





RESEARCH ARTICLE

# Developing and evaluating predictive conveyor belt wear models

Callum Webb<sup>1</sup> , Joanna Sikorska<sup>2</sup> , Ramzan Nazim Khan<sup>3</sup>  and Melinda Hodkiewicz<sup>2</sup> 

<sup>1</sup>WearHawk Pty. Ltd., Western Australia, Australia

<sup>2</sup>Faculty of Engineering and Mathematical Sciences, University of Western Australia, Perth, Western Australia, Australia

<sup>3</sup>Department of Mathematics and Statistics, University of Western Australia, Perth, Western Australia, Australia

Corresponding author. Email: [callum.webb@wearhawk.com](mailto:callum.webb@wearhawk.com)

(Received 16 January 2020; revised 26 January 2020; accepted 28 January 2020)

**Keywords:** Conveyor belt; cross-validation; measurement; prediction; wear

## Abstract

Conveyor belt wear is an important consideration in the bulk materials handling industry. We define four belt wear rate metrics and develop a model to predict wear rates of new conveyor configurations using an industry dataset that includes ultrasonic thickness measurements, conveyor attributes, and conveyor throughput. All variables are expected to contribute in some way to explaining wear rate and are included in modeling. One specific metric, the maximum throughput-based wear rate, is selected as the prediction target, and cross-validation is used to evaluate the out-of-sample performance of random forest and linear regression algorithms. The random forest approach achieves a lower error of 0.152 mm/megatons (standard deviation [SD]=0.0648). Permutation importance and partial dependence plots are computed to provide insights into the relationship between conveyor parameters and wear rate. This work demonstrates how belt wear rate can be quantified from imprecise thickness testing methods and provides a transparent modeling framework applicable to other supervised learning problems in risk and reliability.

## Impact Statement

Conveyor belts are critical components in global supply chains in mining, power, and manufacturing industries. The belt is often the most expensive component of a conveyor system and downtime is costly. Predicting belt wear rate from conveyor design and operational parameters is useful because it allows operators to accurately forecast belt replacements on new conveyors, estimate wear rate on conveyors without adequate thickness data, and improve their understanding of how different variables influence or relate to belt wear. Our work demonstrates how such predictive models can be developed and evaluated.

## 1. Introduction

Conveyor belts are a cost effective method of transporting bulk materials in many industries worldwide. In the mining industry, conveyor belts are critical components of the supply chain, and the ability to estimate belt wear rates is important to ensure that risk of failure and maintenance activities are managed optimally.

Typically, belt thickness measurements are carried out periodically, often with ultrasonic probes while the belt is shut down. A common practice is to fit a straight line to the minimum thickness over calendar

time, producing an estimate of wear rate in millimeters per week. By extrapolating thickness over time in this way, a future belt replacement date can be estimated. This practice, although useful, has several limitations.

- Not all conveyors have periodic thickness measurements due to the cost of downtime and measurement.
- Wear is not linear with time for belts that are utilized inconsistently; belt throughput is a more robust basis for wear rate metrics.
- Wear rate can only be estimated accurately when enough thickness measurements for a belt have been taken.
- Wear rates on similar, new conveyor installations cannot be accurately predicted as their design and operating characteristics usually differ to existing, monitored belts.

In this article, we describe a large dataset assembled from a mining company operating in Western Australia, which includes conveyor design, operational parameters, belt thickness measurements, and conveyor throughput captured by a material tracking system. We propose new wear rate metrics and apply one of these metrics (worst-case wear, millimeters per million tons of throughput) to estimate the wear rates of 165 belt lifetimes from 95 unique conveyor systems based on our data. Finally, we compare linear regression and random forest models for predicting wear rates of out-of-sample conveyors using repeated  $k$ -folds cross-validation. Our objective is to present a framework for evaluating predictive models, and to assess if design and operational factors, commonly available and measured on site, support conveyor belt wear rate prediction.

## 2. Background

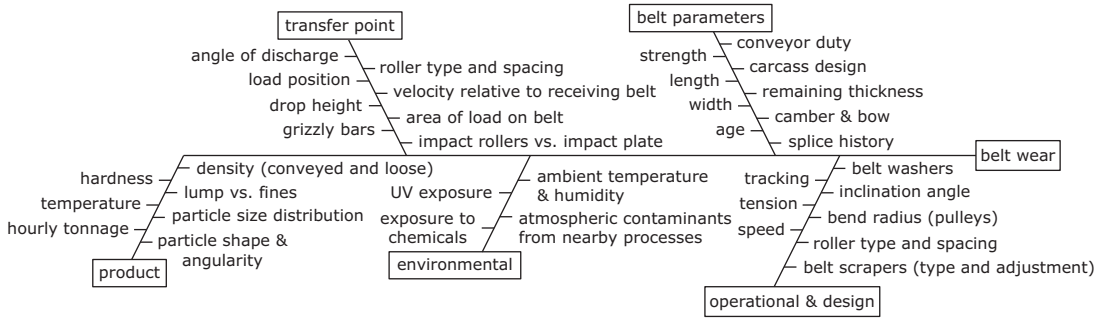
### 2.1. Conveyors

Conveyor belts are made of several layers of different material selected for the operating conditions. Typically, the belt comprises an inner load-bearing carcass of fabric or steel cords, surrounded by top and bottom rubber covers. The carcass is the principal structural component of the belt, providing tensile strength, while the rubber covers protect the carcass from damage. The top cover carries the bulk material and the thinner bottom cover is supported by pulleys, freely rotating idlers, and drive systems. Long belts are often made up of shorter belt segments spliced together, which can be replaced individually or as a set. It is common for the belt to be the most expensive component of a conveyor system.

Wear of the top cover is predominantly due to abrasion by the bulk material, resulting in a reduction in thickness with usage (Schallamach, 1954). In practice, the rate of wear is not constant throughout the life of the belt due to changing operating conditions, bulk material properties, or maintenance practices. Belts are considered worn-out when the thickness of the top cover has reached a threshold (e.g., 3 mm) at any point across the width of the belt. The belt is then either replaced or refurbished by re-coating the worn top of the belt with new rubber. While belts can fail in a variety of modes, this work focuses on belt working life as limited by top cover wear.

Belt working life is generally a function of the belt materials, the conveyor design and operation, maintenance, and the physical properties of the material being transported. As part of our work, a panel of subject matter experts were consulted with a view to extend the list of factors influencing belt wear rate that are available in the literature (All State Conveyors Pty Ltd, 2018; Masaki et al., 2018; Metso, 2016; Molnar et al., 2014; Schallamach, 1954). This extended list is shown in a fishbone diagram of the factors (Figure 1). These factors are grouped by product (the characteristics of the conveyed product), environmental factors (location and operating environment), belt parameters (design and operational history), conveyor operational and design factors, and finally features associated with the transfer point design. The relative importance of these factors was not determined by the subject matter experts, and it is not known a priori which of these factors, or combination of factors, are most influential.

