Machine-Learning Analysis of Factors Influencing Induced Seismicity Susceptibility in the Montney Play Area, Northeastern British Columbia (NTS 093P, 094A, B, G, H)

A. Amini, Department of Earth, Ocean and Atmospheric Sciences, The University of British Columbia, Vancouver, British Columbia, aamini@eoas.ubc.ca

E. Eberhardt, Department of Earth, Ocean and Atmospheric Sciences, The University of British Columbia, Vancouver, British Columbia

Amini, A. and Eberhardt, E. (2021): Machine-learning analysis of factors influencing induced seismicity susceptibility in the Montney play area, northeastern British Columbia (NTS 093P, 094A, B, G, H); *in* Geoscience BC Summary of Activities 2020: Energy and Water, Geoscience BC, Report 2021-02, p. 45–56.

Introduction

Unconventional gas resources represent an emerging low-cost, clean-burning energy source, the export of which presents both a greener transition option to replace more carbon-intensive fossil fuels like coal and a key economic opportunity for British Columbia (BC) and Canada. In northeastern BC, new discoveries and advancements in extraction technologies have led to resource estimates of 94.5 trillion m^3 (3337 tcf) of gas-in-place (BC Oil and Gas Commission, 2018), enough to support development and export operations for more than 150 years. However, with the development of these new resources comes new challenges. Amongst these are public, First Nations and regulator concerns regarding induced seismicity associated with hydraulic fracturing and wastewater injection operations. Both activities involve the injection of large volumes of fluids into deep geological formations, which serve to create localized increases in pore pressures and stress changes acting on critically stressed faults, resulting in fault slip and induced seismicity. Notable induced events in northeastern BC include one magnitude (M) 4.4 and two M 4.6 events between 2014 and 2018.

In response to these events, and other environmental concerns, the BC government appointed a scientific panel to review hydraulic fracturing practices and their impacts (Scientific Hydraulic Fracturing Review Panel, 2019). In their review, a key knowledge gap was identified in relation to induced seismicity susceptibility. In particular, the effects of different geological and operational factors on the spatial and temporal distribution of events are not well understood and vary in importance for different unconventional gas plays. Although separating geological from operational factors is a complex task, it is also recognized that a massive amount of geological, operational and seismic data are being collected from hydraulic fracturing activities for which robust analysis methods are needed. The rapid development of multivariate statistical and machine-learning techniques to analyze large datasets makes the application of these techniques to this problem especially attractive and conducive, although there is not a lot of experience yet in applying these to induced seismicity hazard assessments (i.e., likelihood, severity, etc.), especially in analyzing both operational and geological parameters together. Distinguishing between these factors is of interest as the influence of geological factors on induced seismicity susceptibility for a given formation being targeted cannot be controlled or manipulated (outside of avoidance), whereas many operational factors (i.e., well completion related) can be controlled to some extent offering a means to potentially mitigate induced seismicity hazards for a susceptible formation.

Presented in this paper are the preliminary results of research (Geoscience BC project 2019-014) investigating the development of induced seismicity susceptibility maps to aid decision makers with their planning of hydraulic fracturing activities and managing of induced seismicity hazard. To accomplish this, machine-learning techniques will be integrated with mechanistic validation using controlled laboratory experiments and three-dimensional (3-D) numerical modelling (to account for cause and effect relationships). The results presented here are from the first phase of this work, the application of different machine-learning algorithms to determine the relative importance of several geological and operational parameters (termed feature importance) in relation to the triggering of induced seismicity. This is done for data compiled for the Montney Formation in northeastern BC. The algorithms applied and compared include the decision-tree, random-forest and gradient-boost methods. In addition to testing the robustness of these algorithms through a comparative analysis, guidance is provided in the use of machine learning to identify influencing factors as a step toward developing induced seismicity susceptibility maps.

This publication is also available, free of charge, as colour digital files in Adobe Acrobat® PDF format from the Geoscience BC website: http://geosciencebc.com/updates/summary-of-activities/.

Background

A significant increase in the seismicity rate in western Canada in recent years has been associated with the development of unconventional oil and gas resources, including the related activities of hydraulic fracturing (Bao and Eaton, 2016) and wastewater disposal (Schultz et al., 2014). There are numerous operators conducting these activities, each using different operational parameters (e.g., fluid injection volumes and rates) tailored for the local geological setting and targeted formation, as well as further shaped by in-house objectives, experiences and optimization efforts. This raises the question of what are the cause and effect relationships of these parameters on induced seismicity susceptibility and magnitude distribution? The question of susceptibility addresses the likelihood that a particular well will generate induced seismicity; this can be viewed as a classification problem (seismogenic or not seismogenic). The question of magnitude distribution addresses the potential severity.

With respect to operational parameters, it has been argued that the moment release attributable to induced earthquakes is related to the net volume of the injected fluid, with empirical trends established that link an upper limit for the moment magnitude to injection volume (Hallo et al., 2014; McGarr, 2014). The data analyzed in these studies included a mix of hydraulic fracturing and wastewater disposal activities in sedimentary rocks and enhanced geothermal-development activities in crystalline rocks, combining data from Europe, the United States and Australia. Weingarten et al. (2015) carried out a similar study combining information on injection wells from public databases with available earthquake catalogues and concluded that injection rate is the most important operational parameter affecting induced seismicity. Their study focused on data from hydraulic fracturing and wastewater disposal activities in the eastern and central United States. For the Western Canada Sedimentary Basin (WCSB), where the Montney Formation is situated, different studies have shown that injection volume is associated with induced seismicity (Schultz et al., 2014; Babaie Mahani et al., 2017). Schultz et al. (2018) investigated the relationship between injection parameters and induced seismicity in the Duvernay shale play in Alberta and concluded that events are associated with completions that used larger injection volumes and that seismic productivity scales linearly with injection volume. Their analysis further showed that the wellhead injection pressure and rate have an insignificant association with seismic response, and that geological factors account for the variability in induced seismicity susceptibility observed in the region.

With respect to geological parameters, Göbel (2015) compared several fluid injection operations in California and Oklahoma and examined the temporal and spatial variations in their induced seismicity responses. His results suggest that operational parameters for fluid injection are likely of secondary importance and that the primary controls on seismicity induced by injection are the site-specific geology and geological setting. Van der Baan and Calixto (2017) compared current and historic seismicity rates in six states in the United States and three Canadian provinces to past and present oil and gas production. Their study showed that in addition to injection volumes, local- and regional-scale geology and tectonics influenced earthquake hazard susceptibility. Amini and Eberhardt (2019) similarly compared induced seismicity and well data for several key North American unconventional gas plays, with a focus on magnitude distribution relative to differences in the tectonic in situ stress regime. They found that stress regime has a significant influence on event magnitude with a thrust fault stress regime, as exists in parts of the Montney play area, being more susceptible to large magnitude events compared to a strike-slip or normal fault stress regime.

