sustainability

MDPI

Article

Digital Visualization of Environmental Risk Indicators in the Territory of the Urban Industrial Zone

Ruslan Safarov [1], Zhanat Shomanova [2,*], Yuriy Nossenko [2], Zhandos Mussayev [2] and Ayana Shomanova [2,*]

[1] Department of Chemistry, Faculty of Natural Sciences, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan; safarov_rz@enu.kz
[2] Higher School of Natural Science, Margulan University, Pavlodar 140002, Kazakhstan; nosenko1980@yandex.ru (Y.N.); musajandos@gmail.com (Z.M.)
* Correspondence: zshoman@yandex.ru (Z.S.); ayanashomanova@mail.ru (A.S.)

Abstract: This study focused on predicting the spatial distribution of environmental risk indicators using mathematical modeling methods including machine learning. The northern industrial zone of Pavlodar City in Kazakhstan was used as a model territory for the case. Nine models based on the methods kNN, gradient boosting, artificial neural networks, Kriging, and multilevel b-spline interpolation were employed to analyze pollution data and assess their effectiveness in predicting pollution levels. Each model tackled the problem as a regression task, aiming to estimate the pollution load index (PLI) values for specific locations. It was revealed that the maximum PLI values were mainly located to the southwest of the TPPs over some distance from their territories according to the average wind rose for Pavlodar City. Another area of high PLI was located in the northern part of the studied region, near the Hg-accumulating ponds. The high PLI level is generally attributed to the high concentration of Hg. Each studied method of interpolation can be used for spatial distribution analysis; however, a comparison with the scientific literature revealed that Kriging and MLBS interpolation can be used without extra calculations to produce non-linear, empirically consistent, and smooth maps.

Keywords: urban industrial zone; sustainable city; pollution load index (PLI); machine learning methods; soil contamination

Citation: Safarov, R.; Shomanova, Z.; Nossenko, Y.; Mussayev, Z.; Shomanova, A. Digital Visualization of Environmental Risk Indicators in the Territory of the Urban Industrial Zone. *Sustainability* **2024**, *16*, 5190. https://doi.org/10.3390/su16125190

Academic Editors: Dhiya Al-Jumeily OBE, Jamila Mustafina and Manoj Jayabalan

Received: 10 May 2024
Revised: 25 May 2024
Accepted: 14 June 2024
Published: 18 June 2024

1. Introduction

In today's rapidly evolving world, digitalization plays a crucial role in shaping various aspects of society, including environmental sustainability. As the need to preserve our planet becomes increasingly urgent, leveraging digital technologies has become essential in addressing environmental challenges and promoting sustainable practices. Digitalization refers to the integration of digital technologies into everyday life, transforming how individuals and organizations operate. From automation to data analytics, digitalization has revolutionized industries and sectors, offering new ways to address complex issues such as environmental conservation.

Environmental sustainability is the practice of preserving natural resources and ecosystems for future generations. With climate change, pollution, and deforestation posing significant threats to the planet, achieving environmental sustainability has become a global priority. Digitalization offers innovative solutions to environmental challenges by providing tools and technologies that enable more efficient resource management, data-driven decision-making, and enhanced communication among stakeholders. By harnessing the power of digital tools, organizations and individuals can make informed choices that benefit both the environment and society.

One of the key ways digitalization contributes to environmental sustainability is through the development of tools for monitoring and managing environmental impact. These tools allow organizations to track their resource consumption, waste generation,

and carbon emissions in real time, enabling them to identify areas for improvement and implement targeted solutions. Data analytics plays a crucial role in driving sustainable practices by providing insights into environmental performance and trends. By analyzing data on energy usage, water consumption, and waste generation, organizations can identify patterns, set targets for improvement, and measure the impact of their sustainability initiatives.

Environmental risk assessment is critically important as it allows us to understand and evaluate potential adverse effects caused by chemical substances, industrial activities, or development projects on the natural environment and human health. This systematic process helps to identify and measure risks associated with environmental hazards, leading to informed decision-making and the development of mitigation strategies to prevent or minimize harm [1].

At present, when developing the concept of adopting the Internet of Things, systematic monitoring can be carried out using appropriate environmental sensors. In this context, the search for and development of effective methods for predicting the spatial distribution of pollutants is an extremely urgent task, since only intelligent modeling will make it possible to create chains of balanced decisions for sustainable human life [2]. For instance, the integration of Internet of Things (IoT) technology in agriculture has the potential to revolutionize farming by promoting sustainability, diversification, and high yields while minimizing environmental impact. IoT sensors enable the real-time monitoring of crop health and conditions, optimizing resource use and reducing costs. Additionally, IoT-enabled supply chains enhance efficiency, align crop distribution with consumer demand, and reduce food waste, ensuring high yields and long-term sustainability [3].

Environmental risk indicators are quantitative measures that provide insight into the level of impact that human activities may have on the environment. These indicators are used to assess the potential for damage and to guide policy and management decisions [4]. Some commonly used environmental risk indicators include biological oxygen demand [5], chemical oxygen demand [6], ecological footprint [7], water quality indices [8], air quality index [9,10], and pollution load index [11,12].

The PLI provides a cumulative indication of the overall levels of heavy metal pollution present in a particular area. It is calculated by taking the nth root of the product of the concentration factors of the metals measured, where n is the number of metals.

A PLI value of more than one suggests pollution, while a value less than one indicates no pollution. Rabee [13] used the PLI to evaluate sediment pollution in the Tigris River, finding low PLI values that indicated no pollution. Similarly, Angulo [14] applied the PLI to mussel data, identifying highly contaminated sites with PLI values over 100. Chen [15] used pollution indices (PIs) to assess heavy metal pollution in Beijing's urban parks, with the highest integrated pollution index (IPI) in the historic center district. Jorfi et al. [16] studied soil pollution by heavy metals in Mian-Ab plain as a major agricultural site of Khuzestan province (Iran). The contamination level of selected heavy metals was assessed by single-factor contaminant index (PI) and pollution load index (PLI). Based on the PLI results, 24% of the studied area was moderately polluted and 76% of the area was unpolluted. These studies collectively demonstrate the utility of the PLI in assessing heavy metal pollution in various environments.

Accurately interpolating the spatial distribution of environmental risk indicators presents several challenges. Environmental data are often scarce, especially in remote or less studied regions, making it difficult to interpolate risk across vast areas with confidence [17]. Environmental risk indicators often exhibit complex spatial patterns with local variations and non-linear relationships that might not be effectively captured [18,19]. The interpolation process itself introduces uncertainties due to estimation errors and assumptions made about the underlying spatial processes, leading to potential inaccuracies in the final risk maps [20]. Selecting the most suitable interpolation method depends on various factors like data characteristics, the scale of analysis, and the specific risk indicator under study, making it a crucial but challenging step [21]. These challenges necessitate the careful

consideration of data availability, selection of appropriate methods, and integration of uncertainty considerations to ensure reliable and informative spatial risk maps.

Traditional deterministic approaches used for spatial interpolation include:

- Inverse Distance Weighting (IDW), where weights are assigned to nearby data points based on their distance to the target location, with closer points having greater influence on the interpolated value [22];
- Kriging, a geostatistical method that incorporates spatial autocorrelation and variogram analysis to estimate risk at unsampled locations, considering spatial dependence and uncertainty [23];
- spline interpolation, where splines are a type of piecewise polynomial function used for interpolation. They connect multiple polynomial segments, called splines, to create a smooth curve that passes through or near a set of data points [24].

The Kriging method has been used in many scientific studies for spatial distribution analysis, including estimating CO_2 emissions in Spain [25], assessing contaminant patterns [26], and evaluating soil erosion risk [27]. In one study [28], ordinary Kriging was used with a stable semi-variogram to prepare spatial distribution maps of soil parameters. Another study [29] discusses Kriging models for calculating distances between locations on Earth's surface as part of the interpolation process. Faisal and Jaelani found that Kriging interpolation is very suitable for irregular data, such as the observation of NO_2 [30].

The multiple uses of this method in different research indicate the high appropriateness of the Kriging method for spatial distribution analysis. So, Kriging is known for its precision and is often considered the best linear unbiased estimator for a wide range of spatial interpolation problems. It provides estimates of the uncertainty of predictions, which can be very valuable for decision-making. Kriging allows for the incorporation of both the mean and the spatial covariance structure of the data. Also, the limitations of the method were reflected in scientific works. So, Kriging can be computationally intensive, especially for large datasets, which can limit its practicality in some situations [31]. It assumes the stationarity of the underlying spatial process, which might not be true for all datasets [31]. Kriging requires a well-specified semi-variogram model, which can be difficult to determine and sensitive to outliers [32].

One interesting advanced method of spline interpolation is multilevel b-spline interpolation, which was first described by S. Lee, G. Wolberg, and S. Y. Shin [33]. The algorithm introduced in that paper makes use of a coarse-to-fine hierarchy of control lattices to generate a sequence of bicubic b-spline functions, whose sum approaches the desired interpolation function [34]. A detailed description of the mathematical content of the algorithms with numerical examples can be found in the report "Approximation of Scattered Data with Multilevel b-splines" by Ø. Hjelle, published in 2001 [35]. This report provides a comprehensive overview of the method and its applications. So, the concept of multilevel b-spline interpolation has been around for a few decades and has been refined and expanded upon by various researchers since its inception. It continues to be a valuable tool in fields requiring high-precision interpolation. In the field of scattered data interpolation and approximation, multilevel b-splines are used to compute a continuous surface through a set of irregularly spaced points. Overall, multilevel b-spline interpolation offers substantial improvements in accuracy and its performance is competitive with other methods. It is a serious contender for calculating pairwise interactions in molecular dynamics simulations [36] and provides high-fidelity reconstruction from a selected set of sparse and irregular samples [33].

Despite its versatility in scientific applications, multilevel b-spline interpolation presents both advantages and drawbacks [37]. One study [38] proposed a b-spline-based localization algorithm to extract multidimensional information from single-molecule data. b-Splines allow for the efficient modeling of higher-dimensional point spread functions using less memory than conventional splines. The versatile b-spline approach makes complex, high-dimensional point spread function models more accessible without specialized software. The b-spline PSF modeling method achieved high accuracy in extracting abundant infor-

mation from single molecules, including 3D position, wavelength, and dipole orientation. b-Splines were used in one study [36] to increase the accuracy of the multilevel summation method for non-periodic boundaries without incurring additional computation. b-Spline interpolation plays a crucial role in enhancing the accuracy and performance of the multilevel summation method for calculating pairwise interactions in molecular dynamics simulations. Thus, although there are few works where MLBS interpolation is used for geospatial analysis, there are numerous studies were MLBS interpolation has shown its high accuracy in extracting information from single molecules and improving efficiency in molecular simulations. However, its complexity can pose challenges, requiring specialized knowledge for implementation and careful parameter tuning to avoid overfitting.

Along with traditional deterministic methods, machine learning (ML) methods are also widely used to interpolate spatial distributions. These methods allow for the estimation of values at unobserved locations based on available data points. The most widely used methods are k-nearest neighbors, weighted k-nearest neighbors, gradient boosting, and artificial neural networks.

The k-nearest neighbors (kNN) technique is a powerful non-parametric classifier widely used for spatially contiguous predictions. The kNN method, a simple yet powerful machine learning algorithm, has been widely used and studied across various fields of research. Its simplicity lies in its basic premise—classifying new cases based on the majority vote of its k-nearest neighbors. This non-parametric approach has been favored for its ease of implementation and intuitive understanding. Researchers have explored various aspects of kNN, and here are some key findings. kNN is robust for noisy data and easy to implement. Its simplicity makes it accessible for practitioners. kNN considers spatial proximity, making it suitable for spatial interpolation tasks. It leverages the similarity between neighboring points to predict values at unobserved locations [39]. Beyond classification, kNN has applications in anomaly detection, dimensionality reduction, and missing value imputation [40].

One novel approach combines the local linear model tree (LOLIMOT) network with a k-nearest neighborhood (kNN) search. This method selects data pairs through decile analysis based on distances calculated during kNN data grouping [41]. The positive aspect of this approach lies in its robust performance, demonstrated in both synthetic 3D datasets and real-world micro-gravimetric data and earthquake catalogues. However, it is not without challenges; the integration of LOLIMOT and kNN introduces complexity. Another method, spatial kNN, is used for the non-parametric prediction of real spatial data and the supervised classification of categorical spatial data. It employs a double nearest neighbor rule with random bandwidth to control distances between observations and locations [42]. The positive aspect here is the almost complete convergence rate for prediction and the almost certain convergence for supervised classification. However, it exhibits potential sensitivity to the choice of bandwidth parameter. Lastly, one study compared the spatial linear model (SLM) and kNN approaches for mapping and estimating totals. While both methods possess desirable properties, the SLM stands out for its prediction optimality and robustness, whereas kNN lacks the same level of prediction optimality compared to the SLM [43].

The weighted k-nearest neighbors (WkNN) algorithm is a fundamental non-parametric method in pattern recognition and machine learning. It has garnered significant attention over the years, yet certain aspects remain unsettled. The problem involves determining the optimal number of neighbors (k) and their corresponding weights. Despite extensive study since the 1950s, the practical regime (where the sample size is finite) has often been overlooked. Most theoretical work focuses on asymptotic scenarios with infinite samples, neglecting the specific dataset structure and the properties of the data points for which we seek label estimation [44]. The WkNN algorithm's success hinges on three key choices: number of neighbors (k), weight vector, and distance metric. The empirical results demonstrate superior performance over standard locally weighted methods. In summary,

the WkNN method continues to intrigue researchers, and its practical applicability remains an active area of investigation [45].

Gradient boosting is a powerful ensemble learning technique that constructs an additive approximation of a target function by combining multiple weak learners. It has gained prominence due to its effectiveness in handling complex data and large datasets. Based on the scientific literature, there are different gradient boosting methods. One scalable ensemble method, XGBoost, has demonstrated reliability and efficiency in solving machine learning challenges. It balances speed and accuracy, making it a popular choice [46]. Focused on fast training performance, LightGBM selectively samples high-gradient instances. While it excels in speed, it may not be the most accurate [46,47]. The CatBoost algorithm modifies gradient computation to enhance model accuracy by avoiding prediction shifts. It performs well in terms of generalization accuracy and AUC [46,48]. A comprehensive comparison between XGBoost, LightGBM, CatBoost, random forests, and traditional gradient boosting reveals that CatBoost achieves the best generalization accuracy, while LightGBM is the fastest. XGBoost ranks second in both accuracy and training speed [46]. Thus, gradient boosting methods continue to evolve, with novel variants addressing speed and accuracy trade-offs. These algorithms play a crucial role in modern machine learning applications.

Gradient boosting has been widely used in spatial distribution analysis, with studies demonstrating its effectiveness in various applications. Thus, gradient boosting has been successfully applied in spatial distribution analysis, with Thomas [49] using it to estimate the abundance of common eider in Massachusetts, USA, and Biau [50] introducing an accelerated version of the method. Bailly [51] found that boosting outperformed maximum likelihood in cases of overlapping radiometric variable classes, while Şahin [46] reported that CatBoost, XGBoost, and LightGBM were all effective in landslide susceptibility mapping, with CatBoost showing the highest prediction capability. However, the approach is not without its limitations. Gradient boosting models can be prone to overfitting if not tuned properly (number of trees, depth of trees, and learning rate are all critical parameters). They can be computationally intensive, especially with large datasets or complex spatial relationships. Also, they require careful cross-validation to ensure that they generalize well to new data. Furthermore, the method's performance can be affected by the choice of parameters, as highlighted by Taşpınar [52] in the context of spatial dynamic panel data models.

Artificial neural networks (ANNs), a subset of machine learning techniques, have gained significant prominence due to their ability to model complex relationships and learn from data. These networks are inspired by the human brain's neural architecture, which consists of interconnected neurons. ANNs have been widely applied in various domains, including image recognition, natural language processing, and bioinformatics [53].

One of the fundamental types of ANNs is the multi-layer perceptron (MLP). The MLP is a feed-forward neural network with multiple layers, including an input layer, one or more hidden layers, and an output layer. Each layer contains interconnected neurons (also known as nodes), and information flows from the input layer through the hidden layers to the output layer. The MLP's architecture allows it to capture intricate patterns and non-linear relationships in data [54].

With a larger dataset, ANNs have the potential to learn more complex patterns and relationships within the data. This can lead to better generalization, meaning the ability to perform well on unseen data not encountered during training. ANNs are well-suited for handling high-dimensional data (data with many features). However, they typically require a larger amount of data compared to simpler models to learn the complex relationships effectively in such high-dimensional spaces. In situations where data are scarce, the sensitivity of ANNs to dataset size can be a major drawback. They might not perform well or learn effectively with limited data, hindering their applicability in such scenarios. If the training data are biased, then the ANN might inherit and amplify those biases, leading to unfair or inaccurate predictions. Careful data selection and cleansing are crucial to mitigate this risk.

The northern industrial zone of Pavlodar City is a pronounced example of an urbanized and industrial-developed territory [55,56]. It is a multisectoral complex devoted to the production of electricity and fuel-based energy, oil refining, the production of chemicals and construction materials, machine building, and light and food industries [57,58]. It also contains mercury-storing ponds, used earlier for the electrochemical production of Kaustik soda with a liquid mercury cathode. There is information about mercury contamination in that territory [59–61]. Along with the positive economic effects, this territory affects the residential area of the city of Pavlodar and should be regularly monitored and studied in order to assess environmental risks and plan measures for the remediation and reduction of contamination in order to preserve the well-being of residents. Recent statistics show a significant increase in respiratory diseases and cancer cases in the Pavlodar region [62–64].

A small number of studies have examined the distribution of heavy metals in the northern industrial zone of Pavlodar City [65–67]. In addition, a comprehensive spatial analysis of a wide range of heavy metals and environmental risk indicators such as the PLI was not conducted using mathematical modeling. Furthermore, no comparison was made between different mathematical models in the past. Thus, the objective of this study is to find the most appropriate mathematical models for revealing the spatial distribution of the environmental risk indicator PLI.

To achieve this goal, the following research tasks were defined: (1) conduct soil sampling in the territory of the northern industrial zone of Pavlodar, Kazakhstan; (2) perform elemental analysis of soil samples for heavy metal content and calculate PLI values for each point; (3) select a number of mathematical models for the prediction and visualization of the spatial distribution of the pollution index in the territory of the industrial zone; (4) conduct a comparative analysis of the effectiveness of the selected models for the visualization of the spatial distribution of PLI in the studied territory using cross-validation and self-assessment methods. Additionally, the most optimal methods in terms of the lowest error, the highest empirical consistency, and visualization quality were identified.

