Developing a risk assessment for local groundwater degradation by mine waste storage facilities

Josepha SDR Zielke-Olivier[1], Danie Vermeulen[1]

[1]*University of the Free State, Institute for Groundwater Studies, 205 Nelson Mandela Drive, Park West, 9301 Bloemfontein, South Africa, josephazielke@gmail.com, vermeulend@ufs.ac.za*

Abstract

A linear and nonlinear statistic approach was chosen to develop a risk assessment predicting the groundwater pollution potential and SO_4 concentrations considering twelve environmental and spatial variables at an industrial and coal mining complex. Linear regression models and nonlinear classification and regression trees indicated that the explanatory variables *ln*WLDepth and vadose zone were most significant in predicting the potential pollution risk and SO_4 values. Tree models were able to identify additional correlations between SO_4 and distance to pollution, fault and stream as they recognize nonlinear relationships, and were found to be useful visual tools to develop a site-specific risk assessment.

Keywords: sulfate prediction, linear regression, classification and regression trees

Introduction

Mine waste facilities are often substantial sources of diffuse and point pollution which affects the water quality of the surrounding environment (Morin and Hutt, 2001). Groundwater monitoring is limited at these sites due to financial constraints and the possibility of creating unnecessary pathways between the pollution source and the aquifer due to drilling. This results in incomplete data sets and expensive monitoring programs (Babiker *et al.* 2005). Therefore, risk assessments have become helpful tools and are widely used to delineate areas that are more prone to groundwater pollution due to anthropogenic activities (Babiker *et al.* 2005; Kazakis and Voudouris 2015). Once vulnerable areas have been identified, they can be targeted with refined monitoring programs and individual remediation or prevention techniques (Babiker *et al.* 2005).

A widely accepted aquifer vulnerability assessment is the DRASTIC method which has been studied in a combination of various statistical approaches (eg. Huan *et al.* 2018; Kazakis and Voudouris 2015; Khosravi *et al.* 2017). It is a deterministic approach using order and ranking to evaluate multiple options with specific variables that are not measured but classified according to a rank. This method incorporates the major hydrogeological factors affecting and controlling groundwater movement including **D**epth to water table, net **R**echarge, **A**quifer media, **S**oil media, **T**opography, **E**ffect of the vadose zone and hydraulic **C**onductivity of the aquifer (Aller *et al.* 1985). However, the DRATIC method is not suitable to predict aquifer vulnerability for small, relatively homogenous areas as DRASTIC is designed to assess different locations with spatially variable phenomena (Babiker *et al.* 2005). Therefore, this study considers an alternative approach employing simple linear models and regression and classification trees which can deal with nonlinear relationships and high-order interactions of complex data (De'Ath and Fabricius 2000).

The aquifers of investigation form part of the Karoo Supergroup located in the north-eastern portion of the Karoo Basin, South Africa. At the study site, seepage from different types of discard dumps containing waste rock and by-products from a coal mining and power station facility and mine water storage dams, typically contain elevated sulfate concentrations. High sulfate levels were detected in the local shallow weathered and deeper fractured aquifer near such mine waste storage facilities and local streams.

These sulphate levels exceed both the South African Drinking Water Standard and water use license requirements of 200 and 134 mg/L SO_4, respectively.

This study aimed to develop a risk assessment by employing a linear and nonlinear statistic approach to predict a) groundwater pollution risk by SO_4 and b) SO_4 concentrations in areas without monitoring boreholes based on environmental and spatial variables. This would aid in the decision making of the mining and industrial complex to be able to comply to the water use licence and to protect the local ground- and surface water quality. In addition, the site of investigation was divided into a control and study site involving a fine coal discard dump. A statistic model should determine whether measured elevated SO_4 concentrations were a result of seepage from the discard dump or were to be expected in the area even in the absence of the discard dump.

Methodology

Study site

The industrial and coal mining complex has an approximate size of 209 km^2 and is located in the Mpumalanga Province, South Africa. This area forms part of the north-eastern section of the Karoo Basin and is composed of upward-coarsening cycles of siltstone, mudstone, immature sandstone, carbonaceous shale and coal seams of the Permian Vryheid Formation of the Ecca Group (Johnson *et al*. 2006). Locally, the formation was intruded by late Karoo sub-horizontal dolerite sills with a thickness ranging between 30 and 60 m (Hulley 2013). Several faults were mapped in the area during mining activities, which form part of a larger graben structure with a displacement ranging from 22 to 55 m (Hulley 2013; pers. com. Vermeulen 2015).