In Oklahoma, Shah and Keller (2017) combined geophysical and drillhole data to map subsurface geological features in the Precambrian crystalline basement and found that most induced seismicity events are located where the crystalline basement is likely composed of fractured intrusive or metamorphic rock; areas of extrusive rock or thick sedimentary cover (>4 km) exhibited little induced seismicity. They concluded that the differences in seismicity may be due to variations in permeability structure; within intrusive rocks, fluids can become narrowly focused in fractures and faults, causing a concentrated increase in local pore fluid pressures, whereas more distributed pore space in sedimentary and extrusive rocks may relax pore fluid pressures. Hincks et al. (2018) developed an advanced Bayesian network to model joint conditional dependencies between spatial, operational and seismicity parameters in Oklahoma. They found that injection depth relative to crystalline basement most strongly correlates with seismic moment release and that the combined effects of depth and volume are critical, as injection rate becomes more influential near the basement interface. Similar findings were reported by Skoumal et al. (2015) and Currie et al. (2018) for hydraulic fracturing operations in Ohio. The latter showed that seismicity occurred along faults below the injection interval in the crystalline basement. From seismic reflection lines, they showed that these fault systems intersected the injection interval targeted by the well, providing permeability pathways for fluid pressure increases leading to fault slip.

Specific to the geology of the WCSB, Schultz et al. (2016) found that hypocentres of induced seismicity clusters in Alberta coincided with the margins of the Devonian carbonate reefs and interpreted this spatial correspondence as the result of geographically biased activation potential, possibly as a consequence of reef nucleation preference to paleobathymetric highs associated with Precambrian basement

tectonics. Their work provided evidence that in some areas Paleozoic and Precambrian strata are likely to be in hydraulic communication, which points to the important role of regional- and local-scale geological factors in the nature of induced seismicity. Eaton and Schultz (2018) also suggested natural processes involving the transformation of organic material (kerogen) into hydrocarbons and cracking to produce gas can cause fluid overpressures resulting in an increased susceptibility to induced seismicity. They presented two examples from the WCSB where induced seismicity attributed to hydraulic fracturing is strongly clustered within areas characterized by high pore-pressure gradients.

The above examples highlight the importance of different operational and geological factors on induced seismicity, but do so from the perspective of studying the influence of a single factor. It is unlikely, however, that a single causative factor is solely responsible for an induced seismicity event. Instead, multiple factors can play an influencing role. Therefore, it is important to consider and understand the cause and effect relationships of different operational and geological factors on the spatial and temporal distribution of induced seismicity events. However, this is not a simple task and requires probing a wide variety of linear and nonlinear associations and interaction terms between factors affecting induced seismicity without assuming a priori knowledge on the nature of the relationships between these factors.

The use of machine learning and data analytics are quickly evolving as a means to identifying hidden patterns and extracting information from large datasets. In the geosciences and rock engineering, they have been applied to predicting rockburst potential in deep mines (Ribeiro e Sousa et al., 2017; Pu et al., 2018) and squeezing behaviour in deep tunnels (Sun et al., 2018), as well as developing geological maps using remote sensing data (Cracknell and Reading, 2014) and analyzing data from rock testing (Millar and Clarici, 1994) and blasting (Liu and Liu, 2017). In the context of earthquake seismology, machine learning has been applied to a variety of problems such as laboratory earthquake identification (Rouet-Leduc et al., 2017) and forecasting (Panakkat and Adeli, 2009). Building on these studies, it is recognized that a massive amount of geological, operational and seismic data are being collected with hydraulic fracturing activities, and the size and complexity of these datasets have made traditional empirical and statistical analyses inefficient and ineffective. This has led to recent studies by Pawley et al. (2018) who combined tectonic, geomechanical and hydrological data with induced seismicity data, related to hydraulic fracturing operations in the Duvernay play in Alberta, to train a logistic regression algorithm to map and develop an induced seismicity potential map. Their results suggest that the proximity to basement, formation overpressure, minimum horizontal stress, proximity to reef margins, lithium concentrations and natural seismicity rate are the dominant contributing factors/indicators to triggering induced seismicity within the study area. Zhang et al. (2020) used machine learning on real-time induced seismicity data to locate small events in Oklahoma by accessing seismic waveform data from a regional network. They designed a fully convolutional network (FCN), to predict a 3-D image of the earthquake location probability from a volume of input data recorded at multiple network stations. Their results showed that the designed system is capable of locating small events of local magnitude (M_L) ≥2.0 with a mean epicentre error of 4 to 6 km.

Data Compilation and Preparation

A database of 16 945 hydraulic fracturing stages from 1244 horizontal wells within the Montney Formation (from 2014 till end of 2016) was compiled and analyzed using multiple sources that reported well activities in northeastern BC (BC Oil and Gas Commission, 2018; geoLOGIC systems ltd., 2019). This was combined with a second database that included a comprehensive earthquake catalogue compiled for northeastern BC and western Alberta (Visser et al., 2017). This was produced specifically to study induced seismicity in this region and consists of 4916 events for the period of January 2014 to December 2016 with a magnitude of completeness (M_L) of 1.8.

To prepare the data for analysis using a supervised machine-learning algorithm (as discussed in the next section), it was necessary to determine the output labels. Here, induced seismicity was considered as a binary-classification problem with respect to the observed seismic activity. Wells were classified as being either 'aseismic' or 'seismic' based on spatial and temporal correlations with hydraulic fracturing operations. This was done by cross-correlating the earthquake catalogue with the well database and applying a series of spatial and temporal filters to identify the subset of earthquake events that are likely induced seismicity events related to hydraulic fracturing. The first step was to clip the data to only include earthquakes located within the boundaries of the Montney play area in northeastern BC and to filter out events spatially associated with anthropogenic activities that are not related to oil and gas activities, such as those from mining and construction (e.g., blasting). This step reduced the total number of events being considered from 4916 to 2867. Next, a spatial filter was applied to search for all event locations that were within a 5 km radius of an active hydraulic fracturing well. The 5 km radius represents the uncertainty in the event location accuracy reported for the earthquake catalogue. To this, a three-month temporal filter was applied (see Atkinson et al., 2016). Thus, if the epicentre of an earthquake event was recorded as occurring within 5 km from the surface location of an active well and within three months from the start date of the

hydraulic fracturing activity, it was considered here to be an induced seismicity event and the corresponding well was classified as being seismogenic. This resulted in a subset of 543 events identified as induced seismicity events.

The input parameters for the machine-learning analysis, referred to herein as features (using the term common to machine learning), were selected from available geological and operational data for the Montney Formation. For the geological features such as distance to basement, if data was not available for a given well, values were interpolated using the average of the three closest wells. The operational features were treated differently as only wells that had complete data throughout all features were included. This resulted in data for 11 415 stages out of 16 945 being used for the machine-learning analysis. The input features are described in Table 1.