2. Materials and Methods

2.1. Research Area and Sample Analysis

This research focuses on the northern industrial zone (NIZ) in Pavlodar City, the capital of Pavlodar District in northeast Kazakhstan. Pavlodar is one of the oldest and most scenic cities in the country, situated on the shore of Irtysh, Kazakhstan's largest river. The metallurgical industry in Pavlodar District is dominated by the Pavlodar Aluminum Plant, the Kazakhstan Electrolysis Plant, and the Aksu Ferroalloy Plant. The mining industry includes major companies such as coal miners 'Bogatyr Akses Komir' and 'Maicuben West'. In Pavlodar City, the NIZ hosts the Pavlodar petrochemical factory, one of the top petroleum producers in Kazakhstan, as well as combined heat and power plants TPP-2 and TPP-3, a cardboard and roofing rubber plant, and the caustic soda plant 'Kaustik'.

The region experiences a continental climate due to air currents from the Arctic, temperate, and southern zones. The spring–summer season is dry, while the winter is long and chilly (lasting 5–5.5 months). The summer is brief and warm (lasting 3 months). The annual rainfall is low and erratic, with a peak in summer. Strong winds occur throughout the year. Spring is the windiest season, with the strongest winds occurring in March, April, May, and occasionally June. On average, there are 35 days per year with winds over 15 m/s, and only 5–6% of days are calm. The average wind speed for the year is 4–5 m/s [68]. The soil cover varies greatly because of the different factors that influence soil formation. The rocks that form the soil are mostly sands and sandy loam, and sometimes loam, which makes the soil texture light.

The study area was sampled at 39 locations between 20 September and 5 October 2022. The humus layer was removed and 5 soil samples of about 200 g each were taken from 0–20 cm depth with a hand shovel at every location. The samples from each location were combined to make 39 total samples of around 1 kg each. The samples were kept in

plastic containers to avoid sunlight exposure. They were air dried, passed through a 2 mm sieve, homogenized, and stored in plastic containers at room temperature until lab analysis. Figure 1 shows the sampling map.

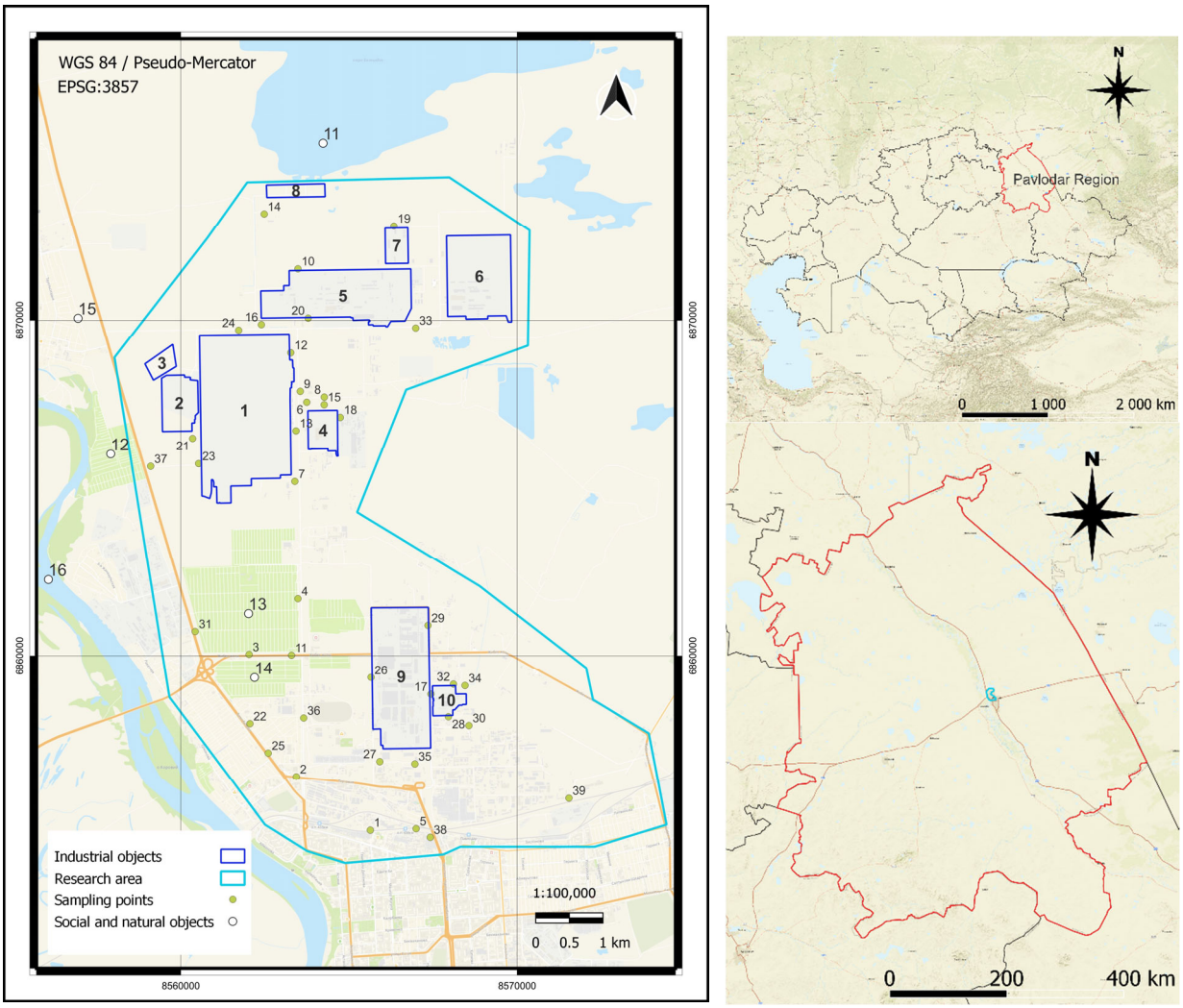


Figure 1. Pavlodar northern industrial zone research area (202.29 km^2) with sampling locations. Industrial, social, and natural objects: 1—Pavlodar Petrochemical Plant, 2—Neftechim LTD, 3—Solid Waste Accumulator, 4—TPP-3, 5—Former Pavlodar Chemical Plant (industrial site 1), 6—Former Pavlodar Chemical Plant (industrial site 2), 7—Calcined Petroleum Coke Production Plant, 8—Ponds for Mercury Waste, 9—KSP-Steel (former Pavlodar Tractor Plant), 10—TPP-2, 11—Balkyldak Lake, 12, 13, 14—Residents' summer houses, 15—Pavlodarskoye Village, and 16—Irtysh River. 1–39 (green circle mark)—sampling points locations.

Each sample was ground with a grinding mortar. Elemental analysis was performed with the XRF method on an X-Supreme 8000 (Hitachi High-Tech Analytical Science, Shanghai, China) in accordance with ST RK 3616-2020 X-ray fluorescence analysis of wastes of mineral origin.

2.2. Environmental Risk Indicator Calculation

This study measured the soil contamination levels and the potential environmental risk using a standard indicator, the pollution load index (PLI). This indicator is widely used to assess the extent of trace element pollution in soils and to contrast the findings with other regions that might be affected by pollution.

The pollution load index is a tool used to assess the overall level of pollution in a particular area, typically in the context of soil or sediment quality. It provides a cumulative indicator of the presence and intensity of different pollutants. The PLI is calculated by taking the n-root of the multiplication of the concentration factors (CFs) of given heavy metals or pollutants, where 'n' represents the number of pollutants assessed (1) [69].

$$PLI = (CF_i \times CF_j \times \cdots \times CF_n)^{1/n}, \quad (1)$$

For example, if you are assessing the levels of four heavy metals, then you would calculate the concentration factor for each metal and then multiply these concentration factors together. The fourth root (since there are four metals) of this product would give you the PLI.

A PLI greater than one indicates pollution, with a value of one suggesting that the concentrations of pollutants are at baseline levels, and a value less than one indicating no pollution. The PLI is frequently utilized by environmental agencies and academic researchers as a means of summarizing the overall quality of the environment with respect to pollutant load.

The concentration factor or contamination factor (CF) was measured by dividing the HM concentrations (mg kg^{-1}) in each soil sample by the background values from a nearby area in Pavlodar City that was not impacted by environmental factors according to the following formula [70]:

$$CF = C_{Soil}/C_{Background}, \quad (2)$$

We calculated the concentration factors for the following HMs: Cr, Mn, Fe, Zn, Sr, Cu, Pb, and Ni. Some samples had levels of V, Co, and Hg below the analytical detection values. In samples where one or more heavy metals were below the analytical detection limit, the 'n' value was determined based on the number of heavy metals with detectable concentrations. Additionally, if the contamination factor (CFi) was less than or equal to 0.7, then the corresponding heavy metal was excluded from the analysis. The contamination levels were classified by their intensities on a scale from 1 to 6 (Table 1).

Table 1. Contamination levels classification for HM [71].

Class Number	CF	PLI	Classification Description
1	$CF \leq 0.7$	$PLI \leq 0.7$	No contamination
2	$0.7 < CF \leq 1$	$0.7 < PLI \leq 1$	Low contamination
3	$1 < CF \leq 3$	$1 < PLI \leq 3$	Moderate contamination
4	$3 < CF \leq 6$	$3 < PLI \leq 6$	Considerable contamination
5	$6 < CF$	$6 < PLI$	Very high contamination

2.3. Methods for Spatial Interpolation and Evaluation Measures

The problem of finding the PLI spatial distribution was solved through regression using well-known and often used computational methods including machine learning, k-nearest neighbors, weighted k-nearest neighbors, gradient boosting (CatBoost), artificial neural networks (multi-layered perceptron), Kriging, and multilevel b-spline interpolation. The review of the used methods is represented in Table 2.

Table 2. Review of used interpolation methods.

Name	Description	Instruments	References
k-Nearest Neighbors (kNN)	Simple, intuitive, and widely used non-parametric algorithm for both classification and regression in machine learning. In regression, it assigns the property value determined by the average of the values of its k-nearest neighbors. k-Nearest neighbors is used in various applications like recommendation systems, pattern recognition, and anomaly detection due to its simplicity, effectiveness, and ease of interpretation.	KNeighborsRegressor() function of Scikit-learn package in Python	[72,73]
Weighted k-Nearest Neighbors (WkNN)	A variant of the basic kNN algorithm where different weights are assigned to the contributions of the neighbors, so the nearest neighbors contribute more to the average than the more distant ones. This approach can be particularly useful when dealing with heterogeneous datasets, where certain data points are more closely clustered together, and outliers could disproportionately affect the result if an unweighted approach were used. Weighted kNN can help to mitigate the influence of outliers and provide a more nuanced classification or regression outcome.	KNeighborsRegressor() function of Scikit-learn package in Python	[74,75]
Gradient Boosting. CatBoost (CB)	CatBoost is an open-source gradient boosting algorithm developed by Yandex, which is designed to work with categorical features without the need for the extensive data preprocessing that is typically required by other machine learning algorithms. CatBoost is particularly powerful for datasets with lots of categorical features and has been successfully used in various applications, including ranking tasks, regression problems, and classification problems. It is known for its robustness, handling of large datasets, and high-quality predictions, along with its ease of use.	CatBoostRegressor() function of Catboost package in Python was utilized with RMSE as loss function	[48,76]
Artificial Neural Networks (ANN)	Computational models consisting of interconnected groups of artificial neurons or nodes that process information using a connectionist approach to computation. They can learn to approximate non-linear functions, which makes them suitable for a wide range of problems including but not limited to classification, regression, and time series prediction. Learning occurs in the network through a process of adjusting the synaptic weights of the connections between neurons, usually executed by a learning algorithm like backpropagation.	MLPRegressor() function from the Scikit-learn package in Python	[77–82]
Kriging	A geostatistical interpolation technique that is widely used for spatial analysis and modeling. Kriging assumes that the spatial variation in the data can be modeled as a stochastic process with a structured dependence captured by the variogram or covariance function. To generate a prediction, ordinary Kriging computes a set of weights for the known data points, ensuring that the sum of these weights equals one to maintain the estimator's unbiasedness. These weights are determined by solving a system of linear equations that arises from the variogram model, incorporating a constraint that enforces the unbiasedness of the predictions. This method is especially favored in fields such as geostatistics and environmental modeling, where understanding and accounting for spatial variability is critical	rk.Krige() function from the PyKrige package in Python. The best Kriging model for the given dataset was selected using the GridSearchCV function from the Scikit-learn package with cross validation.	[83–86]

Table 2. *Cont.*

Name	Description	Instruments	References
Multilevel b-spline interpolation (MLBS)	A sophisticated mathematical method used for smooth curve fitting and surface approximation, which is particularly useful when working with complex data in multiple dimensions. b-Splines, short for basis splines, are a series of piecewise polynomials that are defined by a set of control points that determine the shape of the curve or surface. The benefits of this method include its inherent smoothness, the ability to handle large and potentially irregularly spaced data sets, and the control it provides over the smoothness of the interpolation. Moreover, it is quite robust, reducing the impact of noise in the data, and it is capable of capturing the underlying trend without overfitting to the precise data points. This method serves as a powerful tool in applications that require a blend of accuracy and visual or analytical smoothness. Because MLBS interpolation strictly uses initial points for interpolation, conducting a self-assessment is not useful. This is because the error will be driven to zero, which does not reflect the true accuracy of the prediction.	The function was used with SAGA ver. 9.1.0 GIS software's command line API through PySAGA_cmd package in Python.	[33–36]

2.4. Data Analysis and Visualization

2.4.1. Cross Validation

The small number of obtained data (39 samples) is the main limitation for the unbiased creation and assessment of the models for spatial interpolation. When working with limited available data, and to minimize the risk of overfitting, it is particularly important to employ cross-validation. Cross-validation should be used for assessing the accuracy of a prediction model when you want to evaluate its ability to generalize an independent dataset.

By partitioning the dataset into multiple subsets, cross-validation allows each subset to act as a training set and a validation set. This helps in ensuring that the model's performance is consistent across different samples from the dataset and not just tailored to a specific set of data. During our research, k-fold cross-validation was used, which is a common method where the dataset is split into 'k' number of folds, and the model is trained on 'k−1' folds and validated on the remaining fold. This process is repeated 'k' times with each fold serving as the validation set exactly once [87].

Cross-validation is also advisable when tuning hyperparameters, as it helps in selecting the model parameters that lead to the best generalization performance [88]. It allows for a more objective and comprehensive evaluation of the model performance compared to methods that rely on a single training–validation split.

To implement cross-validation with tuned hyperparameters, we used the GridSearchCV function from the Scikit-learn package. Typically, the number of folds was set equal to the number of points (39) so that we could include in the accuracy evaluation the contribution of each point in the set.

The MSE metric, suitable for regression problems, was used to evaluate the performance of the models during the hyperparameter optimization process. It is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. MSE is highly sensitive to outliers due to the squaring of the errors, which can be both an advantage and a disadvantage depending on the context of the analysis [89].

2.4.2. Self-Assessment

Due to the limited dataset size, all observed points were used to train the models. Consequently, in addition to cross-validation, each model underwent self-assessment. This involved testing the fitted model on the entire training set. It is expected that for models where the input strongly influences the output, the mean squared error (MSE) of the self-

assessment will approach zero. However, some models may lose their direct relationship with the input values after training. For these models, self-assessment can be a valuable tool for making a decision about the appropriateness of the model.

2.4.3. Statistics and Visualization

The quantitative data were processed utilizing QGIS 3.28.6, SAGA 9.1.0 and Python scripts using mainly Scikit-learn library. Basic descriptive statistical analysis of the sample data was performed in MS Excel to obtain the maximum, minimum, mean, standard deviation, and coefficient of variation for the sampled HM content.

The QGIS ver. 3.28.11-Firenze and SAGA ver. 9.1.0 software were utilized for spatial visualization, and the SAGA module's multilevel b-spline function was applied in Python through command line API to analyze the spatial distribution of the PLI in the studied territory.

The data obtained underwent multivariate statistical analysis, including correlation analysis with principal component analysis (PCA). A statistical investigation was conducted using Python script, employing functions 'StandardScaler' and 'PCA' from the Scikit-learn library. Graphical visualization of the statistical data analysis results, such as plots and heatmaps, was performed using Seaborn and Matplotlib libraries.

3. Results

3.1. Statistical Distribution of Heavy Metal Concentrations in Soil

Earlier, we obtained the data on heavy metal content from 39 locations in the territory of the northern industrial zone (NIZ) in Pavlodar City. The average concentrations and statistical information on the detected elements are given in Table 2. The GOST 17.4.1.02-8 [90] classifies the detected HMs as follows: Pb, Hg, and Zn are highly hazardous elements, while Mo, Cu, Co, Ni, and Cr are moderately hazardous. Sr, Mn, and V are classified as low-hazard. Also, Table 3 shows the calculated average values of contamination factors for every metal and the total average PLI value for the entire studied territory.

The analysis shows that each element exhibits distinct concentration ranges, variability, and relationships with the background levels. The concentrations of Fe exhibit a moderate spread with an average close to the background level, suggesting minimal overall enrichment. However, individual samples might show enrichment or depletion. Similar to Fe, Mn displays a moderate spread but with an average below the background, indicating potential enrichment in some samples. A wide spread and an average below the background level point towards the enrichment of Cr concentration, with outliers potentially present. The narrow spread and average exceeding the background level of Sr concentration suggest minimal variation and possible anthropogenic sources. The moderate spread of Zn and an average above the background level imply potential enrichment in some samples. Similar to Cr, Cu exhibits a wide spread and an average below the background, suggesting enrichment with possible outliers. Lead (Pb) and cobalt (Co) show narrow spreads and averages below the background, indicating minimal variation and potential anthropogenic sources. A moderate spread of Ni and an average below the background suggest potential enrichment in some samples. Wide spreads and averages below the background level for molybdenum (Mo) and vanadium (V) point towards high variability and likely outliers. Despite minimal background data, the wide spread and low average of Hg suggest high variability and potential contamination.

The variation coefficient (CV) serves as a quantitative measure of the spatial variability in elemental content across different sampling locations. Li et al. [100] defined three categories of spatial distribution based on CV values: uniform (CV < 33%), random (33% < CV < 64%), and clustered (CV > 64%).

Table 3. Statistics on heavy metal concentrations in the northern industrial zone of Pavlodar.

