Abstract. Significant spatial variability of the accumulation of pollutants in soils can make problems in the determination of the borders defining a zone where pollution, according to the applied legal requirements, is excessive. Particular difficulty is caused by a short-distance variability, disturbing the regularity in a spatial distribution of pollution around the source of emission. The paper presents an alternative, compared to traditional interpolation methods, algorithms based on the optimization and the application non-linear neural networks called mixture density network MDN and feature space mapping network FSM. The benefit from the application of this approach is more information referring to the distribution of pollution. This approach allows the estimation of the local variance of the accumulation of pollutants and approximate local distribution. This allows greater extent of taking into account the uncertainty connected with the spatial variability of soil pollution.

Legal regulations referring to the assessment of the state of soils formally differentiate the classes of areas fulfilling sanitary requirements and those not fulfilling such requirements. Counting soils into proper classes means sampling often there is also a need to demark a proper borderline. Relatively great natural variability in soil properties and morphological factors, lithological factors and land use distort the regularity of the spatial distribution of pollutants emitted from point sources, complicating a simple-looking problem. Some observations indicate that a short-distance variability in the concentration of pollutants in soils, especially in heavy polluted areas, is very large, which can make the task of selection more difficult. In practice, in case of the need for distinguishing two classes of the areas: ‘polluted’ and ‘clean’ ones it is possible to apply one of the interpolation methods, based on a regular grid or a network of dispersed points or to use any model including an inevitable uncertainty connected with the spatial variability in the concentration of pollutants in soils. The investigated problem is significant also in the context of more and more widely applied digital technique of
soil charting, that cannot be based only on the transformation of traditional analogue materials into a digital form, but should also extend the possibilities of variability analyses referring to soil characteristics with the assessment of reliability of these estimations. In this paper, based on the observations in the environs of the industrial plant that makes a significant source of soil pollution with chromium compounds, the analysis of the problem of determination the variability and the border of soil contamination was made and alternative techniques of such assessment were presented.

STUDY AREA

The Chemical Plant in Alwernia was built on the turn of 1923/24, after WW-II, it was enlarged and its production profile was extended. Nowadays the Chemical Plant ‘Alwernia’ Co. specializes in three production profiles: chromium compounds (green chromium oxide, sodium dichromate, potassium dichromate, alkaline chromium sulfate, chromium acid anhydride), phosphorus compounds including phosphates for animal feeds, sulfates: sodium sulfate and magnesium sulfate. The plant is situated in the municipality Alwernia, in a narrow north-south oriented valley of the Regulka River. The hills surrounding the plant from the east are steep and the difference in levels reaches 150 m. Mild slopes are in western direction. From the north the valley is closed by the Tenczyn Ridge (Garb Tenczyński). In the first series of field studies soils were sampled on the surface of about 2.5 km² in the knots of a regular grid of squares 200 x 200 m, in the nearest vicinity of the emission source. The situation of the knot points of the grid was established by a GPS receiver with the accuracy up to about 5 m. Samples were taken from the layer of 0-30 cm. In 5 points sampling was deepened to 150 cm, the every 20 cm. In three areas the observation grid was made denser - to 20 m x 20 m. The increased density was to examine the problem of the variability of the concentration of pollution in short distances. In the samples the content of total chromium as well as many other soil properties were determined with ASA method. The total sample number in this series was 279. In the second series of studies soil samples were collected in a bigger radius around the industrial plant, to observe the area beyond the range of strong pollution and make conclusions referring to the pollution range. The depth of sampling was in this case more differentiated and included selected soil profiles up to 150 cm. The total sample number in this series was about 750.