A limited range of studies on conveyor belt wear and damage has been published. Andrejiova and Marasova (2013) found that belt length and transported quantity of material were associated with belt



**Figure 1.** Fishbone diagram showing the inputs to conveyor belt wear.

service life based on exploration of data from 30 conveyors at a quarry. Further experimental work with impact drop hammers explored the significance of belt damage from falling material at the belt loading point (Andrejiova et al., 2014). Another experimental study by Andrejiova et al. (2016) examined the effects of selected factors including belt storage on belt wear.

While the existing literature establishes understanding of relationships between some factors and belt life, we have neither found clear or rigorous definitions for wear rate metrics that are useful in practice, nor work that focuses on predicting wear rate and quantifying the uncertainty in prediction performance estimates.

## 2.2. Thickness measurement

The current approach to planning belt replacement is to extrapolate minimum top cover thickness over time. Conventionally, a time of flight ultrasonic probe is used to measure the top cover thickness of a belt when the conveyor is shut down. Shorter belts are frequently only measured at one (*longitudinal*) position along the belt length, whereas longer belts made up of spliced segments may have multiple longitudinal measurement positions, often located near splices which serve as landmarks. At each longitudinal position, measurements are taken across the belt width (*transverse direction*) at equally spaced intervals producing a picture of the top cover cross-section.

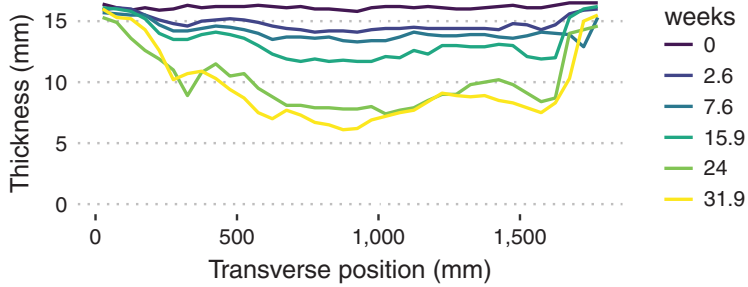
Several factors complicate the analysis of these data.

- It is difficult to guarantee that thickness measurements are taken at the same longitudinal position each time.
- The belt cross-section is known only at the (longitudinal) measurement positions.
- Data from ultrasonic probes are noisy and subject to measurement and calibration errors.
- Belts made up of multiple segments may have splices that were installed at different times.

Figure 2 shows a sample of data from a belt displaying wear patterns that are typical for a conveyor system. The thickness should decrease over time. However, some crossover of the lines is seen in the figure, indicating measurement error.

## 3. Data

The data collected for this work comprise 165 belts on 95 conveyors from two different bulk material handling sites in the north-west of Western Australia over a period of 8 years. The target variable, wear rate (max mm/megatons [Mt]), was derived from 41,652 ultrasonic thickness measurements and 406,572 material movements. Not all of the factors shown in the fishbone diagram had readily available data. A subset of factors for which data could be collected within time and resource constraints was selected for this work with a preference for factors that subject matter experts expressed a prior belief of being more important, or a greater interest in testing.



**Figure 2.** Top cover thickness of an 1,800-mm wide belt over time. Measurements are spaced 50 mm apart, but elapsed time between measurements is irregular.

The eight explanatory variables include seven numeric (belt width, belt length, belt speed, belt strength, drop height, % fines, and loading frequency) and one categorical (conveyor duty). Loading frequency is calculated as belt speed divided by length.

The data were obtained from three distinct sources.

- Ultrasonic belt thickness measurements.
- A material tracking system, recording tonnages of bulk material movements through a supply chain.
- Conveyor specifications, assembled from various asset registers and design drawings.

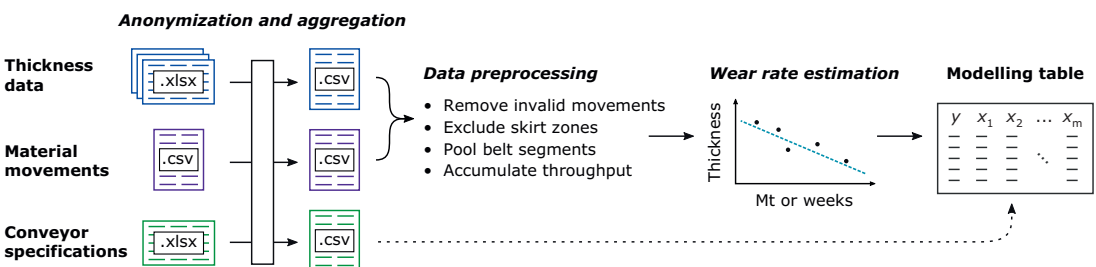
A data preparation pipeline was developed using the R programming language (R Core Team, 2019) to take the data from raw spreadsheets and csv files and produce a *modeling table* with one row for each conveyor belt, a column for wear rate (the prediction target), and columns for each explanatory variable.

This process spanned three main stages which are shown in Figure 3: (a) anonymization and aggregation for de-identifying sensitive commercial information and merging spreadsheets into a single csv file; (b) data preprocessing, where data were cleaned and merged, including necessary transformations to model the relationship between belt thickness and throughput; and (c) wear rate estimation, an intermediate modeling stage where belt wear rate metrics are calculated.

Selected parts of the pipeline that are relevant to understanding the model are discussed herein.

### 3.1. Thickness data

Thickness measurements were collected at irregular intervals by specialist contractors from each conveyor while it was shut down. Measurements were taken at one or more longitudinal positions and at 50 mm intervals across the width of the conveyor. These measurements were stored in spreadsheets; each instance of a conveyor belt life had, in principle, its own file, which was maintained until that belt was



**Figure 3.** Data preparation process overview. Several approaches to wear rate estimation are discussed in this article, but only one is used for predictive modeling.

replaced. Spreadsheets also recorded the date that the belt was installed on the conveyor. A total of 78,557 thickness records were extracted from 243 spreadsheets into a single table and anonymized.

Measurements at points within 400 mm of either belt edge were excluded since excessive wear can occur in these regions if conveyor skirts are improperly adjusted. Records were also removed if any of the following conditions were met: (a) limited material tracking history was available; (b) thickness data were available but tonnage records were not; or (c) the belt had less than three measurements with valid tonnage values. The final dataset comprises 41,652 records.

Plotting the thickness data for each conveyor revealed multiple sets of measurements (i.e., recorded in separate spreadsheets) from the same conveyor on the same date with identical belt installation dates. In total, 79 spreadsheets were affected, attributable to 18 separate belts. As these records had different thickness values, they were identified as separate sets of measurements and not merely duplicates. We made the assumption that multiple measurements had been collected from different longitudinal positions on particularly long or otherwise important belts. We refer to all measurements that have been collected from the same conveyor over the same time frame with the same belt installation date, as a pool. We assume that all measurements in a pool relate to the same conveyor belt. The concept of a pool allows us to accommodate multiple measurements on the same belt as described above.

### 3.2. Throughput data

Tonnages are measured by weightometers positioned throughout the supply chain and were extracted from a material tracking system, producing 475,191 records of bulk material movements over a continuous 8-year period. Each record captures the source and destination of the material, the product type (lump or fines), and IDs of all the equipment in between. Timestamps marking the beginning and end of the movement, the total tons, and variables describing the type of material are also recorded. The equipment IDs were matched with conveyors to calculate throughput between belt thickness tests. These data were cleaned by considering only movements with non-zero duration and a calculated throughput rate between 300 and 15,000 tons per hour. It is worth noting that the frequency of calibration of the weightometers is unknown, but since these material movements are an important value driver for the company we have made the assumption that the error in weightometer readings is negligible in our analysis.

