Groundwater vulnerability mapping of Witbank coalfield in South Africa using deep learning artificial neural networks

Emmanuel Sakala¹, Francois Fourie², Modreck Gomo², Henk Coetzee¹

¹Council for Geoscience ²University of the Free State

DOI: <u>http://dx.doi.org/10.4314/sajg.v8i2.12</u>

Abstract

This study highlights the usage of deep learning artificial neural networks in the assessment of groundwater vulnerability of a coalfield. The network uses the DRIST model with parameters (depth to water level, recharge, impact of the vadose zone, soils and topographic slope) as training inputs and borehole sulphate concentration as training output. This technique was applied to Witbank coalfield, where acid mine drainage emanating from coal mining operations is a huge concern for surrounding environment and groundwater resources. The generated groundwater vulnerability model was validated with another sulphate dataset not used during model training. The deep neural network model with dropout and decaying learning rate regularisers correlated very well with sulphate data from another source as compared to the index and overlay DRIST model. The approach, differentiated areas in terms of vulnerability to acid mine drainage, which can aid policy, and decision makers to make scientifically informed decisions on land use planning. The approach developed in this research can be applied to other coalfields in order to evaluate its robustness to different hydrogeological and geological conditions.

Keywords: Artificial neutral network, acid mine drainage, groundwater vulnerability, coalfield, Modelling.

1. Introduction

South Africa has nineteen coal provinces, of which the current mining activities are largely focused on coalfields in Mpumalanga Province of South Africa (Banks *et al.*, 2011). The coal mining industry has been a fundamental catalyst in the development of the South African economy for over a century (Bell *et al.*, 2001). However, coal mining operations in particular Witbank Coalfield has seriously affected the surrounding environment by massive deterioration of groundwater quality by Acid Mine Drainage (AMD) (Bell *et al.*, 2001). AMD is formed when sulphide-bearing material found in coal discard or ore bodies reacts with water in the presence of oxygen. The generated product is characterised by high concentration of toxic heavy metals and sulphide (Sakala *et al.*, 2018). In the study area, a regional knowledge-driven fuzzy expert system was used by (Sakala *et al.*, 2018), which correlated slightly with concentration of sulphate in boreholes. Due to a large amount of boreholes with sulphide concentration couple with the need to improve this correlation, a data-driven deep neural network approach was used. The production of

more accuracy groundwater vulnerability model of the coalfield would go a long way in demarcating AMD sensitive areas for appropriate land use planning at the same time testing the applicability of deep learning on large aerial groundwater vulnerability assessments.

Assessment of groundwater vulnerability is divided in two types: intrinsic and specific vulnerability (NRC, 1993). The intrinsic type defines only the ease with which pollutants migrates from surface to groundwater. On the other hand, the specific type takes into consideration the pollutant properties and its interaction with the subsurface (Vrba and Zoporozec, 1994; Huan et al., 2012). According to (Huan et al., 2012), in the modern days, the intrinsic vulnerability is considered meaningless as compared to specific type as factors affecting intrinsic vulnerability such as depth to water table, soil and recharge are changing as effects due to humans are increasing. This research hence, focuses on the specific vulnerability assessment in Geographic Information System (GIS) environment taking advantage of the good learning and generalisation capabilities of deep artificial neural networks to establishing the complex relationship between the groundwater vulnerability inputs and AMD indicators.

2. Method and Material

2.1. **Study Area**

The study area (~ 8 000km²) is located between 25°30" and 27°45" south latitude, 28°30" and 30°30" east longitude (Figure 1) and extends from Delmas to Wonderfontein in the west-east direction and between Kromdraai and Hendrina in the north-south direction. The study area is marked to the north by the edge of the Karoo rocks and to the south by the Smithfield Ridge, a palaeohigh. Several abandoned and current mining areas are scattered throughout the coalfield. The Olifants River and its tributaries form the main drainage system within the study area.