Two aquifers control the geohydrological setting, an unconfined shallow weathered and deeper semi-confined fractured Ecca aquifer (Grobbelaar 2001) with an average yield of 0.6 and 0.2 l/s, respectively (King 2003). Although these aquifers are relatively low yielding, bedding planes and secondary structures such as fractures, fissures and faults may form preferential flow paths that would allow seepage from the mine waste storage facilities to reach the aquifer system.

Data acquisition

Data were obtained from boreholes forming part of a continuous monitoring program. Sulfate was treated as the response variable, as it was determined to be the major pollution culprit at the study site. The explanatory variables were based on the DRASTIC model (Aller *et al*. 1985), but parameter *R* was excluded, since only one value was assigned to the relatively small study area, making it statistically irrelevant.

In addition, distance to nearest pollution, study dump, pollution excluding study dump, fault, stream and electrical resistivity tomography (ERT) were considered as explanatory variables (Tab.1). For SO_4 and depth to water level, the latest value was used. The slope % was calculated for small sections of the study site according to Freeze and Cherry (1979), utilizing the coordinates and elevation of three boreholes at a time. The hydraulic conductivity was obtained from a hydrogeological model constructed for the area (IGS Report No. 01/2018/AA). Distances from boreholes to potential pollution sources were calculated using the middle point of each source. Distances to streams and faults were measured in intervals by constructing buffer zones of 10, 50, 100 and 500 m around the linear features. Observations with missing SO_4 data were excluded from the analysis as the remaining information of the dependent variable did not add any value to the statistical model.

Statistical analysis

Different statistical model approaches were chosen to explore relationships between the variables and to predict 1) the probability of high (SO_4>134 mg/L) and low risk (SO_4<134 mg/L) of groundwater pollution, and 2) the sulfate concentration in the groundwater based on the given environmental parameters. Seventy percent of the data set was randomly split into a training sample and the remaining 30% were used as control sample to validate the statistical model. Variables SO_4, K, WLDepth, DistPol, DistToDump and DistWithout were

Table 1 *Description of the study variables; the type of variable is denoted by N=numeric or C=categorical; materials are denoted by ss=sandstone, BKFL=backfill, mudst=mudstone.*

Variable	Short name	Type	Value (min. and max. range)
SO4 concentration (mg/L)	SO4	N	0.25-31071
Distance to pollution (m)	DistPol	N	88-14393
Distance to study dump (m)	DistToDump	N	224-2939244
Distance to closest pollution without study dump (m)	DistWithout	N	88-2935399
Depth to water level (m)	WLDepth	N	0-108
Aquifer geology	Aquifer	C	bedded shale/ss/mudst, BKFL, dolerite, dolerite weathered
Soil	Soil	C	absent, BKFL, clay, clay loam, gravel, loam, sand, sandy loam
Vadose zone	Vadose	C	bedded shale/ss/mudst, BKFL, dolerite, dolerite weathered, clay, siltstone
Hydraulic conductivity (m/d)	K	N	0.05 (discard dump), 0.14 (weathered aquifer), 14 (conductive area), 85 (fault)
Slope %	Slope	N	0-3
Distance to fault (m)	Fault	N	0-12308
Distance to stream (m)	Stream	N	0-4000
ERT (Ωm)	ERT	N	1.8-96

transformed to the natural log scale due to high variations in minimum and maximum values. All statistic methods were run with the freely available software R version 3.4.2 (The R Foundation for Statistical Computing 2017).

At first, a logistic regression model was implemented with the dependent variable SO_4 posing a high (>134 mg/L) or low risk (<134 mg/L), based on the water quality objectives of the water use license. The control sample set excluded data from artesian boreholes. An iterative approach was applied to find combinations of factors that explain the risk of elevated SO_4 ranging from including all to none of the explanatory variables (Tab. 1). The different models were compared by means of the Bayesian Information Criterion (BIC) to try and balance the accuracy of the model fit against the model complexity.

