The Neural Network Assisted Land Use Regression

Jan Bitta 1,2, Vladislav Svozilík 1,2,* and Aneta Svozilíková Krakovská 3,4

1 Laboratory of Information Technologies, Joint Institute for Nuclear Research, Moscow Region, 141980 Dubna, Russia; jan.bitta@vsb.cz
2 Faculty of Materials Science and Technology, VSB—Technical University of Ostrava, 70800 Ostrava-Poruba, Czech Republic
3 Faculty of Mining and Geology, VSB—Technical University of Ostrava, 70800 Ostrava-Poruba, Czech Republic; krakovska@jinr.ru
4 Frank Laboratory of Neutron Physics, Joint Institute for Nuclear Research, Moscow Region, 141980 Dubna, Russia
* Correspondence: vladislav.svozilik@vsb.cz; Tel.: +7-926-676-89-10

Abstract: Land Use Regression (LUR) is one of the air quality assessment modelling techniques. Its advantages lie mainly in a much simpler mathematical apparatus, quicker and simpler calculations, and a possibility to incorporate more factors affecting pollutant concentration than standard dispersion models. The goal of the study was to perform the LUR model in the Polish-Czech-Slovakian Tritia region, to test two sets of pollution data input factors, i.e., factors based on emission data and pollution dispersion model results, to test regression via neural networks and compare it with standard linear regression. Both input datasets, emission data and pollution dispersion model results, provided a similar quality of results in the case when standard linear regression was used, the $R^2$ of the models was 0.639 and 0.652. Neural network regression provided a significantly higher quality of the models, their $R^2$ was 0.937 and 0.938 for the factors based on emission data and pollution dispersion model results respectively.

Keywords: land; use; regression; model; air; pollution; modelling; artificial; neural network

1. Introduction
1.1. Air Pollution

Air pollution is an undesirable state of the environment caused by the emission of pollutants from various pollution sources to the air. A pollutant is defined as any substance that negatively affects human health, ecosystems or properties [1,2]. The overall imbalance in the environment has increased over the years. Sources of pollution can be classified as natural and anthropogenic. Concentrations emitted by natural sources appear as a natural background and are not influenced by human activities [1]. Anthropogenic pollution is caused by human activities, and its origin is bound to human settlements. Anthropogenic pollution, unlike natural pollution, is influenceable. Common air pollution pollutants are particulate matter (PM$_{10}$ and PM$_{2.5}$), nitrogen oxides, carbon monoxide, sulphur dioxide, benzo[a]pyrene, persistent organic pollutants (POPs), etc., [3]. This article deals with PM$_{10}$. PM$_{10}$ are solid or liquid (aerosol) particles of 10 µm or less in diameter, compound from various organics, sulfates, nitrates, ammoniac salts, soot, mineral particles, metals, bacteria, pollen and water [1].

1.2. Air Pollution Modelling

Air pollution monitoring is a standard tool for the air pollution assessment. In the Czech Republic, it is supposed to be the main tool for air monitoring and research in accordance with Czech Act no. 201/2012 Coll. Air pollution concentrations are regulated in respect to both acute effects of pollution exposure (short-term averages) and the effects of chronic air pollution exposure represented by yearly averages. This approach was
implemented from the EU Legislation Act on Ambient Air Quality and Cleaner Air for Europe [4] and is universal for the EU countries. The given study focuses on the chronic part of the air pollution effect represented by yearly averages.

The disadvantage of pure air pollution monitoring is that measurements can provide information about concentrations only at specific measurement points. For effective air pollution management, it is better to have continuous information about the distribution in the study area. Mathematical air pollution dispersion modelling is a suitable method to acquire the concentration distribution of air pollution in the study area.

Mathematical models utilized in the air quality assessment and management can be classified by the model type as [5]:

- empirical models,
- Gaussian models,
- numerical models,
- physical models.

Physical models are a smaller physical representation of a real situation that they represent and allow studying. Numerical models are a category of models based on general formulas and algorithms of computational fluid dynamics. These model categories are not considered in the study.

Empirical models are based on in situ measurements and observations. They use various data analysis techniques to create a model describing the phenomenon under study. Empirical models are usually much simpler, require less computational power, and are easier to work with. Their biggest disadvantage is that they are site and time specific. Their scope is limited by the data they are based on. A difference in data location or a time period results in different models constructed. Land Use Regression (LUR) models belong to the empirical model category.

1.3. Gaussian Models

Gaussian dispersion modelling is based on the assumption of continuous leakage of pollution from a pollution source and subsequent dispersion of pollutants in a constant homogeneous wind speed field without spatial limiting conditions. The dispersion of pollutants occurs in the wind direction due to convection and in the direction perpendicular to it due to diffusion, which is caused by atmospheric turbulence and expressed statistically using the (Gaussian) normal distribution [6]. Spatial limiting conditions, such as the effect of terrain, are included in the calculation using coefficients [7]. Input data of Gaussian models are terrain data, meteorological data, the characterization of pollution sources, and a mesh of reference points. Input data characterize an average situation during the modelled period [8].

Gaussian models work with a reasonable degree of air pollution dispersion process abstraction. Therefore, it is possible to describe the relationship between the concentrations at the reference point and the source using an easily enumerable mathematical formula. Simultaneously, abstraction imposes smaller requirements on the input data and time consumption of evaluation. Therefore, Gaussian models are widely used for air pollution modelling in widespread areas. Among the Gaussian air pollution models preferred by the EPA are Industrial Source Complex (ISC) [9], CALPUFF [10], CALINE3 [11], AERMOD [12], and others. Gaussian models preferred by the Czech legislation are SYMOS’97 [13], AE-OLIUS [14] and ATEM [15]. These models provide steady-state results describing air pollution concentrations in the study area and time interval.

