Overall Equipment Effectiveness Prediction with Multiple Linear Regression for Semi-automatic Automotive Assembly Lines

Péter Dobrá*, János Jósvai

1 Doctoral School of Multidisciplinary Engineering Sciences, Széchenyi István University, Egyetem tér 1., H-9026 Győr, Hungary
2 Department of Vehicle Manufacturing, Széchenyi István University, Egyetem tér 1., H-9026 Győr, Hungary

* Corresponding author, e-mail: dobra.peter@sze.hu

Received: 30 March 2023, Accepted: 01 September 2023, Published online: 07 September 2023

Abstract
In the field of industry, especially in the production areas, it is particularly important that the monitoring of assembly efficiency takes place in real-time mode, and that the related data-based estimation also works quickly and reliably. The Manufacturing Execution System (MES), Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) systems used by companies provide excellent support in data recording, processes, and storing. For Overall Equipment Effectiveness (OEE) data showing the efficiency of assembly lines, there is a regular need to determine expected values. This paper focuses on OEE values prediction with Multiple Linear Regression (MLR) as supervised machine learning. Many factors affecting OEE (e.g., downtimes, cycle time) are examined and analyzed in order to make a more accurate estimation. Based on real industrial data, we used four different methods to perform prediction with various machine learning algorithms, these were the cumulative, fix rolling horizon, optimal rolling horizon and combined techniques. Each method is evaluated based on similar mathematical formulas.

Keywords
OEE, machine learning, multiple linear regression, assembly line, prediction

1 Introduction
In the domain of industry, especially in the production areas, it is particularly important that the monitoring of assembly efficiency takes place in real-time mode, and that the related data-based estimation also works quickly and reliably. The efficiency of the production lines affects the operation of the entire company, therefore, it is important to predict future values as accurately as possible. Efficiency Key Performance Indicators (KPI’s) have an impact on among others financial results (e.g., profits), production scheduling (e.g., assembly sequence), inventory (e.g., finished products), investments (e.g., transfer new machines), maintenance (e.g., required planned repair time) and continuous improvement (e.g., optimization of assembly processes).

Nowadays, Overall Equipment Effectiveness (OEE) is the most common efficiency metric in industrial practice. This standard indicator clearly shows current status of production and also includes different downtimes and scraps [1–3].

The aim of this paper is to predict the OEE values using machine learning. The paper is organized as follows. Section 2 focuses on the relevant scientific work regarding OEE prediction. Then, Section 3 reveals four different Multiple Linear Regression (MLR) ways such as cumulative, fix rolling horizon, optimal rolling horizon, and combined method. Section 4 concludes the paper.

2 OEE prediction with machine learning
OEE is a part of Total Productive Maintenance (TPM) concept and the basic formula for calculation is written as:

\[ \text{OEE} = a \times p \times q, \]

where \( a \) is the availability (%), \( p \) is the performance (%), \( q \) is the quality (%) [4].

Numerous systems, among others Manufacturing Execution System (MES), Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) provide assistance in the automatic recording, processing and storage of OEE values in the assembly lines [5–7].

Machine learning techniques can be divided into three categories: supervised learning, unsupervised learning and reinforcement learning [8, 9]. All types and its components
can be used for OEE prediction in the field of assembly operations [10]. These estimations are aimed at the following areas:

- machine, station or tool failure;
- downtime occurrence;
- process parameters;
- products failure;
- production planning;
- type change errors, etc. [11–14].

The most commonly used machine learning methods and algorithms in OEE and its components prediction from simple to complex are:

- Logistic Regression [15–17];
- Gaussian Naive Bayes [15, 16];
- K-Nearest Neighbor [15, 18, 19];
- Bayesian Ridge Regression [20, 21];
- Decision Tree Regression Algorithm [19, 22];
- Random Forest [15, 16, 22, 23];
- Support Vector Machine [15, 16, 22, 24, 25];
- Support Vector Regression Generic Algorithm [23, 26, 27];
- Extreme Gradient Boost [15, 23, 26];
- Artificial Neural Network [19, 24, 28, 29];
- Deep Learning [23, 30];
- Combined methods [31].