### 3.3. Conveyor specifications

Conveyor and belt design and operational parameters were collated from asset registers, belt thickness files, and design drawings. These included: belt width (mm); belt length (m); belt tensile strength (kN/m); manufacturer's belt material grade; drop height (m); and conveyor duty.

Conveyor duty is a categorical variable that classifies conveyors by a broad collection of design parameters related to the application or *duty* of the conveyor. When conveyors are designed, it is usually assumed that equipment belonging to the same duty-class will have similar operating requirements. Some conveyor duties had only a few observations, so these were merged with other conveyors that were similar in design. Specifically, Wharf and Tunnel duties were combined with Yard, and Car dumper conveyors were combined with Transfer conveyors. After this merging process, conveyor duty has five discrete categories: Yard, Transfer, Stack, Reclaimer, and Shiploader. The number of conveyors and belts across duties is shown in Table 1.

Belt material grade was initially considered as a categorical variable, but was omitted as the data were deemed to be insufficiently trustworthy. Belt grade had been assigned by plant staff retrospectively based on the memory of key engineers, rather than accurately recorded from the markings on the belt at the time it was installed. In more than half of the records, this value did not agree with other information that independently indicated a different grade.

Drop height is the vertical distance between the tail pulley of the conveyor and the head pulley of the upstream loading conveyor. Subject matter experts agreed that the relative velocity of the bulk material and belt at the point of loading is an important factor for belt wear, and this vertical distance was a crude but simple attempt to characterize this. It is crude because the design of the transfer chute is ignored.

**Table 1.** Summary of conveyor and belt counts in dataset by conveyor duty.

Conveyor duty	Conveyor count	Belt count
Reclaimer	5	23
Shiploader	6	27
Stacker	7	22
Transfer	56	60
Yard	21	33

#### 4. Belt Wear Rate Estimation

To enable predictive model development, wear rate metric(s) must be defined and applied to the raw data to produce a target variable for prediction. Any wear rate metric will be the ratio of two quantities: some measure of thickness and another quantity that measures belt utilization.

At the mine site operations in this study, remaining life of belts in service was estimated using the slope of the regression line of minimum belt thickness over time. This corresponds to a wear rate metric with units millimeters/week, and could be applied to historical data to summarize the lifetime wear rate of belts. However, this metric has some limitations.

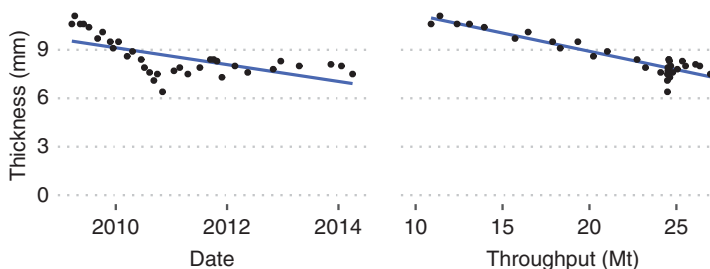
First, conveyor utilization is not necessarily uniform over the life of a belt; periods of inactivity or changes in production plans can result in wear that is only piece-wise linear with time. In this situation, a single estimate of belt lifetime wear rate based on a linear model over the entire time period is not meaningful (see Figure 4).

Second, the minimum thickness at a particular transverse position on a belt can only be obtained by measuring at all points across the belt. Since this is not common practice, the minimum thickness may not be known. In addition, the transverse position of the point of minimum thickness often moves between measurements, and this introduces further error in the measurements.

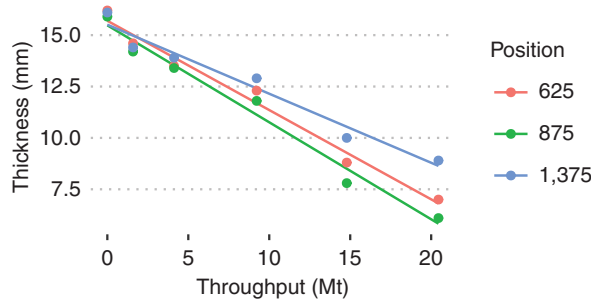
We use the slope of a regression line approach to estimate belt lifetime wear rate, but propose four alternative wear rate metrics by considering two different methods for using the thickness data to calculate the numerator of the ratio, and adding throughput as an alternative to time for the denominator. One metric, maximum throughput-based wear rate, is used for further modeling.

##### 4.1. Maximum wear rate

This metric aims to provide a worst-case estimate of wear rate that avoids loss of physical interpretability and reduces the risk of bias due to random noise.



**Figure 4.** Throughput rate for this conveyor drops at the beginning of 2011 resulting in a reduced rate of wear with time and poor linear fit. Regressing thickness against cumulative throughput produces a good fit (thickness data taken from a single measurement position).



**Figure 5.** Maximum wear rate is estimated to be the slope of the steepest regression line, in this case at position 875. Only three positions are drawn for clarity.

Instead of using minimum thickness, data from a pool are first grouped by transverse position. For each such group, the measured thickness is regressed against either time or throughput, producing an individual estimate of wear rate at each measurement position.

The slope of the steepest line is taken to be the maximum wear rate. Figure 5 illustrates this process for three positions (in practice, an 1,800 mm wide belt will have 20 measurement positions after skirt zones are excluded).

#### 4.2. Mean wear rate

Maximum wear rate effectively discards thickness data from all but one measurement position; the motivation for this metric is to use all the data by estimating the mean wear rate. We first estimate the belt cross-section area *A* using the trapezoidal rule:

$$A \approx \sum_{k=1}^N \frac{t_{k-1} + t_k}{2} \Delta x = \frac{\Delta x}{2} (t_0 + 2t_1 + 2t_2 + \dots + 2t_{N-1} + t_N), \tag{1}$$

where  $\Delta x$  is the transverse (equal) spacing between measurements, *N* is the total number of measurements across the belt width, and  $t_k$  is the thickness at the *k*th measurement position. Removing measurements within 400 mm of the belt edge can be considered equivalent to dropping the first and last eight  $t_k$  values—we will index the remaining measurements by  $k'$  and  $N'$ . Normalizing by remaining belt width then yields:

$$A' = \frac{\Delta x}{2} \frac{1}{N' \Delta x} \sum_{k'} 2t_{k'} = \frac{\sum_{k'} t_{k'}}{N'}, \tag{2}$$

which is the mean thickness. Wear rate estimation proceeds by grouping thickness data from a pool by measurement date, calculating the mean thickness, and regressing against time or throughput. The slope of this line is the rate of change of mean belt thickness.

#### 4.3. Summary of metrics

The four wear rate metrics (maximum and mean wear rates each with respect to time and throughput) were calculated for the set of pooled belt lifetimes and summary statistics from the regression analyses are shown in Table 2. Along with this tabular summary, 116 plots of the four regression models were visually inspected. Generally, a linear model for wear rate provides a good fit for the data and explains a large proportion of the variation in the thickness metric.