Element	Minimum (mg·kg^{-1})	Maximum (mg·kg^{-1})	Average (mg·kg^{-1})	Median (mg·kg^{-1})	Standard Deviation (mg·kg^{-1})	Variation Coefficient, %	Background (mg·kg^{-1})	Average CF	PLI	MPC (mg·kg^{-1}) [1]	Sample over MPC, %
Fe	11,380	34,360	20,406.92	19,860	4810.28	24	20,680	0.99	1.31	40,000 [3]	0
Mn	240	1900	553.85	460	366.24	66	600	0.92		1500	8
Cr	0	820	203.85	150	206.88	101	440	0.46		200 [4]	79
Sr	110	250	173.33	170	23.24	13	90	1.93		600 [5]	0
Zn	30	910	121.28	70	159.10	131	60	2.02		55	79
Cu	0	630	64.87	0	110.06	170	80	0.81		33	46
Pb	0	200	46.92	50	42.56	91	32 [2]	1.47		32	67
Ni	0	150	34.10	0	47.65	140	20 [2]	1.71		20	36
Mo	0	440	11.28	0	69.55	617	50 [2]	0.23		50 [6]	3
V	0	90	7.69	0	23.26	302	150 [2]	0.05		150	0
Hg	0	100	2.56	0	15.81	618	2.1 [2]	1.22		2.1	3
Co	0	80	2.05	0	12.64	617	50 [2]	0.04		50 [7]	3

[1] MPCs were used for gross forms of metals in soil recommended by hygienic standards in [91,92] as well in [93]. [2] Element was not detected in background soil samples. The MPC value was used for calculations. [3] MPC was not established. Approximately permissible concentration (APC) value was used for the study according to [94], [4] [95], [5] [96,97], [6] [98], and [7] [99].

Within the analyzed dataset, strontium (Sr) and iron (Fe) exhibit uniform spatial distributions (CV < 33%), indicating relatively consistent concentrations across the study area. Additionally, their content levels fall below established toxicity thresholds, suggesting minimal environmental concern. Conversely, all other detected heavy metals (HMs) display clustered spatial distributions (CV > 64%), characterized by significant spatial variability and potential localized enrichment. The variation coefficients for these elements range from 66% for manganese (Mn) to 618% for mercury (Hg).

Furthermore, molybdenum (Mo), Hg, and cobalt (Co) exhibit anomalous spatial distributions, evidenced by exceptionally high CV values (617–618%). Notably, these elements also demonstrate very low overall content, with 38 out of 39 samples falling below the analytical detection limit. This pattern suggests the presence of isolated hotspots with elevated concentrations alongside widespread low-level background contamination.

Thus, the spatial distribution of elements varies, with Sr and Fe showing uniformity, other HMs exhibiting clustering, and Mo, Hg, and Co displaying an anomalous pattern. The content levels of Sr and Fe fall below toxicity thresholds, while other HMs present potential localized enrichment. Mo, Hg, and Co have very low overall content with isolated hotspots, suggesting specific contamination sources.

Industrial areas are frequently characterized by significant heterogeneity in the spatial distribution of heavy metals within soil samples [71]. This variability arises from several factors. The haphazard dumping of industrial waste in various locations creates disparate sources of heavy metal contamination with varying compositions and concentrations [101]. Industrial activities, coupled with wind and other meteorological phenomena, can disperse particulate matter containing heavy metals across the environment, leading to a complex and diffuse pattern of deposition. Due to the limitations of the research, we did not study deeper soil layers, so it is difficult to assess the contribution of the leaching and subsurface migration of HMs. However, the downward movement of water through soil layers can mobilize and redistribute heavy metals, further contributing to the heterogeneity in their spatial distribution. Consequently, these combined factors result in a patchy distribution of pollutants lacking a readily discernible structure or clear trend. This poses significant challenges for accurately mapping and characterizing heavy metal contamination in technogenic industrial areas.

3.2. Analysis of PLI Statistical Distribution

Pollution load indexes were calculated for every point in 39 locations. Table 4 shows the calculated data.

Table 4. Spatial distribution of PLI in the northern industrial zone of Pavlodar.

Sample ID	In Degrees		In Metrical Units		PLI
	X coord	Y coord	X coord	Y coord	
1	76.94639	52.30136	8,565,633.076	6,854,799.322	3.38
2	76.92663	52.31008	8,563,432.835	6,856,388.113	1.67
3	76.91402	52.33016	8,562,029.675	6,860,044.77	1.41
4	76.92706	52.33927	8,563,481.57	6,861,703.848	1.05
5	76.95862	52.3016	8,566,994.468	6,854,844.032	1.13
6	76.92942	52.37143	8,563,743.427	6,867,565.049	1.28
7	76.92623	52.3585	8,563,389.086	6,865,207.782	2.76
8	76.9341	52.37226	8,564,265.137	6,867,717.35	1.43
9	76.92768	52.37319	8,563,550.388	6,867,886.849	1.36
10	76.92706	52.39318	8,563,480.713	6,871,532.878	1.8
11	76.92536	52.33002	8,563,291.559	6,860,019.086	1.61
12	76.92515	52.37943	8,563,268.572	6,869,024.966	1.35
13	76.92655	52.3668	8,563,424.619	6,866,720.58	1.64
14	76.91806	52.4022	8,562,479.194	6,873,177.94	2.73

Table 4. *Cont.*

Sample ID	In Degrees		In Metrical Units		PLI
	X coord	Y coord	X coord	Y coord	
15	76.93407	52.37103	8,564,261.018	6,867,492.227	1.08
16	76.9173	52.38409	8,562,394.658	6,869,874.255	2.85
17	76.96258	52.3236	8,567,435.216	6,858,849.493	1.7
18	76.93838	52.36898	8,564,741.061	6,867,118.117	1.33
19	76.95269	52.40022	8,566,334.444	6,872,816.447	1.87
20	76.9298	52.38515	8,563,786.085	6,870,068.148	1.29
21	76.89894	52.36555	8,560,350.899	6,866,494.044	1.1
22	76.91425	52.31876	8,562,054.699	6,857,968.202	1.12
23	76.90053	52.3614	8,560,528.031	6,865,736.674	1.16
24	76.91124	52.38316	8,561,719.672	6,869,705.171	1.54
25	76.91913	52.31395	8,562,597.838	6,857,091.725	1.37
26	76.94659	52.32635	8,565,655.173	6,859,350.008	1.71
27	76.94892	52.31256	8,565,914.903	6,856,838.966	1.29
28	76.96725	52.3199	8,567,954.777	6,858,174.939	1.83
29	76.96179	52.33491	8,567,347.24	6,860,909.126	2.49
30	76.97273	52.31845	8,568,564.663	6,857,911.93	1.88
31	76.89961	52.33391	8,560,425.55	6,860,726.512	1.16
32	76.96859	52.32524	8,568,104.691	6,859,147.102	1.31
33	76.95853	52.38353	8,566,983.938	6,869,771.082	1.46
34	76.97169	52.325	8,568,449.67	6,859,104.846	2.12
35	76.9583	52.31219	8,566,958.88	6,856,770.774	1.56
36	76.9286	52.3197	8,563,652.546	6,858,138.17	1.85
37	76.88773	52.36098	8,559,102.718	6,865,659.803	1.41
38	76.96241	52.3002	8,567,416.403	6,854,588.54	1.64
39	76.99945	52.30655	8,571,539.677	6,855,744.188	1.39

Based on the calculated spatial data for the PLI, key statistical parameters describing its distribution were estimated:

mean: 1.64 (represents the average PLI across all samples);

median: 1.46 (indicates the 'middle' value when data are ordered);

standard deviation: 0.53 (quantifies the average deviation from the mean);

minimum: 1.05 (lowest recorded PLI value);

maximum: 3.38 (highest recorded PLI value);

range: 2.33 (difference between maximum and minimum);

coefficient of variation (CV): 32.5% (measures relative variability, higher value indicates greater spread).

Based on the calculated statistical parameters, the character of the PLI distribution can be characterized by the following. The mean PLI of 1.64 indicates a moderate level of overall pollution across the sampled locations. The standard deviation of 0.53 suggests a relatively even distribution around the mean, with most values falling within a moderate range. However, the range of 2.33 shows that some samples exhibit significantly higher or lower PLI values compared to the average. The coefficient of variation (CV) of 32.5% indicates a moderately high level of variability. This means that while the spread of values is not extreme, there is still a noticeable difference in PLI levels across different sampling points.

Combining the above points, the PLI distribution appears to be moderately skewed with a slight tendency towards higher values. While the average level of pollution is not exceptionally high, some areas exhibit pockets of significant contamination (as evidenced by the maximum value of 3.38). The moderate CV suggests that these variations are not random but might be influenced by specific factors at specific locations. Figure 2 shows the histogram and curve of the normal distribution of the PLI.

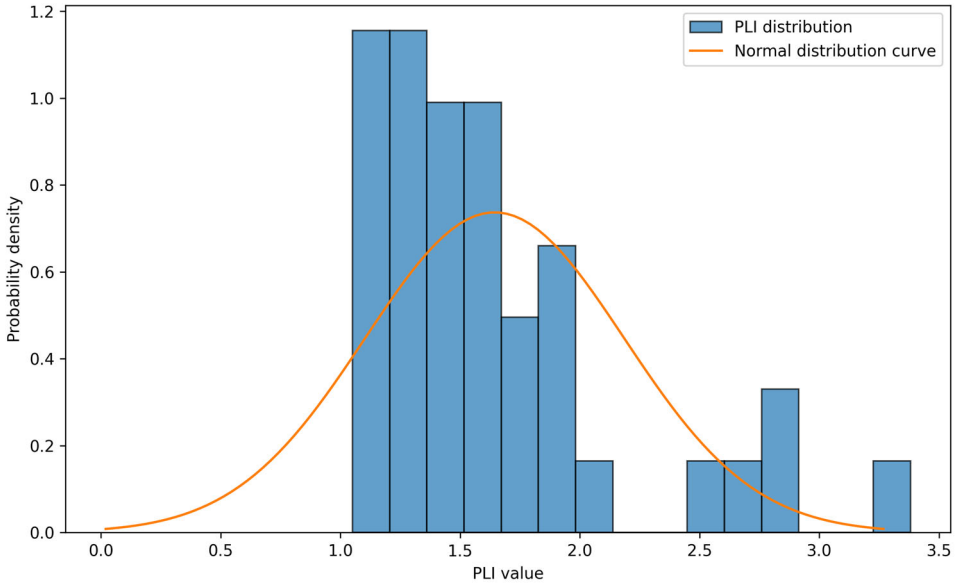


Figure 2. Histogram and normal distribution of PLI.

The PLI distribution appears to be right-skewed, meaning there are more points with lower PLI values compared to higher values. However, a significant number of outliers fall far above the main distribution, indicating areas with considerable to very high contamination. The data are widely spread, suggesting large variability in PLI values across different sampling locations. The deviation from a normal distribution curve further confirms the non-uniformity of PLI levels.

The Shapiro–Wilk test was conducted on the PLI data, resulting in a test statistic of W = 0.8335 with a *p*-value of 0.0000. A low statistic value of 0.8335 indicates that the ordered data deviates significantly from a normal distribution. The smaller the statistic, the stronger the evidence against normality. An extremely low *p*-value of 0.0000 implies that this deviation is highly statistically significant. It means that the probability of observing such a non-normal distribution by chance is virtually zero, assuming a normal null hypothesis. Given a significance level of 0.0, we reject the null hypothesis of normality (p-value < 0.05). This indicates that the PLI distribution significantly deviates from a normal distribution, confirming the visual observations of skewness and presence of outliers.

3.3. Modeling of PLI Spatial Distribution Using Machine Learning Methods

3.3.1. k-Nearest Neighbors and Weighted k-Nearest Neighbors

The most critical parameter influencing both kNN models is the number of neighbors (k). Table 4 presents the mean MSE values obtained during cross-validation with a number of folds equal to the number of observations (n = 39). This implies that 39 MSE values are generated for 39 datasets, where each dataset employs a training set of $39 - 1$ points and utilizes one point for testing. Consequently, each point contributes to the calculation of the average MSE during the cross-validation process. Lower MSE values for individual points indicate a more accurate model. Also, Table 5 shows results of accuracy of the kNN models in self-assessment process.

Table 4 reflects the slight difference between the two kNN models. In general, the WkNN model shows better accuracy. In both cases, the minimum value of the mean MSE calculated by cross-validation corresponds to k = 9. It is noteworthy that for the self-assessment, the number k negatively affects the accuracy in the case of the kNN model, but for the WkNN model, the self-estimation is always absolutely accurate. This is justified by the fact that in this model, the distance weights directly affect the prediction result.

Table 5. Dependence of mean MSE on number of neighbors (k).

k	kNN		WkNN	
	Mean MSE Cross-Validation	**MSE Self-Assessment**	**Mean MSE Cross-Validation**	**MSE Self-Assessment**
2	0.413	0.105	0.378	0.000
3	0.379	0.183	0.362	0.000
4	0.348	0.213	0.343	0.000
5	0.332	0.223	0.329	0.000
6	0.330	0.230	0.326	0.000
7	0.333	0.242	0.327	0.000
8	0.329	0.255	0.324	0.000
9	0.326	0.260	0.318	0.000
10	0.337	0.264	0.323	0.000

Considering the trade-off between higher accuracy in cross-validation and lower accuracy in self-assessment, it is necessary to select a weighted value of k for model building. The optimal value of k can be determined graphically by identifying the intersection of the trends on the segments with significant changes in the error level (Figure 3). The resulting value is approximately 5. Given that k is always an integer, we set the optimal value of k for the kNN model to 5 and for the WkNN model to 9. The two models were trained using the specified hyperparameters, and the predicted PLI values were obtained for each image point. After the prediction, process maps of the PLI spatial distribution were obtained and imported into GIS as georeferenced raster TIFF files.

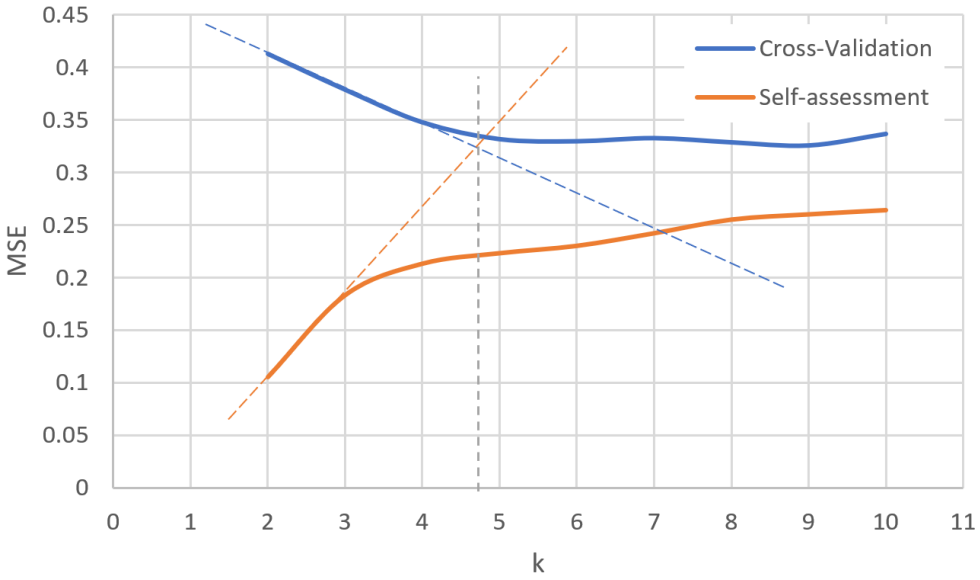


Figure 3. Visualization of error changes at different k-values for optimal k-value selection in kNN model building.

The generated maps are shown in Figures 4 and 5. The map generated through the kNN method displays sharp transitions in the PLI distribution. Areas with varying PLI values do not transition smoothly, but rather discretely with pronounced boundaries.

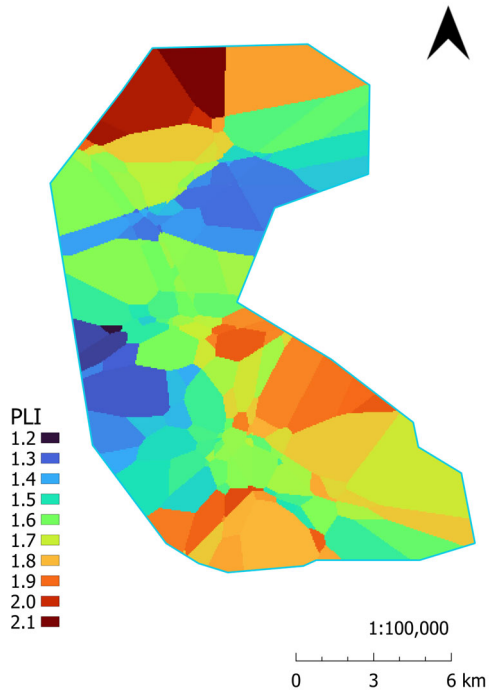


Figure 4. Spatial distribution of PLI computed with the kNN model (k = 5).

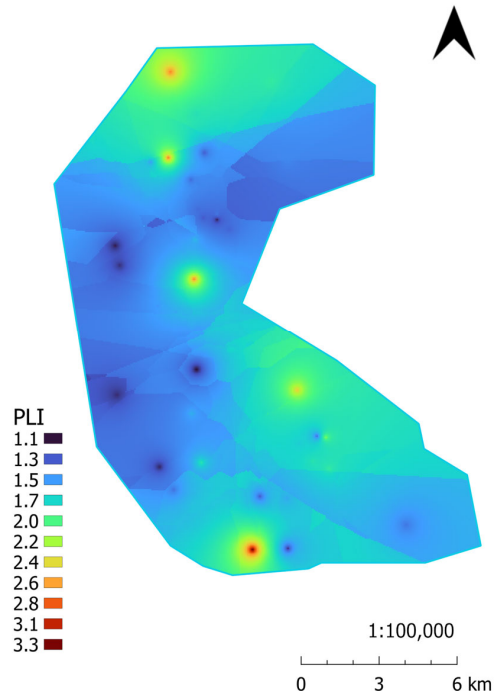


Figure 5. Spatial distribution of PLI computed with the WkNN model (k = 9, weights = 'distance').

When utilizing the WkNN model, the transitions between areas with varying PLI values are not as distinct, but still have noticeable boundaries. The contrast of these transitions is reduced because points with high PLI values have the most significant impact, which does not extend over long distances. This reduction in influence is not abrupt, but rather gradual. Thus, the influence of high values remains limited to the local area and the main map area does not vary significantly with values close to the mean.