### 2.1 Multiple Linear Regression

Among the machine learning methods discussed in the scientific literature, with numerous practical examples, MLR is not included, despite its potential for predicting the value of OEE. MLR is a statistical technique that uses several explanatory variables to estimate the outcome of a response variable. Calculation of MLR:

\[ z_i = A_0 + A_1 x_{i1} + A_2 x_{i2} + ... + A_p x_{ip}, \] (2)

where \( z_i \) is the dependent variable, \( A_0 \) is the \( y \) intercept, \( x_i \) is the explanatory variables, \( A_p \) is the slope coefficients for each explanatory variable, and for \( i = n \) observations \( (n \) observation of one dependent variable and \( p \) independent variables).

This paper presents the prediction of OEE values through Multiple Linear Regression.

### 3 OEE prediction with multiple linear regression

When predicting the OEE, the authors followed the classic data processing and evaluation, which includes the following steps:

- Exploring the data;
- Cleaning the data;
- Data visualization;
- Building the model;
- Training the model;
- Predicting with the model;
- Evaluating the performance of the model.

Each step is presented in detail in Sections 3.1 to 3.7.

#### 3.1 Exploring the data

Data from the seat structure hybrid assembly line of a Central European automotive supplier from the years of 2021 and 2022 were used. The original real data were extracted from the MES system and an SQL database. Four different main databases were used, such as OEE data, downtime data, products timestamp data, and quality data. In this article, eight hour (one shift) data is used as a record during machine learning.

#### 3.2 Cleaning the data

The analyzed production data comes from a completely automatic source, so there was no need for major data cleaning. Only those items were excluded where there was no data for the entire shift, for example assembly operations were completed earlier (the workers continued production on another assembly line in that shift). It is important to note that the authors tried to model real production conditions and environment, so extreme values were not excluded.

#### 3.3 Data visualization

When visualizing the data, the most important thing was the representation of the original OEE values, where it can be seen that the examined semi-automatic assembly line is in a slow growing phase. Fig. 1 shows the original OEE data, one data represent an 8-hours shift.

#### 3.4 Building the model

The following key independent variables were considered for multiple linear regression: process failure downtime, break downtime, technical downtime, changeover downtime, quality reason downtime, logistics reason downtime, not planned downtime, other downtime reason, number of changeovers, average cycle time, number of assembled units, and number of scrap units. Generally, OEE forecasting models can be used for either production planning or industrial investment purposes. In this article, for the sake of clarity, the authors present the OEE forecast for
the case of an investment analysis. From a practical point of view, in the case of the industrial investment analysis, the prediction goal is to determine with the help of machine learning how the OEE values develop and, based on these, to decide whether a new assembly line is needed or whether new customer demands can be accepted.

R and R Studio program was used for the entire analysis, in addition to the rolling horizon data, cumulated data series was applied. During the rolling horizon approach, the size of the time windows can change, both for the length of the period and for the starting time, while in the cumulative approach, the cycles are counted from the first shift.

3.5 Training the models
The data shown in Table 1 was used as a basis for the prediction models.

The selected sample period and the predicting period follow industrial practice, data for the next three months are predicted monthly based on the last three months.

3.6 Predicting with the models
During the prediction, the following four models were run and evaluated:

- Cumulative method of multiple linear regression;
- Fix rolling horizon (50 records) method of multiple linear regression;
- Optimal rolling horizon (training records changes) method of multiple linear regression;
- Combined method (cumulative and optimal rolling horizon) of multiple linear regression.

After running the analysis, numerous patterns and correlations were revealed thanks to the many factors taken into account. Fig. 2 shows a scatter plot example where the not planned downtime and OEE value presented as a revealed pattern.

Fig. 3 shows an example for the cumulative method of MLR where training set was the data of records 1–200 and test set was 201–350. The blue line represents the predicted values, and the red line represents the actual OEE values.

3.7 Evaluating the performance of the models
The predicted values were evaluated using three metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE) based on Eqs. (3) to (5):

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}),
\]

\[
MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{y_i} \right),
\]

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2,
\]

where \(n\) is the number of fitted points, \(y_i\) is the actual value, \(\bar{y}\) is the predicted value.
In addition, the main measure of prediction accuracy was the adjusted R-squared. Detailed evaluation results are shown in Tables 2 to 5 for each model. Based on the obtained values and metrics, a decision can be made as to whether a new investment is necessary for the given assembly line or whether an additional order can be accepted. (In reality, no new investments were made, but additional customer orders were accepted.)

