www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

Arch. Min. Sci. **67** (2022), 4, 661-680

Electronic version (in color) of this paper is available: http://mining.archives.pl

DOI: https://doi.org/10.24425/ams.2022.143680

CHENGKAI FAN[1], NA ZHANG[2], BEI JIANG[2*], WEI VICTOR LIU[2*]

PREPROCESSING LARGE DATASETS USING GAUSSIAN MIXTURE MODELLING TO IMPROVE PREDICTION ACCURACY OF TRUCK PRODUCTIVITY AT MINE SITES

The historical datasets at operating mine sites are usually large. Directly applying large datasets to build prediction models may lead to inaccurate results. To overcome the real-world challenges, this study aimed to handle these large datasets using Gaussian mixture modelling (GMM) for developing a novel and accurate prediction model of truck productivity. A large dataset of truck haulage collected at operating mine sites was clustered by GMM into three latent classes before the prediction model was built. The labels of these latent classes generated a latent variable. Two multiple linear regression (MLR) models were then constructed, including the ordinary-MLR (O-MLR) and the hybrid GMM-MLR models. The GMM-MLR model incorporated the observed input variables and a latent variable in the form of interaction terms. The O-MLR model was the baseline model and did not involve the latent variable. The GMM-MLR model performed considerably better than the O-MLR model in predicting truck productivity. The interaction terms quantitatively measured the differences in how the observed input variables affected truck productivity in three classes (high, medium, and low truck productivity). The haul distance was the most crucial input variable in the GMM-MLR model. This study provides new insights into handling massive amounts of data in truck haulage datasets and a more accurate prediction model for truck productivity.

Keywords: Oil sands mining; Mine truck productivity; Gaussian mixture model; Latent variable; Prediction accuracy; Relative importance

1. Introduction

Oil sand mining plays a vital role in Canada's economy [1]. In 2017 alone, it contributed CAD$13 billion to the national revenues and created more than 228,000 direct and indirect

[1] UNIVERSITY OF ALBERTA, EDMONTON, DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING, ALBERTA T6G 2E3, CANADA
[2] UNIVERSITY OF ALBERTA, DEPARTMENT OF MATHEMATICAL AND STATISTICAL SCIENCES, EDMONTON, ALBERTA T6G 2G1, CANADA

* Corresponding authors: bei1@ualberta.ca; victor.liu@ualberta.ca

jobs [2]. In oil sands mining, truck haulage is a dominant means of transporting ores and wastes [3,4]. The productivity of truck haulage (or referred to as truck productivity), defined as truck payload per unit time in each truck cycle, is directly related to a mine's overall productivity [5]. Therefore, it is of great significance to predict truck productivity, which affects a mine's production, planning, income, and expenditure [5,6].

To predict truck productivity, researchers attempt to establish data-driven prediction models based on historical datasets [7]. The datasets may originate from various sources, such as sensor networks [8], remote sensing [9], wireless communication [10], a Vital Information Management System (VIMS) [11], and Michelin Earthmover Management System (MEMS) [12]. Regardless of the source, datasets at mine sites are usually large. For example, Baek and Choi [13] obtained two large datasets from limestone quarries, including 16,217 and 16,005 data points, respectively. The datasets were used to build prediction models for morning and afternoon ore production over two months. Likewise, a dataset collected from oil sand mines in this study was even larger, with more than 300,000 data points covering truck haulage information for an entire year.

Large datasets are usually preprocessed by clustering techniques [14,15]. Clustering is a data mining technique that assigns each data point into a specific class [16]. In each class, the assigned data points share more similarities than those in the other classes [17]. Commonly used clustering techniques include *k*-means [16], hierarchical clustering [18], density-based spatial clustering [19], and Gaussian mixture modelling (GMM) [20]. Of these, GMM is the superior technique for preprocessing large datasets, showing potential for handling massive amounts of data from mine sites. GMM is a probability distribution-based clustering technique that identifies latent classes from a large dataset [21]. In GMM, each class is assumed to follow a Gaussian distribution. Together, these classes form a mixture of Gaussian distributions, which are also known as multi-peak Gaussian distributions [22]. According to the central limit theorem [23], large datasets observed in engineering often present multi-peak Gaussian distributions. This applies to truck haulage data obtained from oil sand mines [24]. For instance, in Fig. 1, the haul distance, ranging from 0 to 10 km, is plotted in a column chart. Each range of haul distance falls under a density ranging from 0 to 0.4. The density refers to the fraction of a range divided by the total size of data. As shown in Fig. 1, the column can either be described by superimposed density curves of a single Gaussian distribution (Fig. 1(a)) or a multi-peak Gaussian distribution (Fig. 1(b)). The multi-peak Gaussian distribution presents two peaks of haul distance, which includes additional information. Relying on these peaks, GMM has the ability to identify latent classes [25], thereby increasing model predictability. For example, Lu et al. [26] used GMM to identify four classes from multi-peak heating load data and then built prediction models separately based on the datasets included in each class. The research showed that the accuracy of prediction models was enhanced by at least 20% based on the identification results. Similar to the research by Lu et al. [26], Ni et al. [27] obtained large streamflow datasets with multi-peak Gaussian distributions and used GMM to divide them into several classes. Each class was then fitted with a single model, and the final prediction was a weighted sum of these models. The results showed that the proposed model's accuracy for streamflow was improved by about 11% compared with the prediction models built based on the original large datasets. In addition, GMM can generate latent variables; the latent variable is defined as the labels of classes, which can be involved in modelling to improve prediction accuracy [28,29]. From the above studies, GMM has advantages for in-depth data mining with multi-peak Gaussian distributions [30]. Thus, GMM may be a more suitable option to improve prediction models because large datasets of truck haulage are usually under multi-peak Gaussian distributions. However, according to the current

literature, no research has reported the application of GMM to preprocess large datasets obtained from mine sites; it is still unknown if GMM can be used to improve the model predictability of truck productivity at mine sites.

To this end, this study was designed to handle large datasets of truck haulage using GMM for developing a novel and accurate prediction model of truck productivity. The large dataset had 303,712 groups of data, which were collected from active oil sands mines in Northern Alberta, Canada. GMM was first used to cluster the large dataset. After that, a latent variable was extracted to build the prediction model in conjunction with other input variables [31]. This is because the multiple linear regression (MLR) method is a computationally efficient tool and can provide explicit formulae for engineers [32]. It was adopted to build the prediction models. The main contribution of this study was the first application of GMM to preprocess massive amounts of data to improve model predictability of truck productivity.

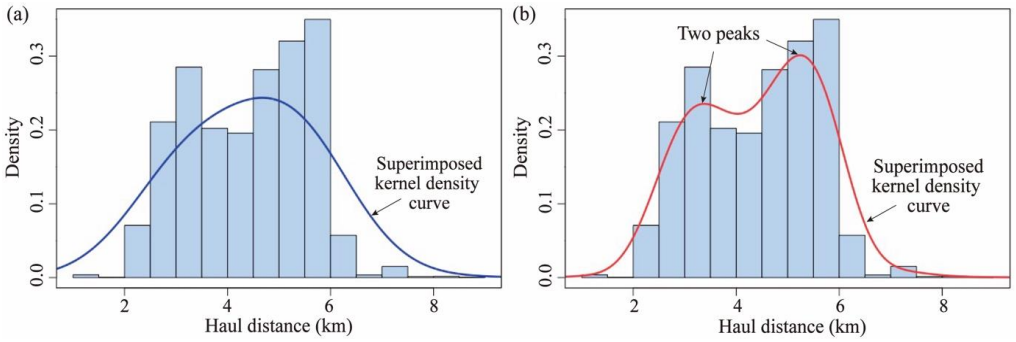

Fig. 1. Data distributions from oil sands mines (using the haul distance as an example). (a) Haul distance is described by a single Gaussian distribution; (b) Haul distance is described by a multi-peak Gaussian distribution

2. Methodology

Fig. 2 illustrates the flowchart of the overall methodology. A large dataset from the mine data management system was split into a training dataset and a testing dataset for model training and evaluation. Before the modelling, the training dataset was clustered by GMM into three latent classes, and a latent variable was generated by the labels of these classes. Two MLR models were then built on the training dataset, including the ordinary-MLR (O-MLR) model and the GMM-MLR model. The GMM-MLR model was the proposed model for predicting truck productivity, incorporating the latent variable. The O-MLR model was the baseline model without involving the latent variable. The testing dataset was used to evaluate the performance of two MLR models. The performance of each model was quantified by four commonly used parameters in statistics [33]: the adjusted R^2, the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). Finally, the Lindeman, Merenda, and Gold (LMG) method was selected to determine the relative importance of input variables to the GMM-MLR model since LMG is a simple and efficient method when an MLR model contains few input variables [34]. The abovementioned training process was implemented in RStudio software using the *R* (version 4.1.3) language environment.

