Using Data Science and Machine Learning to Improve Site Hydrogeological Conceptual Models

Tim R. Ezzy[1], John Fortuna[1]

[1]Principal Hydrogeologist, Golder Associates, Milton, Queensland, Australia, TEzzy@golder.com.au, JFortuna@golder.com.au

Abstract

A key goal of many mining groundwater investigations is to identify the main geological features, hydraulic boundaries and connection pathways that will materially influence: a) operations of a project, and b) the natural resources connected to the groundwater system. Exploratory data science techniques such as machine learning provide the experienced mining hydrogeologist opportunities to accelerate understanding of the role of key features within a site hydrogeological conceptual model (HCM) that may affect groundwater management. This has implications for both regulatory approval processes and operational efficiency.

Keywords: hydrogeological conceptual model, data science, machine learning, groundwater management

Introduction

Data science comprises three overlapping disciplines: 1) statistical modelling and analysis; 2) computer science skills necessary to efficiently store, process and visualise data; and 3) domain expertise in terms of classical training in a subject (VanderPlas 2017). The domain expertise (in this case mining hydrogeology) is necessary to pose the right questions and contextualise the outputs of the analysis.

In this paper we will review a recent mining project where data science techniques were implemented to identify which key geological features that have potential to influence future mine dewatering rates. The authors were engaged by our mining client to undertake a feasibility study (FS) for water management and river realignment for extension of several open cut pits. The mine site is underlain by sedimentary rock aquifers made up of calcareous sediments and dolomite with a porphyry intrusion aquitard. The aquifers are intersected by a dense network of subsurface faults that are situated within a broader pull-apart basin.

Faults have potential to control groundwater flow dynamics via cross-fault juxtaposition and modification of rock properties in the vicinity of the fault (e.g. Caine et al. 1996; Scibek et al. 2016; McCallum et al. 2018; OGIA 2019). Conceptually, there are three categories of permeability structure within a fault zone: barriers, conduits and barrier-conduits (Caine et al. 1996). Barriers tend to reduce groundwater flux, for example, by reducing the permeability along the central fault core. Conduits tend to increase flux either across faults or along fault planes, often through enhanced permeability of the fault damage zone. Barrier-conduits tend to increase fluxes parallel to the fault plane (along the damage zone corridor) while limiting flux across the fault core.

A key water management concern for future development is the potential for connections between the open cut pits and a nearby river through faults and related structural features. The null hypothesis at the mine site was that faulting has no influence on aquifer hydraulic properties. Previous groundwater modelling studies at the site (e.g. the mine prefeasibility study) adopted this hypothesis, and faults were not explicitly differentiated in the model domain. Given the stress regime and geological setting, an alternative hypothesis is that the normal NE-SW faults are oriented favourably with the principal horizontal stresses and are therefore expected to be dilated and could act as regional conduits for groundwater fluxes. Ten regional-scale NE-SW faults have been mapped in close vicinity to the proposed mine pits. If some, or all of these faults were conduits, then it would have implications for pit dewatering requirements. Further, a number of these faults intersect a major regional perennial river, and the degree of groundwater-surface water connectivity may be a controlling factor in the feasibility of mining in this area.

Aquifer testing undertaken during the FS development was designed to specifically challenge these hypotheses. Aquifer tests were conducted in different rock types that were situated at a range of distances from variably-oriented faults. The objective of the data science approach was to interrogate the thousands of data points that were generated by the field investigations and to challenge the null hypothesis.

Methods

There are 75 estimates of hydraulic conductivity (K) generated from six constant rate pumping tests, including recovery tests, and 37 slug tests undertaken during various stages of project development. These aquifer tests were undertaken at three main hydrostratigraphic units and key geological contact zones. Twenty-three groundwater samples were analysed for major ion chemistry to provide an independent line of evidence on the influence of faults on regional groundwater flow processes. The K data and chemistry data were interrogated using an integrated data science approach that included geological modelling, geospatial analysis and statistical analysis. The integrated data science approach incorporated:

- Geological modelling to visualise the three-dimensional architecture of the hydrostratigraphic units and faults, using Leapfrog software (Seequent).
- Analysing the aquifer testing datasets including time-drawdown derivative analysis to further characterise system boundary conditions and flow system types, using AQTESOLV software (HydroSOLVE, Inc.).
- K data and chemistry data were embedded with geospatial attributes including host rock type and distances to the nearest fault types. QGIS 3.8.3 is an open-source geographic information system (GIS) platform that was used for spatially analysing the outputs from Leapfrog and AQTESOLV.
- Exploratory data analysis (EDA) including statistical analysis and machine learning (ML) were then used to challenge the null hypothesis and seek support for the alternative hypothesis: that the stress regime has enhanced the rock permeability along NE-SW normal faults. Pandas, Numpy, Seaborn and Scikit-Learn are all open source data analysis libraries built on top of the Python programming language. Jupyter Notebook was used to access these Python libraries and clean the datasets and undertake statistical analysis, data visualisation and machine learning.