Among the individual models, the combined method of MLR performed the best in terms of adjusted R-squared (0.8422), while the fixed rolling horizon method performed best in terms of the MAPE value (4.7723). Out of the many prediction variations, only one is presented in this article, but there is certainly an optimal prediction period and sampling.

### 4 Conclusion

This paper presented OEE prediction techniques using MLR as supervised machine learning at the domain of semi-automatic assembly lines. Prediction with real and validated industrial data analysis and compares four different ways such as cumulative, fix rolling horizon, optimal rolling horizon and combined method. Regarding industrial investments analysis, the fixed rolling horizon

#### Table 2 Cumulative method of MLR

<table>
<thead>
<tr>
<th>Sample period</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
<td>450</td>
<td>500</td>
<td>550</td>
<td>600</td>
<td></td>
</tr>
<tr>
<td>Predicted period</td>
<td>151</td>
<td>201</td>
<td>251</td>
<td>301</td>
<td>351</td>
<td>401</td>
<td>451</td>
<td>501</td>
<td>551</td>
<td>601</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0397</td>
<td>0.0301</td>
<td>0.0338</td>
<td>0.0372</td>
<td>0.0454</td>
<td>0.0437</td>
<td>0.0401</td>
<td>0.0439</td>
<td>0.0432</td>
<td>0.0485</td>
</tr>
<tr>
<td>MSE</td>
<td>0.0031</td>
<td>0.0018</td>
<td>0.0022</td>
<td>0.0022</td>
<td>0.0030</td>
<td>0.0027</td>
<td>0.0023</td>
<td>0.0028</td>
<td>0.0028</td>
<td>0.0036</td>
</tr>
<tr>
<td>Multiple R-squared</td>
<td>0.7101</td>
<td>0.8336</td>
<td>0.8225</td>
<td>0.8098</td>
<td>0.8137</td>
<td>0.7944</td>
<td>0.7935</td>
<td>0.7854</td>
<td>0.7780</td>
<td>0.7745</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.6847</td>
<td>0.8229</td>
<td>0.8135</td>
<td>0.8019</td>
<td>0.8071</td>
<td>0.7880</td>
<td>0.7879</td>
<td>0.7801</td>
<td>0.7731</td>
<td>0.7699</td>
</tr>
<tr>
<td>p-value</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
</tr>
</tbody>
</table>

#### Table 3 Fix rolling horizon method of Multiple Linear Regression

<table>
<thead>
<tr>
<th>Sample period</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
<td>450</td>
<td>500</td>
<td>550</td>
<td>600</td>
<td></td>
</tr>
<tr>
<td>Predicted period</td>
<td>151</td>
<td>201</td>
<td>251</td>
<td>301</td>
<td>351</td>
<td>401</td>
<td>451</td>
<td>501</td>
<td>551</td>
<td>601</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0397</td>
<td>0.0325</td>
<td>0.0386</td>
<td>0.0376</td>
<td>0.0467</td>
<td>0.0393</td>
<td>0.0334</td>
<td>0.0288</td>
<td>0.0326</td>
<td>0.0394</td>
</tr>
<tr>
<td>MSE</td>
<td>0.0031</td>
<td>0.0034</td>
<td>0.0037</td>
<td>0.0023</td>
<td>0.0032</td>
<td>0.0022</td>
<td>0.0019</td>
<td>0.0013</td>
<td>0.0019</td>
<td>0.0030</td>
</tr>
<tr>
<td>Multiple R-squared</td>
<td>0.7101</td>
<td>0.8325</td>
<td>0.8532</td>
<td>0.8784</td>
<td>0.8078</td>
<td>0.7133</td>
<td>0.6880</td>
<td>0.6257</td>
<td>0.5789</td>
<td>0.6858</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.6847</td>
<td>0.8229</td>
<td>0.8135</td>
<td>0.8019</td>
<td>0.8071</td>
<td>0.7880</td>
<td>0.7879</td>
<td>0.7801</td>
<td>0.7731</td>
<td>0.7699</td>
</tr>
<tr>
<td>p-value</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
</tr>
</tbody>
</table>