**Table 2.** Summary statistics after calculating wear rate metrics using the set of pooled belt lifetimes.

Metric	$\overline{R^2}$	SD ( $R^2$ )	$\overline{SE}$
Throughput-based			
Max (mm/Mt)	0.908	0.124	$4.48 \times 10^{-2}$
Mean (mm/Mt)	0.895	0.152	$3.94 \times 10^{-2}$
Time-based			
Max (mm/week)	0.904	0.125	$1.50 \times 10^{-2}$
Mean (mm/week)	0.892	0.151	$1.29 \times 10^{-2}$

The sample mean  $R^2$  value, sample SD of  $R^2$ , and sample mean standard error of regression slope are evaluated on the sample of 165 belt lifetimes.

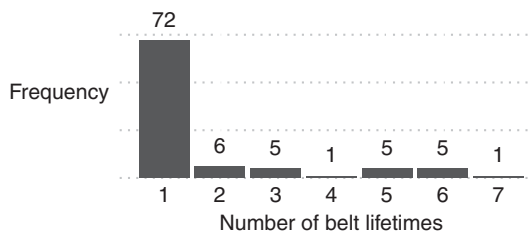
There is little difference in model fit between time- and throughput-based metrics, and this is a reflection of the fact that in our data most conveyors are utilized consistently. Mean wear rates have a slightly lower mean  $R^2$  value, but are still a strong fit to the data. However, we select maximum wear rate as our metric, because if any point over the cover wears out the belt needs replacement. Further, throughput-based metrics are less sensitive to changes in production plans and non-uniform conveyor utilization, and are therefore more useful for forward planning. That is, we select throughput-based maximum wear rate as the appropriate metric to model belt wear rate.

**4.4. Exploratory data analysis**

Explanatory variables were explored to identify any relationships in the data, inform the modeling process, and confirm data fidelity. The range of thickness measurement dates is known to span multiple belt lifetimes for some conveyors. Figure 6 shows the distribution of the number of belt lifetimes in the thickness data over the 95 conveyors.

While most conveyors only have a single belt (i.e., a single row in the modeling table), 23 conveyors have more than one belt lifetime. The modeling process will be designed to eliminate a potential source of bias by ensuring belt lifetimes from the same conveyor do not appear simultaneously in both data used to fit models and data used to evaluate the models.

A correlation matrix for the continuous explanatory variables is shown in Figure 7. Belt length and load frequency are strongly inversely correlated, which is expected given that loading frequency is equal to the quotient of belt speed and length. Belt length and strength are positively correlated, and this may reflect belt tension requirements, which increase with belt length for optimal conveyor operation. Speed and width are negatively correlated, again an expected result. Wider belts have greater carrying capacities and throughput rate for a fixed speed, and volumetric flow rates of ore across a chain of belts must be



**Figure 6.** Bar plot of number of lifetimes per conveyor. Most conveyors (72) in the data only have a single belt lifetime.



	% Fines				
Drop height					0
Load freq				-0.4	-0.1
Belt length			-0.8	0.3	0.1
Belt speed		0	0.1	-0.2	0.1
Belt strength	-0.1	0.4	-0.3	0	0
Belt width	0.2	-0.5	-0.1	0.2	0.1

**Figure 7.** Pair-wise Pearson correlation (rounded to one decimal place) between continuous explanatory variables.

consistent to avoid overloading or starving equipment. All variables are hypothesized to contribute in some way to explaining wear rate, and therefore will be included in modeling.

The relationship between conveyor duty, a categorical variable, and the other continuous explanatory variables is shown in Figure 8. Some observations from these plots include:

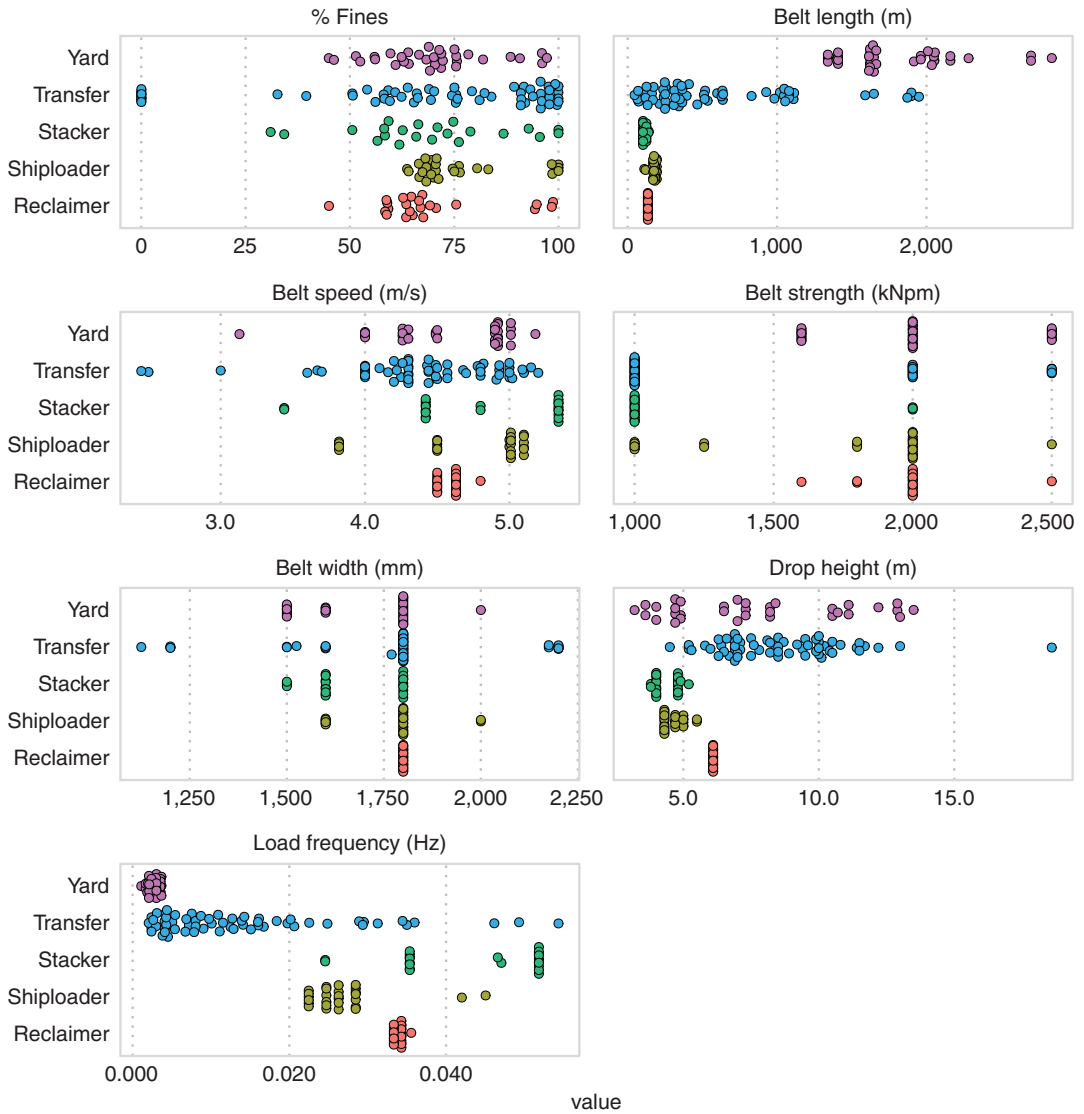
- stacker, shiploader, and reclaimer conveyors have similar short lengths;
- while belt width and strength are in principle continuous quantities, only a few fixed values are present in our data (perhaps specific to the operation);
- all reclaimer belts are 1,800 mm wide; and
- transfer (including car dumper) and yard (including wharf and tunnel) belts have the greatest range of design parameters.