We also undertook unsupervised, multivariate machine learning analyses of the Project major ion water chemistry dataset. The objective of these analyses was to identify any indicators of groundwater-surface water connectivity during the three constant rate discharge (CRD) tests that targeted specific faults. Two different unsupervised ML algorithms were utilised, namely:

- K-means cluster analysis (KCA); and
- Principal component analysis (PCA).

Data Science Workflow and Results

A geological resource model was built in Leapfrog, using data from the exploration drilling program and targeted hydrogeological investigations. This model contained the main dolomitic and calcareous sedimentary rock formations, as well as the rhyodacite porphyry unit. All of the main geological formations are heavily faulted. In general, the faults tend to be oriented in four main directions: a conjugate pair of N-S striking faults and E-W striking faults; and a separate conjugate pair of NE-SW striking faults and NW-SE striking faults. In total, over 100 faults have been mapped within 10 km of the mine site.

To explore the influence of faults on groundwater inflows to open cut pits, the FS field investigation undertook targeted test pumping at three main fault locations. At each of these three locations, the test pumping bore and observation bores were constructed to intersect these mapped and modelled fault structures in the near vicinity of the proposed pits. These three CRD tests along faults complemented three previous Pre-FS CRD tests that tested the bulk rock transmissivity and storage properties away from faults.

Confined aquifer analytical radial flow solutions were primarily used in the analysis of the pumping test data. Generally, confined behaviour was observed through the timing and nature of the observation bore responses to pumping, which reflected low values of elastic storage. The approach to the analysis was sequential, starting with a simplified assumption of infinite acting radial flow to the pumped well and adding complexity where necessary. Table 1 summarises the K estimates that were derived for each major rock formation with respect to each aquifer test type. This aquifer testing forms the basis for our subsequent interpretation.

The diagnostic flow plots (time drawdown and drawdown derivative plotted together) aided in confirming the aquifer flow regimes and identifying key boundary conditions at two aquifer test locations (Figure 1). At one test location, barrier boundaries were identified in multiple observation bores at approximately 550 m from the pumped bore, which is the approximate distance to the nearby porphyry intrusion, a site aquitard. At the other test pumping location, a constant source recharge boundary was identified at 250-300 m from the pumping bore, which was the approximate distance to the regional perennial river. The HCM was modified to reflect the aquitard barrier boundary and potential for hydraulic connection with the river.

***Table 1** Summary of K estimates per HSU and Aquifer Test Type*

Test type	HSU	Count	Mean m/day	St.Dev m/day	25th Percentile m/day	75th Percentile m/day	Min m/day	Max m/day
CRD test drawdown	HSU1	1	34	-	-	-	-	-
	HSU2	24	14	17	0.7	19	0.01	56
	HSU3	3	2	3	0.7	3	0.3	5
CRD test recovery	HSU1	1	12	-	-	-	-	-
	HSU2	7	29	34	4	50	1	82
	HSU3	2	2	0	2	2	2	2
Slug test	HSU2	31	12	17	1	14	$3x10^{-6}$	60
	HSU3	6	3	2	2	4	0.	4