Figures 6 and 7 show histograms and normal distribution curves for PLI values predicted with the kNN and WkNN models. The visual distribution of the predicted

PLI values shows distinctive differences for models. The histogram of the WkNN model looks closer to a normal distribution of values. However, according to the Shapiro–Wilk test, the statistic (W) of 0.9758 (kNN), 0.9682 (WkNN), and the *p*-value of 0.0000 for both models, along with the visual distribution of the data, suggest that the PLI distribution is not normally distributed either model. However, the PLI value distribution after the modeling process has no gaps in data and looks more balanced and closer to a normal distribution than the initial data.

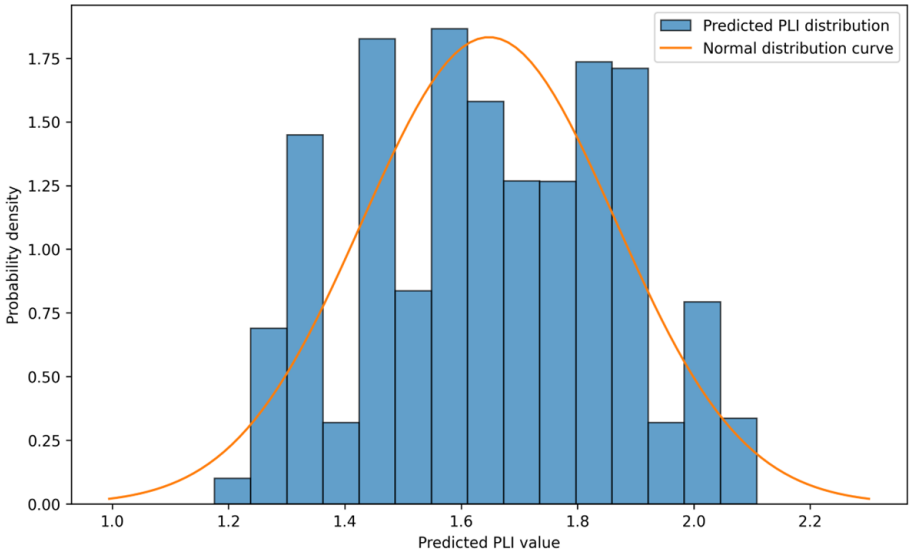


Figure 6. Histogram and normal distribution curve of PLI values predicted by the kNN model.

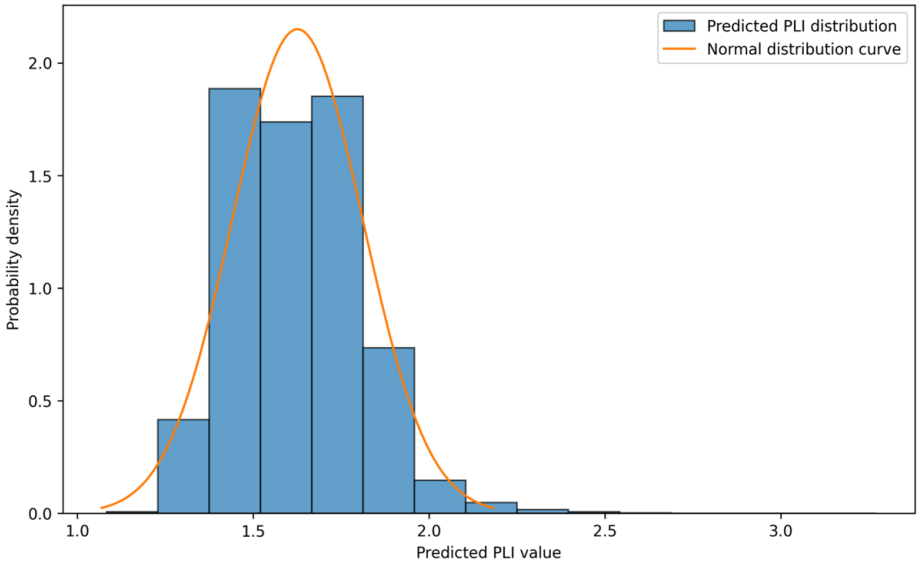


Figure 7. Histogram and normal distribution curve of PLI values predicted by the WkNN model.

3.3.2. Gradient Boosting (CatBoost Regression)

The hyperparameter values for the gradient boosting model (GB) were selected using the GridSearchCV optimizer function from the Scikit-learn library. This function is designed to automatically search for optimal hyperparameters for machine learning models by evaluating various combinations of hyperparameter values and their impact on model performance. The evaluation function used was MSE. The main hyperparameters for gradient boosting models are as follows:

- depth: the term 'depth' refers to the maximum number of splits allowed in each individual decision tree used within the ensemble;
- learning rate: a crucial hyperparameter that controls the magnitude of the update each new tree makes to the overall model prediction. Smaller learning rates lead to smaller updates, potentially slower learning, but better generalization and reduced overfitting;
- iterations: the number of individual decision trees built in the ensemble model. More iterations (more trees) can lead to higher complexity and a potentially better fit to the training data. However, too many iterations can also lead to overfitting, where the model becomes too specific to the training data and does not generalize well to unseen data;
- loss function: a critical component that defines how the algorithm measures the discrepancy between the model's predictions and the actual target values. It plays a fundamental role in driving the learning process and influencing the final model's performance.

The following values were used to select the optimal set of hyperparameters:
{'depth': [4, 5, 6, 7, 8, 9, 10],
'learning_rate': [0.01, 0.02, 0.03, 0.04],
'iterations': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100],
'loss_function': ['RMSE', 'MAE', 'Quantile', 'Poisson']}.

After searching for parameters using cross-validation (39 folds), the following best set of parameters was selected: 'depth': 10, 'iterations': 60, 'learning_rate': 0.03, and 'loss_function': 'RMSE'. The best score (MSE) across all searched parameters equals 0.2873. This value is better than accuracy for both kNN and WkNN models. The best GB model parameters were used to predict the spatial distribution of the PLI in the studied territory. After the prediction, the output TIFF file was imported into GIS. The obtained map is presented in Figure 8.

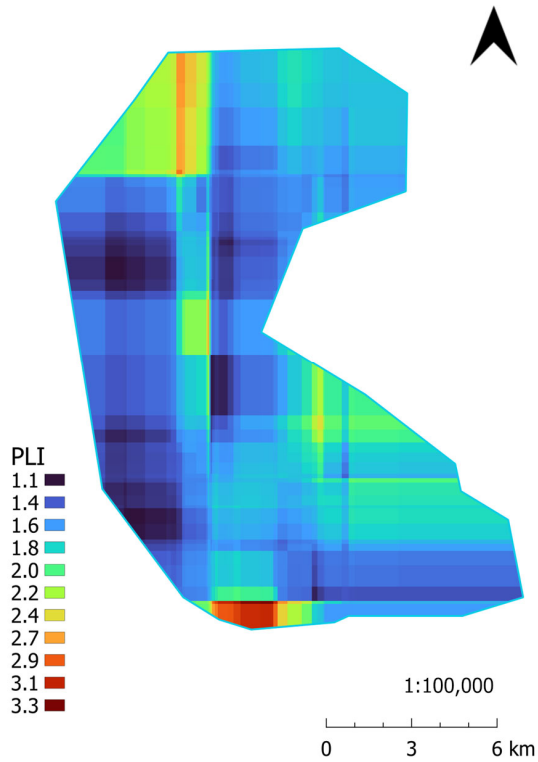


Figure 8. Spatial distribution of PLI computed with the GB1 model ('depth': 10, 'iterations': 60, 'learning_rate': 0.03, and 'loss_function': 'RMSE').

The map constructed with the GB model is characterized by a cluster distribution of PLI (Figure 8). The clusters have a pronounced rectangular shape, which may be due to the fact that only coordinate values (X, Y) were input when training the model. As in the previous maps, there are local PLI maxima in the northern section (the most extensive) and in the southern section. The clustering of increased PLI values in the central zone is more pronounced than for the WkNN model, but not as pronounced as for the kNN model. Contrary to the WkNN model, the map obtained by the gradient boosting method is more contrasted due to the absence of extreme transitions from high to low values of the ecological indicator.

Figure 9 shows the histogram and normal distribution curve for PLI values predicted with the GB1 model. The histogram of predicted values looks closer to a normal distribution in comparison to the initial data. However, according to the Shapiro–Wilk test, the highly significant *p*-value (0.0000) indicates that we can strongly reject the null hypothesis of normality. This means the data are very unlikely to be normally distributed. While the Shapiro–Wilk statistic of 0.9486 suggests non-normality, it is not as extreme as values closer to 0. This indicates moderate non-normality, not necessarily a severe deviation. With only 39 observations, the moderately high statistic suggests more substantial non-normality. Thus, the results of the Shapiro–Wilk test, along with the visual distribution of the data, suggest that the PLI distribution is not normally distributed, which confirms the visual clustered distribution of PLI values on the map (Figure 8).

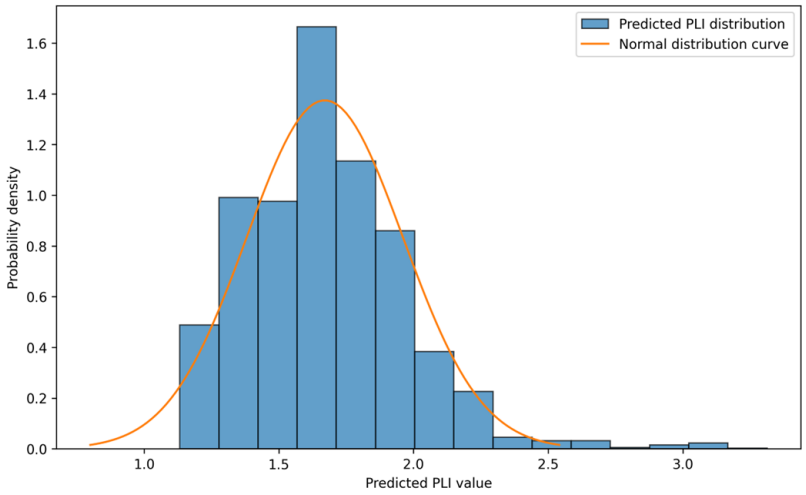
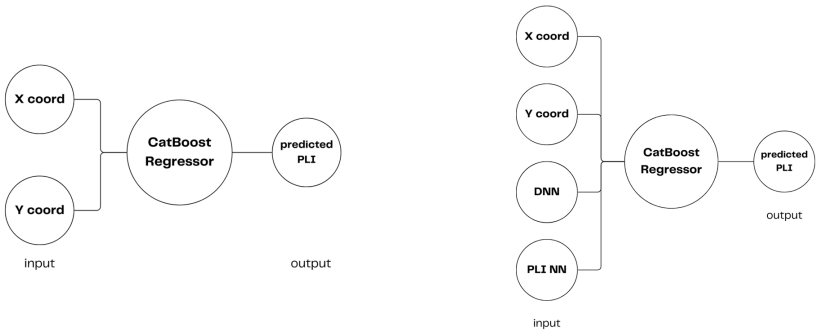


Figure 9. Histogram and normal distribution curve of PLI values predicted by the GB1 model.

The second gradient boosting model (GB2) was created by including two additional input parameters that express dependence on the closest neighboring point. This approach can improve the model by accounting for distances to the closest known points. The two parameters are the distance to the closest known location and the measured known value of the PLI in that location.

The training set consists of coordinates for each point in the initial set, as well as the distance to the closest location from the initial set (excluding the point itself to avoid zero distances). The test set includes 134,561 points with coordinates calculated using geographical data with a step of 50 m and a grid of 409 rows and 329 columns. It also includes distances to the closest locations from the initial set and the measured PLI value at the closest known location from the initial set.

The schemes for two gradient boosting models are presented in Figure 10.



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

Figure 10. Architectures for gradient boosting models: (**a**)—GB1: 'depth': 10, 'iterations': 60, 'learning_rate': 0.03, 'loss_function': 'RMSE'; (**b**)—GB2: 'depth': 5, 'iterations': 100, and 'learning_rate': 0.03, 'loss_function': 'MAE'; X coord and Y coord—geographical coordinates in metric system; DNN—distance to the nearest known neighbor from the initial set; PLI NN—PLI of the nearest known neighbor location from the initial set.

When implementing the GB2 model (Figure 10b), the best parameters from the grid search were obtained ({'depth': 5, 'iterations': 100, 'learning_rate': 0.03, and 'loss_function': 'MAE'}). Using the GB2 model (Figure 10b), the mean MSE score across all the searched parameters using cross-validation was 0.289, which is slightly worse than the score for the GB1 model (Figure 10a). The resulting map is presented in Figure 11. In Figure 11, the dependence on distances to the nearest known neighbors are expressed with radial patterns located in the areas of observations. However, the character of spatial distribution does not look as normal, and the distribution is not smooth. Near the observed locations there are radial patterns, but in the distant locations the distribution has distinct borders. The calculated Shapiro–Wilk test statistic of 0.9839 and the extremely low *p*-value of 0.0000 indicate that the PLI distribution for the GB2 model does not appear to be normally distributed.

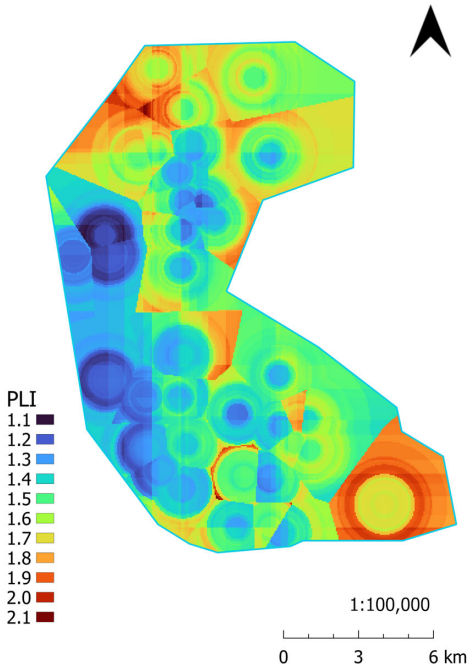


Figure 11. Spatial distribution of PLI computed using the GB2 model ('depth': 5, 'iterations': 100, 'learning_rate': 0.03, and 'loss_function': 'MAE').

3.3.3. Artificial Neural Network (MLP Regression)

To solve the set problem to find the spatial distribution of the environmental index, an artificial neural network known as the multi-layered perceptron (MLP) was used. For building the models, the MLPRegressor() function of the Scikit-learn package was used. The basic architecture of the ANN consists of:

- input layer, including two neurons, taking geographical X- and Y-coordinates in the metric system or four neurons when added to the distance to the nearest known neighbor and the PLI at the nearest known neighbor location;
- one or two hidden layers, including *n* neurons;
- output layer, consisting of one neuron, yielding a predicted value of the PLI as a result of the regression.

Architectures of ANN models with two and four input neurons are presented in Figure 12.

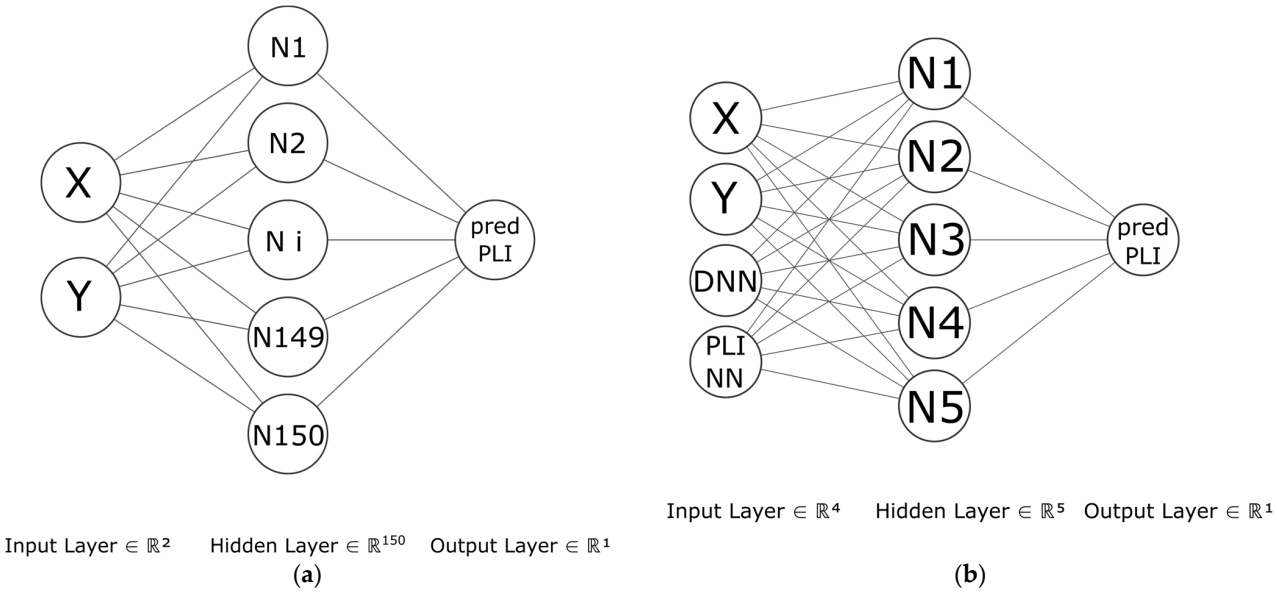


Figure 12. Architectures of the ANN models: (**a**) architecture for models ANN1 and ANN2 with two input neurons; (**b**) architecture for model ANN3 with four input neurons. X—x geographic coordinate, Y—y geographic coordinate, DNN—distance to the one nearest neighbor, PLI NN—PLI value of the nearest neighbor.

The hyperparameters of the models were optimized during the grid search process. The main hyperparameters for optimization were:

- the function of activation ('activation'), which plays a crucial role in introducing non-linearity into the network. This is vital because without non-linearity, a network would simply be performing linear regressions at each layer, ultimately leading to a limited ability to learn complex patterns in the data.
- The number of neurons in the hidden layer ('hidden_layer_sizes') is a crucial parameter that defines the architecture of the hidden layers. These layers lie between the input and output layers and play a vital role in learning complex relationships between the data and the target variable. Experimenting with different 'hidden_layer_sizes' values through grid search and cross-validation helps find the optimal architecture for the specific problem.
- The optimization algorithm ('solver') used to train the model and adjust its internal parameters (weights and biases) during the learning process. These algorithms aim to minimize a specific loss function, which measures the discrepancy between the model's predictions and the true target values. Different solvers come with various

strengths and weaknesses, making the choice crucial for achieving optimal performance.