Conveyor duty acts as a proxy for a host of design and operational factors that make conveyors unique. Duties that have a large range of continuous variable values (e.g., transfer belts) could be scrutinized more closely in future work to identify aspects that better discriminate conveyors and potentially produce more accurate predictions. If data were available, it may be better to eliminate duty and instead enumerate the underlying specific factors that they reflect.

## 5. Model Evaluation Framework

The goal is to develop a model capable of predicting maximum wear rate (mm/Mt) from the explanatory variables and estimate its performance on out-of-sample conveyors. Prediction performance in this article is measured as the root mean square error (RMSE) for predictions based on the model for conveyor belts unseen in model training. We rank models by the percentage improvement (decrease) in RMSE over the uninformative *null* models (which simply predict the mean wear rate). RMSE was chosen for ease of interpretation (its units are mm/Mt), and because it is a common metric in regression analysis. The percentage improvement over the null algorithm is included to provide a baseline level of performance and to quantify the additional predictive power of a model.

We use a cross-validation framework to test two modeling algorithms: linear regression (ordinary least squares) and random forest. The framework presented in this article can be applied to estimate the performance of any type of algorithm for building predictive models. Linear regression was chosen for its simplicity and prevalence in prediction problems, and random forest was included to test whether such a nonlinear, more flexible modeling approach would produce better predictions for this problem. The intent is neither to make definitive statements about the comparative performance of linear regression and random forests, nor to find the “best” model through extensive tuning, but rather to demonstrate how this



**Figure 8.** Swarm plot showing the distribution of continuous explanatory variables by conveyor duty.

model evaluation framework can be used to compare modeling algorithms, and to show how insights into the relationship a model has “learned” from the data can be extracted.

**5.1. Repeated *k*-fold cross-validation**

To estimate out-of-sample prediction error, a model should be evaluated on data that were not used to build the model (Hastie et al., 2009; Hurvich and Tsai, 1990). We use repeated *k*-fold cross-validation, which involves splitting the data into *k* subsets of roughly equal size. A single subset is set aside as the test set, and the remaining *k* – 1 subsets are used as a training set to fit a model. The *loss function* (in our case, RMSE) is evaluated on the test set. This is repeated for each subset to produce *k* loss function values. For higher precision, the entire process is repeated *n* times with different splits, resulting in *nk* loss function values. The arithmetic mean of this set is taken to be the estimate of out-of-sample prediction error for a

model-fitting algorithm; we refer to this value as the *cross-validation statistic*. The standard deviation (SD) of the  $nk$  loss function values is also computed to provide a measure of model performance stability.

Stated more precisely for regression, let  $D$  denote the sampled data,

$$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \quad y_i \in \mathbb{R}, \quad (3)$$

where  $y_i$  is the prediction target value and  $\mathbf{x}_i$  is the corresponding vector of explanatory variables. Let  $h_D: \mathbf{x} \mapsto \hat{y}$  be a trained model instance that maps new observations  $\mathbf{x}$  to predicted values  $\hat{y}$ , obtained by inputting data  $D$  to an algorithm  $\mathcal{A}: D \mapsto h_D$ . The loss function  $L: (y, \hat{y}) \mapsto \mathbb{R}$  evaluates predictions  $\hat{y} = h_D(\mathbf{x})$  against known values  $y$ .

After partitioning  $D$  into  $k$  subsets, let  $D^i$  denote the data in the  $i$ th subset, and  $D^{-i} = (D \setminus D^i)$  denote the data with the  $i$ th subset removed. Let  $\mathbf{x}^i, y^i \in D^i$  and  $\mathbf{x}^{-i}, y^{-i} \in D^{-i}$  refer to the explanatory variable vectors and corresponding prediction target values respectively of these mutually exclusive subsets of  $D$ . If the number of repeats  $n = 1$ , the out-of-sample prediction error or cross-validation statistic for the algorithm  $\mathcal{A}$  is calculated as

$$\frac{1}{k} \sum_{i=1}^k L(y^i, h_{D^{-i}}(\mathbf{x}^i)). \quad (4)$$

When  $n > 1$ , the entire process is repeated with  $n$  different sets of  $k$  partitions of  $D$ . The resulting  $nk$  loss function evaluations are pooled and the arithmetic mean calculated. The optimal number of folds is often reported as being  $5 \leq k \leq 10$  (Hastie et al., 2009; Kohavi, 1995); in this article, we use  $k = 10$  and  $n = 100$ .

We expect wear rates of different belts from the same conveyor to be highly correlated. This could lead to a potential source of bias in the results if different belts from the same conveyor appear in both training and test sets within the cross-validation process. To eliminate this, we form the  $k$  subsets by shuffling conveyor IDs and pairing each with a value from the sequence  $\{1, 2, \dots, k, 1, 2, \dots\}$  until the conveyor IDs are exhausted. Belts from each conveyor are assigned to the corresponding subset, producing  $k$  subsets. With 95 conveyors and  $k = 10$ , the result is five subsets containing nine conveyors and five subsets of 10 conveyors. Most subsets will also be roughly equal in size (number of belt lifetimes), with a small number of larger subsets containing conveyors with multiple belt lifetimes. The resulting subsets are then examined to verify that each level of the conveyor duty variable appears in at least two distinct subsets, to ensure that the modeling algorithm is always trained with data that contains every conveyor duty. This is particularly important for linear regression, where at least one observation of each level of categorical variable is required to estimate all of the coefficients. If this condition is not met, the random split is rejected, conveyor IDs reshuffled, and the process repeated until a valid set of subsets are produced. This process of generating subsets was automated and run  $n$  times, and the same random splits were used for assessing both linear regression and random forest algorithms.

In practice, after cross-validation the same algorithm  $\mathcal{A}$  is applied to the entire dataset to produce a model instance that could be used in a production setting. In this sense, we use repeated  $k$ -fold cross-validation as a framework for estimating the performance of a model fitting *process* (algorithm), rather than any particular model instance. Because the cross-validation framework estimates loss using a model trained on a subset of the available sample, and model performance typically increases with the size of training data, we expect our estimate to include some *pessimistic bias* (Kohavi, 1995).

