water

MDPI

Article

An Automated Approach to Groundwater Quality Monitoring—Geospatial Mapping Based on Combined Application of Gaussian Process Regression and Bayesian Information Criterion

Dmitrii Shadrin [1,*,†], Artyom Nikitin [1,†], Polina Tregubova [1], Vera Terekhova [2], Raghavendra Jana [1], Sergey Matveev [1] and Maria Pukalchik [1]

[1] Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, 121205 Moscow, Russia; artem.nikitin@skolkovotech.ru (A.N.); p.tregubova@skoltech.ru (P.T.); R.Jana@skoltech.ru (R.J.); s.matveev@skoltech.ru (S.M.); m.pukalchik@skoltech.ru (M.P.)

[2] Faculty of Soil Science, Lomonosov Moscow State University, 119991 Moscow, Russia; vterekhova@gmail.com

* Correspondence: Dmitry.Shadrin@skolkovotech.ru

† These authors contributed equally to this work.

Abstract: Sustainable management of the environment is based on the preservation of natural resources, first of all, freshwater—both surface and groundwater—from exhaustion and contamination. Thus, development of adequate monitoring solutions, including fast and adaptive modelling approaches, are of high importance. Recent progress in machine learning techniques provide an opportunity to improve the prediction accuracy of the spatial distribution of properties of natural objects and to automate all stages of this process to exclude uncertainties caused by handcrafting. We propose a technique to construct the weighted Water Quality Index (WQI) and the spatial prediction map of the WQI in tested area. In particular, WQI is calculated using dimensionality reduction technique (Principal Component Analysis), and spatial map of WQI is constructed using Gaussian Process Regression with automatic kernel structure selection using Bayesian Information Criterion (BIC). We validate our approach on a new dataset for groundwater quality in the New Moscow region, where groundwater is mostly used for drinking purposes. According to estimated WQI values, groundwater quality across the study region is relatively high, with few points, less than 0.5% of all observations, severely contaminated. Estimated WQIs then were used to construct spatial distribution models, GPR-BIC approach was compared with ordinary Kriging (OK), Universal Kriging (UK) with exponential, Gaussian, polynomial and periodic kernels. Quality of models was assessed using cross-validation scheme, according to which BIC-GPR approach showed better performance on average with 15% higher R^2 score comparing to other Kriging models. We show that the proposed geospatial interpolation is a potentially powerful and adaptable tool for predicting the spatial distribution of properties of natural resources.

Keywords: water quality; kriging; PCA-loading index; gaussian process regression; bayesian information criterion

Citation: Shadrin, D.; Nikitin, A.; Tregubova, P.; Terekhova, V.; Jana, R.; Matveev, S.; Pukalchik, M. An Automated Approach to Groundwater Quality Monitoring—Geospatial Mapping Based on Combined Application of Gaussian Process Regression and Bayesian Information Criterion. *Water* **2021**, *13*, 400. https://doi.org/10.3390/w13040400

Received: 23 December 2020
Accepted: 28 January 2021
Published: 4 February 2021

1. Introduction

It is hard to overestimate the importance of role of freshwater resources for our planet. Global climate change and the rise of human population lead to extensive land-use changes, expanding urbanisation and industrial activity, have already heightened the risk of pollution of freshwater sources [1–3]. Assessment and monitoring of fresh water resources are necessary for decision-makers at all levels regarding economic, health, scientific, and ecological spheres of life [4–7].

Two principal approaches to the assessment of natural water bodies' status can be distinguished. The first one operates the concept of water reservoirs' "vulnerability". It

considers features of the environment, that can affect the quality of water resources. Vulnerability assessment takes into account both natural characteristics (geological, hydrological) of the environment, their ability to provide "protection" to water resources, as well as possible contamination scenarios due to properties of pollutants [8–10]. This approach allows to understand underlying processes and, according to them, to predict the possible fate of contaminants. Such outlook can be treated as "preventive", or "explainable", and relates more to the potential of the territory. The second approach is the assessment of the water quality using the real existing data of water composition itself. Such data usually include various chemical, physical, biological characteristics of water that are important for consumption and usually measured by local governance and residential consumers. This work is dedicated to this second type of approach and the type of available data.

Water Quality Indices (WQIs) are composite indicators of water quality that combine complex data into an aggregate value that can be quickly and easily communicated to its intended audience (CCME 2001). There are two main ways how to pool all measured parameters into one: (i) via expert bias, when the determinant set of water parameters in WQI is established based on expert opinion, [11]; (ii) via data-driven approaches, helping to reveal the weights assigned to each parameter [12,13]. The existing body of research on water quality assessment suggests that classical expert bias approaches poorly response to outer dynamic processes, whether it changes in water state or in law regulation. Meanwhile, the aggregated nature of WQIs is very attractive to both scientists and policy-makers. Application of different multivariate statistical techniques, such as Principal Component Analysis (PCA) and factor analysis (FA) could help to reveal the priority contaminants in water more effectively, and, at the same time, this approaches make the process less subjective.

Another important stage in environmental assessment is the prediction of the distribution of natural systems' patterns and characteristics. The last decade, although not being novel by itself, machine learning (ML) techniques—modelling approaches based on data learning—have become more popular in environmental issues. ML has been successfully employed in various environmental modelling studies, in terms of both spatial and temporal interpolation. For example, support vector machine (SVM) and artificial neural network (ANN) were implemented for soil pollution index prediction [14], hybrid solution based on SVM and spatial statistics methods successfully predicted and simulated suspended sediment load [15], another hybrid approach, based on support vector regression, principal component analysis and back propagation ANN was developed to establish relationships between water quality changes and lake water surface temperature [16], integrating approach based on ANN and kriging was proposed for spatial soil organic matter prediction [17], single-output and simultaneous ANNs showed their effectiveness for spatial multiparameter interpolation at single- and multi-active monitoring sites [18].

Gaussian Process regression (GPR), known as kriging in geostatistics, is one of the most popular ML interpolation approaches for environmental studies including soil [19,20], water [21] and air media [22,23]. GPR is considered as a powerful and flexible predictive tool, allows to catch the spatial distribution of data, to handle complicated non-linear problems, and quantify the reliability of predictions [24–26]. Central concept in kriging—is a variogram model, or, correspondingly, kernel, a plausible interpolation function for the spatial covariances, need to be computed and fitted from the data [27,28]. In terms of computation speed and reproducibility, kriging techniques are restrained by the necessity to select a variogram, as well as a list of model hyperparameters, such as nugget, sill, range, while their variations may change the interpolation results significantly [29,30]. Another bottleneck is the computation time in case of handling data-sets with large number of observation points [27]. Recent advances in kriging now consider the unification demand—with that approaches to avoid the uncertainty of variogram hyperparameters, not the data, have been proposed, mostly being based on the idea of choosing one optimal kernel. One of the possible solutions is the suggestion and evaluation of multiple semivariogram models with subsequent cross-validation [21,31] to choose the most suitable one, which is even have been introduced already in GIS environments or proposed as a part of an

independent interpolation decision tree [32]. More comprehensive solutions were proposed, for example, Approximate Bayesian Computation (ABC), which allows to automatically deal with kernel selection and estimation of hyper-parameters and to avoid the dependence on dimensionality when operating large numbers of kernels with different dimensions [33], with simultaneous comparing of competing models.

Therefore, an automated kernel selection is highlighted as a promising research direction to avoid manual operations, raise robustness to obtain optimal results, speed up calculations, especially when most of the existing solutions are still balancing between the accuracy of interpolation and computational speed.

In this paper, we discuss the approach to water quality spatial prediction based on the modified GPR and exploratory assessment via overall variance, on the example of groundwater monitoring net. We believe that successful implementation of this approach is especially important to subsurface water resources. Groundwater is directly connected with both surface water resources and terrestrial ecosystems at the same, but hidden from rapid observation [34]. Moreover, due to restricted access to direct measurements, groundwater monitoring systems are often represented by relatively small observation nets, where kriging techniques are also the most powerful. The core idea is to introduce the Water Quality Index (WQI) that will first include the most influential environmental parameters representing environmental pollution and second, take into account the variability characteristics of these parameters that can serve as the measure for risk assessment. To observe the patterns of spatial distribution, we developed an extension to the standard Gaussian process regression (GPR) model based on automated composite kernel search implemented on a greedy algorithm with the usage of Bayesian Information Criterion; hyper-parameters selection for each elementary kernel in the optimal composite kernel was done by standard approaches. By doing this, we exclude handcrafting in the selection of potentially effective kernels. BIC criteria allows to find a simple kernel structure among possible with the fewest number of parameters but with highest accuracy at the same time according to the validation sample, compared to other variants. In contrast, without BIC criteria there will be much more parameters which will lead to overfitting and much more complex kernels.

The paper is organised as follows: Section 2 is devoted to the description of the utilised experimental and computational methods which we exploit for the organisation and processing of the dataset. In particular, we describe the details about the target region and the available dataset in Section 2.1. The details of data processing and machine learning techniques (PCA, Gaussian processes, and Bayesian methods) are discussed in Sections 2.3 and 2.4. The main results are presented in Section 3 and discussed in Section 4. To be more specific, we present an automated approach for optimisation of parameters for the kriging procedure which allows to improve significantly the quality of reconstructed freshwater maps in comparison with popular standard tools widely used by community (the details can be found therein). The final discussion sums up the advantages and drawbacks of the proposed approach for freshwater quality map reconstruction and discovers the possible directions of future research.