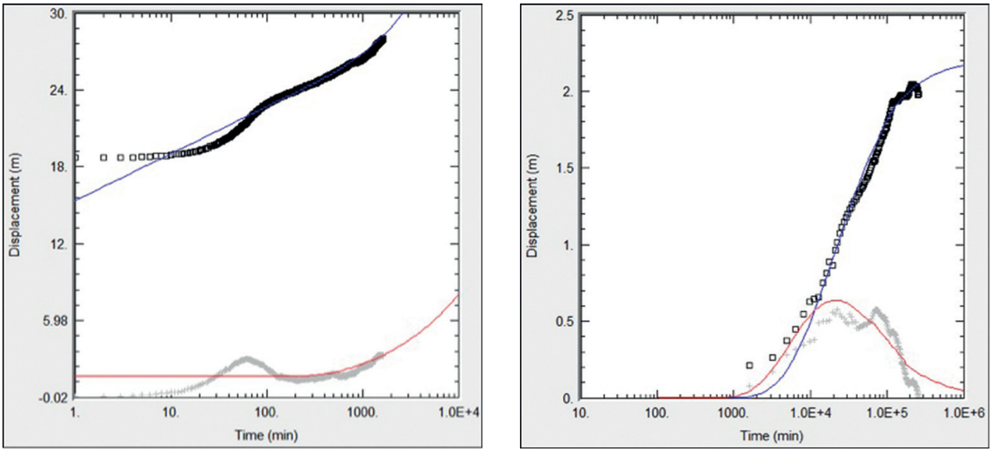

***Figure 1** Characteristic diagnostic flow plots showing: a) presence of no flow barrier associated with porphyry intrusion (Left); and b) presence of constant head recharge boundary associated with hydraulic connection to a regional-scale perennial river (Right).*

To enable spatial analysis of the K data, the solid geology and fault architecture was extracted from the Leapfrog geology model and imported into QGIS. Geospatial analysis included:

- Each fault was characterised based on its broad orientation into one of the four main spatial types and converted from a line to an array of 1m-spaced points.
- The distance between each K data point and the nearest fault point from each of the four fault groups was calculated.

The corresponding K-fault distance dataset was imported into Pandas to create a dataframe for hypothesis evaluation. Figure 2 presents the K distributions for each type of hydraulic test.

Figure 2a presents the K data with respect to the host bedrock geology. This plot shows that the K distributions are similar across different geological formations, except for the contact with the porphyry intrusion which acts as a regional aquitard. Geological formations that are dominantly fine-grained (siltstones and mudstones) have similar K to formations that are expected to be more permeable which highlights the strong influence of secondary porosity through fractures and karst. Also, of note is that the similar ranges of K values from different test types suggests that localised permeability features encountered in slug tests translate to macro-scale features encountered at the pump test scale (i.e. local-scale fracturing is interconnected at the macro scale). Figure 2b presents the K data with respect to geographic location and shows that higher K features are encountered more often at the South Pit area.

Figure 3 presents a series of jointplots showing the relationship of K to distance from each of the four fault type groups. These plots consider the probability distribution for the K data and the distance of that K data from the given fault type. Each of the four fault types have a good representation of nearfield and farfield K data to provide spatial context on whether the faults are influencing the permeability structure within the aquifers. The top two plots present the first conjugate pairs of faults being N-S and E-W orientations. Despite having a high number of data points close to these fault features, the probability distribution shows that a large number of the particularly high K points are situated away from these faults, and only a small number of high K points are within 50-75 m of the fault. It does not appear that these N-S and E-W faults are influencing permeability distribution.

In stark contrast, the NE-SW and NW-SE faults (the bottom two plots in Figure 3) show a much stronger relationship between proximity to faults and higher K. Figure 4 evaluates this relationship by area, and demonstrates that the NE-SW oriented faults (left plot) have higher K values in the South

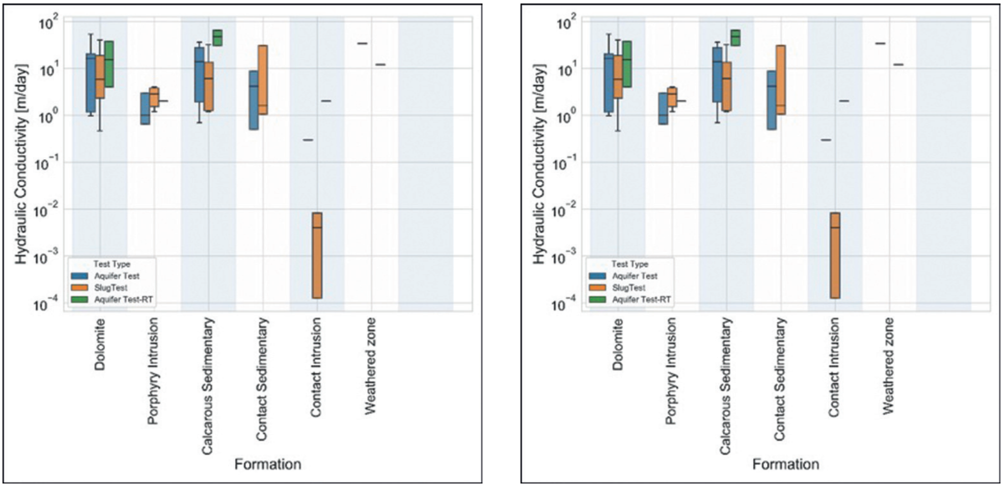

***Figure 2** Boxplots of hydraulic conductivity for each test type: a) categorised by type of bedrock geology (Left); and b) by geographic location with respect to vicinity to the nearest open cut pit.*