## 5.2. Modeling algorithms

The first algorithm tested is ordinary least squares linear regression, which was chosen for its simplicity and wide applicability. The second algorithm tested is a random forest, implemented in the R package *randomForest* (Liaw and Wiener, 2002). A random forest is an ensemble model consisting of a collection of decision tree predictors, each trained on a random subset of the training data sampled with replacement to a size equal to that of the training data (a *bootstrap sample*). Decision trees are grown by recursively

splitting each terminal node of the tree, where the optimal split is found by searching over  $mtry$  randomly selected predictor variables, until a tree depth limiting condition is met (minimum node size or maximum number of terminal nodes). Predictions from a random forest are made by passing the vector of explanatory variables into each decision tree making up the forest and averaging the terminal node values.

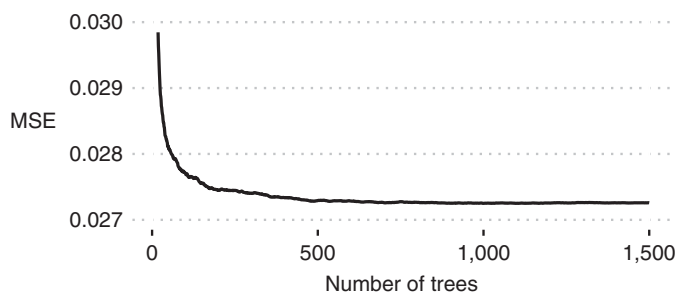
A feature of random forests is the *out-of-bag* error estimate. In a bootstrap sample of size  $n$ , the probability of any particular observation being selected in the first draw is  $1/n$ . Therefore, the probability that an observation is *not* selected is  $1 - 1/n$ . It follows that for the  $n$  independent draws making up the bootstrap sample, the expected proportion of data excluded is  $(1 - 1/n)^n$ . It can be shown that this proportion approaches  $1/e \approx 1/3$  as  $n$  grows to infinity. These data are “out-of-bag” (OOB), and therefore can be used to estimate prediction performance similarly to  $k$ -fold cross-validation.

There are several parameters of a random forest that can impact performance, including:  $ntree$ , the number of trees to grow;  $mtry$ , the number of randomly chosen candidate variables for splitting; and  $nodesize$ , the minimum size of terminal nodes. The  $ntree$  parameter is understood to increase model performance for regression with more trees at the cost of computation time, converging to a steady maximum beyond which additional computational effort does not afford any improvement (Breiman, 2001). The OOB error as a function of  $ntree$  for our data is shown in Figure 9, demonstrating that a plateau is reached after roughly 500–1,000 trees. We set  $ntree$  to 1,000 for which the computational cost was not prohibitively high for our data and hardware.

The *randomForest* package provides a function *tuneRF* for tuning the  $mtry$  parameter to minimize the OOB error, which we use inside the training step of the cross-validation process. It is important to note that the cross-validation statistic is itself a random variable, and algorithm or tuning parameter selection based on its value can introduce optimistic bias, resulting in underestimating the out-of-sample prediction error and effectively over-fitting to the test data (Cawley and Talbot, 2010). Because the OOB data are still limited to the  $k - 1$  training parts and do not overlap with the test data, tuning  $mtry$  by minimizing the OOB error inside the cross-validation process does not introduce an optimistic bias in this way, and can be considered part of the training process. We will use the cross-validation estimate to compare linear regression and random forest algorithms. However, as this involves only a single pre-specified comparison, we assume any optimistic bias associated with selection using the cross-validation statistic is negligible in this instance.

A limitation of using the OOB error in training to tune the random forest is that the bootstrap sampling process is completely random and does not respect the same constraints established in the cross-validation splitting. Specifically, two belts from the same conveyor can be both in-bag and out-of-bag, potentially resulting in sub-optimal selection of tuning parameters for out-of-sample prediction performance. Furthermore, we will not attempt to tune the value of  $nodesize$ , and instead use the package default value of 5. Implementing a modified bootstrap sampling process and tuning  $nodesize$  are two potential areas for future work.

The optimal value of  $mtry$  found in each  $nk$  trained random forest is shown in Table 3. For our data, a value of 1 was optimal in roughly 80% of forests, which is equivalent to randomly selecting a variable at



**Figure 9.** Out of bag prediction error as a function of the number of trees in the random forest. At each point, 100 forests were grown and errors averaged to produce a smooth curve. Error decreases monotonically and reaches a plateau.

**Table 3.** Stability of optimal *mtry* values, the number of variables randomly sampled as candidates at each split when growing trees.

Optimal <i>mtry</i> value	Count
1	795
2	204
4	1

This parameter is tuned to minimize out-of-bag error on training data in every cross-validation fold. The value is 1 almost 80% of the time.

each node for splitting when growing regression trees, and minimizes correlation between trees in the forest. The function *tuneRF* starts with  $mtry = \lfloor n_x/3 \rfloor = 2$ , (where  $n_x$  is the number of explanatory variables and  $\lfloor \cdot \rfloor$  is the floor function which returns the integer part of a real number) and searches by inflating and deflating this value by a factor of 2, which is why 3 does not appear in the results.

Because tuning *mtry* is a part of the model training process, in practice when training a random forest in a production setting using all of the data, *mtry* should not be fixed to 1; instead the same tuning process should be repeated. Analyzing the distribution of optimal values is still informative and provides a measure of model stability.

## 6. Results and Discussion

This section presents the results of the model evaluation framework and provides some insights into the learnt relationship between wear rate and the explanatory variables of the random forest.

### 6.1. Prediction error

The sample mean loss function value (cross-validation statistic), sample SD of loss values, and the percent improvement in RMSE over the uninformative null model are summarized in Table 4.  $\overline{\text{RMSE}}$  were calculated over mean loss values for each test set, pooled over each repeat of the cross-validation process (a total of  $kn$  values) to measure performance stability.

Both random forest and simple linear regression outperformed the null model, demonstrating that the explanatory variables contain predictive information for wear rate that generalizes to out-of-sample conveyors. Random forest performed better than simple linear regression ( $\overline{\text{RMSE}}$  of 0.152 vs. 0.160), though the difference ( $7.75 \times 10^{-3}$ ) is much smaller than the SD of the difference ( $2.97 \times 10^{-2}$ , calculated over  $kn$  RMSE pairs), suggesting no difference between the performance of the two methods.

The random forest has a median  $R^2$  value of 0.58, calculated against test data over each cross-validation split (SD = 0.29). This shows that while the model is effective in using information from the explanatory variables to narrow the likely range of belt wear rate ( $R^2 > 0$ ), a relatively large amount of unexplained variance remains, suggesting that there are possibly other factors not in our data that could be gathered to produce better predictions.

**Table 4.** Prediction performance results.

Model type	$\overline{\text{RMSE}}$	SD ( $\overline{\text{RMSE}}$ )	% decrease over null
Null	0.284	0.0976	–
Linear regression	0.160	0.0557	43.6
Random forest	0.152	0.0648	46.3

Because the prediction target  $y$  (belt wear rate measured as slope of regression line of measured thickness over throughput) used to train the model is a variable with nonconstant variance, the accuracy of the predictive model should be interpreted carefully. That is, an assumption of linear regression is homoscedasticity, and this is likely to be violated here. Also, a thorough approach to estimating the overall prediction error with respect to the true physical wear rate (estimated by  $y$ ) would include information about the uncertainty of both the predictive model and the regression process used to generate wear rates. The mean standard error of the regression slopes is sufficiently small ( $4.48 \times 10^{-2}$ ) relative to the RMSE of the linear regression ( $1.60 \times 10^{-1}$ ) and random forest ( $1.52 \times 10^{-1}$ ) models, so we can ignore the effect of the intermediate regression modeling in estimating the wear rates.