The distribution of the content of Cr in the collected samples is characterized by the mean value of about 885 mg kg⁻¹, median is 390 mg kg⁻¹ and standard deviation reaches 2122, lower quartile makes the value of 26 mg kg⁻¹ and the upper quartile makes 890 mg kg⁻¹, so the concentration is slightly above the average. A characteristic feature of a spatial distribution of Cr is the translocation of points fulfilling and not fulfilling sanitary state of soils.
THE ANALYSIS OF THE SPATIAL DISTRIBUTION OF CR CONTENT IN SOILS, IN THE REGION OF THE CHEMICAL PLANT ‘ALWERNIA’

The assessment of the spatial distribution of soil pollution in the surroundings of an emission source requires field studies and laboratory studies providing certain number of the observations of the pollution concentration $t$ in soils, with their respective vectors $x$, covering: horizontal coordinates of points and the depth of taking a soil sample. Based on the set of spatial information, a model allowing the estimation of pollution concentration in each point (also the one that was not sampled) should be constructed. Traditionally, to the imaging of the distribution of pollution in soils one of interpolation algorithms is applied. In this case three-dimensional interpolation is useful. Software packs allowing the interpolation of the variable in a three-dimensional space and the visualization of its distribution are available (e.g. Voxler by Golden Software, GRASS etc.). A good side of interpolation algorithms is that the obtained iso-surface is close to the values resulting from empirical observations; the disadvantage is including all the irregularities and casual deviations to the model.

The alternative to the interpolation is a regression model combining for example value $t(x)$ with the vector $x$. The obtained estimation $\langle t_i | x \rangle$ refers to the expected value of the concentration of pollution $t_i$ in the point described by the vector of coordinates $x_i$. The choice between two kinds of modelling depends on the purpose of documentation and value of local variance in the observed pollution concentration. For the estimation of the degree of pollution in a defined area is done based on data biased in a different way, there is a problem of the confidence of such estimation. In such conditions, apart from the estimator of central trend of the concentration of pollution in a certain region an estimator of the degree of its credibility is necessary. In other words estimation $p(t_i | x_i)$, of the distribution of conditional probability of a given pollution level would be necessary. This implies that any more complete picture of the environmental risk must include the component defining the trend of distribution (conditional expected value, depending on the situation of the point) and random dispersion, expected in this situation. One of the tools of making such a picture of variability can be an artificial neural network following the model of mixture density - so called mixture density network-(MDN) [1, 8].

**MDN networks**

MDN algorithm, based on the observation making a training set, provides estimators of conditional probability distribution of the observed variable $p(t|x)$ [2]. Unlike other algorithms of regression, the obtained model allows the estimation of conditional distribution of the variable, as superposition of a certain number of normal distributions. MDN model has a local character. The advantage
of the algorithm is the avoidance of the inconvenient necessity of choosing conditional distribution of random variable *a priori*. Usually its disadvantage is a large size of processing structure, which in case of the lack of a sufficient number of training examples, makes problems with the generalization of the description of the phenomenon model.

The network combines Gaussian Mixture Model (and the optimization of its parameters with the application of perceptron) with a hidden layer [12, 13]. Optimization means minimization of error by the manipulation with mean and local participations of model’s variances - in GMM. The combination of neural optimization with GMM the principle of iteration in constructing a model is based on the criterion of maximum likelihood – unlike in neural networks fulfilling regression task. The procedure is an alternative for the regression models, enabling the estimation of the distribution of the deviations from regression in cases where they are asymmetric or multi-modal [2]. Conditional density of probability in GMM is defined by the equation [1]:

$$p(t|x) = \sum_{i=1}^{m} \alpha_i(x) \phi_i(t|x)$$  \hspace{1cm} (1)

where \( m \) is the number of model components. Parameter \( \alpha_i(x) \) is a mixing coefficient. Function \( (t|x) \) represents conditional density of the goal vector \( t \) - this kernel. Kernel function can come to different forms, while the authors of the network MDN prefer Gaussian function:

$$\phi_i(t|x) = \frac{1}{(2\pi)^{\frac{m}{2}} \sigma_i(x)^{\frac{m}{2}}} \exp \left\{ -\frac{1}{2} \frac{||t - \mu_i(x)||^2}{\sigma_i(x)^2} \right\}$$  \hspace{1cm} (2)

