Forecasting the Deformation of Quay Walls in Amsterdam

A Model Fusion Approach

submitted in partial fulfillment for the degree of master of science

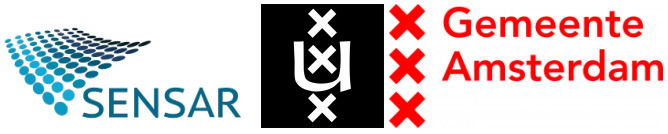
Julian El-Fasih
11302380

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 30.06.2023

UvA Supervisor
dr. F. Nack
Universiteit van Amsterdam
f.m.nack@uva.nl

External Supervisor
P. Karamitopoulos
Gemeente Amsterdam
p.karamitopoulos@amsterdam.nl



Forecasting the Deformation of Quay Walls in Amsterdam

A Model Fusion Approach

Julian El-Fasih | 11302380

ABSTRACT

The state of the quay walls in Amsterdam has deteriorated. In the past years, they have been monitored using tacheometry. The aim of this study is to forecast the time-series of the tacheometry data based on CPT and InSAR data. This was done by developing different configurations of BiLSTM models. The configurations conformed to three types of model fusion: early fusion, incremental fusion, and late fusion. These were compared to four baselines: a naive method and models taking input from each dataset separately. The models were compared based on their RMSE, MAE, and MASE scores. Incremental fusion yielded the significantly best results both on itself and in combination with late fusion. Early fusion did not benefit the performance. This is in line with previous studies using incremental and late model fusion, and contrary to a study using early fusion to predict deformation.

KEYWORDS

Time Series Forecasting, Quay Walls, Deformation, Tacheometry data InSAR data, CPT data, Model Fusion, BiLSTM

GITHUB REPOSITORY

https://github.com/Amsterdam-Internships/QuayWallDeformation

1 INTRODUCTION

Amsterdam and water are inseparable, as the canals and bridges provide for the city's character. Alongside the canals, 600 kilometers of quay walls support the streets and sidewalks [13]. The walls are historic and were designed and built for less load than they suffer today. They were built on a foundation of wooden piles, and instead of horses and carriages before, trucks are passing and cars are parked on top of the walls nowadays. As a result, their state has deteriorated [12]. A typical construction of a quay wall is depicted in Figure 1.

In the past years, the worrying state of the quay walls came to attention. Sinkholes arose at several walls [25, 26, 39], and parts of the Nassaukade and the Grimburgwal even collapsed [1, 27]. After the collapse of the Grimburgwal in 2020, various studies investigated the cause. Korff et al. concluded in 2022 that this was mainly caused by two mechanisms. First, the piles bent horizontally, causing a sinkhole and reduced stability. Second, the wooden base collapsed, inducing the failure of the entire quay [21]. A rapid assessment after the collapse of the Grimburgwal in 2020 concluded the horizontal bending of the piles to be a result of a deeper water bottom [20]. The cause of the deeper bottom was deemed to be the turning of boats. These would collide with the wall repeatedly, damaging the quay. The trigger for the collapse was considered to be extra load caused temporarily by the renewal of the street on top of the quay. The renewal was necessary because of the already occurred deformation of the quay wall [20].

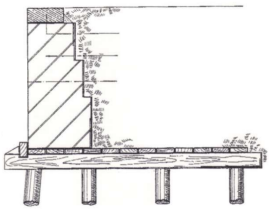


Figure 1: Cross-section of a quay wall built on a wooden platform supported by wooden piles.
Note. **Adapted from "Amsterdamse Risicobeoordeling: Kademuren: Constructieve staat Amsterdamse kademuren (Kademuur op houten palen)" by Neijzing & Westein, 2022, p. 11. Copyright 2022 by Neijzing & Westein.**

The municipality of Amsterdam started a large-scale investigation of the quay walls and bridges in order to gain a better insight into the actual state of the structures and prevent similar incidents. Additionally, the municipality started monitoring the behavior of the construction. This resulted in a better understanding and a lot of data regarding the state of the constructions [13]. With this data and inspections, the Amsterdamse Risicobeoordeling Kademuren (ARK) was created to quantify the risk of every quay wall. The risk is divided into a chance score indicating the risk of the wall collapsing and a consequence score representing the risk for the city if the wall were to collapse. The chance score consists of six factors, each with a weight of 1.0, 2.0, or 3.0 [24].

As a part of monitoring, tacheometry was used to measure the deformation of quay walls in the vertical, longitudinal, and lateral directions over time [3]. These were used to define a deformation score, part of the chance score of ARK with a weight of 2.0. Moreover, with these time-series, Dahmen (2022) tested various models to forecast the deformation of quay walls [10]. His results showed a non-parametric Bidirectional Long Short-Term Memory (BiLSTM) model to yield the best performance. The input consisted of the time series gathered by tacheometry and a boolean indicating the presence of a parking spot on top of the quay wall. This paper continues the work of Dahmen to obtain more comprehensive knowledge of the subject [10].

No other papers are available on forecasting the deformation of quay walls. Consequently, for inspiration on how the analysis and predictions can be performed, similar domains such as dam deformation are considered. Dams are similar because they form a border for water, as do quay walls, and both are affected by erosion and traffic load [11].

Xie et al. (2019) developed bootstrap aggregating regression trees to predict dam deformation. They used various predictor variables and the time series of displacement as input and concluded their model to outperform traditional regression trees [40].

Cao et al. (2020) attempted to predict dam deformation as well. They combined two decomposition models with an autocorrection

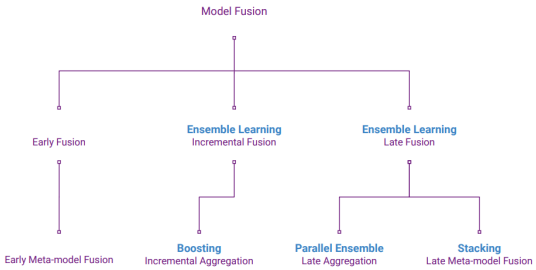


Figure 2: Taxonomy of Model Fusion.
Note. **Adapted from "Evaluating State-of-the-Art, Forecasting Ensembles and Meta-Learning Strategies for Model Fusion" by P. Cawood and T. van Zyl, 2022, *Forecasting, 4*(3), p. 734. Copyright 2022 by P. Cawood and T. van Zyl.**

analysis and extreme learning machines based on parameter optimization. The conclusion was drawn that their method performed best compared to other tested models [5].

Moreover, Ren et al. (2020) predicted dam deformation using a support vector machine (SVM). A modified fruit fly optimization algorithm (MFOA) was used to optimize the hyper-parameters of the SVM. The results of their model showed better accuracy and robustness than conventional models [29].

Furthermore, Chen et al. (2021) used an ensemble learning model-based on clustering to predict dam deformation. They concluded their model to yield outstanding performance in terms of prediction accuracy [8].

The aforementioned studies all combined models to yield better predictions. This is called model fusion [6]. As these studies show, model fusion is used widely and successfully when predicting dam and bridge deformation. Cawood and van Zyl defined model fusion as follows in 2022:

> We define model fusion as integrating base learners to produce a lower biased and variance; and more robust, consistent, and accurate model than the learners individually. These base learners can either be homogeneous from the same hypothesis class (e.g. decision trees in a random forest) or be heterogeneous from different hypothesis classes (e.g. neural network with a support vector machine for the classification). [6, p. 733]

It is further categorized into early model fusion, late model fusion, and incremental model fusion. Early model fusion combines base learners before the training phase, enabling the integrated model to be trained as one. Late model fusion comprises base learners being trained individually before being integrated. Incremental model fusion integrates models incrementally during the training phase. The integrated base learners keep their fixed parameters after training.

Furthermore, Cawood and van Zyl argue that the process of model integration can be divided into two categories. For meta-model fusion, the integration process is carried out utilizing model-based meta-learning. Differently, model integration by aggregation fusion is accomplished by employing a straightforward aggregation technique such as weighted averaging [6]. These definitions result in the taxonomy shown in Figure 2.

Accordingly, we can categorize the aforementioned studies on the deformation of bridges and dams into these classes. Firstly, the study by Xie et al. (2019) is an example of late fusion as they use bootstrapped aggregated regression trees, which are characteristic to late aggregation. Secondly, the model developed by Cao et al. (2020) combines multiple sub-models to form a single model before the training phase. Therefore, it is an example of early fusion. Thirdly, Ren et al. (2020) can be categorized as incremental fusion because of the hyperparameter optimization of the SVM by the MFOA. Fourthly, the clustering-based ensemble learning method proposed by Chen et al. (2021) consists of base learners that were trained individually before being integrated using a weighted average method. Hence, this model is another example of late fusion.

Even though model fusion has been widely used for predicting deformation, mostly concerning time series forecasting and yielding promising results, model fusion has not yet been used to predict the deformation of quay walls yet. Therefore, this study will fill the research gap by predicting the deformation of quay walls, the case of Amsterdam serving as the example, by comparing different categories of model fusion based on temporal and static data. This will be done by answering the following research question:

How do models conforming to early, incremental, and late model fusion perform when predicting the deformation of quay walls based on temporal and static data?

To examine the influence of the specific data sources on the performance of the models and to assess whether the models could potentially substitute the tacheometry measurements, the following subquestions are formulated:

- How do the different datasets influence the performance of the models?
- What is the effect of leaving out the preceding deformation values as a predictor?

The hypothesis is that the fused models perform better than separate models since model fusion yielded better results than conventional methods in previous deformation studies [5, 8, 29, 40].

In the subsequent sections, the paper is structured as follows. Section 2 introduces the datasets, along with their exploration and processing. Section 3 explains the implementation of the models. In Section 4, the results are presented. These are discussed in Section 5. The paper is concluded in Section 6.

2 DATA

For this study, five datasets were made available. These comprised tacheometry measurements of quay walls and of buildings - both provided by the municipality of Amsterdam [4], results of cone penetration testing (CPT) and of bore analysis - both obtained through BROLoket [30], and interferometric synthetic aperture radar (InSAR) measurements of surface settlement provided by Sensar [35]. A summary of their nature and size is provided in Table 1. All datasets were loaded into a Jupyter Notebook [19] using the Python [37] library Pandas [23]. Visualizations were made with Seaborn [38] and Matplotlib [15]. The notebooks can be found in the Github repository, containing the data exploration

Table 1: Summary of the datasets, indicating their nature and size.