Figure 1. Geological setting and location of the study area

2.2. Geology and Hydrogeology

Geologically, the study area fall on the northern tip of the Main Karoo Basin which is a sedimentary basin which the first-order depositional sequence is generally placed in the Late Carboniferous, around 300 Ma. The study area consists of six major lithologies (Figure 1). The geological descriptions of the rocks are displayed in Table 1. The geology was extracted from the 1:250 000 scale geology. (du Toit and Sonnekus, 2014) subdivided lithologies into four aquifer types, viz. Intergranular - Fractured aquifers consisting of the Rooiberg Group, Loskop Formation, Bushveld Complex, Ecca Group, Karoo dolerite dykes and sills and the Pretoria Group quartzite. The intergranular aquifers correspond to the quaternary alluvial following major rivers. The dolomite of Chuniespoort Group of the Transvaal Supergroup forms karst aquifers found around Delmas Town. The fractured rock aquifers consists of the Black Reef Formation, Dwyka Group, Magaliesberg Formation, Wilger River Formation and the sedimentary rocks of the Pretoria Group with the bulk belonging to the intergranular - fractured aquifer type (Xws and Götz, 2014). The aquifers within the Witbank Coalfield are generally shallow (with distance less than 12 m to the water table) making them highly vulnerable to pollution and warranting a groundwater assessment to help manage and protect the water resources (Vrba and Zoporozec, 1994).

Lithology group (youngest to oldest)	Description
Quaternary - Alluvial	Unconsolidated sediments along rivers
The Karoo Supergroup	
Ecca Group	Sandstones, shales and coal
Dwyka	Diamictite (Tillite)
Intrusive rocks	Dolerite (Dykes and Sills)
Waterberg Group	Sandstone with conglomerate and shale
The Bushveld Igneous Complex	
Rustenburg, Lebowa Suite	Gabbro, norite, anorthosite; Felsic granites
Transvaal Supergroup	
Chuniespoort Group	Dolomite, chert
Pretoria Group	Shale, quartzite
Intrusive rocks	Diabase
Rooiberg Group	Rhyolite and Felsite
Loskop Formation	Sandstone, conglomerate
Witwatersrand Supergroup	×
Hospital Hill Formation, Central and West Rand	Shale, conglomerate and quartzite
Basement Complex	Granite and gneiss

Table 1. Geological description of rocks in the study area (Sakala et al., 2018)

2.3. DRIST Method

In this study, assessment of groundwater vulnerability is based on the overlay and index DRIST method (Chenini *et al.*, 2015), an improvement of the popular DRASTIC method by (Aller *et al.*, 1987). The method is based on five parameters of the DRASTIC method namely: depth to water level, net recharge, impact of the vadose zone, soil media and topography. The aquifer media and hydraulic conductivity present in the DRASTIC method are not used in the DRIST method (Chenini *et al.*, 2015). The vulnerability values are calculated similarly to the DRASTIC method.

2.3.1. Input - Depth to water level layer (D)

This refers to the distance between the ground level and the water table, which determines the passage through which water and pollutant have to travel to get to the aquifer (Aller *et al.*, 1987). The longer the distance, the higher the possibility of nature attenuating the pollutants as compared to shorter distances. The layer was calculated from monitoring borehole data obtained from the South African Department of Water and Sanitation. The water level data within the study area were interpolated using the kriging technique to generate the 30m resolution gridded raster layer which was reclassified into three classes as proposed by (Chenini *et al.*, 2015) (Figure 2a). The depth to water level varies between 4 to 21m and increases in the western direction.



Figure 2. DRIST inputs (a) Depth to water level (b) Recharge (c) Impact of the vadose zone (d) Soils and (e) Topographic slope

2.3.2. Input - Rainfall (R)

The rate of pyrite oxidation in the mining waste, tailings or subsurface is mainly controlled by the availability of oxygen and water at the mineral grain surface (Anawar, 2013). Mining operations expose the pyritic material to the atmosphere, making them prone to water and oxygen. Rainfall is often a major factor which supplies water for AMD reactions and which acts as a ligand for the transportation of products of the AMD reactions as surface runoff or as infiltration. From a hydrogeologist point of view, the greater the amount of rainfall, the greater the amount of

groundwater recharge, thus increase the potential for groundwater aquifer pollution. In the study area, the average long-term rainfall data (Figure 2b) which was obtained from historical data published by the South African Weather Services (SAWS, 2016).