2. Materials and Methods

2.1. Site Description and Available Dataset

The data used in the current study was collected in the territory of The New Moscow district, located adjacent to the city of Moscow in the Central European part of Russia. The area of New Moscow district extends over 1480 km^2, latitude and longitude ranges are approximately from 55°09′ to 55°40′ N and 36°48′ to 37°36′ E, respectively.

The climate of the region is moderately continental. According to the data from local weather stations (available at https://www.ncdc.noaa.gov/cdo-web/datatools/findstation) through the period of last five years, from 2015 to 2020, across the territory of study, the mean daily temperature in the coldest month of the year, January, was −5.4 °C, while in the warmest month, July, the mean daily temperature was +17.8 °C. The average

annual precipitation was approximately 440 mm, with circa two thirds rainfall and the rest snow. The territory is mostly located on the verge of the southern taiga change to the zone of temperate broad-leaf forests.

New Moscow is located in the central part of East European Plain. It is composed by the thick sedimentary rocks covering Precambrian cristallic basement. Sedimentary rocks include dolomite, lime, marl, sandstone. Bedrock rarely outcrops and mostly covered by Quaternary deposits of glacial and fluvioglacial genesis with occurrence of alluvial deposits [35]. Thus, across the region there are two main aquifers: the first is in Quaternary sediments, with thickness in a range from 2 to 12 m. This is the main source of the drinking water for the cities and households. The artesian basin is consists of the aquifers of coal age composed by limestone and dolomite. In general, across the New Moscow region aquifers are composed by different-grained, well-permeable sands and sandy loams, periodically covered by the clay of different ages [36]. Main land-use types are as follows: forests (50% from total area), arable land (21%) and low density discontinuous urban fabric (19%) [37]. The predominant types of natural vegetation are coniferous and broad-leaved forests, while agricultural lands include pastures and arable land, mostly growing feed crops and cereals. The main specific of New Moscow is that it has been rapidly urbanising during the last decade.

Sample net, using in this study, covers almost all territory of New Moscow. A total of 1600 water samples were collected during 2017–2018, mostly from May to September, from wells (1215 samples), rivers (225 samples) and springs (160 samples) in the region (see Figure 1). Water samples were collected from the wells, rivers, or springs by using a 2-L stainless-steel container. The samples were bottled and then immediately transported to the laboratory for chemical analysis, eliminating the need for conservation methods.

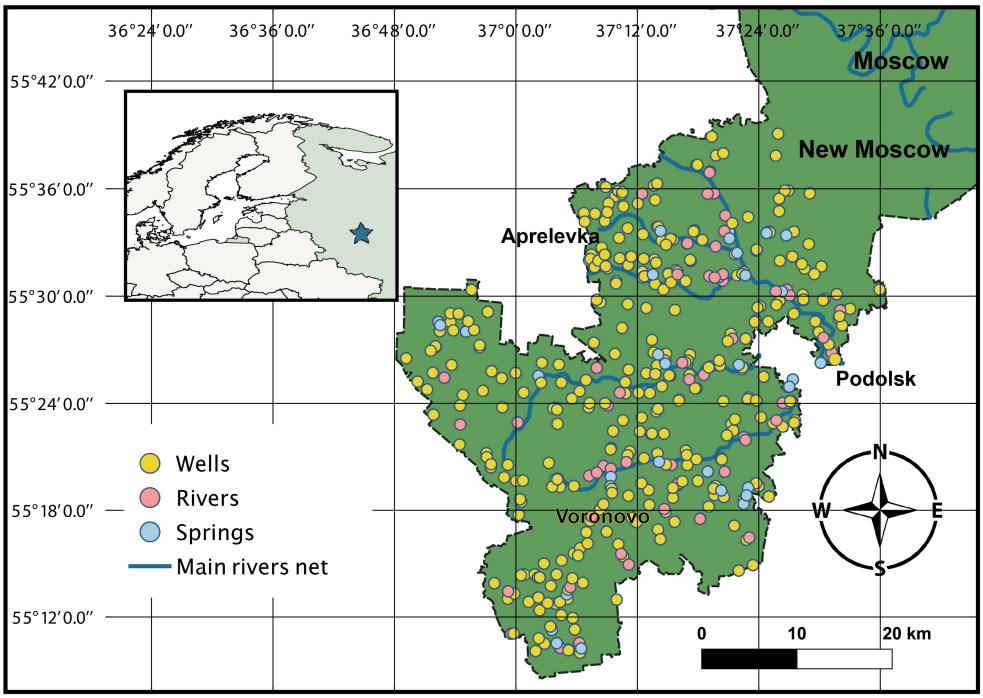


Figure 1. Location map of the study area. Different colors mark source of collected water samples—wells colored in yellow; rivers colored in pink; and springs colored in blue. Blue lines represent main river streams from open-source 10-m resolution data.

For each water sample, 25 parameters were measured. The pH was measured by using a HANNA pH-meter 213. Anions (NO_3, NO_2, PO_4, SO_4 and Cl) were measured by ion chromatography using a Dionex 1100 instrument. NH4 content was obtained on an HACH DR2800 using colorimetric determination with Nessler's reagent. Cation (K, Cr, Ni, Ca, Zn, Fe, Mn, Na, Cu, Mg) contents were obtained by inductively coupled plasma

atomic emission spectroscopy on an ICP-OES Agilent 5110 spectroscope. Mineralization was measured by gravimetric analysis consisting of evaporation at 105 °C in a drying chamber. Alkalinity was obtained by titration with 0.05 NHCl. Hardness was measured by titration with Trilon B and eriochrome black. Despites being minor but not the least the important advantage of this work is relatively large size of the dataset which contains more than 1600 samples (each with 25 measured chemical parameters). We hope that it might be useful for validation and other methodological research in community and share it as open data [38].

2.2. Data Preparation and Methodology

In this study, an end-to-end solution for geospatial water quality assessment using ML methods such as Gaussian Process Regression and Bayesian Information Criterion was proposed and evaluated. Figure 2 presents a brief summary of the steps involved in this procedure.

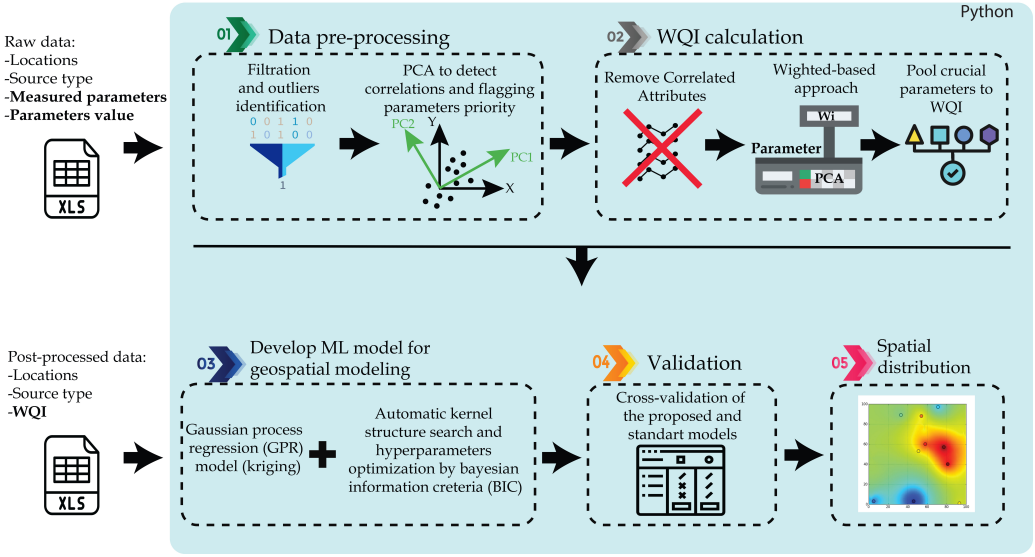


Figure 2. Our approach use machine learning methods for weighted WQI calculation (Steps 1 and 2) and geospatial WQI prediction by using Gaussian process regression with automatic kernel search (Steps 3–5).

2.3. Water Quality Index Calculation Based on PCA and Weighted Factors

In order to construct WQI we performed PCA and used the following outputs: principal components (PCs), eigenvalues, loadings, described variance. The brief description of the theory behind PCA is given in the Section 2.3.1. Description of the approach to WQI calculation is given in the Section 2.3.2.

2.3.1. PCA Theory

The full theory behind the PCA is quite comprehensive, so we address the reader to the relevant literature sources for the full explanation [39–41]. In order to reveal the main ideas of using of PCA for current study, we discuss the main outputs extracting in the analysis. PCA is the linear unsupervised method of dimensionality reduction of multivariate data, proposed by Hotelling [42], where original dimensions are initial (measured) variables. PCA projects the data into new coordinate system and representing it in lower dimensions set in a way retaining the maximum of the variation of the original data, so keeping the most of the information.

