www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

PROBABILISTIC FORECASTING OF CONSTRUCTION LABOR PRODUCTIVITY METRICS

SUBMITTED: August 2023
REVISED: January 2024
PUBLISHED: February 2024
EDITOR: Robert Amor
DOI: 10.36680/j.itcon.2024.004

Emil L. Jacobsen, Ph.D. Candidate
Department of Civil and Architectural Engineering, Aarhus University, Denmark
elj@cae.au.dk

Jochen Teizer, Professor
Department of Civil and Mechanical Engineering, Technical University of Denmark, Denmark
teizerj@dtu.dk

Søren Wandahl, Professor
Department of Civil and Architectural Engineering, Aarhus University, Denmark
swa@cae.au.dk

Ioannis Brilakis, Professor
Department of Engineering, University of Cambridge, United Kingdom
ib340@cam.ac.uk

SUMMARY: *This study investigates the possibility of doing probabilistic forecasting of construction labor productivity metrics for both long-term and short-term estimates. The research aims to evaluate autoregressive forecasting models, which may help decision-makers with information currently unavailable in construction projects. Unlike point forecasts, the proposed method employs probabilistic forecasting, offering additional valuable insights for decision-makers. The distributional information is obtained by updating the moments of the distribution during training. Two datasets are used to evaluate the models: one collected from an entire construction site for long-term forecasting and one from an individual worker for short-term forecasting. The models aim to predict the state of direct work, indirect work, and waste. Several models are trained using different hyperparameters. The models are tuned on the number of trees and the regularization used. The presented method gives estimates of future levels of direct work, indirect work, and waste, which will add value to future processes.*

KEYWORDS: *Probabilistic forecasting, Construction labor productivity, Progress monitoring, Work sampling*

REFERENCE: *Emil L. Jacobsen, Jochen Teizer, Søren Wandahl, Ioannis Brilakis (2024). Probabilistic forecasting of construction labor productivity metrics. Journal of Information Technology in Construction (ITcon), Vol. 29, pg. 58-83, DOI: 10.36680/j.itcon.2024.004*

1. INTRODUCTION

This paper is about probabilistic forecasting of metrics for Construction Labor Productivity (CLP). Probabilistic forecasting concerns the estimate of a future distribution of the given parameter rather than just an estimate of a point. The metrics focused on in this paper are from the work sampling classification scheme: direct work, indirect work, and waste. The forecasting is done for construction labor, which in this project are construction workers from several different trades. The forecasted metrics correlate to productivity, with direct work being an indicator of the level of productivity for the specific worker or trade. This paper addresses the problem that construction projects are inherently difficult to monitor due to their transitory, non-repetitive, and complex nature. This problem is significant because monitoring construction projects is essential to ensure the projects stay within time and budget (Kopsida et al., 2015). Historically, the construction industry has seen slow adoption of disruptive technologies, which, alongside other factors, has led to construction falling behind other industries in terms of productivity increase. Monitoring construction projects is almost exclusively done by manually collecting information that makes it possible to infer past performance (Hui et al., 2015). The construction managers who monitor the construction projects would benefit from predictions or estimates about the future. By determining a way to compare as-planned to estimates of the future as-performed states of a construction project, mitigation strategies can be made based on actual data. Optimization of the as-planned can then be made through simulations. Changes could be regarding site layout, activity sequences, and the definition of work zones. In these situations, future performance would serve as the reward parameter in an objective function. A deeper understanding of construction labor productivity is needed for these analyses and simulations. Current methods make it possible to understand past performances but not future states of the production system. Having information about the future states of the production system can enable future research with added features and help the decision-makers on the construction projects with information that could form the basis for altering their production system based on forecasted values.

CLP is an inherently important metric for construction projects, as labor costs are a significant part of the total budgets of projects (Buchan et al., 2003; Kazaz et al., 2008). Optimizing the projects based on construction workers is difficult, as construction workers are independent resources of the project and, therefore, challenging to alter in terms of working style. A socio-technical system introduces several complications, as the interactions between workers, machines, the environment, and external forces can happen in several different ways (Wandahl et al., 2022). This makes a prediction system more complex, as several external factors can impact how construction workers utilize the time spent on construction projects (Kazerooni et al., 2021). In practice, monitoring is currently done by looking at the past. The closest current practice gets to forecasting is through linear regression when estimating durations of tasks (For instance, time spent on 1 m^2 of bricks times total m^2 of bricks). Several critical assumptions are made in this estimate, including assuming a constant or mean productivity. Not accounting for the fluctuations in productivity makes right-time interference impossible. Predictions that account for fluctuations make information available that will be valuable in decision-making regarding items such as the sequencing of tasks, procurement, and site layout or in further analyses, for instance, through production system simulations.

Current methods for predicting productivity in the construction domain use independent variables to create models that can estimate the general level of productivity from those variables. The estimation or prediction of productivity is not temporal, which means that in most cases, only one value is given as the prediction. Most research estimates a single value for the entire project, which makes the prediction challenging to use for monitoring purposes. Because of this, the question that will be the focus of this paper can be asked: How can the construction labor productivity metrics be used not only to evaluate past performance but also to get future projections of the production system? The paper aims to develop forecasting models that can estimate future values of construction labor productivity metrics in an autoregressive fashion. Autoregression is used, as this will minimize the efforts needed for data collection. CLP is a complex metric affected by numerous variables, such as weather, experience of workers, details of the specific project (size, number of floors, complexity, etc.), use of technology, and the managerial environment. However, autoregression has been proven successful in similar complex systems such as exchange rates (So et al., 1999), volatility of oil futures (He et al., 2021), and residential energy consumption (Fan et al., 2023). These have many independent variables that will affect the output, but autoregression has been proven successful in these domains. As this is the case, it is deemed apparent to investigate the potential applicability of such methods for forecasting CLP. The forecasting process is explored through the method of boosted trees. Here, several models are trained on two datasets to mimic the extreme use cases: individual productivity metric forecasting and project productivity metric forecasting. To ensure the methods can be deployed on construction projects, a short analysis of the importance of frequency is done to find an optimum between performance and computational efficiency. The output will help decision-makers such as construction managers to understand the

next individual cycles (from the individual forecast) and the future of the overall construction site (from the project forecast). As mentioned, the method does not use external factors such as weather, workforce information (age, experience, etc.), or project details (project size, complexity, number of levels, etc.). The reason for this is that the motivation of the research is to create a simplistic system that can be used in a general setting without the need for extensive data collection.

The objectives of the paper are (1) to give an overview of current productivity prediction research, (2) to develop autoregressive models for productivity metric forecasting, (3) to evaluate probabilistic forecasting methods on both the individual and project level of construction projects, and lastly (4) to evaluate the importance of sampling frequency when working with high-frequency datasets for CLP forecasting.

2. RELATED WORKS

Productivity can be defined as the ratio between the output volume and the volume of the inputs (OECD, 2023). Because of this definition, data can be collected in many forms, and no standard format or data collection method is available, even though it has long been claimed as an essential step (Thomas and Yiakoumis, 1987; Yi and Chan, 2014). Best practices such as work sampling exist, but implementations still differ. CLP monitoring is an essential process in construction projects, as the CLP is a good indication of the overall project status (Wandahl et al., 2022). Monitoring methods, which will be presented in this section, generally fall into two categories: manual monitoring through direct on-site observations and prediction through machine learning using independent variables. The first will only allow for information regarding past CLP, as the method requires on-site data collection and analysis. The second can, as mentioned earlier, predict an average CLP for the entire project. It could be argued that doing so can retrieve future CLP for the project. However, this would not give an understanding of the fluctuation of productivity, but rather an average or momentarily understanding. Alongside monitoring methods, the use of autoregression in construction will be reviewed. This will establish why and how this form of forecasting is applied to construction problems.

2.1 On-site monitoring

The monitoring of CLP has been extensively studied throughout the literature. Work sampling, the most widely used work study-based method to assess labor time utilization (Yi and Chan, 2014), has been used as the data collection method for numerous studies in construction (Allmon et al., 2000; Dai et al., 2007; Dai et al., 2009; Kalsaas et al., 2014; Gong et al., 2011; Wandahl et al., 2021). The distribution between or occurrence of the classes collected through work sampling for construction workers is an essential piece of information to measure and reduce labor waste and, therefore, optimize the time that is spent on construction sites (Liou and Borcherding, 1968; Gouett et al., 2011; Neve et al., 2020). Optimizing the time spent on construction is an obvious objective, but the methods vary greatly. An essential theory regarding this optimization is the TFV theory (Koskela, 2000) and the concept of flow (Koskela, 1992). A system such as the Last Planner System, which is based on these concepts, will help increase efficiency. However, such well-established systems primarily focus on planning rather than on-site monitoring.

The studies that manually collect project data through, for instance, work sampling, have irregular and inconsistent datasets often limited to a small timeframe relative to the duration of a construction project. This is not an issue for most projects that analyze the data, but it creates difficult datasets to use in more complex algorithms. As work sampling is a time-consuming and labor-intensive task, two weeks of data is already an extensive task. However, consistent work sampling studies, which follow the entirety of a project, are still needed as this would allow for more elaborate studies with continuous data from all phases. Furthermore, it is important to note that work sampling is still considered among the best methods for understanding the current status of CLP. As the data from the work sampling studies are starting to be further used for machine learning models, statistical time-series studies, or other correlation studies, it is also important to consider the irregular temporal nature of many work sampling studies, where the time between data points can vary greatly, when collected through random walks. Several research projects have attempted to automate the monitoring of construction projects and CLP specifically (Barbosa and Costa, 2021). Some of these studies use the classes of work sampling, meaning that when generalized, these frameworks could be used to automatically collect the data usually collected through work sampling studies (Jacobsen et al., 2023). The automation is done using several different methods, for instance computer vision (Gong and Caldas, 2011; Liu and Golparvar-Fard, 2015; Luo et al., 2018), sensors (Joshua and Varghese, 2011; Cheng et al., 2011; Joshua and Varghese, 2014; Ryu et al., 2019; Jacobsen et al., 2023), and audio (Rashid and Louis, 2020; Cheng et al., 2017).

2.2 Autoregressive models in construction

Autoregressive models- models where the output variable depends on its own previous values- are represented in many domains, including construction research. In the construction domain, autoregression has primarily been used for cost forecasting (Xu and Moon, 2013; Hwang et al., 2012; Ashuri and Lu, 2010; Joukar and Nahmens, 2016; Ilbeigi et al., 2017; Cao and Ashuri, 2020; Omar, 2020), but has also been applied in the productivity domain to estimate changes in productivity on an industry-level (Assaad and El-Adaway, 2021). Similarly, Wong et al. (2005) use an autoregressive approach to model labor productivity on a macro-level for the Hong Kong construction industry over a period of 20 years. They use an autoregressive integrated moving average (ARIMA) model to forecast labor productivity with a test set spanning one year with quarterly data points. A test-set performance of 6.6 mean absolute percentage error (MAPE) is obtained. The performance of autoregressive models for macro-level productivity forecasting is deemed satisfactory (Assad and El-Adaway, 2021; Wong et al., 2005). However, this only shows that autoregression is a suitable method for forecasting productivity metrics for the industry on a macro-level. This application is suitable for strategic decisions on a company level, but for individual projects, it is difficult to use this information. On a project level, the research on cost forecasting can be useful, especially for procurement and the timing thereof. For instance, Ilbeigi et al. (2017) forecast asphalt-cement prices using ARIMA, among other methods, which can potentially help optimize budgets in transportation projects. Even with both project-level applications and industry-wide applications, no research on CLP forecasting using autoregressive methods is present. By using autoregression for project-level productivity metrics, similarly to how the industry-level productivity metrics were forecasted, it would be possible to estimate future levels of productivity without the need for extensive data collection. However, a closer examination of the possibilities of autoregression for project-level and individual-level forecasting needs to be conducted. Other areas, for instance, tunnel boring, have seen significant research in forecasting the advance rate of the tunnel boring machine (Shangxin et al., 2021; Gao et al., 2019; Gao et al., 2021). However, methods concerning the forecasting of advance rate have also been critiqued for the short horizon and time-delayed predictions of the forecast and, therefore, minor benefit if applied in the industry, as the predictions do not contain any valuable information (Erharter and Marcher, 2021). This important conclusion should also be considered in the CLP forecasting domain.