2.3.3. Input – Impact of the vadose zone (I)

A laboratory column-leach experiment was conducted as part of this study to differentiate study area rocks in terms of their reactivity and removal of pollutants (toxics) from an AMD solution under unsaturated conditions. The results show that dolomite had the highest reactivity followed by mudrock, shales, dolerite, diabase, then felsite, rhyolite, and quartzite and lastly alluvial, sandstone, diamictite being the least reactive. The results are in agreement with findings by (Lapakko, 1994; Yager *et al.*, 2006). The vadose zone impact layer (Figure 2c) shows classification of rocks in terms of their ability to lessen or remove AMD pollutants load. Areas with lithologies having lesser ability to remove pollutants could be more vulnerability to AMD as less or no reactivity can mean that the AMD will pass through the vadose zone chemically unaltered and reach the groundwater. Generally, the largest area is marked by rocks with lower ability of reactivity which can translates to lack of attenuation of AMD pollution.

2.3.4. Input – Soil layer (S)

Soil data at a scale of 1:250 000 from Agriculture Research Council - Institute for Soil, Climate and Water (ARC-ISCW) was reclassified according to clay content. Clay materials are known to create a barrier zone restricting water and pollutant migration, thus the more the clay content, the better the barrier effect hence lessening the chances of groundwater pollution. Figure 2d shows the clay content layer of the study area, where the northern part is marked by low clay content. The clay content ranges from 9 to 51%.

2.3.5. Input – Topographic slope layer (T)

SRTM data of 30m resolution was used to generate the percentage slope layer (Figure 3e). Surface flatness increases the resident time for water and pollutants to react and infiltrate. Generally, the slope varies between 0 to 18% with the biggest portion of the study area being flat to gently sloping as marked by grey colour (Figure 2e).

2.3.6. Output – DRIST model

The five layers (Figure 3) were combined in a GIS environment using equation (2) to generate the DRIST groundwater vulnerability model. The equation is given by (Chenini *et al.*, 2015) as:

$$DRIST \ Index = D_r D_w + R_r R_w + I_r I_w + S_r S_w + T_r T_w$$
(2)

Where r is the rating of the parameter and w is the importance weight of the parameter shown on Figure 3.

The results were divided into four zones (very low, low, moderate and high) using the natural breaks method (Ioannou *et al.*, 2010). The output result is a map (Figure 3) showing the DRIST model, where grey areas are less vulnerability and the red coloured areas as most vulnerable. The

eastern areas are marked by moderate to high groundwater vulnerability whereas the central and western areas are very low to moderate.

2.4. Deep Artificial Neural Network

Artificial neural networks (ANNs) belong to the data-driven branch of Artificial Intelligence (AI) which is inspired by the biological neural system in terms of which the computer is trained to do the functions which, at the moment, are best handled by humans, such as learning (Shigidi and Garcia, 2005). Artificial neural networks (ANN) form a class of non-linear parallel distributed information processing and adaptive systems originally based on studies of the brains of living species (McCulloch and Pitts, 1943). An ANN consists of a layer of neurons that accept various inputs (Input layer) then fed them into further hidden layers with neurons and ultimately to the output layer, which produces an output response (Figure 4). The aim of the technique is to train the network such that its response to a given set of inputs is as close as possible to a desired output (McCulloch and Pitts, 1943). A number of algorithms are available for training a neural network of which the back propagation with gradient decent which is the most popular training algorithm (Manoj and Nagarajan, 2003) and was used in the present study. The five DRIST parameters were used as training datasets. ANN training, validation and classification was done using the Tensorflow® library in Python® programming language.



Figure 3. DRIST model showing the four groundwater vulnerability classes within Witbank coalfield



Figure 4. ANN architecture used in the study (Fijani et al., 2013)

3. Results and Discussion

The processes of formulating the five DRIST parameters (depth to water level, rainfall, impact of the vadose, soils and topography) and combining them using the traditional DRIST method was done in the previous section. In this section, results of building an ANN and resultant results will be discussed and model compared with the DRIST model.

3.1. ANN Training and validation

To build an ANN, the network has to be fed with labelled data (set of inputs and an output label). In this study, 350 boreholes with sulphate concentration values were extracted from the South African national groundwater archive database for building the ANN model. Boreholes with high sulphate values (>200mg/l) were used as polluted sites (output label -1) and those with values below 200mg/l as non-polluted training sites (output label -0). DRIST input values at the borehole locality for the five parameters together with associated borehole output label were fed into the ANN. The input-output pairs were partitioned into training, validation and test sets using a stratified 10-fold cross-validation approach (Kohavi, 1995).