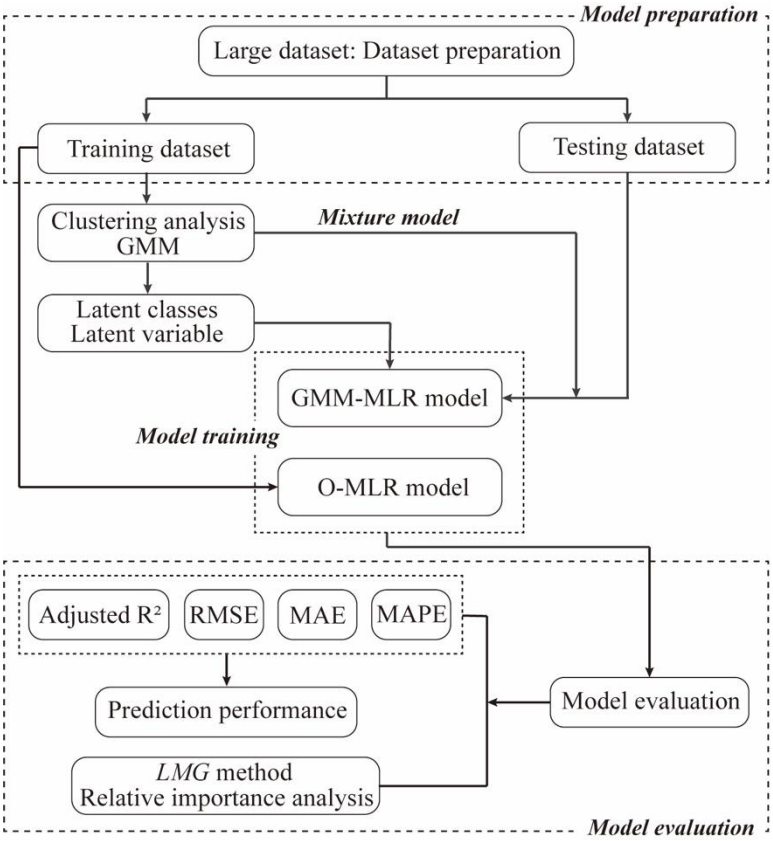

Fig. 2. Flowchart showing the execution process of methodology

2.1. Multiple linear regression (MLR)

MLR is a common statistical technique for building prediction models [32]. It has been widely applied in the fields of agriculture [35], environment [36], and energy [37] because of its simple structure and efficient calculation [32]. In addition, mining companies often utilise MLR to build prediction models because it can provide explicit expressions for engineers to use easily [7,24]. MLR obtains the best-fitting line by minimising the square sum of vertical deviations from data points to a fitted line [37]. This line describes the linear relationship between an output variable and a set of input variables. Suppose that $x = \{x_1, x_2, ..., x_M\}$ is the input vector, where M is the number of input variables, and y is the output variable. The linear relationship can be expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \ldots + \beta_M x_M + \epsilon \tag{1}$$

where β_0 is the constant term that denotes the intercept, β_m is the regression coefficients linked to the m^{th} input variable, and ϵ is the random error term. Equation (1) represents a prediction model based on the MLR method.

2.2. Gaussian mixture modelling (GMM)

GMM is an unsupervised clustering technique that identifies several latent classes from a data population [22]. A set of data points in each class adheres to a specific Gaussian distribution. Statistically, GMM generates a mixture model, which is defined as the weighted combination of k Gaussians, representing the probability density function of the data population. The description of the mixture model is written as follows [38]:

$$H\left(x,\varnothing\right)=\sum_{k=1}^{K}\pi_k f_k\left(y\mid x,\theta_k\right) \tag{2}$$

where $f_k(y\mid x,\theta_k)$ denotes the probability density function of the k^{th} class; θ_k is the parameter vector, which is defined as (μ_k,Σ_k); μ_k and Σ_k are the mean vector and the covariance matrix, respectively; the parameter π_k is the weight of the kth class, also known as the mixture coefficient, which is non-negative together with $\sum_{k=1}^{K}\pi_k=1$; and $\varnothing$ indicates the parameter set of the mixture model, which is written as $\{\pi_k,\theta_k\}$.

To determine the mixture model, GMM first estimates the parameter set $\{\pi_k,\theta_k\}$ from all data points. This estimation can be conducted using the expectation-maximisation (EM) algorithm to maximise log-likelihood (log L) [39]:

$$\log L=\sum_{n=1}^{N}\log\left(H(y\mid x,\phi)\right)l=\sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k f_k(y\mid x,\theta_k)\right) \tag{3}$$

where N is the number of data points. The EM algorithm determines the parameter set $\{\pi_k,\theta_k\}$ through an iterative process, mainly including the E-step and the M-step. In the E-step, data points are assigned to a class with the maximum posterior probability [38]. Based on the Bayes' theorem [40], the posterior probability that a data point (x_i, y_i) belongs to each class is given by

$$\gamma_{ik}=\frac{\pi_k f_k\left(y_i\mid x_i,\theta_k\right)}{\sum_{k=1}^{K}\pi_k f_k\left(y_i\mid x_i,\theta_k\right)} \tag{4}$$

The data point is assigned to the kth class when

$$\lambda_i=\underset{k\in\{1,2,\ldots,K\}}{\arg\max}\ \gamma_{ik} \tag{5}$$

where λ_i represents a set of data points that has the maximum posterior probability, γ_{ik}. Later, in the M-step, with the γ_{ik} , the parameter set $\{\pi_k,\theta_k\}$ can be further estimated by the likelihood setting in Equation (3). These two steps are repeated until the maximum log-likelihood is reached. As a result, the parameter set can be acquired from the EM process.

After the parameters set is estimated, GMM starts to determine the optimal number of latent classes. In this study, the Bayesian information criterion (BIC) was selected as a metric to opti-

mise the number because it has been commonly used in engineering and proved to be superior to other methods in a rigorous study [41]. The BIC formula is shown below:

$$BIC = -2\log L + C\log N \tag{6}$$

where C means the mixture model complexity. The criterion for the optimal number is to minimise the BIC value to achieve a more proper mixture model of the data population [42].

2.3. Dataset preparation and preprocessing

The large dataset contained 303,712 groups of data covering a full year of truck haulage cycles. Before the prediction models were built, the dataset was randomly and proportionally split into training (75%) and testing datasets (25%). Both the training and testing datasets had five input variables observed from the mine sites. These five input variables were chosen because they have been noted by practising engineers at mine sites and are all associated with truck cycle time. They were related to haulage operations, haul routes, and meteorological factors, which were also selected with reference to the research by Chanda and Gardiner [7]. The observed input variables included haul distance (x_1, km), empty speed (x_2, km/h), destination (x_3), ambient temperature (x_4, °C), and precipitation (x_5, mm/h). The first three variables were monitored and identified by the installed sensors on trucks. The remaining two variables were obtained from the local meteorological observatory [43]. Table 1 shows these five input variables, of which the haul distance, empty speed, and ambient temperature were continuous variables. The destination and precipitation were categorical variables, which means that they had several distinct categories. For example, there were three destinations at the mine sites, denoted as D_1, D_2, and D_3. Fig. 3 shows the statistical information about these observed input variables (x_m) and the output variable (y). In Fig. 3, the superimposed density curves represent the distribution of these variables. The continuous variables, including the haul distance, empty speed, and ambient temperature, were represented by the skewed Gaussian and multi-peak Gaussian distributions. Remarkably, the multi-peak Gaussian distributions shown by the haul distance and ambient temperature indicated that the original dataset had a mixture of Gaussians, which provided the rationale for selecting GMM to preprocess the dataset [44].