where \( \mu_i(x) \) represents the center of this kernel, \( \sigma_i(x) \) are variances. The model allows the estimation of a conditional mean value, conditional distribution depending on the input to the model and conditional variance. This information improves the knowledge of the modeled phenomena, especially when the knowledge referring to the degree of the risk of transgressing some border values is necessary, while the assumption referring to homoscedasticity is not fulfilled. Detailed description of the MDN algorithm, computation the error in parameter space and the gradient of the error function in parameter space contains reports [1, 2, 6]. The program implementation of – among others – this algorithm is made available in the form of macro set called NETLAB, used in the MATLAB environment.

More limited than in the model of regression, but in some cases sufficient information referring to spatial relationship between the pollution of individual
fragments of the area, can also be provided by classification algorithms. In the classification we look for a function of explaining variables, optimising the division of objects into defined homogenous groups, called classes. We are looking for estimation \( k_i x_i \), that is the estimation of the belonging of a given fragment of surface, with the vector of coordinates \( x_i \) to one of \( k \) class \( k_1...k_k \). In case of soil pollution two classes would be interesting: \( k_S \) class - grounds fulfilling cleanness standards and \( k_N \) class - grounds failing to fulfil such standards. Similarly to the problem of regression, one can accept that also in this case the assessment of the reliability of estimation would be desirable. For example, it would be useful to assess \( p(k_i \mid x_i) \) of the conditional probability of subsequent classes. There are numerous possibilities of the construction of a proper model: e.g. the application of the probabilistic neural network (PNN) can be considered. A competitive to PNN algorithm can be Feature Space Mapping (FSM) included in GhostMiner pack. An algorithm of this kind can be useful especially in cases of the occurrence in data of relatively large dislocation in the space of explaining variables (in this case, simply in 3D space of soils) of points belonging to different classes. In such cases classification techniques based on the nearest neighbour can fail, while techniques that more generalize this phenomenon, such as diluted inference such as FSM may be useful.

RESULTS

The reference point for the assessment of adaptation algorithms can be the assessment of the accuracy of spatial projection of the distribution of pollution obtained by the interpolation procedure. It is necessary to apply 3D interpolation, because data on Cr concentration in soils come from the samples of different depths, while our purpose also knows the trend in vertical distribution of pollutants. The applied here module of 3D interpolation, contained in GRASS pack is called \textit{v.vol.rst} and it implements this procedure using Regularized Spline with Tension (RST) algorithm. The basic parameter of the module is the size of tension deciding on the degree of ‘elasticity’ of iso-surfaces separating respective volumes. The module can be put in motion from the option of cross validation that is made by LOO technique. This means multiple (the number of times equals the size of the set treated as base for the interpolation) putting in motion the interpolation procedure, each time one of points is excluded. Based on the obtained interpolation model the approximation error is calculated in the point excluded from interpolation procedure. This rather time-consuming procedure, apart of the assessment of the interpolation algorithm itself, provides a good material to the assessment of local variability of a feature. Table 1 contains some statistics of the distribution of rests obtained in LOO procedure at different values of tension parameter.
The statistics of the rests confirm great irregularity of the distribution of Cr in soils. One should assume that this is a short distance irregularity, characterized by a strong differentiation of the level of pollution accumulation in points situated quite close to one another in all the sections. It is also difficult to indicate the optimal value of tension parameter, because generally closer to the observations course of iso-surfaces (great value of tension, small inter-quartile interval) is accompanied by great extreme values. One should also emphasize that these results prove more the character of spatial variability of the accumulation of pollutants rather than the disadvantages of the interpolation method. Indirectly they strengthen the thesis that less deterministic approach to the problem of the documentation of the assessment of the sanitary state of soils is necessary.

MDN algorithm, based on the training set, adjusts the conditional distribution of the probability of the observed variable in the space of explaining variables. This binds the pollution assessment with the estimation of its uncertainty. Distribution can be characterised by smaller or greater concentration, multi-modality and asymmetry.