2.3 Construction labor productivity prediction

In this paper, productivity prediction is defined as the process in which productivity is estimated for a point in time where, at least for the model, the productivity is unknown. When examining current research regarding predicting CLP metrics, most publications focus on automating the process through prediction to find average productivity. This comes in different granularities, from daily to project averages. These predictions are often based on various project or crew parameters. The information used in these prediction algorithms is often static (at least for the project at hand), such as the crews' experience, the project's size, and contractual agreements. As the information is static, temporal granularity in the prediction is impossible to obtain, as the independent variables used for the prediction will never or rarely change during the construction phase. One way to obtain such granularity would be through autoregressive models, which ideally have many timesteps each day, giving a granularity of the forecast that would be actionable for the construction managers.

The field of productivity prediction is dominated by machine learning methods, which have been successfully used not only for predicting productivity but in the construction domain as a whole (Jacobsen et al., 2022). The machine learning methods often utilize independent variables to predict CLP. They are motivated by the difficulty of obtaining information regarding the productivity of a construction project (Jacobsen et al., 2023). However, many models have input data requiring manual data collection with information scattered across the project. The models use information from interviews with workers, project documents, or data that needs to be collected through walks on the construction site. Several publications focus on predicting productivity through construction project variables (Adebowale and Agumba, 2022). The data used is a mix of intangible (motivation, well-being) and tangible parameters (project size, floors, experience), where the intangible parameters especially require extensive work to collect. An overview of research within the field is presented in Table 1.

The models predicting productivity utilize various factors to do so. The prediction method ranges from simple regression models (Sanders and Thomas, 1993; Smith, 1999) to Artificial Neural Network (ANN) models (Golnaraghi et al., 2019; Golnaraghi et al., 2020; Nasirzadeh et al., 2020; Ebrahimi et al., 2022; Mirahadi and Zayed, 2016; Tsehayae and Fayek, 2016; Heravi and Eslamdoost, 2015; Muqeem et al., 2011; Florez-Perez et al., 2022; Dissanayake et al., 2005; Goodarzizad et al., 2023), Support Vector Machines (SVMs) (Cheng et al., 2021; Momade et al., 2020; Florez-Perez et al., 2022), random forest (RF) (Ebrahimi et al., 2021), Self-organizing maps

(SOM) (Oral and Oral, 2010), and K-nearest neighbors (KNN) (Florez-Perez et al., 2022; Ebrahimi et al., 2022). Cheng et al. (2021) introduce a combination of dynamic feature selection, least square support vector machine, and symbiotic organisms search to predict construction productivity for formwork activities. The model uses 12 input variables to predict productivity for formwork activities. The prediction model was trained and tested on a dataset of 220 observations collected through various methods, such as work sampling, reports filled out by foremen, and online accessible data records. Predicting the overall productivity from a sample of 220 observations can be dangerous, as this could be unrepresentative of the overall construction project. Further studies into the sample size are essential to ensure the sample is statistically representative of the overall project. The number of features used in the models ranges from 4 (Oral and Oral, 2010) to 43 (Tsehayae and Fayek, 2016), with both publications predicting productivity for several activities. The input features used throughout the studies in Table 1 have been extensively studied (Caldas et al., 2015), with Dai et al. (2009) introducing 83 factors that influence productivity. Picking the correct features can be challenging because the importance can vary based on the country or region of the projects examined and because the importance of features can vary based on the trade. As this is the case, a method such as autoregression can alleviate these challenges.

The literature mainly focuses on concreting activities, with a few publications focusing on other trades, such as masonry work (Sanders and Thomas, 1993; Florez-Perez et al., 2022). This indicates that the methods are not being generalized yet, which is a limitation if needed for an entire construction project. A wide range of evaluation metrics are used to evaluate the methods in the studies. Most of them are well-established metrics in machine learning, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Florez-Perez et al. (2022) approach the problem of productivity prediction differently, using a classification algorithm with three classes (low, medium, and high productivity). This approach gives a quick overview of the predictions, but the approach suffers when outputs are close to the classification threshold. Two points classified into medium productivity could be 21% and 59% productive, which is a big span and often very relevant for construction projects. The productivity prediction problem is likely not suited for a classification algorithm, as the intraclass variability is important. Nasirzadeh et al. (2020) use prediction intervals (giving a lower- and upper bound of the prediction), introducing probabilistic prediction rather than a single value of labor productivity, evaluating the models using the prediction interval normalized average failure distance (PINAFD). This method gives an interval at a given confidence level, which makes the output more reliable and, thereby, easier for the project managers to use for improvement strategies.

The datasets collected to train and test the models are primarily between 84 and 570 observations, with two publications using significantly more data: 1977 observations (Florez-Perez et al., 2022) and 4915 observations (Bai et al., 2019). Most datasets are not described in detail and are not readily accessible for reproduction of the models. Most datasets are also with low frequency, for instance, the dataset first used by Mirahadi and Zayed (2016), which was later used by Golnaraghi et al. (2019; 2020), Nasirzadeh et al. (2020), and Cheng et al. (2021). The dataset was collected on Montreal construction projects over 30 months, and a total of 221 data points were collected. This means that if the dataset was collected with regular intervals, the monthly frequency would be 7.37 observations. This is deemed sufficient for a single prediction, as similar dataset sizes are found in the other studies, as shown in Table 1. However, the low frequency and relatively small dataset are insufficient for continuous productivity forecasting with a sub-daily granularity, as predicting with a higher frequency than the collected dataset is unfeasible. Notably, most publications in Table 1 collect substantially more data than displayed in the column Dataset size. However, this data is pre-processed, ending with the number of observations in Table 1. This number of observations is the number of data points or lines of data on which the models are trained and tested. The dataset used by Mirahadi and Zayed (2016) amongst others, use work sampling data, but the work sampling data is processed into three features (a percentage distribution of direct work, support work, and delay) from an entire day of work sampling. A similar process was used by Tsehayae and Fayek (2016), who collected 15,306 points of work sampling data, which was not used in its raw format. The dataset collected by Tsehayae and Fayek (2016) is further used by Ebrahimi et al. (2021; 2022). Even though two datasets are shared among several publications, as seen in Table 1, the number of observations used for the studies differ. This indicates that a general and accepted method has not yet been established, neither for the pre-processing nor for the actual prediction, as numerous methods are applied for both. All models except for Bai et al. (2019) use manual data collection efforts (for instance, interviews, questionnaires, or work sampling), invalidating the efforts toward a fully autonomous system. Manual data collection is used because most parameters the authors use for the prediction are difficult to obtain automatically without developing extensive collection systems that require on-site people to report the parameters (Oral and Oral, 2007).

Table 1: Overview of productivity prediction methods.

Reference	Activity	Max input	Model type	Model performance	Dataset size	Data collection methods	Data access
(Sanders and Thomas, 1993)	Masonry work	6	Additive regression	0.411 R2	570 observations	On-site data collection	Not disclosed
(Smith, 1999)	Earthmoving	11	Stepwise regression	0.906 R2	141 observations	On-site data collection	Not disclosed
(Dissanayake et al., 2005)	Pipe fabrication	7	ANN	0.006 MSE, 0.94 R2	164 observations	On-site data collection	Not disclosed
(Muqeem et al., 2011)	Formwork	5	ANN	0.000182 MSE	84 observations	Questionnaire	Not disclosed
(Heravi and Eslamdoost, 2015)	Concreting	15	ANN	4.8% error	93 observations	Questionnaire and interviews	Not disclosed
(Mirahadi and Zayed, 2016)	Concrete pouring	9	ANN	0.00932 MSE	131 observations	Work sampling, foremen's reports, online data records	Available on request
(Golnaraghi et al., 2019)	Formwork	9	ANN	0.0215 MSE, 0.949 R2	221 observations	Same dataset as above	Same dataset as above
(Golnaraghi et al., 2020)	Formwork	9	ANN	0.0419 MSE, 0.9902 R2	221 observations	Same dataset as above	Same dataset as above
(Nasirzadeh et al., 2020)	Concrete pouring	9	ANN	11-21.4% PINAFD	221 observations	Same dataset as above	Same dataset as above
(Cheng et al., 2021)	Formwork	12	SVM	3.67% MAPE, 0.0563 MAE, 0.0721 RMSE	220 observations	Same dataset as above	Same dataset as above
(Tsehayae and Fayek, 2016)	Concreting, electrical, shutdown	43	ANN, fuzzy rule-based models	0.3042 accuracy*	399 observations	On-site data collection	Not disclosed
(Ebrahimi et al., 2021)	Concreting	14	RF	0.112 MAE. 0.137 RMSE	85 observations	Same dataset as above	Same dataset as above
(Ebrahimi et al., 2022)	Concreting	19	ANN, KNN, RF, ANFIS	0.668 RMSE, 0.516 MAE	82 observations	Same dataset as above	Same dataset as above
(Oral and Oral, 2010)	Concrete pouring	4**	SOM	25.68% MAPE, 0.12 MAE, 0.03 MSE	144 observations	Time study sheets	Not disclosed
(Oral and Oral, 2010)	Formwork	4**	SOM	38.04% MAPE, 0.011 MAE, 0.00023 MSE	101 observations	Time study sheets	Not disclosed
(Oral and Oral, 2010)	Reinforcement	4**	SOM	25.05% MAPE, 0.19 MAE, 0.06 MSE	101 observations	Time study sheets	Not disclosed
(El-Gohary et al., 2017)	Carpentry and reinforcement	29	ANN	0.0032 MSE	640 observations	Questionnaires, online data records	Available on request
(Bai et al., 2019)	Cutter suction dredgers	9	XGBBoost	9% MAPE, 218 MAE, 0.75 R2	4915 observations	Real-time monitoring	Available on request
(Momade et al., 2020)	Multiple trades	19	SVM and RF	83.5% accuracy	220 observations	Questionnaire and interviews	Not disclosed
(Florez-Perez et al., 2022)	Masonry	14	ANN, KNN, SVM	97.7% accuracy	1977 observations	On-site data collection	Not disclosed
(Goodarzizad et al., 2023)	Concrete pouring	6	ANN	0.9015 R2	107 observations	Questionnaires	Not disclosed

* Accuracy is in this paper a combination of 3 measures. ** The four features are engineered from several raw features.

To summarize CLP prediction, most publications use time-consuming methods to collect the data used to train and test the prediction models. The literature examined shares most of the parameters used for the prediction, which shows a consensus on which parameters are important for productivity prediction across multiple activities. The publications use several different evaluation metrics, and the results are shown in Table 1. As the researchers do not share datasets openly (some datasets are available on request, and some do not disclose any information

regarding data availability), comparing performance is impossible, as some datasets may be easier to predict than others. Many of the publications in Table 1 use time study methods such as work sampling as a part of their feature set. The work sampling data could be used in an autoregressive model by using past work sampling distributions to predict the future work sampling distribution. This would potentially allow for a more granular view of productivity metrics by enabling frequent outputs rather than one output for an entire project or phase. To enable these insights, probabilistic autoregressive models could be used.

The authors have not found any examples of autoregressive forecasting techniques in CLP research. Significant work has been done regarding CLP prediction and autoregression in construction, which is presented throughout this section. Combining the two could alleviate some of the continuous efforts needed for data collection, which is still required in most CLP predictions. Suppose an automated sampling technique is applied, such as the one presented in the author's previous work (Jacobsen et al., 2023). In that case, a fully autonomous system giving future levels of productivity is possible. Therefore, an autoregressive system for both short- and long-term forecasting is needed in the body of knowledge. This objective entails several technical research questions: (1) Is it possible to forecast productivity metrics with a higher temporal granularity? (2) Is it possible to forecast productivity metrics using auto-regression rather than independent variables? And (3) How can the uncertainty be incorporated into the productivity forecast?

3. METHODOLOGIES

Probabilistic forecasting has been used for several decades, especially in models with binary probabilities; for instance, the chance of it raining tomorrow could be 30% (Gigerenzer et al., 2005). The transition from point forecasts to probabilistic forecasts can significantly impact construction research as the probability of each output will be known, which can create a better understanding of the future and, therefore, enable better responses.

The process from data collection to forecasting spans widely. In this section, the subprocesses and their methods will be presented. Figure 1 gives an overview of the process and the relation between subprocesses.

