Data-driven modelling of groundwater vulnerability to nitrate pollution in Slovenia

Podatkovno vodeno modeliranje ranljivosti podzemne vode na nitratno onesnaženje v Sloveniji

Jože Uhan[1, *]

[1]Agencija RS za okolje, Vojkova 1b, SI-1000 Ljubljana, Slovenia

*Corresponding author. E-mail: joze.uhan@gov.si

Received: October 4, 2012 **Accepted:** October 19, 2012

Abstract: This paper describes the case study of statistical data-driven models implementation to assess groundwater vulnerability to nitrate pollution of alluvial aquifers in Slovenia. The aim of the research was spatial prediction of the relative probability for increased groundwater nitrate concentration in order to plan the groundwater nitrate reduction measures and optimize the programme for monitoring the effects of these measures. For the selection of possibly optimal statistical model and comparison with the one of point count system methods PCSM, receiver operating characteristic method ROC was used. Results of the probabilistic classifier from the weights-of-evidence model WofE and neuro-fuzzy model NEFCLASS has in the case of groundwater nitrate pollution a significant better average performance than the widespread used SINTACS parametric point count relative rating as groundwater contamination potential.

Izvleček: Članek opisuje študijski primer uporabe statističnih podatkovno vodenih modelov za ocenjevanje ranljivosti podzemne vode na nitratno onesnaženje v aluvialnih vodonosnikih Slovenije. Namen raziskave je bil prostorsko napovedati relativno verjetnost zvišane vsebnosti nitrata v podzemni vodi za potrebe načrtovanja ukrepov za zmanšanje nitratnega onesnaženja in optimiranja programov merilnega nadzora učinkov teh ukrepov. Za izbor optimalnega statističnega modela in primerjavo z rezultati večparametrske metode razvrščanja in tehtanja je bil uporabljen pokazatelj karakteristike delovanja klasifikacijskih metod ROC. Verjetnostni klasifikatorji modela teže evidenc WofE in

Original scientific paper

nevronske mehke logike NEFCLASS izkazujejo v primeru nitratnega onesnaženja podzemne vode značilno boljše povprečne klasifikacijske lastnosti kot sicer zelo razširjena metoda razvrščanja in tehtanja parametrov SINTACS.

Key words: groundwater vulnerability, nitrate, data-driven modelling, Slovenia
Ključne besede: ranljivost podzemne vode, nitrat, podatkovno vodeno modeliranje, Slovenija

Introduction

Groundwater nitrate pollution in Slovenian shallow alluvial aquifers has been a major concern in recent years, and more than a third of the groundwater in these aquifers has poor chemical status according to Water Framework Directive (Directive 2000/60/ES) criteria, most frequently due to a high concentration of nitrate (UHAN et al., 2010). The operative programme of measures requires identification of the potentially vulnerable priority areas within groundwater bodies for cost-effective measures planning. Groundwater vulnerability maps are an important tool of the water management decision-making process. Most of the previous groundwater vulnerability assessments of shallow alluvial aquifers in Slovenia (JANŽA & PRESTOR, 2002; BRAČIČ ŽELEZNIK et al., 2005; MALI & JANŽA, 2005; UHAN et al., 2008) used a variety of parametric point count methods with a relative rating for the potential of groundwater contamination, e.g. the SINTACS index, adapted to conditions in the Mediterranean region (CIVITA, 1990). These methods require validation with field measurements, such as a tracer test, groundwater residence studies or investigation of pollution processes, e.g. denitrification. GOGU & DASSARGUES (2000) identified the integration of results from process-based models in vulnerability mapping techniques as a new research challenge in groundwater vulnerability assessment. Data-driven modelling offer the possibility of analysing the relevant data about a groundwater system, in fact, learning from available data, which incorporates the so far unknown dependencies between a system's inputs and outputs (MITCHELL, 1997; PRICE & SOLOMATINE, 2000).

Materials in methods

Modelling of the groundwater vulnerability to pollution is generally understood as probability modelling or a mathematical representation of a random phenomenon. Most com-

monly used methods of the classification modelling are neural networks and fuzzy logic methods, statistical method of logistic regression and closely related Bayesian approaches to classification. In the Lower Savinja Valley case study, we have used the neuro-fuzzy approach for the data classification NEFCLASS-J (Nauck & Kruse, 1995), and weights-of-evidence method for combining evidence in support of a hypothesis Arc-WofE (Kemp et al., 1999). The results of these two data-driven methods were compared with the results of the SINTACS parametric point count relative rating as groundwater contamination potential (Uhan et al., 2008), coupling with the results of the agricultural nitrate hazard index IPNOA (Pehan, 2008).