1.4. Land Use Regression

Models based on Land Use Regression belong to the category of empirical models. LUR can be treated as an independent air pollution modelling method. This kind of model is based on the principle that concentrations at a specific location depend on the characteristics of the surrounding environment, especially the characteristics that affect the intensity of emissions and rate of dispersion and deposition. Modelling itself is performed
on the basis of the regression model describing the influence of relevant environmental and spatial characteristics, i.e., input factors [16,17]. The model is composed of regression equations that include the relationship between input factors and substance concentrations at monitoring sites. The resulting formula can be used to predict concentrations over the whole area represented by point measurements.

LUR models are not tied to any time or spatial resolution. They are used in a variety of time intervals from short-term campaigns to long-term averages at fixed pollution monitoring sites. The spatial resolution of LUR models can vary from hyper-local urban areas to regional or country level models.

Most LUR models deal with linear regression [18]. Linear regression describes the relationship between the variables x and y, where the values of x (x can be a scalar or a vector representing a set of input values) are assumed to be independent variables and measured without errors. Linear regression consists of two variables connected by a linear function [19]. The dependent random variable represents the value of y under investigation. It is assumed that the variable y is a linear function of x. The coefficient of determination ($R^2$) is used to evaluate the relationship between these variables [20].

LUR itself is especially useful in areas that are difficult to monitor or have an unsatisfactory density of the monitoring station mesh. Therefore, LUR air pollution modelling is used for assessment all over the world, especially in Europe, America, Australia, and Asia.

Van den Bossche et al. [21] used mobile monitoring to gather data at a high spatial resolution in order to build LUR models for the prediction of annual average concentrations of black carbon (BC). The overall prediction was low due to the input uncertainty and lack of predictive variables. The authors highlighted the use of independent data to validate and exclude those data during variable selection in the model building procedure and the importance of using an appropriate cross-validation scheme to estimate the predictive performance of the model. LASSO, the regularized linear model, performed slightly better than the classical supervised approach, and the nonlinear SVM technique did not show significant improvement over the linear model. The generalization of the LUR model to areas where no measurements were made was limited, especially when predicting absolute concentrations.

Lee et al. [22] developed LUR models for particulate matter in the Taipei Metropolis with a high density of roads and strong activities of industry, commerce, and construction. It was possible to achieve $R^2$ values of 95% (PM$_{2.5}$), 96% (PM$_{2.5}$ absorbance), 87% (PM$_{10}$), and 65% (PM$_{coarse}$). Local traffic, construction, residential land use, and industrial sources were identified as the causes of PM$_{2.5}$ pollution. A variable representing the river vicinity decreased PM$_{2.5}$ pollution. PM$_{2.5}$ absorbance levels were boosted by local traffic, commercial and industrial land use. Increased concentrations of PM$_{coarse}$ were caused by elevated motorways.

LUR input data have a spatial character. Geographic information systems are a suitable tool for processing and management, especially when using remote sensing data. Hsu et al. [23] used GIS and remote sensing to develop ten regression models for the PM$_{2.5}$ bound compound concentration based on measurements of a six-year period. The regression models included NH$_4^+$, SO$_2^-$, NO$_3^-$, OC, EC, Ba, Mn, Cu, Zn, and Sb. The authors managed to explain the variance ($R^2$) of the LUR models in the range between 0.60 and 0.92. In the course of the study, they were able to successfully estimate the fine spatial variability of PM$_{2.5}$ and its compounds in Taiwan. Traffic distribution, industrial areas, greenness, and culture-specific PM$_{2.5}$ sources, such as temples, were used as inputs. The main variables determined by the LUR model that affect PM$_{2.5}$-bound concentrations are traffic, industrial areas, and greenness.

The study Wu et al. [24] assesses the influence of surrounding greenness on the concentrations caused by local culture-specific emission sources (Chinese restaurants and temples) within a city of Taipei using LUR modelling. Correlation analysis of the LUR PM$_{2.5}$ model was carried out. A strongly negative correlation ($r$: −0.71 to −0.77) between NDVI was detected. Temples ($r$: 0.52 to 0.66) and Chinese restaurants ($r$: 0.31 to 0.44)
were positively correlated with PM$_{2.5}$ concentrations. The result was confirmed using a cross-validation test with the result $R^2$ of 0.90 and external validation $R^2$ of 0.83, and with the adjusted model $R^2$ of 0.89.

1.5. Artificial Neural Networks

Artificial neural networks (ANNs) are a mathematical abstraction of biological processes that constitute animal brains. An ANN consists of a set of nodes and neurons, which are interconnected. It loosely simulates the neurons of animal brains. Synapses are represented by these connections. Each neuron receives signals via connections, processes them, and signals further via connections. The signal is a real number. In each neuron, the input signals are summed and transformed using some nonlinear function (activation function). The result is an output signal of the neuron. Each input connection has a weight set, which increases or decreases the importance of the signal. The neuron can also have a bias property that represents the sensitivity of the neuron to the signal. The bias value is added to the input signal. Neurons are usually aggregated into layers. The layers can have different activation functions. In each neural network, the signal is propagated from the input receiving layer (input layer) to the output layer, which provides the output signal through one or more layers of neurons. The signal can be one-directional or contain loops McLachlan et al. [25]. Neural networks can be trained via sets of examples consisting of pairs that combine inputs and desired results. From a mathematical point of view, training is an optimization process that optimizes the performance of the neural network. The performance is usually defined as the difference between the desired results and neural network outputs. Independent variables of the optimization are the weights and biases of the neural network. The adjustment of the weights and biases gives increasingly more accurate results. The training of the network is terminated after a sufficient number of adjustments. This process is called supervised learning. There are also types of neural networks that are trained on data not containing the desired results. This process is called unsupervised learning.