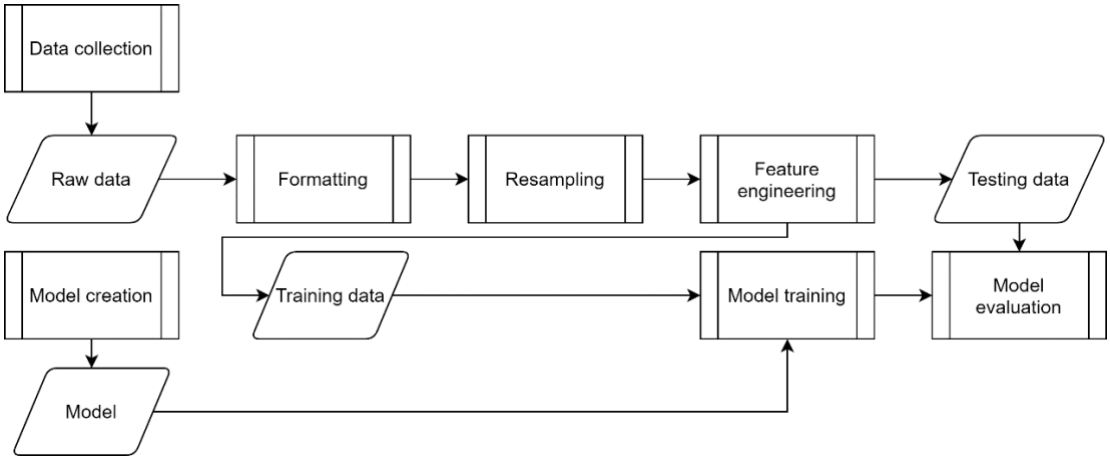

Figure 1: Process diagram of the framework.

As mentioned in the introduction, both a short-term (individual productivity) and long-term (project productivity) forecasting is done. To do so, two datasets are collected. One high-frequency kinematic dataset from a painting job used for the individual productivity forecasting, and one dataset collected through work sampling for the project productivity forecasting.

3.1 Data collection

As forecasting is to be done on both project and individual worker levels, two datasets are needed. A dataset of 90 minutes of automated time and motion studies of painters is used for the forecasting models for short-term. The dataset has been used and presented in Jacobsen et al. (2023), where the entire collection process is explained. The dataset presents a micro-level view of productivity metrics, differing from most earlier publications, which tackles a macro-level problem by examining the overall productivity of an entire construction project or task. The dataset is collected through a laboratory experiment consisting of four individual data collections. The data used in this

research is the labeling that was done through camera footage in Jacobsen et al. (2023), which consists of the three classes: direct work, indirect work, and waste, which are categories taken from classic construction work sampling research (CII 2010; Wandahl et al., 2021).

The second dataset, which is used for long-term forecasting, is a work sampling dataset collected over 29 days of work sampling on four Danish construction projects. The four work sampling studies are all general work sampling studies, including all on-site workers. This means the dataset consists of data from carpentry, masonry, electricity, scaffolding, painting, ventilation, and demolition. The four construction projects are all renovation, three being renovation of apartment-complexes and the fourth being lab and teaching facilities. The work sampling is done by doing random walks through the site, and whenever a worker is seen, their action is noted as either direct work, indirect work, or waste. The long-term dataset is irregular and far less frequent than the short-term dataset, with only 208 data points per day on average (Compared to the 60 Hz of the short-term dataset). Furthermore, the long-term dataset fluctuates more, as seen when comparing Figure 3 and Figure 4. This is due to the nature of the data collection, where the short-term dataset is one worker at a time, and the long-term dataset is multiple workers simultaneously. Because of this, the long-term dataset can easily find a worker doing direct work and one doing indirect work immediately after each other. This does not happen in the short-term dataset where the observations are continuous for a single worker and follow their cycles. This means that the observations will often have many data points after each other for the same class and that the cycle of the classes persists.

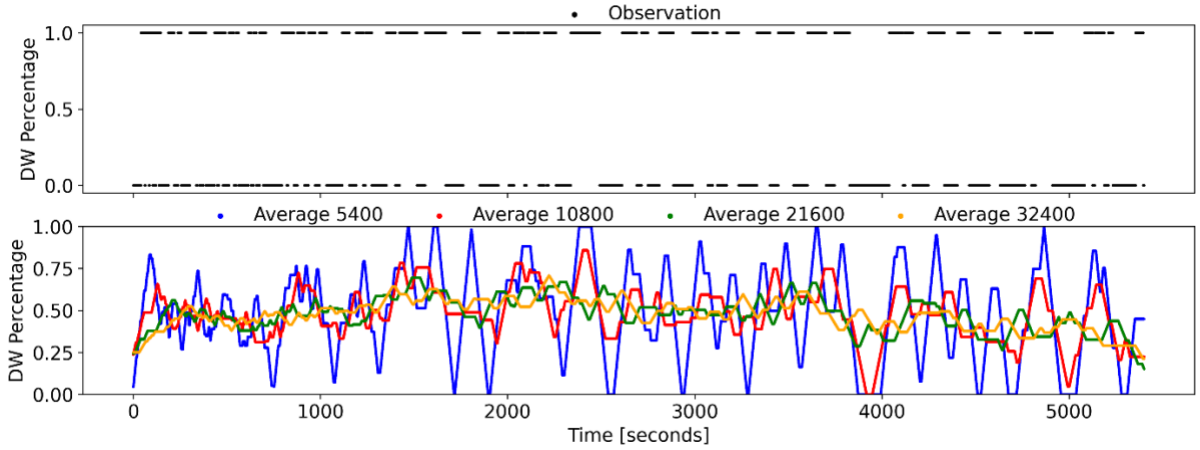

Figure 2: Visualization for the class of direct work. The encoded kinematic observations (top) and the four moving averages of the short-term dataset (bottom).

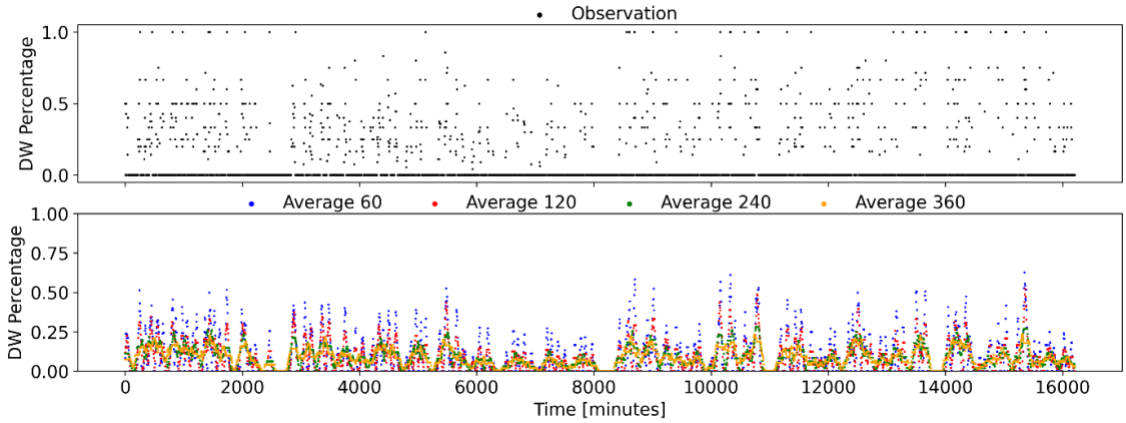

Figure 3: Visualization for the class of direct work. The encoded work sampling observations (top) and the four moving averages of the long-term dataset (bottom).

As work sampling is a probabilistic method, it is important to understand the underlying statistical features. For the dataset to have a high accuracy (resembling the population correctly), a higher number of observations is

needed. To calculate the number of observations needed, a preliminary estimate of the distribution between the classes is needed. This will show how often it is expected to observe each of the three categories. Based on past literature, the expected distribution is set to have 40% direct work, 32% indirect work, and 28% waste. Given equation 1 for standard error, the number of observations needed to have a relative accuracy of ±5% with a confidence level of 95% for each class is presented in Table 2.

$$\sigma_p = z\sqrt{\frac{p(1-p)}{n}} \tag{1}$$

Where σ_p is the standard error, p is the percentage occurrence of the activity (either direct work, indirect work, or waste), z is the Z score (1.96 for 95% confidence level), and n is the number of observations.

Table 2: Overview of the work sampling dataset and its classes.

Class	Total number of observations required
Direct work	2,400
Indirect work	3,400
Waste	4,115

As can be seen from Table 2, the class with the maximum requirement is Waste, which means that the minimum number of observations in the dataset will be 4,115 to meet the requirement of relative accuracy and confidence level. As the total number of observations in the dataset is 6,059, it can be concluded that the work sampling study has enough observations to give the required statistical accuracy.

3.2 Data processing

As the two datasets are collected differently, they have separate processes to prepare them for the training and testing of the model. For the long-term dataset, a resampling is needed to make the dataset regular. As work sampling is a stochastic method that heavily relies on the number of data points collected, the resampling does not reduce the number of data points but pushes them to the nearest timestamp that will be part of the regulated dataset, as shown in Figure 4. Before resampling, the dataset is one hot encoded. The data is initially in an observational format, which means a category from work sampling is noted for every timestamp (given by the observation). The encoding turns these observations into numerical features corresponding to one of the three classes (direct work, indirect work, or waste). An example could be the observation [1,0,0], which corresponds to an observation of direct work.

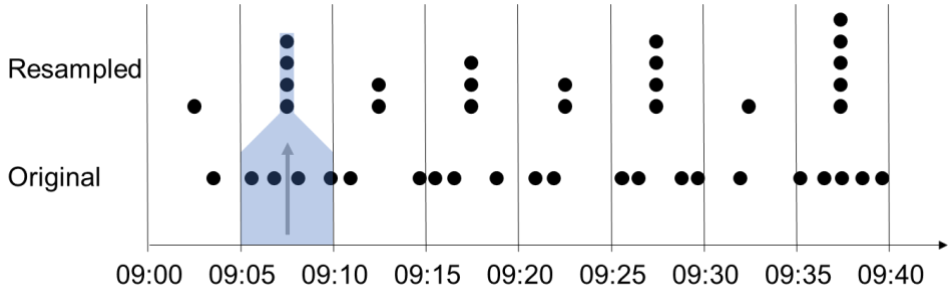

Figure 4: Concept of the resampling process for work sampling data illustrated.

By resampling, every observation is set within bins of 5 minutes instead of at one point in time. As the work sampling dataset is used to examine the entire construction project, a granularity of 5 minutes is deemed sufficient. As some bins will have more observations than others, as seen in the example in Figure 4, a weight parameter is added to each period, which is used to signal how many points are present in the period. This weight parameter is used when computing moving averages of the time series, as a period with more points should have more influence

on the moving average than periods with fewer points. This also means that the forecasting will give predictions of future states in intervals of 5 minutes.

For the short-term dataset, resampling is unnecessary, as the dataset is collected through sensors with 60 Hz, and it is, therefore, already a regular time series. Comparing the size of the datasets to the previously mentioned research showcased in Table 1, the number of data points is higher than average. However, as this data is to be used for forecasting in autoregressive models rather than predictions from independent variables, the sizes of the datasets are not a justifiable comparison.

After both datasets have been fixed to regularly spaced intervals, the features to be used for forecasting are computed. The developed process utilizes three types of features: moving averages, lagged observations, and gradients (not to be confused with the gradients used to estimate distributional parameters later in the method). The encoded dataset on its own does not provide much value and is difficult to forecast due to the stochastic nature of work sampling, primarily when the dataset is obtained through random walks. To get a useful metric out, a moving average is used to get the average occurrence of classes for several different window sizes. This data transformation is shown in Figure 3 and Figure 4. A summary of the feature engineering is given in Table 3.

Table 3: Overview of datasets and their features.

Features	**Kinematic dataset**	**Work sampling dataset**
Forecasting horizon	90 seconds	60 minutes
Moving averages	90, 180, 360, and 540 seconds	30, 60, 120, and 180 minutes
Lagged values	From 60 to 180 seconds in the past	From 30 to 150 minutes in the past
	17 values evenly spaced	24 values evenly spaced
Gradients	1 for each lagged value	1 for each lagged value
Classes	Direct work, indirect work, waste	Direct work, indirect work, waste

The moving averages are calculated with four different subset lengths. For all four moving averages, a number of lagged values are taken for each prediction (17 for the short-term dataset and 24 for the long-term dataset). For each lagged value, a gradient is calculated by taking the two values surrounding the point and calculating the gradient from the two values. The gradient gives information about the trend around the point, which can be valuable information when forecasting. These calculations are done for all three classes, resulting in 408 features for the short-term dataset and 576 features for the long-term dataset.