The SINTACS scheme of aquifer pollution vulnerability mapping incorporates seven parameters, relevant for the contaminant attenuation and vertical flow capacity (Table 1). The grid square cell structure of the SINTACS input data has been designed in order to use several weight strings in order to satisfactorily describe the effective hydrogeological and impacting situation as set up by the sum of data (Civita & De Maio, 2000). For each grid squares, element normalized SINTACS index was calculated and coupled with the agricultural nitrate hazard index (IPNOA). The IPNOA method integrates two categories of parameters (Table 1): the hazard factors representing farming activities and the control factors which adapt the hazard factors to the characteristics of the site (Padovani & Trevisan, 2002).

Neuro-fuzzy system is an identification method that combines the methods of neural networks and fuzzy logic. The neural networks classify among the »black box« methods, where the model is set solely on the basis of measured data without an insight into the dynamics of the process. Fuzzy logic on the other hand classifies among the »grey box« methods, where the model structure is given as a parameterized mathematical function that is at least partially based on the laws of physics. Both systems have been developed independently, and only later great advantages have been recognised in their joint use, especially for the classificatory purposes (Nauck & Kruse, 1995), as well as in the area of groundwater vulnerability to pollution (Dixon, 2001).

Bayesian classifier uses attribute independence assumption and estimates the conditional probabilities (coefficients in the model) on the basis of counting the cases in a particular class. Although the Bayesian classifier probabilistic model is based on the assumption, which the practice does not support, the empirical evidence shows that this has no major impact on its classificatory accuracy (Domingos & Pazzani,

Table 1. List of used evidential themes in groundwater nitrate pollution vulnerability assessment case study

SINTACS	IPNOA	NEFCLASS	WofE
Deep to the groundwater	Use of fertilizers	Hydrogeological homogeneous units	Hydrogeological homogeneous units
Effective infiltration	Application of livestock and poultry manure	Irrigation and drainage areas	Irrigation and drainage areas
Unsaturated zone attenuation capacity	Food industry wastewater and urban sludge	Development of the river networks	Development of the river networks
Soil/overburden attenuation capacity	Topographic slope	Long-term groundwater recharge	Long-term groundwater recharge
Hydrogeological characteristics of the aquifer	Climatic conditions	Nitrogen load in seepage water	Nitrogen load in seepage water
Coefficient of hydraulic conductivity	Agronomic practices	Groundwater flow velocity in saturated zone	Groundwater flow velocity in saturated zone
Topographic slope			

