
9. Sun, S.; Barraud, S.; Castebrunet, H.; Aubin, J.B.; Marmonier, P. Long-term stormwater quantity and quality analysis using continuous measurements in a french urban catchment. *Water Res.* **2015**, *85*, 432–442. [CrossRef] [PubMed]
10. Storey, M.V.; van der Gaag, B.; Burns, B.P. Advances in on-line drinking water quality monitoring and early warning systems. *Water Res.* **2011**, *45*, 741–747. [CrossRef] [PubMed]
11. Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; Glance, N. Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, USA, 12–15 August 2007; pp. 420–429.
12. Lillesand, T.; Kiefer, R.W.; Chipman, J. *Remote Sensing and Image Interpretation*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
13. Demars, B.O.; Manson, J.R. Temperature dependence of stream aeration coefficients and the effect of water turbulence: A critical review. *Water Res.* **2013**, *47*, 1–15. [CrossRef] [PubMed]
14. Ng, R.T.; Han, J. Efficient and effective clustering methods for spatial data mining. In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, Chile, 12–15 September 1994; Morgan Kaufmann Publishers Inc.: Santiago; pp. 144–155.
15. Chang, H. Spatial analysis of water quality trends in the han river basin, South Korea. *Water Res.* **2008**, *42*, 3285–3304. [CrossRef] [PubMed]
16. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR, USA, 2–4 August 1996; pp. 226–231.
17. Tran, T.N.; Drab, K.; Daszykowski, M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chem. Intell. Lab. Syst.* **2013**, *120*, 92–96. [CrossRef]
18. Tan, P.N.; Steinbach, M.; Kumar, V. *Data Mining Cluster Analysis: Basic Concepts and Algorithms*; Addison-Wesly: Reading, MA, USA, 2013.
19. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2016.
20. Satsangi, J.; Silverberg, M.S.; Vermeire, S.; Colombel, J.F. The montreal classification of inflammatory bowel disease: Controversies, consensus, and implications. *Gut* **2006**, *55*, 749–753. [CrossRef] [PubMed]
21. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
22. Chung, Y. A simulation of the oxygen profile in the han river. *Yonsei Med. J.* **1975**, *16*, 29–39. [CrossRef] [PubMed]
23. Ministry of Environment, R.o.K. *Water Policies & Innovative Practices Republic of Korea 2004*; Ministry of Environment: Seoul, Korea, 2004; p. 14.
24. Ministry of Environment, Republic of Korea. *Ministry of Environment*; Ministry of Environment: Sejong City, Korea, 2015; p. 40. Available online: http://eng.me.go.kr/eng/file/readDownloadFile.do?fileId=115224&fileSeq=1&openYn=Y (accessed on 15 March 2017).
25. An, Y.-J.; Lee, J.-K.; Cho, S. Korean water quality standards for the protection of human health and aquatic life. In Proceedings of the 2nd International Forum on Water Environment Partnership in Asia, Beppu City, Japan, 3–4 December 2008.
26. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011.
27. Rousseeuw, P.J. Silhouettes—A graphical aid to the interpretation and validation of cluster-analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]

28. Mcquitty, L.L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educ. Psychol. Meas.* **1957**, *17*, 207–229. [CrossRef]
29. Ng, R.T.; Han, J.W. Clarans: A method for clustering objects for spatial data mining. *IEEE Tran. Knowl. Data Eng.* **2002**, *14*, 1003–1016. [CrossRef]
30. Ankerst, M.; Breunig, M.M.; Kriegel, H.-P.; Sander, J. Optics. In *ACM Sigmod Record*; ACM: New York, NY, USA, 1999; Volume 28, pp. 49–60.
31. Daszykowski, M.; Serneels, S.; Kaczmarek, K.; Van Espen, P.; Croux, C.; Walczak, B. TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chem. Intell. Lab. Syst.* **2007**, *85*, 269–277. [CrossRef]
32. Eaton, J.G.; Scheller, R.M. Effects of climate warming on fish thermal habitat in streams of the United States. *Limnol. Oceanogr.* **1996**, *41*, 1109–1115. [CrossRef]