In this paper, the most extensively used Rectified Linear Unit (ReLU) activation function (Epelbaum, 2017) was used as it does not activate all the neurons at the same time making the network sparse which in turn improves efficiency as compared with other functions such as the sigmoid, tanh and others (Gupta, 2017). Using the ReLU as an activation function with a gradient descent optimiser (Lv *et al.*, 2017), several experiments with various parameter settings (learning rate, number of training iterations, optimisation function and the architecture (number of neurons and hidden layers)) were done to choice the optimum parameters with the highest training and validation accuracy. Overfitting is a problem in building an ANN model, in this paper, the principle of early stopping and regularisers were used to minimise overfitting. Early stopping involves stopping the training process once performance of the validation dataset stops increasing (i.e. the cost begins increasing steadily instead of decreasing). In this study, the training process was stopped earlier at 430 epochs (Figure 5a). Figure 5b presents the number of hidden layers with the

performance measured using mean-squared error (MSE) at 430 iterations. The results obtained after this evaluation shows that the best number of the hidden layers which corresponds to the smallest MSE, and good ANN outputs is a three hidden layer system. From Figure 5c, the MSE shows a minimum value when the hidden neurons number is 35, indicating that an ANN with 35 hidden neurons provides a better optimum number of neurons, where increasing the number has not effect on the training process. Therefore, an ANN with architecture 5-35-35-35-1 (five inputs, 35 hidden neurons in each of the three hidden layers and one output) was used.



Figure 5. Parameters setting (a) number of epochs (b) number of hidden layers (c) number of neurons

The ANN trained without regularisation at a constant learning rate of 0.5 was able to reach 83% accuracy in training and 81% in validation, but the training results show erratic changes in the accuracy (Figure 6a). When, the dropout technique (Gupta, 2017) which involves stochastically dropping some of the hidden neurons was used together with an exponential decay learning rate (Brownlee, 2016) the accuracy improves drastically to 92.3% accuracy for training and 95.7% for validation and the erratic variations disappear (Figure 6b). Thus, adding regularisers greatly improved the training process and the network can now generalise better, and considerably reducing the effects of overfitting.



Figure 6. Training process (a) without regularisers (b) with dropout and learning rate

The trained networks were used to predict the output values (termed 'groundwater vulnerability values') at each of the 30m x 30m grid points within the study area. The groundwater vulnerability model values were reclassified into four classes using natural break method as shown on Figure 7. The model shows the groundwater vulnerability model generated using sulphate training where grey coloured areas are less vulnerability and the red coloured areas as the most vulnerable areas.



Figure 7. ANN groundwater vulnerability model

The ANN model shows high groundwater vulnerability for the area surrounding Witbank, Ogies, Hendrina and south of Middelburg. These areas are marked by soils with less clay content, shallow water table, moderate rainfall amount and underlain by rocks with poor neutralising potential showing lack of the natural attenuation of AMD which makes the groundwater highly vulnerability. Several authors (de Villiers and Mkwelo, 2009; Dabrowski and de Klerk, 2013) did studies on the Olifants River catchment and reported that high content of sulphate concentrations and low pH in water resources downstream of current and abandoned mining. The pollution is reported to be affecting farmers, people living in the catchment as well as tourism and wildlife of the Kruger National Park (de Villiers and Mkwelo, 2009). The conditions which makes the groundwater highly vulnerable to AMD as explained in present study could be one of the reasons for results recorded by (de Villiers and Mkwelo, 2009; Dabrowski and de Klerk, 2013). In terms of possible physicalchemical reasoning, the vulnerability results obtained is a combination of presence or lack of physical trapping mechanisms (i.e. distance to water level determines the residence time for which the subsurface can attenuate AMD and interstitial void size determines AMD movement) and the soils and lithology controls the migration by pollution sorption processes, chemical reactions which are depended on the clay content and AMD-rocks reactivity.