Dataset	Nature	Locations	Time-span
Tacheometry QW	Temporal	5817	11-2018 | 12-2021
Tacheometry B	Temporal	12937	11-1948 | 09-2021
CPT	Static	5371	-
Bore	Static	222	-
InSAR	Temporal	448	01-2017 | 01-2022

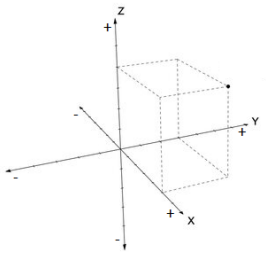


Figure 3: X-, Y-, and Z-direction. The cube represents a quay wall. The direction of negative values is indicated by a (-), and the direction of positive values by a (+). Negative values for Y indicate movement towards the water; negative values for Z indicate sinking of the wall.

and processing, training, and evaluation of models. It does not contain the data, as it is in most cases not public. The bore data and the building tacheometry data were omitted from this study because of measurement sparsity. The building tacheometry data contained only 28 measurements after 2016, whereas the bore data had insufficient locations, as shown in Table 1.

2.1 Tacheometry

The first dataset contained the target variable of this study and was downloaded from a platform of the municipality of Amsterdam [4]. The platform is not publicly accessible, and the data is confidential.

Tacheometry is a surveying method to measure distances and elevations using a telescopic instrument. Based on angular measurements, the coordinates of a point can be derived [14]. This method was used to repeatedly determine the coordinates of bolts, inserted into the quay walls for this purpose. Over time, these indicate the deformation of the walls in X-, Y-, and Z-directions. An illustration of the directions is given in Figure 3. Failure mechanisms do not occur in the X-direction, indicating the Y- and Z-directions to be most important [24]. Based on the thresholds shown in Table 2, the measurements in the Y- and Z-directions make up the deformation score, contributing to the chance score part of the quantified risk in ARK [24]. Important to note is that the margin of error for tacheometry measurements is 2.5mm in every direction [34].

The dataset consisted of 126 Excel files, one per rack. A rack is the part of a quay wall between two bridges. In total, the files covered about 25 out of 600 kilometers of quay walls [12] and contained the initial coordinates of 5817 measuring bolts and the difference in mm with regard to the initial position over time. The locations of the bolts are shown in Figure 12 in Appendix B. Each bolt had between 2 and 28 measurements in each direction, ranging from November

Table 2: Thresholds for the deformation score in mm. The rate is per month. Both thresholds are only for the direction Y and Z. The score is obtained when either or both the cumulative deformation and deformation rate are met.

Score	Cumulative (mm)	Rate (mm)
1	<15	<5
2	15 - 25	5 - 10
3	25 - 35	10 - 15
4	>35	>15

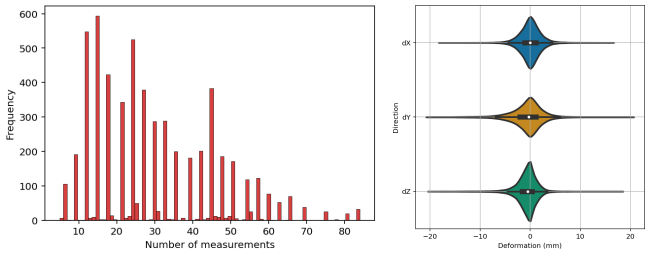


Figure 4: The histogram at the left shows the distribution of measurements per bolt, including every direction. The violin plot at the right shows the distribution of the deformation values for each direction.

2018 to December 2021. Figure 4 shows the distribution of measurements per bolt. Most bolts had between 4 and 15 measurements in each direction. The figure also shows that the bolts had measurements for all three directions most of the time. Apart from NaN, missing values were indicated by 'NULM', 'n.g.', 'n.v.t.', 'ng', 'n.v.t..', 'g.v.m', 'n.g', 'vervallen', or 'n,g,'. Table 3 provides summary statistics of the tacheometry measurements in X-, Y-, and Z-directions. It stands out that the Z-direction contained fewer measurements. The maximum and minimum values seem unrealistic, given the unit of millimeters and the fact that quay walls are usually not higher than one or two meters. This is also reflected in the standard deviation. The Z-direction does not show any clear unrealistic value. Figure 4 shows the distribution of the values between -20mm to 20mm for every direction. The deformation in the X-direction shows a near-perfect normal distribution, whereas the Y- and Z-direction show a negatively skewed distribution. This makes sense as the X-direction is parallel to the wall, and the distribution of the Y- and Z-direction indicate the walls to move more downwards and towards the water, which is linked to failure mechanisms [24].

From every file, the ID, initial coordinates, dates of measurement, and measurement values were extracted. Bolts without measurement were removed, as were bolts that had no initial coordinates. All non-numeric values indicating the absence of a measurement were substituted by NaN. All values exceeding 1000mm were considered unrealistic because a quay wall is usually not more than 2 meters high. A deformation value of more than half the height of the wall would probably indicate the collapse of a quay wall, an event that did not occur for the monitored walls during the timespan of the data. Accordingly, these values were removed. This amounted to 138, 91, and 0 values for the X-, Y-, and Z-direction,

Table 3: Summary statistics of the deformation values in mm. Minimum values show clear outliers for the direction X and Y. This is reflected in the difference between the mean and median as well.

Stat	X	Y	Z
count	53239	53239	50,788
mean	-3536.18	-8544.47	-1.51
std	85467.94	206484.90	17.08
min	-2115280.00	-4999433.00	-722.70
25%	-1.10	-2.10	-1.60
50%	0.00	-0.30	-0.50
75%	1.20	1.10	0.40
max	6146.30	587.70	96.30

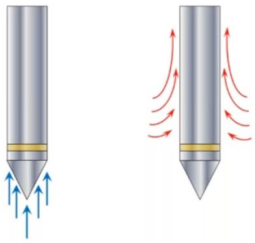


Figure 5: Cone probe measures resistance (left) and friction (right). *Note.* Adapted from "Cone Penetration Testing (CPT) for Geotechnical Investigations" by E. Wassenaar, 2022, *Numac webinar, 7*. Copyright 2022 by E. Wassenaar.

respectively. Subsequently, the dataset was split into three, one for each of the directions X, Y, and Z. The resulting subsets consisted of columns representing the dates of measurement, and rows representing the values per measuring bolt. The values were aggregated per month to obtain consistency in time intervals. In case multiple measurements were registered for a bolt within one month, the mean value of these measurements was used. However, several bolts remained without measurements for certain months. Linear interpolation was applied for missing values of a bold to address this. To be able to do this for missing values at the beginning of a time-series, a month was added before the first measurement in the dataset, and the corresponding value for all bolts was set to zero. All three resulting datasets consisted of measurements of 5187 bolts over 38 months.

2.2 CPT

The second dataset was obtained through BROLoket [30], which is publicly available. After selecting a polygon on the map, the data within can be requested. The requested selection for this study corresponded to the polygon the tacheometry datapoints constructed.

The data regarded the subsurface of the quay walls in Amsterdam and was gathered using Cone Penetration Testing (CPT). This method comprises a cone-shaped probe being inserted into the ground. It measures resistance and friction at different depths, as shown in Figure 5. The behavior of the soil at a certain depth can be derived from these two variables, which can be considered an indicator of the strength of the subsurface [31]. This is expected to be a suitable predictor for the deformation of quay walls.

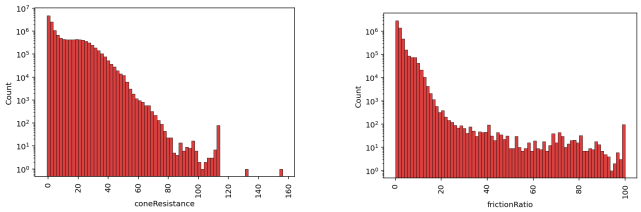


Figure 6: Histograms of Cone Resistance and Friction Ratio

The data consisted of 5442 XML files, containing a total of 1.4 million points with 12 variables. Every file corresponded to a CPT of one location and contained the cone resistance and friction ratio of several depths, besides metadata regarding the sampling conditions. The files were read using Geotexxx [36]. The datapoints were registered per depth. Depth values ranged from 0 to 74 meters. The variables cone resistance and friction ratio consisted of 0.43% and 63.4% of NaN, respectively. Figure 6 shows the distribution of values for cone resistance and friction ratio. For both variables, as the value increased from zero, its occurrence generally became more sparse.

Duplicate rows and rows containing only NaN-values were removed from the dataset. To obtain a single row per CPT, for each row, the cone resistance and friction ratio were averaged per depth interval of 0.5m, and the maximum elapsed time was taken as this is a cumulative variable. Missing values for the friction ratio and cone resistance were filled with the last valid value for a previous depth, and any remaining missing values were filled with the first valid value of the following depth. Missing values for the elapsed time were filled with the median elapsed time. The metadata was removed, except for the coordinates. With the coordinates, it could be checked whether the CPT was conducted at a quay wall. The polygons of the quay walls were constructed by their coordinates, with a buffer of 2 meters. Using the spatial join function of GeoPandas [17], the quay wall ID was added to the CPT datapoint if it was contained by the polygon of the quay wall. The remaining data covered 30 quay walls, with the remaining features cone resistance at depth, friction ratio at depth, and elapsed time. These were scaled to range between zero and one. The locations of the datapoints can be seen in Figure 12 in Appendix B.