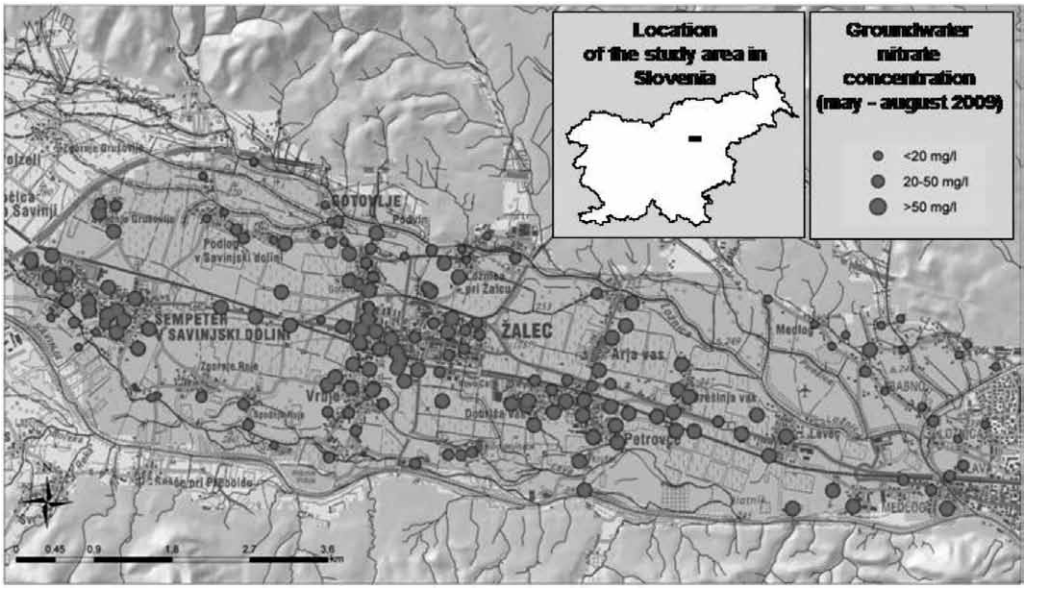

Figure 1. Groundwater nitrate measurements in central part of Lower Savinja Valley, used as a training dataset

1997; Kononenko, 2001). Advantages of this method are in better response for problems with a small number of training points and missing attribute values, and in greater clarity of the resulting model. Mainly due to an easier interpretation of the results, the weights-of-evidence method, which is based on the Bayesian theorem, has been successfully used also for the assessment of groundwater vulnerability to pollution (Arthur et al., 2005; Baker et al., 2006; Masetti et al., 2007; Sorichetta et al., 2008).

In both cases of data-driven modelling of groundwater vulnerability to nitrate pollution, NEFCLASS-J modelling and Arc-WofE modelling, we have used the same evidential themes, including also process-based modelling outputs of groundwater recharge, groundwater flow velocity and nitrate leached from the soil profile (Table 1). In the vulnerability assessment procedure, the central part of the Lower Savinja Valley (30.8 km^2) was discretised with a regular mesh grid of 100 m × 100 m. Randomly chosen 173 groundwater nitrate in-situ measurements have been used as a training points dataset. Monitoring sites have been classified for further analysis into two or three groups on the basis of distribution of groundwater nitrate concentration with 20 mg/l as antopogenic impact concentration or 50 mg/l as EU threshold value.

Results and discussion

Groundwater intrinsic vulnerability assessment of Lower Savinja Valley shallow aquifer using SINTACS parametric method (Uhan et al., 2008) identified two classes with different vulnerability degrees. The first zone with higher vulnerability is characterised mainly by the lower terrace with shallow groundwater, high surface/groundwater interaction and a thin protective soil layer. The second zone with medium vulnerability is characterised mainly by the upper terraces with deeper groundwater and thick soil layer with increased clay component. The most sensitive parameters are depth to the groundwater and effective infiltration action. The results of single-parameter sensitivity analysis enable better understanding of the vulnerability model results, enable consistent evaluation of the analytical result and give a new orientation for further methodological contamination research by using statistical and numerical model results with selected SINTACS groundwater vulnerability parameters. It is pointed out that detailed vulnerability mapping, including analysis of hydrochemical data, especially nitrate concentration in groundwater, linked to the assessment of pressures and impacts, is a very good basis for establishing detailed monitoring programmes and programmes of measurement to achieve the WFD objectives of good groundwater status for groundwater bodies at risk.

The intrinsic nitrate contamination risk from agricultural sources, assessed using the IPNOA methodology, is high for the 87 % of the study area, whereas the 11 % of the southern parts of the area indicate a diffuse and extremely high potential risk. The greatest discrepancies between the estimates of the potential risk of groundwater contamination by nitrates from agricultural activities and the results of groundwater nitrate field measurements have been identified on the northern part of the study area. Here the nitrate levels in groundwater are in many places markedly below the expectations, given the high level of potential risk for groundwater contamination by nitrates from agricultural activities (PEHAN, 2008). These findings have highlighted the need for further study of spatial variability in conditions of nitrogen cycle processes, which affect the reduction processes in groundwater.

In the Lower Savinja Valley groundwater vulnerability model, we have, in light of the results of an extensive sensitivity analysis of the NEFCLASS-J model (DIXON, 2004), used the triangular membership function and three fuzzy sets structure. The model discovered a total of 36 possible learning rules, of which five of the best rules for a particular classificatory range were used for the grid. Additional optimization of the network resulted in the 93.02 % accuracy for the classification of the learning data patterns, 75.58 % accuracy for the validation, and 84.30 % accuracy for the classification of the entire data series into two classes of groundwater nitrate level. When modelling the three-class fuzzy grid (<20 mg/l, 20–50 mg/l, >50 mg/l), the model accuracy was somewhat lowered, yet the classificatory accuracy improved. The model classified all of the 3,079 spatial cells, namely: 978 in the first group (31.76 %), 689 in the second group (22.38 %), and 1412 in the third group (45.86 %) of the spatial cells. The hydrogeological boundary between the middle and the highest terrace markedly stood out at this classification (Figure 2).