The value to be forecasted is chosen to be the 90-second moving average for short-term forecasting and the 60-minute moving average for long-term forecasting. As presented in Table 3, the forecasting horizon is 90 seconds for the short-term dataset and 60 minutes for the long-term dataset. The moving averages are chosen as the forecasting value because of the nature of work sampling data. Figure 3 and Figure 4 visualize the two datasets with their encoded observations and moving averages. The frequency of the forecast will, however, follow the frequency of the two datasets. This means that for the short-term forecasting, the forecasting will happen at 60Hz, and for the long-term forecasting, a point will predicted every 5 minutes, as this is the resampled long-term dataset frequency.

As can be seen, the raw observations are difficult to use without any processing. This is due to how the datasets are collected, where it is not unusual for the long-term dataset to not see one specific class for a 5-minute interval. For the short-term dataset, it is impossible to get values that are not either 0 or 1, as the observations are already regular before processing. Furthermore, a moving average over a large subset flattens out as it covers more values.

3.3 Tree boosting and probabilistic forecasting

When working with forecasting, regression, or function estimation, a system with an output y and input x is given. The system can be split into a training and testing set, where the training set $\{y_i, x_i\}_1^N$ is used to approximate a function $\hat{F}(x)$ of the real unknown function $F(x)$ that maps all given x-values to the output variable y. This approximate function is found by minimizing a loss function $L(y, F(x))$.

Tree boosting has seen state-of-the-art results in several areas of applications (Zhang et al., 2017; Zhang et al., 2020; Chen et al., 2015). The method excels in time-series data or other tabular datasets, consistently outperforming deep learning models across several datasets (Shwartz-Ziv and Armon, 2022). Extreme Gradient

Boosting (XGBoost) is chosen as the tree-boosting model for this project. XGBoost makes parallel tree learning possible through sparsity-aware algorithms, introduces an improved regularization term to the objective function, and develops a cache-aware pre-fetching algorithm to improve the computational speed. It furthermore uses an out-of-core computation principle, which uses the disk to store data blocks and then pre-fetches it during run-time into the main memory buffer. These model features make it fast to run, which is seen as an important metric if the model is extended to an entire construction project, where there will be many individuals whose future metrics potentially need to be estimated. Furthermore, by using out-of-core computations, it is possible to work with very large datasets, which, with other methods, would see bottlenecks in the available memory on the computers usually available on construction projects. A visual representation of the algorithm is given in Figure 5.

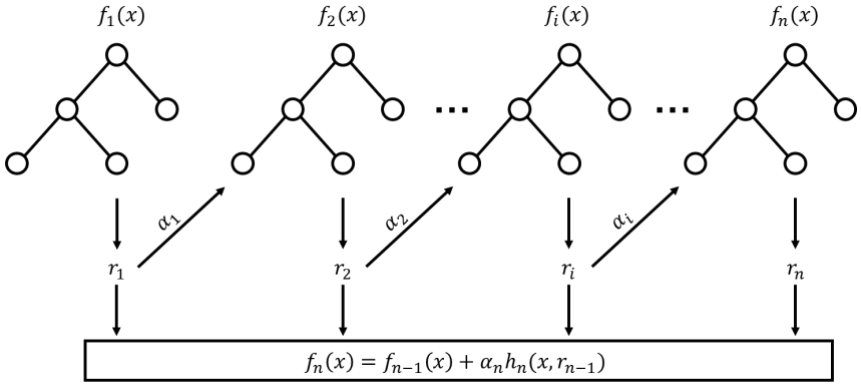

Figure 5: The estimation of an output through gradient tree boosting. Here, α_n is the regularization parameter computed from the nth tree, r_n is the residuals computed from the nth tree, and h_n is the residuals' prediction function.

The XGBoost model utilizes an ensemble of trees to estimate the output, as showcased in Figure 5. The model prediction can be expressed mathematically as shown in Equation 2, where K is the number of trees, f_k is a function in the space of **F**, where **F** is the set of all possible decision trees. For a more detailed explanation, the reader is referred to Chen and Guestrin (2016).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \boldsymbol{F} \tag{2}$$

This estimation of the output is optimized given the objective learning function in Equation 3.

$$\begin{aligned} \mathcal{L}^{(t)} &= \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(t)}\right) + \Omega(f_t) \\ &= \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \end{aligned} \tag{3}$$

The first part of Equation 3 is the training loss, and the second part is the regularization term. ℓ is the loss function, which finds the difference between the prediction at the given iteration and the given instance ($\hat{y}_i^{(t-1)} + f_t(x_i)$) and the true value y_i. The second term $\Omega(f_t)$ is a regularization term that ensures the model does not overfit. This is done by penalizing the complexity of the tree by setting a minimum gain to the model before it can split a leaf into two leaves. Equation 3 is the general regularized objective function. However, this function has a second-order approximation, allowing for faster objective optimization. The second-order approximation is explained in detail by Chen and Guestrin (2016). As most XGBoost implementations use the approximation, this is also the case in this research.

Instead of training a model to forecast a point, in this case, a point displaying the amount of direct work, indirect work, or waste, at a given point in time, probabilistic forecasting attempts to learn a distribution of the possible future values and returns this distribution rather than a point. By having the distribution, more information is

available to assess situations. Rather than only knowing the certainty of a point prediction (there is 75% certainty this prediction is correct), the model can give information about the underlying distribution of possible predictions, which makes it possible to assess risk and ensure that decisions are made on a statistically significant basis.

The distribution's location and scale are estimated through the model training. The distribution of a univariate variable is estimated through the higher moments, which are found through the optimization of a loss function. Most regression models assume higher moments as fixed parameters and only focus on the distribution's estimated mean value (the location). This research investigates the use of the normal distribution, but in future research, more distributions could be examined to understand how different distributions affect the forecast. The normal distribution is chosen as it only has two moments to estimate. Therefore, it is computationally cheaper than using a distribution with higher-order moments that need to be estimated. The estimate of the distributional parameters is done independently by calculating the negative gradient and hessian of the distributional parameter, as presented in equations 4 and 5, where m is the current iteration and θ_k is the current distributional parameter (März, 2019). While the current distributional parameter (θ_k) from the parameter space (θ) is estimated, the other distributional parameters are kept constant. The gradient and hessian of the loss are used for a second-order approximation of the loss in each iteration.

$$\hat{g}_{\theta_k}^{m} = -\left[\frac{\partial \ell\big(y, f(x)\big)}{\partial f(x)}\right]_{f(x)=\hat{f}_{\theta_k}^{(m-1)}(x)} \tag{4}$$

$$\hat{h}_{\theta_k}^{m} = -\left[\frac{\partial^2 \ell\big(y, f(x)\big)}{\partial f(x)^2}\right]_{f(x)=\hat{f}_{\theta_k}^{(m-1)}(x)} \tag{5}$$

When all (in this case the location and scale) distributional parameters are estimated, they are used in conjunction to incorporate the information from the other parameters one by one. The distributional parameters are then updated by incorporating information from the other parameters, so the final output becomes a single function in which all the distributional parameters are present. This distributional estimate is done using the Python framework XGBoostLSS (März, 2019).

The method does not take the logical constraints of the experiment into account yet. This means that the quantiles can become greater than 1 or below 0, but this is not physically possible (having a direct work percentage higher than 100% or lower than 0%). Therefore, for all experiments, the points and the quantiles are constrained to be between 0 and 1. In practice, this is done by replacing all negative values with 0 and those greater than 1 with 1. This is done for both the quantiles and the predicted values. The datasets have three classes, so the problem is defined as a single distributional output. This means that several models will be trained for each class. In practice, this means that three models, as a minimum, would be needed to estimate the levels of all three classes.

To summarize, the forecasting method is split into short-term and long-term forecasts, which differ in feature sets. The long-term forecast is resampled to a regular time series, which is more suited for a forecasting problem. The models all estimate the gradients and hessians, and rather than training one tree per iteration, a tree is trained for both the location and the scale of the distribution. In practice, this will lead to a higher runtime than XGBoost, as double the number of trees are trained. As the framework gives a distribution for the timesteps, 500 samples are drawn from the distribution to estimate the quantiles.

3.4 Performance evaluation

To evaluate each model created against the other, four evaluation metrics are chosen for the assessment. Two metrics focusing on the point forecast, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), as well as two distributional metrics, Percentage of Points between the 25th and 75th percentile (PP50) and between the 5th and 95th percentile (PP90). The MAE was chosen over MAPE, as both datasets have true values close to 0 and, therefore, will have relatively large percentage errors. Furthermore, as the test sets of some classes do not range from 0 to 1, comparing the MAPE between models could be misleading. By using MAE, the comparison should be easier to make, as the absolute error will transfer well to the understanding as the maximum range is from 0 to 1. The four metrics are presented in order in equations 6-9. In equations 8 and 9, *card* is the cardinality of the set of true values that fall within the range of the percentiles of the predicted normal distribution.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\hat{y}_i^{(t)} - y_i\right)^2}{n}} \quad (6)$$

$$MAE = \frac{\sum_{i=1}^{n} \left|\hat{y}_i^{(t)} - y_i\right|}{n} \quad (7)$$

$$PP50 = \frac{card(y_{50})}{n} \quad (8)$$

$$PP90 = \frac{card(y_{90})}{n} \quad (9)$$

The two different approaches to evaluating the models are made because of the two aspects that are important for the overall objective. The first is that the forecast is as accurate as possible, meaning that the forecast has the smallest residuals possible. Apart from this obvious objective, it is also essential to understand what range the productivity might fall into in the future, as this is a more reliable metric than a point forecast. The distributional accuracy of the models is evaluated using the metrics PP50 and PP90, which find the percentage of true values that fall within the two intervals.

4. RESULTS

The results of this research can be split into two: The long-term forecasting and the short-term forecasting. As mentioned earlier, the two datasets represent the two extremes of a construction project, one at the project level and the other at the individual level. Both datasets are split into training and testing datasets. As they are both collected from multiple instances (several painters for the short-term dataset and several projects for the long-term dataset), the split between training and testing data is done by leaving out a painter and a project. This means that the testing of the models, which will be presented in the following subsections, is done on a subsample of the datasets that the models have never seen. For the short-term dataset, this is a painter the model has not seen before, which means that the timing of the classes is unique to this specific part of the dataset. Similarly, the long-term test set is a project the model has never seen. This dataset split ensures that the model evaluation is done on a generalizable basis. For each dataset, three different classes are to be forecasted. To do so, the class is isolated from the dataset, which means that when forecasting direct work, this class is excluded from the dataset and used to evaluate the model estimate. The same process is then done for indirect work and waste. This means that the distributional output of the models is for one specific class at a time.

To evaluate the method on the presented datasets, several experiments are run to investigate the optimal hyperparameters of the model. Three different numbers of boost rounds are examined (500, 1500, 3000), as well as the use of stabilization of the derivative when computing the gradients and hessians. The stabilization is done through a regularization term by taking the square root of the mean derivative squared. If this is smaller than 0.0001, the gradient is replaced by 0.0001; similarly, if it is larger than 10,000, it is replaced by 10,000.

All other hyperparameters are kept constant but could, for future research, be evaluated. The learning rate is set to 0.001, the maximum depth of the trees to 15, the minimum loss reduction required to partition leaves to 9.745e-06, the subsampling ratio to 0.927 and the minimum sum of hessian needed in a child is set to 2. When the distributions have been estimated, 500 samples are drawn to estimate the point value (mean of the sample), the percentiles, and the distributional parameters.

In total, 36 models are trained, 18 for the short-term dataset and 18 for the long-term dataset. For both datasets, the 18 models fall into the three classes of the work sampling study (direct work, indirect work, and waste). This means that for each unique forecasting objective, six models are created.

4.1 Short-term forecasting

When examining the performance of the models presented in Table 4, the models that use L2 stabilization of the derivatives are very similar in performance. This is expected as stabilization is implemented to ensure a faster convergence. The slower convergence when not using stabilization is due to the variability of the ranges of the distributional parameters, which can make the model struggle with convergence due to very large or very small gradients. This means that, for most of the cases analyzed, a model with 500 trees with stabilization is enough to get the best performance, or very close to it, for the given hyperparameters.

Table 4: Forecasting models of the short-term dataset split into the three classes: direct work, indirect work, and waste (for all three classes, the best-performing model(s) in each metric are marked in bold).