2.3 InSAR

The third dataset was provided by Sensar [35], a company that processes InSAR data and is not publicly accessible. It exclusively contained the time-series of InSAR measurements regarding the surface deformation of the quay walls in the vertical and horizontal direction, corresponding to the Y- and Z-direction in Figure 3, respectively, and of the hinterland in the vertical direction. InSAR is a technique to measure the deformation of the Earth's surface using satellite sensors [41]. The sensors send radar signals toward the surface and receive the echoed signals. These signals make up a radar image of the surface. As each satellite passes the target region once every eleven days, the differences in these radar images can be analyzed to obtain deformation on a millimeter scale. The satellites either have an ascending or a descending orbit, sending signals in either the east or the west direction. The detection of movement is particularly effective in this direction. Therefore, for quay walls oriented north-south, horizontal deformation can be

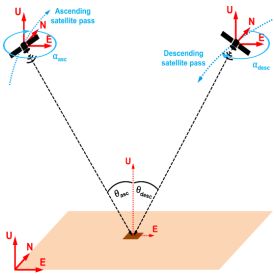


Figure 7: Decomposition is only possible with the signals of two opposite viewing angles.
Note. **Reprinted from "De AmsterScan: Efficiënte monitoring van kademuren en bruggen" by Sensar, 2022. Copyright 2022 by Sensar.**

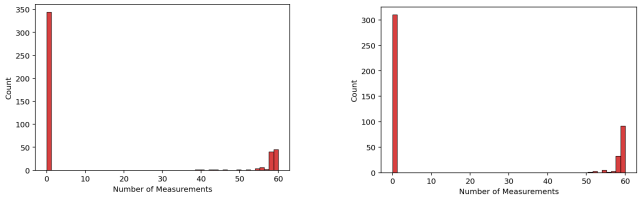


Figure 8: Distributions of measurement per time-series for the vertical and horizontal direction

measured most accurately. For quay walls oriented east-west, only vertical deformation can be measured. To be able to measure both vertical and horizontal deformation at the same time, decomposition has to be applied. This is only possible when signals from opposite viewing angles are available as illustrated in Figure 7.

Reinders et al. (2021) concluded that InSAR data serves as a valuable complement to conventional monitoring when looking at surface deformation due to shield tunneling [28]. Accordingly, the InSAR data is expected to be a suitable predictor for quay wall deformation.

The dataset consisted of three subsets containing the deformation of the quay wall's surface over time, one for the vertical direction, one for the horizontal direction, and one for the deformation of the surface of the hinterland. Each subset contained 448 time-series divided over the quay wall racks, ranging from 5 to 50 per rack, as the surface of each rack was subdivided into segments of 10 to 15 meters wide. The data ranged from January 2017 to January 2022. Along with metadata of the satellites, the sensors, and the quality of measurements, each subset contained the data of the 30 quay walls which had datapoints for both the tacheometry data and the CPT data, as seen in Figure 12 of Appendix B. The histograms in Figures 8 and 9 show that a lot of the time-series consisted of only NaN-values. The time-series that did contain measurements had at least 40 values. The violin plot in Figure 9 shows the distribution of the values for all three subsets. All three had a negative median value. The subset of the hinterland had the widest range, the horizontal subset was the most spread out, and the vertical subset had the most values around zero.

From the subsets, the metadata was removed, except for the coordinates. For some of the quay walls, the deformation could be measured in both the horizontal and the vertical direction at the same time. Consequently, the vertical subset contained columns

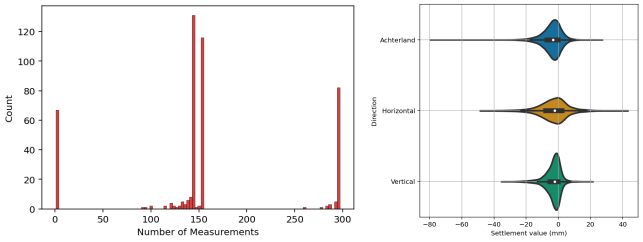


Figure 9: Distribution of measurement per time-series for hinterland and violin plot showing the distribution of values for the three datasets

for the horizontal deformation and vice versa. These were removed as the information was already present in the other subset. The data was aggregated by Sensar for every 30 days. As a result, on one occasion, a month consisted of two columns. For consistency with the tacheometry data, the data was aggregated per month in the same way. Because of missing measurements for several bolts for certain months, the data was also interpolated in the same way. The measurements of the segments were aggregated to obtain one time-series per quay wall rack.

The intersection of the three datasets covered 30 quay wall racks and time-series of 1662 bolts between November 2018 and December 2021. Accordingly, this study focused on the following racks:

- BDG0101
- GRW0101
- KVV0602
- LLG0202
- PRG0301
- SIN0601
- BDG0202
- HEG0302
- KVV0702
- LYG0501
- PRG0401
- SIN0602
- DCG0101
- HEG0401
- KZG0301
- LYG0601
- PRG0402
- SIN0701
- DCG0202
- HEG0802
- KZG0302
- NHG0201
- SIN0501
- WEG0201
- DCG0302
- HEG0901
- LLG0102
- OVW0601
- SIN0502
- WKN0101

The summary statistics for the tacheometry data of these racks are displayed in Tables 7, 8, and 9 in Appendix A.

3 MODELS

The data was split into a training and a test set. 70% of the quay wall racks were used for training, 30% were used for testing. This split was done whilst preserving clusters of racks along the same canal to be able to evaluate the generalizability of the models to unseen racks. This was done for all three datasets. For the tacheometry and the InSAR data, the first 70% of the months were used for training and the last 30% for testing. The order of the months was preserved to be able to evaluate how the models predict the future. This split was not done for the CPT dataset because of its static nature. Subsequently, cross-validation was applied to the training set by creating multiple folds. For each fold, a training window of 10 months and a label window of 5 months were created, as this setting was also used in the work of Dahmen [10]. The training window was created for the tacheometry data and the InSAR data. The CPT data was repeated to obtain the same shape. The label window was created for the tacheometry data, as this was the target variable. The split yielded 10 windows for training, 4 for validation, and 1 for testing.

To answer the research question, models conforming to early, incremental, and late fusion were developed. In every case, a BiLSTM model was used, as Dahmen showed in 2022 that BiLSTM models yielded the best results when forecasting the deformation of quay walls in Amsterdam [10].

Similar to a Long Short Term Memory (LSTM) model, a BiLSTM model consists of memory cells allowing the model to remember and forget information over time, enabling it to capture temporal dependencies [33]. Differently, it contains two layers to process sequential data in both forward and backward directions simultaneously, allowing the model to also incorporate future context in the predictions. Accordingly, BiLSTM models are explicitly suited for handling time series.

To determine whether the tacheometry measurements could potentially be substituted by the predictions, all developed models were trained with and without the preceding tacheometry measurements as predictors.

Figure 13 in Appendix B shows that the tacheometry values of every bolt were correlated, indicating the measurements of one bolt could be predictors for the measurements of another. A small experiment was conducted to see whether the models should be developed per quay wall rack, per canal, or all at once. The experiment regarded only the tacheometry data. The results are shown in Table 10 in Appendix A. Developing models per rack yielded the best results. Therefore, all further experiments were conducted by training models for every rack separately.

3.1 Baselines

To assess if the developed models perform better than a simple method, a baseline was construed by the naive forecasting method (BL1). This method takes the last period of the training data and returns this as the prediction for the next period. This period is equal to the label window.

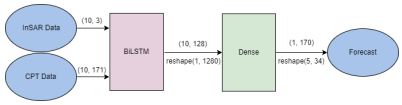
To evaluate the performance of fusion in general, the fusion models were compared to models trained on tacheometry (BL2), CPT (BL3), and InSAR (BL4) individually. The data was passed to sequential models consisting of a BiLSTM layer with 64 hidden units and a ReLu activation function [2], as this was used in the optimal model of Dahmen [10]. The output was flattened and passed to a dense layer, after which the original shape was restored.

3.2 Early Fusion

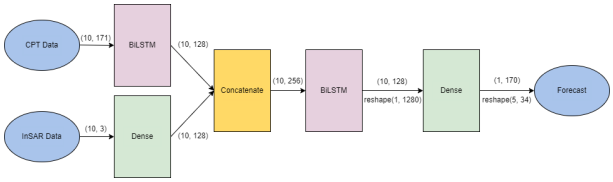
For early fusion, models integrating the datasets before the training phase were constructed. Accordingly, both the CPT and InSAR data were passed to a sequential model (EF1) with the configuration as stated in Section 3.1. The architecture can be seen in Figure 10a.

3.3 Incremental Fusion

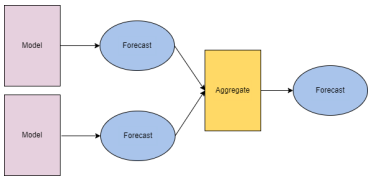
For incremental fusion, models integrating the datasets during the training phase were developed. This was done by first passing the CPT data to a sequential model (IF1) with the configuration as stated in Section 3.1. At the same time, the InSAR data was passed to a dense layer. The two outputs were concatenated and passed to another BiLSTM layer. The output was flattened and passed to a dense layer, after which the original shape was restored. The architecture is shown in Figure 10b. Similarly, the process was repeated with the InSAR data being passed to the sequential model and the CPT data to the dense layer initially (IF2). When including tacheometry data as a predictor, in both configurations, this is passed to the first BiLSTM layer.



(a) The architecture of the Early Fusion model.



(b) The architecture of the Incremental Fusion model.



(c) The architecture of the Late Fusion model.

Figure 10: Architectures of the Early Fusion model (a), Incremental Fusion model (b), and Late Fusion model (c). The numbers in brackets indicate the shape of the output.

3.4 Late Fusion

For late fusion, the models should be fused after the training. This is typically done by aggregating the predictions. Correspondingly, the mean values of the predictions yielded by the models BL3 and BL4 were taken (LF1). The same was done for IF1 and IF2 (LF2).

All models were trained, validated, and tested for the X-, Y-, and Z-directions separately, meaning the corresponding subset of the tacheometry was used. The CPT and InSAR data were passed to the models for every direction equally. The models were compiled using Adam optimization [18] and were trained for a maximum of 100 epochs. If the mean absolute error on the validation data had not improved for 20 epochs, the optimal weights were restored. The models were implemented using Tensorflow Keras [9].

3.5 Evaluation

The applicability of the models is important to the municipality, meaning the models should be accurate and generalizable. Consequently, the test set does not only contain unseen parts of the time-series but also contains unseen racks. To test the generalizability, the data of the unseen racks will be passed to the developed models. Because of the excessiveness of doing this for every model and every configuration, the two best-performing models and the worst-performing model for the two best-performing configurations and the worst-performing configuration were assessed. This way, a reasonable overview of the generalization of the models to unseen racks was obtained.