WofE modelling technique combines known occurrences of phenomenon (training points) with available spatial data (predictor evidence) in a predictive response (phenomena occurrences conditional probability map). Six evidential themes were applied to generate the response theme with posterior probability values ranging from 0 to 0.312 (Figure 3). The response theme values describe the relative probability that a 100 m × 100 m spatial unit will have a groundwater nitrate concentration higher than the training points threshold values with regard to the prior probability value of 0.057. Based on the definition of the training point, higher posterior probability values correspond with more ground-

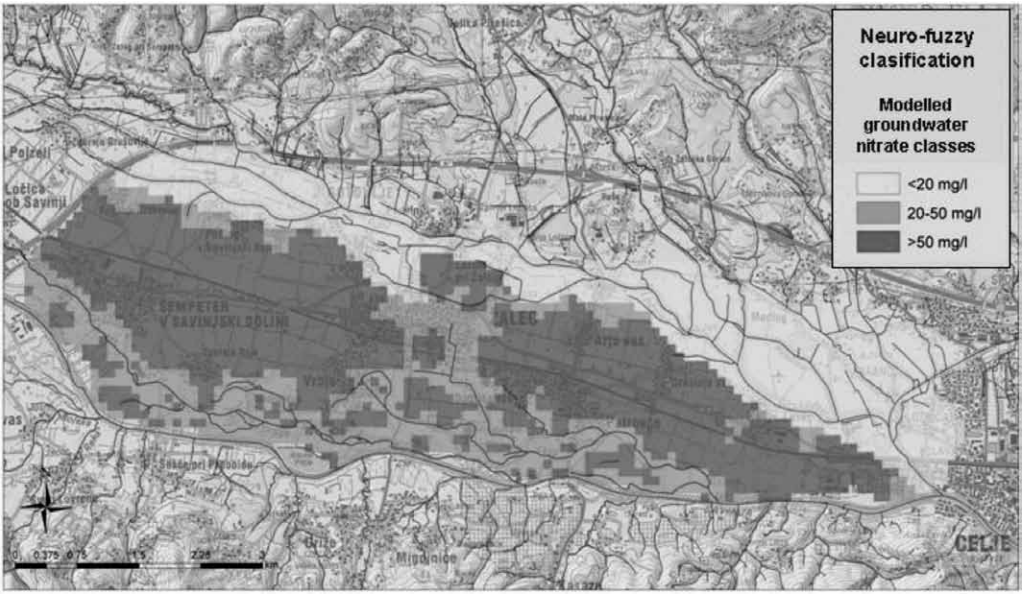

Figure 2. Neuro-fuzzy prediction of groundwater nitrate pollution in central part of Lower Savinja Valley (threshold values: <20 mg/l, 20–50 mg/l and >50 mg/l)