As a result of multiple samples it was established that the structure of MDN model consisted of 8 Gauss units and 6 sigmoid units in a hidden layer is optimal in terms of the minimization of its size and the decrease of regression error (minimization of the negative log predictive density). To estimate the expected generalization error the procedure of the cross validation method 10-Folds-Out was applied. The criteria of comparing the errors of the training set and test set included: negative log predictive density, mean-square error and absolute mean

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Tension=80</th>
<th>Tension=40</th>
<th>Tension=10</th>
<th>Tension=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic mean</td>
<td>0.73</td>
<td>0.94</td>
<td>20.5</td>
<td>52.8</td>
</tr>
<tr>
<td>Median</td>
<td>13.6</td>
<td>8.4</td>
<td>10.7</td>
<td>100</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1797</td>
<td>1753</td>
<td>1659</td>
<td>1656</td>
</tr>
<tr>
<td>Minimum</td>
<td>-22966</td>
<td>-22384</td>
<td>-21973</td>
<td>-22076</td>
</tr>
<tr>
<td>Maximum</td>
<td>18241</td>
<td>18226</td>
<td>8629</td>
<td>2964</td>
</tr>
<tr>
<td>Lower quartile</td>
<td>-32</td>
<td>-51</td>
<td>-72</td>
<td>-37</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>191</td>
<td>205</td>
<td>251</td>
<td>506</td>
</tr>
</tbody>
</table>

TABLE 1. THE SET OF SOME STATISTICS OF REST VALUES OBTAINED AS A RESULT OF CROSS VALIDATION (LEAVE-ONE-OUT) PROCEDURE OF 3D INTERPOLATION WITH ALGORITHM v.vol.rst OF GRASS PACK, DATA REFERRING TO SOIL POLLUTION WITH Cr COMPOUNDS IN THE SURROUNDING OF CHEMICAL PLANT ‘ALWERNIA’
Negative Log Predictive Density (NLPD) is calculated according to the formula:

\[ \text{NLPD} = \frac{1}{N} \sum_{i=1}^{N} \log(p(y_i|x_i)) \]  

(3)

The great value of NLPD indicates considerable range in the conditional distribution, which means great degree of uncertainty accompanying the assessment of the expected value of a variable and greater range in the confidence interval. The other criterion to compare the results of the processing of a training set and test set by the model are root mean-square error and absolute mean error.

The data presented here confirm the remarks referring to great spatial variability of Cr concentration in soils. Great differences between actual values and values interpolated in MDN model are observed in the form of a great range of the conditional distribution. In a similar way errors in the values of means confirm a great dispersion of results. In case of NLPD the difference between the mean of the training and test sets is statistically significant, unlike the differences of: root mean-square error and absolute mean error for the training set and test set. MDN presenting expected value of Cr concentration in soil layer 0-30 cm, in the surrounding of the Chemical Plant ‘Alwernia’ is shown in Fig. 1.

Figure 2 presents the estimation of local standard deviation, calculated based on the model. This model allows the estimation of local distribution, which consequently, allows more advanced analysis of spatial differentiation.

Figure 3 presents the obtained from MDN estimated values of the probabilities of a normative level of Cr contents in thirty-centimeter soil layer.

Table 3 illustrates basic statistics calculated from conditional distributions obtained from MDN model consisted of 8 Gauss units and 6 sigmoid units, on the background of the observed values.
Fig. 1. Expected values of the contents of Cr in soil layer 0-30 cm in the surrounding of the Chemical Plant ‘Alwerna’: estimated with MDN. Hypothetic source of emission has co-ordinates (0, 0).

Fig. 2. Estimation of local standard deviation mean values of MDN of Cr content in 0-30 cm layer of soils in the surrounding of Chemical Plant ‘Alwerna’.
Fig. 3. Estimation of the probability of Cr concentration in 0-30 cm layer in soils, below the upper value of the accepted level of concentration, based on MDN, in the surrounding of Chemical Plant ‘Alwernia’.