The new set of dimensions — principal components, PCs — is mutually orthogonal, so PCs are uncorrelated to each other. PCs are the linear combinations of the original variables and ranked by the variance of the data along. To obtain PCs, at first, the covariance matrix of the original data is estimated. This covariance matrix is symmetric and consists of set

of orthogonal components—eigenvectors, or "principal directions", expressed by their own eigenvalues reflected to the measure of the variance related to the PC. Thus, the first PC reflects to the maximum of the data variance and so on in the descending order. The contributions of the each initial variables into the PC are loadings. To strengthen the interpretability of the PCs an additional manipulation can be performed, such as Varimax rotation [43,44]. In this purpose the variance of squared loadings is seeking to be maximized, thus each PC has only few, non-overlapping variables with the highest loadings.

2.3.2. Construction of WQI Based on PCA

As was discussed previously, it might be found very useful to express all of the multivariate water quality complexity in one conscientious parameter, which is WQI, keeping as much information as possible.

A PCA model was used to assess the pollutant loads integral to water quality and to avoid data redundancy. Raw data was filtered to eliminate anomalies: missing coordinates, incorrect record type etc. After this initial pre-processing step, the total number of useful samples decreased from 1600 to 1569. We decided to remove Hg, Cd, Co and Pb from the dataset for further analysis, as their concentrations were insignificant (much lower than toxic levels) and did not exceed the required water quality standards in Russia. Twenty-one water quality parameters were included in the PCA model. Only those components for which the corresponding eigenvalue was higher than or equal to 1 following *Varimax* rotation, and PCs that explained at least 5% of the observed data variation were considered for further examination. Moreover, those parameters that were correlated with other significant parameters (correlation between two particular parameters is more than 0.6) were eliminated only if they had the smallest loadings among correlated parameters. The weight scores (w_i) derived from PCA were used as weighted factors for the significant variables (indicators) from the respective PCs, and the WQI was calculated by using the Equation (1):

$$\mathrm{WQI} = \sum_{i=1}^{S} L_i \cdot w_i, \tag{1}$$

where S is the number of significant principal components, L_i denotes the loading values of each selected water property included in the particular principal component, and w_i denotes the weight of the corresponding component, which is defined as the part of the described variance by each component. To scale WQI to the [0,1] range, we normalized the weight scores (w_i) to a summarized score value by using Equation (2):

$$w_i := \frac{w_i}{\sum_{i=1}^{S} w_i} \tag{2}$$

2.4. Machine Learning Approach for Geospatial Modelling of WQI with Automatic Kernel Detection

2.4.1. Gaussian Process Regression: General Overview of the Methodology

To perform geospatial modeling of multiple water properties from the collected dataset, we referred to the *Gaussian Process Regression* (GPR) framework [45], more commonly known as *kriging* in geostatistics. A Gaussian process is completely determined by its *mean* $\mu(\cdot)$ and *covariance* (kernel) $k(\cdot,\cdot)$ functions:

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x},\mathbf{x}')), \\ \mu(\mathbf{x}) &= \mathbb{E}\, f(\mathbf{x}), \\ k(\mathbf{x},\mathbf{x}') &= \mathbb{E}\,[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))], \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^2$ is a vector of d input parameters. In our case, $d = 2$ and $\mathbf{x}$ represents a vector of spatial coordinates. Let us consider a simple GPR model with additive Gaussian noise:

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Given the training data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\intercal \in \mathbb{R}^{n \times d}$, $\mathbf{y} = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$, where n is the number of samples and $(\cdot)^\intercal$ denotes a transpose, the predictive distribution at the unobserved point $\mathbf{x}^*$ is given by

$$\begin{aligned} \hat{f}(\mathbf{x}_*) &\sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2), \\ \hat{\mu}(\mathbf{x}_*) &= \mu(\mathbf{x}_*) + k_* \Sigma (\mathbf{y} - \mu(\mathbf{X})), \\ \hat{\sigma}^2(\mathbf{x}_*) &= k(\mathbf{x}_*, \mathbf{x}_*) - k_*^T \Sigma^{-1} k_*, \\ \Sigma &= K + \sigma^2 I, \end{aligned} \tag{3}$$

where I is an identity matrix, $K = k(\mathbf{X}, \mathbf{X}) = k(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, N$ is a spatial covariance matrix between all of the training points, $k_* = k(\mathbf{X}, \mathbf{x}_*)$ is a spatial covariance between training points and the single prediction point and $\mu(\mathbf{X}) = \mu(\mathbf{x}_i)$, $i = 1, \ldots, n$ is the mean function calculated at the training points. The particular choice of the kernel function depends on the assumptions about the model and a particular application, e.g., widely used *Gaussian kernel* (corresponding to Gaussian variogram). Kernel hyper-parameters are usually optimized using *Maximum Likelihood Estimation* (MLE) [46] or its variations.

Figure 3 shows an example of GPR using Gaussian kernel with a constant mean function over the observations sampled from the sigmoid function with random noise. The predictive variance increases at points with missing observations, and increases significantly outside of the interpolation region with the mean failing to capture the true function trend. This emphasizes the need for a better method to select kernel hyper-parameters.

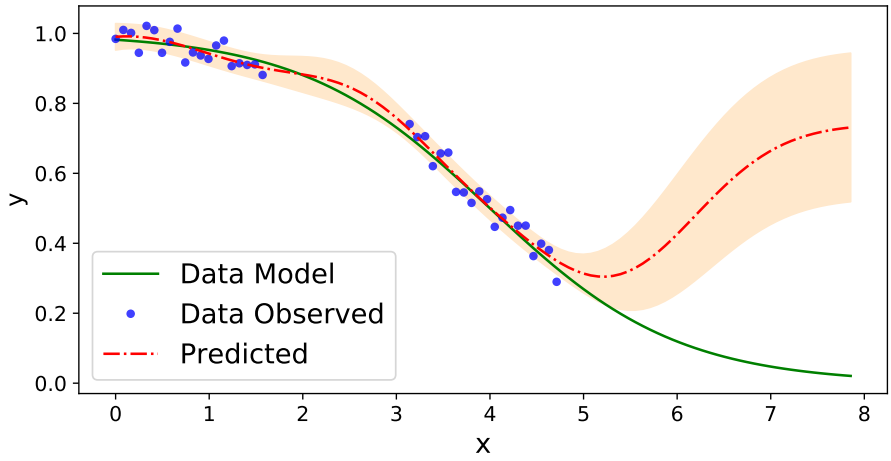


Figure 3. Gaussian Process Regression (red dashed line depicts the predictive mean and orange fill depicts the standard deviation intervals) with noisy measurements (blue dots) of the sigmoid function (solid green line) using Gaussian kernel and constant mean function.

2.4.2. Hyper-Parameters Selection Using Bayesian Information Criterion

Common approaches to hyper-parameter optimisation are *Maximum Likelihood Estimation* (known model, continuous parameters), and *Cross-Validation* (model is unknown, discrete parameters). Typically, one could select multiple combinations of different kernels, perform MLE for each of them and then compare the models using cross-validation to choose the best overall model. In our work we follow the approach from [47] using Bayesian Information Criterion (BIC), which considers kernel function as a combination of a small number of base covariance functions using sum and product operations, and can be represented as:

$$\begin{aligned} \text{BIC} &= -2 \cdot \text{Log-likelihood} + m \cdot \log n, \\ \text{Log-likelihood} &= -\frac{n}{2} \cdot \log 2\pi - \frac{n}{2} \cdot \log |\Sigma| - \frac{1}{2} \cdot (\mathbf{y} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu) \end{aligned} \tag{4}$$

where n is the number of samples, m is the total number of optimised parameters, and Σ is defined as in the Equation (3). To construct the optimal kernel, we consider a basic set of operations, such as *plus* and *multiplication*, and apply them to the following kernel functions: polynomial (Equation (5)), Gaussian (Equation (6)), periodic (Equation (7)) and exponential (Equation (8)). Thus, the final automatically constructed kernel, for example, can be the multiplication of the polynomial kernel on Gaussian, plus periodic, etc. Optimal kernel structures can include the multiplication of the same types of elementary kernels. The best kernel is a combination (structure) of the elementary kernels with optimized parameters that gives the minimal BIC value. This way, we are able to model a variety of stationary kernels and control the accuracy by selecting basic kernels and boundary values for their hyper-parameters. The main goal for introducing such boundary values is to avoid over-fitting and ensure the robustness of the performance of the obtained optimal kernel (composition of the basic kernels). Moreover, we aim to reduce the model complexity by decreasing the number of tuned hyper-parameters in the optimal kernel.

$$k_{poly}(\mathbf{x}, \mathbf{x}' | \, \theta_1, \theta_2, \theta_3) = \theta_1 \left(\sum_{i=1}^{d} \theta_2 \mathbf{x}_i \mathbf{x}'_i + \theta_3 \right)^{deg} \tag{5}$$

$$k_{gaussian}(\mathbf{x}, \mathbf{x}' | \, \theta_4, \ell) = \theta_4 \exp \left(-\frac{1}{2} \sum_{i=1}^{d} \frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{\ell_i^2} \right) \tag{6}$$

$$k_{periodic}(\mathbf{x}, \mathbf{x}' | \, \theta_5, s, T) = \theta_5 \exp \left(-\frac{1}{2} \sum_{i=1}^{d} \frac{1}{s_i} \sin^2 \left(\frac{\pi}{T_i} (\mathbf{x}_i - \mathbf{x}'_i) \right) \right) \tag{7}$$

$$k_{exp}(\mathbf{x}, \mathbf{x}' | \, \theta_6, l) = \theta_6 \exp \left(-\sqrt{\frac{1}{2} \sum_{i=1}^{d} \frac{(\mathbf{x}_i - \mathbf{x}'_i)^2}{l_i^2}} \right) \tag{8}$$

where $d = 2$ for our task (coordinates), polynomial was taken of degree 2; θ_1, θ_4, θ_5, θ_6 are the variances; θ_2, ℓ, l and s—length scales; θ_3—bias; T—period. In our experiments all of the kernels were considered isotropic and we applied the following constraints during hyper-parameter optimization: $T_1 = T_2 = T \in [1, 10]$, $s_1 = s_2 = s \in [0.1, 10]$ and $\ell_1 = \ell_1 = \ell \in [0.1, 10]$, other parameters were left unconstrained.