TABLE 1

The input variables (x_m), characteristics and their descriptions

Input variable	Unit	Type	Description
Haul distance (x_1)	km	Continuous	Listing haul distance for each cycle from a loading area to a dumping area
Empty speed (x_2)	km/h	Continuous	Listing running speed of empty truck for each cycle
Destination (x_3)	—	Categorical	Listing three destinations of truck haulage: D_1, D_2, and D_3
Ambient temperature (x_4)	°C	Continuous	Listing ambient temperature per hour at the local mining area
Precipitation (x_5)	mm/h	Categorical	Listing precipitation per hour at mine sites with three categories: no precipitation (P_1), 0-1 mm/h (P_2), and larger than 1 mm/h (P_3)

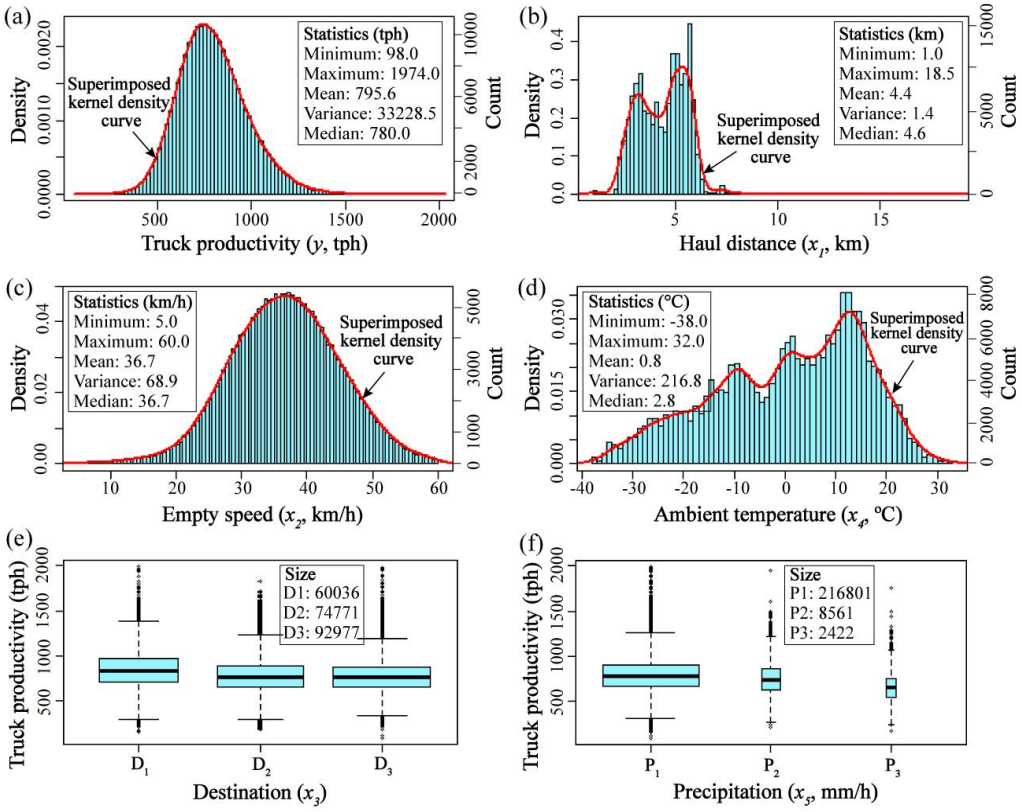

Fig. 3. The output variable and observed input variables in the training dataset. (a) The output variable (y): truck productivity (unit: tph, tonne per hour); (b)-(d) show the histograms of the continuous variables: haul distance (x_1), empty speed (x_2), and ambient temperature (x_4); (e)-(f) show the boxplots of two categorical variables with three categories: destination (x_3) and precipitation (x_5)

By preprocessing the training dataset using GMM, several latent classes were identified from all data points, and a latent variable was generated by the labels of classes. This latent variable was also a categorical variable with several distinct categories; it was in conjunction with other observed input variables to establish the GMM-MLR model. As for the testing dataset, the data points were grouped into several classes based on the mixture model obtained in GMM. The results of the GMM analysis and the number of latent classes will be explained and discussed in detail in Section 3.1.

2.4. Performance criteria for prediction models

To investigate the effect of GMM on prediction performance, two MLR models were built for comparison. One was the GMM-MLR model, which was considered a latent variable generated from the GMM analysis. The other was the O-MLR model, serving as the baseline model without involving the latent variable. To assess the performance of these two models, four performance

parameters were adopted in this study [33]. These parameters were RMSE, MAE, MAPE, and the adjusted R^2. They are calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(A_n - P_n)^2} \tag{7}$$

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|A_n - P_n\right| \tag{8}$$

$$MAPE = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{A_n - P_n}{A_n}\right| \tag{9}$$

where A_n is the actual values, indicating the measured truck productivity in the testing dataset; P_n is the predicted truck productivity. RMSE shows the standard deviation of the residuals between actual and predicted values; MAE is used to characterise the absolute error between actual and predicted values, while MAPE denotes the relative error [33]. The adjusted R^2 is calculated based on R^2. Both are shown, respectively, as Equation (10)-(11):

$$R^2 = 1 - \frac{\sum_n^N \left(A_n - P_n\right)^2}{\sum_n^N \left(A_n - E_n\right)^2} \tag{10}$$

$$R_{adj}^2 = 1 - \frac{\left(1 - R^2\right)\left(N - 1\right)}{N - M - 1} \tag{11}$$

where E_n is the mean of actual values and M is the number of input variables. Both R^2 and the adjusted R^2 represent the degree to which data points fit a curve, ranging from 0 to 1. The adjusted R^2 is generally smaller than R^2 because input variables unrelated to the output variable are screened when calculating the adjusted R^2; therefore, the adjusted R^2 indicates the goodness of fit more accurately than R^2 [45]. The prediction model with a higher adjusted R^2 and a lower RMSE, MAE, and MAPE have better prediction accuracy.

2.5. The Lindeman, Merenda, and Gold (LMG) method

To evaluate the contributions of input variables to the proposed GMM-MLR model, a quantitative method was introduced to calculate the relative importance of each input variable. This method is called the LMG method [46]. It is straightforward and efficient when an MLR model contains few input variables [34]. The LMG method can consider all sequences of an input variable entering an MLR model. The relative importance of this input variable is calculated by averaging the R^2 of all possible orderings, which can be determined according to Equation (12):

$$LMG = \frac{1}{M!}\sum_{p\ permutation} seq\{R^2(x_m \mid p)\} \tag{12}$$

where $M!$ is the factorial of M; p represents the permutation of input variables before entering x_m,

and $seq\{R^2(x_m|p)\}$ refers to the R^2 of the prediction model after entering x_m in the permutation p. The relative importance of x_m is the average value of R^2 under all permutations.

3. Results and discussion

3.1. GMM analysis

In this study, GMM was applied to cluster the training dataset under the principle of minimising BIC. As a result, the training dataset was clustered into three latent classes, as shown in Fig. 4. In Fig. 4(a), taking truck productivity as an example, the number of data points was different in each class. The boxplot showed that Class 1 (C_1) had the lowest number of data points (6,684), while Classes 2 and 3 (C_2 and C_3) had 119,145 and 101,955 data points, respectively. Q_1 and Q_3 were the 25th and 75th percentiles in each class, depicting the distribution interval of data points [47]. Fig. 4(b) shows the frequency histogram of truck productivity in each latent class. According to the definition of GMM [22], the data points in each latent class are described by a Gaussian distribution. The mean values of each Gaussian were around 1,166 tph, 865 tph, and 670 tph. As shown in Fig. 4, the training dataset was well partitioned into three latent classes. Amid these classes, the value of truck productivity varied significantly, in the order of $C_1 > C_2 > C_3$. This can be known as the high, medium, and low truck productivity at mine sites. This is similar to Ni et al. [27]; in their research, the streamflow data were also clustered into three latent classes using GMM. A prediction model was then developed for monthly low flow forecasting based on the GMM analysis; the R^2 of this model was increased from 0.59 to 0.66 compared to the baseline model without the GMM analysis. This suggests that implementing GMM may improve the model accuracy of truck productivity in this study.