Pit and West Pit areas within 50 m of faults, as compared to the background fracture K which tended to be <10 m/day. The NW-SE transfer faults do not exhibit this same relationship and the proximity to these faults is not an explanatory variable for higher K within a reasonable fault damage zone corridor (i.e. <50 m).

As an independent line of evidence, both KCA and PCA were conducted sequentially to interrogate water quality data. Three main groups of groundwater quality were obtained through the KCA, and these groups broadly aligned with the two main hydrochemical facies being Ca-SO4 type water north of the porphyry intrusion and Ca-HCO3 type water south of the intrusion. There are a number of potential surface water-groundwater processes that can be observed from the PCA analysis:

- South Pit observation bore chemistry is similar to the nearby perennial river sample prior to and after the CRD test.
- The Northeast Pit and West Pit bores transition towards the perennial river sample chemistry towards the ends of the tests. This indicates potential for pumping-induced river water recharge, suggesting that the fault systems may be connected to the river.
- The South Pit pumping bore transitions from a water quality that resembles a

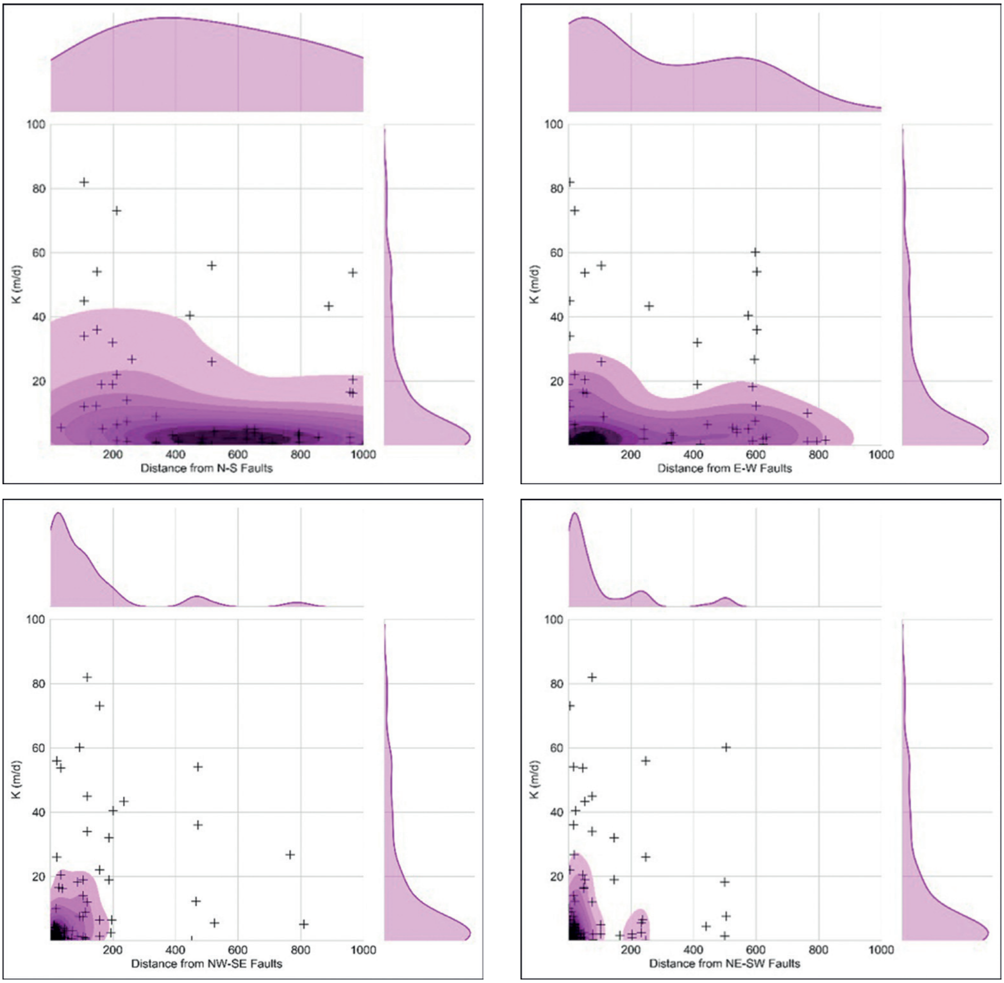

Figure 3 *Jointplots showing the relationship of K to distance from each of the four fault types: a) Distance from N-S faults (Top Left); b) Distance from E-W faults (Top Right); Distance from NE-SW faults (Bottom Left); and d) Distance from NW-SE faults (Bottom Right).*

nearby pit lake to groundwater that is observed in two bores situated more than 1 km to the west.

Conclusions