The procedure of fitting Gaussian Process is quite expensive O(n^3), where n is the number of training data points. Hence, instead of brute-force search of the best kernels, a greedy search was implemented in the current study. Greedy search in general means that the extension with the lowest BIC is selected for each extension of the current kernel. The main advantage of this approach is that it does not require any handcrafting of potentially effective kernels but instead enables an automatic search for the best kernel structure and hyper-parameter optimisation.

2.4.3. Universal and Ordinary Kriging

To compare our method with baseline geospatial modelling techniques we performed Ordinary Kriging (OK) and Universal Kriging (UK) using the *GPy* library. Since it only allows to perform Gaussian Process Regression and not kriging as is, we draw a connection between these methods as follows: (a) basic kernel functions k_{poly}, $k_{gaussian}$, $k_{periodic}$, k_{exp} correspond to respectively named variograms, (b) GPR with a constant mean function $\mu(\mathbf{x}) = \mu$ corresponds to OK, and (c) GPR with linear mean function $\mu(\mathbf{x}) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$ corresponds to UK with linear trend. Hyper-parameters of the kernel and mean functions are optimized using MLE approach during the training phase.

2.5. Approach to Geospatial Modelling

First, we converted the spatial coordinates from EPSG:4326 (latitude, longitude) format to EPSG:32637 (UTM zone 37 N) format and scaled them down to the [0,10] range. Some measurements of the water quality measurements were taken spatially far from the main investigated area (Moscow region). Thus, to filter outliers, we performed clustering

of the water sampling locations using the density-based DBSCAN method [48] with neighbourhood size $\epsilon = 1.0$. The parameter ϵ allows to tune the size of a cluster and serves as the measure of the permissible distance to the point to be included into the cluster. After clustering and removing of the outliers, we again re-scaled the coordinates to the [0,10] range. Finally, it was decided to use the data only from the major class (wells, 1215 data points). Totally, 391 data points were removed from the dataset (37 data points out of them were removed by DBSCAN) and 1178 data points left for further investigation. WQI was calculated (methodology is presented in Section 2.2) for each data sample and a rectangular 100×100 grid was used for geospatial modelling and mapping. The boundaries of the selected grid were defined by the minimum and maximum coordinates of the kept water sampling locations.

2.6. Validation Procedure

To validate our model, we applied a standard cross-validation scheme with 5 random splits into training and testing data sets of relative size 90% and 10%, respectively. For each training/testing split, we (a) fit the model to the training data, then, (b) predict the values of WQI for the test data point locations and (c) subsequently calculate the root mean square error (RMSE) and the coefficient of determination R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{10}$$

where $\hat{y}_i$, y_i are the predicted and observed values, respectively, and $\overline{y}$ is the average across the observed values.

The RMSE is a good comparative statistic for assessing model output, as it provides a global indication of how similar the interpolated values are to the observed or measured data point values [49]. When analysing the RMSE statistics, a small RMSE value indicates that the interpolated values for the output model are more similar to the observed data point values, whereas a large RMSE value suggests that the interpolated model values are less similar to the observed data points. Thus, RMSE was used here to determine how well the model fits the observed data values, with low RMSE values indicting a high degree of model accuracy [50,51].

2.7. Software

All the calculations were carried out in Python programming language using the following libraries: scikit-learn, [52], GPy, [53] and Folium.

3. Results and Discussion

3.1. PCA-Based Weighted Water Quality Index

We applied PCA to reveal the significant contaminants among samples and calculate weighted-loads of tested parameters in WQI. In total, we observed five PCs with loads above 1 and a cumulative variance of about 61% (see Table 1). Then, we eliminated the parameters of each PC that were correlated significantly with others and had the lowest loading's among them (see Figure 4). Finally, our WQI includes only non-correlated parameters with loadings greater than 0.3 to the contributed PCs. Varimax rotation was used for PCA calculation and helped us to reveal the PCs with the exact chemical properties of water, which were clearly interrelated and signalled specific types of pollution. As an example, the chemical indicators usually linked with organic pollution were coupled to PC3, whereas parameters of water mineralization were coupled to PC1 (please see Table 1).

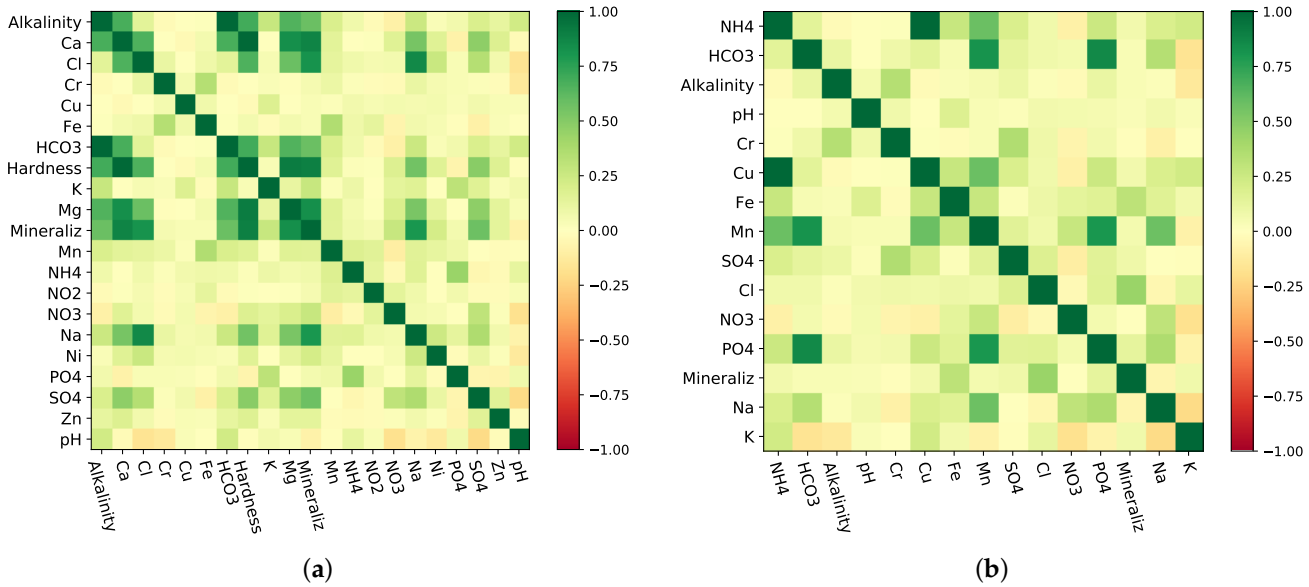
Table 1. Chemical components loading attributed to each PCs based on the PCA with Varimax rotation.

Principal Components	**Comp1**	**Comp2**	**Comp3**	**Comp4**	**Comp5**
Eigenvalues	6.116	2.057	1.856	1.543	1.237
Variance (%)	29.12	9.79	8.84	7.35	5.89
Cumulative variance (%)	29.12	38.92	47.76	55.10	61.00
Parameters loadings					
NH_4	0.0794	0.0041	**0.5602**	0.0279	−0.0603
HCO_3	−0.0363	**0.5385**	0.0041	0.0229	0.0137
Alkalinity	−0.0364	**0.5386**	0.0041	0.0228	0.0136
pH	−0.1731	**0.3074**	0.2065	−0.0889	−0.1959
Hardness of water	0.2960	0.2583	−0.1245	−0.0123	0.0035
Cr	0.0076	−0.0764	−0.0718	**0.5049**	0.1270
Cu	−0.1188	0.0103	0.0489	0.2093	**0.4262**
Fe	−0.0179	0.0199	−0.0408	**0.6504**	−0.0269
Mn	0.0557	0.0913	0.1145	**0.4557**	−0.1452
Ni	0.2217	−0.1376	−0.0030	−0.1010	−0.0475
Zn	−0.0368	0.1017	−0.1915	0.0638	0.1721
SO_4	0.1987	−0.0145	−0.1570	−0.0894	**0.3695**
Cl	**0.5033**	−0.1380	0.0726	0.0079	−0.1002
NO_3	0.0666	−0.1398	−0.0800	−0.1048	**0.5048**
NO_2	0.0518	−0.0645	0.1705	0.1495	0.0442
PO_4	0.0223	−0.0059	**0.6047**	−0.0642	0.1163
Mineralization	**0.3729**	0.1255	0.0228	−0.0215	0.1407
Ca	0.2973	0.2457	−0.1414	−0.0098	−0.0152
Mg	0.2552	0.2604	−0.0634	−0.0169	0.0540
Na	**0.4440**	−0.0817	0.1863	0.0101	−0.0330
K	−0.1235	0.1455	0.2777	−0.0010	**0.5150**