During the evaluation, the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) were used to measure accuracy. The RMSE is frequently used to assess the quality of a model's prediction and calculates the deviations between predicted and actual values [7]. The formula is the following:

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

Because of the exponential factor in the error function, the metric penalizes large errors. In predicting deformation, large errors are not desired. Therefore, the RMSE is a well-suited metric for this study. The MAE is a good metric to take into account as well because it provides a measure of the absolute magnitude of the error [7]. The formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Additionally, because of varying magnitudes of the tacheometry data, as illustrated in Tables 7, 8, and 9 in Appendix A, a relative error metric could provide valuable information. Ideally, the Mean Absolute Percentage Error (MAPE) would be used. However, this was not suitable because the tacheometry data contains values equalling zero. This would result in infinite values for MAPE. A solution would be to add a neglectable number to every value. Yet the values for MAPE would become extremely high. Therefore, the Mean Absolute Scaled Error (MASE) was used. The MASE is widely used to evaluate forecasts and encompasses the ratio of the MAE of a model's forecast to a naive forecast [16]. The naive forecast, in this case, is the average magnitude of the differences between the actual values in the label window. The formula is depicted below:

$$MASE = \frac{\text{MAE}}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|}$$

The RMSE, MAE, and MASE scores of the models were assessed using paired t-tests to see whether a model's performance was significantly better than the others. The three metrics for the three directions together, made up the population. This was done using a sequential approach. Firstly, if the scores of one model were significantly better than the scores of another model, the scores of the next model were tested for significant differences against that particular model. Secondly, if no significant difference was found, the scores of the next model were compared to the other models for which no significant difference was found. Accordingly, the best-performing models could be identified.

4 RESULTS

The performance of the developed models on the test set is shown in Table 4 on the next page. This concerns the mean scores of all the measuring bolts. The performance on the X-, Y-, and Z-direction is displayed in separate columns, as is the performance for including and excluding tacheometry as a predictor. The different configurations of models, as illustrated in Section 3, are indicated by a separate row. The best scores for every column are highlighted. It is evident from the table that, for the models including tacheometry as a predictor, IF2 and LF2 achieved the best scores. Together with IF1, these are the only models consistently scoring better than the naive baseline. What strikes, for the models excluding tacheometry, is that the scores seem very similar for all configurations. The models trained on only the CPT data seem to have performed best overall. Remarkably, the MASE scores for the models excluding tacheometry are near the best scores obtained when including tacheometry. This is not the case for the RMSE and MAE scores.

Table 11 in Appendix A shows the results of the models on the validation set. Evidently, these scores are all better than on the test set.

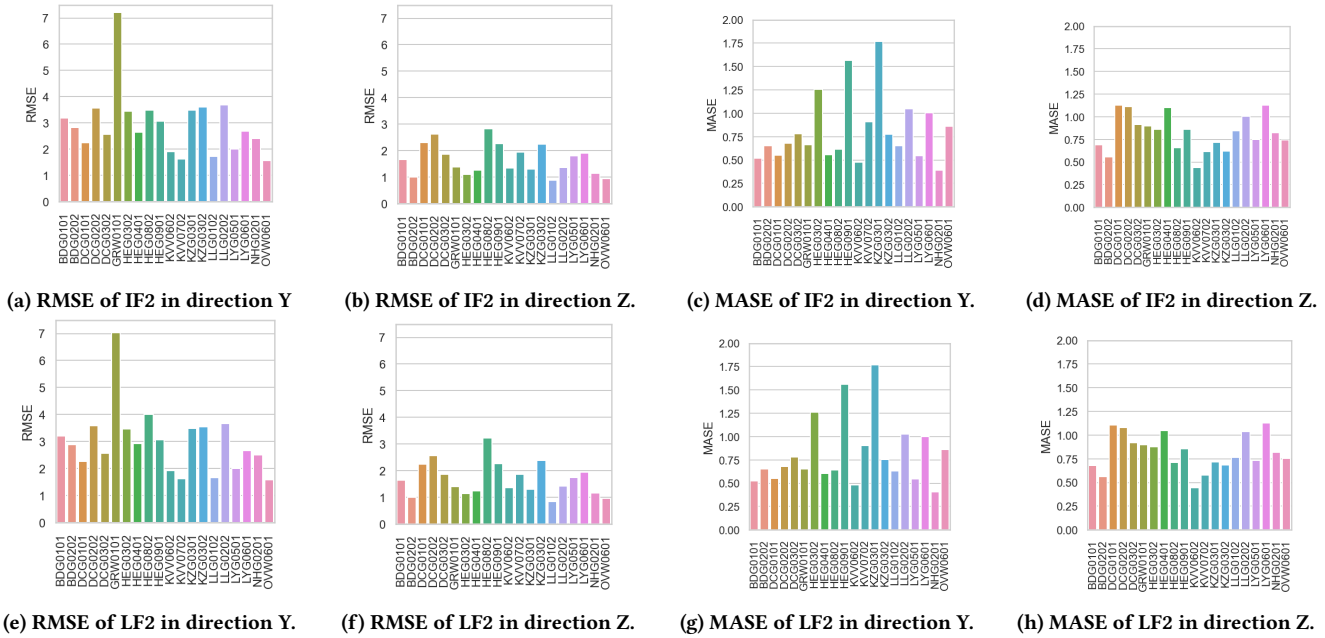
Table 5 on page 9 exhibits the results of the paired t-tests on the scores of the models. The t-tests were only done for the models including the tacheometry data as a predictor. As the population of both sides of the test consisted of 180 scores, the Central Limit Theorem implies the populations to be approximately normally distributed [22]. The scores of the different models are independent. Figures 16 to 42 in Appendix B show that some of the models might contain outliers in their scores. Accordingly, the first two assumptions of the paired t-test are met, but the third might be violated [32]. A negative value of the t-statistic and a p-value smaller than 0.05 indicate that the scores of the configuration on the left side of the test were significantly better than the configuration on the right side of the test. A positive value of the t-statistic and a p-value smaller than 0.05 indicate the opposite [32]. The tests followed the same order as the models listed in Table 4. The table reveals that LF2 and IF2 have significantly better scores than the other models. There is no significant difference between the two configurations. Furthermore, the scores for LF1 and EF1 are only significantly better than BL2 and BL3. The scores obtained by IF1 are significantly better than the early fusion models and all baseline models.

The scores were also averaged per rack. This was only done for the Y- and Z-direction, as the X-direction is less relevant due to the lack of failure mechanisms in this direction [24]. The rack-specific RMSE and MASE of IF2, including tacheometry data as a predictor, are displayed per direction in Figures 11a, 11b, 11c, and 11d. Those of LF2 are shown in Figures 11e, 11f, 11g, and 11h. Strikingly, the figures show only slight differences between the scores for IF2 and LF2. A notable observation is that the RMSE of both IF2 and LF2 for GRW0101 is almost double the second-highest score in the Y-direction. In the Z-direction, the worst RMSE is obtained for HEG0802. For both configurations, the best RMSE is obtained for KVV0702, LLG0102, and OVW0601 in the Y-direction and for LLG01012 in the Z-direction. Evidently, in direction Y, the worst MASE of both IF2 and LF2 is obtained for KZG0301. In the Z-direction, the differences are smaller for the MASE, as DCG0101, DCG0202, HEG0401, LLG0202, and LYG0601 show the worst MASE for both configurations. Both IF2 and LF2 achieved the best MASE for NHG0201 in direction Y and for KVV0602 in the Z-direction. The rack-specific RMSE, MAE, and MASE of all models are shown in Figures 16 to 63 in Appendix B.

The tests for generalization were done using IF2, LF2, and BL2, as these were the best, second-best, and worst configurations, respectively. LLG0102 showed the better RMSE for the Z-direction and one of the best for the Y-direction. KVV0602 achieved the best and second-best MASE for the Z- and Y-direction, respectively. GRW0101 showed the highest RMSE for direction Y, and the obtained MASE in direction Z was among the worst. Consequently, the models trained for these racks were chosen for generalization. The results are displayed in Table 6. For KVV0602 it was impossible to generalize to any other rack because of the large number of measuring bolts used in training. On the contrary, the models trained

Table 4: This table shows the results on the test set. The table is divided into rows corresponding to the configuration of the models, subdivided into the category of fusion or baseline they belong to. The columns show the metrics per direction, subdivided into including tacheometry and excluding tacheometry.

	Including Tacheometry									Excluding Tacheometry								
	X			Y			Z			X			Y			Z		
Model	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
Baseline																		
(BL1) Naive	1.73	1.73	1.14	3.34	3.34	1.27	1.73	1.73	1.14	-	-	-	-	-	-	-	-	-
(BL2) Tacheometry Only	4.54	3.72	2.26	11.28	8.93	2.18	5.03	4.15	2.42	-	-	-	-	-	-	-	-	-
(BL3) CPT	4.50	3.66	2.23	6.89	5.57	1.78	3.75	3.02	1.77	**3.09**	**2.23**	0.86	**4.61**	**3.39**	**0.83**	**3.10**	**2.24**	**0.86**
(BL4) InSAR	3.15	2.57	1.47	4.31	3.51	1.25	3.24	2.66	1.47	3.20	2.31	0.89	4.75	3.48	**0.83**	3.17	2.30	0.87
Early Fusion																		
(EF1) CPT + InSAR	2.69	2.20	1.32	5.31	4.36	1.46	3.40	2.78	1.49	3.15	2.28	0.86	4.74	3.48	**0.83**	3.17	2.29	0.87
Incremental Fusion																		
(IF1) CPT ← InSAR	1.73	1.40	0.83	3.05	2.40	0.82	1.73	1.40	0.83	3.10	**2.23**	**0.85**	4.76	3.49	0.84	3.13	2.26	**0.86**
(IF2) InSAR ← CPT	**1.69**	**1.36**	0.83	**2.95**	**2.33**	**0.81**	**1.66**	**1.34**	**0.82**	3.11	2.42	0.86	4.70	3.44	0.84	3.12	2.25	**0.86**
Late Fusion																		
(LF1) CPT | InSAR	3.47	2.81	1.68	4.92	3.92	1.33	3.13	2.52	1.46	3.12	2.26	0.86	4.66	3.42	**0.83**	3.11	2.25	**0.86**
(LF2) CPT ← InSAR | InSAR ← CPT	**1.69**	1.37	**0.82**	2.99	2.36	**0.81**	1.68	1.36	**0.82**	3.10	**2.23**	**0.85**	4.72	3.46	0.84	3.12	2.25	**0.86**



(a) RMSE of IF2 in direction Y. (b) RMSE of IF2 in direction Z. (c) MASE of IF2 in direction Y. (d) MASE of IF2 in direction Z.

(e) RMSE of LF2 in direction Y. (f) RMSE of LF2 in direction Z. (g) MASE of LF2 in direction Y. (h) MASE of LF2 in direction Z.