#### Table 4 Optimal rolling horizon method of Multiple Linear Regression

<table>
<thead>
<tr>
<th>Sample period</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
<td>450</td>
<td>500</td>
<td>550</td>
<td>600</td>
<td></td>
</tr>
<tr>
<td>Predicted period</td>
<td>151</td>
<td>201</td>
<td>251</td>
<td>301</td>
<td>351</td>
<td>401</td>
<td>451</td>
<td>501</td>
<td>551</td>
<td>601</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0397</td>
<td>0.0325</td>
<td>0.0386</td>
<td>0.0376</td>
<td>0.0467</td>
<td>0.0393</td>
<td>0.0334</td>
<td>0.0288</td>
<td>0.0326</td>
<td>0.0394</td>
</tr>
<tr>
<td>MSE</td>
<td>0.0031</td>
<td>0.0034</td>
<td>0.0037</td>
<td>0.0023</td>
<td>0.0032</td>
<td>0.0022</td>
<td>0.0019</td>
<td>0.0013</td>
<td>0.0019</td>
<td>0.0030</td>
</tr>
<tr>
<td>Multiple R-squared</td>
<td>0.7101</td>
<td>0.8325</td>
<td>0.8532</td>
<td>0.8784</td>
<td>0.8078</td>
<td>0.7133</td>
<td>0.6880</td>
<td>0.6257</td>
<td>0.5789</td>
<td>0.6858</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.6847</td>
<td>0.8229</td>
<td>0.8135</td>
<td>0.8019</td>
<td>0.8071</td>
<td>0.7880</td>
<td>0.7879</td>
<td>0.7801</td>
<td>0.7731</td>
<td>0.7699</td>
</tr>
<tr>
<td>p-value</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
</tr>
</tbody>
</table>
approach performed best among the four methods according to the MAPE, while the combined method provide the best according to the adjusted R-squared.

Table 5 Combined method of Multiple Linear Regression

<table>
<thead>
<tr>
<th>Sample period</th>
<th>1</th>
<th>1</th>
<th>101</th>
<th>151</th>
<th>151</th>
<th>151</th>
<th>151</th>
<th>151</th>
<th>151</th>
<th>151</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>151</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>Predicted period</td>
<td>151</td>
<td>201</td>
<td>251</td>
<td>301</td>
<td>351</td>
<td>401</td>
<td>451</td>
<td>501</td>
<td>551</td>
<td>601</td>
</tr>
<tr>
<td>MAE</td>
<td>0.0397</td>
<td>0.0301</td>
<td>0.0386</td>
<td>0.0376</td>
<td>0.0420</td>
<td>0.0446</td>
<td>0.0462</td>
<td>0.0546</td>
<td>0.0547</td>
<td>0.0606</td>
</tr>
<tr>
<td>MSE</td>
<td>0.0031</td>
<td>0.0018</td>
<td>0.0037</td>
<td>0.0023</td>
<td>0.0027</td>
<td>0.0029</td>
<td>0.0030</td>
<td>0.0041</td>
<td>0.0042</td>
<td>0.0051</td>
</tr>
<tr>
<td>Multiple R-squared</td>
<td>0.7101</td>
<td>0.8336</td>
<td>0.8532</td>
<td>0.8784</td>
<td>0.8784</td>
<td>0.8784</td>
<td>0.8784</td>
<td>0.8784</td>
<td>0.8784</td>
<td>0.8546</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.6847</td>
<td>0.8229</td>
<td>0.8404</td>
<td>0.8677</td>
<td>0.8677</td>
<td>0.8677</td>
<td>0.8677</td>
<td>0.8677</td>
<td>0.8677</td>
<td>0.8422</td>
</tr>
<tr>
<td>p-value</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
<td>2.2e-16</td>
</tr>
</tbody>
</table>

The authors plan to further compare the mentioned four methods in the field of production planning. Future research could be the application of other machine learning methods in the prediction of OEE.

References