In fact, each PC contributed to a series of chemical parameters in the tested dataset. For example, the PC1 was linked to the chloride content, overall mineralization and sodium content of water (with loading's greater than 0.3). However, all three of these parameters were correlated: r(Na & Cl) = 0.856; r(Cl & Mineralization) = 0.819; and r(Na & Mineralization) = 0.800. Thus, the final shortlisted parameters from these PCs were a subset of the co-correlated parameters to prevent overlooked results and include only Cl. A similar case with co-correlated parameters was observed in parameters attributed to PC2. The PC2 revealed three main characteristics of water pollution: hydrocarbonates (HCO_3), alkalinity and pH. At the same time, only HCO_3 & Alkalinity were characterized by r as 1.0 , while two other parameters revealed low values of co-correlaton R^2(pH & HCO_3) = 0.227, r(pH & Alkalinity) = 0.228 and were included in the shortlisted parameters. All correlations among significant parameters for PC3, PC4 and PC5 were low (see Figure 4); thus, all parameters with $loads > 0.3$ (Table 1) were used for the WQI calculation. In detail, r(NH_4 & PO_4) = 0.437 in PC3, r(Cr & Fe) = 0.336, r(Cr & Mn) = 0.097 and r(Mn & Fe) = 0.353. The last PC, PC5, consisted of four significant parameters with extra low co-correlations: r(Cu & SO_4) = 0.0642, r(K & NO) = 0.1376, r(K & SO_4) = 0.1637, r(Cu & NO_3) = 0.0571, r(SO_4 & NO_3) = 0.2970 and r(Cu & K) = 0.1769.

Finally, our WQI was presented as a combination of 12 parameters with different normalized weighted factors:

$$\begin{aligned} \mathrm{WQI} = 0.2912 \cdot (\mathrm{Cl}) + 0.0979 \cdot (\mathrm{pH} + \mathrm{Alkalinity}) + 0.0884 \cdot (\mathrm{NH_4} + \mathrm{PO_4}) + \\ + 0.0735 \cdot (\mathrm{Cr} + \mathrm{Fe} + \mathrm{Mn}) + 0.0589 \cdot (\mathrm{Cu} + \mathrm{SO_4} + \mathrm{K} + \mathrm{NO_3}) \end{aligned} \quad (11)$$



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

Figure 4. The correlation heatmap for chemical parameters in tested freshwater samples. Figure (**a**) present correlation coefficient) between all measured chemical parameters, while figure (**b**) present correlation coefficient only for parameters with significant PCA loading. Initial number of water quality parameters for WQI constriction was reduced from twenty-one to fifteen after PCA.

All the parameters (concentrations) in the Equation (11) should be scaled to [0,1]. The resulting WQI after applying the Equation (11) should be again re-scaled to [0,1] before using it for the geospatial modeling. The distribution of the calculated WQIs among the tested samples is presented in the Figure 5. The mean WQI was 0.24 in the tested locations, and the median was 0.22. These values signalled that less than 0.4% of the tested samples were actually characterized as highly polluted, with a WQI > 0.75. Distribution of WQIs across the spatial coordinates—latitude and longitude—does not show any significant trends.

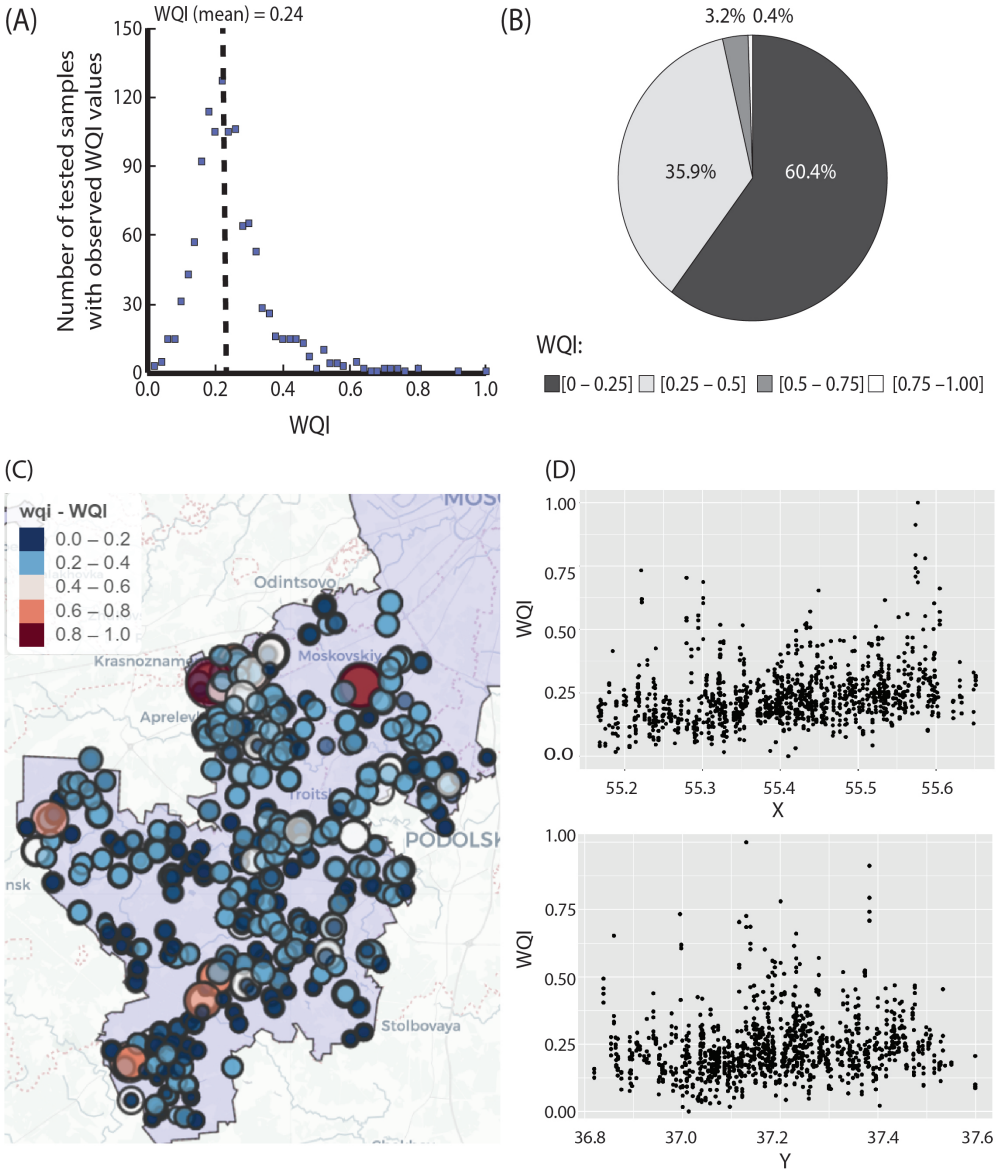


Figure 5. The overall distribution of WQI in sample points. (**A**) The graph presents the number of sample points with observed WQI and the mean value of the WQI. (**B**) Pie chart of statistical distribution of WQI for tested samples. (**C**) Distribution of points with estimated WQI across the study area, lower WQI values are corresponding to good groundwater quality, and higher—to poor groundwater quality. (**D**) Ratio of WQI to spatial coordinates: X—Latitude, Y—Longitude.

3.2. Geospatial Modeling

In this research, we proposed and validated technique based on GPR and kernel structure selection using BIC. The optimal kernel structure obtained by the BIC method was found to be a sum of Gaussian and periodic kernels (see Equation (12)). The optimized hyper-parameters can be found in Table 2 with $\ell_1 = \ell_2 = \ell$, $s_1 = s_2 = s$, $T_1 = T_2 = T$ (isotropic case).

$$k^*(x, x') = \theta_4 \exp\left(-\frac{1}{2} \sum_{i=1}^{2} \frac{(x_i - x_i')^2}{\ell^2} \right) + \theta_5 \exp\left(-\frac{1}{2} \sum_{i=1}^{2} \frac{1}{s} \sin^2\left(\frac{\pi}{T}(x_i - x_i') \right) \right) \quad (12)$$

Table 2. The optimal kernel parameters for the tested Gaussian kernel with periodical kernels.