## 6.2. Variable importance and effects

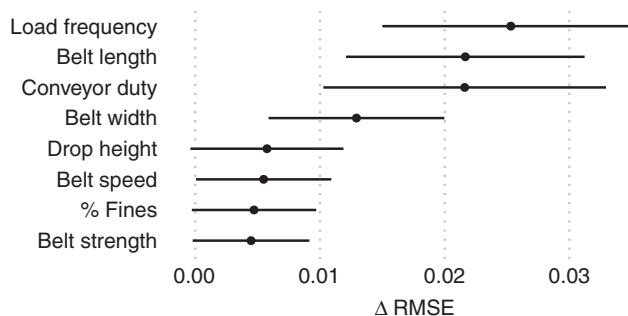
While our primary purpose of modeling in this problem is prediction and not inference, it is often desirable or even essential for model predictions to be explainable. It is also useful to understand the relative importance of input variables and how they relate to the predicted values. One method for assessing the importance of variables that can be applied regardless of the choice of algorithm is the permutation method.

The permutation method takes a trained model instance and proceeds as follows.

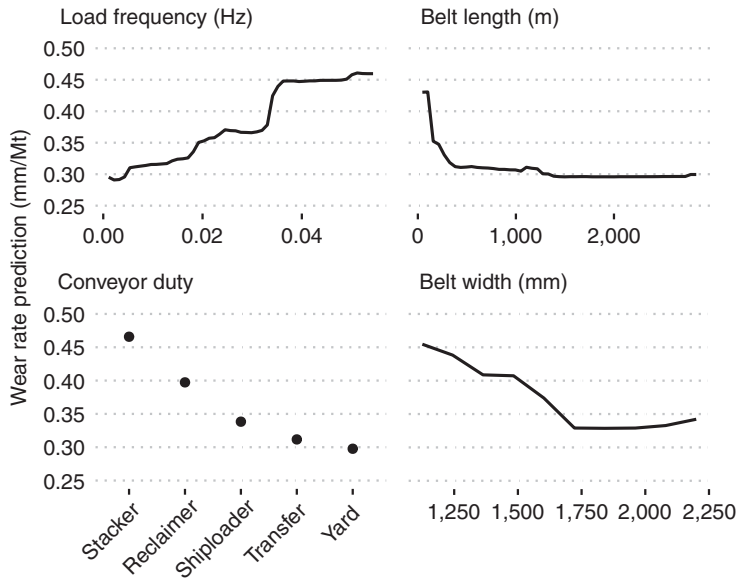
1. Calculate the loss (RMSE) on a dataset unseen in training.
2. Shuffle (permute) the values of a single explanatory variable, and recalculate the loss on the same data using the shuffled values.
3. Store the deterioration in loss after shuffling, restore the original ordering of values, and repeat the process for remaining explanatory variables.

A version of this method is implemented by the *randomForest* package that makes use of OOB data. However, as previously discussed, the OOB data can still include conveyors seen in training which could bias results. Instead of using this method, we reuse the repeated cross-validation procedure and demonstrate an approach that is compatible with any algorithm. The permutation method steps are repeated in each cross-validation partition and the difference in RMSE before and after permuting each explanatory variable is stored resulting in  $nk$  values for each variable. The mean and SD of the change in model RMSE for the random forest is shown in Figure 10.

The limitations of this technique should be considered when interpreting these results. Most importantly, they should not be taken to be a direct answer to questions about the importance of factors for the underlying physical process of belt wear. Controlled experiments are a better approach for such questions. Instead, the results reflect how important the variables are for prediction performance for a specific choice of algorithm and performance metric. Second, permutation importance tends to spread importance across collinear variables and shows bias toward categorical variables with many levels (Altmann et al., 2010; Strobl, 2007).



**Figure 10.** Permutation importance results, shown as the deterioration in RMSE as a result of shuffling each variable. Larger values indicate the variable is more important to prediction accuracy. Lines represent one SD of  $\Delta$ RMSE over each test set in cross-validation.



**Figure 11.** Partial dependence plots of random forest model with top four variables.

These results suggest that load frequency, belt length, and conveyor duty are the most important variables for prediction accuracy. The necessarily strong correlation between length and load frequency makes it difficult to separate the importance of these two variables; Strobl et al. (2009) and Strobl (2007) suggest a conditional permutation approach which may provide greater insight. Conveyor duty is the next most important variable, indicating that further efforts in “unpacking” duty into individual variables characterizing the conveyor may provide clearer insight into the effect of conveyor duty on wear rate. Belt width was expected to be an important variable due to its effect on belt loading and capacity, though it is only moderately important relative to other variables. Belt speed, strength, drop height, and % fines are similarly low in importance in these models.

Finally, to provide some insight into the relationship between the most important variables and predicted wear rate, partial dependence plots shown in Figure 11 were generated using the *pdp* package for R (Greenwell, 2017). These figures correspond to a random forest model with fixed parameters  $mtry = 1$  and  $ntree = 1,000$ . The joint effects of conveyor duty paired with loading frequency, belt length, and belt width were inspected, along with individual conditional expectation plots (Goldstein et al., 2013), which did not suggest the presence of any remarkable interactions. Therefore, only single predictor partial dependence plots are shown.

Figure 11 shows that predicted wear rates increase with greater load frequencies and decrease for longer and wider belts. Stacker conveyors have the highest wear rate among the duty levels, and yard conveyors have the lowest. The range of effect size for load frequency and conveyor duty are similar and larger than for belt length and width. The relatively large effect size of load frequency and its reasonably linear relationship to wear rate, coupled with the lack of remarkable interactions may explain why the performance difference between the random forest and linear regression models is small.

## 7. Conclusions

The ultrasonic belt thickness data, while relatively low in resolution, show wear patterns that are typically strongly linear over the belt lifetime. Combining these data with material tracking records support throughput-based metrics, which are more robust in cases where conveyor utilization is inconsistent.

We trained linear regression and random forest models to learn a relationship between conveyor specifications and worst-case wear rate that generalizes to out-of-sample conveyors. The random forest approach performed slightly better, with  $\overline{\text{RMSE}} = 0.152$  (46.3% better than an uninformative model with no explanatory variables), and a median  $R^2$  value of 0.58 on unseen conveyors. This delivers an improved ability to predict wear rates of new conveyor installations, or for those conveyors where thickness data are unavailable.

Variable importance measures derived from the permutation method indicate that belt length (both directly and indirectly through the calculated loading frequency variable) and conveyor duty contribute the most to the random forest's predictive accuracy, followed by belt width, with belt speed, strength, drop height and % fines being relatively less important. It is unsurprising that load frequency is important, as the loading zone is known by subject matter experts to be where most belt wear occurs. Conveyor duty is a catch-all variable that describes a broad range of conveyor design and application attributes, so its high importance is also consistent with prior expectations. Drop height displayed little predictive value, and we hypothesize that the naive approach of taking the vertical pulley distance while neglecting the design of the transfer chute and loading area is overly crude. More careful characterization of the velocity of material at the loading point relative to the belt would be an interesting area for future work. Finding that the composition of the conveyed material (% fines) is relatively unimportant was unexpected, though the limitations of the permutation importance method means that this should not be equated to concluding that it is irrelevant to the underlying belt wear process.