The interpolated pore-pressure gradients ranged from 5.4 to approximately 18 kilopascals per metre (kPa/m). In addition to the reservoir pore pressure, information regarding the maximum horizontal stress (S_{Hmax}) direction was included. This was calculated for each well based on S_{Hmax} azimuths extracted from the World Stress Map database (Heidbach et al., 2018) and interpolated for each well. Two different values were investigated: 1) the difference between the local S_{Hmax} and the horizontal well azimuths; and 2) the difference between the local and regional S_{Hmax} azimuths, where N45°E was assumed to be the regional S_{Hmax} direction in the Montney play area. Injection depth (total vertical depth [TVD]) was also considered as a proxy for the magnitude of stresses in this region. Specific to the local geology, the vertical distance between the injection depth and the top of the Montney and Debolt formations were considered, together with the distance to the Precambrian crystalline basement. For these, a negative value indicates an injection depth above the formation top/basement and a positive value refers to below the formation top. It should be noted that there is a high degree of uncertainty in the interpolated values for the top of the basement due to a lack of direct borehole measurements (from vertical wells). Lastly, the two-dimensional (2-D) distance from the well to the closest mapped fault (Hayes et al., 2021) was included. This was taken as the shortest horizontal distance between the wellhead and closest fault. In this analysis, no cutoff value for distance to fault was considered.

Machine-Learning Algorithm Development

Machine learning can be undertaken using supervised or unsupervised algorithms. Supervised learning is where the input features and an output result are given, and an algorithm is used to learn the mapping function between these. The goal is to approximate the mapping function so well that for any new input data, the output can be predicted for that specific data. This contrasts with unsupervised learning where only the input data is known, and no corresponding output variables are given. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. For this study, supervised learning was used for the initial data analysis to identify which wells were associated with induced seismicity and which were not. These represent the correct answers to the classification problem for training the mapping function; the corresponding data associated with each set of wells is referred to as the training data.

Three different supervised machine-learning algorithms were used that are generally considered to be robust for classification problems: decision tree, random forest and gradient boost (Hastie et al., 2017). These methods were chosen because of the ease of interpretability of their results and also because they are not sensitive to the scale of input data. The objective of the algorithm is to iteratively make predictions on the training data and to correct these until the algorithm achieves an acceptable level of performance.

Decision trees are a nonparametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision trees have two advantages: the resulting model can easily be visualized and understood by non-experts, and the algorithms are completely invariant to scaling of the data. As each feature is processed separately, and the possible splits of the data do not depend on scaling, no preprocessing of features is needed for decision-tree algorithms. The main limitation of decision trees is that they tend to over fit the data and provide poor generalization performance.

A random forest is essentially a collection of decision trees, where each tree is slightly different from the others. In a random forest each tree might do a relatively good job of predicting but will likely over fit part of the data. To reduce the amount of overfitting, many trees are built, all of which work well and over fit the data in different ways, and the results are averaged.

The gradient-boost regression tree is another ensemble method that combines multiple decision trees to create a more powerful model. This can be used for both regression and classification. In contrast to the random-forest approach, gradient boosting works by building trees in a serial manner, where each tree tries to correct the mistakes of the previous one. The main idea is to combine many simple models (known as weak learners) that can provide good predictions on parts of the data, and so more and more trees are added to iteratively improve performance.

All three machine-learning models were built using scikit-learn, a Python library for machine learning (Pedregosa et al., 2011). The data was divided into training and validation

Table 1. Input features used for machine-learning analysis of induced seismicity in the Montney play area, northeastern British Columbia.

Feature category	Feature name (units)	Abbreviation	Source	Number of available data points	Lower limit	Upper limit	Median
Geological	Pore-pressure gradient (kPa/m)	PP_grad	BC Oil and Gas Commission (2018)	2252	5	18	12
Geological	Local and regional S_{Hmax} azimuth difference (deg)	Az_diff_L_R	Heidbach et al. (2018)	601	0	12	5
Geological	Distance from injection depth to top of Montney Fm. (m)	D_Mont	BC Oil and Gas Commission (2018), geoLOGIC Systems ltd. (2019)	8232	-100	420	78
Geological	Distance from injection depth to top of Debolt Fm. (m)	D_Deb	BC Oil and Gas Commission (2018), geoLOGIC Systems ltd. (2019)	2183	-478	9	-247
Geological	Distance from injection depth to basement (m)	D_Base	BC Oil and Gas Commission (2018), geoLOGIC Systems ltd. (2019)	28	-2450	-130	-1660
Geological	Distance from wellhead to faults (m)	Dist_F	Hayes et al. (2021)	16945	2	70400	1770
Operational	Local S_{Hmax} and horizontal well azimuth difference (deg)	Az_diff_L_W	BC Oil and Gas Commission (2018), geoLOGIC Systems ltd. (2019)	16945	0	88	42
Operational	Injection depth (m)	Inj. Depth	geoLOGIC Systems ltd. (2019)	16945	1312	3416	2052
Operational	Well completion length (m)	Comp_Len	geoLOGIC Systems ltd. (2019)	16945	90	3830	1610
Operational	Maximum injection pressure (MPa)	Max_P	geoLOGIC Systems ltd. (2019)	13889	4	99	56
Operational	Average injection rate (m^3/min)	Rate	geoLOGIC Systems ltd. (2019)	15472	1	22	8
Operational	Stage injection volume (m^3)	Volume	geoLOGIC Systems ltd. (2019)	16603	1	5640	592

Abbreviations: kPa, kilopascal; Mpa, megapascal; S_{Hmax}, maximum horizontal stress

sets accounting for 75% and 25% of the full dataset, respectively. The training dataset was further divided into training and test sets, which were used to train the algorithm using 50-fold cross validations, with the training set accounting for 98% of the training dataset and the test set for 2% at each cross-validation run.

The training data was used to train and evaluate the optimum tree depths of the three algorithms using 50-fold cross validations. Figure 1 shows the results of 50-fold cross validations for each algorithm. At each run, the accuracy of the model for a specific tree depth is calculated. In these figures the orange line represents the accuracy of the training set. The blue line shows the mean cross-validation accuracy and the shaded area represents the confidence interval (±2 standard deviations) for the calculated means. For these plots, an accuracy of 1 represents 100% accuracy. This determines if the training set is over fitted, and alongside this, it determines the optimal tree depth based on the confidence interval. Based on the cross-validation results, the tree depths of 16, 12 and 8 were chosen for the decision-tree, random-forest and gradient-boost algorithms, respectively.