Parameter	Value
Gaussian kernel variance, θ_4	0.0367
Gaussian kernel length scale, l	4.86
Periodic kernel variance, θ_5	0.0204
Periodic kernel period, T	5.67
Periodic kernel length scale, s	0.1

To validate our approach and compare it to baseline methods (OK and UK with different kernels), we applied the cross-validation scheme with 5 different train/test splits (90% and 10% train/test split). Table 3 shows corresponding R^2 and RMSE values obtained for different cross-validation splits. From the Table 3 it can be noticed that on the 2-nd split our proposed Kriging with BIC method gives much better results comparing to other approaches. Results, close to our Kriging with BIC approach were obtained by the exponential kernel in average, however on the second split Kriging with BIC outperformed exponential kernel, thus being more robust. The worst result represented by the negative coefficient of determination values was obtained by using the polynomial kernel. In general, the coefficient of determination can also be negative. The negative values of the coefficient of the determination indicate an inappropriate model which means that a simple averaging will give better interpolation results than by using the proposed model. The length scale of the kernels fitted in the standard methods is 1, while the variance for Gaussian kernels varies in range from 0.0032 to 0.0054. The variance for exponential kernel is 0.04. It can be seen, that our approach with optimal kernel selection gave the best R^2 compared to the standard kriging methods. RMSE for our model was comparable to other methods, however, the standard deviation of errors on the different validation data subsets was minimal compare to the other approaches which make our proposed method beneficial. In the case of RMSE assessment, it is also important to compare the obtained RMSE values with the average value for WQI in the tested dataset. As can be seen, the average value of WQI was 0.24 (Figure 5A); therefore, the calculated RMSE 0.065 indicated that proposed GPR model coupled with BIC is suitable for modelling. Finally, Figure 6 shows the results of geospatial modelling of WQI values and corresponding uncertainty maps, obtained by different approaches. The results clearly demonstrate the advantages of automatic kernel selection using BIC, allowing to recognize the local pollutant areas.

Table 3. Performance evaluation of selected models. Results of cross-validation of the obtained models on 5 different train/test splits.

		1	2	3	4	5	Mean	std
Kriging with BIC approach	R^2	0.729	0.487	0.609	0.641	0.702	0.637	0.098
	RMSE	0.060	0.072	0.071	0.062	0.059	0.065	0.0063
Ordinary Kriging Gaussian kernel	R^2	0.580	−0.075	0.599	0.625	0.575	0.461	0.300
	RMSE	0.068	0.076	0.056	0.060	0.059	0.064	0.0085
Universal Kriging Exponential kernel	R^2	0.610	0.014	0.604	0.646	0.622	0.499	0.271
	RMSE	0.070	0.077	0.056	0.060	0.058	0.064	0.0088
Universal Kriging Gaussian kernel	R^2	0.544	−0.052	0.600	0.631	0.590	0.463	0.289
	RMSE	0.071	0.076	0.055	0.059	0.058	0.064	0.0093
Universal Kriging Polynomial kernel	R^2	−11.205	−9.316	−11.042	−6.693	−9.860	−9.623	1.820
	RMSE	0.129	0.113	0.109	0.097	0.103	0.110	0.0122
Universal Kriging Periodic kernel	R^2	0.415	−0.038	0.579	0.637	0.593	0.437	0.278
	RMSE	0.080	0.076	0.057	0.059	0.058	0.066	0.0114

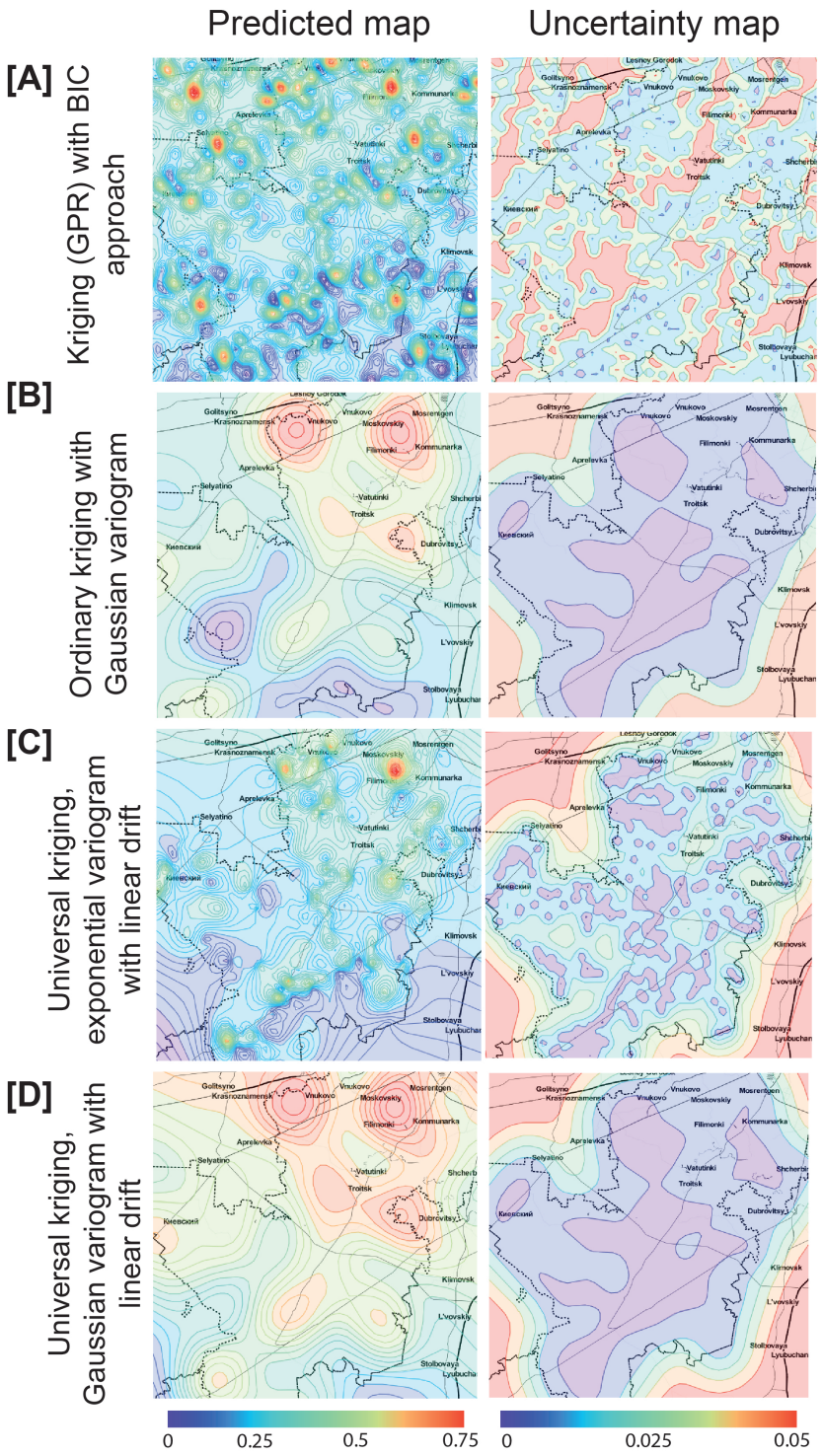


Figure 6. Geospatial prediction of Water quality index and uncertainty maps based on different techniques: (**A**)—GPR coupled with BIC; (**B**)—Ordinary kriging with Gaussian variogram; (**C**)—Universal kriging, Exponential variogram + linear drift; (**D**)—Universal kriging, Gaussian variogramm + linear drift.

From obtained results, we can see, that the spatial distribution of estimated WQI is not uniform, which can be explained by two main reasons: natural and artificial. e.g., as was described in Section 3.1 the ions of Cl and Na, Mineralization, and HCO_3, Alkalinity, pH mainly contribute to overall variance according to the PCA. We can suggest, that the driving factor of high variability of these characteristics is the aquifer composition itself

and basic rocks interacting with waters. The occurrence of trace elements (heavy metals), phosphate, nitrates (PC3, PC4, PC5 loadings) can also significantly contribute to low WQI. Their presence in waters can be explained by various reasons, but more likely, they are associated with the agricultural activity since all of the listed parameters are contained in macro- and micronutrients, and livestock wastes. Additionaly, Quaternary horizons, in general, are highly permeable, so they are exposed to the high risk of contamination by filtrates and pollutants' further spread.

3.3. PCA-Weighted Approach in WQI Construction

The WQI was proposed for the first time in 1965 by Horton [54]; nowadays, its variations are widely used in ecological studies. The implementation of weighted factors for quality index construction is currently becoming very popular in environmental science. This procedure was applied earlier in the environmental sustainability index [55] and the Langat River water quality index [56]. Nevertheless, the large diversity of approaches to WQI development shows a list of vulnerabilities of the idea, and many details remain unclear: the high diversity of types of water resources on the global scale that cannot be described by the same measure, the consequent diverse number of parameters used [12], and, finally, the high level of subjectivity. For these reasons, most existing WQIs are not universal and may be used only in case studies [57]. As a suggestion, the WQI should be based on an algorithm, thus excluding subjectivity, including parameters with maximum loads into general variability, thus being adaptive, so policy-makers, water users and managers may be able to implement it according to local object features and purposes.

Application of different multivariate statistical techniques, such as PCA, helps to summarise complex or multi-dimensional issues in view of supporting decision-makers, and make the process less subjective [58]. The goal of the PCA is to reveal how different variables change in relation to each other, or how they are associated. This is achieved by transforming correlated original variables into a new set of uncorrelated variables using the covariance matrix, or its standartized form—the correlation matrix. The new variables are linear combinations of the original ones and are sorted into descending order according to the amount of variance they account for in the original set of variables. The proposed PCA-weighted WQI, which involves the most influential parameters, allows us to model the comprehensive environmental situation in the region. Obviously, this simplification is a logical step toward the description such complicated object as water resources, convenient for use in both scientific and practical applications. In the case of the New Moscow area with 1569 sampling points, our PCA-based approach helps to reveal the 12 crucial parameters of water quality (Cl, pH, Alkalinity, NH_4, PO_4, Cr, Fe, Mn, Cu, SO_4, K, and NO_3) instead of the 25 parameters initially measured. All selected parameters, alone and in mixtures, may negatively affect human health. For example, pH is a crucial water-quality parameter that affects water chemistry, including alkalinity, speciation and solubility. Alkalinity in groundwater exceeding 200 mg/L gives an unpleasant taste and, thus, limits acceptance as potable water. Health concerns regarding sulphates in drinking water have been raised because of the reports that diarrhoea, catharsis, dehydration and gastro-intestinal irritation may be associated with the ingestion of water containing higher levels of sulphate.