As a next step, a linear regression model was developed to predict future SO_4 values using the natural log (due to a wide range of values) and a normal distribution. Similarly, to the logistic regression, an iterative approach was applied using a combination of factors. The models were then fitted against the log of SO_4 with a significance level of 0.05 (P=0.05) to determine the most applicable model.

Classification and regression trees were applied to fit non-linear models by means of decision trees that try to maximally differentiate the sample with each succeeding split in a branch. Regression and classification tress are useful tools when dealing with nonlinear relationships and high-order interactions of complex data sets (De'Ath and Fabricius 2000). Splitting of tree branches was performed with the aid of the Gini index for the classification tree, and with the ANOVA method (sums of squares) for the regression tree. Other than the previous models, all variables were included to populate the trees. The trees were then pruned utilising the default complexity parameter (CP) method.

As final approach, a robust regression model was utilized to try and predict the natural log of SO_4 ($ln SO_4$) based on the observed sulfate measurement of the

trainings sample. It differs from the linear regression, that samples with a depth to water level between 0 and 0.14 m were included. A Bayesian framework was employed with a student -t distribution as it is more robust to outliers but more difficult to fit.

Results and discussion

Predicting sulfate concentrations with linear functions

Results of the logistic regression model indicated that simple models including only the significant variables were preferred. Based on the BIC value (393.26), the second logistic regression model was the most favourable with only the two explanatory variables vadose zone and *ln*WLDepth being statistically significant (for P<0.05). The resulting linear formula is $Y \approx -0.4868 \times lnWLDepth + 0.0695$ (add -2.0739 of vadose=clay, 1.4054 if vadose=dolerite and 1.4957 if vadose=dolerite weathered).

Similar to the logistic regression model, the linear regression model considered the two variables vadose zone and *ln*WLDepth to be statistically most significant. According to the lowest BIC value, the model that could most reliably predict future sulfate concentrations is $SO_4 \approx e^{-0.458 \times \ln WLDepth + 4.919}$ (but add to 4.919 a) -2.804 if the vadose zone=clay, b) 1.137 if the vadose zone=dolerite, or c) 1.22 if the vadose zone=dolerite weathered). This prediction is based on the control sample given that the water level is below the surface (not artesian). Considering the control and study area around the discard dump, the model suggested that the study area has a generally higher SO_4 concentration with an increase factor of 95.5% than predicted by the control area.

For the robust regression model, the same linear model was fitted using a Bayesian framework with a more robust student-t distribution including artesian borehole measurements. The resulting formula for the model with the best fit predicting future sulfate concentrations was $SO_4 \approx e^{-0.334 \times \ln WLDepth}$, adding a) 4.782 to the exponent if the vadose zone is composed of shale, b) 2.007 for clay, c) 5.834 for dolerite or d) 6.045 for weathered dolerite. This model proposed that the study area has a generally higher SO_4 concentration with an increase factor of 88.3% compared to the predictions of the control area.

All linear model approaches were consistent in that they only considered the two explanatory variables depth to water level (as natural log) and vadose zone as the major environmental factors that control the risk of sulfate pollution at the investigated site. Both linear and robust regression formulas showed similar results for the prediction of future sulfate concentrations in the control area. Furthermore, both models indicated that the observed sulfate concentrations in the study area were much higher than predicted for the control area. This suggests that there are additional environmental factors that were not considered in the model influencing the sulfate concentration in the groundwater. For example, the study site contains a backfilled area, wetlands and some observed groundwater flow paths at depth, possibly due to anthropogenic alterations of the geology caused by undermining which could play a role in the SO_4 distribution. It was surprising that the distance to the closest pollution point statistically did not have a significant effect on the sulfate concentration in the groundwater, although the sulfate distribution suggested otherwise. Boreholes closest to the tailings dumps and mine water dams had the highest sulfate concentrations. It is therefore recommended to revisit the methodology on how the distance of boreholes to the closest pollution source was calculated.