Machine-Learning Results

Feature Importance

The results of the cross validations using each algorithm were further analyzed to investigate the importance of each feature on the classification outcome. Figure 2 shows the feature importance calculated using each of the three classification algorithms. The bars are colour coded to differenti-

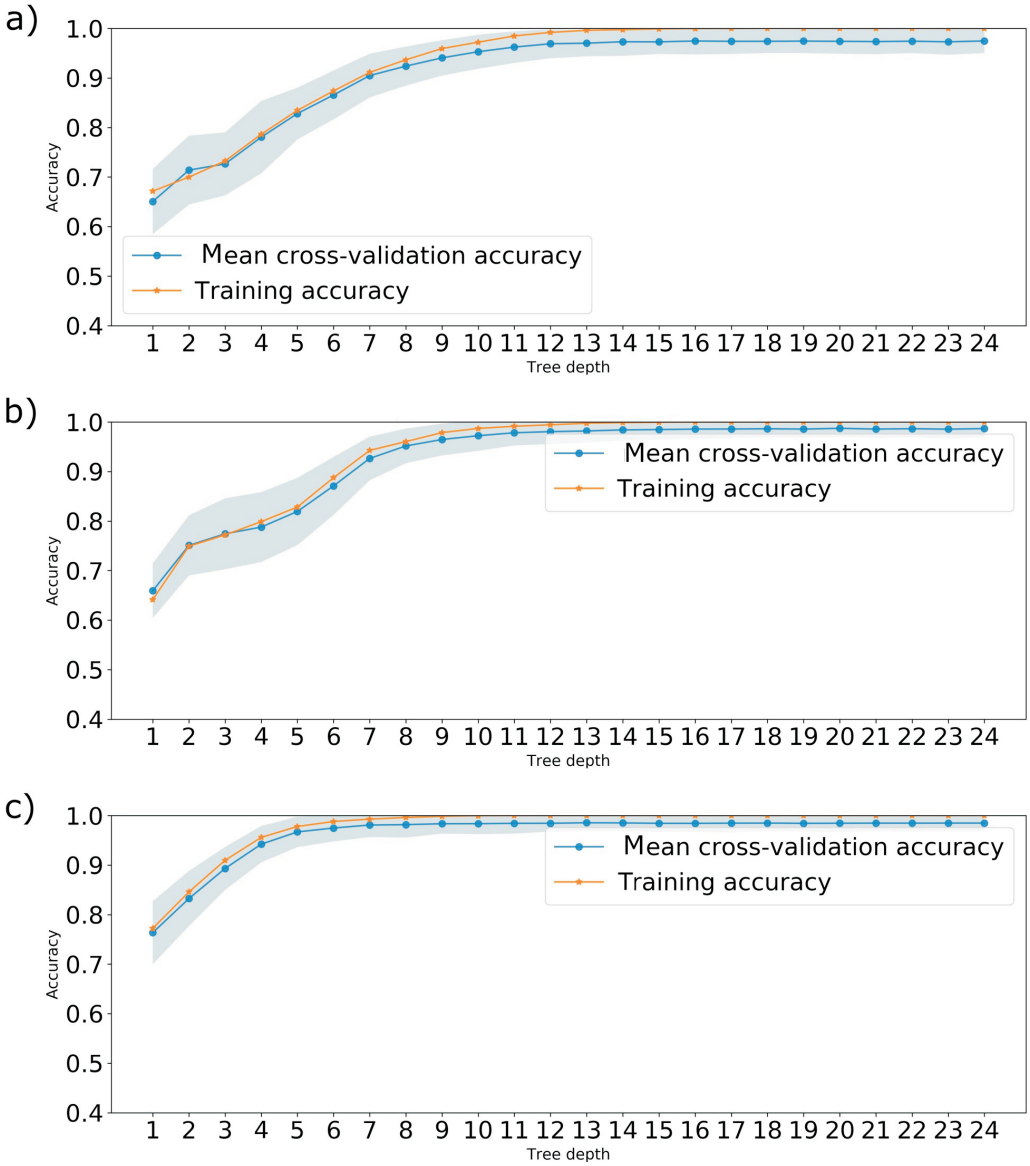

Figure 1. Results of the 50-fold cross validations to determine the optimum tree depth for each algorithm: **a)** decision tree, **b)** random forest and **c)** gradient boost. An accuracy of 1 represents 100% accuracy.

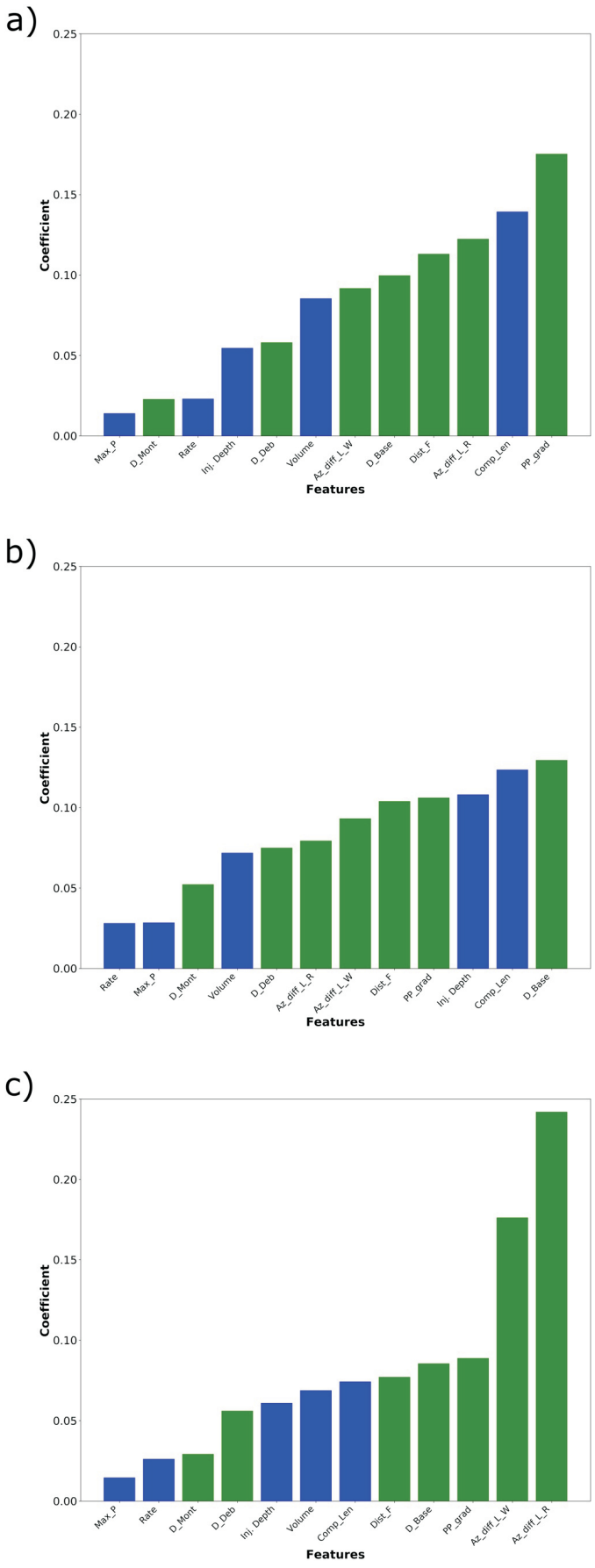

Figure 2. Feature importance calculated using three different supervised machine-learning algorithms: **a)** decision tree, **b)** random forest and **c)** gradient boost. Blue indicates operational features and green indicates geological features. The coefficient does not have a physical meaning and is compared based on the relative values. Feature abbreviations: Az_diff_L_R, local and regional maximum horizontal stress azimuth difference; Az_diff_L_W, local maximum horizontal stress and horizontal well azimuth difference; Comp_Len, well completion length; D_Base, distance from injection depth to basement; D_Deb, distance from injection depth to top of Debolt Fm.; D_Mont, distance from injection depth to top of Montney Fm.; Dist_F, distance from wellhead to faults; Inj. Depth, injection depth; Max_P, maximum injection pressure; PP_grad, pore-pressure gradient; Rate, average injection rate; Volume, stage injection volume.