Figure 11: The rack-specific scores. The top row exhibits the late fusion of the individual models. The bottom row represents the late fusion of the incrementally fused models. The left half shows the RMSE scores. The MASE scores are shown in the right half. The scores for the Y- and Z-direction are given for the twenty racks.

for GRW0101 have the possibility to generalize to all other racks. Interestingly, the mean RMSE and MAE for the Y-direction are in some cases better than that of the rack-specific models in Table 4. The MASE scores, however, are consistently better. The optimal scores are obtained by IF2 and LF2. The results per rack are shown in Tables 12, 13, and 14 for directions X, Y, and Z, respectively.

5 DISCUSSION

The significantly better performance of the incrementally fused models, and their late fusion, as revealed by Table 5, is in line with previous deformation studies by Xie et al. (2019), Ren et al. (2020), and Chen et al. (2021) as they showed incremental or late fusion to yield the best results [8, 29, 40]. Contrarily, early fusion did not result in better performance than all the baselines. This is discordant with the study of Cao et al. (2020), who concluded early fusion to yield the best results [5].

5.1 Findings

Tables 4 and 5 showed that models conforming to incremental fusion and their late fusion performed better than the other assessed configurations. As these are the most complex models, the results suggest the data used is best handled by more complex models. Early fusion and late fusion using the models trained separately on the CPT and InSAR data, respectively, performed better than the

Table 5: Results of Paired T-Tests TEST. The column Test indicates which two models are compared. The abbreviations are as indicated in Table 4.

Test	t-stat	p-value	power
(BL2) < (BL1)	3.59	<0.001***	0.00
(BL3) < (BL1)	5.67	<0.001***	0.00
(BL3) < (BL2)	-1.99	0.048*	0.63
(BL4) < (BL1)	4.04	<0.001***	0.00
(BL4) < (BL3)	-3.68	<0.001***	0.98
(EF1) < (BL1)	4.43	<0.001***	0.00
(EF1) < (BL4)	0.96	0.34	0.00
(EF1) < (BL3)	-4.45	<0.001***	1.00
(IF1) < (BL1)	-4.95	<0.001***	1.00
(IF2) < (IF1)	-3.34	0.001**	0.95
(LF1) < (IF2)	7.57	<0.001***	0.00
(LF1) < (IF1)	7.58	<0.001***	0.00
(LF1) < (BL1)	5.16	<0.001***	0.00
(LF1) < (EF1)	0.21	0.837	0.32
(LF1) < (BL4)	1.42	0.158	0.00
(LF1) < (BL3)	-5.18	<0.001***	1.00
(LF2) < (LF1)	-7.63	<0.001***	1.00
(LF2) < (IF2)	1.71	0.09	0.00
(LF2) < (IF1)	-4.64	<0.001***	1.00

Note: * p < 0.05, ** p < 0.01, *** p < 0.001

second and third baselines. They did not perform better than the baseline set by the naive method and by the model trained on the InSAR data. This suggests that InSAR data is a better predictor than the CPT data, when combined with the tacheometry data, although it did not outperform a naive baseline.

The performance of the models excluding tacheometry was worse compared to the best-performing models, including tacheometry as a predictor, when looking at the RMSE and the MAE in Table 4. For the MASE, however, the difference was smaller. As the MASE is a scaled error, this indicates the models excluding tacheometry as a predictor to yield bigger actual errors and smaller relative errors. Moreover, the scores of these models were similar for the various configurations. Out of these, surprisingly, the baseline model trained only on static data performed best. Since the naive method performed better than most of the developed models, an explanation could be that the models without tacheometry as a predictor behave more similar to a naive method. This should be further investigated by analyzing the behavior of the models and comparing it to that of the naive method. Moreover, the effect of the inclusion of multiple static predictors in the models should be assessed.

Furthermore, the combination of the two temporal predictors, InSAR and tacheometry, showed the most potential even though it did not outperform the naive baseline. The combination of the static and temporal predictors seems to work best when both temporal predictors were used in an incrementally fused model. It could be further assessed how the incremental fusion of InSAR and tacheometry data, excluding CPT data, would perform. The use of a single temporal predictor, in general, did not demonstrate favorable performance. However, the scores for the models based on only one predictor, static in particular, show a promising relative error.

The results of the tests for generalization in Table 6 showed promising scores for the incrementally fused model in the Y-direction. The RMSE and MAE were better than the scores obtained for the rack-specific models. The MASE was nearly as good. For the baseline configuration trained only on tacheometry, the scores were worse for all metrics in this direction. The fact that predictions could be produced for only four racks by the model based on LLG0102 and a different distribution of values for these racks may have impacted the scores. Furthermore, not all measuring bolts were taken into account to fit the shape of the model. Moreover, the worse MASE suggests a higher relative error. This should be assessed further. Nonetheless, the results in Table 6 could indicate the more complex model configurations to benefit from the tacheometry data of other racks. As the experiment for determining the models to be trained per rack, canal, or all at once was conducted using only the tacheometry data, this was not assessed in this study and should be investigated further.

5.2 Implications

Should the municipality use the models developed for this study, they ought to use the incrementally fused models including tacheometry as a predictor. Their predictions can easily be aggregated to see if their late fusion is more optimal. For racks that were not contained in this study, they should use the code made available to create models. Alternatively, they could use the models trained on fewer measuring bolts to predict that part of the measuring bolts of a new rack. However, as this only includes part of the rack, this is suboptimal. A possibility could be to manipulate the output layer of the model to enable it to predict for the number of measuring bolts of the new rack. The effect of this should be assessed first. When tacheometry measurements are not available for certain racks, the behavior of the models excluding tacheometry as a predictor should be investigated further, before choosing a model.

As the X-direction is deemed to be irrelevant by the municipality because of the absence of failure mechanisms in this direction, it should be considered to stop measuring deformation in this direction. This could save time and expenses in both measuring and registering deformation data.

The predictions yielded by the models can provide value to the Amsterdamse Risicobeoordeling Kademuren (ARK). The current weight of the deformation score - based on the tacheometry data - is 2.0. As the predictions are calculated from the tacheometry data, to prevent circular reasoning, their influence on the score should not be added on top of this weight. Instead, the advice is to share the weight between the predictions and the deformation values. The thresholds for the deformation score should be applied to the predictions to make up a prediction score with a weight of 1.0. The weight of the deformation score would then be set to 1.0 as well. This way, the predictions are an addition to the ARK without substituting anything. The prediction scores for five months after the span of the tacheometry data and the deformation scores of the tacheometry data used in this study are shown in Figures 14 and 15 in Appendix B, respectively. In addition, these plots could be shown on an automatically updating dashboard for easier monitoring.

5.3 Limitations

The tacheometry measurements were relative to the first measurement, causing a measuring or notation error to propagate through

Table 6: This table shows the results for generalizations to racks from the test set. The mean scores for all ten racks are displayed. The model trained on KZG0301 is left out of the table because of its impossibility to generalize to any of the objects.

	LLG0102									GRW0101								
	X			Y			Z			X			Y			Z		
Model	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
(BL2) Tacheometry Only	3.20	2.66	1.50	6.94	6.65	2.25	3.31	2.69	1.52	32.60	26.46	12.49	20.85	15.78	5.82	2.46	2.09	1.39
(IF2) InSAR ← CPT	1.93	**1.63**	0.94	2.74	2.06	**0.85**	1.91	**1.61**	0.93	1.94	1.65	1.09	**1.96**	**1.66**	1.10	2.33	1.98	1.31
(LF2) CPT ← InSAR | InSAR ← CPT	**1.90**	1.65	**0.91**	2.76	2.01	0.86	**1.82**	1.72	**0.92**	2.09	1.78	1.18	3.21	2.64	1.09	2.15	1.83	1.21

the time-series. In the case of notation errors, an improvement could be to induce what the values should have been, based on preceding values. As mentioned in Section 2, the values of this dataset were averaged per month when there were multiple for that month and interpolated for missing months. As a consequence, the actual values of Y were not always the measured values but could be the interpolated values. This may have affected the scores. Its influence should be assessed. Moreover, some of the measuring bolts had fewer measurements than others, as depicted in Figure 4 in Section 2. This may have impacted the outcome.

The InSAR data did not always contain values in both the horizontal and vertical directions for every rack. This was substituted by either the values for the other directions of the rack or the values of a rack along the same canal, as stated in Section 2. The effect of this on the scores should be inspected.

The CPT data contained one value per variable per rack. Consequently, given the development of models per rack, every model only received one value per variable, which prevented the model from learning patterns. This could explain the poor performance of the CPT model, including tacheometry, and it urges the suggestion to assess the performance of an incrementally fused model using InSAR and tacheometry without CPT. The results for models excluding tacheometry as a predictor are contrary, as the CPT model yielded the best performance. It could be that because of the lack of learning in this model it behaves more similar to a naive method, which obtained good results in this study. This should be further assessed.

Some of the models might have contained outliers in their scores, which would violate the third assumption of the paired t-tests. This may have affected the outcome.

The difference between the scores in the validation set and the test set could be due to their size. The validation set contained four windows and the test set one window. Moreover, because the values of the target variable are cumulative and the splits of the data were time-based, the values in the test set were mostly larger. The effect of this should be further assessed.

The architecture of the incrementally fused models leads part of the data through two BiLSTM layers, while the other configurations only contain one BiLSTM layer. It could be that this provided the incrementally fused models with an advantage. This could be investigated by ensuring the other models are passed to two BiLSTM layers as well.

Because of the development per rack, and the differing number of measuring bolts per rack, the models have different input and output shapes. Therefore, only data from racks with at least as many measuring bolts as the rack the model was trained on could be passed to the models. Accordingly, the models were not entirely suitable to test for generalization to unseen racks. Since the tests that were performed did show promising results, the possibility of generalization should be further assessed.

Due to the lack of data, only 30 of the 126 monitored racks were included in this study. The results might be different if more racks could be assessed. When incorporating the data from 2022 and later in training, the results might differ as well.

6 CONCLUSION

This study aimed to fill the research gap concerning the usage of fusion techniques to forecast the deformation of quay walls. This was done by assessing the performance of models conforming to early, incremental, and late model fusion when predicting the deformation of quay walls in Amsterdam based on temporal and static data. Furthermore, the influence of CPT and InSAR data on the performance of the models and the effect of excluding the preceding deformation values as a predictor was judged.