Network training is a time-consuming process demanding a lot of computational power. Once the training process is completed, the neural network is able to recall the output value when provided with an input dataset. Mathematically, this is an enumeration of an explicit mathematical formula, a quick and simple operation.

2. Methodology

2.1. Gaussian Models

The Analytical Dispersion Modelling Supercomputer System (ADMOSS) was developed at the VSB—Technical University of Ostrava (VSB-TUO) to perform air pollution modelling in widespread areas [26]. The ADMOSS is based on a combination of geographic information systems (GIS), a mathematical model of air pollution dispersion, and parallel computer clusters. The Modelling methodology implemented in the ADMOSS is the Gaussian model SYMOS’97 recommended by the Czech legislation [27]. The ADMOSS is independent of the modelling methodology and able to run with other models of the same class. Studies (Air Silesia [28], AIR PROGRES CZECHO-SLOVAKIA [29], AIR TRITIA [30]) focusing on air pollution modelling and assessment in the Czech-Polish-Slovak borderland were carried out using the ADMOSS. The study is based on the Gaussian modelling results of the AIR TRITIA project.

2.2. Input Factors of the Land Use Regression

Three groups of factors, i.e., emission factors, results of air pollution dispersion modelling, and land cover factors, were considered in the study. Each modelling result factor was read out as a value at pollution monitoring sites. Factors of land cover and factors representing emission were calculated in a similar manner. Based on the experience from Bitta et al. [31], all factors representing land cover and emission were enumerated as
weighted averages, where the weights were defined by the estimated probability of the wind direction.

The whole area of interest was divided into 46 areas, the division was performed with respect to the terrain configuration. Each area had its own unique meteorological dataset. Buffer zones around the monitoring stations were divided into eight slices representing eight wind directions. The factors were calculated as wind direction probability weighted averages of the factors calculated in the slices. The outlines of division polygons and wind direction probability graphs are visualized in Figure 1.

Figure 1. Wind direction probability distribution, Basemap: OpenStreetMap (CC BY-SA 4.0).
All factors considered in the study are listed in Table 1. Tests which buffer perimeters are the most representative for the LUR were performed in the study in Section 4.1.

Table 1. Considered regression factors.

| Factor Group         | Factor               | Code | Unit    |
|----------------------|----------------------|------|---------|
| Emission factors     | Industrial sources   | IN   | t/y     |
|                      | Car traffic          | CAR  | t/y     |
|                      | Domestic heating     | DH   | t/y     |
| Land cover factors   | Grass covered        | GRS  | %       |
|                      | Built-in             | BLD  | %       |
|                      | Forested             | FRS  | %       |
|                      | Open soil            | SOIL | %       |
|                      | Water bodies         | WTR  | %       |
| Gaussian model results| Pollution from industrial sources | MOD\textsubscript{IN} | µg/m\textsuperscript{3} |
|                      | Pollution from domestic heating | MOD\textsubscript{DH} | µg/m\textsuperscript{3} |
|                      | Pollution from car traffic | MOD\textsubscript{CAR} | µg/m\textsuperscript{3} |

2.3. Neural Network-Based Regression

Neural networks with supervised learning can be utilized as a universal regression technique, an alternative to standard statistical methods. In the case of the LUR model, the set of input factors represents input data, and the pollutant concentration is the desired output. The neural network is trained to provide the desired results from a given list of input factor values.

When neural networks are used for such a task, there is a severe risk of overfitting. Overfitting is the product of analysis that corresponds too closely or exactly to a particular set of data and, therefore, may fail to fit additional data or reliably predict future observations (Figure 2).

![Figure 2. Under- and overfitting of the prediction model.](image)

We limited the maximal number of input factors to five and used the five-fold cross-validation technique to assess the suitability of the current dataset selection and neural network configuration to avoid overfitting. \(k\)-fold cross-validation is the technique that splits the initial data sample into \(k\) of equally sized subsamples. One subsample is kept for
model testing and the remaining $k-1$ subsamples are training data. The training process is repeated $k$-times, each data subsample is once used as test data. The performance of the neural network is represented by the average performance of $k$ trained neural networks [25].

For the sake of analysis, we used multilayer perceptron (MLP) neural networks. An MLP is a class of feedforward ANNs. It is a one-directional neural network consisting of three layers of neurons—the input layer, the hidden layer, and the output layer [32]. In the experiments, we used MLPs with the hidden layer, in which the number of neurons ranged from 1 to 30, and there were two possible activation functions—the logistical function and hyperbolic tangent.

To determine the best neural network configuration for each group of the selected factors, we tested all possible neural network configurations with five different random five-fold cross-validations. That resulted in the need to train 1500 neural networks and compare their performance for each combination of the hidden layer size and activation function. The performance parameter was the $R^2$ estimate based on average $R^2$ of cross-validation neural networks. When the optimal neural network configuration was determined, the final network with those parameters was trained on the whole dataset. The $R^2$ performance value of this neural network represents the quality of prediction based on the selected input factors.

A neural network-based LUR model was developed to estimate spatial and temporal variability of nitrogen dioxide ($NO_2$). At first, the standard LUR model was elaborated to identify significant variables. Secondly, the deep neural network algorithm was applied on the LUR results to fit the model for predicting concentrations. Lautenschlager et al. [33] focus on the development of the OpenLUR platform. This platform consists of the LUR modelling technique combined with machine learning and open datasets.

The study performed by Alam and McNabola [34], extending the typical LUR approach with ANNs. At first, the average daily PM$_{10}$ concentrations of Vienna and Dublin were delivered using the concept of multiple linear regression (MLR) modelling. Secondly, an ANN was used to manage the input variable nonlinearity. The best result of $R^2 = 66\%$ for Vienna and $51\%$ for Dublin was reached owing to the ANN.