ate features that relate to the geology from those that are operational. The importance of each feature is indicated as a coefficient; these coefficients do not have a physical meaning and are compared based on the relative values and not the absolute values.

Based on the results from the decision-tree analysis, the high importance features were determined to be pore-pressure gradient, distance to basement, well completion length, azimuth difference between the local and regional S_{Hmax} orientation and distance to faults. Four of these five features also form the top five ranked features from the random-forest analysis, although in a slightly different order and with injection depth replacing azimuth difference between the local and regional S_{Hmax} as being of higher importance. For the gradient-boost analysis, again four of these features were ranked in the top five, the exception being that this model showed a higher sensitivity to the horizontal well direction than the completion length. The gradient-boost model also showed very high sensitivity to azimuth difference between the local and regional S_{Hmax} and azimuth difference between local S_{Hmax} and horizontal well direction compared to other features.

Overall, the features consistently ranked as being highly influential by all three machine-learning algorithms were pore-pressure gradient, distance to faults and distance to basement. In all models the same groupings of operational features were observed; injection rate and maximum injection pressure were ranked lowest in importance, and injection volume ranked in the middle. This is an interesting result because injection rate and volume are often cited as operational features that have a significant influence on induced seismicity (McGarr, 2014; Schultz et al., 2018). This appears to hold partly true in the case of injection volume, but operational features such as well completion length and injection depth, which have not been thoroughly studied, appear to have a stronger correlation with induced seismicity.

Model Validation

To validate the trained algorithms, they were next applied to the test data that was set aside to evaluate each model's performance. The test data comprised 25% of the full dataset and was not previously used to train the algorithms. To evaluate the performance of each algorithm, a confusion matrix was calculated. Also known as an error matrix, the confusion matrix allows visualization of the performance of a supervised machine-learning algorithm by reporting the number of true and false positives and true and false negatives. These are based on predictions using the test data relative to the mapping functions determined from the training data. In this case, a true positive would be a correct prediction that a well is associated with induced seismicity and a true negative would be a correct prediction that the well is not. Similarly, a false positive would be the incorrect prediction of a well being associated with induced seismicity where there was none, and a false negative would be an incorrect prediction of a well not being associated with induced seismicity when it was.

The results from calculating a confusion matrix for each algorithm are shown in Figure 3. Comparing these, the random-forest and gradient-boost classifiers performed slightly better than the decision-tree classifier. However, all three models performed with a very high accuracy (97–98%).

To further interpret the results, a SHapley Additive exPlanations (SHAP) analysis was run. The SHAP is a game theory approach used to help interpret predictions from complex models, for example the output from machine-learning models. The SHAP assigns each feature an importance value for a particular prediction and shows there is a unique solution for each class of additive feature importance that adheres to desirable properties (Lundberg and Lee, 2017). The SHAP TreeExplainer tool is a subcategory of SHAP that is specifically built for interpreting tree models, such as decision trees and random forests. The SHAP value plot can show the positive and negative relationships of the predictors with the target variable. The analysis presented here is for the random-forest model as it performed slightly better than the decision-tree model.

Figure 4 presents the summary plot from the SHAP analysis, which combines feature importance with feature impact. Based on this plot, the following information can be gained. First, each feature is ordered according to its importance (starting with the most important at the top). Note that the SHAP plot is calculated for one instance of the random-forest model, whereas the ranking of feature importance in

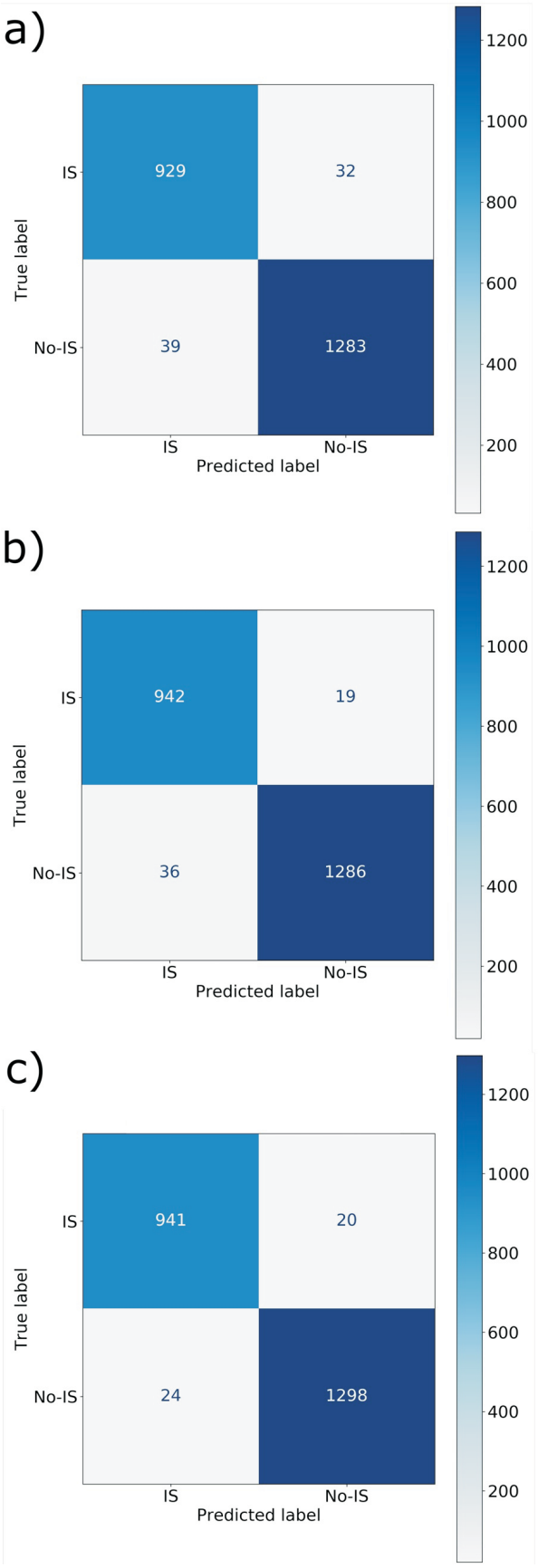

Figure 3. Confusion matrices comparing predicted versus actual results for the **a)** decision-tree, **b)** random-forest and **c)** gradient-boost trained models. Abbreviations: IS, induced seismicity events; No-IS, no induced seismicity events.