Class	Model number	Number of trees	Derivative stabilization	RMSE	MAE	PP50	PP90
Direct work	1	500	None	0.234	0.192	**0.352**	**0.734**
	2	1500	None	0.183	0.153	0.165	0.451
	3	3000	None	0.173	0.141	0.054	0.130
	4	500	L2	**0.153**	**0.126**	0.150	0.331
	5	1500	L2	**0.153**	**0.126**	0.148	0.329
	6	3000	L2	**0.153**	**0.126**	0.149	0.328
Indirect work	7	500	None	0.111	0.090	**0.333**	**0.558**
	8	1500	None	0.095	0.079	0.152	0.353
	9	3000	None	0.094	0.079	0.026	0.100
	10	500	L2	0.092	0.073	0.157	0.366
	11	1500	L2	0.092	0.073	0.156	0.356
	12	3000	L2	**0.091**	**0.071**	0.179	0.373
Waste	13	500	None	0.278	0.231	**0.302**	**0.639**
	14	1500	None	0.183	0.147	0.164	0.466
	15	3000	None	**0.148**	**0.113**	0.052	0.129
	16	500	L2	0.182	0.139	0.160	0.345
	17	1500	L2	0.182	0.140	0.161	0.345
	18	3000	L2	0.182	0.139	0.161	0.346

Models with the least number of trees all excel in the PP90 interval metric, meaning that a significant amount of the true values are within the boundaries of the interval (73.4% for model 1). Without stabilization, the models converge slower, meaning the distributions' scales are larger. This is shown in Figure 6, where models 1-3 are compared. Here, the model with the least trees has a significantly larger scale than the model with the most trees.

The comparison in Figure 6 shows the importance of understanding the evaluation metrics to choose a suitable model. If model 1 is selected, this will mean that 73.4% of the values fall within the range from the 5^{th} to the 95^{th} percentile of the distribution. However, this range is much larger in this model compared to models 2 and 3, as depicted in Figure 6. The formatting of Figure 6 is done to visually show the relative confidence of the models, depicted by the gradient of the sampled mean prediction (from red to white). As the distribution scale becomes smaller, this gradient shifts from red to white, indicating a smaller scale and a more confident estimate from the model. The shift in the colored area is used to depict the percentiles of the distribution, with the light blue interval being between 5% and 95% and the dark blue being between 25% and 75%.

When comparing the performance across models for each of the three classes (see Table 5), the mean direct work model performs better than indirect work and waste in the two distributional metrics (PP50 and PP90) but is significantly worse than Indirect Work in the two residual-based metrics (RMSE and MAE). The two metrics that use the residuals to estimate accuracy are lower for the Indirect work class due to the smaller range in data, with only around half of the full range (0-1) used. Overall, the distributional performances of models with a high number of trees are suspected to suffer from overfitting, as the scale of the distribution becomes significantly smaller in

these models. This indicates that fewer trees are sufficient for the given forecasting problem.

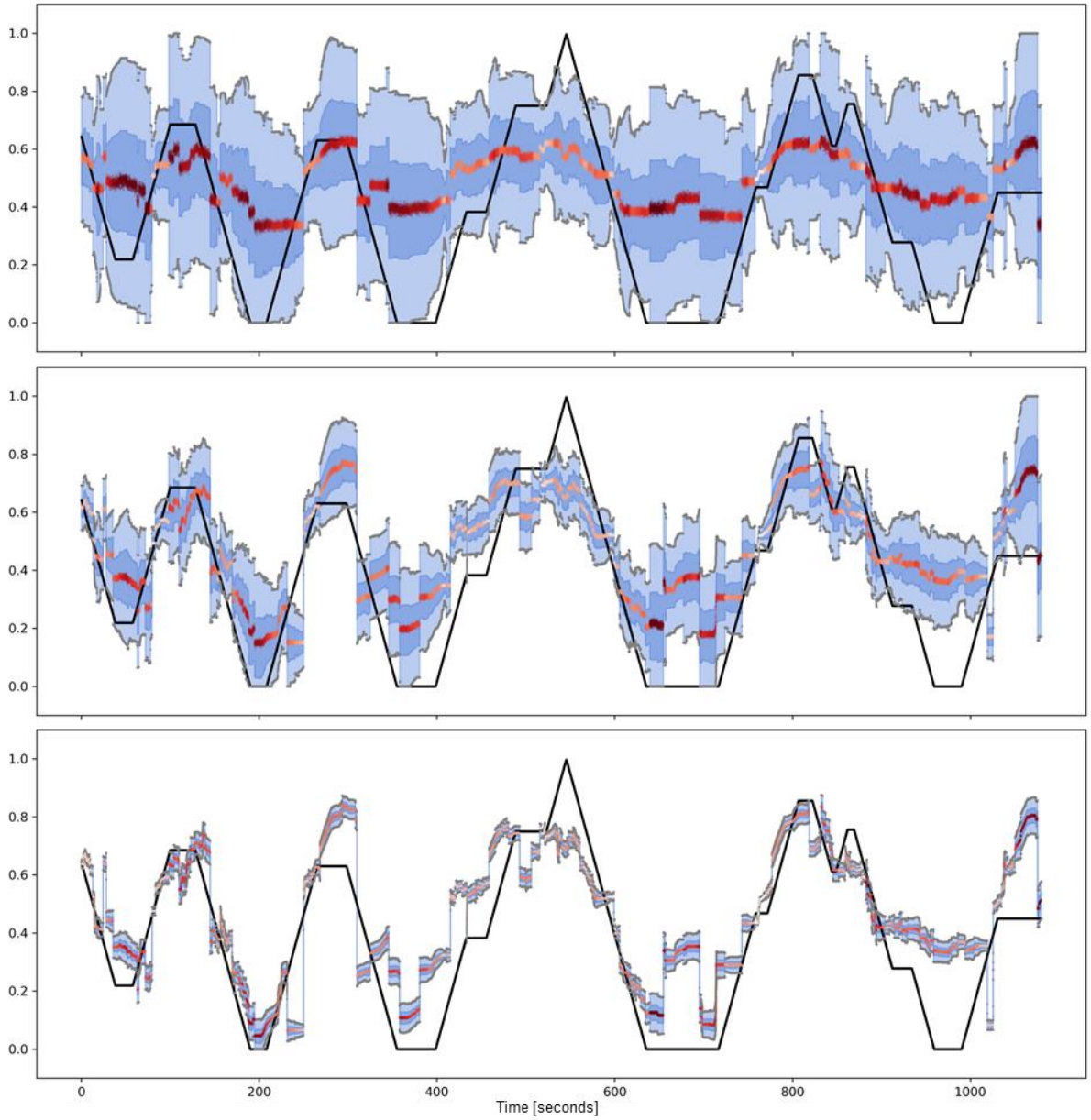

Figure 6: The probabilistic forecasting of direct work using three different models (top: 500 trees, middle: 1500 trees, bottom: 3000 trees). The light-blue area is the PP90 interval, and the dark-blue area is the PP50 interval. The black line is the true value.

Table 5: Mean value of metrics for each of the three classes for the short-term dataset (direct work, indirect work, and waste).

Class	RMSE	MAE	PP50	PP90
Direct work	0.175	0.144	0.170	0.395
Indirect work	0.096	0.078	0.167	0.351
Waste	0.193	0.152	0.167	0.378

4.2 Long-term forecasting

For the long-term dataset, the 18 models are presented in Table 6. Just as for the kinematic dataset (short-term forecasting), the forecasting models with stabilization in the long-term forecasting have similar scores across evaluation metrics. The main difference between the models trained for short-term forecasting and those trained for long-term forecasting is the distributional performance of the stabilized models. In the work sampling dataset, the stabilized models perform best or very close to the top performer across all metrics. This is different when compared to short-term forecasting, as the stabilized model had a generally lower distributional performance.

Table 6: Forecasting models of the long-term dataset split into the three classes: direct work, indirect work, and waste (for all three classes, the best-performing model(s) in each metric is marked in bold).

Class	Model number	Number of trees	Derivative stabilization	RMSE	MAE	PP50	PP90
Direct work	19	500	None	0.064	0.051	0.528	0.687
	20	1500	None	0.032	0.025	0.446	0.673
	21	3000	None	**0.021**	**0.014**	0.159	0.389
	22	500	L2	0.022	0.018	**0.713**	**0.728**
	23	1500	L2	0.023	0.019	**0.713**	**0.728**
	24	3000	L2	0.022	0.019	**0.713**	**0.728**
Indirect work	25	500	None	0.109	0.090	0.546	0.781
	26	1500	None	0.059	0.047	0.434	0.750
	27	3000	None	0.043	**0.029**	0.191	0.392
	28	500	L2	**0.039**	0.030	0.755	**0.801**
	29	1500	L2	**0.039**	0.030	0.755	**0.801**
	30	3000	L2	**0.039**	0.030	**0.756**	**0.801**
Waste	31	500	None	0.071	0.059	0.366	0.628
	32	1500	None	0.039	0.029	0.357	0.633
	33	3000	None	**0.027**	**0.016**	0.168	0.375
	34	500	L2	**0.027**	0.020	**0.691**	**0.716**
	35	1500	L2	**0.027**	0.020	**0.691**	**0.716**
	36	3000	L2	**0.027**	0.019	**0.691**	**0.716**

The long-term forecasting from the work sampling dataset generally performs better than the short-term forecasting when examining the evaluation metrics. The best models perform well in both the point prediction evaluation, meaning how far the mean value of the sample is to the actual point, and when examining the distributional evaluation metrics, where the number of actual points within the distribution quantiles is evaluated. The models with stabilization and 500 trees from each class in the work sampling dataset are shown in Figure 7. These models are chosen from the three classes as they are computationally cheap to train and perform either the best or very close to the best. As seen in Figure 7, the overall ranges of true values are smaller in the long-term dataset than those in the short-term dataset, making the residuals smaller in absolute terms, leading to a lower MAE.

Similarly to the short-term dataset, the overall trend is caught by all the models portrayed in Figure 7. Only the most extreme trends (significant increase or decrease in values over a short time) are difficult for the model to forecast, with even extreme values being forecasted with acceptable residuals. The main difference between the two forecasting tasks (short- and long-term) is the performance in the probabilistic metrics. It is expected that the significant difference in the distributional parameters is due to the difference in the datasets. The work sampling dataset is more fluctuating, as it contains multiple trades and workers, making it difficult for the models to overfit the distributional moments. This is not the case for the short-term dataset, which could be why the distributional scale is generally smaller.