3. Data Sources

2015 data were used in the study. The study area is the area of interest of the Air Tritia project [35]. The Air Tritia project focuses on air quality modelling and assessment. Emission and pollution monitoring data were collected and published in this project. The Tritia region consists of two Polish voivodships (Silesian, Opole), the Moravian-Silesian region in the Czech Republic and the Žilina region in Slovakia. Figure 3 demonstrates the position and size of the study area.

The level of air pollution was determined by vast hard coal deposits of the Upper Silesian Basin, which has been mined since the 18th century. The presence of coal enabled the growth of industries by using coal as a source of energy or as feedstock (coal mining, coal processing, steel production, industrial chemistry), downstream industries (e.g., machine industry) and coal heat and electricity production (utility, C&I and domestic scale). The combination of industrial production services, required by densely populated surrounding urban settlements, and the unfavorable basin-like terrain configuration are the cause of severe air pollution. The EEA marks the Tritia region with a population of approximately 7.5 million inhabitants as one of the most air polluted regions of the EU (Figure 4). The most problematic pollutants are particulates (PM$_x$), PAH (polyaromatic hydrocarbons), and heavy metals (Hg, Cd, As) [36].
Figure 3. Tritia region and pollution monitoring sites, Basemap: OpenStreetMap (CC BY-SA 4.0).

Figure 4. Annual average PM$_{10}$ concentrations in the EU, 2015 (European Environment Agency (EEA)) [37].
3.1. Air Pollution and Meteorological Data

Pollution monitoring data consist of yearly averages of PM$_{10}$ (Figure 3) in 2015, collected from all pollution monitoring sites in the Tritia region. The values and site locations were obtained from yearbooks of the Czech Hydrometeorological Institute [38], the Slovakian Hydrometeorological Institute [38] and Inspectorates for Environmental Protection of the Silesian and Opole Voivodships [39,40]. There were 47 air pollution monitoring stations measuring the PM$_{10}$ concentration in the study area. Yearly average values are presented in Table A1.

Meteorological data were obtained from the Air Tritia project dataset [41].

3.2. Emission Data

Emission data were obtained from the emission database provided by the Air Tritia project. The data were divided by the country of origin (Czech–Polish-Slovak) and type of pollution source (industrial, domestic heating, car traffic). Industrial sources were represented as point sources, and the input data contained following parameters: position, average emission flow, height, exhaust diameter, speed, volume flow, and the temperature of exhaust gases. Domestic heating was represented by area sources, which were squares of a 200-m size containing a position, average emission flow, and height. Car traffic sources were modelled as linear sources with a length $\leq 50$ m containing a line description, average emission flow, and height. Brief emission statistics are presented in the following table (Table 2) and emission squares (Figure 5) [30].

Table 2. Emission of PM$_{10}$ in the area of interest.

| Country | Pollution Sources   | No. of Sources | Emission [t/y] |
|---------|---------------------|----------------|----------------|
| Czechia | Industrial sources  | 4728           | 1325           |
|         | Domestic heating    | 52,669         | 1225           |
|         | Car traffic         | 59,651         | 514            |
| Poland  | Industrial sources  | 17,485         | 9962           |
|         | Domestic heating    | 131,741        | 20,603         |
|         | Car traffic         | 403,451        | 1965           |
| Slovakia| Industrial sources  | 1366           | 233            |
|         | Domestic heating    | 33,576         | 2686           |
|         | Car traffic         | 34,009         | 301            |
| Total   |                     | 738,676        | 38,814         |

3.3. Gaussian Model Results

The results of the Gaussian model were obtained from the Air Tritia project. The long-term SYMOS’97 model was used in the project. Meteorological conditions were standardized by the wind speed (low, medium, strong), wind direction (eight directions + calm), and atmospheric stability (Bubník–Koldovský classification, five classes) in the long-term model. Each combination of the meteorological parameters was calculated separately, and the annual average concentration was calculated as a mean of those values weighted by the probability of occurrence of such weather conditions. The emission data entered modelling in the form described in the section above [13,42].
3.4. Land Use Data

Land use data were represented by data from vector topographical datasets, namely, ZABAGED (CZ), BDOT (PL), ZBGIS (SK), which are all available at a 1:10,000 map scale CUZK [43], GUGiK [44] and GKU [45]. There were five kinds of land cover used for analysis, i.e., built-up areas, forested areas, grass-covered areas, water bodies, and open soil agricultural areas.
4. Experiments and Results

The modelling experiment was structured according to the following flowchart (Figure 6).

![Flowchart of the modelling experiment.](image)

The first step of the modelling experiment was to calculate possible input factors based on emission data, land cover data, and Gaussian model results. The number of possible input factors for LUR needed to be reduced. For this reason, the second step of the experiment was factor preselection. When a proper set of possibly eligible factors was selected, LUR modelling followed.

Four sets of LUR models were constructed. Standard LUR models using linear regression were constructed for each combination of five or less input factors. The first set of input factors consisted of emission factors and land cover factors, and the second set of input factors comprised Gaussian model factors and land cover factors. Neural network-assisted LUR models were also developed for the same two groups of input factors mentioned above. Those models replaced the linear regression of LUR models with nonlinear regression provided by the neural network.

The best models were selected from each of the four calculated model groups, which represented the best available result of the combination of the input factor group and regression technique. The $R^2$ value of the models was chosen as the quality parameter for the result of each model.