The following set was used to grid search for the best combination of hyperparameters:
{'activation': ['identity', 'logistic', 'tanh', 'relu'],
'hidden_layer_sizes': [(5), (10), (50), (100), (150), (5,5), (10,10), (50,50), (100,100), (150,150)],
'solver': ['lbfgs', 'sgd', 'adam']}.

After searching for parameters using cross-validation (39 folds) for ANN1 with two inputs, the following best set of parameters was selected: 'activation': 'tanh', 'hidden_layer_sizes': 150, and 'solver': 'sgd'. The best MSE score across all the searched parameters equals 0.2852 (Table 5). This value is better than the accuracy for both the kNN and WkNN models and almost equal to the GB model.

The spatial distribution of the PLI in the studied area was predicted using the best ANN model parameters. The resulting output TIFF file was imported into GIS, and the obtained map is presented in Figure 13a. The map shows no changes in the studied area, with the predicted PLI values closely resembling the mean value of the initial data at every point. Therefore, in this model, the function that depends on the distance from the observed points is equal to zero. Although the model has a lower error rate, it cannot be utilized for spatial prediction with only two input parameters. To improve its accuracy, an additional parameter that expresses dependence on the distance from observed locations should be introduced into the input layer. Also, the obtained result can be explained with overfitting effects.

The activation function ReLU and solver lbfgs are the parameters that lead to differences in the PLI distribution. The set of hyperparameters that resulted in the highest accuracy with different predicted PLIs across the territory was used: activation = 'relu', hidden_layer_sizes = (5), and solver = 'lbfgs'. Although the level of error is higher, with an MSE of 0.3128, it is a compromise to avoid the negative effects of overfitting. However, the map calculated using the ANN2 model (activation = 'relu', hidden_layer_sizes = (5), and solver = 'lbfgs') shown in Figure 13b is still inadequate for predicting the spatial distribution of the environmental indicator. The PLI values are gradually distributed diagonally across the studied area. The accuracy calculated during the testing of both the ANN1 and ANN2 models is shown in Table 5.

To improve the previous neural network models, we included two parameters: distance to the closest known location and the measured known value of the PLI in that location. These parameters were given to additional two input neurons. This approach is similar to the one used with gradient boosting models. The parameters for the ANN3 were selected after a grid search procedure, where the best parameters were: 'activation': 'relu', 'hidden_layer_sizes': (100), and 'solver': 'lbfgs'. The ReLu activation function was left as the only choice due to its ability to provide a non-zero spatial distribution. Although other activation functions may result in lower overall error, they produce a zero distribution and a PLI value for the entire area that is close to the mean value. The lowest error level achieved was 0.346. Additionally, the self-assessment results showed a positive effect from the inclusion of two additional input parameters. The MSE for the self-assessment was the lowest with this model compared to the other ANN models, reaching 0.278 (Table 6).

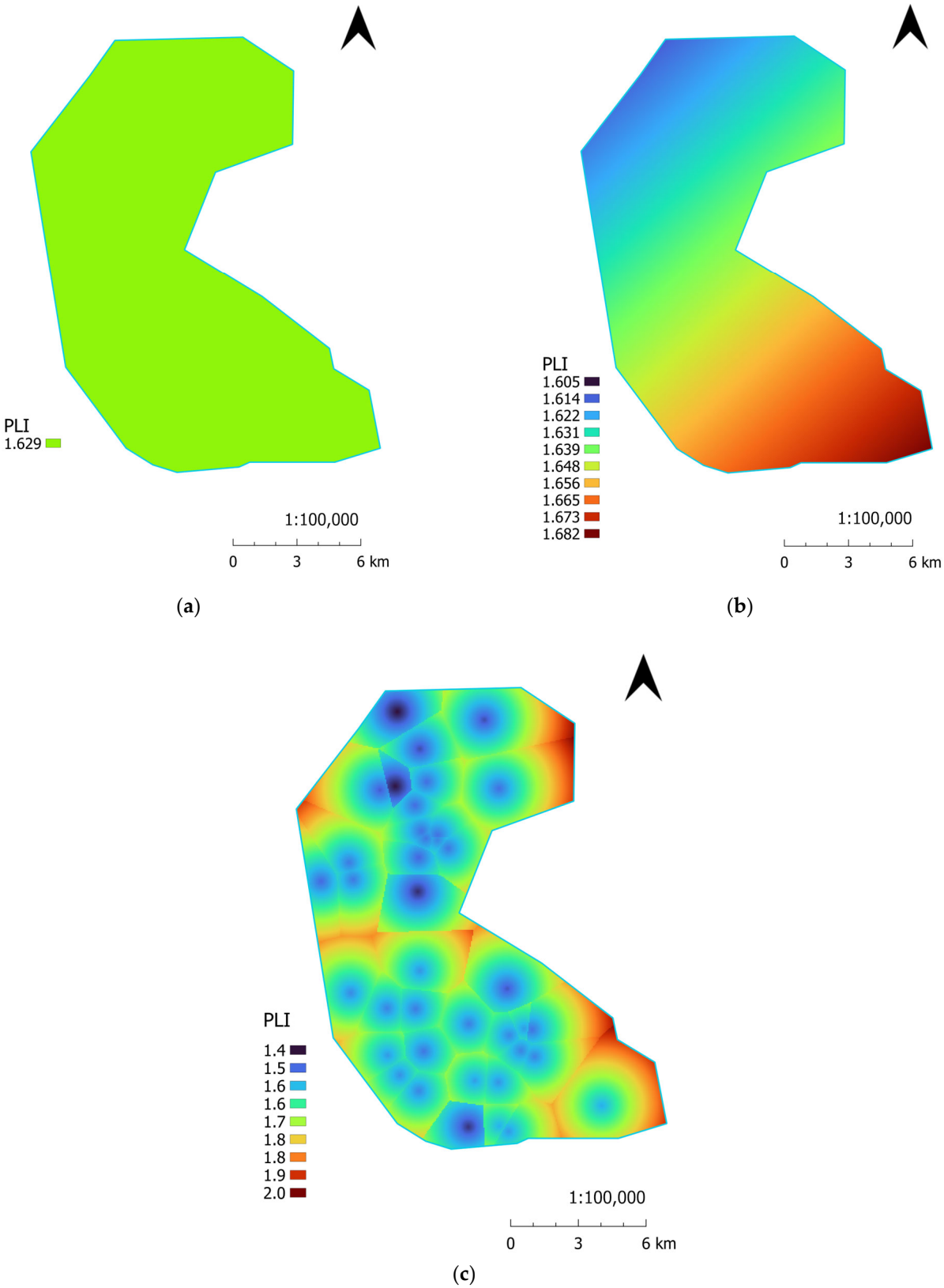


Figure 13. Spatial distribution of PLI computed with the ANN models: (**a**)—ANN1: two inputs ('activation': ' tanh ', 'hidden_layer_sizes': (150), and 'solver': 'sgd'); (**b**)—ANN2: two inputs (activation = 'relu', hidden_layer_sizes = (5), and solver = 'lbfgs'); (**c**)—ANN3: four inputs (activation = 'relu', hidden_layer_sizes = (5), and solver = 'lbfgs').

Table 6. Accuracy of ANN models.

ANN1 2 Input ('activation': ' tanh ', 'hidden_layer_sizes': (150), 'solver': 'sgd')		ANN2 2 Input (activation = 'relu', hidden_layer_sizes = (5), solver = 'lbfgs')		ANN3 4 Input (activation = 'relu', hidden_layer_sizes = (5), solver = 'lbfgs')	
Mean MSE Cross-Validation	**MSE Self-Assessment**	**Mean MSE Cross-Validation**	**MSE Self-Assessment**	**Mean MSE Cross-Validation**	**MSE Self-Assessment**
0.285	0.286	0.313	0.285	0.346	0.278

Based on the visual estimation of the three maps obtained using the ANN models, it appears that these models only occasionally predict the PLI values (Figure 13). The 'best' predictions are summarized as the output of the mean value of the PLI. However, the ANN3 model with four inputs (Figure 13c) shows a distribution that is more similar to the natural distribution due to the addition of a function used to account for the distances to known points.

Figure 14 shows histograms and normal distribution curves for the PLI values predicted by the three ANN models. The visual distribution of the predicted PLI values shows distinctive differences between the models. The first histogram (Figure 14a) is absolutely abnormal because all values in the predicted data are represented by the same number, close to the mean value of the initial PLI values. This result is confirmed with a very low Shapiro–Wilk test statistic: 0.0012. The histogram of the second ANN2 model (Figure 14b) looks closer to a normal distribution of values. For this model, the Shapiro–Wilk test statistic is quite high and reaches 0.9911. However, the *p*-value of 0.0000 for both models, along with the visual distribution of the data, suggest that the PLI distribution is not normally distributed for either model. The histogram of the third ANN3 model is not normal either, and has a skewed form (Figure 14c), a Shapiro–Wilk test statistic of 0.9065, and a *p*-value of 0.0000, which indicate that the PLI distribution does not appear to be normally distributed.

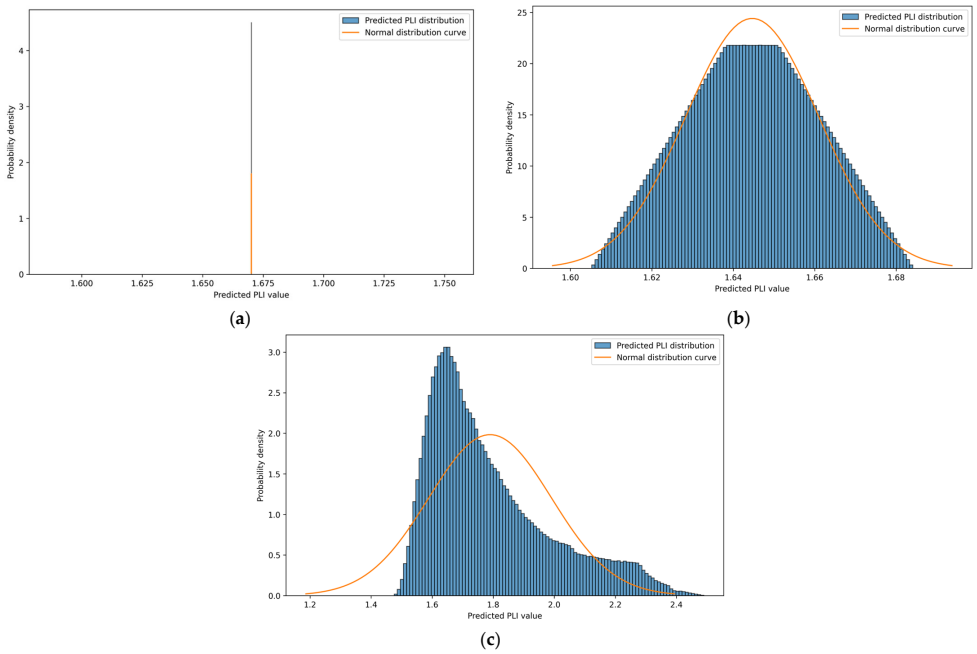


Figure 14. Histogram and normal distribution curves of PLI values predicted by the ANN models: (**a**)—ANN1: two inputs ('activation': ' tanh ', 'hidden_layer_sizes': (150), and 'solver': 'sgd'); (**b**)—ANN2: two inputs ('activation' = 'relu', 'hidden_layer_sizes' = (5), and 'solver' = 'lbfgs'); (**c**)—ANN3: four inputs, 'activation': 'relu', 'hidden_layer_sizes': (100), and 'solver': 'lbfgs').

3.3.4. Kriging Model

To solve the set problem and determine the spatial distribution of the environmental index, the Kriging method was used. For building the models, the rk.Krige() function of the PyKrige package was used. The hyperparameters of the models were optimized during the grid search process. The main hyperparameters for optimization were:

- method: ordinary or universal;
- variogram model: 'linear', 'power', 'gaussian', 'spherical', 'hole-effect', 'exponential';
- number of lags (nlags): specifies the number of lags used in the variogram calculation. This determines the level of detail captured in the spatial relationship;
- calculation of weights (weight): we should choose to use or not perform calculations. Each weight reflects the influence of a specific sampled data point on the prediction at a particular unsampled location. Points closer to the target location and points that exhibit similar values tend to have higher weights, contributing more significantly to the prediction.

The following values were used to select the optimal set of hyperparameters:
{'method': ['ordinary', 'universal'],
'variogram_model': ['linear', 'power', 'gaussian', 'spherical', 'hole-effect', 'exponential'],
'nlags': [4, 6, 8, 16],
'weight': [True, False]}.

After searching for parameters using cross-validation (39 folds) the following best set of parameters was selected: 'method': 'universal', 'nlags': 6, 'variogram_model': 'hole-effect', 'weight': True. The best MSE score across all the searched parameters equals 0.282, which is the best value across all of the above-mentioned models.

The spatial distribution of the PLI in the studied area was predicted using the best Kriging model parameters. The resulting output TIFF file was imported into GIS, and the obtained map is presented in Figure 15. The map shows that the distribution is in accordance with the initial data, taking into account distance weights and the influence of the PLI value at every known point, which is radially distributed from the location. Areas without observations are derived from the mean PLI value from the initial data set. Transitions from one known point to another, as well as from known points to unknown areas, are smooth without any sharp borders. Additionally, this prediction method does not produce predicted PLI values that exceed the real maximum and minimum PLI values. The self-assessment calculation shows a predictable result of zero for the MSE. This is due to the function used to weigh distances in the calculation. The PLI value has a greater influence on the prediction result for points closer to the known point, ultimately leading to the known values.

Figure 16 shows the histogram and normal distribution curve for the PLI values predicted by the Kriging model. The histogram looks more or less balanced, with a high peak for the average value. This is logically explained by the peculiarities of the model, because the unknown areas come down to the mean value of the PLI. The limitations of real observation numbers lead to difficulties in predictions, that is, we cannot absolutely accurately assert the veracity of the calculated prediction, but at least we can suppose that the obtained distribution is more probable than simply using the average values for every point, which does not correspond to the initial dataset. For the obtained Kriging model, the Shapiro–Wilk test statistic achieved 0.8629 and the p-value was 0.0000, which indicates that the PLI distribution does not appear to be normally distributed.

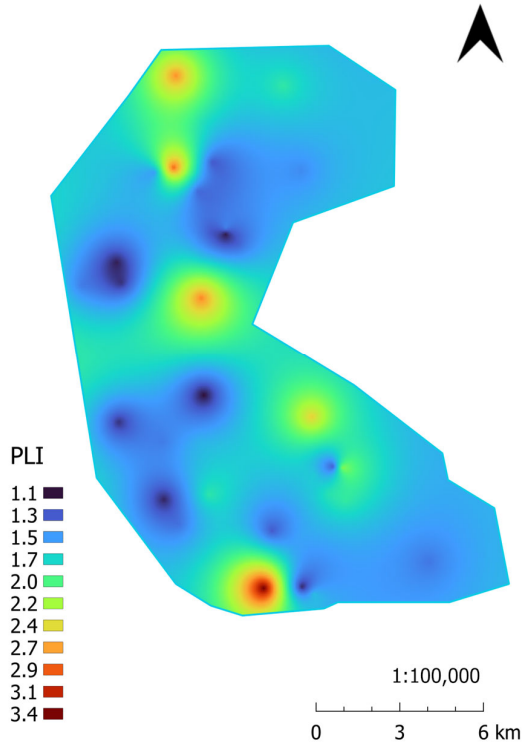


Figure 15. Spatial distribution of PLI computed with the Kriging model: 'method': 'universal', 'nlags': 6, 'variogram_model': 'hole-effect', and 'weight': True.

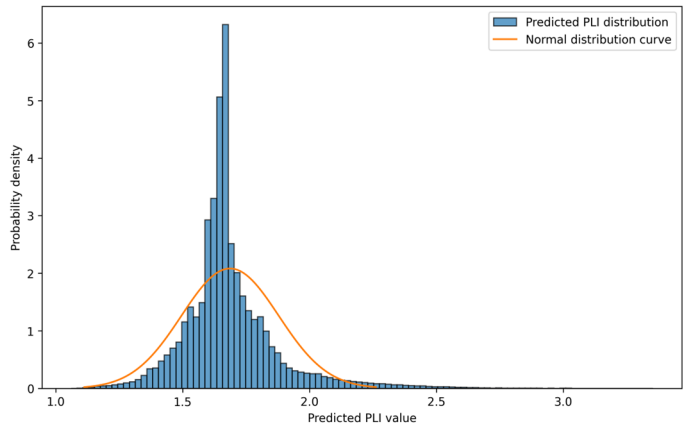


Figure 16. Histogram and normal distribution curves of PLI values predicted by the Kriging model: 'method': 'universal', 'nlags': 6, 'variogram_model': 'hole-effect', and 'weight': True.

3.3.5. Multilevel b-Spline Interpolation

As MLBS interpolation is not a machine learning method in a direct sense; this method has no specific parameters to adjust. However, taking in to account the appropriate area of implementation, this method was used for comparison with the machine learning methods. Thus, this method was used without searching for any parameters using cross-validation (39 folds). The mean MSE score after a full round of cross-validation equals 0.404, which is not the best value across all of the above-mentioned models. However, the map obtained is smooth with a distinctive distribution of the observed index. Taking into account the low amount of initial data, the mean MSE cannot absolutely precisely predict the accuracy of the model. The results from each model are more or less approximated, which is proved by the extremely low Shapiro–Wilk statistics and *p*-values for every model. Thus, the level of approximation is set by the observer, but it has a several limits based on the calculated distribution:

- the calculated value cannot be equal to the average (or close to) value in every predicted point;
- the distribution cannot follow a simple linear dependence (like simple gradients), at least for difficult dependencies such as the distribution of pollutants in environmental objects, which are characterized by numerous reciprocal influences with many factors;
- the calculated distribution has to reflect the observed values, assuming the method of observation and observed data are correct.

The MLBS interpolation model was used to determine the spatial distribution of the PLI in the studied area. The resulting output TIFF file was imported into GIS, and the obtained map is presented in Figure 17. The map shows that the distribution is in accordance with the initial data, taking into account distance weights and the influence of the PLI value at every known point, which is radially distributed from the location. The areas between the observations do not approach the mean PLI value from the initial data set. In this model, the PLI values from known points transform smoothly in all directions. Due to mathematical peculiarities, this method produces predicted PLI values that slightly exceed the real maximum and minimum PLI values. This effect is explained by the increase in spline curves on the edges. The self-assessment calculation shows a predictable result of zero for the MSE. The reason for this is the weighting function used in distance calculation. The PLI value has a stronger impact on the prediction outcome for points that are closer to the known point, resulting in the known values.