Based on their significantly better RMSE, MAE, and MASE scores, the models conforming to incremental fusion, and their late fusion, performed significantly better than baseline methods, early fusion, and late fusion of baseline methods when including a combination of two temporal predictors and a static predictor. Early fusion and the late fusion of individual models only performed significantly better than three out of four baselines.

The impact of the InSAR data was promising in combination with the preceding deformation values as a predictor even though a naive baseline was not outperformed. The CPT data, in combination with the tacheometry data, did not seem to influence the performance of the models. The combination of the three in a complex configuration performed yielded optimal results.

Excluding the preceding deformation values, the models performed surprisingly well in terms of relative error but not as good as the models including them. There was an even clearer difference for the RMSE and MAE. Remarkably, out of these, the model trained only on CPT data showed the best results, although it might be that this is because the model behaved more similar to a naive method.

It should be taken into account that the interpolation of tacheometry measurements and the substitution of time-series for the InSAR data may have affected the scores of the models. Furthermore, the models received only one value per variable of the CPT data, which may have masked the influence of this predictor. Additionally, the presence of two BiLSTM layers in the architecture of the incrementally fused models may have provided them with an advantage.

The study is in line with Xie et al. (2019), Ren et al. (2020), and Chen et al. (2021) [8, 29, 40] and discordant with Cao et al. (2020) [5]. The paper provided more comprehensive knowledge of the subject, contributing to the work of Dahmen (2022) [10]. The advice for

the municipality is to develop incrementally fused models for new racks based on the available code and to use them to predict the deformation. The thresholds of the deformation score of the ARK should be applied to the predictions to obtain a prediction score. This prediction score should be incorporated in the deformation score by sharing the weight equally.

The findings and limitations of this study call for further research. Several directions have been mentioned in Section 5. The most promising way forward would be to assess the performance of models incrementally fusing the tacheometry data and the InSAR data, as this combination and configuration were most promising. It would also provide more information regarding the effect of the CPT data. Furthermore, the advantage of the extra BiLSTM layer in the architecture of the incrementally fused models should be assessed, by adding an extra BiLSTM layer to the other configurations, for example. Another potential research could be to assess the effect of manipulating the output layer of the models to enable them to predict for the number of measuring bolts of a new rack. This could potentially enhance the generalization of the models to new racks. Moreover, it would be interesting to see the results when incorporating the most recent data. In this case, the effect of a bigger test set should also be investigated. Finally, it might be useful for the municipality to develop a dashboard incorporating the visualizations of the deformation and prediction scores to aid the monitoring.

REFERENCES

[1] M. Kruyswijk A. Ranzijn, M. van der Beek. 2020. Kade ingestort in centrum Amsterdam, situatie 'stabiel'. *Het Parool* (1 9 2020). https://www.parool.nl/amsterdam/kade-ingestort-in-centrum-amsterdam-situatie-stabiel~b2f44086/ visited on 2023-01-26.

[2] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[3] Gemeente Amsterdam. [n. d.]. *Geografische Data.* https://data.amsterdam.nl/data/geozoek/?legenda=true&lagen=hgte-mbz%7Condrgd-mbz&zoom=12

[4] Gemeente Amsterdam. 2022. Amsterdam Inspectie Portaal. Dashboard. https://aip.amsterdam.nl/dashboard.

[5] Enhua Cao, Tengfei Bao, Chongshi Gu, Hui Li, Yongtao Liu, and Shaopei Hu. 2020. A novel hybrid decomposition—ensemble prediction model for dam deformation. *Applied Sciences* 10, 16 (2020), 5700.

[6] Pieter Cawood and Terence van Zyl. 2022. Evaluating State of the Art, Forecasting Ensembles-and Meta-learning Strategies for Model Fusion. *arXiv preprint arXiv:2203.03279* (2022).

[7] Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development* 7, 3 (2014), 1247–1250.

[8] Wenlong Chen, Xiaoling Wang, Zhijian Cai, ChangXin Liu, YuShan Zhu, and Weiwei Lin. 2021. DP-GMM clustering-based ensemble learning prediction methodology for dam deformation considering spatiotemporal differentiation. *Knowledge-Based Systems* 222 (2021), 106964.

[9] Francois Chollet et al. 2015. *Keras.* https://github.com/fchollet/keras

[10] Casper Dahmen. 2022. *Forecasting the deformation of quay walls in Amsterdam.* Master's thesis. Universiteit van Amsterdam, Amsterdam. https://drive.google.com/file/d/12elGCNOj0IzUK2DNNMr_N0pM-cKAcNZA/view?usp=sharing

[11] Mark Denny. 2010. *Super structures: the science of bridges, buildings, dams, and other feats of engineering.* JHU Press.

[12] Bruggen en Kademuren. 2020. *Herstellen en Verbinden. Bouwen aan het Fundament van de Stad. Programmaplan Bruggen en Kademuren.* Technical Report. Gemeente Amsterdam, Amsterdam.

[13] Bruggen en Kademuren. 2022. *Actieplan Bruggen en Kademuren 2023-2026.* Technical Report. Gemeente Amsterdam, Amsterdam.

[14] Maike Grün. 2007. Coordinates and Plans: Geodetic Measurement of Room Installations. *Inside Installations: Theory and Practice in the Care of Complex Artworks* (2007), 185–195.

[15] John D Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering* 9, 3 (2007), 90–95.

[16] Rob J Hyndman and Anne B Koehler. 2006. Another look at measures of forecast accuracy. *International journal of forecasting* 22, 4 (2006), 679–688.

[17] K Jordahl. 2014. GeoPandas: Python tools for geographic data. *URL: https://github.com/geopandas/geopandas* (2014).

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt (Eds.). IOS Press, 87 – 90.

[20] M Korff, MJ Hemel, and R Esposito. 2021. Bezwijken Grimburgwal: Leerpunten voor het Amsterdamse Areaal. (2021).

[21] Mandy Korff, Mart-Jan Hemel, and Dirk Jan Peters. 2022. Collapse of the Grimburgwal, a historic quay in Amsterdam, the Netherlands. *Proceedings of the Institution of Civil Engineers-Forensic Engineering* 175, 4 (2022), 96–105.

[22] Sang Gyu Kwak and Jong Hae Kim. 2017. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology* 70, 2 (2017), 144–156.

[23] Wes McKinney et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. Austin, TX, 51–56.

[24] L. Neijzing and R. Wesstein. 2022. *Amsterdamse Risicobeoordeling: Kademuren: Constructieve staat Amsterdamse kademuren (Kademuur op houten palen).* Technical Report. Programma Bruggen en Kademuren (PBK), Verkeer & Openbare Ruimte (V&OR), Amsterdam. Kenmerk: [20210401 - LNE - ARK].

[25] Het Parool. 2017. Tramverkeer ontregeld door sinkhole Marnixstraat. *Het Parool* (1 11 2017). https://www.parool.nl/nieuws/tramverkeer-ontregeld-door-sinkhole-marnixstraat~b76c2249/ visited on 2023-01-26.

[26] Het Parool. 2022. Zinkgat in kade Lijnbaansgracht, straat afgezet voor onderzoek: 'Het is een diepe leegte, echt krankzinnig'. *Het Parool* (21 9 2022). https://www.parool.nl/amsterdam/zinkgat-in-kade-lijnbaansgracht-straat-afgezet-voor-onderzoek-het-is-een-diepe-leegte-echt-krankzinnig~bada1b2d/ visited on 2023-01-26.

[27] H. Pen. 2018. Geen water door gesprongen leiding Nassaukade. *Het Parool* (3 3 2018). https://www.parool.nl/nieuws/geen-water-door-gesprongen-leiding-nassaukade~b74c1cc1/ visited on 2023-01-26.

[28] Kristina J Reinders, Ramon F Hanssen, Freek J van Leijen, and Mandy Korff. 2021. Augmented satellite InSAR for assessing short-term and long-term surface deformation due to shield tunnelling. *Tunnelling and Underground Space Technology* 110 (2021), 103745.

[29] Qiubing Ren, Mingchao Li, Lingguang Song, and Han Liu. 2020. An optimized combination prediction model for concrete dam deformation considering quantitative evaluation and hysteresis correction. *Advanced Engineering Informatics* 46 (2020), 101154. https://doi.org/10.1016/j.aei.2020.101154

[30] Rijksoverheid. [n. d.]. *Basisregistratie Ondergrondgegevens.* https://www.broloket.nl/ondergrondgegevens

[31] Peter K Robertson. 2009. Interpretation of cone penetration tests—a unified approach. *Canadian geotechnical journal* 46, 11 (2009), 1337–1355.

[32] Amanda Ross, Victor L Willson, Amanda Ross, and Victor L Willson. 2017. Paired samples T-test. *Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures* (2017), 17–19.

[33] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[34] Sensar. 2022. *De AmsterScan: Efficiënte monitoring van kademuren en bruggen (Eindrapportage).* Technical Report. Sensar.

[35] Sensar. 2023. SENSAR. Website. https://www.sensar.nl/?lang=nl.

[36] T. van der Linden. 2023. Geotexxx. https://pypi.org/project/geotexxx/

[37] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA.

[38] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. 2017. *mwaskom/seaborn: v0.8.1 (September 2017).* https://doi.org/10.5281/zenodo.883859

[39] J. Wolthuizen. 2017. Sinkhole in kade Entrepotdok: 'Ik dacht dat we zouden zinken'. *Het Parool* (2017). https://www.parool.nl/nieuws/sinkhole-in-kade-entrepotdok-ik-dacht-dat-we-zouden-zinken~b0ab2e90/ visited on 2023-01-26.

[40] Jiemin Xie, Jun Zhang, Xuan Xie, Zhiwei Bi, and Zhuoheng Li. 2019. Ensemble of bagged regression trees for concrete dam deformation predicting. In *IOP Conference Series: Earth and Environmental Science*, Vol. 376. IOP Publishing, 012040.

[41] Howard A Zebker, Scott Hensley, Piyush Shanker, and Cody Wortham. 2010. Geodetically accurate InSAR data processor. *IEEE Transactions on Geoscience and Remote Sensing* 48, 12 (2010), 4309–4321.