All calculations and data analysis were performed in the Python 3.4 programming language. The pandas module was used for table data handling, the statmodels module was applied for basic statistical analysis, the matplotlib module was used for graph generation, and sklearn was applied for neural network analysis. All the modules above are parts of the Anaconda distribution of the Python language Anaconda [46], which was utilized for calculations. arcpy is the API of the ArcGIS Pro 2.6 software, which enables spatial analysis and map generation in the Python environment.

The modelling experiment entailed a large number of mathematical computations, which were significantly accelerated due to parallel computing on the CESNET [47] Metacentrum and the “Govorun” supercomputer JINR [48], where hundreds of neural networks could be trained simultaneously.

4.1. Factor Preselection

It was necessary to determine the best buffer zone perimeter for each emission and land cover factor. Analysis based on Spearman correlation coefficients was used. Factor values were calculated for distance perimeters ranging from 200 to 10,000 m with a 200 m step. Spearman’s correlation between each factor and the 2015 yearly average PM$_{10}$ concentration at the monitoring sites was calculated (Figure 7). The best perimeters used in further analysis were the distances at which Spearman’s correlation had its local or global maximum.
Table 3 lists the selected factors and the best perimeters. There are also factors representing the model results included. The factors were read out from the model results at locations of the monitoring stations.

### Table 3. Selected factors.

| Factor Group          | Factor                  | Code | Unit   | Perimeter(s) [m] |
|-----------------------|-------------------------|------|--------|------------------|
| Emission factors      | Industrial sources      | IN   | t/y    | 2000/8000        |
|                       | Car traffic             | CAR  | t/y    | 1000/10,000      |
|                       | Domestic heating        | DH   | t/y    | 1000/5000        |
| Land cover factors    | Grass covered           | GRS  | %      | 5000             |
|                       | Built-in                | BLD  | %      | 400/8000         |
|                       | Forested                | FRS  | %      | 400              |
|                       | Open soil               | SOIL | %      | 5000             |
|                       | Water bodies            | WTR  | %      | 1000             |
| Gaussian model results| Pollution from industrial sources | MOD\(_{IN}\) | µg/m\(^3\) | - |
|                       | Pollution from domestic heating | MOD\(_{DH}\) | µg/m\(^3\) | - |
|                       | Pollution from car traffic | MOD\(_{CAR}\) | µg/m\(^3\) | - |

### 4.2. Linear Regression-Based Land Use Regression

Linear regression LUR models were developed using three different input factor groups, namely, emission factors, land cover factors (Table 3), and Gaussian model result factors. All the following models were limited to a maximum of five input factors to avoid overfitting during linear regression calculations on the dataset containing 47 monitoring sites. At first, models containing any subset of emission and land cover factors were tested (Figure 8). There were 1585 linear models tested. The quality of the models was determined by the coefficient of determination of the model (R\(^2\)).
Figure 8. Performance comparison of linear models with emission + land cover factors.

The best model found was a model containing the following factors: industrial sources in a 2000 m distance, car traffic in a 1000 m distance, built-in land in an 8000 m distance, forested area in a 400 m distance and grass covered land in a 5000 m distance. This model (Figure 9) has $R^2 = 0.639$ and the formula:

$$PM_{10} = 32.6237 + 0.0212 \times IN_{2000} + 0.4823 \times CAR_{1000} + 0.2294 \times BLD_{8000} - 0.1133 \times GRS_{5000} - 0.0965 \times FRS_{400}$$

Figure 9. Observed PM$_{10}$ concentrations vs. best model predictions.

Secondly, the linear model was selected among linear models that were constructed using Gaussian model results for the industrial source ($MOD_{IN}$), domestic heating ($MOD_{DH}$),
and car traffic (MOD\textsubscript{CAR}) and land cover factors. 465 models containing up to six input factors were constructed. The quality of the models was also determined by their coefficient of determination ($R^2$) (Figure 10).

![Figure 10](image)

**Figure 10.** Performance comparison of linear models with Gaussian model results + land cover factors.

The best model found was a model containing the following factors: built-in areas in an 8000 m distance, forested area in a 400 m distance, grass covered land in a 5000 m distance, and all three Gaussian model results (Figure 11). This model has $R^2 = 0.652$ and the formula:

$$PM_{10} = 28.7064 + 0.2457 \times BLD_{8000} - 0.0870 \times FRS_{400} + 0.0459 \times SOIL_{5000} + 2.9565 \times MOD\textsubscript{IN} + 2.6991 \times MOD\textsubscript{CAR}$$

![Figure 11](image)

**Figure 11.** Observed PM\textsubscript{10} concentrations vs. best model predictions.
4.3. Neural Network-Assisted Land Use Regression

Two experiments with neural networks were conducted. All possible combinations of emission and land cover factors were tested, with the number of selected factors being \( \leq 5 \) in the first experiment. A total of 1585 combinations were tested. In the following graph, each point represents the performance of neural network models. The most efficient models for each number of the input parameters are highlighted (Figure 12).

The best performing model was a neural network model containing the \( BLD_{400} \), \( BLD_{8000} \), \( GRS_{5000} \), \( IN_{2000} \) and \( CAR_{1000} \) factors. The \( R^2 \) of the model was 0.937 (Figure 13).

![Figure 12. Performance comparison of neural network models with emission + land cover factors.](image)

![Figure 13. Observed PM\textsubscript{10} concentrations vs. best model predictions.](image)
All possible input factor combinations with Gaussian model results and land cover factors were tested in the second experiment. The number of input factors was limited to five again. 381 combinations were tested. In the following graph, each point represents the performance of neural network models. The most efficient models for each number of the input parameters are highlighted (Figure 14).

**Figure 14.** Performance comparison of neural network models with Gaussian model results + land cover factors.

The best performing model was a neural network model containing the $BLD_{400}$, $FRS_{400}$, $GRS_{5000}$, $MOD_{IN}$ and $MOD_{CAR}$ factors. The $R^2$ of the model was 0.938 (Figure 15).