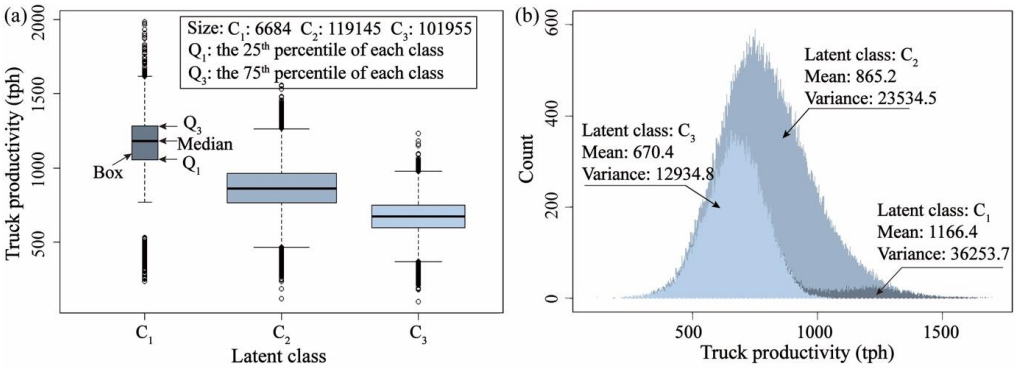

Fig. 4. Extraction of three latent classes from the training dataset. (a) Boxplots of three classes; (b) Histogram: truck productivity corresponds to three latent classes, which are described by Gaussian distributions

3.2. Establishment, interpretation, and comparison of prediction models

3.2.1. O-MLR model

The O-MLR model was a baseline model established on the training dataset without the implementation of GMM. The explicit equation of this model can be written as

$$y = \beta_0 + \sum_{m=1}^{5} \beta_m x_m \tag{13}$$

where y was the output variable. β_0 was the intercept of the equation, and β_m was the regression coefficients linked to the mth observed input variable (x_m). The regression parameters of Equation (13) can be seen in Table 2 presents the observed input variables (x_m), regression coefficients (β_m), significance test results (p-values), and intercept (β_0). The observed input variables included the haul distance (x_1), empty speed (x_2), destination (x_3), ambient temperature (x_4), and precipitation (x_5). The regression coefficients describe the mathematical relationship between each input variable and the output variable [48]. For example, the haul distance's regression coefficient (β_1) was a negative value (–62.70), indicating that the truck productivity was reduced by 62.70 tph when the haul distance increased by 1 km. The same result was found by Schexnayder et al. [49]; their data proved that the truck productivity dropped by 374 tph when the haul distance rose by 1.3 km. Hence, the truck productivity had a negative relationship with the haul distance. The p-values for regression coefficients represent whether these relationships are statistically significant [50]. In statistics, if a p-value is smaller than a significance level (usually 0.05), the relationship between the input and output variables is significant [51]. As shown in Table 2, the relationships between three continuous variables (haul distance, empty speed, and ambient temperature) and truck productivity were statistically significant because their p-values were less than 0.05. Also, two categorical variables (destination and precipitation) were significantly related to truck productivity, except for the second category (D_2) of destination, as its p-value

TABLE 2

The regression parameters and significance test results for the O-MLR model

Input variable		Regression coefficient		p-value	Significance test
x_1	Haul distance	β_1	–62.70	$<2\times10^{-16}$	Reject
x_2	Empty speed	β_2	4.91	$<2\times10^{-16}$	Reject
x_4	Ambient temperature	β_4	–1.44	$<2\times10^{-16}$	Reject
x_3	Destination (D_2)	β_3	–0.57	0.546	Accept
	Destination (D_3)		–11.71	$<2\times10^{-16}$	Reject
x_5	Precipitation (P_2)	β_5	–34.31	$<2\times10^{-16}$	Reject
	Precipitation (P_3)		–75.51	$<2\times10^{-16}$	Reject
Intercept		β_0	900.20	$<2\times10^{-16}$	Reject

Note: If the p-value is less than 0.05, the null hypothesis that x and y are not significantly related will be rejected; otherwise, it will be accepted. For example, the p-value (0.546) for the second category (D_2) of x_3 is larger than 0.05; as a result, the null hypothesis is accepted.

(0.546) was larger than 0.05. In short, almost all the input variables had a significant relationship with truck productivity, suggesting the trained O-MLR model can be used to predict truck productivity.

3.2.2. GMM-MLR model (incorporation of a latent variable and its interaction terms)

After the implementation of GMM, the training dataset was employed to build the GMM-MLR model. The explicit expression of this model can be given by

$$y = \beta_0 + \sum_{m=1}^{5} \beta_m x_m + \beta_6 x_6 + \sum_{m=1}^{5} \beta_{m+6} (x_m \times x_6) \tag{14}$$

where β_6 was the regression coefficients of the latent variable (x_6), and β_{m+6} was the regression coefficients of interaction terms ($x_m \times x_6$) between the five observed input variables (x_m) and the latent variable (x_6). Compared with Equation (13), two more terms were incorporated in Equation (14), including an independent term and a set of interaction terms. The independent term was constituted by a latent variable (x_6) and its regression coefficient (β_6). The latent variable was a categorical variable with three categories (C_1, C_2, and C_3), and the GMM analysis showed that it was related to the five observed input variables. Hence, a set of interaction terms was considered in the GMM-MLR model between the five observed input variables and the latent variable. The interaction term refers to the product of two or more input variables in a regression equation [52]. For instance, in Equation (14), the haul distance (x_1) had an interaction term ($x_1 \times x_6$) with the latent variable (x_6).

Table 3 lists the detailed regression parameters of Equation (14), including the input variables, interaction terms, regression coefficients, *p*-values, and intercept. As shown in Table 3, the GMM-MLR model incorporated the five observed input variables, a latent variable and five sets of interaction terms. The regression coefficients in Table 3 will be explained in detail in Section 3.2.3. As for the *p*-values, almost all the input variables and interaction terms had a significant relationship with the truck productivity since their *p*-values were smaller than 0.05. Thus, the established GMM-MLR model can also be applied for predicting truck productivity.

TABLE 3

The regression parameters and significance test results for the GMM-MLR model

Input variable and interaction term		Regression coefficient		*p*-value	Significance test
x_1	Haul distance	β_1	–105.92	$<2\times10^{-16}$	Reject
x_2	Empty speed	β_2	0.52	4.29×10^{-5}	Reject
x_4	Ambient temperature	β_4	–4.23	$<2\times10^{-16}$	Reject
x_3	Destination (D_2)	β_3	–42.20	$<2\times10^{-16}$	Reject
	Destination (D_3)		–40.58	$<2\times10^{-16}$	Reject
x_5	Precipitation (P_2)	β_5	–44.28	6.26×10^{-15}	Reject
	Precipitation (P_3)		–71.90	6.58×10^{-10}	Reject