Risk evaluation and sulfate prediction with nonlinear functions

Nonlinear regression and classification trees were applied to relate SO_4 concentrations to spatial and physical environmental variables and to predict SO_4 concentrations as well as the risk of groundwater pollution by SO_4 (>134 mg/L) based on a random training sample. The regression tree was overfitting using all variables, but showed better results considering only the variables *ln*WLDepth, vadose zone and distance to pollution, fault, and stream (Fig. 1A). According to the splitting of branches, the variable *ln*WLDepth was most significant to determine the risk of sulfate pollution for both the regression and classification tree, followed by the vadose zone (Fig. 1). This corresponds with the

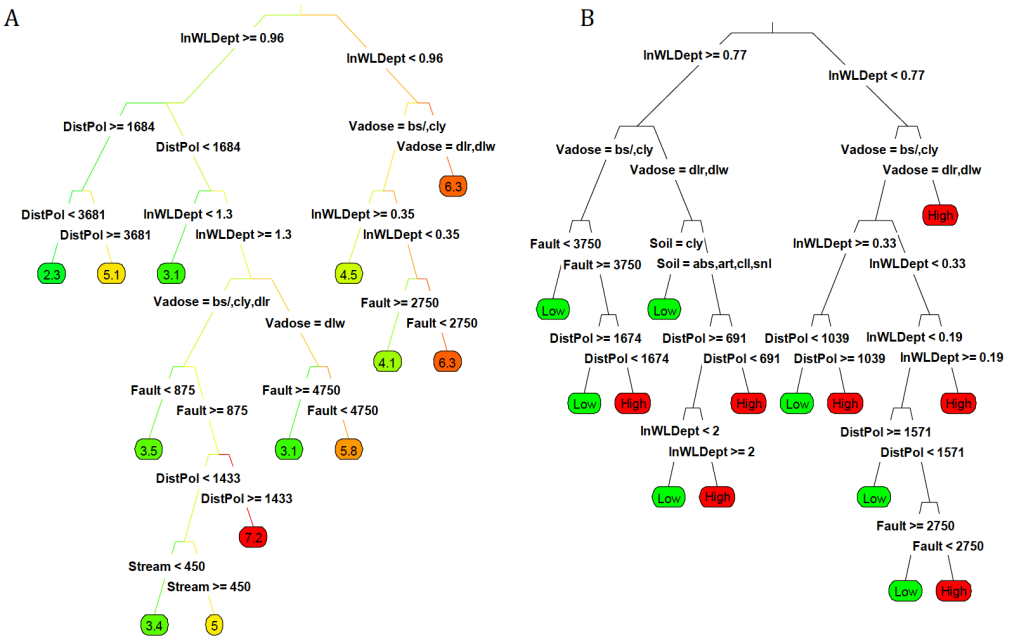

Figure 1 Regression (A) and classification tree (B) of a random training sample with SO_4 as dependent variable, SO_4 predictions expressed as natural log; vadose zone materials are denoted as bs=bedded shale, sandstone, mudstone, cly=clay, dlr=dolerite, dlw=dolerite weathered; soil types are denoted as cly=clay, art=artificial, cll=clay loam, snl=sandy loam.

findings of the linear regression models.

In the regression tree, distance to pollution was also considered as a significant variable, followed by distance to fault and stream (Fig. 1A). Although the regression tree was able to model complex data to determine environmental characteristic associated with SO_4 groundwater distribution, it showed discrepancies when trying to explain the distance to pollution with SO_4 concentrations. For example, it could not reasonably explain an elevated SO_4 concentration at a distance to pollution≥3681 m compared to a shorter distance to pollution with lower SO_4 concentration. This contradicts the field observations and could be explained by either additional factors that were not considered in this risk assessment or due to inconsistencies in the data set.

Other than the regression tree, the classification tree also considered the soil type to be a significant branch to distinguish between high and low pollution risks of SO_4 but did not include distance to stream (Fig. 1B). Like the regression tree, one branch split indicated that a low SO_4 risk is expected for distances closer to the pollution source compared to distances further away (>1039 m) which disagrees with the field observations and could be related to other factors not being considered in the risk assessment. Overall, the regression and classification trees were a useful and visual tool to predict SO_4 concentrations and potential risk of SO_4 pollution for unknown areas in the field. Furthermore, this nonlinear approach was able to identify additional relationships between variables which were not addressed with linear functions due to the complexity of the data.