Partial dependence plots provide a window into the relationship of any “black box” algorithm in a supervised learning problem. Applied to the random forest, the most remarkable finding was that *stacker* conveyors have significantly higher predicted wear rates than other duties. This was surprising when compared against the *shiploader* duty, which is similar to the stacker in many aspects. Closer inspection of these two duties is recommended to understand the reasons for this difference, potentially leading to a better understanding of belt wear. More data for each conveyor duty will allow the individual duty types to be used in the modeling instead of aggregated types in our data. This could also produce greater insights into the effect of conveyor duty and potentially better prediction performance.

The cross-validation model evaluation framework, permutation method for variable importance, and partial effects plots described in this work can be applied to any supervised learning problem. These techniques provide a more complete picture of predictive models than point statistics from a single data split where data are limited, but are under-represented in reliability literature. We have used cross-validation to estimate out-of-sample conveyor prediction performance; a limitation of this procedure is that “out-of-sample” is still limited to the universe of conveyors (real or conceptual) that are similar to the population of conveyors represented by the data. We expect that the results would generalize to other conveyors owned or planned to be built by the mining company that provided the data, assuming they are similar in kind and application. Radical changes to conveyor design or operation, or conveyors in different bulk materials handling operations are expected to behave differently. The modeling process presented is portable and could be applied to other populations of conveyors.

The objective of this work was not to eliminate the need for regular belt thickness testing; this would be a risky proposition even if the models provided better prediction performance. However, benefits of these predictive models go beyond estimating remaining useful life. For example, the model could be used to identify and study belts wearing unusually fast or slow compared to a prediction, to improve understanding of belt wear and replicate best practices, thus lifting plant performance. Additionally, building predictive models that benefit from easily-accessible, high-quality data motivates best practices for data collection and governance. We believe there is a lot of low-hanging fruit to improve prediction performance. Future work could focus on: (a) new variables, for example, velocity of material relative to belt at loading zone, belt rubber grade, conveyor rise/inclination angle, and other variables from Figure 1; (b) higher resolution and more frequent belt thickness data; (c) time series conveyor operational data from SCADA systems; (d) better characterization of the load, for example, tons per meter, tons per belt cycle; and (e) testing other algorithms.



**Funding Statement.** This work would not been possible without funding from the BHP Fellowship for Engineering for Remote Operations—supporting community projects in areas in which BHP operates.

**Competing Interests.** The author declares no competing interests exist.

**Authorship Contributions.** C.W. and J.S.: conceptualization, methodology, data curation, visualization, software, validation, writing—original draft, writing—review and editing, approved the final submitted draft; R.N.K.: conceptualization, project administration, supervision, writing—review and editing, approved the final submitted draft; M.H.: conceptualization, project administration, funding acquisition, supervision, writing—review and editing, approved the final submitted draft.

**Data Availability Statement.** The data used in this work is proprietary and could not be made publicly available. The modeling code is available at <https://github.com/callumwebb/conveyor-wear-modelling>.

## References

- All State Conveyors Pty Ltd** (2018) Definitive Guide for Choosing the Right Conveyor Belt. Available at <https://allstateconveyors.com/solutions/useful-tips-for-choosing-the-right-conveyor-belt/> 2010-09-30.
- Altmann A, Tolosi L, Sander O and Lengauer T** (2010) Permutation importance: A corrected feature importance measure. *Bioinformatics* **26**, 1340–1347. <https://doi.org/10.1093/bioinformatics/btq134>.
- Andrejiova M, Grincova A and Marasova D** (2016) Measurement and simulation of impact wear damage to industrial conveyor belts. *Wear* **368**, 400–407.
- Andrejiova M Grincova A, Marasova D, Fedorko G and Molnar V** (2014) Using logistic regression in tracing the significance of rubber–textile conveyor belt damage. *Wear* **318**(1–2), 145–152.
- Andrejiova M and Marasova D** (2013) Using the classical linear regression model in analysis of the dependences of conveyor belt life. *Acta Montanistica Slovaca* **18**, 77–84.
- Breiman L** (2001) Random forests. *Machine Learning* **45**(1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Cawley GC and Talbot NLC** (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079–2107. Available at <http://dl.acm.org/citation.cfm?id=1756006.1859921>.
- Goldstein A, Kapelner A, Bleich J and Pitkin E** (2013) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation *Journal of Computational and Graphical Statistics* **24** 44–65. <https://10.1080/10618600.2014.907095>.
- Greenwell BM** (2017) pdp: An R package for constructing partial dependence plots. *The R Journal* **9**(1), 421–436. Available at <https://journal.r-project.org/archive/2017/RJ-2017-016/index.html>.
- Hastie T, Tibshirani R and Friedman J** (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edn. Springer. Available at <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.
- Hurvich CM and Tsai C-L** (1990) The impact of model selection on inference in linear regression. *The American Statistician* **44**(3), 214–217. Available at <http://www.jstor.org/stable/2685338>.
- Kohavi R** (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 2. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers, Inc., pp. 1137–1143. Available at <http://dl.acm.org/citation.cfm?id=1643031.1643047>.
- Liaw A and Wiener M** (2002) Classification and regression by random forest. *R News* **2/3**, 18–22. Available at <https://CRAN.R-project.org/doc/Rnews/>.
- Masaki M, Zhang L and Xia X** (2018) A design approach for multiple drive belt conveyors minimizing life cycle costs. *Journal of Cleaner Production* **201** 526–541. <https://10.1016/j.jclepro.2018.08.040>.
- Metso** (2016) *Metso Conveyor Solutions Handbook*. Trelleborg, Sweden: Trelleborg.
- Molnar W, Varga M, Braun P, Adam K and Badisch E** (2014) Correlation of rubber based conveyor belt properties and abrasive wear rates under 2- and 3-body conditions. *Wear* **320**, 1–6. <https://10.1016/j.wear.2014.08.007>.
- R Core Team** (2019) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at <https://www.R-project.org/>.
- Schallamach A** (1954) On the abrasion of rubber. *Proceedings of the Physical Society. Section B* **67**(12), 883–891. Available at <https://doi.org/10.1088%2F0370-1301%2F67%2F12%2F304>.
- Strobl C** (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**(1), 25. Available at <https://doi.org/10.1186/1471-2105-8-25>.
- Strobl C, Hothorn T and Zeileis A** (2009) Party on! A new, conditional variable importance measure for random forests available in the party package. *The R Journal* **1**, 14–17. <https://10.32614/RJ-2009-013>.

**Cite this article:** Webb, C. Sikorska, J. Khan, R. N. and Hodkiewicz, M. 2020. Developing and evaluating predictive conveyor belt wear models. *Data-Centric Engineering*, 1: e3. doi:<https://doi.org/10.1017/dce.2020.1>