TABLE 3. Continued

x_6	Latent variable (C_2)	β_6	–643.08	$<2\times10^{-16}$	Reject
	Latent variable (C_3)		–973.95	$<2\times10^{-16}$	Reject
$x_1\times x_6$	Haul distance × latent variable (C_2)	β_7	8.48	2.28×10^{-16}	Reject
	Haul distance × latent variable (C_3)		75.91	$<2\times10^{-16}$	Reject
$x_2\times x_6$	Empty speed × latent variable (C_2)	β_8	9.68	$<2\times10^{-16}$	Reject
	Empty speed × latent variable (C_3)		4.99	$<2\times10^{-16}$	Reject
$x_4\times x_6$	Ambient temperature × latent variable (C_2)	β_{10}	2.22	$<2\times10^{-16}$	Reject
	Ambient temperature × latent variable (C_3)		3.23	$<2\times10^{-16}$	Reject
$x_3\times x_6$	Destination (D_2) × latent variable (C_2)	β_9	51.92	$<2\times10^{-16}$	Reject
	Destination (D_2) × latent variable (C_3)		32.95	$<2\times10^{-16}$	Reject
	Destination (D_3) × latent variable (C_2)		41.95	$<2\times10^{-16}$	Reject
	Destination (D_3) × latent variable (C_3)		20.83	6.78×10^{-13}	Reject
$x_5\times x_6$	Precipitation (P_2) × latent variable (C_2)	β_{11}	11.65	0.046	Reject
	Precipitation (P_2) × latent variable (C_3)		14.41	0.015	Reject
	Precipitation (P_3) × latent variable (C_2)		–4.74	0.690	Accept
	Precipitation (P_3) × latent variable (C_3)		25.21	0.037	Reject
Intercept		β_0	1,616.21	$<2\times10^{-16}$	Reject

Note: If the *p*-value is less than 0.05, the null hypothesis that *x* and *y* are independent will be rejected; otherwise, it will be accepted. For example, the *p*-value (4.29×10^{-5}) for x_1 is less than 0.05; as a result, the null hypothesis is rejected.

3.2.3. Interpretation of interaction terms

The interaction term implies that the effect of an input variable on an outcome depends not only on that particular input variable but on other input variables [53]. For instance, in the GMM-MLR model, the effect of haul distance on truck productivity depended on both the haul distance and the latent variable. Furthermore, the GMM analysis demonstrated that the latent variable could represent three classes of truck productivity: C_1 (high values), C_2 (medium values), and C_3 (low values). This means that the interaction terms can further characterise the effects of the five observed input variables on each class of truck productivity. Also, these effects can be quantitatively measured through the regression coefficients of the established GMM-MLR model.

In Table 3, there are 11 sets of regression coefficients. Among them, the regression coefficients (β_1 to β_5) for each observed input variable (x_m) indicated the effect of the input variable on the truck productivity belonging to C_1. The regression coefficients (β_7 to β_{11}) of each interaction term ($x_m\times x_6$) suggested the effect of the input variable on the truck productivity belonging to C_2 and C_3. As shown in Fig. 5, the haul distance and precipitation were used as examples to interpret the regression coefficients. In Fig. 5(a), there were three negative values: –105.92 tph, –97.44 tph, and –30.01 tph. Of these values, –105.92 was the β_1, indicating that the high truck productivity (C_1) was reduced by 105.92 tph when the haul distance increased by 1 km. The values of –97.44 tph and –30.01 tph were calculated from the sum of the β_1 (–105.92) and β_7 (8.48 and 75.91), meaning that the medium (C_2) and low (C_3) truck productivity decreased by 97.44 tph and 30.01 tph when the haul distance rose by 1 km. Likewise, the effects of the precipitation (P_2 and P_3) on three classes of truck productivity are illustrated in Fig. 5(b)-(c). In Fig. 5(b), the

high, medium, and low truck productivity were reduced by 44.28 tph, 32.63 tph, and 29.87 tph when the precipitation (P_2) increased by 1 mm/h. In Fig. 5(c), the effect of the precipitation (P_3) on the medium truck productivity (C_2) was ignored as the *p*-value of this term was larger than 0.05. The high and low truck productivity dropped by 71.90 tph and 46.69 tph, respectively, when the precipitation (P_3) rose by 1 mm/h. Thus, the interaction terms revealed that the effect of each observed input variable on truck productivity was significantly different between the three classes. The finding was similar to that in studies by Kyburz et al. [54] and Lunt [55], who were interested in the effect of treated time on a radiographic damage score for subjects in an early or late treated group. To evaluate the difference between the groups, Kyburz et al. [54] and Lunt [55] constituted an interaction term in a regression model. The results proved that the interaction term could also measure the different effects between groups.

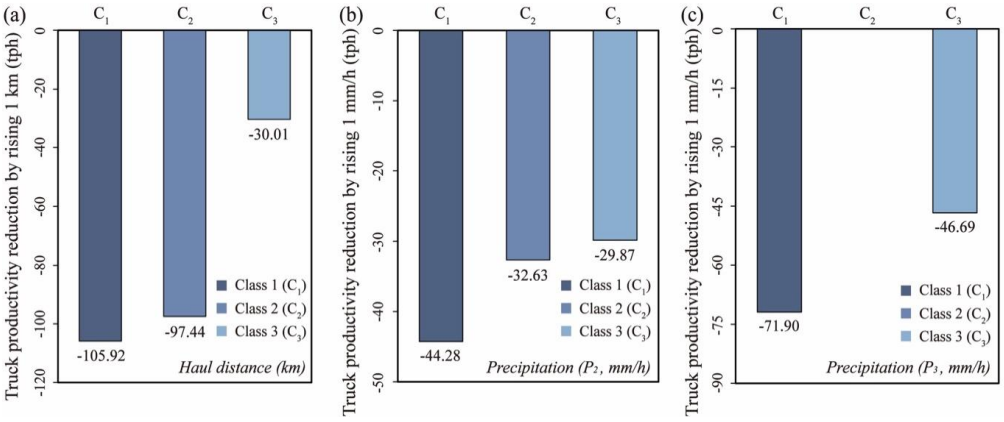

Fig. 5. The effects of the observed input variables on each class of truck productivity (C_1, C_2, and C_3 represented the high, medium, and low truck productivity, respectively). (a) The effects of the haul distance. (b) The effects of the precipitation (P_2). (c) The effects of the precipitation (P_3)

3.2.4. Comparison between O-MLR and GMM-MLR models

Fig. 6 shows the scatterplots of the actual (on the vertical axis) and predicted (on the horizontal axis) truck productivity. The $y = x$ is a 45-degree diagonal line. The closer the scatters along the $y = x$ line, the better the prediction [56]. As shown in Fig. 6, the scatters generated by the GMM-MLR model were closer along the line, which means that the GMM-MLR model performed better than the O-MLR model. To quantitatively evaluate the performance of the established models, four parameters were calculated for each model from the testing dataset. The results are listed in Table 4, which shows that the GMM-MLR model was more accurate than the O-MLR model. The GMM-MLR model had a lower RMSE, MAE, and MAPE, and a higher adjusted R^2, with values of 91.87, 72.58, 0.10, and 0.75. Accordingly, these four performance parameters of the O-MLR model were 160.27, 124.31, 0.17, and 0.23. In terms of the adjusted R^2 alone, the accuracy of the GMM-MLR model (the adjusted $R^2 = 0.75$) was three times higher than the O-MLR model (the adjusted $R^2 = 0.23$). In other words, the GMM-MLR model performed well in predicting

truck productivity. After using GMM to preprocess the large dataset, the model predictability was considerably enhanced by incorporating the latent variable and its interaction terms. This provides new insights and inspiration for engineers to handle massive amounts of engineering data in their future work. Similar findings were also noted in the research by Ho Park et al. [57], who incorporated seven input variables and constituted 11 sets of interaction terms in a linear regression model for post-event flood waste estimation. The results showed that the adjusted R^2 of the prediction model was increased from 0.36 to 0.59 when the model was added with these input variables and interaction terms.