3.2. Model validation

The sulphate parameter which is a good indicator for AMD was used for model validation and for comparing the performance of the ANN and DRIST groundwater vulnerability models. Groundwater samples collected in July 2013 were analysed by Waterlab testing laboratory for cations, anions and alkalinity using Inductively Coupled Plasma Mass Spectrometry (ICP-MS), Ion Chromatography (IC) and Spectrophotometry respectively. Correlation scatter plots between the DRIST and ANN models with measured sulphate values were produced and displayed on Figure 8(a) and (b) respectively. The correlation value of 0.4158 was obtained for the DRIST model as compared to the 0.8392 values for the ANN model.



Figure 8. Correlation scatter plot for testing dataset for (a) DRIST (b) ANN models

The low correlation value obtained for DRIST model could be because the DRIST method lacks the connection with measureable AMD parameter basing only on the surface and subsurface conditions without considering any specific aspects of AMD. On the other hand, the ANN model also based on the same parameters as the DRIST method but training using AMD indicators improves the correlation dramatically because it is more specific to AMD pollution with which sulphate is a sensitive indicator. This significant improvement means that the ANN model is better in terms of highlighting the vulnerability status of the groundwater system. Comparison of the deep ANN model with the fuzzy model (Sakala *et al.*, 2018), the correlation results of the ANN are higher. Thus, the deep ANN, which was purposefully built from complex relationship the network learnt from the input-output pairs, improves the correlation.

4. Conclusions

The purposefully built deep ANN using the ReLU, gradient descent optimiser with dropout and an exponential decay-learning rate was able to generate a groundwater vulnerability model, which correlates very well with physically measured field data. The results significantly improves the quality of the output model when compared with the use of the index and overlay DRIST method and knowledge-driven model from literature. Results can be used help policy and decision makers

to make scientifically informed decisions on land use planning. Based on findings of this study, management and protection of groundwater resources recommendations can be made:

- Land use activities that generate sulphates like coal mining should be avoided on highly vulnerable zones or done in a strict manner that minimises pollutants from entering the subsurface.
- Rehabilitation exercise over abundant coal mining areas, which are still generating AMD within or near the highly vulnerable zones.

The datasets used in this study are readily available from various governmental agencies making the approach cost-effective in evaluating larger areas. The AMD groundwater vulnerability approach developed in this research need to be applied to other pollutants with similar or different hydrogeological settings in order to determine the robustness of the methodology.

5. Acknowledgements

The authors would like to thank the Department of Mineral Resources, Council for Geoscience, Department of Water Affairs and the Agricultural Research Council Institute for Soil, Climate and Water for access to the various datasets used in this study.

6. References

- Aller, L, Bennett, T, Lehr, JH, Pretty, RJ. and Hackett, G 1987, DRASTIC: A standardized system for evaluating ground water pollution potential using hydrogeologic settings. US Environmental Protection Agency, Ada, Oklahoma. Report No. EPA/600/ 2-87-036.
- Anawar, HM 2013, Impact of climate change on acid mine drainage generation and contaminant transport in water ecosystems of semi-arid and arid mining areas, Physics and Chemistry of the Earth. Elsevier, Ltd, 58(60), 13–21. doi: 10.1016/j.pce.2013.04.002.
- Banks, VJ, Palumbo-Roe, Van Tonder, DB, Davies, J, Fleming, C and Chevrel, S 2011, Conceptual models of the Witbank coalfield, South Africa. Earth observation for monitoring and observing environmental and societal impacts of mineral resources exploration and exploitation. CEC FP7 Project EO-MINERS, Deliverable D3, 107.
- Bell, FG, Bullock, SET, Hälbich, TFJ and Lindsay, P 2001, Environmental impacts associated with an abandoned mine in the Witbank Coalfield, South Africa, International Journal of Coal Geology, 45(2), 195–216. doi: 10.1016/S0166-5162(00)00033-1.
- Brownlee, J 2016, Using Learning Rate Schedules for Deep Learning Models in Python with Keras.
- Chenini, I, Zghibi, A and Kouzana, L 2015, Hydrogeological investigations and groundwater vulnerability assessment and mapping for groundwater resource protection and management: State of the art and a case study, Journal of African Earth Sciences. Elsevier Ltd, 109(February 2016), 11–26. doi: 10.1016/j.jafrearsci.2015.05.008.
- Dabrowski, JM and de Klerk, LP 2013, An assessment of the impact of different land use activities on water quality in the upper Olifants River catchment, Water SA, 39(2), 231–244.
- Epelbaum, T 2017, Deep learning: Technical introduction. Mediamobile.
- Fijani, E, Nadiri, AA, Asghari Moghaddam, A, Tsai, FTC and Dixon, B 2013, Optimization of drastic method by supervised committee machine artificial intelligence to assess groundwater vulnerability for

maragheh-bonab plain aquifer, Iran, Journal of Hydrology, 503, 89–100. doi: 10.1016/j.jhydrol.2013.08.038.