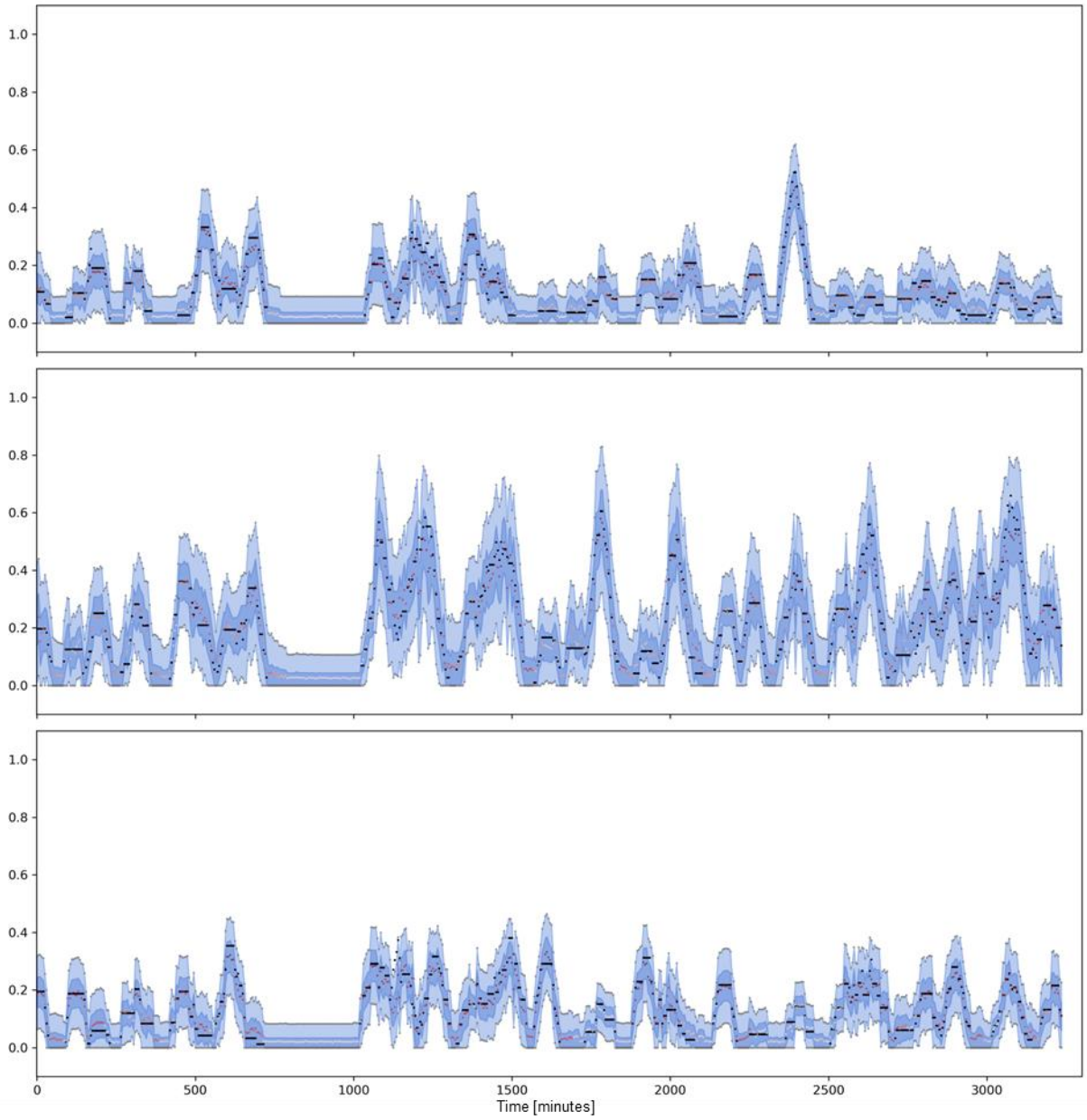

Figure 7: The best model for each class from the long-term forecasting dataset (top: Model 21 - direct work, middle: Model 29 - indirect work, bottom: Model 33 - waste). The figure follows the formatting of Figure 6.

4.3 Sampling frequency

To examine the effect of sampling frequency, three models with the same hyperparameters are trained from the short-term dataset, as the models trained from the short-term dataset have worse performance across all metrics compared to the long-term dataset. Model 4 is chosen as the baseline (containing 17 lagged values), and two extra models are trained with more features. One uses every 50th observation from the raw dataset as lagged values (containing 108 lagged values), and the other uses every 20th observation as lagged values (containing 270 lagged values). This is done for the moving averages shown in Table 3. A visual comparison of the three models can be seen in Figure 8.

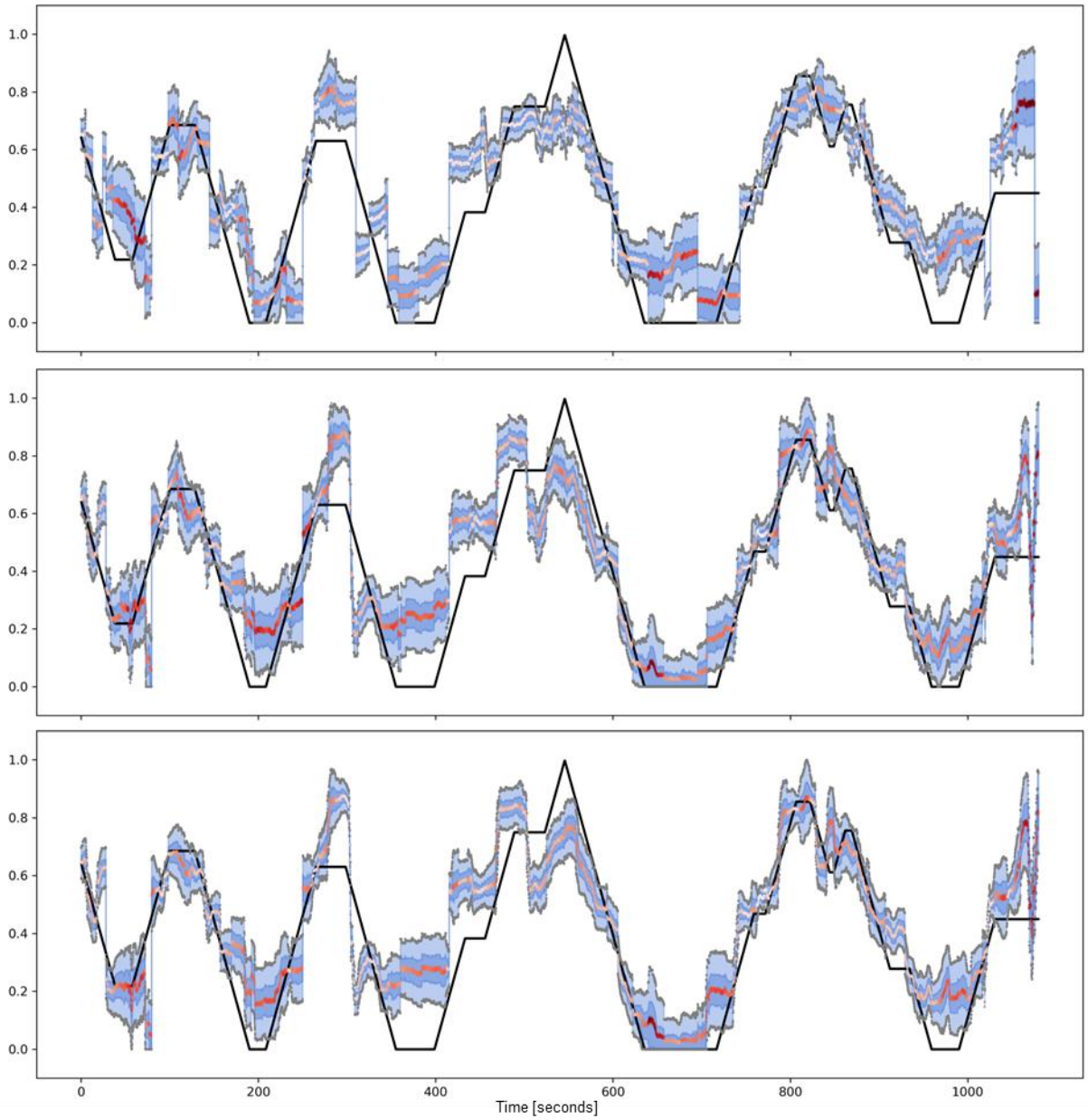

Figure 8: The comparison of the three models for Direct Work from the kinematic dataset (top: Original model, middle: Every 50th observation, bottom: Every 20th observation). The figure follows the formatting of Figure 6.

The experiment results are presented in Table 7, in which the model with the original number of features (Model 4) is not the best-performing model. However, it can also be seen that adding more features is only a good performance enhancer up to a given point. Going from using every 50th observation to every 20th decreases the performance across all four metrics. The decrease in performance is very little and could, therefore, be interpreted as a stagnation in development. If it is the case that the performance is decreasing, the problem could be an overfitting issue, where the models will put too much attention on the local trend. A more comprehensive analysis should be made to find the optimal frequency, which could be different for the long-term dataset compared to the short-term dataset. Furthermore, the stagnation of performance when increasing the frequency should be investigated.

Table 7: Comparison of models with original features, lagged values every 50th observation, and lagged values every 20th observation.

Model	RMSE	MAE	PP50	PP90
Original (Model 4)	0.153	0.126	0.150	0.331
Every 50th observation	0.137	0.107	0.219	0.460
Every 20th observation	0.138	0.109	0.213	0.441

The three models presented in Figure 8 and Table 7 catch the trends of the data but struggle with the extremes (both in terms of extreme point values and extreme short-term changes in values showcased by a sudden increase or decrease). With more feature engineering, the models could potentially increase their robustness and, this way, catch the extreme values of the trends presented in the test set better as more information would thereby be available for the prediction. However, the results presented in Table 7 indicate that lagged values are not the right features to focus on for increasing performance as a stagnation of the improvement is quickly met but could be used in combination with more sophisticated feature engineering or independent variables. Independent variables as extra features could make the model performance increase. As this data has not been available for these experiments due to the information not being recorded or made available by the projects, it has not been possible to examine it. Independent variables in the form of categorical data have been successfully implemented features in previous studies, presented in the background section of this paper, which signals a potential group of variables that could be further investigated. However, as the objective of this research is to suggest a simpler model in terms of data collection, the independent variables that could be used alongside the historical values, such as weather information, should be possible to collect automatically.

When looking at PP50 and PP90 for the three models in Table 7, they have significantly lower performance than Models 1 (PP50: 0.352; PP90: 0.734), 7 (PP50: 0.333; PP90: 0.558), and 13 (PP50: 0.302; PP90: 0.639). The highest performing model in Table 7, when considering the distributional metrics, is the model with every 50th observation, which is worse than all the three models mentioned above, with PP50 of 0.219 (27.5% worse than model 13) and PP90 of 0.460 (18% worse than model 7). However, as all the models presented in Table 7 and Figure 8 are with stabilization, they converge to distributions with a relatively small range. This means that the two intervals become very narrow, sometimes only covering 0.05 (5% direct work, indirect work, or waste). By having a small range in the distribution, misplacing it quickly leads to the true value falling outside the distribution. In these cases, the two metrics using the residuals say more about the performance, with the MAE being 0.107 for the model with every 50th observation as part of the feature set. Even though the distribution often obtains a small range when using stabilization, there is a significant increase from the original model (Model 4) to the model using every 50th observation when looking at the distributional performance. This indicates a more frequent dataset is preferred. However, as the model using every 20th observation performs slightly worse than the model with every 50th observation, there is a stagnation point, where the model's performance does not increase or even loses accuracy.

As showcased through visualizations and evaluation metrics, the two datasets have different performances in the forecasting models. The long-term dataset is easier to forecast, even though this dataset is used for long-term forecasting. A reason for the difference in performance could be the number of points between the known value and the value to be forecasted. Due to the higher frequency, this range is significantly larger in the short-term forecasting dataset than in the long-term forecasting dataset. This means that the number of unknown points, which can affect the trend, is greater in the short-term dataset. It is important to note that the two datasets differ in collection method, processing, and applicability. Comparing model performance between the two datasets is not a good indication of which area is the easiest for forecasting. As the short-term dataset has a wider range (0-1), the evaluation metrics have the chance of being larger.

The overall results of the forecasting models make it possible to understand future levels of direct work, indirect work, and waste with satisfactory accuracy. The probabilistic nature of the forecasts makes it possible to obtain information regarding the point that is predicted and the scale of the underlying distribution estimated by the algorithm. This is true for both short-term and long-term forecasts, which will give new insights into both the individual worker and their cycles and the overall project and the distribution of work sampling classes on the project.

5. DISCUSSION

The paper presents a framework for short- and long-term forecasting of CLP metrics. The paper utilizes a probabilistic version of XGBoost, which allows for the estimation of distributional parameters rather than a point forecast. The number of features used in the autoregressive model is evaluated through a case study where the lagged features are altered, and it becomes clear that for the current setup, a stagnation in performance happens before the entirety of the past values are used.

The models are all capable of forecasting the trend of the classes, with the long-term dataset having an MAE below 3% and the short-term dataset having an MAE below 13%. The 10 percent point difference in performance is due to the rapid fluctuation in the short-term forecast, as this dataset concerns the work of a single worker. The fluctuation of productivity will be significantly less when looking at a large sample of workers combined, as done in the long-term dataset. However, even with the 10 percent point difference, both extremes (single worker and entire construction site) are valid areas for applying auto-regressive forecasting. The method is still not readily applicable, as the horizon is relatively short. Extending the horizon would enable decision-makers to make informed decisions, which is deemed impossible when only having a horizon of 60 minutes.

The proposed method for forecasting construction labor productivity metrics still has several limitations. The method, alongside the rest of the prediction algorithms for productivity, needs more data, preferably a continuous stream of data, but this data needs to be collected. The long-term forecasting presented in this paper is currently based on work sampling data, which was manually collected for 29 days on a construction site. The data is irregular (when done through random walks) and cumbersome to collect, so the data collection method needs to be changed. Using the output of classification algorithms would enable a regular time-series dataset that could be directly streamed from the on-site workers.