**Figure 15.** Observed PM$_{10}$ concentrations vs. best model predictions.
It was required to train 1501 neural networks for each tested combination of the input factors. It means that in total, approximately 2.95 million neural networks needed to be trained.

5. Discussion

Two sets of input data were selected for the LUR model construction; the first dataset consisted of emission factors and land cover factors, while the second dataset comprised Gaussian model results and land cover factors. The LUR model construction was performed via two techniques, the first one was standard linear regression, the second one was the multilayer perceptron neural network. A set of LUR models was created for different selections of input factors ($\leq 5$) in each of the four experiments, determined by the input dataset and model construction technique. The best model in each experiment was defined by the $R^2$ score of the result (Table 4).

Table 4. $R^2$ values of the best models.

| Model Performance Measure | Emission + Land Cover Factors | Gaussian Model + Land Cover Factors |
|---------------------------|-------------------------------|-----------------------------------|
| Linear regression model   | 0.639                         | 0.652                             |
| Neural network model      | 0.937                         | 0.938                             |

A comparison of different possible model performance measures of all four best models is presented in Table 5.

Table 5. Model performance measures.

| Model Performance Measure | Emission + Land Cover Factors | Emission + Land Cover Factors | Gaussian Model + Land Cover Factors | Gaussian Model + Land Cover Factors |
|---------------------------|-------------------------------|-------------------------------|-----------------------------------|-----------------------------------|
|                           | Linear Regression Model       | Neural Network Model          | Linear Regression Model           | Neural Network Model              |
| Correlation               | 0.807                         | 0.968                         | 0.824                             | 0.968                             |
| $R^2$                     | 0.639                         | 0.937                         | 0.652                             | 0.938                             |
| Normalized mean bias      | $2.5 \times 10^{-16}$         | 0.00145                       | $4 \times 10^{-16}$               | $-0.00172$                        |
| Normalized root mean square error | 0.1143                     | 0.048322                      | 0.1098                             | 0.048258                          |

Nearly all model performance measures indicate a better performance of neural network models. The only exception is the normalized mean bias, which shows the low overestimation or underestimation of neural network models, while linear regression models show almost zero values. This is the natural behaviour of linear models.

All the best models provide significantly improved results over the pure Gaussian model. The $R^2$ of the Gaussian model was 0.485, and its results compared to the measurements are shown in Figure 16. There are several monitoring stations where the Gaussian model significantly overestimated or underestimated the measured values. The values were overestimated at the monitoring sites in forested areas and underestimated in rural areas with a high percentage of open soil near the monitoring stations. This clearly shows the limitations of pure Gaussian models, which do not include the influence of land cover factors.
Both linear regression-based LUR models provided a similar performance, the $R^2$ of the models was 0.639 for the emission factor linear model and 0.652 for the linear model based on the Gaussian model results. The results of these LUR models are similar to those of the other LUR models, which are 0.65 in Bitta et al. [31], 0.58 in Liu et al. [49] or 0.66–0.76 in Masiol et al. [50]. A more detail comparison of these studies and the model of the study is not possible, since each model used a different set of variables, different time scales, etc. Gaussian model results were more accurate at the monitoring sites representing urban or industrial sites. The LUR model, which included land cover factors, provided better results in rural and natural environments.

Both neural network-based LUR models performed significantly better than their linear regression-based counterparts. Both models showed similar results with $R^2$ of 0.937 and 0.938. A better performance of neural network-based models is the expected result, since nonlinear regression provided by neural networks can better reflect a generally nonlinear nature of pollution dispersion.

One can also compare the performance of linear regression and neural network models with the same input data. This comparison is provided in the following two tables (Tables 6 and 7).

Table 6. $R^2$ values of the best linear regression models and neural network models with the same inputs.

|                      | Emission + Land Cover Factors | Gaussian Model + land Cover Factors |
|----------------------|------------------------------|-------------------------------------|
| Linear regression    | 0.639                        | 0.652                               |
| Neural network model | 0.854                        | 0.810                               |
Table 7. $R^2$ values of the best neural network models and linear regression models with the same inputs.

| Emission + Land Cover Factors | Gaussian Model + Land Cover Factors |
|-------------------------------|-----------------------------------|
| Linear regression model       | 0.516                             |
| Neural network model          | 0.937                             |

It is clear that neural network models provide significantly more accurate results than linear regression models. Nonlinear regression provided by neural network models performed best with input factors that had slightly worse results in the linear regression model than the best possible linear regression model inputs. Conversely, inputs that provided the best linear regression model results slightly underperformed when they were used as inputs of the neural network model. There are factors that better reflect the nonlinear behaviour of air pollution dispersion.

Such high $R^2$ values can have several reasons. The selected method of the model construction may provide high quality results or be distorted due to overfitting (even though we tried to avoid it) or the monitoring station mesh dominated by urban background stations (42 out of 47 stations). Higher variability among monitoring stations would definitively bring higher confidence in the results.

For each LUR model construction technique, both input factor data sets provided a similar quality of results. The question arises whether Gaussian dispersion modelling is a necessary step. The Gaussian dispersion model in this case was highly computationally intensive. It took several processor years of processing, which could have been avoided. This hypothesis also requires further investigation on different datasets.

LUR models are empirical models that are constructed as statistical models based on the data provided. Each LUR model represents only a specific time and space interval, which is reflected in input data. The model formula and coefficients implicitly represent the effects of general phenomena that affect air pollution. Meteorological factors, which demonstrably influence pollution dispersion, such as precipitation, temperature, thermal stability, etc., are typical examples of factors that are implicitly reflected in LUR model coefficients. It also means that a different time period or different area of interest produces its own set of LUR model coefficients. The goal of LUR modelling should not be to provide one universally applicable model. The goal is to provide an algorithm that generates the LUR model fitting the best selected time period and area of interest.