TABLE 3. THE COMPARISON OF THE STATISTICS OF OBSERVED Cr CONCENTRATIONS IN SAMPLES WITH THE PARAMETERS OF CONDITIONAL DISTRIBUTIONS OBTAINED FROM MDN MODEL

<table>
<thead>
<tr>
<th>Statistics in the set</th>
<th>Cr - observed</th>
<th>Parameters of conditional distributions from MDN model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Mean</td>
<td>872</td>
<td>858</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Maximum</td>
<td>24600</td>
<td>11603</td>
</tr>
<tr>
<td>Lower quartile</td>
<td>26</td>
<td>72</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>890</td>
<td>917</td>
</tr>
<tr>
<td>SD</td>
<td>65</td>
<td>48</td>
</tr>
<tr>
<td>r</td>
<td></td>
<td>0.73</td>
</tr>
</tbody>
</table>
The value of CRPS (Continuous Ranked Probability Score) \([3, 7]\) is the mean value of the index used in the assessment of the modelled statistic distributions, according to the formula:

$$
CRPS = \frac{1}{N} \sum_{i=1}^{N} crps(p_i, y_i),
$$

where \(N\) is the number of the analysed cases, while \(crps(p_i, y_i)\) is calculated according to the formula:

$$
crps(p, t) = \int_{-\infty}^{\infty} (p(y|x) - H(y, t))^2 \, dy,
$$

where \(H(x, t) = \mathbb{1}_{\{x \geq t\}}\) is the Heavside function and takes value 1 if \(x\) is greater than \(t\), in other cases it takes value 0. This index is expressed in the estimated variable, which highly facilitates its interpretation, because it can be directly compared to the observed values.

The review of the presented data allows the following conclusions:

- when it is necessary to choose a single value as a characteristic of the level of Cr concentration in soils, probably the mean value of the model distribution would give the best approximation,
- like in the case of the spatial interpolation, rather great extreme values are observed, signalising relatively big rest values,
- in constructing a conditional distribution with the application of a definite number of Gauss components there is a risk of its multi-modality, which can cause significant interpretation difficulties connected with the estimation referring to the meaning of several modal values,
- quite an important notice can be a significant linear correlation between observed values and CRPS index, which suggests greater dispersion of a conditional distribution at greater level of pollution.

**Classification approach**

Algorithm MDN generates conditional distribution of an expected level of pollution, expressed on the units of its concentration. It is based on the observation data. It can, however be analysed to which degree the modelled conditional distribution allows correct separation between the volume complying to the standard (NORMA) and the one not complying to the standard (NNORMA). Standards of the cleanness of soil define the limit value between both classes as 150 mg/kg. In both normative soil layers the same value is compulsory (apart from industrial areas). For each value \(x_i\) being a vector of coordinates in the accepted system it is possible to calculate value \(p(y < 150|x_i)\). It can then be confronted with...
the fact of belonging the sample \( i \) to one of two differentiated classes of soil cleanness. Like in each case of classification there is a problem of the optimisation of the limit of the discriminative function that is (in this case) common for all the vectors \( x \) the value of conditional probability \( p(y<150|x_i) \), the exceeding of which would mean a reasonable suspicion that the concentration in a concrete point (volume) meets the quality standard, i.e. belongs to the NORMA class. This limit value can be estimated with the analysis of the behaviour of the \( \kappa \) coefficient at different limit values for \( p(y<150|x_i) \). The \( \kappa \) coefficient is widely applied to the assessment of the correctness of the classification; it regards both correct and incorrect indications of the classifier. According to this criterion, in case of the discusses model, according to the expectations the value in within the range \( p(y<150|x_i) = 0.5-0.6 \). The value of the \( \kappa \) coefficient is about 0.76-0.77 (according to common interpretation this means a good quality of classification), while the number of wrong indications is about 0.11%. The system of conditions made that this limit encourages rather ‘conservative’ decisions, because in the group of 111 wrongly classified cases as much as 91 refers to the situation when the algorithm classified the points fulfilling the standards to NNORMA, while in only 20 cases (less than 2% of the whole set) points were wrongly put into NORMA class, although the soil cleanness standards were exceeded.