A similar PCA-weighted approach has also been proposed by [59] for the water quality index. The authors applied the PCA with Varimax rotation to select the most important features of water quality and reduced the original dataset from 13 parameters to 9. These authors used all important features of water quality; however, unlike our approach, included even co-correlated parameters, which in practice meant overestimation of the final values.

In addition, we can highlight at least one possible disadvantage of the proposed approach, which may be connected with the required data sizes for PCA. For example, ref. [60] recommended that at least 150 cases are needed to obtain satisfactory results by using this method. At the same time, not every study of water quality assessment

includes more than 150 collection points (as examples, [61,62]) due to high installation, operational, and maintenance costs for each sampling representative of the whole water system conditions.

3.4. Automatic Approach to Geospatial Mapping

It is a well-known fact that water legislation worldwide requires adequate and rigorous monitoring on different spatial and temporal scales, including different ecosystem components; usually, these monitoring programs are time-consuming and costly. A process and tool that can be used to perform an accurate and automatic geospatial interpolation of locally collected data are in high demand.

In this research, we propose an advanced method for geospatial modelling based on Gaussian Process Regression and Bayesian Information Criterion. The proposed approach allowed us to detect multiple local foci of environmental contamination, compared with the commonly used ordinary kriging and universal kriging, which were less accurate and effective for solving our particular problem (Figure 6).

Our approach permits determination of the most realistic spatial distribution of the WQI due to the application of an automatically constructed kernel algorithm, which consists of basic non-linear kernels. Actually, when it comes to the end-to-end implementation in operational data processing chains, like geospatial modelling, it is mandatory to invest in models that are both accurate and robust but also require minimal user intervention for fitting parameters. An automatic kernel search helps to solve the problem of manual hyper-parameter and kernel structure selection. According to the applied cross-validation, our model showed lack of over-fitting and provided an accurate prediction on the test dataset according to the used metrics. Recently, this approach of automatic kernel selection was used successfully in several cases, e.g., the estimation of chlorophyll-a concentrations from remote sensing data, [63], delineation referents of city centres from topographic data [64], soft-sensor modelling for algal bloom monitoring [65]. However, to date, it has not been transferred to geological modelling.

4. Conclusions

We developed an end-to-end framework that allowed us to automatically reconstruct the geospatial distribution of WQI with high accuracy. Our approach states the clear methodology from the step of initial data pre-processing and construction of the generalizing Water Quality Index using PCA to the automatic kernel structure search for geospatial mapping. We apply and show the feasibility and robustness of our proposed methodology in the case of WQI estimation in the New Moscow region. The novel approach of an automatic kernel structure search was adapted and applied in this framework, and this approach allowed us to achieve detailed results for geological modelling, compared with ordinary and universal kriging methods. Overall, our proposed methodology opens up wide possibilities for solving similar problems in which it was demonstrated in this paper, in the most accurate and efficient way. Despite being a minor but not the least, important advantage of this work is the relatively large size of the dataset which contains more than 1600 samples (each with 25 measured chemical parameters). We hope that it might be useful for validation and other methodological research in community and share it as open data [38].

Author Contributions: Conceptualization, D.S., A.N., M.P.; methodology, software, validation, D.S., A.N., S.M.; formal analysis, investigation, data curation, D.S., A.N., M.P., P.T.; writing—original draft preparation, D.S., A.N., M.P., P.T., R.J., V.T.; visualization, M.P., D.S., P.T.; supervision, project administration, funding acquisition, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Russian Science Foundation (project No. 20-74-10102).

Data Availability Statement: The data that support the findings of this study are openly available at https://doi.org/10.6084/m9.figshare.10283225.

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

Abbreviations

The following abbreviations are used in this manuscript:

WQI	Water Quality Index
PCA	Principal Component Analysis
GPR	Gaussian Process Regression
BIC	Bayesian Information Criterion
OK	Ordinary Kriging
UK	Universal Kriging
ML	Machine Learning
FA	Factor Analysis
SVM	Support Vector Machine
ANN	Artificial Neural Network
ABC	Approximate Bayesian Computation
MLE	Maximum Likelihood Estimation
RMSE	Root Mean Square Error
DBSCAN	Density-based spatial clustering of applications with noise

References

1. Álvarez, X.; Valero, E.; Santos, R.M.; Varandas, S.; Fernandes, L.S.; Pacheco, F.A.L. Anthropogenic nutrients and eutrophication in multiple land use watersheds: Best management practices and policies for the protection of water resources. *Land Use Policy* **2017**, *69*, 1–11. [CrossRef]
2. Brahney, J.; Mahowald, N.; Ward, D.S.; Ballantyne, A.P.; Neff, J.C. Is atmospheric phosphorus pollution altering global alpine Lake stoichiometry? *Glob. Biogeochem. Cycles* **2015**, *29*, 1369–1383. [CrossRef]
3. Kashulin, N.A.; Dauvalter, V.A.; Denisov, D.B.; Valkova, S.A.; Vandysh, O.I.; Terentjev, P.M.; Kashulin, A.N. Selected aspects of the current state of freshwater resources in the Murmansk region, Russia. *J. Environ. Sci. Health Part A* **2017**, *52*, 921–929. [CrossRef]
4. Dudgeon, D.; Arthington, A.H.; Gessner, M.O.; Kawabata, Z.I.; Knowler, D.J.; Lévêque, C.; Naiman, R.J.; Prieur-Richard, A.H.; Soto, D.; Stiassny, M.L.; et al. Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biol. Rev.* **2006**, *81*, 163–182. [CrossRef]
5. Foley, J.A.; DeFries, R.; Asner, G.P.; Barford, C.; Bonan, G.; Carpenter, S.R.; Chapin, F.S.; Coe, M.T.; Daily, G.C.; Gibbs, H.K.; et al. Global consequences of land use. *Science* **2005**, *309*, 570–574. doi:10.1126/science.1111772. [CrossRef] [PubMed]
6. Tietenberg, T.H.; Lewis, L. *Environmental and Natural Resource Economics*; Routledge: London, UK, 2016.
7. Tscheikner-Gratl, F.; Bellos, V.; Schellart, A.; Moreno-Rodenas, A.; Muthusamy, M.; Langeveld, J.; Clemens, F.; Benedetti, L.; Rico-Ramirez, M.A.; de Carvalho, R.F.; et al. Recent insights on uncertainties present in integrated catchment water quality modelling. *Water Res.* **2019**, *150*, 368–379. [CrossRef] [PubMed]
8. Zwahlen, F. *Vulnerability and Risk Mapping for the Protection of Carbonate (Karst) Aquifers*; Office for Official Publications of the European Communities: Luxembourg, 2003.
9. Hamdan, I.; Margane, A.; Ptak, T.; Wiegand, B.; Sauter, M. Groundwater vulnerability assessment for the karst aquifer of Tanour and Rasoun springs catchment area (NW-Jordan) using COP and EPIK intrinsic methods. *Environ. Earth Sci.* **2016**, *75*, 1474. [CrossRef]
10. Daly, D.; Dassargues, A.; Drew, D.; Dunne, S.; Goldscheider, N.; Neale, S.; Popescu, I.; Zwahlen, F. Main concepts of the "European approach" to karst-groundwater-vulnerability assessment and mapping. *Hydrogeol. J.* **2002**, *10*, 340–345. [CrossRef]
11. Ramakrishnaiah, C.; Sadashivaiah, C.; Ranganna, G. Assessment of water quality index for the groundwater in Tumkur Taluk, Karnataka State, India. *J. Chem.* **2009**, *6*, 523–530. [CrossRef]
12. Sun, W.; Xia, C.; Xu, M.; Guo, J.; Sun, G. Application of modified water quality indices as indicators to assess the spatial and temporal trends of water quality in the Dongjiang River. *Ecol. Indic.* **2016**, *66*, 306–312. [CrossRef]
13. Tripathi, M.; Singal, S.K. Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India. *Ecol. Indic.* **2019**, *96*, 430–436. [CrossRef]
14. Sakizadeh, M.; Mirzaei, R.; Ghorbani, H. Support vector machine and artificial neural network to model soil pollution: A case study in Semnan Province, Iran. *Neural Comput. Appl.* **2017**, *28*, 3229–3238. [CrossRef]
15. Nourani, V.; Alizadeh, F.; Roushangar, K. Evaluation of a two-stage SVM and spatial statistics methods for modeling monthly river suspended sediment load. *Water Resour. Manag.* **2016**, *30*, 393–407. [CrossRef]
16. Yang, K.; Yu, Z.; Luo, Y.; Yang, Y.; Zhao, L.; Zhou, X. Spatial and temporal variations in the relationship between lake water surface temperatures and water quality—A case study of Dianchi Lake. *Sci. Total. Environ.* **2018**, *624*, 859–871. [CrossRef] [PubMed]