Further research is needed to go from individual collection to a general construction site dataset. This would be necessary to support both short- and long-term forecasting. A suggestion for such a data collection system is presented by Jacobsen et al. (2023) and used as the short-term dataset in this research. If this method was implemented in the entire project construction phase duration, it would be possible to use the same dataset for both the short- and long-term forecasting. However, more work would be needed to transform the individual workers' data stream into a dataset representing the entire construction site (required for long-term forecasting).

As this is among the first attempts at using autoregressive models for monitoring construction productivity, several initiatives are still not fully explored. As mentioned, the system should work in run-time when the data availability issue is resolved. Implementing a model that can utilize streamed data would give information that would allow for more frequent cycles, including decisions or corrections to the current production system, which could be valuable.

The area of autoregressive modeling or forecasting generally has numerous methods not examined in this paper. To ensure optimal performance is found, other methods than boosted trees should be explored. This could be deep learning methods such as DeepAR, which has seen state-of-the-art accuracy on several datasets. DeepAR could also help generalization, as boosted trees could potentially struggle in the transferability between trades, which is one of the areas in which DeepAR excels (Salinas et al., 2019).

Boosted trees can tend to overfit in the deeper trees. An investigation of the timing of this occurrence would make it possible to get even better performance from the models, which especially could push the performance of the distributional parameters. In general, the hyperparameters of the models have only been slightly tuned. A deeper examination of the hyperparameters would potentially show a significant performance boost. This could be done using a hyperparameter optimization framework to search the large spaces of hyperparameter combinations. As mentioned earlier, independent variables or features could also be used alongside the historical values. To ensure this research's objective is kept the same, these variables should be easily collectible.

In practice, the proposed framework is a first step towards a forecasting system that can assist in monitoring the construction site and be a foundation for decisions. To move the system into practice, the horizon of the forecast needs to be moved further ahead, making the system predict days ahead rather than the 60 minutes that is currently used. The short-term forecasting is not yet proven generalized between trades, currently only having data from a painting job. This needs to be done before the method can be implemented on a construction project. To do so, a high-quality dataset from a construction site consisting of several trades needs to be collected. The methodology's potential is, in practice, a better-monitored construction project. Understanding future states of the metrics presented in this paper will make planning of activities, procurement, and layout planning more informed processes, where the cycles of workers and the overall productivity of future states can be taken into account.

6. CONCLUSION

Being able to forecast metrics that tie to the progress of a construction project allows for better insights and the possibility of making informed decisions regarding changes to the production system. The information from the long-term forecasting about the project's future states can be used to evaluate decisions regarding the logistical planning, the site layout, and the timing of work packages. The individual productivity metric forecasting makes it easier to understand cycle times and could be used to predict the finishes of activities. The forecast could, if combined with location data from the workers, also give insights into the spatial differences in productivity (some areas might be easier to be highly productive in).

The major contribution of this research is the ability to do probabilistic productivity metric forecasting based on historical data. To the authors' knowledge, this is the first research that forecasts CLP metrics in an autoregressive fashion without the need for other variables. By only using the past performance of the workers, the forecasting method is temporally more granular and requires less effort in data collection. This method differs from similar papers by only using the previously observed data points and by estimating the additional moments of the distribution. By doing so, forecasting becomes an easily accessible method for understanding the construction project. The use of boosted trees has been successful in numerous fields of forecasting and is an effective model for large datasets, which is expected if an entire construction project is to be monitored. The developed models for individual productivity monitoring (short-term forecasting) range in performance from 0.278 to 0.091 RMSE, 0.231 to 0.071 MAE, 0.026 to 0.352 PP50, and 0.100 to 0.734 PP90.

Long-term forecasting models based on work sampling data have also been developed. The combination of short-term forecasting of individual workers and long-term forecasting of the entire construction project is seen as a valuable combination. This gives insight into the general state of the project and also the cycle times of the activities. The two aspects cover the two extreme granularities of a construction project. The developed models for construction site productivity monitoring from the long-term dataset range in performance from 0.109 to 0.021 RMSE, from 0.090 to 0.014 MAE, from 0.159 to 0.756 PP50, and from 0.375 to 0.801 PP90.

The two different datasets have different optimal models. For the short-term dataset, models with fewer trees perform better when looking at distributional performance but have bigger residuals on average. Having derivative stabilization makes the convergence faster and makes the smallest models, in terms of trees, similar in performance compared to the bigger models with and without stabilization, even when looking at residuals. For the work sampling dataset, models with the highest numbers of trees perform the best, even when looking at the distributional performance. Models with stabilization tend to perform very similarly due to the fast convergence. The difference between the largest models with and without stabilization is significant, with the best option being with stabilization. Overall, most of the models developed fall within acceptable ranges for the evaluation metrics. As the application of the method proposed throughout this project is to provide information for informed decision-making, the acceptable range can change from project to project. However, for ordinary projects, it is deemed reasonable to forecast with an absolute error of 10%. This will be enough for the decision-makers to intervene when drastic changes in productivity are predicted, as they will fluctuate significantly more than 10%. This means that the models can successfully forecast individual short-term productivity metrics with a 90-second horizon and general long-term productivity metrics with a 60-minute horizon. As the datasets have not been used in other forecasting research, comparing results to the research of Table 1 is deemed illogical both due to the difference in objective and the difference in data used for the evaluation.

A study of extra features was done to investigate the importance of the number of features presented to the models. As the short-term dataset has 60 data points available every second, an investigation was done into how often a sample is needed. The algorithms can quickly become computationally expensive, so resampling the raw dataset could become necessary. The overall conclusion for this investigation is that maximum performance is found before the full frequency of the dataset is used. More experiments should be done by resampling the raw features and adding feature engineering or including independent variables.

With the ongoing development of the construction industry, a continuous forecast of productivity for each worker will be a valuable added feature for many processes. Both for continuous analyses, such as simulations of production systems, but also for the decision makers on construction sites, that can use an estimate of future productivity metrics to make decisions before the productivity starts to decrease. This paper sets a direction with the introduction of forecasting productivity using autoregression. With larger productivity datasets becoming available, the forecasting horizon can be increased to cover not only a single day but potentially forecast days or even weeks into the future.

REFERENCES

Adebowale O. J. and Agumba J. N. (2022). A scientometric analysis and review of construction labour productivity research, *International Journal of Productivity and Performance Management*, Vol. ahead-of-print, Issue ahead-of-print, https://doi.org/10.1108/IJPPM-09-2021-0505.

Allmon E., Haas C. T., Borcherding J. D. and Goodrum, P. M. (2000). U.S. construction labor productivity trends, 1970–1998, *Journal of Construction Engineering and Management*, Vol. 126, Issue 2, https://doi.org/10.1061/(ASCE)0733-9364(2000)126:2(97).

Ashuri B. and Lu, J. (2010). Time series analysis of construction cost index, *Journal of Construction Engineering and Management*, Vol. 136, Issue 11, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000231.

Assaad R. and El-adaway I. H. (2021). Impact of dynamic workforce and workplace variables on the productivity of the construction industry: new gross construction productivity indicator, *Journal of Management in Engineering*, Vol. 37, Issue 1, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000862.

Bai S., Li M., Kong R., Han S., Li H. and Qin L. (2019). Data mining approach to construction productivity prediction for cutter suction dredgers, *Automation in Construction*, Vol. 105, 102833, https://doi.org/10.1016/j.autcon.2019.102833.

Barbosa A. S. and Costa D. B. (2021). Productivity monitoring of construction activities using digital technologies: a literature review, *Proceedings of the 29th Annual Conference of the International Group for Lean Construction (IGLC29)*, pp. 707-716, https://doi.org/10.24928/2021/0141.

Buchan R. D., Fleming F. W. and Grant F. (2003). Estimating for builders and surveyors, *Routledge*, ISBN 9780750642712.

Caldas C. H., Kim J., Haas C. T., Goodrum P. M. and Zhang, D. (2015). Method to assess the level of implementation of productivity practices on industrial projects, *Journal of Construction Engineering and Management*, Vol. 141, Issue 1, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000919.

Cao Y. and Ashuri B. (2020). Predicting the volatility of highway construction cost index using long short-term memory*, Journal of Management in Engineering*, Vol. 36, Issue 4, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000784.

Chen T. and Guestrin C. (2016). XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

Chen T., Singh S., Taskar B. and Guestrin C. (2015). Efficient second-order gradient boosting for conditional random fields, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015,* pp. 147-155, https://proceedings.mlr.press/v38/chen15b.

Cheng C. F., Rashidi A., Davenport M. A. and Anderson D. V. (2017). Acoustical modeling of construction jobsites: hardware and software requirements, *ASCE International Workshop on Computing in Civil Engineering 2017*, https://doi.org/10.1061/9780784480847.044.

Cheng T., Venugopal M., Teizer J. and Vela P. A. (2011). Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments, *Automation in Construction*, Vol. 20, Issue 8, https://doi.org/10.1016/j.autcon.2011.05.001.

Cheng M., Cao M. and Mendrofa A. Y. J. (2021). Dynamic feature selection for accurately predicting construction productivity using symbiotic organisms search-optimized least square support vector machine, *Journal of Building Engineering*, Vol. 35, 101973, https://doi.org/10.1016/j.jobe.2020.101973.

CII. (2010). Guide to Activity Analysis, *Construction Industry Institute*

Dai J., Goodrum P. M. and Maloney W. F. (2007). Analysis of craft workers' and foremen's perceptions of the factors affecting construction labour productivity, *Construction Management and Economics*, Vol. 25, Issue 11, https://doi.org/10.1080/01446190701598681.

Dissanayake M., Fayek A. R., Russell A. D. and Pedrycz W. (2005). A Hybrid Neural Network for Predicting Construction Labour Productivity, *Computing in Civil Engineering*, https://doi.org/10.1061/40794(179)78.

Dai J., Goodrum P. M. and Maloney W. F. (2009). construction craft workers' perceptions of the factors affecting their productivity, *Journal of Construction Engineering and Management*, Vol. 135, Issue 3, https://doi.org/10.1061/(ASCE)0733-9364(2009)135:3(217).

Ebrahimi S., Fayek A. R. and Sumati V. (2021). Hybrid artificial intelligence HFS-RF-PSO model for construction labor productivity prediction and optimization, *Algorithms*, Vol. 14, Issue 7, https://doi.org/10.3390/a14070214.

Ebrahimi S., Kazerooni M., Sumati V., and Fayek A. R. (2022). Predictive Model for Construction Labour Productivity Using Hybrid Feature Selection and Principal Component Analysis, *Canadian Journal of Civil Engineering*, Vol. 49, Issue 8, https://doi.org/10.1139/cjce-2021-0248.

El-Gohary K. M., Aziz R. F. and Abdel-Khalek H. A. (2017). Engineering approach using ANN to improve and predict construction labor productivity under different influences, *Journal of Construction Engineering and Management*, Vol. 143, Issue 8, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001340.

Erharter G. H. and Marcher T. (2021). On the pointlessness of machine learning based time delayed prediction of TBM operational data, *Automation in Construction*, Vol. 121, https://doi.org/10.1016/j.autcon.2020.103443.

Fan G., Zheng Y., Gao W., Peng L., Yeh Y., Hong W. (2023). Forecasting residential electricity consumption using the novel hybrid model, *Energy and Buildings*, Vol. 290, 113085, https://doi.org/10.1016/j.enbuild.2023.113085.

Gao B., Wang R., Lin C., Guo X., Liu B. and Zhang W. (2021). TBM penetration rate prediction based on the long short-term memory neural network, *Underground Space*, Vol. 6, Issue 6, https://doi.org/10.1016/j.undsp.2020.01.003.

Gao X., Shi M., Song X., Zhang C. and Zhang H. (2019). Recurrent neural networks for real-time prediction of TBM operating parameters, *Automation in Construction*, Vol. 98, https://doi.org/10.1016/j.autcon.2018.11.013.

Gigerenzer G., Hertwig R., Van Den Broek E., Fasolo B. and Katsikopoulos K.V. (2005). "A 30% chance of rain tomorrow": how does the public understand probabilistic weather forecasts?, *Risk Analysis*, Vol. 25, Issue 3, https://doi.org/10.1111/j.1539-6924.2005.00608.x.

Golnaraghi S., Moselhi O., Alkass S. and Zangenehmadar Z. (2020). Predicting construction labor productivity using lowerupper decomposition radial base function neural network, *Engineering Reports*, Vol. 2, Issue 2, https://doi.org/10.1002/eng2.12107.