In our examples, EDA and ML techniques facilitated by Python scripting were used to integrate data from aquifer testing, geological modelling, structural interpretation, and hydrochemistry analysis to achieve a more wholistic understanding of groundwater flow dynamics along structural features and improve both the HCM and resulting estimates of pit inflows and groundwater drawdown distribution. This improved aquifer characterisation has identified which styles of faults are likely to be more transmissive and in doing so provided a more robust basis for estimating ranges of likely groundwater inflows and testing of mitigation measures such as advanced dewatering strategies

Proper application of these techniques is only possible when directed by an experienced hydrogeologist with a keen understanding of the HCMs, through problem formulation directed at specific data gaps.

Acknowledgements

The authors thank fellow Golder hydrogeologists Dr Scott Weeks, Jaco Breytenbach and Nick Coulson for support during the analysis and interpretation of the data.

References

Caine JS, Evans JP, Forster CB (1996) Fault zone architecture and permeability structure. Geology 24(11): 1025-1028.

McCallum J, Simmons C, Mallants D, Batelaan O (2018) Simulating the groundwater flow dynamics of fault zones. Report submitted to Australian Department of Environment and Energy. National Centre for Groundwater Research and Training. 52pp.

Office of Groundwater Impact Assessment (2019) Underground Water Impact Report for the Surat Cumulative Management Area. Department of Natural Resources, Mines and Energy, State of Queensland. 272pp.

Scibek J, Gleeson T, McKenzie JM (2016) The biases and trends in fault zone hydrogeology conceptual models: global compilation and categorical data analysis. Geofluids (2016) 16: 782-798.

Van der Plas J (2017) Python data science handbook: Essential tools for working with data. O’Reilly, Boston, 548 pp.

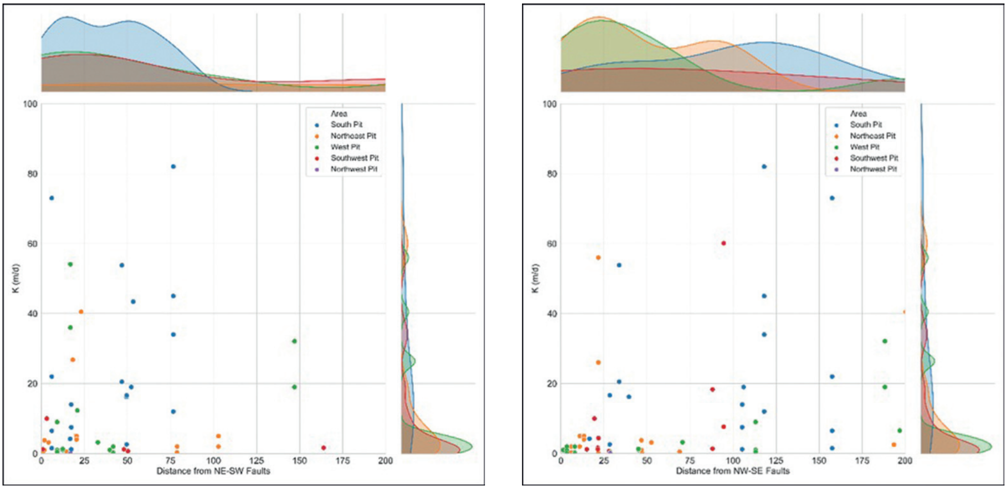

Figure 4 Jointplot showing the relationship of K to distance from: a) NE-SW faults (Left), and b) NW-SE faults (right). Note: only shows K data within 200m of this fault type