FSM belongs to constructivist algorithms of neural networks, where beside optimisation parameters of the function of transfer of processing units, also their number is optimised, thus the process of network constructing regards the degree of complexity of the task and the assumed efficiency of identification. The advantage of GostMiner pack is the presence of the cross validation algorithm, allowing the assessment of the degree of the generalization of the phenomenon by the model [5].

The essence of the FSM algorithm is making the size of the model dependent on the degree of complexity of the classification task. In the course of the training, combined with continuous modification of the network structure, the units processing to/from the hidden layer are attached or removed. The training stops at the moment of achieving the indicated by the user degree of identification correctness. Because in very complicated classification tasks the algorithm has the tendency to make very large processing structures, the authors of the commercial pack added the option of the cross validation method 10-Folds to facilitate the estimation of generalization error.

The best of FSM models was generated based on data in 150 training steps, while the hidden layer consisted of 70 Gauss units. Cross validation indicates that this model, with the ability of the identification of training data of on average 95.7% for test data is 92%, with the errors of mean values: 4.3 and 8.5%, respectively. Like in MDN algorithm, FSM network enables the assessment of conditional probability of the distribution of belonging to the defined classes.
CONCLUSIONS

1. The analysis of the model allows stating that apart from considerable spatial variability of Cr concentration in soils in the surroundings of chemical plants, also extremely large variability in the variance of this value on a relatively small surface is characteristic for this phenomenon. This indicates the need of a certain modification of traditional approach to the problem of estimating the situation of the borders between the polluted and clean area, regarding this fact.

2. Particularly characteristic is the asymmetry in the distribution of the concentrations within a small surface, which means the probability of occurring high concentrations in the area of relatively low level of contamination.

3. Subsequent popularization of digital soil cartography [10] increases the requirements concerning the way of the documentation and presentation of their spatial differentiation. In particular, traditional algorithms of transferring information obtained in one point into a certain surrounding require a critical analysis. This especially refers to interpolation algorithms applied to imaging the values of a large variability, observed at relatively short distances. More useful can be regression algorithms, including those allowing full characteristic of the deviations of the values from the mean.

4. Starting from the statement about a problematic usefulness of interpolation techniques in the studies on the spatial distribution of variables on small distances, it is easy to conclude that the algorithms of artificial neural networks would be useful for this purpose. They usually work in non-linear problems and provide additional information on statistical deviations from the regularity of the model.

5. The information concerning the variability of the value, its distribution and variance allows more complete estimation of the risk of transgressing the acceptable limit values. One should notice that this is a considerable progress compared to regression models assuming the uniformity of the distribution of deviations in the whole area of the examined relation.

REFERENCES

PROBLEM OF THE ESTIMATION OF THE INDUSTRIAL SOIL POLLUTION

Znaczną przestrzenno zmienność akumulacji zanieczyszczeń w glebach może sprawiać problemy przy wyznaczaniu zasięgu strefy nadmiernej nimi obciążonej, zgodnie ze stosownymi wymogami prawnymi. Szczególnie utrudnienie sprawia zmienność krótkodystansowa, zakłócająca oczekiwanej regularność rozkładu przestrzennego zanieczyszczeń wokół źródła emisji. W pracy zaprezentowano zastosowanie alternatywnych, wobec tradycyjnych metod interpolacji, algorytmów sieci neuronów MDN i FSM. Korzyścią płynącą z zastosowania tego podejścia jest pogłębienie informacji dotyczących rozkładu zanieczyszczenia. Podejście pozwala na oszacowanie lokalnej wariancji akumulacji zanieczyszczeń oraz przybliżonego lokalnego rozkładu. Pozwala to na pełniejsze uwzględnienie niepewności związanej z przestrzenną zmiennością obciążenia gleb zanieczyszczeniami.