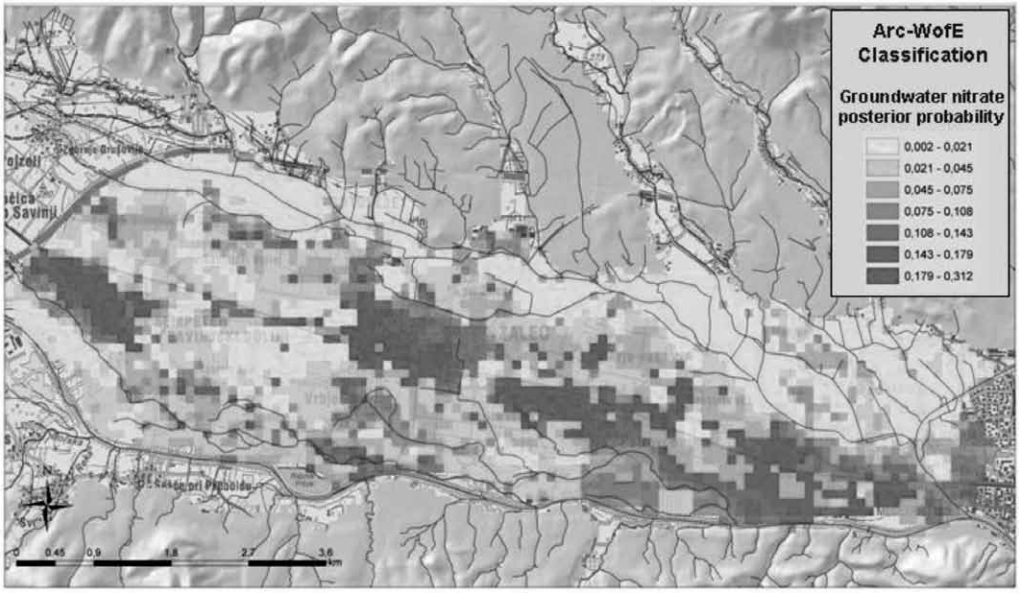

Figure 3. WofE posterior probability prediction of groundwater nitrate pollution in central part of Lower Savinja Valley (threshold value: 20 mg/l; prior probability = 0.057)

Table 2. Statistics of ROC analysis

Model output	Area under the ROC curve (%)	Standard error	95 % confidence interval
SINTACS x IPNOA	75.7	0.041	0.685–0.820
WofE	82.9	0.034	0.756–0.877
NEFCLASS (3 classes)	84.4	0.031	0.779–0.895

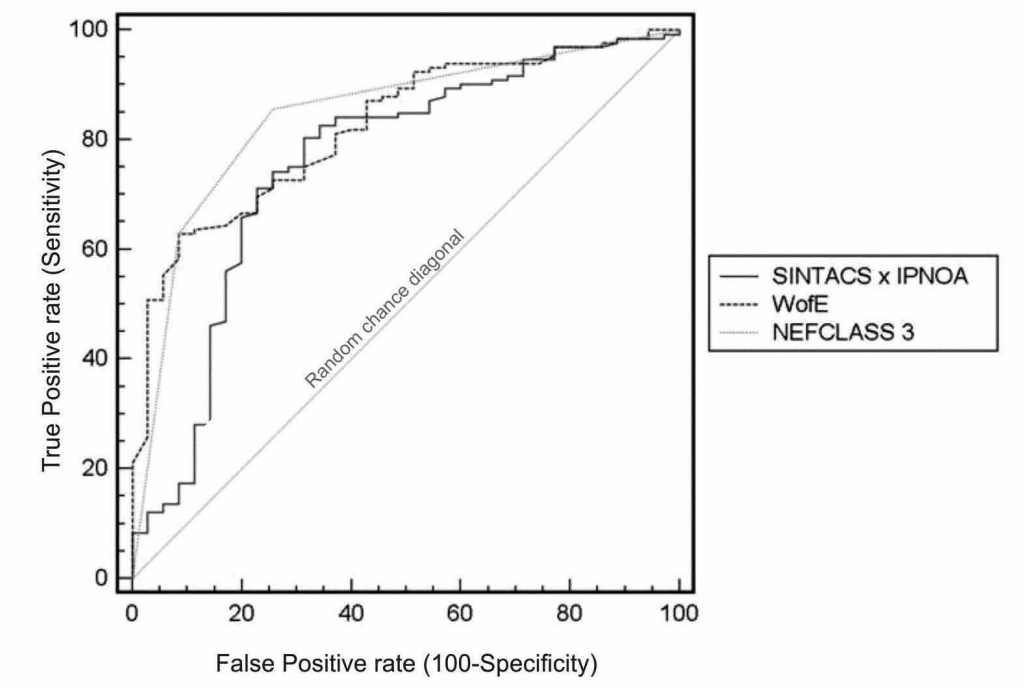

Figure 4. Predictive reliability of different classification schemes in ROC diagram

water vulnerable cells and lower posterior probability values correspond to less vulnerable areas. The highest probability of groundwater nitrate vulnerability zones has been found to be generally in the central part of the study area. According to the calculated confidence value, the most important contribution to the final response theme was assessed for the groundwater flow velocity evidential theme, followed by the groundwater recharge evidential theme. Conditional independence as an important assumption of the WofE model was within the range that generally indicates no dependence amongst evidential themes (BAKER et al., 2007).