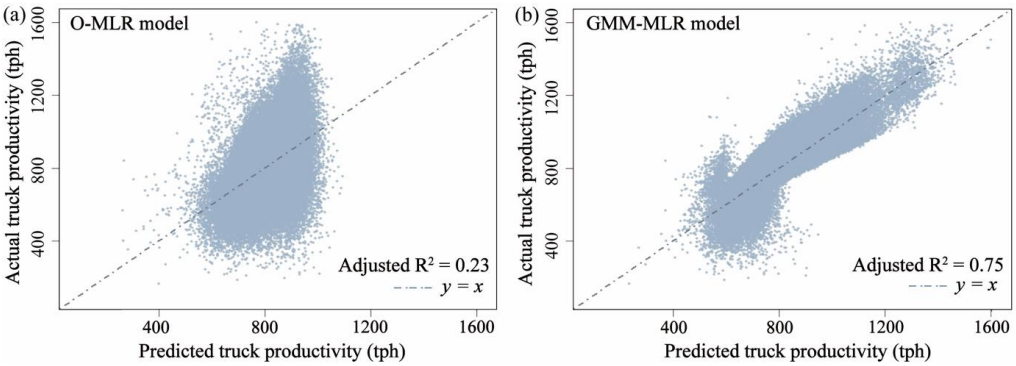

Fig. 6. Scatterplots of the actual truck productivity in the testing dataset and the predicted truck productivity generated by the O-MLR and GMM-MLR models. (a) The O-MLR model; (b) The GMM-MLR model

TABLE 4

Performance evaluation by four parameters for the trained models

Prediction model	RMSE	MAE	MAPE	Adjusted R^2
GMM-MLR	91.82	72.58	0.10	0.75
O-MLR	160.27	124.31	0.17	0.23

3.3. Relative importance analysis of observed input variables

In this study, the LMG method was adopted to determine the relative importance of each observed input variable. Fig. 7 shows the relative importance of these observed input variables in the GMM-MLR model. The vertical axis represented the five observed input variables; the horizontal axis was the relative importance proportion (in percentage) of each one. The relative importance for the input variables was ranked as haul distance (54.65%) > empty speed (23.14%) > ambient temperature (13.82%) > destination (6.22%) > precipitation (2.18%). Among these variables, the haul distance had the highest relative importance, indicating its effect on truck productivity was greater than that of other input variables. Cervantes et al. [24] reported that mining companies often plotted a fitted line between haul distance and truck productivity because the increase in haul distance directly affects the increase in cycle time, thereby reducing truck productivity. Similar to the study by Cervantes et al. [24], the results from the relative importance analysis also proved that the haul distance was a critical input variable in predicting truck productivity.

After the haul distance, the analysis showed that the empty speed had the second-highest relative importance, with a value of 23.14%. According to Schexnayder et al. [49], the empty speed determined the travel time from dumping sites to loading sites, affecting truck productivity. The relative importance of the destination was 6.22%, indicating its effect on truck productivity was not significant. This is reasonable since the destination cannot directly affect the payload weight and cycle time length [58]. The sum of the relative importance of the ambient temperature and precipitation was 16.01%, showing that the meteorological factors had a certain contribution to the GMM-MLR model. Similar to the research by Sun et al. [59], the prediction accuracy was enhanced by 5.13% when considering the effect of meteorological factors. To summarise, the observed input variables contributed differently to the GMM-MLR model, with haul distance being the most crucial input variable. The relative importance analysis can help mine engineers to gain a comprehensive understanding of the real-world influences affecting truck productivity, thus providing appropriate suggestions and methods to improve truck productivity.

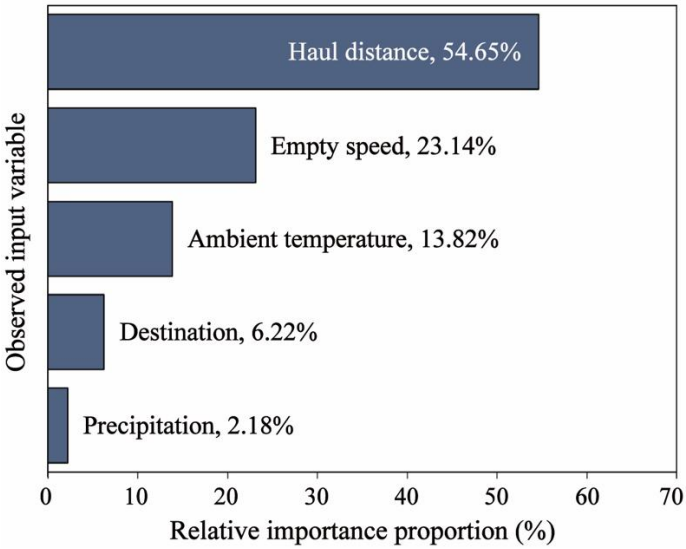

Fig. 7. The relative importance of the observed input variables in the GMM-MLR model

3.4. Advantage, limitation, and future improvement of the proposed model

In this study, the GMM-MLR model was the proposed model for predicting truck productivity. Unlike previous studies [7,13,59], this proposed model not only considered input variables observed at mine sites but involved unobserved variables (i.e., latent variables) obtained from the GMM analysis. Due to the involvement of latent variables, the model accuracy of truck productivity was considerably enhanced (e.g., the R^2 was increased from 0.23 to 0.75). Despite its better performance, the proposed GMM-MLR model had limitations in this study. Much research will be required to further the prediction model. For instance, although GMM has advantages in dealing with large datasets with multi-peak Gaussian distributions, it is not the only clustering

technique [60]. Previous studies have shown that clustering techniques such as k-means and fuzzy c-means improved model accuracy [56,61]. A comparative study of clustering techniques may be helpful to improve prediction models. In addition, more input variables, such as tire temperature, wind speed, and elevation, can be considered in the future to build prediction models. According to Ma et al. [62], high tire temperature may cause rubber failure, affecting truck speed and cycle time. Likewise, wind speed and elevation over the haul route may impact truck speed and the driver's vision [7,59]. However, these parameters are not included in the currently proposed model. Furthermore, the modelling approach used in this study was the MLR method, whilst more robust algorithms, such as support vector machine [63], random forest [64], and artificial neural network [65], can also provide accurate prediction models. In the future, these algorithms will be used to increase model predictability.

4. Conclusions

This study aimed to handle large datasets of truck haulage at mine sites using Gaussian mixture modelling (GMM) for developing a novel and accurate prediction model of truck productivity based on multiple linear regression (MLR). The main conclusions are listed below:

(1) GMM significantly improved the predictability of the truck productivity prediction model by preprocessing large truck haulage datasets. For example, the adjusted R^2 of the ordinary-MLR (O-MLR) model was only 0.23, whereas the GMM-MLR improved the predictability more than three times, with an adjusted R^2 of 0.75. This information can provide new insights and inspiration for engineers to deal with massive amounts of engineering data in their future work.

(2) Interaction terms quantitatively measured the significant differences in the effect of an observed input variable on truck productivity between classes. For instance, when the haul distance increased by 1 km, the high (Class 1), medium (Class 2), and low (Class 3) truck productivity dropped by 105.92 tph, 97.44 tph, and 30.01 tph, respectively. Hence, the effect of the haul distance on high truck productivity was more significant than that on medium and low truck productivity, showing the significant differences between the classes revealed by the interaction terms.

(3) Among the observed input variables, the haul distance was the most crucial input variable of the GMM-MLR model. The relative importance of the haul distance was 54.65%, which was higher than that of the empty speed (23.14%), destination (6.22%), ambient temperature (13.82%), and precipitation (2.18%). The relative importance analysis helps mine engineers to gain a comprehensive understanding of the real-world influences affecting truck productivity, thus providing appropriate suggestions and methods to improve truck productivity.

(4) The GMM-MLR model with higher accuracy is expressed as an explicit and straightforward equation, which can help mine engineers predict truck productivity at mine sites.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Collaborative Research Project (RES0043251) and the Pilot Seed Grant (RES0049944) from the University of Alberta.