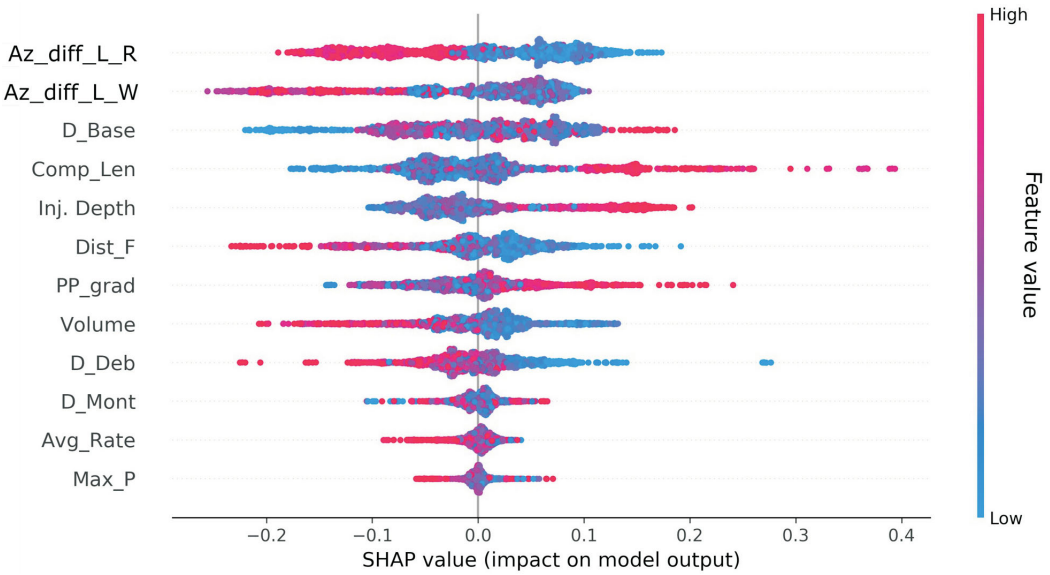

Figure 4. Results for a SHapley Additive exPlanations (SHAP) feature importance analysis using the random-forest trained model. Feature abbreviations: Az_diff_L_R, local and regional maximum horizontal stress azimuth difference; Az_diff_L_W, local maximum horizontal stress and horizontal well azimuth difference; Comp_Len, well completion length; D_Base, distance from injection depth to basement; D_Deb, distance from injection depth to top of Debolt Fm.; D_Mont, distance from injection depth to top of Montney Fm.; Dist_F, distance from wellhead to faults; Inj. Depth, injection depth; Max_P, maximum injection pressure; PP_grad, pore-pressure gradient; Rate, average injection rate; Volume, stage injection volume.

Figure 2 is based on an averaging of 50-fold cross-validation runs. Thus, the order of feature importance between the two is slightly different. Next, points are plotted to show the distribution of the SHAP values using colour to represent the feature value and stacking of overlapping points in the y-axis direction to give a sense of the distribution of the SHAP values. From this, the impact (both positive and negative) is shown through the horizontal location of stacking, which shows whether the effect of that value is associated with a higher or lower prediction. This can be compared to whether the value for that variable/observation is high (red) or low (blue). For example, it can be seen that high values of completion length have a high positive impact on the quality rating. The high values related to this feature are indicated by the red colour of the points, and the high positive impact is shown by its extent on the x-axis.

The results of the SHAP feature importance analysis of the random forest model help to validate the meaningfulness of the algorithm results. Inspecting both Figures 2 and 4, it can be seen that key influencing features such as completion length, pore-pressure gradient and injection depth have a positive correlation with induced seismicity. The influence of pore pressures in the Montney Formation has been studied by Eaton and Schultz (2018), who demonstrated a positive relationship between overpressured areas and induced seismicity. The positive correlation of completion length is also valid as higher completion lengths correspond with larger stimulated volumes and therefore a higher probability of adversely interacting with a critically stressed fault. Features such as distance from the basement or distance to a known fault have negative correlations, meaning shorter distances between the injection point and basement or fault increase the likelihood of triggering an induced seismicity event.

Discussion

Machine-learning models are highly dependent on the quality and quantity of the input data. For the analyses presented here, for a feature where data was either limited or the spatial distribution and/or coverage of the data was sparse relative to the distribution of the wells, this was compensated for by using linear interpolation. However, large distances between points can reduce the accuracy of interpolation, as can the interpolation method itself (e.g., assigning linear versus nonlinear weightings). With time, as new data becomes available, including that for features not included in this study, the induced seismicity susceptibility model can be updated to improve its predictive capabilities.

Based on the results obtained, an interesting observation is the correlation of injection volume and SHAP values. As can be seen in Figure 4, high injection volumes have a negative correlation with triggering induced seismicity. This might be interpreted as high injection volumes reduce the risk of induced seismicity, which is counter to general experience. Thus, empirical analyses and machine-learning data correlations for feature analysis do have their limitations and should be constrained by an understanding of the physics of fault slip and induced seismicity mechanisms.

When comparing the ranking of feature importance, the decision-tree and random-forest models provided similar results. In these models, high importance features included the geological parameters of pore-pressure gradient, distance to basement and distance to faults. The operational features of well completion length and injection depth were also highly ranked by the random-forest model. For the gradient-boost method, the ranking was slightly different with the stress field showing greater influence, both with respect to the geological feature of local and regional S_{Hmax} azimuth difference and related operational feature of the horizontal well and local S_{Hmax} azimuth difference. These differences in ranking are related to how each algorithm works. Gradient-boost models are based on shallow trees (high bias, low variance) and they reduce error mainly by reducing bias. Bias is the simplifying assumptions made by a model to make the target function easier to learn. In contrast, decision-tree and random-forest models use fully grown trees (low bias, high variance) and they reduce the model's error by reducing variance. Variance is the amount that the estimate of the target function will change if different training data were used. For this problem, the source of bias is the number of features that are used for classification, and by including both operational and geological features, the overall bias tends to be less than that if just considering one or the other. The variance of the data can be calculated, and as shown in Table 2, it is higher for parameters such as injection volume and distance to faults whereas it is lower for the two features related to the in situ stress. Thus, this explains the differences between the gradient-boost results and those from the decision-tree and random-forest models.