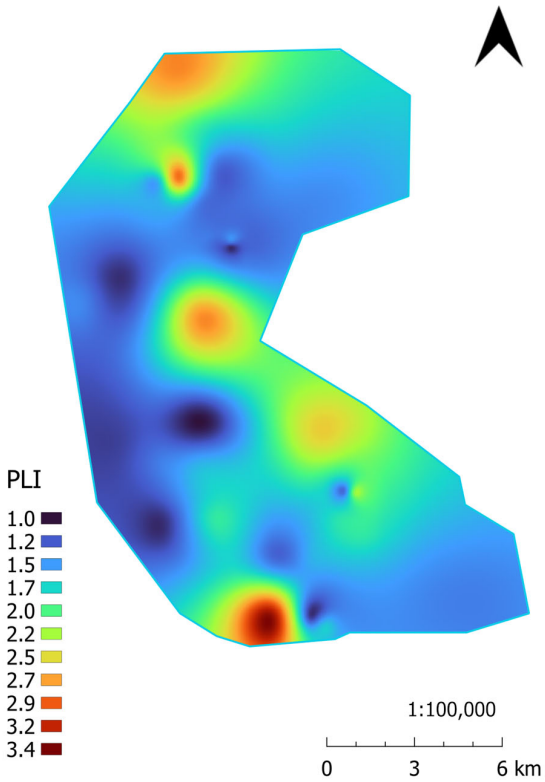


Figure 17. Spatial distribution of PLI computed with the MLBS interpolation model.

Figure 18 shows the histogram and normal distribution curve for the PLI values predicted by the Kriging model. The histogram looks skewed to the right, with two high peaks close to the average value (in areas of PLI, about 1.3 and 1.7). For the obtained MLBS interpolation model, the Shapiro–Wilk test statistic achieved 0.9332 and the *p*-value was 0.0000, which indicates that the PLI distribution does not appear to be normally distributed.

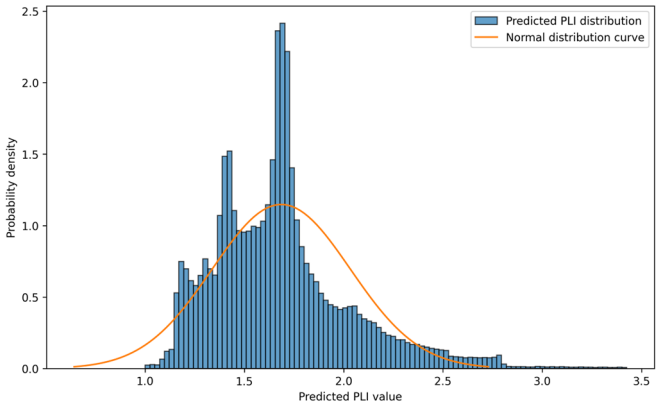


Figure 18. Histogram and normal distribution curves of PLI values predicted by the MLBS interpolation model.

4. Discussion

The research conducted in the northern industrial zone (NIZ) of Pavlodar City, Kazakhstan, utilized computational methods to predict the spatial distribution of the environmental risk indicator PLI, particularly focusing on heavy metal contamination in the region. Identifying areas in terms of PLI can be a challenging task due to the complex nature of industrial pollution. Heavy metal contamination levels and consecutively PLI can vary extensively among cities, countries, continents, and time periods [102]. Additionally, co-contamination of heavy metals is common in industrial areas due to the presence of heavy metals in industrial waste [103,104]. In our research area, the concentration coefficients for each metal and pollution index are distributed unevenly, with a noticeable clustering of pollutant distribution. Due to the large number of plants in the area, determining trends is challenging. However, it is possible to identify that the maximum PLI values are mainly located to the southwest of the TPPs over some distance from their territories according to the average wind rose for Pavlodar City (Figure 19). The text is in agreement with the theory of the distribution of smoke heavy components that are transported by air [105]. Furthermore, there is another area of high PLI in the northern part of the studied region, near the Hg-accumulating ponds. The high PLI level is generally attributed to the high concentration of Hg. While only three locations have Hg concentrations above the MPC, the extreme toxicity of this element results in a wide zone of contamination.

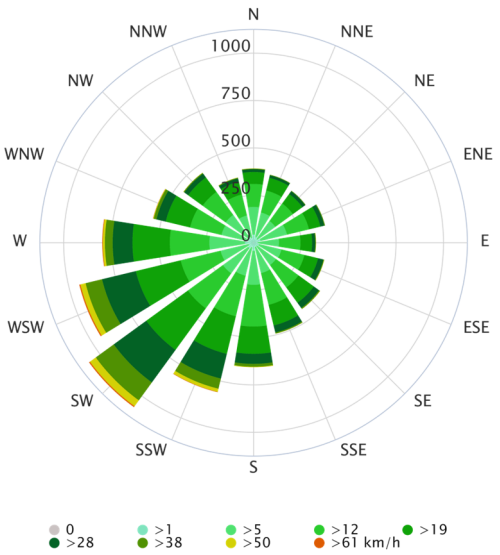


Figure 19. Average wind rose for Pavlodar City [106].

The study utilized nine models to analyze the data and evaluate their performance in predicting pollution levels in the area. Each model was used to calculate the regression problem and obtain the PLI values at the given points. Optimal parameters of the models were selected with GridSearchCV using cross-validation and self-assessment.

The primary metric used to assess accuracy was mean squared error (MSE), specifically the mean MSE in the cross-validation process. The number of folds for cross-validation was equal to the number of observed points (39) to assess the contribution of each point to the mean error level. The final model was created by fitting all observed values. For several models where the initial data was not used directly for creation, self-assessment was also implemented. Self-assessment involves estimating the mean squared error (MSE) for predicted values using the initial data as a testing set after training the model. In cases where the initial data directly influences the prediction or weighted distances to the nearest known values are taken into account, the error level in self-assessment was approaching zero.

Furthermore, the Shapiro–Wilk statistic was likely used to test the normality of the residuals in the models. This statistic assesses whether a dataset follows a normal distribution, which is crucial for ensuring the validity of statistical analyses and predictions [107]. By examining the Shapiro–Wilk statistic across the models, we can determine the appropriateness of the model assumptions and the reliability of the predictions.

In evaluating the usefulness of the models with lower mean MSE values and higher Shapiro–Wilk statistics are indicative of better predictive capabilities and a closer fit to the actual data. Also, the models were assessed visually for accordance to several principles of visual appropriateness for spatial distribution of pollutants in environment.

The benchmark results (Table 7) were used to conduct a correlation analysis with principal components analysis (PCA) to identify the features that most strongly influence the appropriateness conclusion. The Shapiro–Wilk *p*-value was excluded from the analysis because it was consistently zero across all models. Boolean values of visual parameters assessed subjectively were transposed to 1 and 0, for 'Yes' and 'No', respectively. 'StandardScaler' transformation from Scikit-learn package was implemented to data to standardize the range of features in a dataset. This scaler removes the mean and scales the data to unit variance (from −1 to 1).

Figure 20 shows obtained scree plot visualizing explained variance ratio. From the scree plot we can see that first two principal components capture over 74.3% of total variance in the dataset. So, we can concatenate the principal components with the original data for conducting correlational analysis.

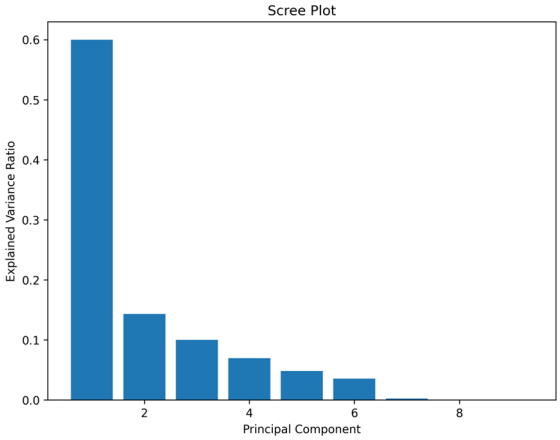


Figure 20. Scree plot of principal components analysis of model benchmarking results.

Table 7. Benchmarking Model Appropriateness.

Model	kNN	WkNN	CB1	CB2	ANN1	ANN2	ANN3	Kriging	MLBS
Parameters	k: 5	k: 9, weights: 'distance'	2 features: (X coord, Y coord) 'depth': 10, 'iterations': 60, 'learning_rate': 0.03, 'loss_function': 'RMSE'	4 features: (X coord, Y coord, DNN, PLI NN) 'depth': 5, 'iterations': 100, 'learning_rate': 0.03, 'loss_function': 'MAE'	2 inputs (X coord, Y coord) 'activation': 'tanh', 'hidden_layer_sizes': (150), 'solver': 'sgd'	2 inputs (X coord, Y coord) activation: 'relu', hidden_layer_sizes: (5), solver: 'lbfgs'	4 inputs (X coord, Y coord, DNN, PLI NN) activation: 'relu', hidden_layer_sizes: (5), solver: 'lbfgs'	'method': 'universal', 'nlags': 6, 'variogram_model': 'hole-effect', 'weight': True	n.a.
Mean MSE Cross-Validation	0.332	0.318	0.287	0.289	0.285	0.313	0.346	0.282	0.404
MSE Self-Assessment	0.223	0.000	0.090	0.000	0.286	0.285	0.278	0.000	0.000
Shapiro–Wilk statistic (W)	0.9758	0.9682	0.9486	0.9839	0.0012	0.9911	0.9065	0.8629	0.9332
Shapiro–Wilk *p*-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Non-Uniform Prediction	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Non-Linearity	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes
Empirical Consistency	Yes	Yes	Yes	No	No	No	No	Yes	Yes
Smoothness	No	No	No	No	No	Yes	No	Yes	Yes
Min	1.176	1.083	1.132	1.148	1.629	1.605	1.432	1.061	1.001
Max	2.108	3.272	3.312	2.068	1.629	1.684	2.479	3.353	3.425
Min–Max Difference	1.398	0.141	0.150	1.410	2.330	2.251	1.283	0.038	0.094
Appropriateness Conclusion	Yes	Yes	Yes	No	No	No	No	Yes	Yes

To assess correlations between different parameters of predicted PLI data and subjective accessibility of the model, we used a correlation matrix with Pearson's coefficients (Figure 21). The Pearson correlation coefficient indicates the level of linear association between two variables on a scale from −1 to 1. A score of −1 indicates a fully negative linear correlation, while 0 signifies a lack of linear correlation, and 1 represents a complete positive linear correlation [108].

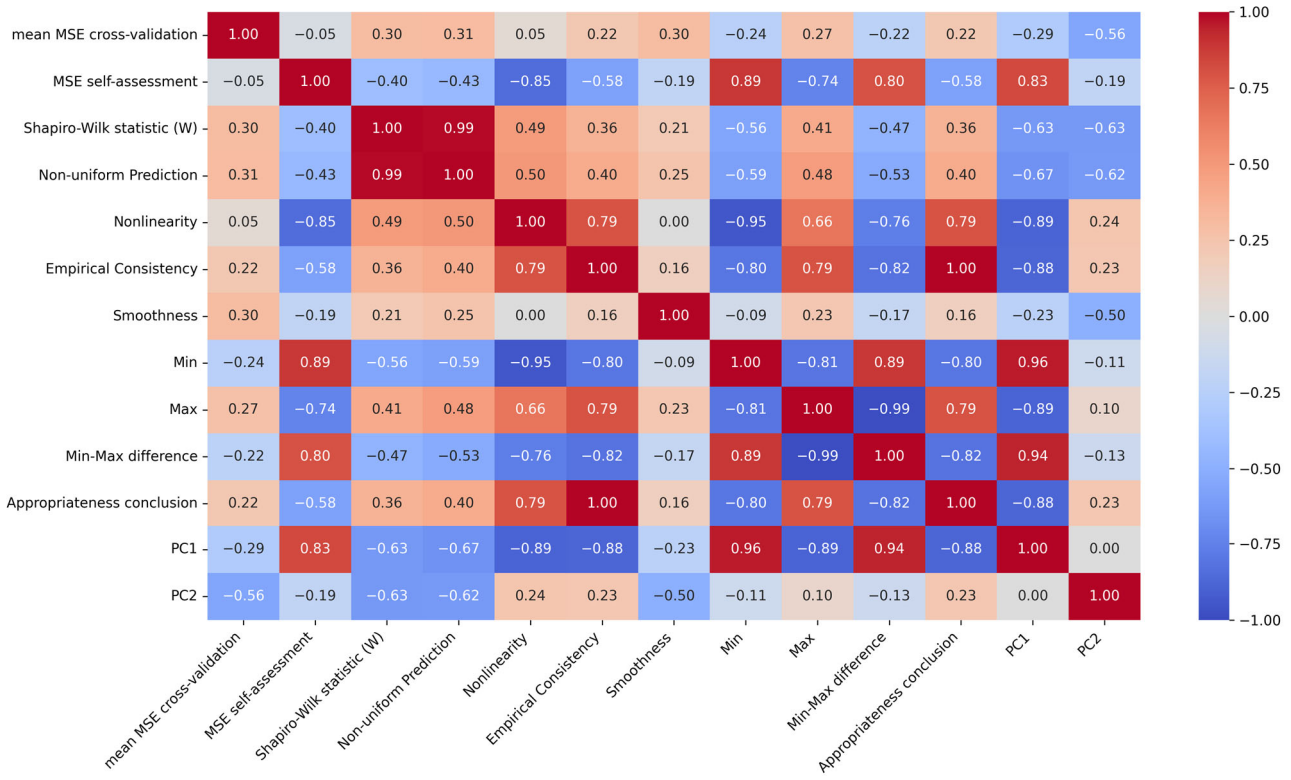


Figure 21. Heatmap of correlation coefficients between parameters of assessment of predicted datasets and appropriateness conclusion including principal components.

The correlation matrix reveals patterns and trends in the data. Specifically, two principal components explain the majority of the data. Figure 22 illustrates the relationship between PC1 and PC2 and the appropriateness of the model. Notably, changes in PC2 do not affect the model quality. However, decreasing PC1 leads to a positive indication of the model's appropriateness. That can be seen from correlation matrix as well. So, the relation between PC1 and the appropriateness conclusion is strongly negative with a Pearson's coefficient of −0.88. The main significant features contributing to PC1 are non-linearity (−0.89), empirical consistency (−0.88), MSE self-assessment (0.83). Features like Min, Max, and Min–Max difference have a significant influence, but they were excluded logically. For example, despite the Min value showing strong positive relation to PC1 (0.96), which means that minimal value of Min expresses more adequate model. However, the minimum and maximum values are limited to certain real values that are expected to fall within the range of the observed minimum and maximum values. So, constantly decreasing the Min value will not improve the adequacy of the model.

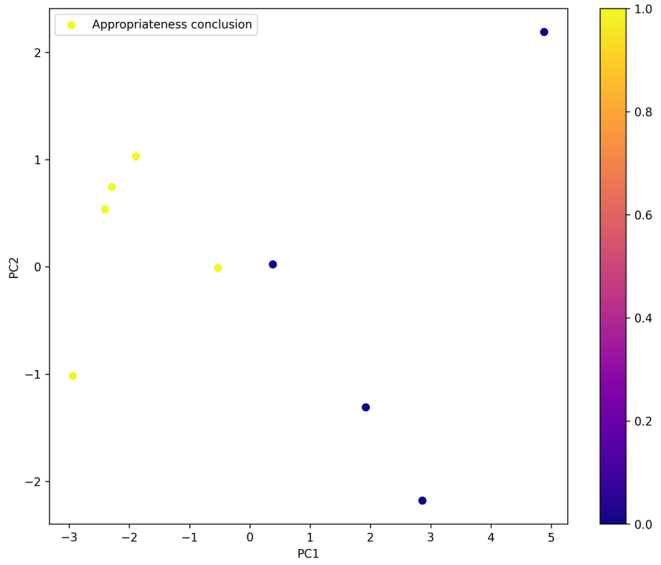


Figure 22. Dependence of appropriateness rate on PC1 and PC2.

The model's appropriateness is most strongly correlated with empirical consistency (1.00), which is evident. The non-linearity feature has a strong positive correlation (0.79). It is interesting to note that smoothness has a weak correlation (0.16), and the MSE characteristics for both cross-validation and self-assessment do not have a significant influence. Additionally, self-assessment is a more correlating measure than cross-validation. This can be explained by the small size of the observed data and the non-uniform distribution of the observed PLI values.

In this study, we evaluate nine distinct models for predicting spatial distribution. Our selection process is guided by both visual analysis and empirical consistency. Among the considered methods, we identify k-nearest neighbors (kNN), weighted k-nearest neighbors (WkNN), gradient boosting, Kriging, and multi-layered b-spline interpolation as appropriate candidates. Notably, we prioritize visual expressiveness, favoring smoother maps. Consequently, we opt for the Kriging model and multi-layered b-spline (MLBS) interpolation models. These choices enable a more accurate representation of the spatial distribution of environmental indicators.

The kNN method is straightforward and easy to implement, requiring no training phase and adaptable to both classification and regression tasks. However, it is computationally intensive, requires significant memory, and can be sensitive to irrelevant features and noise. WkNN improves accuracy by assigning weights to neighbors, reducing the influence of outliers and handling varying data densities better. Despite these benefits, WkNN adds computational complexity and requires careful tuning of the weighting function, sharing many of the same limitations as kNN, such as high memory usage and sensitivity to the size of dataset. In general, both kNN and WkNN reflect the spatial distribution of environmental index, but the visualization is not smooth, locations with different predicted PLI levels have sharp borders. This effect decreased with WkNN.

Gradient boosting offers high predictive accuracy and effectively models complex spatial relationships, but it requires significant computational resources and careful parameter tuning to avoid overfitting. Additionally, the method's complex models may pose challenges in interpretation compared to simpler methods. Interestingly, both models shown similar accuracy with MSE 0.287 and 0.289 in cross-validation, but the model with only two inputs was presentable visually. The four-input gradient boosting model resulted concentrical patterns related to sample points.