- Gupta, D 2017, Fundamentals of Deep Learning Activation Functions and When to Use Them? viewed 30 April 2019, https://www.analyticsvidhya.com/blog/2017/10/fundamentals-deep-learning-activation-functions-when-to-use-them/>.
- Huan, H, Wang, J and Teng, Y 2012, Assessmen and validation of groundwater vulnerability to nitrate based on a modified DRASTIC model: A case study in Jilin City of northeast China, Science of the Total Environment, 440, 14–23.
- Ioannou, K, Myronidis, D, Lefakis, P and Stathis, D 2010, The use of artificial neural networks (ANNs) for the forecast of precipitation levels of Lake Doirani (N. Greece), in Fresenius Environmental Bulletin 19(9a), 921–1927.
- Kohavi, R 1995, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection 2 Methods for Accuracy Estimation, in International Joint Conference on Artificial Intelligence, 1137– 1145. doi: 10.1067/mod.2000.109031.
- Lapakko, KA 1994, Comparison of Duluth Complex rock dissolution in the laboratory and field, in International Land Reclamation and Mine Drainage Conference and 3rd International Conference on the Abatement of Acidic Drainage. Pittsburgh, PA, 419–428.
- Lv, K, Jiang, S and Li, J 2017, Learning Gradient Descent: Better Generalization and Longer Horizons, in 34th International Conference on Machine Learning. Sydney, Australia.
- Manoj, C and Nagarajan, N 2003, The application of artificial neural networks to magnetotelluric time-series analysis, Geophys. J. Int, 153, 409–423.
- McCulloch, WS and Pitts, W 1943, A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics, 5, 115–133.
- National Research Council (NRC) 1993, Groundwater vulnerability assessments: Contamination potential under conditions of uncertainty. Washington, DC: National Academy Press.
- Sakala, E, Fourie, F, Gomo, M and Coetzee, H 2018, GIS-based groundwater vulnerability modelling: A case study of the Witbank, Ermelo and Highveld Coalfields in South Africa, Journal of African Earth Sciences, 137, 46–60. doi: 10.1016/j.jafrearsci.2017.09.012.
- Shigidi, A and Garcia, LA 2005, Parameter estimation in groundwater hydrology using artificial neural networks, Journal of computing in civil engineering, 31, 281–289.
- South African Weather Service (SAWS) 2016, Historical Rain Maps. Veiwed 23 March 2016, http://www.weathersa.co.za/climate/historical-rain-maps.
- du Toit, WH and Sonnekus, CJ 2014, An explanation of the 1:500 000 general hydrogeological map, Nelspruit 2530. Pretoria.: Department of Water and Sanitation.
- de Villiers, S and Mkwelo, ST 2009, Has monitoring failed the Olifants River, Mpumalanga?, Water SA, 35(5), 671–676. doi: 10.4314/wsa.v35i5.49193.
- Vrba, J and Zoporozec, A 1994, Guidebook on mapping groundwater vulnerability, in International Contributions to Hydrogeology. VerlagHeinzHeiseGmbHand Co. KG.
- Xws, PJ and Götz, AE 2014, South Africa's coalfields A 2014 perspective, International Journal of Coal Geology, 132, 170–254. doi: 10.1016/j.coal.2014.06.019.
- Yager, DB, McCafferty, AE, Stanton, MR, Diehl, SF and Driscoll, RL 2006, Net Acid Production, Acid Neutralizing Capacity and Associated Geophysical, Mineralogical, and Geochemical Characteristics of Animas River Watershed Rocks, Silverton, Colorado, 7th International Conference on Acid Rock Drainage, 2540–2541.