The comparison of feature importance between the geological and operational parameters show that, overall, the geological parameters generally ranked higher in importance. In all models, the operational parameters of average injection rate and maximum injection pressure consistently ranked as being the least influential. It should be noted that the maximum injection pressure data analyzed was limited to the pressure values measured at the wellheads. Another parameter that is worth investigating is the bottom hole pressure (BHP), which is more applicable to the influence of injection pressure on triggering induced seismicity. The compilation and analysis of BHP data is the subject of ongoing research as part of this project. The only operational features ranked as being of high importance were the completion length and the horizontal well direction relative to the S_{Hmax} azimuth. As was shown in Figure 4, completion length had a positive correlation with seismicity and can be thought of in terms of increasing the volume of the formation being stimulated by hydraulic fracturing. The larger the stimulated volume, the higher the probability of intersecting a fault (directly or indirectly).

Table 2. Variance of features that differ in rank based on the algorithm used. The highly important features of the gradient-boost results (where local and regional maximum horizontal stress azimuth difference [Az_diff_L_R] and local maximum horizontal stress and horizontal well azimuth difference [Az_diff_L_W] are ranked in the top five) are compared with distance to faults (Dist_F) and stage injection volume (Volume). The values of variances reported are divided by the mean for each feature in order to make them unitless for comparison.

Feature	Az_diff_L_R	Az_diff_L_W	Dist_F	Volume
Variance/mean	17	2	19110	370

Conclusions

The application of machine learning was investigated for the purpose of ranking the influence of geological and operational parameters on the classification problem of induced seismicity susceptibility (i.e., distinguishing between wells that are associated with induced seismicity and those that are not). Three different algorithms, decision tree, random forest and gradient boost, were tested using data related to hydraulic fracturing activities in the Montney play area in northeastern British Columbia. All models were initially trained on a subset of 75% of the total data compiled using a 50-fold cross-validation analysis. The remaining 25% of the data was used as a validation set to test the trained models. The validation results showed a high accuracy of successful predictions (97–98%) for all three models.

The classification results were used to calculate the relative importance of all features on whether a well had or had not been associated with induced seismicity. Geological features were differentiated from operational features as the latter are of particular interest as they can be controlled or manipulated to mitigate induced seismicity hazards. However, it was the geological features that generally rated higher with respect to correlation with wells associated with induced seismicity. In all models, pore-pressure gradient (hydrostatic versus overpressured) ranked highly as having a major influence. For the decision-tree and random-forest trained models, distance to basement and distance to known faults also ranked highly, whereas for the gradient boost, the maximum horizontal stress azimuth was a key geological feature that ranked highly. For the operational features, the completion length was the feature most consistently ranked as being of high importance.

Overall, the results of these analyses agree with the current understanding of features that influence induced seismicity susceptibility, such as reservoir overpressure, stress regime and injection volume to stimulate well productivity. These point to the importance of understanding the geology of the Montney Formation including the three-dimensional seismic mapping of faults and taking in situ stress measurements. The machine-learning algorithms investigated here can be used to better understand induced seismicity by determining and ranking the factors that influence induced

seismicity susceptibility and therefore further improve industry practices and regulator oversight.

However, it is also recognized that machine-learning analyses focus exclusively on prediction, bypassing the need for explanations of causality that can add reasoning and confidence to the results. The next steps in this research program will be to add a step of refining the machine-learning output through mechanistic validation using a combination of controlled laboratory experiments and three-dimensional numerical simulations to account for known cause and effect relationships. This will help to increase the reliability of the results and deliver a more robust susceptibility map to help decision makers with their planning of hydraulic fracturing activities and induced seismicity hazard management, as well as identifying areas requiring additional focused research.

Acknowledgments

The authors thank Geoscience BC for their financial support and for facilitating access to several key datasets. This includes use of geoLOGIC systems ltd.'s Well Completion and Frac Database, which was key to this work. The authors would also like to thank to M. Hayes, M. Gaucher, S. Venables, M. Cooper and B. Hayes for their help with compiling geology, pore-pressure and fault location data for the Montney play area, and M. Bustin and P. McLellan for their constructive comments on aspects of this work.

References

Amini, A. and Eberhardt, E. (2019): Influence of tectonic stress regime on the magnitude distribution of induced seismicity events related to hydraulic fracturing; Journal of Petroleum Science and Engineering, v. 182, URL <https://doi.org/10.1016/j.petrol.2019.106284>.

Atkinson, G.M., Eaton, D.W., Ghofrani, H., Walker, D., Cheadle, B., Schultz, R., Shcherbakov, R., Tiampo, K., Gu, J., Harrington, R.M., Liu, Y., Van Der Baan, M. and Kao, H. (2016): Hydraulic fracturing and seismicity in the Western Canada Sedimentary Basin; Seismological Research Letters, v. 87, no. 3, p. 631–647, URL <https://doi.org/10.1785/0220150263>.

Babaie Mahani, A., Schultz, R., Kao, H., Walker, D., Johnson, J. and Salas, C. (2017): Fluid injection and seismic activity in the northern Montney play, British Columbia, Canada, with special reference to the 17 August 2015 Mw 4.6 induced earthquake; Bulletin of the Seismological Society of America, v. 107, issue 2, p. 542–552, URL <https://doi.org/10.1785/0120160175>.

Bao, X. and Eaton, D.W. (2016): Fault activation by hydraulic fracturing in western Canada; Science, v. 354, issue 6318, p. 1406–1409.

BC Oil and Gas Commission (2018): BC Oil and Gas Commission data & reports; BC Oil and Gas Commission, URL <https://www.bcogc.ca/data-reports/> [September 2018].

BC Oil and Gas Commission (2019): British Columbia's Oil and Gas Reserves and Production Report 2018; Reservoir Engineering Department, BC Oil and Gas Commission, 29 p., URL <https://www.bcogc.ca/files/reports/Technical-Reports/2018-oil-and-gas-reserves-and-production-reportfinal.pdf> [November 2020].

Cracknell, M.J. and Reading, A.M. (2014): Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information; Computers and Geosciences, v. 63, p. 22–33.

Currie, B.S., Free, J.C., Brudzinski, M.R., Leveridge, M. and Skoumal, R.J. (2018): Seismicity induced by wastewater injection in Washington County, Ohio: influence of preexisting structure, regional stress regime, and well operations; Journal of Geophysical Research: Solid Earth, v. 123, issue 5, p. 4123–4140, URL <https://doi.org/10.1002/2017JB015297>.

Eaton, D.W. and Schultz, R. (2018): Increased likelihood of induced seismicity in highly overpressured shale formations; Geophysical Journal International, v. 214, issue 1, p. 751–757.

geoLOGIC systems ltd. (2019): geoSCOUT version 8.12; geoLOGIC systems ltd., URL <https://www.geologic.com/products/geoscout/> [March 2020].

Göbel, T. (2015): A comparison of seismicity rates and fluid-injection operations in Oklahoma and California: implications for crustal stresses; The Leading Edge, v. 34, issue 6, p. 640–648, URL <https://doi.org/10.1190/tle34060640.1>.