The predictive reliability of the applied models has been verified by the receiver operating characteristic analysis (Metz, 1978), which through the sensitivity and specificity assessment provides the area under curve (AUC) in receiver operating characteristic diagram (ROC). Receiver operating characteristic curves were developed in the field of statistical decision theory and assess the value of diagnostic/prediction tests by providing a standard measure of the ability of a test to correctly classify subjects or phenomena. The ROC curve reflects the probability of correct and incorrect positive findings of the phenomenon and can be illustrated in space with the coordinates for sensitivity and specificity. Sensitivity is defined as the probability that the highly vulnerable spatial aquifer cell is correctly classified, whereas specificity is defined as the probability of the correct classification of the moderately vulnerable spatial cell. The rate of false negative value is given by 1-specificity. The discrete three-class classification of the model of neuro-fuzzy network NEFCLASS-J and the linear distribution of the WofE posterior probability of increased groundwater nitrate levels in the studied area of Lower Savinja Valley has been compared also with the results of the SINTACS and IPNOA analysis (Figure 4). The area under ROC curve (AUC) was the lowest for SINTACS x IPNOA prediction model (75.7). According to the ROC analysis, the best results were achieved by the neuro-fuzzy model, within which the highest value of the parameter AUC (84.4) was achieved (Table 2). When comparing it to the weights-of-evidence model through the κ statistical comparison of matching between the measured and the predicted categories (Jenness & Wynne, 2007), the differences were, however, very small.

Conclusion

When comparing the results of classification schemes, the neuro-fuzzy method was proven somewhat more effective for predicting the groundwater nitrate concentration and thereby predicting the groundwater vulnerability in Lower Savinja Valley. However, the discrete character of this model result has to be emphasised, whereas the weights-of-evidence method enables the assessment of the probability of groundwater nitrate pollution while not sacrificing much the quality of the results. The assessment of the probability of groundwater nitrate pollution can be of great service for mapping the groundwater vulnerability to nitrate pollution. Data-driven models cover the relationships between the relevant input and output variables and are very effective if it is difficult or not possible to build knowledge-driven simulation models. Case study in Lower Savinja Valley aquifer indicates the possibilities and the directions of incorporation of data-driven models into the decision support frameworks.

Acknowledgment

The author gratefully dedicates this article to B. Sc. Mentor prof. dr. Dušan Kuščer, to M. Sc. Mentor prof. dr. Simon Pirc and to Ph. D. Mentor prof. dr. Jožef Pezdič and returns thanks for their invaluable pedagogical work and a wonderful example of professionalism.

References

Arthur J. D., Kaker, A. E., Cichon, J. R., Wood, H. A. R. & Rudin, A. (2005): Florida Aquifer Vulnerability Assessment (FAVA): Contamination potential of Floridas principal aquifer systems. Florida Deaprtment od Environmental Protection : Report submitted to Division of Water Resouce Management, 148 p.

Baker, A. E., Wood, A. R. & Cichon, J. R. (2007): The Marion County Aquifer Vulnerability assessment. Marion County Project No. SS06-01, 42 p.

Bračič-Železnik, B., Frantar, P., Janža, M. & Uhan, J. (2002): Ranljivost podzemne vode. In: Podtalnica Ljubljanskega polja. Rejec Brancelj, I. (ed.), Smrekar, A. (ed.), Kladnik, D. (ed.), Perko, D. (ed.). Geografija Slovenije, 10, Založba ZRC, Ljubljana, pp. 61–72.

Civita, M. (1990): La valutazione della vulnerabilità degli acquiferiall'inquinamento = Assessment of aquifer vulnerability to contamination. Proc 1st Conv. Naz. Protezione e Gestione delle Acqe Sotterranee: Metodologie, Technologie e Obiettivi. Marano sul Panaro, V.3, pp. 39–86.

Civita, M., De Maio, M. (2000): SINTACS R5. Quaderni di tecniche di protezione ambientale. Bologna: Pitagora Editrice 72, 226 p.

Domingos, P., Pazzani, M. (1997): On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29, pp. 103–130.

Dixon, B. (2001): Application of Neuro-fuzzy techniques to predict ground water vulnerability in NW Arkansas. Doctoral Dissertation, University of Arkansas, Environmental Dynamics, 265 p.