ANNs excel in modeling complex spatial relationships and handling large datasets, providing high predictive accuracy for spatial distribution analysis. However, they require substantial computational resources, extensive training time, and extremely depend on the

size of dataset. ANNs with only X- and Y- coordinates as inputs shown results of extra generalization: in one case it was the one value close to average PLI for the entire map; in other case it was a slight gradient with limits again close to average PLI from 1.605 to 1.682. These results are not representable and do not reflect the real distribution of the environmental index. Adding another two inputs made the obtained map more uneven, but it was similar to map obtained from four-input gradient boosting model and still was non-representable visually. Moreover, the MSE value was higher (0.346).

Kriging offers high accuracy and provides a measure of prediction uncertainty, making it a robust method for spatial distribution analysis, especially with well-correlated data. However, it can be computationally intensive and requires a thorough understanding of the spatial correlation structure. Multilevel b-spline interpolation is more computationally efficient and flexible, handling large datasets and varying spatial resolutions well. Nonetheless, it may lack the predictive accuracy and uncertainty quantification that Kriging provides, particularly in regions with complex spatial variability. Both of these models resulted in visually acceptable maps. However, MSE in self-assessment for Kriging was minimal (0.282) and for MLBS it was maximal (0.404).

It should be noted that all of the proposed approaches can be successfully used in spatial analyses to a greater or lesser extent. However, the low results of the evaluation of the effectiveness of predicting the distribution of the environmental index, lying within the MSE 0.28–0.4, are a consequence of the limitations of the study. The first and most significant limitation is related to the small sampling size of 39 points. This limitation is crucial for models using machine learning techniques, which require large data sets for correct training, validation, and to avoid overfitting effects. The small sample size has even affected the gradient boosting model, which usually show greater efficiency on small data compared to neural networks. However, this limitation is very common in ecological studies. Since large scale monitoring with processing of at least 1000 points, is a very difficult task even at the government level. Collecting and processing such a large number of samples requires considerable resources, both material and human. The aim of the study was to find the most efficient methods of spatial visualization of the ecological index on small samples. This task is difficult in itself, but also significant not only for this study, but also for other spatial analysis tasks.

The second important limitation is related to the nature of pollution level distribution. The distribution of environmental risk in industrial zones is non-linear, difficultly predictable, and has a pronounced cluster character. The level of pollution can vary greatly from point to point, causing difficulties in the interpretation of intermediate locations that have not been physically studied. Therefore, any calculations in this field are always performed with some approximations and assumptions. Industrial zones are characterized by a wide range of production facilities, active construction, and the presence of waste storage areas and technogenic materials. This introduces considerable uncertainty in the search for sources of pollution. At the same time in our study the mercury pollution zone and the zones of influence of ash emissions from TPPs, which correspond to the average annual wind rose direction, were confidently identified.

5. Conclusions

As a result of the research, the following main results were obtained:

1. The study focused on soil samples collected from 39 locations in the research area, with elemental analysis of heavy metals including Pb, Hg, Zn, Mo, Cu, Co, Ni, Cr, Sr, Mn, and V. The data revealed significant variability and potential contamination in several elements, particularly chromium (Cr), zinc (Zn), and lead (Pb), where a substantial percentage of samples exceeded their respective MPCs. For example, 79% of chromium samples exceeded the MPC of 200 $mg \cdot kg^{-1}$, with concentrations ranging from 0 to 820 $mg \cdot kg^{-1}$ and an average of 203.85 $mg \cdot kg^{-1}$. Zinc also showed high levels of contamination, with 79% of samples exceeding the MPC of 55 $mg \cdot kg^{-1}$, and concentrations ranging from 30 to 910 $mg \cdot kg^{-1}$ and an average of 121.28 $mg \cdot kg^{-1}$. Ad-

ditionally, 67% of lead samples exceeded the MPC of 32 $mg \cdot kg^{-1}$, with concentrations ranging from 0 to 200 $mg \cdot kg^{-1}$ and an average of 46.92 $mg \cdot kg^{-1}$. Conversely, elements such as iron (Fe), strontium (Sr), and vanadium (V) showed minimal exceedance of MPCs. All iron, strontium, and vanadium samples were below the corresponding MPCs. Overall, the analysis underscored the importance of continuous monitoring and targeted intervention strategies to address the identified contamination issues, particularly for chromium, zinc, and lead. Efforts should focus on remediation and stricter regulatory measures to manage and mitigate the environmental and health risks associated with these contaminants.

2. Nine mathematical models based on kNN, gradient boosting, artificial neural networks, Kriging, and multilevel b-spline interpolation methods were employed to predict pollution levels, with the primary accuracy metric being MSE. In this study, we were focused on visual expressiveness and empirical consistency. As a result of the model's benchmarking, Kriging and MLBS interpolation were ultimately chosen for their ability to create smooth, visually appealing maps that accurately represent the spatial distribution of environmental indicators, with MSE values of 0.282 and 0.404, respectively. While kNN and WkNN are straightforward and adaptable, they are computationally intensive and less visually smooth. Gradient boosting and ANNs offer high predictive accuracy, with MSE values around 0.287 and 0.289 for gradient boosting, but are resource-heavy, complex, and visually unacceptable from the point of view of empirical consistency.
3. Taking into account the impossibility to accurately assess the accuracy of the model because of the small initial dataset, the following four visual parameters were used for the assessment of the appropriateness of the maps obtained with the computational models: non-uniform prediction, non-linearity, empirical consistency, and smoothness. Based on the conducted correlation analysis, the most important features were empirical consistency and non-linearity.

Each studied method of interpolation can be used for spatial distribution analysis; however, a comparison with the scientific literature revealed that Kriging and MLBS interpolation can be used without extra calculations to produce non-linear, empirically consistent, and smooth maps.

Kriging produces maps with a greater contribution from mean values, while MLBS interpolation produces maps with a greater contribution from observed points. These methods can assist in identifying contamination hotspots and guiding environmental remediation efforts.

After reviewing the benchmarking of the nine mathematical models, the Kriging model was the most effective, taking into account the limitations of a small dataset and a highly uneven distribution of PLI values. This model yielded the lowest level of MSE with a smooth and trustful visual representation of the environmental index distribution map.

Looking ahead, the future of environmental sustainability is closely intertwined with digital innovation. Emerging technologies such as artificial intelligence, blockchain, and the Internet of Things hold great potential for transforming how we monitor, manage, and protect the environment. By embracing these technologies and staying ahead of trends, organizations can drive positive change and create a more sustainable future for all.

Digitalization plays a pivotal role in advancing environmental sustainability by offering tools, strategies, and opportunities to address pressing environmental challenges. By harnessing the power of digital technologies, we can make significant strides towards a greener, more sustainable future for generations to come. It is imperative that we continue to embrace digital solutions, collaborate across sectors, and innovate for a better tomorrow.

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

Funding: This research and publication were funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP14872294).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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.

References

1. Thompson, P.B. Ethics and Environmental Risk Assessment. In *Food and Agricultural Biotechnology in Ethical Perspective*; Thompson, P.B., Ed.; The International Library of Environmental, Agricultural and Food Ethics; Springer International Publishing: Cham, Switzerland, 2020; pp. 137–165, ISBN 978-3-030-61214-6. [CrossRef]
2. Wang, J.; Kim, H.-S. Visualizing the Landscape of Home IoT Research: A Bibliometric Analysis Using VOSviewer. *Sensors* **2023**, *23*, 3086. [CrossRef]
3. Bahari, M.; Arpaci, I.; Der, O.; Akkoyun, F.; Ercetin, A. Driving Agricultural Transformation: Unraveling Key Factors Shaping IoT Adoption in Smart Farming with Empirical Insights. *Sustainability* **2024**, *16*, 2129. [CrossRef]
4. Rodier, D.J.; Zeeman, M.G. Ecological Risk Assessment. In *Basic Environmental Toxicology*; Cockerham, L.G., Shane, B.S., Eds.; CRC Press: Boca Raton, FL, USA, 2019; pp. 581–604, ISBN 978-1-315-13809-1. [CrossRef]
5. Latif, U.; Dickert, F.L. Biochemical Oxygen Demand (BOD). In *Environmental Analysis by Electrochemical Sensors and Biosensors*; Moretto, L.M., Kalcher, K., Eds.; Nanostructure Science and Technology; Springer: New York, NY, USA, 2015; pp. 729–734, ISBN 978-1-4939-1300-8. [CrossRef]
6. Latif, U.; Dickert, F.L. Chemical Oxygen Demand. In *Environmental Analysis by Electrochemical Sensors and Biosensors*; Moretto, L.M., Kalcher, K., Eds.; Nanostructure Science and Technology; Springer: New York, NY, USA, 2015; pp. 719–728, ISBN 978-1-4939-1300-8. [CrossRef]
7. Wackernagel, M. Measuring Ecological Footprints. In *Measuring Sustainable Production*; OECD Sustainable Development Studies; OECD: Paris, France, 2008; pp. 49–59, ISBN 978-92-64-04412-8. [CrossRef]
8. Sivaranjani, S.; Rakshit, A.; Singh, S. Water Quality Assessment with Water Quality Indices. *Int. J. Bioresour. Sci.* **2015**, *2*, 85. [CrossRef]
9. Horn, S.A.; Dasgupta, P.K. The Air Quality Index (AQI) in Historical and Analytical Perspective a Tutorial Review. *Talanta* **2024**, *267*, 125260. [CrossRef]
10. Safarov, R.Z.; Shomanova, Z.K.; Kopishev, E.E.; Nossenko, Y.G.; Bexeitova, Z.B. Spatial Distribution of PM2.5 and PM10 Pollutants in Residential Area of Pavlodar, Kazakhstan. *News Nat. Acad. Sci. Repub. Kazakhstan Ser. Chem. Technol.* **2023**, *457*, 181–200. [CrossRef]
11. Ferreira, S.L.C.; da Silva, J.B.; dos Santos, I.F.; de Oliveira, O.M.C.; Cerda, V.; Queiroz, A.F.S. Use of Pollution Indices and Ecological Risk in the Assessment of Contamination from Chemical Elements in Soils and Sediments—Practical Aspects. *Trends Environ. Anal. Chem.* **2022**, *35*, e00169. [CrossRef]
12. Rinklebe, J.; Antoniadis, V.; Shaheen, S.M.; Rosche, O.; Altermann, M. Health Risk Assessment of Potentially Toxic Elements in Soils along the Central Elbe River, Germany. *Environ. Int.* **2019**, *126*, 76–88. [CrossRef] [PubMed]
13. Rabee, A.M.; Al-Fatlawy, Y.F.; Abd, A.-A.-H.N.; Nameer, M. Using Pollution Load Index (PLI) and Geoaccumulation Index (I-Geo) for the Assessment of Heavy Metals Pollution in Tigris River Sediment in Baghdad Region. *Nahrain J. Sci.* **2011**, *14*, 108–114. [CrossRef]
14. Angulo, E. The Tomlinson Pollution Load Index Applied to Heavy Metal, 'Mussel-Watch' Data: A Useful Index to Assess Coastal Pollution. *Sci. Total Environ.* **1996**, *187*, 19–56. [CrossRef]
15. Chen, T.-B.; Zheng, Y.-M.; Lei, M.; Huang, Z.-C.; Wu, H.-T.; Chen, H.; Fan, K.-K.; Yu, K.; Wu, X.; Tian, Q.-Z. Assessment of Heavy Metal Pollution in Surface Soils of Urban Parks in Beijing, China. *Chemosphere* **2005**, *60*, 542–551. [CrossRef]
16. Jorfi, S.; Maleki, R.; Jaafarzadeh, N.; Ahmadi, M. Pollution Load Index for Heavy Metals in Mian-Ab Plain Soil, Khuzestan, Iran. *Data Brief* **2017**, *15*, 584–590. [CrossRef]
17. Gorgoglione, A.; Castro, A.; Chreties, C.; Etcheverry, L. Overcoming Data Scarcity in Earth Science. *Data* **2020**, *5*, 5. [CrossRef]
18. Grunwald, S. Disaggregation and Scientific Visualization of Earthscapes Considering Trends and Spatial Dependence Structures. *New J. Phys.* **2008**, *10*, 125011. [CrossRef]

19. James, P.M.A.; Fortin, M.-J. Ecosystems and Spatial Patterns. In *Encyclopedia of Sustainability Science and Technology*; Meyers, R.A., Ed.; Springer: New York, NY, USA, 2012; pp. 3326–3342, ISBN 978-0-387-89469-0. [CrossRef]
20. Phillips, D.L.; Marks, D.G. Spatial Uncertainty Analysis: Propagation of Interpolation Errors in Spatially Distributed Models. *Ecol. Modell.* **1996**, *91*, 213–229. [CrossRef]
21. Igaz, D.; Šinka, K.; Varga, P.; Vrbičanová, G.; Aydın, E.; Tárník, A. The Evaluation of the Accuracy of Interpolation Methods in Crafting Maps of Physical and Hydro-Physical Soil Properties. *Water* **2021**, *13*, 212. [CrossRef]
22. Wong, D.W.S. Interpolation: Inverse-Distance Weighting. In *International Encyclopedia of Geography*; Richardson, D., Castree, N., Goodchild, M.F., Kobayashi, A., Liu, W., Marston, R.A., Eds.; Wiley: Hoboken, NJ, USA, 2017; pp. 1–7, ISBN 978-0-470-65963-2. [CrossRef]
23. Calder, C.A.; Cressie, N. Kriging and Variogram Models. In *International Encyclopedia of Human Geography*; Elsevier: Amsterdam, The Netherlands, 2009; pp. 49–55, ISBN 978-0-08-044910-4. [CrossRef]
24. Cuevas, E.; Luque, A.; Escobar, H. Spline Interpolation. In *Computational Methods with MATLAB®*; Synthesis Lectures on Engineering, Science, and Technology; Springer Nature: Cham, Switzerland, 2024; pp. 151–177, ISBN 978-3-031-40477-1. [CrossRef]
25. Huete-Morales, M.D.; Villar-Rubio, E.; Galán-Valdivieso, F. Spatial Distribution of CO2 Verified Emissions: A Kriging-Based Approach. *Emiss. Control Sci. Technol.* **2021**, *7*, 63–77. [CrossRef]
26. Gilbert, R.O.; Simpson, J.C. Kriging for Estimating Spatial Pattern of Contaminants: Potential and Problems. *Environ. Monit. Assess.* **1985**, *5*, 113–135. [CrossRef] [PubMed]
27. Wang, C.; Zhu, H. Combination of Kriging Methods and Multi-Fractal Analysis for Estimating Spatial Distribution of Geotechnical Parameters. *Bull. Eng. Geol. Environ.* **2016**, *75*, 413–423. [CrossRef]
28. Khadka, D.; Lamichhane, S.; Giri, R.K.; Chalise, B.; Amgain, R.; Joshi, S. Geostatistical Based Soil Fertility Mapping of Horticultural Research Station, Rajikot, Jumla, Nepal. *J. Agric. Nat. Res.* **2020**, *3*, 257–275. [CrossRef]
29. Krivoruchko, K.; Gribov, A. Distance Metrics for Data Interpolation over Large Areas on Earth's Surface. *Spat. Stat.* **2020**, *35*, 100396. [CrossRef]
30. Faisal, M.; Jaelani, L.M. Spatio-Temporal Analysis of Nitrogen Dioxide (NO2) from Sentinel-5P Imageries Using Google Earth Engine Changes during the COVID-19 Social Restriction Policy in Jakarta. *Nat. Hazards Res.* **2023**, *3*, 344–352. [CrossRef]
31. Gribov, A.; Krivoruchko, K. Empirical Bayesian Kriging Implementation and Usage. *Sci. Total Environ.* **2020**, *722*, 137290. [CrossRef]
32. Griffin, T.; Brown, J.; Lowenberg-DeBoer, J. Yield Monitor Data Analysis Protocol: A Primer in the Management and Analysis of Precision Agriculture Data. *SSRN J.* **2007**. [CrossRef]
33. Lee, S.; Wolberg, G.; Shin, S.Y. Scattered Data Interpolation with Multilevel B-Splines. *IEEE Trans. Vis. Comput. Graph.* **1997**, *3*, 228–244. [CrossRef]
34. Conrad, O. Tool Multilevel B-Spline Interpolation/SAGA-GIS Tool Library Documentation (v6.1.0). Available online: https://saga-gis.sourceforge.io/saga_tool_doc/6.1.0/grid_spline_4.html (accessed on 18 February 2024).
35. Hjelle, H. MBA: Multilevel B-Splines Reference Manual. Available online: https://www.sintef.no/globalassets/upload/ikt/9011/geometri/mba/mba_doc/index.html (accessed on 18 February 2024).
36. Hardy, D.J.; Wolff, M.A.; Xia, J.; Schulten, K.; Skeel, R.D. Multilevel Summation with B-Spline Interpolation for Pairwise Interactions in Molecular Dynamics Simulations. *J. Chem. Phys.* **2016**, *144*, 114112. [CrossRef] [PubMed]
37. Maximov, G.A.; Larichev, V.A.; Lesonen, D.N.; Derov, A.V. Geospline: Mathematical Model of 3D Complex Geological Medium. In Proceedings of the SPE Arctic and Extreme Environments Technical Conference and Exhibition, Moscow, Russia, 15–17 October 2013; SPE: Moscow, Russia, 2013; p. SPE-166834-MS. [CrossRef]
38. Li, M.; Shi, W.; Liu, S.; Fu, S.; Fei, Y.; Zhou, L.; Li, Y. Fast and Universal Single Molecule Localization Using Multi-Dimensional Point Spread Functions. *bioRxiv* **2023**. [CrossRef]
39. Cunningham, P.; Delany, S.J. K-Nearest Neighbour Classifiers—A Tutorial. *ACM Comput. Surv.* **2022**, *54*, 1–25. [CrossRef]
40. Syriopoulos, P.K.; Kalampalikis, N.G.; Kotsiantis, S.B.; Vrahatis, M.N. kNN Classification: A Review. *Ann. Math. Artif. Intell.* **2023**. [CrossRef]
41. Tabatabaei, M.; Kimiaefar, R.; Hajian, A.; Akbari, A. Robust Outlier Detection in Geo-Spatial Data Based on LOLIMOT and KNN Search. *Earth Sci. Inform.* **2021**, *14*, 1065–1072. [CrossRef]
42. Ahmed, M.-S.; N'diaye, M.; Attouch, M.K.; Dabo-Niange, S. K-Nearest Neighbors Prediction and Classification for Spatial Data. *J. Spat. Econom.* **2023**, *4*, 12. [CrossRef]
43. Ver Hoef, J.M.; Temesgen, H. A Comparison of the Spatial Linear Model to Nearest Neighbor (k-NN) Methods for Forestry Applications. *PLoS ONE* **2013**, *8*, e59129. [CrossRef] [PubMed]
44. Hechenbichler, K.; Schliep, K. *Weighted K-Nearest-Neighbor Techniques and Ordinal Classification*; LMU: Munich, Germany, 2004; Paper 399. [CrossRef]
45. Mladenova, T. A Feature-Weighted Rule for the K-Nearest Neighbor. In Proceedings of the 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 21–23 October 2021; pp. 493–497. [CrossRef]
46. Sahin, E.K. Comparative Analysis of Gradient Boosting Algorithms for Landslide Susceptibility Mapping. *Geocarto Int.* **2022**, *37*, 2441–2465. [CrossRef]