6. Conclusions

We managed to significantly improve the performance of the standard linear regression-based LUR model by replacing the linear regression algorithm with the multilayer perceptron neural network. This is an innovative approach that has the potential of future improvement of LUR modelling. The modelling technique described in the study requires further investigation and development. The goal should be a standardized modelling algorithm and standardized input datasets that provide credible results for decision makers.

Author Contributions: Conceptualization, J.B. and V.S.; Data curation, J.B. and V.S.; Formal analysis, J.B., V.S. and A.S.K.; Methodology, J.B.; Supervision, J.B.; Validation, V.S. and A.S.K.; Writing—original draft, J.B. and V.S.; Writing—review & editing, V.S. and A.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the project solved in cooperation between the Czech Republic and JINR (3+3 project) “Development of novel techniques of air pollution distribution estimation”, 2020, No. 05-6-1118-2014/2023, main investigators: V. Svozilik, J. Bitta.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: Modelling data are available at https://bit.ly/3cAuMfi (accessed on 31 March 2021), for further info please contact authors.

Acknowledgments: Access to the computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum under the programme “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042) is greatly appreciated. The research was supported in part by the “Govorun” computational cluster provided by the Laboratory of Information Technologies of the Joint Institute for Nuclear Research in Dubna, Moscow Region, Russia.

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

Abbreviations
The following abbreviations are used in this manuscript:

| Abbreviation | Description |
|--------------|-------------|
| ADMOSS | Analytical Dispersion Modelling Supercomputer System |
| ANN | Artificial Neural Network |
| API | Application Programming Interface |
| CESNET | Czech Education and Scientific NETwork |
| CUZK | Český urad zeměměřický a katastrální/ Czech State Administration of Land Surveying and Cadastre |
| EEA | European Environment Agency |
| EU | European Union |
| GKiK | Geodetic and Cartographic Institute Bratislava |
| GUGiK | Main Office of Geodesy and Cartography/Główny Urząd Geodezji i Kartografii |
| JINR | Joint Institute for Nuclear Research |
| LUR | Land Use Regression |
| m | Meter |
| MLP | Multi-Layer Perceptron |
| PAH | Polycyclic Aromatic Hydrocarbon |
| PM10 | Particulate Matter with a diameter of 10 microns |
| PM2.5 | Particulate Matter with a diameter of 2.5 microns |
| PMcoarse | Particulate Matter Coarse |
| POPs | Persistent Organic Pollutants |
| R2 | Coefficient of determination |
| SYMOS | Modelling System for Stationary Sources |
| VSB-TUO | VSB—Technical University of Ostrava |
| ZABAGED | Feature Catalog ZABAGED® |
| ZBGIS | Základná báza údajov pre geografický informačný systém/ Basic database for geographic information systems |

Appendix A

Table A1. Annual concentration of PM$_{10}$ in 2015 in the AIR TRITIA project area.

| Monitoring Station | Latitude  | Longitude | Concentration |
|---------------------|-----------|-----------|---------------|
| Bílý Kříž           | 49.50260917 | 18.53856083 | 15.5          |
| Čeladná              | 49.55921556  | 18.34835444 | 22.9          |
| Frýdek-Místek       | 49.67179111  | 18.35107028 | 29.9          |
| Návsí u Jablunkova   | 49.59419444  | 18.74396389 | 28.1          |
| Trinec-Kanada       | 49.67237861  | 18.64303778 | 26.3          |
| Trinec-Kosmos       | 49.66811361  | 18.67779917 | 29.8          |
| Český Těšín         | 49.74895861  | 18.60972583 | 36.5          |
| Havířov             | 49.79097775  | 18.40683556 | 36.2          |
Table A1. Cont.