Appendix A SUPPLEMENTARY TABLES

Table 7: Summary of values in X-direction

Index	Count	Mean	Std	Min	25%	50%	75%	Max
BDG0101	34	-2.45	2.56	-10.0	-4.10	-2.50	-0.90	7.8
BDG0202	52	0.33	1.54	-4.0	-0.60	0.20	1.30	6.1
DCG0101	36	1.25	2.16	-5.2	0.10	1.20	2.60	6.8
DCG0202	30	-0.42	2.31	-7.8	-1.80	-0.40	0.98	6.4
DCG0302	44	-1.04	2.13	-7.1	-2.40	-1.10	0.20	5.8
GRW0101	22	0.16	2.32	-7.4	-1.00	0.10	1.40	9.1
HEG0302	52	0.10	1.01	-4.5	-0.40	0.10	0.60	3.9
HEG0401	33	-0.47	1.13	-4.7	-1.00	-0.40	0.20	3.3
HEG0802	87	-0.33	3.81	-9.4	-2.60	-0.70	1.00	19.1
HEG0901	58	-0.51	2.49	-11.8	-1.92	-0.40	0.80	11.3
KVV0602	96	0.63	2.60	-9.0	-0.70	0.50	2.00	14.0
KVV0702	185	-0.30	3.09	-13.5	-1.60	-0.30	1.30	9.8
KZG0301	102	0.56	1.40	-3.8	-0.10	0.60	1.30	4.4
KZG0302	46	-1.15	2.68	-10.2	-2.40	-0.70	0.20	11.8
LLG0102	46	-0.21	1.20	-4.3	-0.60	0.00	0.30	3.6
LLG0202	40	-0.24	1.38	-4.3	-1.20	-0.30	0.60	4.0
LYG0501	40	0.68	2.24	-6.2	-0.60	0.80	2.10	7.2
LYG0601	26	-0.49	1.39	-5.9	-1.10	-0.40	0.20	2.6
NHG0201	89	1.06	1.30	-3.3	0.10	0.90	1.80	5.9
OVW0601	25	-0.04	1.27	-4.8	-0.60	0.15	0.67	2.6
PRG0301	43	-0.67	2.05	-11.5	-2.00	-0.40	0.63	7.8
PRG0401	54	-0.65	1.44	-5.0	-1.43	-0.40	0.30	4.4
PRG0402	49	0.79	1.78	-6.1	-0.20	1.00	2.00	6.9
SIN0501	40	-0.79	1.60	-7.1	-1.50	-0.60	0.25	3.6
SIN0502	36	0.85	2.66	-9.3	-0.70	0.80	2.90	6.9
SIN0601	37	-1.57	1.64	-6.9	-2.70	-1.70	-0.60	3.9
SIN0602	38	0.86	1.65	-3.1	-0.20	0.70	1.90	6.0
SIN0701	79	0.31	1.64	-5.2	-0.70	0.40	1.50	5.3
WEG0201	38	0.04	1.66	-4.4	-1.00	0.00	1.20	9.1
WKN0101	66	-0.21	2.23	-7.2	-1.60	-0.10	1.28	8.4

Table 8: Summary of values in Y-direction

Object ID	Count	Mean	Std	Min	25%	50%	75%	Max
BDG0101	34	-0.02	4.20	-13.6	-2.00	0.00	1.80	14.1
BDG0202	52	0.90	2.53	-11.7	-0.30	0.90	2.00	8.0
DCG0101	36	-0.68	3.15	-9.9	-2.52	-1.15	1.23	9.0
DCG0202	30	-1.48	3.23	-20.7	-3.10	-0.90	0.80	6.6
DCG0302	44	-1.11	3.24	-13.9	-2.60	-1.10	0.65	9.3
GRW0101	22	-2.21	6.60	-25.0	-5.40	-2.15	0.23	16.7
HEG0302	52	-0.16	2.07	-8.3	-0.60	0.10	0.60	4.7
HEG0401	33	-1.33	3.03	-13.1	-2.40	-0.30	0.30	4.1
HEG0802	87	-1.05	5.02	-33.2	-0.90	0.10	1.10	11.8
HEG0901	58	-0.61	2.97	-11.2	-2.00	-0.80	0.40	20.1
KVV0602	96	0.12	3.01	-26.3	-0.80	0.30	1.50	7.4
KVV0702	185	-0.48	2.46	-14.9	-1.20	-0.10	0.80	5.3
KZG0301	102	-2.97	2.20	-9.5	-4.50	-3.20	-1.30	2.6
KZG0302	46	1.04	3.05	-16.1	-0.30	0.40	2.62	11.3
LLG0102	46	-0.50	1.62	-9.0	-1.30	-0.20	0.40	3.4
LLG0202	40	-1.62	4.61	-25.0	-2.10	-0.30	0.90	4.9
LYG0501	40	0.04	2.74	-7.3	-1.53	0.50	1.60	8.8
LYG0601	26	0.44	2.37	-8.4	-0.60	0.20	1.90	9.3
NHG0201	89	-1.34	3.27	-17.4	-3.05	-1.00	0.80	9.3
OVW0601	25	-0.17	2.01	-6.1	-1.50	0.00	1.40	4.1
PRG0301	43	-2.46	5.99	-43.1	-4.52	-1.55	0.50	6.7
PRG0401	54	-0.84	1.82	-5.1	-2.12	-1.10	0.52	3.7
PRG0402	49	-0.23	2.67	-14.6	-1.38	0.00	1.30	5.7
SIN0501	40	1.06	2.95	-12.1	-1.10	1.20	3.20	8.6
SIN0502	36	-1.39	2.27	-8.7	-3.00	-1.20	0.10	4.4
SIN0601	37	-0.83	2.17	-18.3	-1.90	-0.60	0.50	4.6
SIN0602	38	-0.61	1.66	-6.1	-1.60	-0.50	0.50	3.3
SIN0701	79	-1.71	5.59	-30.0	-3.70	-0.40	1.70	9.8
WEG0201	38	-3.45	5.94	-51.9	-6.00	-2.50	0.10	8.1
WKN0101	66	0.70	1.81	-9.2	-0.30	0.55	1.70	7.8

Table 9: Summary of values in Z-direction

Object ID	Count	Mean	Std	Min	25%	50%	75%	Max
BDG0101	34	-0.54	2.18	-5.7	-2.40	-0.50	1.20	6.9
BDG0202	52	-0.27	1.19	-2.8	-1.10	-0.30	0.40	4.3
DCG0101	36	-2.88	2.20	-7.1	-4.90	-2.80	-1.05	1.6
DCG0202	30	-0.71	5.84	-32.2	-0.90	0.20	1.10	9.7
DCG0302	44	-1.10	3.37	-6.4	-2.10	-1.30	-0.50	22.2
GRW0101	22	0.12	1.65	-4.4	-0.60	0.20	1.30	3.6
HEG0302	52	-0.55	3.59	-17.4	-2.08	-0.60	1.10	17.6
HEG0401	33	0.15	1.30	-3.9	-0.30	0.00	0.40	5.7
HEG0802	87	-0.80	1.78	-7.4	-1.50	-0.60	0.10	4.6
HEG0901	58	2.30	4.01	-10.5	0.40	3.00	4.80	12.8
KVV0602	96	-0.48	1.65	-8.6	-1.70	-0.60	0.60	8.6
KVV0702	185	0.14	2.18	-5.0	-1.10	0.00	1.10	15.5
KZG0301	102	0.25	0.98	-2.9	-0.40	0.30	0.80	3.1
KZG0302	46	-2.32	4.21	-21.9	-2.20	-0.70	0.00	1.8
LLG0102	46	-2.02	2.60	-9.5	-3.53	-1.05	0.00	0.9
LLG0202	40	-0.73	3.11	-15.0	-1.60	-0.10	1.30	3.6
LYG0501	40	-2.20	1.52	-6.9	-3.23	-2.00	-1.20	1.2
LYG0601	26	-0.45	1.46	-4.6	-1.10	-0.20	0.40	3.1
NHG0201	89	-0.21	1.76	-7.6	-1.20	-0.10	0.80	5.2
OVW0601	25	-0.34	1.78	-7.9	-1.10	0.00	0.90	3.1
PRG0301	43	-0.91	3.45	-30.7	-1.20	-0.20	0.30	4.2
PRG0401	54	-0.02	1.06	-2.1	-0.90	-0.20	0.90	2.8
PRG0402	49	-5.54	5.33	-29.4	-7.40	-3.80	-2.30	2.5
SIN0501	40	-1.98	1.63	-13.6	-2.65	-1.70	-1.00	0.8
SIN0502	36	0.06	1.71	-5.3	-0.90	0.00	0.80	8.3
SIN0601	37	-0.88	2.12	-14.1	-1.50	-0.60	0.20	4.2
SIN0602	38	0.37	1.94	-3.2	-0.70	0.00	1.00	10.9
SIN0701	79	-3.63	3.36	-20.8	-4.70	-2.80	-1.50	1.7
WEG0201	38	-1.08	4.71	-26.6	-1.50	0.00	1.40	6.2
WKN0101	66	-0.53	1.23	-3.9	-1.30	-0.70	0.10	3.2

Table 10: This is a table showing the results of the grouping experiment

	X			**Y**			**Z**		
Grouped By	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
Rack	**1.11**	**0.79**	**0.46**	**1.24**	**0.91**	**0.40**	**1.09**	**0.78**	**0.46**
Canal	1.39	1.02	0.60	1.83	1.35	0.56	1.41	1.04	0.61
All	1.70	1.26	0.60	2.01	1.50	0.69	1.65	1.23	0.59

Table 11: This table shows the results on the validation set set. The table is divided into rows corresponding to the configuration of the models, subdivided into the category of fusion or baseline they belong to. The columns show the metrics per direction, subdivided into including tacheometry and excluding tacheometry.