References

[1] S. Sleep, I.J. Laurenzi, J.A. Bergerson, H.L. MacLean, Evaluation of variability in greenhouse gas intensity of Canadian oil sands surface mining and upgrading operations. Environ. Sci. & Technol. **52** (20), 11941-11951 (2018). DOI: https://doi.org/10.1021/acs.est.8b03974

[2] CAPP, A strong energy sector is key to ensure Canada's prosperity for the future. Canadian Association of Petroleum Producers (CAPP) (2018). https://www.capp.ca/economy/canadian-economic-contribution/

[3] A.K. Katta, M. Davis, V. Subramanyam, A.F. Dar, M.A.H. Mondal, M. Ahiduzzaman, A. Kumar, Assessment of energy demand-based greenhouse gas mitigation options for Canada's oil sands. J. Clean. Prod. **241**, 118306 (2019). DOI: https://doi.org/10.1016/j.jclepro.2019.118306

[4] P. Bodziony, Z. Kasztelewicz, P. Sawicki, The problem of multiple criteria selection of the surface mining haul trucks. Arch. Min. Sci. **61** (2), 223-243 (2016). DOI: http://doi.org/10.1515/amsc-2016-0017

[5] S. Alarie, M. Gamache, Overview of solution strategies used in truck dispatching systems for open pit mines. Int. J. Surf. Min. Reclam. Environ. **16** (1), 59-76 (2002). DOI: https://doi.org/10.1076/ijsm.16.1.59.3408

[6] P.J. Bartos, Is mining a high-tech industry?: Investigations into innovation and productivity advance. Resour. Policy **32** (4), 149-158 (2007). DOI: https://doi.org/10.1016/j.resourpol.2007.07.001

[7] E.K. Chanda, S. Gardiner, A comparative study of truck cycle time prediction methods in open-pit mining. Eng. Constr. Archit. Manag. **17** (5), 446-460 (2010). DOI: https://doi.org/10.1108/09699981011074556

[8] Y. Gui, Z. Tao, C. Wang, X. Xie, Study on remote monitoring system for landslide hazard based on wireless sensor network and its application. J. Coal Sci. Eng. **17**, 464-468 (2011). DOI: https://doi.org/10.1007/s12404-011-0422-8

[9] Q. Gu, C. Lu, J. Guo, S. Jing, Dynamic management system of ore blending in an open pit mine based on GIS/GPS/GPRS. Min. Sci. Technol. **20** (1), 132-137 (2010). DOI: https://doi.org/10.1016/S1674-5264(09)60174-5

[10] V. Sabniveesu, A. Kavuri, R. Kavi, V. Kulathumani, V. Kecojevic, A. Nimbarte, Use of wireless, ad-hoc networks for proximity warning and collision avoidance in surface mines. Int. J. Min., Reclam. Environ. **29** (5), 331-346 (2015). DOI: https://doi.org/10.1080/17480930.2015.1086550

[11] E. Siami-Irdemoosa, S.R. Dindarloo, Prediction of fuel consumption of mining dump trucks: A neural networks approach. Appl. Energy **151**, 77-84 (2015). DOI: https://doi.org/10.1016/j.apenergy.2015.04.064

[12] K. Zhang, S. Ji, Y. Zhang, J. Zhang, R. Pan, MEMS inertial sensor for strata stability monitoring in underground mining: An experimental study. Shock Vib. **2018**, 4895862 (2018). DOI: https://doi.org/10.1155/2018/4895862

[13] J. Baek, Y. Choi, Deep neural network for predicting ore production by truck-haulage systems in open-pit mines. Appl. Sci. **10** (5), 1657 (2020). DOI: https://doi.org/10.3390/app10051657

[14] M.A. Shahin, H.R. Maier, M.B. Jaksa, Data division for developing neural networks applied to geotechnical engineering. J. Comput. Civ. Eng. **18** (2), 105-114 (2004). DOI: https://doi.org/10.1061/(ASCE)0887-3801(2004)18:2(105)

[15] S.R. Dindarloo, E. Siami-Irdemoosa, Data mining in mining engineering: results of classification and clustering of shovels failures data. Int. J. Min. Reclam. Environ. **31** (2), 105-118 (2017). DOI: https://doi.org/10.1080/17480930.2015.1123599

[16] M.S. Alam, S. Paul, A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh. J. Appl. Stat. **47** (8), 1460-1481 (2020). DOI: https://doi.org/10.1080/02664763.2019.1675606

[17] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S.E. Lee, C. Sekhar, K.W. Tham, K-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy Build. **146**, 27-37 (2017). DOI: https://doi.org/10.1016/j.enbuild.2017.03.071

[18] L. Tu, Y. Lv, Y. Zhang, X. Cao, Logistics service provider selection decision making for healthcare industry based on a novel weighted density-based hierarchical clustering. Adv. Eng. Inform. **48**, 101301 (2021). DOI: https://doi.org/10.1016/j.aei.2021.101301

[19] X. Wang, H.J. Hamilton, A comparative study of two density-based spatial clustering algorithms for very large datasets, in: B. Kégl, G. Lapalme (Eds.) Advances in Artificial Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg (2005). https://doi.org/10.1007/11424918_14

[20] L. Zhang, S.-K. Oh, W. Pedrycz, B. Yang, Y. Han, Building fuzzy relationships between compressive strength and 3D microstructural image features for cement hydration using Gaussian mixture model-based polynomial radial basis function neural networks. Appl. Soft Comput. **112**, 107766 (2021). DOI: https://doi.org/10.1016/j.asoc.2021.107766

[21] J. Diaz-Rozo, C. Bielza, P. Larrañaga, Machine-tool condition monitoring with Gaussian mixture models-based dynamic probabilistic clustering. Eng. Appl. Artif. Intell. **89**, 103434 (2020). DOI: https://doi.org/10.1016/j.engappai.2019.103434

[22] C.M. Bishop, Pattern Recognition and Machine Learning. Springer, Verlag New York (2006).

[23] J.A. Rice, Mathematical Statistics and Data Analysis. Duxbury Press, Belmont, CA (1995).

[24] E.G. Cervantes, S.P. Upadhyay, H. Askari-Nasab, Improvements to production planning in oil sands mining through analysis and simulation of truck cycle times. Mining Optimization Laboratory (MOL), University of Alberta, 142-156 (2019).

[25] F. Ge, Y. Ju, Z. Qi, Y. Lin, Parameter estimation of a Gaussian mixture model for wind power forecast error by Riemann L-BFGS optimization. IEEE Access. **6**, 38892-38899 (2018). DOI: https://10.1109/ACCESS.2018.2852501

[26] Y. Lu, Z. Tian, P. Peng, J. Niu, W. Li, H. Zhang, GMM clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system. Energy Build. **190**, 49-60 (2019). DOI: https://doi.org/10.1016/j.enbuild.2019.02.014

[27] L. Ni, D. Wang, J. Wu, Y. Wang, Y. Tao, J. Zhang, J. Liu, Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. J. Hydrol. **586**, 124901 (2020). DOI: https://doi.org/10.1016/j.jhydrol.2020.124901

[28] G.H. Lubke, J. Luningham, Fitting latent variable mixture models. Behav. Res. Ther. **98**, 91-102 (2017). DOI: https://doi.org/10.1016/j.brat.2017.04.003

[29] O.E. Parsons, A Gaussian mixture model approach to classifying response types, in: N. Bouguila, W. Fan (Eds.) Mixture Models and Applications. Springer International Publishing, Cham (2020). DOI: https://doi.org/10.1007/978-3-030-23876-6_1

[30] L. Ye, Y. Zhang, C. Zhang, P. Lu, Y. Zhao, B. He, Combined Gaussian mixture model and cumulants for probabilistic power flow calculation of integrated wind power network. Comput. Electr. Eng. **74**, 117-129 (2019). DOI: https://doi.org/10.1016/j.compeleceng.2019.01.010

[31] K.S. Berlin, N.A. Williams, G.R. Parra, An introduction to latent variable mixture modeling (part 1): Overview and cross-sectional latent class and latent profile analyses. J. Pediatr. Psychol. **39** (2), 174-187 (2013). DOI: https://doi.org/10.1093/jpepsy/jst084

[32] G. Ciulla, A. D'Amico, Building energy performance forecasting: A multiple linear regression approach. Appl. Energy. **253**, 113500 (2019). DOI: https://doi.org/10.1016/j.apenergy.2019.113500

[33] L. Wu, C. Hu, W.V. Liu, Forecasting the deterioration of cement-based mixtures under sulfuric acid attack using support vector regression based on Bayesian optimization. SN Appl. Sci. **2**, 1970 (2020). DOI: https://doi.org/10.1007/s42452-020-03778-9

[34] W. Tian, Y. Liu, Y. Heo, D. Yan, Z. Li, J. An, S. Yang, Relative importance of factors influencing building energy in urban environment. Energy **111**, 237-250 (2016). DOI: https://doi.org/10.1016/j.energy.2016.05.106