47. Zhang, Z.; Wu, C.; Qu, S.; Chen, X. An Explainable Artificial Intelligence Approach for Financial Distress Prediction. *Inform. Process Manag.* **2022**, *59*, 102988. [CrossRef]
48. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for Big Data: An Interdisciplinary Review. *J. Big Data* **2020**, *7*, 94. [CrossRef] [PubMed]
49. Thomas, J.; Mayr, A.; Bischl, B.; Schmid, M.; Smith, A.; Hofner, B. Gradient Boosting for Distributional Regression: Faster Tuning and Improved Variable Selection via Noncyclical Updates. *Stat. Comput.* **2018**, *28*, 673–687. [CrossRef]
50. Biau, G.; Cadre, B.; Rouvière, L. Accelerated Gradient Boosting. *Mach. Learn.* **2019**, *108*, 971–992. [CrossRef]
51. Bailly, J.S.; Arnaud, M.; Puech, C. Boosting: A Classification Method for Remote Sensing. *Int. J. Remote Sens.* **2007**, *28*, 1687–1710. [CrossRef]
52. Taşpınar, S.; Doğan, O.; Bera, A.K. GMM Gradient Tests for Spatial Dynamic Panel Data Models. *Reg. Sci. Urban Econ.* **2017**, *65*, 65–88. [CrossRef]
53. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
54. Taud, H.; Mas, J.F. Multilayer Perceptron (MLP). In *Geomatic Approaches for Modeling Land Change Scenarios*; Camacho Olmedo, M.T., Paegelow, M., Mas, J.-F., Escobar, F., Eds.; Lecture Notes in Geoinformation and Cartography; Springer International Publishing: Cham, Switzerland, 2018; pp. 451–455, ISBN 978-3-319-60800-6. [CrossRef]
55. Yalaltdinova, A.; Baranovskaya, N.; Rikhvanov, L.; Matveenko, I. Geochemical Peculiarities of Black Poplar Leaves (*Populus nigra* L.) in the Sites with Heavy Metals Intensive Fallouts. Available online: https://ui.adsabs.harvard.edu/abs/2013EGUGA..15..258Y (accessed on 5 October 2023).
56. Mimura, N.T.; Alibekova, A.; Abisheva, A.K.; Aldungarova, M. Study of Geotechnical Conditions of the City of Pavlodar. In *Smart Geotechnics for Smart Societies*; CRC Press: Boca Raton, FL, USA, 2023; pp. 505–510, ISBN 978-1-00-329912-7.
57. Alimbaev, T.; Beksultanova, C.; Mazhitova, Z.; Choybekova, G.; Zhunushalieva, G.; Tentigul Kyzy, N. The Beginning of Virgin Lands Development in Pavlodar Region (in 1954). *E3S Web Conf.* **2023**, *371*, 06017. [CrossRef]
58. Kakabayev, A.; Yermekova, A.; Baikenova, G.; Abdurahmanov, I.; Baituk, G. Technogenic Impact Assessment on the Environment of Pavlodar Region Using GIS Technologies. *E3S Web Conf.* **2023**, *386*, 06001. [CrossRef]
59. Guney, M.; Akimzhanova, Z.; Kumisbek, A.; Beisova, K.; Kismelyeva, S.; Satayeva, A.; Inglezakis, V.; Karaca, F. Mercury (Hg) Contaminated Sites in Kazakhstan: Review of Current Cases and Site Remediation Responses. *Int. J. Environ. Res. Public Health* **2020**, *17*, 8936. [CrossRef] [PubMed]
60. Kismelyeva, S.; Khalikhan, R.; Torezhan, A.; Kumisbek, A.; Akimzhanova, Z.; Karaca, F.; Guney, M. Potential Human Exposure to Mercury (Hg) in a Chlor-Alkali Plant Impacted Zone: Risk Characterization Using Updated Site Assessment Data. *Sustainability* **2021**, *13*, 13816. [CrossRef]
61. Woodruff, S.; Dack, S. Analysis of Risk from Mercury Contamination at the Khimprom Plant in Kazakhstan. *Land Contam. Reclam.* **2004**, *12*, 213–218. [CrossRef]
62. Yessenbayev, D.; Khamidullina, Z.; Tarzhanova, D.; Orazova, G.; Zhakupova, T.; Kassenova, D.; Bilyalova, Z.; Igissinova, G.; Sayakov, U.; Dzhumabayeva, F.; et al. Epidemiology of Lung Cancer in Kazakhstan: Trends and Geographic Distribution. *Asian Pac. J. Cancer Prev.* **2023**, *24*, 1521–1532. [CrossRef] [PubMed]
63. Safarov, R.; Berdenov, Z.; Urlibay, R.; Nossenko, Y.; Shomanova, Z.; Bexeitova, Z.; Kulak, A.; Varga, I.; Balog, A.; Domjánné, R.N.; et al. Spatial Distribution of Elements, Environmental Effects, and Economic Potential of Waste from the Aksu Ferroalloy Plant [Kazakhstan]. *PLoS ONE* **2023**, *18*, e0283251. [CrossRef]
64. Zhakhina, G.; Zhalmagambetov, B.; Gusmanov, A.; Sakko, Y.; Yerdessov, S.; Matmusaeva, E.; Imanova, A.; Crape, B.; Sarria-Santamera, A.; Gaipov, A. Incidence and Mortality Rates of Strokes in Kazakhstan in 2014–2019. *Sci. Rep.* **2022**, *12*, 16041. [CrossRef] [PubMed]
65. Mukataeva, Z.; Dinmukhamedova, A.; Kabieva, S.; Baidalinova, B.; Khamzina, S.; Zekenova, L.; Aizman, R. Comparative Characteristics of Developing Morphofunctional Features of Schoolchildren from Different Climatic and Geographical Regions. *J. Pediatr. Endocrinol. Metab.* **2023**, *36*, 158–166. [CrossRef] [PubMed]
66. Guney, M.; Akimzhanova, Z.; Kumisbek, A.; Kismelyeva, S.; Guney, A.; Karaca, F.; Inglezakis, V. Assessment of Distribution of Potentially Toxic Elements in Different Environmental Media Impacted by a Former Chlor-Alkali Plant. *Sustainability* **2021**, *13*, 13829. [CrossRef]
67. Kanibolotskaya, Y.; Listkov, W.; Shmidt, N. Heavy Metals in Soil and Plants (Agropyron Pectiniforme Roem. et Schult.) of the Pavlodar Region (Kazakhstan). In Proceedings of the International Conference on Sustainable Development of Cross-Border Regions, Barnaul, Russia, 19–20 April 2019; IOP Publishing: Bristol, UK, 2019; Volume 395, p. 012037. [CrossRef]
68. Pogodaiklimat.ru Pavlodar Weather Archive. Available online: http://www.pogodaiklimat.ru/weather.php?id=36003 (accessed on 4 February 2024).
69. Esshaimi, M.; Ouazzani, N.; Avila, M.; Perez, G.; Valiente, M.; Mandi, L. Heavy Metal Contamination of Soils and Water Resources Kettara Abandoned Mine. *Am. J. Environ. Sci.* **2012**, *8*, 253–261. [CrossRef]
70. Sánchez-Donoso, R.; García Lorenzo, M.L.; Esbrí, J.M.; García-Noguero, E.M.; Higueras, P.; Crespo, E. Geochemical Characterization and Trace-Element Mobility Assessment for Metallic Mine Reclamation in Soils Affected by Mine Activities in the Iberian Pyrite Belt. *Geosciences* **2021**, *11*, 233. [CrossRef]

71. Wang, Y.; Duan, X.; Wang, L. Spatial Distribution and Source Analysis of Heavy Metals in Soils Influenced by Industrial Enterprise Distribution: Case Study in Jiangsu Province. *Sci. Total Environ.* **2020**, *710*, 134953. [CrossRef] [PubMed]
72. Chan, S.; Reddy, V.; Myers, B.; Thibodeaux, Q.; Brownstone, N.; Liao, W. Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. *Dermatol. Ther.* **2020**, *10*, 365–386. [CrossRef] [PubMed]
73. Taneja, S.; Gupta, C.; Goyal, K.; Gureja, D. An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering. In Proceedings of the 2014 Fourth International Conference on Advanced Computing & Communication Technologies, Rohtak, India, 8–9 February 2014; pp. 325–329. [CrossRef]
74. Li, P.; Gou, J.; Yang, H. The Distance-Weighted K-Nearest Centroid Neighbor Classification. *J. Inf. Hiding Multimed. Signal Process* **2017**, *8*, 611–622. Available online: https://www.semanticscholar.org/paper/The-Distance-Weighted-K-nearest-Centroid-Neighbor-Li-Gou/6055e3c54be1d095029b37cf2bd7586bf9125546#citing-papers (accessed on 1 May 2024).
75. Yan, X. Weighted K-Nearest Neighbor Classification Algorithm Based on Genetic Algorithm. *TELKOMNIKA* **2013**, *11*, 6173–6178. [CrossRef]
76. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Brooklyn, NY, USA, 2018; Volume 31.
77. Benuskova, L.; Kasabov, N. Artificial Neural Networks (ANN). In *Computational Neurogenetic Modeling*; Springer: Boston, MA, USA, 2007; pp. 81–106, ISBN 978-0-387-48353-5. [CrossRef]
78. Almeida, J.S. Predictive Non-Linear Modeling of Complex Data by Artificial Neural Networks. *Curr. Opin. Biotechnol.* **2002**, *13*, 72–76. [CrossRef] [PubMed]
79. Corazza, M.; Fasano, G.; Mason, F. An Artificial Neural Network-Based Technique for On-Line Hotel Booking. *Procedia Econ. Financ.* **2014**, *15*, 45–55. [CrossRef]
80. Dnyaneshwar, K.; Tushar, J.; Shivam, K.; Sagar, T. Analysis of Car Selling Prediction Based On AIML. *Int. J. Innov. Eng. Sci.* **2023**, *8*, 14–17. [CrossRef]
81. Ma, T.; Wang, F.; Cheng, J.; Yu, Y.; Chen, X. A Hybrid Spectral Clustering and Deep Neural Network Ensemble Algorithm for Intrusion Detection in Sensor Networks. *Sensors* **2016**, *16*, 1701. [CrossRef]
82. Cerezo, M.; Arrasmith, A.; Babbush, R.; Benjamin, S.C.; Endo, S.; Fujii, K.; McClean, J.R.; Mitarai, K.; Yuan, X.; Cincio, L.; et al. Variational Quantum Algorithms. *Nat. Rev. Phys.* **2021**, *3*, 625–644. [CrossRef]
83. Qiao, P.; Lei, M.; Yang, S.; Yang, J.; Guo, G.; Zhou, X. Comparing Ordinary Kriging and Inverse Distance Weighting for Soil as Pollution in Beijing. *Environ. Sci. Pollut. Res.* **2018**, *25*, 15597–15608. [CrossRef] [PubMed]
84. Berrocal, V.J.; Guan, Y.; Muyskens, A.; Wang, H.; Reich, B.J.; Mulholland, J.A.; Chang, H.H. A Comparison of Statistical and Machine Learning Methods for Creating National Daily Maps of Ambient PM2.5 Concentration. *Atmos. Environ.* **2020**, *222*, 117130. [CrossRef] [PubMed]
85. Paci, L.; Gelfand, A.E.; Holland, D.M. Spatio-Temporal Modeling for Real-Time Ozone Forecasting. *Spat. Stat.* **2013**, *4*, 79–93. [CrossRef]
86. Vural, A. Evaluation of Soil Geochemistry Data of Canca Area (Gümüşhane, Turkey) by Means of Inverse Distance Weighting (IDW) and Kriging Methods-Preliminary Findings. *Bull. Min. Res. Exp.* **2019**, *158*, 195–216. [CrossRef]
87. Zhang, Y.; Li, M.; Han, S.; Ren, Q.; Shi, J. Intelligent Identification for Rock-Mineral Microscopic Images Using Ensemble Machine Learning Algorithms. *Sensors* **2019**, *19*, 3914. [CrossRef] [PubMed]
88. Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D.C.; Kording, K.P. The Need to Approximate the Use-Case in Clinical Machine Learning. *GigaScience* **2017**, *6*, gix019. [CrossRef]
89. Jin, H.; Montufar, G. Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks. *J. Mach. Learn. Res.* **2023**, *24*, 137. Available online: https://www.webofscience.com/wos/woscc/full-record/WOS:001111702800001 (accessed on 3 May 2024).
90. Vodyanitskii, Y.N. Standards for the Contents of Heavy Metals and Metalloids in Soils. *Eurasian Soil. Sci.* **2012**, *45*, 321–328. [CrossRef]
91. *GN 2.1.7.2041-06*; Hygienic Standards of Russian Federation "GN 2.1.7.2041-06 Maximum Permissible Concentrations (MPC) of Chemicals in the Soil". Goskomsanepidnadzor of Russia: Moscow, Russia, 2006.
92. adilet.zan.kz Regulatory Legal Act of the Republic of Kazakhstan No. 11755 "Hygienic Standards for the Safety of the Environment". Available online: https://adilet.zan.kz/rus/docs/V2100022595 (accessed on 6 August 2023).
93. Mineev, V.G. *Practical Course in Agrochemistry*, 2nd ed.; M.V. Lomonosov Moscow State University: Moscow, Russia, 2001; p. 689, ISBN 5-211-04265-4.
94. Kazantsev, I.V.; Matveyeva, T.B. Contents of heavy metals in the soil cover in the conditions of technogenesis. *Samara J. Sci.* **2016**, *5*, 34–37. (In Russian) [CrossRef]
95. Vodyanitskii, Y.N. Chromium and Arsenic in Contaminated Soils (Review of Publications). *Eurasian Soil. Sci.* **2009**, *42*, 507–515. [CrossRef]
96. Semendyayeva, N.V.; Dobrotvorskaya, N.I.; Morozova, A.A. The Elemental Composition of Soils of Saline Agrolandscapes and Sanitary-Hygienic Conditions of the Southern Part of the Prichanovskaya Depression. *Open Access J. Environ. Soil Sci.* **2019**, *2*, 166–174. [CrossRef]

97. Timofeeva, T.A.; Chebotar, V.K.; Demidov, D.V.; Gaidukova, S.E.; Yakovleva, I.V.; Kamionskaya, A.M. Effects of Apatite Concentrate in Combination with Phosphate-Solubilizing Microorganisms on the Yield of Ryegrass Cultivar Izorskiy. *Agronomy* **2023**, *13*, 1568. [CrossRef]
98. Noskova, L.N. Justification of maximum permissible concentration of molybdenum in soil. *Hyg. Sanit.* **1988**, *1*, 69–70. Available online: https://cyberleninka.ru/article/n/obosnovanie-predelno-dopustimoy-kontsentratsii-molibdena-v-pochve (accessed on 3 May 2024). (In Russian).
99. Borovskaya, N.E. Materials for justification of maximum permissible concentration of cobalt in soil. *Hyg. Sanit.* **1986**, *5*, 70–71. Available online: https://cyberleninka.ru/article/n/materialy-k-obosnovaniyu-predelno-dopustimoy-kontsentratsii-kobalta-v-pochve (accessed on 3 May 2024). (In Russian)
100. Li, Q.; Pin, M.; Zhongyang, G. Research on the Spatial Distribution of Rural Settlements Based on GIS. In *Computer Science and Applications: Proceedings of the 2014 Asia-Pacific Conference on Computer Science and Applications (CSAC 2014), Shanghai, China, 27–28 December 2014*; CRC Press: Shanghai, China, 2014; pp. 111–116, ISBN 978-1-138-02811-1.
101. Chidunchi, I.; Kulikov, M.; Safarov, R.; Kopishev, E. Extraction of Platinum Group Metals from Catalytic Converters. *Heliyon* **2024**, *10*, e25283. [CrossRef]
102. Roy, S.; Gupta, S.K.; Prakash, J.; Habib, G.; Kumar, P. A Global Perspective of the Current State of Heavy Metal Contamination in Road Dust. *Environ. Sci. Pollut. Res.* **2022**, *29*, 33230–33251. [CrossRef]
103. Li, K.; Wang, F. Global Hotspots and Trends in Interactions of Microplastics and Heavy Metals: A Bibliometric Analysis and Literature Review. *Environ. Sci. Pollut. Res.* **2023**, *30*, 93309–93322. [CrossRef]
104. Sadyrova, G.; Tanybayeva, A.; Bazarbaeva, T.; Mukanova, G.; Jamilova, S.; Nurmakhanova, A. Analysis of the Ecological State of Urban Green Spaces in the Medeu District of Almaty. *Bull. L.N. Gumilyov Eurasian Natl. Univ. Chem. Geogr. Ecol. Ser.* **2023**, *145*, 83–92. [CrossRef]
105. Wang, J.; Christopher, S.A.; Nair, U.S.; Reid, J.S.; Prins, E.M.; Szykman, J.; Hand, J.L. Mesoscale Modeling of Central American Smoke Transport to the United States: 1. "Top-down" Assessment of Emission Strength and Diurnal Variation Impacts. *J. Geophys. Res.* **2006**, *111*, 2005JD006416. [CrossRef]
106. Meteoblue.com Modeling Historical Climate and Weather Data for Pavlodar. Available online: https://www.meteoblue.com/en/weather/historyclimate/climatemodelled/pavlodar_kazakhstan_1520240 (accessed on 20 February 2024).
107. Ahad, N.; Sin Yin, T.; Othman, A.; Yaacob, C. Sensitivity of Normality Tests to Non-Normal Data. *Sains Malays.* **2011**, *40*, 637–641. Available online: https://www.ukm.my/jsm/english_journals/vol40num6_2011/vol40num6_2011pg637-641.html (accessed on 7 May 2024).
108. Liu, X.S. A Probabilistic Explanation of Pearson's Correlation. *Teach. Stat.* **2019**, *41*, 115–117. [CrossRef]