17. Dai, F.; Zhou, Q.; Lv, Z.; Wang, X.; Liu, G. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. *Ecol. Indic.* **2014**, *45*, 184–194. [CrossRef]
18. Mitrović, T.; Antanasijević, D.; Lazović, S.; Perić-Grujić, A.; Ristić, M. Virtual water quality monitoring at inactive monitoring sites using Monte Carlo optimized artificial neural networks: A case study of Danube river (Serbia). *Sci. Total. Environ.* **2019**, *654*, 1000–1009. [CrossRef] [PubMed]
19. Ballabio, C.; Lugato, E.; Fernández-Ugalde, O.; Orgiazzi, A.; Jones, A.; Borrelli, P.; Montanarella, L.; Panagos, P. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. *Geoderma* **2019**, *355*, 113912. [CrossRef]
20. Keskin, H.; Grunwald, S. Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma* **2018**, *326*, 22–41. [CrossRef]
21. McLeod, L.; Bharadwaj, L.; Epp, T.; Waldner, C.L. Use of principal components analysis and kriging to predict groundwater-sourced rural drinking water quality in Saskatchewan. *Int. J. Environ. Res. Public Health* **2017**, *14*, 1065. [CrossRef]
22. Keshtegar, B.; Mert, C.; Kisi, O. Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs. RSM, MARS and M5 model tree. *Renew. Sustain. Energy Rev.* **2018**, *81*, 330–341. [CrossRef]
23. Liu, H.; Yang, C.; Huang, M.; Wang, D.; Yoo, C. Modeling of subway indoor air quality using Gaussian process regression. *J. Hazard. Mater.* **2018**, *359*, 266–273. [CrossRef]
24. Cressie, N. The origins of kriging. *Math. Geol.* **1990**, *22*, 239–252. [CrossRef]
25. Ebden, M. Gaussian processes: A quick introduction. *arXiv* **2015**, arXiv:1505.02965.
26. Van Stein, B.; Wang, H.; Kowalczyk, W.; Emmerich, M.; Bäck, T. Cluster-based Kriging approximation algorithms for complexity reduction. *Appl. Intell.* **2020**, *50*, 778–791. [CrossRef]
27. Chiles, J.P.; Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 731.
28. Oliver, M.; Webster, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* **2014**, *113*, 56–69. [CrossRef]
29. Aalto, J.; Pirinen, P.; Heikkinen, J.; Venäläinen, A. Spatial interpolation of monthly climate data for Finland: Comparing the performance of kriging and generalized additive models. *Theor. Appl. Climatol.* **2013**, *112*, 99–111. [CrossRef]
30. Chica-Olmo, M.; Luque-Espinar, J.A.; Rodriguez-Galiano, V.; Pardo-Igúzquiza, E.; Chica-Rivas, L. Categorical Indicator Kriging for assessing the risk of groundwater nitrate pollution: The case of Vega de Granada aquifer (SE Spain). *Sci. Total. Environ.* **2014**, *470*, 229–239. [CrossRef]
31. Al-Mudhafar, W.J. Bayesian kriging for reproducing reservoir heterogeneity in a tidal depositional environment of a sandstone formation. *J. Appl. Geophys.* **2019**, *160*, 84–102. [CrossRef]
32. Pebesma, E.; Cornford, D.; Dubois, G.; Heuvelink, G.B.; Hristopulos, D.; Pilz, J.; Stöhlker, U.; Morin, G.; Skøien, J.O. INTAMAP: the design and implementation of an interoperable automated interpolation web service. *Comput. Geosci.* **2011**, *37*, 343–352. [CrossRef]
33. Abdessalem, A.B.; Dervilis, N.; Wagg, D.J.; Worden, K. Automatic kernel selection for gaussian processes regression with approximate bayesian computation and sequential monte carlo. *Front. Built Environ.* **2017**, *3*, 52. [CrossRef]
34. Megdal, S.B. Invisible water: The importance of good groundwater governance and management. *npj Clean Water* **2018**, *1*, 1–5. [CrossRef]
35. Shishov, L.; Voinovich, N. *Soils of Moscow Region and Their Use*; Dokuchaev Soil Science Institute: Moscow, Russia, 2002.
36. Dzhamalov, R.; Medovar, Y.A.; Yushmanov, I. Principles of MSW Landfill Sites' Placement Depending on Geological and Hydrogeological Conditions of Territories (Based on Moscow Region). *Water Resour.* **2019**, *46*, S51–S58. [CrossRef]
37. Klimanova, O.; Kolbowsky, E.; Illarionova, O. Impacts of urbanization on green infrastructure ecosystem services: The case study of post-soviet Moscow. *Belg. Rev. Belg. de Géogr.* **2018**. [CrossRef]
38. Pukalchik, M.; Shadrin, D.; Nikitin, A.; Jana, R.; Tregubova, P.; Matveev, S. Freshwater chemical properties for New Moscow region. 2020. Available online: https://figshare.com/articles/dataset/freshwater_chemical_properties_for_New_Moscow_region/10283225 (accessed on 2 February 2021).
39. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [CrossRef] [PubMed]
40. Richardson, M. Principal Component Analysis. 2009. Available online: http://www.dsc.ufcg.edu.br/~hmg/disciplinas/posgraduacao/rn-copin-2014.3/material/SignalProcPCA.pdf (accessed on 2 February 2021).
41. Wall, M.E.; Rechtsteiner, A.; Rocha, L.M. Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 91–109.
42. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417. [CrossRef]
43. Cattell, R. *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1978.
44. Kaiser, H.F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200. [CrossRef]
45. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2.
46. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.
47. Duvenaud, D.; Lloyd, J.R.; Grosse, R.; Tenenbaum, J.B.; Ghahramani, Z. Structure discovery in nonparametric regression through compositional kernel search. *arXiv* **2013**, arXiv:1302.4922.

48. 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; Volume 96, pp. 226–231.
49. MacCormack, K.E.; Brodeur, J.J.; Eyles, C.H. Evaluating the impact of data quantity, distribution and algorithm selection on the accuracy of 3D subsurface models using synthetic grid models of varying complexity. *J. Geogr. Syst.* **2013**, *15*, 71–88. [CrossRef]
50. MacCormack, K.; Arnaud, E.; Parker, B.L. Using a multiple variogram approach to improve the accuracy of subsurface geological models. *Can. J. Earth Sci.* **2018**, *55*, 786–801. [CrossRef]
51. Mueller, T.; Pusuluri, N.; Mathias, K.; Cornelius, P.; Barnhisel, R.; Shearer, S. Map quality for ordinary kriging and inverse distance weighted interpolation. *Soil Sci. Soc. Am. J.* **2004**, *68*, 2042–2047. [CrossRef]
52. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
53. GPy. GPy: A Gaussian Process Framework in Python. Since 2012. Available online: http://github.com/SheffieldML/GPy (accessed on 2 February 2021).
54. Horton, R.K. An index number system for rating water quality. *J. Water Pollut. Control. Fed.* **1965**, *37*, 300–306.
55. Esty, D.C.; Levy, M.; Srebotnjak, T.; De Sherbinin, A. *Environmental Sustainability Index: Benchmarking National Environmental Stewardship*; Yale Center for Environmental Law & Policy: New Haven, CT, USA, 2005; pp. 47–60.
56. Mohd Ali, Z.; Ibrahim, N.A.; Mengersen, K.; Shitan, M.; Juahir, H. The Langat River water quality index based on principal component analysis. *AIP Conf. Proc.* **2013**, *1522*, 1322–1336. [CrossRef]
57. Tyagi, S.; Sharma, B.; Singh, P.; Dobhal, R. Water quality assessment in terms of water quality index. *Am. J. Water Resour.* **2013**, *1*, 34–38. [CrossRef]
58. Nardo, M.; Saisana, M.; Saltelli, A.; Tarantola, S. Tools for composite indicators building. *Eur. Com. Ispra* **2005**, *15*, 19–20.
59. Tripathi, M.; Singal, S.K. Allocation of weights using factor analysis for development of a novel water quality index. *Ecotoxicol. Environ. Saf.* **2019**, *183*, 109510. [CrossRef]
60. Hutcheson, G.D.; Sofroniou, N. *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*; Sage: Thousand Oaks, CA, USA, 1999; [CrossRef]
61. Ouyang, Y. Evaluation of river water quality monitoring stations by principal component analysis. *Water Res.* **2005**, *39*, 2621–2635. [CrossRef]
62. Chen, Y.; Han, D. Water quality monitoring in smart city: A pilot project. *Autom. Constr.* **2018**, *89*, 307–316. [CrossRef]
63. Gómez-Chova, L.; Muñoz-Marí, J.; Laparra, V.; Malo-López, J.; Camps-Valls, G. A review of kernel methods in remote sensing data analysis. In *Optical Remote Sensing*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 171–206. [CrossRef]
64. Lüscher, P.; Weibel, R. Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Comput. Environ. Urban Syst.* **2013**, *37*, 18–34. [CrossRef]
65. Wang, Z.; Zhao, Z.; Li, D.; Cui, L. Data-driven soft sensor modeling for algal blooms monitoring. *IEEE Sens. J.* **2014**, *15*, 579–590. [CrossRef]