Golnaraghi S., Zangenehmadar Z., Moselhi O. and Alkass S. (2019). Application of artificial neural network(s) in predicting formwork labour productivity, *Advances in Civil Engineering*, Vol. 2019, 5972620, https://doi.org/10.1155/2019/5972620.

Gong J., Borcherding J. D. and Caldas C. H. (2011). Effectiveness of craft time utilization in construction projects, *Construction Management and Economics*, Vol. 29, Issue 7, https://doi.org/10.1080/01446193.2011.595013.

Gong J. and Caldas C. H. (2011). An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations, *Automation in Construction*, Vol. 20, Issue 8, https://doi.org/10.1016/j.autcon.2011.05.005.

Gouett M. C., Haas C. T., Goodrum P. M. and Caldas C. H. (2011). Activity analysis for direct-work rate improvement in construction, *Journal of Construction Engineering and Management*, Vol. 137, Issue 12, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000375.

He Y., Han A., Hong Y., Sun Y. and Wang S. (2021). Forecasting crude oil price intervals and return volatility via autoregressive conditional interval models, *Econometric Reviews*, Vol. 40, Issue 6, https://doi.org/10.1080/07474938.2021.1889202.

Heravi G. and Eslamdoost E. (2015). Applying artificial neural networks for measuring and predicting construction-labor productivity, *Journal of Construction Engineering and Management*, Vol. 141, Issue 10, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001006.

Hwang S., Park M., Lee H. and Kim H. (2012). Automated time-series cost forecasting system for construction materials, *Journal of Construction Engineering and Management*, Vol. 138, Issue 11, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000536.

Hui L., Park M. and Brilakis I. (2015). Automated brick counting for façade construction progress estimation, *Journal of Computing in Civil Engineering*, Vol. 29, Issue 6, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000423.

Ilbeigi M., Ashuri B. and Joukar A. (2017). Time-Series analysis for forecasting asphalt-cement price, *Journal of Management in Engineering*, Vol. 33, Issue 1, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000477.

Jacobsen E. L. and Teizer J. (2022). Deep learning in construction: review of applications and potential avenues, *Journal of Computing in Civil Engineering*, Vol. 36, Issue 2, https://doi.org/10.1061/(ASCE)CP.1943-5487.0001010.

Jacobsen E. L., Wandahl S. and Teizer J. (2023). Work estimation of construction workers for productivity monitoring using kinematic data and deep learning, *Automation in Construction, Vol. 152*, https://doi.org/10.1016/j.autcon.2023.104932.

Joshua L. and Varghese K. (2011). Accelerometer-Based activity recognition in construction, *Journal of Computing in Civil Engineering*, Vol. 25, Issue 5, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000097

Joshua L. and Varghese K. (2014). Automated recognition of construction labour activity using accelerometers in field situations, *International Journal of Productivity and Performance Management*, Vol. 63, Issue 7, https://doi.org/10.1108/IJPPM-05-2013-0099.

Joukar A. and Nahmens I. (2016). Volatility forecast of construction cost index using general autoregressive conditional heteroskedastic method, *Journal of Construction Engineering and Management*, Vol. 142, Issue 1, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001020.

Kalsaas B. T., Gundersen M. and Berge T. O. (2014). To measure workflow and waste. A concept for continuous improvement, *Proceedings of the 22nd Annual Conference of the International Group for Lean Construction (IGLC22)*, pp. 835-846, https://www.iglc.net/Papers/Details/1078.

Kazaz A., Manisali E. and Ulubeyli S. (2008). Effect of basic motivational factors on construction workforce productivity in Turkey, *Journal of Civil Engineering and Management*, Vol. 14, Issue 2, https://doi.org/10.3846/1392-3730.2008.14.4.

Kazarooni M., Nguyen P. and Fayek A. R. (2021). Prioritizing construction labor productivity improvement strategies using fuzzy multi-criteria decision making and fuzzy cognitive maps, *Algorithms*, Vol. 14, Issue 9, https://doi.org/10.3390/a14090254.

Kopsida M., Brilakis I. and Vela P. A. (2015). A review of automated construction progress monitoring and inspection methods, *Proceedings of the 32nd CIB W78 Conference "Applications of IT in the Architecture, Engineering and Construction Industry,* pp. 421-431, https://itc.scix.net/pdfs/w78-2015-paper-044.pdf.

Koskela L. (1992). Application of the new production philosophy to construction, *CIFE Technical report 72. Stanford University.*

Koskela L. (2000). An exploration towards a production theory and its application to construction, *VTT Publications, VTT Technical Research Centre of Finland.*

Liou F. and Borcherding J. D. (1986). Work sampling can predict unit rate productivity, *Journal of Construction Engineering and Management*, Vol. 112, Issue 1, https://doi.org/10.1061/(ASCE)0733-9364(1986)112:1(90).

Liu K. and Golparvar-Fard M. (2015). Crowdsourcing construction activity analysis from jobsite video streams, *Journal of Construction Engineering and Management*, Vol. 141, Issue 11, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001010.

Luo X., Li H., Cao D., Yu Y., Yang X. and Huang T. (2018). Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks, *Automation in Construction*, Vol. 94, https://doi.org/10.1016/j.autcon.2018.07.011.

März A. (2019). XGBoostLSS - An extension of xgboost to probabilistic forecasting, *arXiv preprint*, https://doi.org/10.48550/arXiv.1907.03178.

So M. K. P., Lam K. and Li W. K. (1999). Forecasting exchange rate volatility using autoregressive random variance model, *Applied Financial Economics*, Vol. 9, Issue 6, https://doi.org/10.1080/096031099332032.

Mirahadi F. and Zayed T. (2016). Simulation-based construction productivity forecast using Neural-Network-Driven Fuzzy Reasoning, *Automation in Construction*, Vol. 65, pp. 102-115, https://doi.org/10.1016/j.autcon.2015.12.021.

Momade M. H., Shahid S., Hainin M. R. B., Nashwan M. S. and Umar A. T. (2020). Modelling labour productivity using SVM and RF: a comparative study on classifiers performance, *International Journal of Construction Management,* Vol. 22, Issue 10, https://doi.org/10.1080/15623599.2020.1744799.

Muqeem S., Idrus A., Khamidi M. F., Ahmad J. B. and Zakaria S. B. (2011). Construction labor production rates modeling using artificial neural network, *Journal of Information Technology in Construction*, Vol. 16, pp. 713-726, https://www.itcon.org/2011/42.

Nasirzadeh F., Kabir H. M. D., Akbari M., Khosravi A., Nahavandi S. and Carmichael D. G. (2020). ANN-based prediction intervals to forecast labor productivity, *Engineering, Construction and Architectural Management*, Vol. 27, Issue 9, https://doi.org/10.1108/ECAM-08-2019-0406.

Neve H., Wandahl S., Lindhard S., Teizer J. and Lerche J. (2020). Learning to see value-adding and non-value-adding work time in renovation production systems, *Production Planning & Control*, Vol. 33, Issue 8, https://doi.org/10.1080/09537287.2020.1843730.

OECD (2023). Defining and measuring productivity, *OECD*, https://www.oecd.org/sdd/productivity-stats/40526851.pdf [visited 14/02/2023].

Oral E. L. and Oral M. (2010). Predicting construction crew productivity by using Self Organizing Maps, *Automation in Construction*, Vol. 19, Issue 6, https://doi.org/10.1016/j.autcon.2010.05.001.

Oral M. and Oral E. L. (2007) A computer based system for documentation and monitoring of construction labour productivity, *Proceedings of the 24th CIB W78 conference "Bringing ICT knowledge to work"*, pp. 345-350, https://doi.org/10.13140/2.1.3862.5283.

Potočnik P., Škerl P. and Govekara E. (2021). Machine-learning-based multi-step heat demand forecasting in a district heating system, *Energy and Buildings*, Vol. 233, Issue 15, https://doi.org/10.1016/j.enbuild.2020.110673.

Rashid K. M. and Louis J. (2020). Activity identification in modular construction using audio signals and machine learning, *Automation in Construction*, Vol. 119, https://doi.org/10.1016/j.autcon.2020.103361

Ryu J., Seo J., Jebelli H. and Lee S. (2019). Automated Action Recognition Using an Accelerometer-Embedded Wristband-Type Activity Tracker, *Journal of Construction Engineering and Management*, Vol. 145, Issue 1, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001579

Salinas D., Flunkert V., Gasthaus J. and Januschowski T. (2020). DeepAR: probabilistic forecasting with autoregressive recurrent networks, *International Journal of Forecasting*, Vol. 36, Issue 3, https://doi.org/10.1016/j.ijforecast.2019.07.001.

Sanders S. R. and Thomas H. R. (1993). Masonry productivity forecasting model, *Journal of Construction Engineering and Management*, Vol. 119, Issue 1, https://doi.org/10.1061/(ASCE)0733-9364(1993)119:1(163).

Shangxin F., Zuyu C., Hua L., Shanyong W., Yufei Z., Lipeng L., Daosheng L. and Liujie J. (2021). Tunnel boring machines (TBM) performance prediction: A case study using big data and deep learning, *Tunnelling and underground Space Technology*, Vol. 110, https://doi.org/10.1016/j.tust.2020.103636.

Shwartz-Ziv R. and Armon A. (2022). Tabular data: deep learning is not all you need, *Information Fusion*, Vol. 81, pp. 84-90, https://doi.org/10.1016/j.inffus.2021.11.011.

Smith S. D. (1999). Earthmoving productivity estimation using linear regression techniques, *Journal of Construction Engineering and Management*, Vol. 125, Issue 3, https://doi.org/10.1061/(ASCE)0733-9364(1999)125:3(133).

Swei O. (2020). Forecasting infidelity: why current methods for predicting costs miss the mark, *Journal of Construction Engineering and Management*, Vol. 146, Issue 2, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001756.

Taylor S. J. and Letham B. (2018). Forecasting at scale, *The American Statistician*, Vol. 72, Issue 1, https://doi.org/10.1080/00031305.2017.1380080.

Thomas H. R. and Yiakoumis I. (1987). Factor model of construction productivity, *Journal of Construction Engineering and Management*, Vol. 113, Issue 4, https://doi.org/10.1061/(ASCE)0733-9364(1987)113:4(623).

Tixier A. J., Hallowell M. R., Rajagopalan B. and Bowman D. (2016). Application of machine learning to construction injury prediction, *Automation in Construction*, Vol. 69, pp. 102-114, https://doi.org/10.1016/j.autcon.2016.05.016.

Tsehayae A. A. and Fayek A. R. (2016). System model for analysing construction labour productivity, *Construction Innovation*, Vol. 16, Issue 2, https://doi.org/10.1108/CI-07-2015-0040.

Yi W. and Chan A. P. C. (2014). Critical review of labor productivity research in construction journals, *Journal of Management in Engineering*, Vol. 30, Issue 2, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000194.

Wandahl S., Pérez C. T., Salling S. and Lerche J. (2022). Robustness of work sampling for measuring time waste, *Proceedings of the 30th Annual Conference of the International Group for Lean Construction (IGLC30)*, pp. 247-258, https://doi.org/10.24928/2022/0127.

Wandahl S., Neve H. H. and Lerche J. (2021). What a waste of time, *Proceedings of the 29th Annual Conference of the International Group for Lean Construction (IGLC29)*, pp. 157-166, https://doi.org/10.24928/2021/0115.

Wong J. M. W., Chan A. P. C. and Chiang Y. H. (2005). Time series forecasts of the construction labour market in Hong Kong: the Box-Jenkins approach, *Construction Management and Economics*, Vol. 23, Issue 9, https://doi.org/10.1080/01446190500204911.

Xu J. and Moon S. (2013). Stochastic Forecast of Construction Cost Index Using a Cointegrated Vector Autoregression Model, *Journal of Management in Engineering*, Vol. 29, Issue 1, https://doi.org/10.1061/(ASCE)ME.1943-5479.0000112.

Zhang C., Liu C., Zhang X. and Almpanidis G. (2017). An up-to-date comparison of state-of-the-art classification algorithms, *Expert Systems with Applications*, Vol. 82, pp. 128-150, https://doi.org/10.1016/j.eswa.2017.04.003.

Zhang W., Zhang R., Wu C., Goh A. T. C., Lacasse S., Liu Z. and Liu H. (2020). State-of-the-art review of soft computing applications in underground excavations, *Geoscience Frontiers,* Vol. 11, Issue 4, https://doi.org/10.1016/j.gsf.2019.12.003.