Hallo, M., Oprsal, I., Eisner, L. and Ali, M.Y. (2014): Prediction of magnitude of the largest potentially induced seismic event; Journal of Seismology, v. 18, issue 3, p. 421–431, URL <https://doi.org/10.1007/s10950-014-9417-4>.

Hastie, T., Tibshirani, R. and Friedman, J. (2017): The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd edition); Springer-Verlag, New York, New York, corrected 12th printing January 2017, 745 p., URL <https://web.stanford.edu/~hastie/ElemStatLearn/> [April 2020].

Hayes, B.J., Anderson, J.H., Cooper, M., McLellan, P.J., Rostron, B. and Clarke, J. (2021): Wastewater disposal in the maturing Montney play fairway, northeastern British Columbia (NTS 093P, 094A, B, G, H); *in* Geoscience BC Summary of Activities 2020: Energy and Water, Geoscience BC, Report 2021-02, p. 91–102, URL <http://geosciencebc.com/updates/summary-of-activities/> [January 2021].

Heidbach, O., Rajabi, M., Cui, X., Fuchs, K., Müller, B., Reinecker, J., Reiter, K., Tingay, M., Wenzel, F., Xie, F., Ziegler, M.O., Zoback, M.L. and Zoback, M. (2018): The World Stress Map database release 2016: crustal stress pattern across scales; Tectonophysics, v. 744, p. 484–498.

Hincks, T., Aspinall, W., Cooke, R. and Gernon, T. (2018): Oklahoma's induced seismicity strongly linked to wastewater injection depth; Science, v. 359, issue 6381, p. 1251–1255, URL <https://doi.org/10.1126/science.aap7911>.

Liu, K. and Liu, B. (2017): Optimization of smooth blasting parameters for mountain tunnel construction with specified control indices based on a GA and ISVR coupling algorithm; Tunnelling and Underground Space Technology, v. 70, p. 363–374, URL <https://www.sciencedirect.com/science/article/pii/S0886779817301761> [November 2020].

Lundberg, S.M. and Lee, S.-I. (2017): A unified approach to interpreting model predictions; *in* Advances in Neural Information Processing Systems 30 (NIPS 2017), URL <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf> [April 2020].

McGarr, A. (2014): Maximum magnitude earthquakes induced by fluid injection; Journal of Geophysical Research: Solid Earth, v. 119, p. 1008–1019.

Millar, D. and Clarici, E. (1994): Investigation of back-propagation artificial neural networks in modelling the stress-strain behaviour of sandstone rock; *in* Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), June 28–July 2, 1994, Orlando, Florida, p. 3326–3331, URL <https://doi.org/10.1109/ICNN.1994.374770>.

Panakkat, A. and Adeli, H. (2009): Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators; Computer-Aided Civil and Infrastructure Engineering, v. 24, issue 4, p. 280–292.

Pawley, S., Schultz, R., Playter, T., Corlett, H., Shipman, T., Lyster, S. and Hauck, T. (2018): The geological susceptibility of induced earthquakes in the Duvernay play; Geophysical Research Letters, v. 45, p. 1786–1793, URL <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017GL076100> [November 2020].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011): Scikit-learn: machine learning in Python; Journal of Machine Learning Research, v. 12, p. 2825–2830.

Pu, Y., Apel, D.B. and Lingga, B. (2018): Rockburst prediction in kimberlite using decision tree with incomplete data; Journal of Sustainable Mining, v. 17, issue 3, p. 158–165.

Ribeiro e Sousa, L., Miranda, T., Leal e Sousa, R. and Tinoco, J. (2017): The use of data mining techniques in rockburst risk assessment; Engineering, v. 3, issue 4, p. 552–558.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C.J. and Johnson, P.A. (2017): Machine learning predicts laboratory earthquakes; Geophysical Research Letters, v. 44, no. 18, p. 9276–9282.

Schultz, R., Atkinson, G., Eaton, D.W., Gu, Y.J. and Kao, H. (2018): Hydraulic fracturing volume is associated with induced earthquake productivity in the Duvernay play; Science, v. 359, issue 6373, p. 304–308.

Schultz, R., Corlett, H., Haug, K., Kocon, K., Maccormack, K., Stern, V. and Shipman, T. (2016): Linking fossil reefs with earthquakes: geologic insight to where induced seismicity occurs in Alberta; Geophysical Research Letters, v. 43, issue 6, p. 2534–2542, URL <https://doi.org/10.1002/2015GL067514>.

Schultz, R., Stern, V. and Gu, Y.J. (2014): An investigation of seismicity clustered near the Cordel Field, west central Alberta, and its relation to a nearby disposal well; Journal of Geophysical Research: Solid Earth, v. 119, no. 4, p. 3410–3423.

Scientific Hydraulic Fracturing Review Panel (2019): Scientific review of hydraulic fracturing in British Columbia; BC Ministry of Energy, Mines and Low Carbon Innovation, final report, 220 p., URL <https://www2.gov.bc.ca/assets/gov/farming-natural-resources-and-industry/natural-gas-oil/responsible-oil-gas-development/scientific_hydraulic_fracturing_review_panel_final_report.pdf> [May 2019].

Shah, A.K. and Keller, G.R. (2017): Geologic influence on induced seismicity: constraints from potential field data in Oklahoma; Geophysical Research Letters, v. 44, issue 1, p. 152–161, URL <https://doi.org/10.1002/2016GL071808>.

Skoumal, R.J., Brudzinski, M.R. and Currie, B.S. (2015): Earthquakes induced by hydraulic fracturing in Poland township, Ohio; Bulletin of the Seismological Society of America, v. 105, issue 1, p. 189–197, URL <https://doi.org/10.1785/0120140168>.

Sun, Y., Feng, X. and Yang, L. (2018): Predicting tunnel squeezing using multiclass support vector machines; Advances in Civil Engineering, v. 2018, art. 4543984, 12 p.

Van der Baan, M. and Calixto, F.J. (2017): Human-induced seismicity and large-scale hydrocarbon production in the USA and Canada; Geochemistry, Geophysics, Geosystems, v. 18, issue 7, p. 2467–2485, URL <https://doi.org/10.1002/2017GC006915>.

Visser, R., Smith, B., Kao, H., Babaie Mahani, A., Hutchinson, J. and McKay, J.E. (2017): A comprehensive earthquake catalogue for northeastern British Columbia and western Alberta, 2014-2016; Geological Survey of Canada, Open File 8335, 28 p., URL <https://doi.org/10.4095/306292>.

Weingarten, M., Ge, S., Godt, J.W., Bekins, B.A. and Rubinstein, J.L. (2015): High-rate injection is associated with the increase in U.S. mid-continent seismicity; Science, v. 348, issue 6241, p. 1336–1340, URL <https://doi.org/10.1126/science.aab1345>.

Zhang, X., Zhang, J., Yuan, C., Liu, S., Chen, Z. and Li, W. (2020): Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method; Scientific Reports, v. 10, no. 1, p. 1–12, URL <https://doi.org/10.1038/s41598-020-58908-5>.