| Monitoring Station | Latitude      | Longitude     | Concentration |
|---------------------|---------------|---------------|---------------|
| Karviná             | 49.86379611   | 18.5514525    | 36.6          |
| Orlová              | 49.87566      | 18.43360722   | 36.1          |
| Šunychl             | 49.92756667   | 18.36184694   | 34.0          |
| Věřňovice           | 49.92467889   | 18.4228725    | 41.6          |
| Červená hora        | 49.77714167   | 17.54194639   | 18.7          |
| Opava-Kateřinky     | 49.94498833   | 17.90953056   | 27.2          |
| Ostrava-Fifejdy     | 49.8391875    | 18.26368917   | 33.9          |
| Ostrava-Mariánské Hory | 49.82485972 | 18.26365472   | 31.5          |
| Ostrava-Poruba/CHMÚ | 49.82529444   | 18.159275     | 29.1          |
| Ostrava-Prívoz      | 49.85625833   | 18.26974111   | 36.3          |
| Ostrava-Radvanice OZO | 49.81853861 | 18.34034361   | 33.7          |
| Ostrava-Radvanice ZÚ | 49.80705639 | 18.33913806   | 42.2          |
| Ostrava-Zábřeh      | 49.79603944   | 18.24718083   | 31.8          |
| Ostrava-Českobratrská | 49.83985     | 18.289975     | 33.7          |
| Bielsko-Biała, ul. Kossak-Szczechuckie | 49.813464   | 19.027318     | 35.5          |
| Cieszyn, ul. Mickiewicza | 49.738136  | 18.639069     | 33.0          |
| Częstochowa, ul. Baczynskiego | 50.836389  | 19.130111     | 31.9          |
| Dąbrowa Górnicza, ul. Tysiąclecia | 50.329111  | 19.231222     | 41.4          |
| Godów, ul. Gliniki  | 49.921875     | 18.471278     | 44.3          |
| Katowice, ul. Kossutha | 50.264611  | 18.975028     | 39.0          |
| Katowice, ul. Plebiscytowa | 50.246795 | 19.019469     | 46.5          |
| Knurów, ul. Jedności Narodowej | 50.233167 | 18.655722     | 43.8          |
| Lublinec, ul. Piaskowa | 50.658357   | 18.69622      | 38.2          |
| Rybnik, ul. Borki   | 50.11118      | 18.516139     | 47.0          |
| Tarnowskie Góry, ul. Litewska | 50.444736 | 18.829639     | 37.8          |
| Zabrze, ul. M. Curie-Sklodowskiej | 50.3165  | 18.772375     | 43.9          |
| Zawiercie, ul. M. Skłodowskiej-Curie | 50.47954  | 19.43301      | 38.8          |
| Żory, Os. Gen. Władysława Sikorskiego | 50.028681 | 18.691222     | 41.3          |
| Żywiec, ul. Kopernika | 49.671602   | 19.234446     | 43.8          |
| Martin, Jesenského | 49.06694444  | 18.92194444   | 26.0          |
| Rużomberok, Riadok  | 49.07916667  | 19.3025       | 31.0          |
| Žilina, Obežná      | 49.21194444  | 18.77111111   | 30.0          |
| Głubczyce, ul. Kochanowskiego | 50.209042 | 17.816453     | 33.0          |
| Kędzierzyn-Koźle, ul. B. Śmiiałego | 50.349608 | 18.236575     | 31.0          |
| Kluczbork, ul. Mickiewicza | 50.972181 | 18.207575     | 33.0          |
| Nysa, ul. Rodziewiczný | 50.458989  | 17.331906     | 34.0          |
| Opole, os.im.Armii Krajowej | 50.676856 | 17.950278     | 31.0          |
| Opole, ul. Minorytów | 50.666961   | 17.922797     | 33.0          |
| Zdzieszowice, os. Piastów | 50.423533 | 18.120739     | 36.0          |
32. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*; Springer: New York, NY, USA, 2009.

33. Lautenschläger, F.; Becker, M.; Kobs, K.; Steininger, M.; Davidson, P.; Krause, A.; Hotho, A. OpenLUR: Off-the-shelf air pollution modeling with open features and machine learning. *Atmos. Environ.* 2020, 233, 117535. [CrossRef]

34. Alam, M.S.; McNabola, A. Exploring the modeling of spatiotemporal variations in ambient air pollution within the land use regression framework: Estimation of PM10 concentrations on a daily basis. *J. Air Waste Manag. Assoc.* 2015, 65, 628–640. [CrossRef] [PubMed]

35. Air Tritia. Available online: https://www.interreg-central.eu/Content.Node/AIR-TRITIA.html (accessed on 1 November 2020).

36. European Environment Agency. *Air Quality in Europe: 2015 Report*; Publications Office of the European Union: Luxembourg, 2015. [CrossRef]

37. European Environment Agency. Annual Mean PM10 Concentrations in 2015. Available online: https://www.eea.europa.eu/data-and-maps/figures/annual-mean-pm10-concentrations-in (accessed on 1 November 2020).

38. Macoun, J.e.a. *Air Pollution and Atmospheric Deposition in Data, the Czech Republic, 2015*; Czech Hydrometeorological Institute: Prague, Czech Republic, 2016.

39. Sadowski, T.E.A. *Stan Środowiska w Województwie Śląskim w 2015 Roku*; PIONS Katowice: Katowice, Poland, 2016.

40. Gaworski, K.E.A. * Wyniki Pomiarów Uzyskanych w 2015 Roku na Stacjach Jakości Powietrza w Województwie Opolskim*; PIOS Opole: Opole, Poland, 2016.

41. Tritia, A. Air Quality Management System. Available online: https://aqms.vsb.cz/ (accessed on 27 February 2021).

42. Modeling System for Stationary Resources (SYMOS’97). Czech Hydrometeorological Institute: Prague. Available online: https://www.idea-envi.cz/symos-97-en.html (accessed on 27 February 2021).

43. CUZK. ZABAGED. Available online: https://geoportal.cuzk.cz/(S(dlueg0h03ufdmtu4thyb1a5f))/default.aspx?lng=EN&mode=TextMeta&text=dSady_zabaged&side=zabaged&menu=24 (accessed on 1 November 2020).

44. GUGiK. Baza Danych Obiektów Topograficznych (BDOT). Available online: https://www.geoportal.gov.pl/dane/baza-danych-obiektow-topograficznych-bdot (accessed on 1 November 2020).

45. GKU. ZBGIS. Available online: https://www.geoportal.sk/sk/zbgis/ (accessed on 1 November 2020).

46. Anaconda. Anaconda. Available online: https://www.anaconda.com (accessed on 1 November 2020).

47. CESNET. Metacentrum NGI. Available online: https://www.metacentrum.cz/en/index.html (accessed on 1 November 2020).

48. JINR. Govorun Supercomputer. Available online: http://hlit.jinr.ru/en/about_govorun_eng/ (accessed on 1 November 2020).

49. Liu, W. Land use regression models coupled with meteorology to model spatial and temporal variability of NO2 and PM10 in Changsha, China. *Atmos. Environ.* 2015, 116, 272–280. [CrossRef] [PubMed]

50. Masiol, M.; Ziková, N.; Chalupa, D.; Rich, D.; Ferro, A.; Hopke, P. Hourly land-use regression models based on low-cost PM monitor data. *Environ. Res.* 2018, 167, 7–14. [CrossRef] [PubMed]