Dixon, B. (2004): Prediction of ground water vulnerability using integrated GIS-based neuro-fuzzy techniques. Journal od Spatial Hydrology, Vol. 4, No. 2, 38 p.

Gogu, C., Dassargues, A. (2000): Current trends and future challenges in groundwater vulnerability assesment using overlay and index methods. Environmental Geology, 39 (6), pp. 549–559.

Janža, M., Prestor, J. (2002): Ocena naravne ranljivosti vodonosnika v zaledju izvira Rižane po metodi SINTACS. Geologija 15/2, Ljubljana, pp. 401–406.

Jenness, J., Wynne, J. J. (2007): Cohens Kappa and Classification Table Metrics 2.1: An ArcView Extension for Accuracy Assessment of Spatially-Explicit Models. Dostopno na svetovnem spletu: <http://www.jennessent.com/arcview/kappa_stats.htm>

Kemp, L. D., Bonham-Carter, G. F., Raines, G. L. (1999): Arc-WofE: Arcview ex-

tension for weights of evidence mapping. Dostopno na svetovnem spletu: <http://www.ige.unicamp.br/wofe/>

Kononenko, I. (2001): Machine learning for medical diagnosis : history, state of the art and perspective. Artif. intell. med., Vol. 23, No. 1, pp. 89–109.

Masetti, M., Poli, S., Sterlacchini, S. (2007): The Use of the Weights-of-Evidence Modeling Technique to Estimate the Vulnerability of Groundwater to Nitrate Contamination. Natural Resources Research, 16/2, pp. 109–119.

Mali, N., Janža, M. (2005): Ocena naravne ranljivosti vodonosnika s SINTACS modelom v GIS okolju. Geologija 48/1, Ljubljana, pp. 127–140.

Metz, C. E. (1997): Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, pp. 283–298.

Mitchell, T. M. (1997): Machine Learning. McGraw-Hill, New York, 417 p.

Nauck, U., Kruse, R. (1995): NEFCLASS - A neuro-fuzzy approach for the classification of data. V Proceedings of the 1995 ACM symposium on Applied computing, Nashville, pp. 461–465.

Padovani, L., Trevisan, M. (2002): I nitrati di origine agricola nelle acque sotterranee. Un indice parametrico per l´individuazione di aree vulnerabili. Bologna, Pitagora Editrice, 103 p.

Pehan, S. (2008): Ocena vpliva antropogenega vnosa dušika na kakovost podzemne vode v Spodnji Savinjski dolini : diplomsko delo. Ljubljana, 67 p.

Price, R. K., Solomatine, D. P. (2009): A brief guide to hydroinformatics. UNESCO-IHE Institute for water education, Delft, 28 p.

Sorichetta, A., Sterlacchini, S., Blahut, J., Poli, S., Masetti, M. (2008): Groundwater vulnerability assessment: influence of using two subsets of wells, respectively with nitrate concentration above and below an established threshold, as training points in the weight of evidence (WofE) model. *Geophysical Research Abstracts*, 2 p. Dostopno na spletnem naslovu: <http://meetings.copernicus.org/www.cosis.net/>.

Uhan, J., Pezdič, J. & Civita, M. (2008): Assessing groundwater vulnerability by SINTACS method in the Lower Savinja Vally, Slovenia = Ocenjevanje ranljivosti podzemne vode z metodo SINTACS v Spodnji Savinjski dolini, Slovenija. RMZ, št. 55/3, Ljubljana, str. 363–376.

Uhan, J. (ed), Dobnikar Tehovnik, M. (ed) & Pavlič, U. (ed) (2010): Vode v Sloveniji. Ocena stanja voda za obdobje 2006–2008 po določilih okvirne direktive o vodah. Agencija RS za okolje, Ljubljana, 62 p.

Uhan, J. (2011): Ranljivost podzemne vode na nitratno onesnaženje v aluvialnih vodonosnikih Slovenije. Univerza v Ljubljani, Naravoslovnotehniška fakulteta, Ljubljana, doktorska disertacija, 163 p.

Uhan, J. (2012): Modeliranje ranljivosti podzemne vode in določanje nitratno ranljivih območij v Sloveniji. Slovenski vodar, Ljubljana, št. 25, pp. 107–110.

Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy, 2000, OJ L 327, 22.12.2000, pp. 1–73.