Conclusions

Sulfate predictions and groundwater pollution risk were assessed with linear and nonlinear functions to determine the environmental and spatial relationships with SO_4 at an industrial and mining site. All statistical approaches indicated that the explanatory variables *ln*WLDepth and vadose zone were most significant in predicting the potential pollution risk and future SO_4 values. No relationship was found between

the dependent variable SO_4 and the response variables ERT, slope%, hydraulic conductivity and aquifer geology. This can be explained by a limited ERT data set and a small data range of the remaining response variables. The simple linear and robust regression model showed that the study area around the discard dump had a generally higher SO_4 concentration than predicted by the control area with a factor of 95.5 and 88.3%, respectively. This could be caused by additional environmental factors not being considered for the study site such as a backfilled area, wetlands and groundwater flow paths possibly related to underground mining activities. Compared to the logistic and linear regression analysis, regression and classification tress were able to identify additional relationships between SO_4 and distance to pollution, fault and stream as they consider nonlinear relationships and high-order interactions of complex data sets. Regression and classification trees also provided a useful visual tool to predict SO_4 for areas not tested at the study site and to evaluate the pollution potential of the aquifer. It is recommended to re-evaluate the method applied to determine the distance of boreholes to closest pollution points as no relationship was determined with the linear regression models, although the field observations suggested otherwise.

Acknowledgements

The authors thank Dr. Sean van der Merwe for his time to generate the statistic models with the software R and for his consultation regarding different statistical approaches. In addition, the authors would like to thank the petrochemical company for providing essential data for this research project.

References

Aller L, Bennett T, Lehr JH, Petty RJ (1985) DRASTIC: A Standardized System for Evaluating Ground Water Pollution Potential Using Hydrogeological Settings. US EPA, Washington, DC, 622 pp

Babiker IS, Mohamed MAA, Hiyama T, Kato K (2005). A GIS-based DRASTIC model for assessing aquifer vulnerability in Kakamigahara Heights, Gifu Prefecture, central Japan. Sci. Total Environ. 345:127—140

De'Ath G, Fabricius K (2000) Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology 81:3178—3192

Freeze RA, Cherry JA (1979). Groundwater. Prentice Hall, Englewood Cliffs, New Jersey, 604 pp

Grobbelaar R (2001) The Long-Term Impact of Intermine Flow from Collieries in the Mpumalanga Coalfields. Master Thesis, Univ of the Free State, 136 pp (unpublished)

Huan H, Zhang B-T, Kong H, Li M, Wang W, Xi B, Wang G (2018). Comprehensive assessment of groundwater pollution risk based on HVF model: A case study in Jilin City of northeast China. Sci. Total Environ. 628—629:1518—1530

Hulley V (2013). In situ source characterisation of dense non-aqueous phase liquids (DNAPLs) in a fractured rock environment. PhD Thesis, Univ of the Free State, 210 pp (unpublished)

Institute for Groundwater Studies (IGS) (2018). The Influence of Different SASOL Operations in the Graben Structure at SASOL, Secunda. Report No. 01/2018/AA, 67 pp (unpublished)

Johnson MR, van Vuuren CJ, Visser JNJ, Cole DI, de V Wickens H, Christie ADM, Roberts DL, Brandl G (2006) Sedimentary Rocks of the Karoo Supergroup. In: Johnson MR, Anhaeusser CR, Thomas RJ (Eds), 2006 The Geology of South Africa, Geological Society of South Africa, Johannesburg/Council for Geoscience, Pretoria, p 461—500

Kazakis N, Voudouris KS (2015). Groundwater vulnerability and pollution risk assessment of porous aquifers to nitrate: Modifying the DRASTIC method using quantitative parameters. J. Hydrol. 525:13—25

Khosravi K, Sartaj M, Tsai FT-C, Singh VP, Kazakis N, Melesse AM, Prakash I, Bui DT, Pham BT (2017). A comparison study of DRASTIC methods with various objective methods for groundwater vulnerability assessment. Sci. Total Environ. 642:1032—1049

King GM (2003) An Explanation of the 1:500 000 General Hydrogeological Map: Vryheid 2730. Department of Water Affairs and Forestry, Pretoria, 41 pp

Morin KA, Hutt NM (2001) Environmental Geochemistry of Minesite Drainage: Practical Theory and Case Studies. MDAG Publishing, Vancouver, British Columbia, 333 pp

Vermeulen PD (2015) Personal communication, Institute for Groundwater Studies, Univ of the Free State, South Africa