[35] S. Dhulipala, G.R. Patil, Freight production of agricultural commodities in India using multiple linear regression and generalized additive modelling. Transp. Policy **97**, 245-258 (2020). DOI: https://doi.org/10.1016/j.tranpol.2020.06.012

[36] Q. Tan, Y. Wei, M. Wang, Y. Liu, A cluster multivariate statistical method for environmental quality management. Eng. Appl. Artif. Intell. **32**, 1-9 (2014). DOI: https://doi.org/10.1016/j.engappai.2014.02.007

[37] M. Maaouane, S. Zouggar, G. Krajačić, H. Zahboune, Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods. Energy, **225**, 120270 (2021). DOI: https://doi.org/10.1016/j.energy.2021.120270

[38] F. Leisch, FlexMix: A general framework for finite mixture models and latent class regression in R.J. Stat. Softw. **11** (8), 1-18 (2004). DOI: https://doi.org/10.18637/jss.v011.i08

[39] Y. Fu, X. Liu, S. Sarkar, T. Wu, Gaussian mixture model with feature selection: An embedded approach. Comput. Ind. Eng. **152**, 107000 (2021). DOI: https://doi.org/10.1016/j.cie.2020.107000

[40] Y. Li, E. Schofield, M. Gönen, A tutorial on Dirichlet process mixture modeling. J. Math. Psychol. **91**, 128-144 (2019). DOI: https://doi.org/10.1016/j.jmp.2019.04.004

[41] J.S. Russell, A.E. Raftery, Performance of Bayesian model selection criteria for Gaussian mixture models. Despartment of Statistics, University of Washington (2009).

[42] G.J. McLachlan, S.X. Lee, S.I. Rathnayake, Finite mixture models. Annu. Rev. Stat. Appl. **6**, 355-378 (2019). DOI: https://doi.org/10.1146/annurev-statistics-031017-100325

[43] MEP, Current and historical Alberta weather station data viewer. Ministry of Environment and Parks (MEP), Government of Alberta (2019). https://acis.alberta.ca/weather-data-viewer.jsp

[44] Z. Ma, H. Li, Q. Sun, C. Wang, A. Yan, F. Starfelt, Statistical analysis of energy consumption patterns on the heat demand of buildings in district heating systems. Energy Build. **85**, 464-472 (2014). DOI: https://doi.org/10.1016/j.enbuild.2014.09.048

[45] M. Mittlböck, Calculating adjusted RP^2P measures for Poisson regression models. Comput. Methods Programs Biomed. **68** (3), 205-214 (2002). DOI: https://doi.org/10.1016/S0169-2607(01)00173-0

[46] U. Groemping, Relative importance for linear regression in R: The package relaimpo. J. Stat. Softw., **17** (1), 1-27 (2006). DOI: https://doi.org/10.18637/jss.v017.i01

[47] K. Patil, N.K. Nagwani, S. Tripathi, A parametric study of partitioning and density based clustering techniques for boxplot generation. 2018 3rd International Conference for Convergence in Technology (I2CT), 1-5 (2018). DOI: https://doi.org/10.1109/I2CT.2018.8529468

[48] L. Wei, Empirical Bayes test of regression coefficient in a multiple linear regression model. Acta Math. Appl. Sin. **6**, 251-262 (1990). DOI: https://doi.org/10.1007/BF02019151

[49] C. Schexnayder, S.L. Weber, B.T. Brooks, Effect of truck payload weight on production. J. Constr. Eng. Manag. **125** (1), 1-7 (1999). DOI: https://doi.org/10.1061/(ASCE)0733-9364(1999)125:1(1)

[50] Z. Ge, Effectiveness of the T-test in multiple linear regression modeling of environmental systems. Environ. Eng. Sci. **26** (2), 377-384 (2008). DOI: https://doi.org/10.1089/ees.2008.0014

[51] K. Iqbal, D. Sun, Development of thermo-regulating polypropylene fibre containing microencapsulated phase change materials. Renew. Energy **71**, 473-479 (2014). DOI: https://doi.org/10.1016/j.renene.2014.05.063

[52] J. Jaccard, R. Turrisi, J. Jaccard, Interaction Effects in Multiple Regression. Sage, Thousand Qaks, CA (2003).

[53] R.L. Moy, L.S. Chen, L.J. Kao, Multiple Linear Regression, in: R.L. Moy, L.S. Chen, L.J. Kao (Eds.) Study Guide for Statistics for Business and Financial Economics: A Supplement to the Textbook by Cheng-Few Lee. John C. Lee and Alice C. Lee, Springer International Publishing, Cham, 223-240 (2015). DOI: https://doi.org/10.1007/978-3-319-11997-7_15

[54] D. Kyburz, C. Gabay, B.A. Michel, A. Finckh, The long-term impact of early treatment of rheumatoid arthritis on radiographic progression: a population-based cohort study. Rheumatology **50** (6), 1106-1110 (2011). DOI: https://doi.org/10.1093/rheumatology/keq424

[55] M. Lunt, Introduction to statistical modelling 2: categorical variables and interactions in linear regression. Rheumatology **54** (7), 1141-1144 (2015). DOI: https://doi.org/10.1093/rheumatology/ket172

[56] Y. Liu, J. Wang, Z. Wang, X. Lu, M. Avdeev, S. Shi, C. Wang, T. Yu, Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning. Acta Mater. **195**, 454-467 (2020). DOI: https://doi.org/10.1016/j.actamat.2020.05.001

[57] M. Ho Park, M. Ju, S. Jeong, J. Young Kim, Incorporating interaction terms in multivariate linear regression for post-event flood waste estimation. Waste Manag. **124**, 377-384 (2021). DOI: https://doi.org/10.1016/j.wasman.2021.02.004

[58] V.F. Navarro Torres, J. Ayres, P.L.A. Carmo, C.G.L. Silveira, Haul productivity optimization: An assessment of the optimal road grade. In: E. Widzyk-Capehart, A. Hekmat, R. Singhal (Eds.) Proceedings of the 27th International Symposium on Mine Planning and Equipment Selection - MPES 2018, Springer International Publishing, Cham, 345-353 (2019). DOI: https://doi.org/10.1007/978-3-319-99220-4_28

[59] X. Sun, H. Zhang, F. Tian, L. Yang, The use of a machine learning method to predict the real-time link travel time of open-pit trucks, Math. Probl. Eng. **2018**, 4368045 (2018). DOI: https://doi.org/10.1155/2018/4368045

[60] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, T. Herawan, Big data clustering: A review, in: B. Murgante, S. Misra, A.M.A.C. Rocha, C. Torre, J.G. Rocha, M.I. Falcão, D. Taniar, B.O. Apduhan, O. Gervasi (Eds.) Computational Science and Its Applications – ICCSA 2014. Springer International Publishing, Cham, 707-720 (2014). DOI: https://doi.org/10.1007/978-3-319-09156-3_49

[61] C. Wu, K.W. Chau, Y. Li, Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour. Res. **45**, W08432 (2009). DOI: https://doi.org/10.1029/2007WR006737

[62] S. Ma, G. Huang, K. Obais, S.W. Moon, W.V. Liu, Hysteresis loss of ultra-large off-the-road tire rubber compounds based on operating conditions at mine sites. Proc. Inst. Mech. Eng. D: J. Automob. Eng. **236** (2-3), 439-450 (2022). DOI: https://doi.org/10.1177/09544070211015525

[63] K. Drosou, C. Koukouvinos, Proximal support vector machine techniques on medical prediction outcome. J. Appl. Stat. **44** (3), 533-553 (2017). DOI: https://doi.org/10.1080/02664763.2016.1177499

[64] M. Cakir, M.A. Guvenc, S. Mistikoglu, The experimental application of popular machine learning algorithms on predictive maintenance and the design of IIoT based condition monitoring system. Comput. Ind. Eng. **151**, 106948 (2021). DOI: https://doi.org/10.1016/j.cie.2020.106948

[65] R. Tadeusiewicz, Neural networks in mining sciences – general overview and some representative examples. Arch. Min. Sci. **60** (4), 971-984 (2015). DOI: https://doi.org/10.1515/amsc-2015-0064