	Including Tacheometry									Excluding Tacheometry								
	X			Y			Z			X			Y			Z		
Model	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
Baseline																		
(BL1) Naive	1.25	1.19	0.85	1.78	1.70	1.02	1.25	1.19	0.85	-	-	-	-	-	-	-	-	-
(BL2) Tacheometry Only	1.11	0.79	0.46	1.24	0.91	0.40	1.09	0.78	0.46	-	-	-	-	-	-	-	-	-
(BL3) CPT	1.11	0.80	0.47	1.22	0.89	0.39	1.09	0.78	0.46	1.20	0.88	0.51	1.36	0.98	0.42	1.20	0.87	0.51
(BL4) InSAR	1.06	0.75	0.45	1.15	0.83	0.37	1.07	0.77	0.45	1.10	0.78	0.48	1.16	0.83	0.36	1.10	0.79	0.48
Early Fusion																		
(EF1) CPT + InSAR	1.03	0.74	0.44	1.09	0.78	0.34	1.03	0.74	0.44	1.10	0.79	0.47	1.16	0.83	0.36	**1.08**	**0.77**	**0.46**
Incremental Fusion																		
(IF1) CPT ← InSAR	1.06	0.75	0.44	1.06	0.75	0.33	1.03	0.73	0.43	1.11	0.80	0.47	1.13	0.81	**0.35**	1.12	0.80	0.47
(IF2) InSAR ← CPT	1.03	0.73	0.43	1.05	0.75	**0.32**	1.05	0.74	0.43	1.14	0.82	0.48	1.25	0.91	0.39	1.14	0.81	0.47
Late Fusion																		
(LF1) CPT | InSAR	**0.99**	**0.69**	**0.41**	1.07	0.75	0.33	**0.99**	**0.69**	**0.41**	1.10	0.80	0.48	1.20	0.88	0.38	1.11	0.81	0.48
(LF2) CPT ← InSAR | InSAR ← CPT	1.04	0.73	0.43	**1.04**	**0.73**	**0.32**	1.03	0.72	0.42	**1.09**	**0.77**	**0.45**	**1.12**	**0.80**	**0.35**	1.11	0.78	**0.46**

Table 12: This table shows the results of the test for generalization in direction X. The scores for the predictions of the racks per model are shown in every column. At the bottom, the mean of the scores for all racks is given per model. The best scores per rack are highlighted, as well as the best overall score per metric..

	(IF2) InSAR ← CPT									(LF2) CPT ← InSAR | InSAR ← CPT									(BL2) Tacheometry Only								
	LLG0201			KVV0602			GRW0101			LLG0201			KVV0602			GRW0101			LLG0201			KVV0602			GRW0101		
Rack	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
PRG0301	-	-	-	-	-	-	**1.66**	**1.40**	**0.67**	-	-	-	-	-	-	1.83	1.57	0.76	-	-	-	-	-	-	9.20	7.53	3.89
PRG0401	1.69	1.37	1.07	-	-	-	**1.38**	**1.14**	1.01	1.65	1.33	1.03	-	-	-	1.52	1.29	1.13	1.65	1.35	1.05	-	-	-	1.39	**1.14**	**0.72**
PRG0402	2.03	1.71	0.88	-	-	-	1.80	1.62	**0.63**	**1.65**	**1.35**	1.05	-	-	-	1.91	1.67	0.66	4.04	3.40	1.80	-	-	-	2.11	1.85	0.72
SIN0501	-	-	-	-	-	-	**1.37**	**1.12**	**1.09**	-	-	-	-	-	-	1.50	1.22	1.26	-	-	-	-	-	-	1.85	1.51	1.66
SIN0502	-	-	-	-	-	-	**3.49**	**3.00**	**1.27**	-	-	-	-	-	-	3.72	3.21	1.37	-	-	-	-	-	-	3.78	3.27	1.39
SIN0601	-	-	-	-	-	-	**2.37**	**2.11**	**2.27**	-	-	-	-	-	-	2.44	2.18	2.31	-	-	-	-	-	-	4.48	3.76	3.96
SIN0602	-	-	-	-	-	-	**1.77**	**1.47**	**1.41**	-	-	-	-	-	-	1.85	1.54	1.49	-	-	-	-	-	-	2.06	1.71	1.69
SIN0701	1.75	1.53	1.04	-	-	-	**1.54**	**1.31**	1.04	1.73	1.51	**1.03**	-	-	-	1.64	1.38	1.11	2.70	2.27	1.65	-	-	-	1.73	1.43	1.16
WEG0201	-	-	-	-	-	-	**1.75**	**1.47**	**0.65**	-	-	-	-	-	-	1.93	1.65	0.73	-	-	-	-	-	-	11.72	9.45	4.52
WKN0101	2.26	1.91	0.78	-	-	-	2.30	1.89	0.84	**2.24**	**1.88**	**0.77**	-	-	-	2.60	2.13	0.95	4.39	3.60	1.51	-	-	-	32.60	26.46	12.59
Overall	1.93	1.63	0.94	-	-	-	1.94	1.65	1.09	**1.90**	**1.60**	**0.91**	-	-	-	2.09	1.78	1.18	3.20	2.66	1.50	-	-	-	32.60	26.46	12.59

Table 13: This table shows the results of the test for generalization in direction Y. The scores for the predictions of the racks per model are shown in every column. At the bottom, the mean of the scores for all racks is given per model. The best scores per rack are highlighted, as well as the best overall score per metric.

	(IF2) InSAR ← CPT									(LF2) CPT ← InSAR | InSAR ← CPT									(BL2) Tacheometry Only								
	LLG0201			KVV0602			GRW0101			LLG0201			KVV0602			GRW0101			LLG0201			KVV0602			GRW0101		
Rack	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
PRG0301	-	-	-	-	-	-	4.71	3.97	**1.23**	-	-	-	-	-	-	**4.70**	**3.96**	**1.23**	-	-	-	-	-	-	20.60	15.90	5.64
PRG0401	**1.69**	**1.43**	**0.84**	-	-	-	2.78	2.29	1.38	1.72	1.45	0.85	-	-	-	2.36	1.98	1.20	1.91	1.61	0.95	-	-	-	2.93	2.35	1.41
PRG0402	**2.75**	**1.43**	**0.78**	-	-	-	3.28	2.30	1.09	2.80	2.18	0.80	-	-	-	3.21	1.99	1.03	8.78	7.05	2.94	-	-	-	5.38	4.01	1.87
SIN0501	-	-	-	-	-	-	3.27	2.69	0.89	-	-	-	-	-	-	**3.18**	**2.64**	**0.86**	-	-	-	-	-	-	4.60	3.82	1.33
SIN0502	-	-	-	-	-	-	3.31	2.83	0.98	-	-	-	-	-	-	**3.02**	**2.70**	**0.92**	-	-	-	-	-	-	55.01	41.25	15.16
SIN0601	-	-	-	-	-	-	2.98	2.22	1.25	-	-	-	-	-	-	**2.59**	**1.99**	**1.10**	-	-	-	-	-	-	3.72	2.75	1.54
SIN0602	-	-	-	-	-	-	**1.57**	**1.29**	**1.06**	-	-	-	-	-	-	1.69	1.32	1.10	-	-	-	-	-	-	10.27	7.77	6.06
SIN0701	4.34	3.57	**0.75**	-	-	-	**3.44**	2.97	0.92	4.35	**2.57**	**0.75**	-	-	-	3.38	2.91	0.89	11.76	9.93	2.33	-	-	-	21.84	16.64	5.36
WEG0201	-	-	-	-	-	-	**5.06**	**4.32**	**1.33**	-	-	-	-	-	-	5.18	4.47	1.37	-	-	-	-	-	-	79.50	59.65	17.82
WKN0101	**2.16**	**1.81**	**1.04**	-	-	-	2.62	2.16	1.11	2.18	1.83	1.05	-	-	-	2.74	2.23	1.17	5.30	4.47	2.76	-	-	-	4.67	3.67	2.00
Overall	**2.74**	2.06	**0.85**	-	-	-	1.96	1.66	1.10	2.76	**2.01**	0.86	-	-	-	3.21	2.64	1.09	6.94	6.65	2.25	-	-	-	20.85	15.78	5.82

Table 14: This table shows the results of the test for generalization in direction Z. The scores for the predictions of the racks per model are shown in every column. At the bottom, the mean of the scores for all racks is given per model. The best scores per rack are highlighted, as well as the best overall score per metric.

	(IF2) InSAR ← CPT									(LF2) CPT ← InSAR | InSAR ← CPT									(BL2) Tacheometry Only								
	LLG0201			KVV0602			GRW0101			LLG0201			KVV0602			GRW0101			LLG0201			KVV0602			GRW0101		
Rack	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
PRG0301	-	-	-	-	-	-	**1.64**	**1.37**	**0.65**	-	-	-	-	-	-	1.75	1.48	0.71	-	-	-	-	-	-	1.65	1.38	0.66
PRG0401	1.66	1.34	1.04	-	-	-	**1.38**	**1.14**	1.01	1.62	1.29	**1.00**	-	-	-	1.52	1.29	1.13	1.76	1.42	1.10	-	-	-	1.37	1.14	1.01
PRG0402	1.99	1.67	0.86	-	-	-	1.76	**1.58**	**0.62**	**1.69**	2.18	0.87	-	-	-	1.98	1.75	0.69	4.16	3.39	1.77	-	-	-	1.93	2.19	0.65
SIN0501	-	-	-	-	-	-	**1.38**	**1.11**	**1.10**	-	-	-	-	-	-	1.52	1.23	1.28	-	-	-	-	-	-	1.89	1.54	1.67
SIN0502	-	-	-	-	-	-	**3.43**	**2.94**	**1.24**	-	-	-	-	-	-	3.73	3.21	1.38	-	-	-	-	-	-	6.69	5.70	2.54
SIN0601	-	-	-	-	-	-	**2.36**	**2.08**	**2.25**	-	-	-	-	-	-	2.41	2.16	2.28	-	-	-	-	-	-	2.57	2.22	2.38
SIN0602	-	-	-	-	-	-	**1.85**	**1.53**	**1.48**	-	-	-	-	-	-	2.14	1.76	1.71	-	-	-	-	-	-	2.63	2.21	2.26
SIN0701	1.72	1.50	**1.02**	-	-	-	1.55	1.31	1.05	**1.71**	1.50	**1.02**	-	-	-	1.70	**1.41**	1.15	2.94	2.48	1.73	-	-	-	1.64	1.43	1.14
WEG0201	-	-	-	-	-	-	**1.77**	**1.50**	**0.66**	-	-	-	-	-	-	2.00	1.73	0.76	-	-	-	-	-	-	1.88	1.58	0.70
WKN0101	2.27	1.91	**0.78**	-	-	-	2.45	2.01	0.90	**2.24**	**1.89**	**0.78**	-	-	-	2.71	2.22	1.00	4.36	3.48	1.48	-	-	-	2.37	1.97	0.90
Overall	1.91	1.61	0.93	-	-	-	2.33	1.98	1.31	**1.82**	1.72	**0.92**	-	-	-	2.15	1.83	1.21	3.31	2.69	**1.52**	-	-	-	2.46	2.09	1.39

Appendix B SUPPLEMENTARY FIGURES

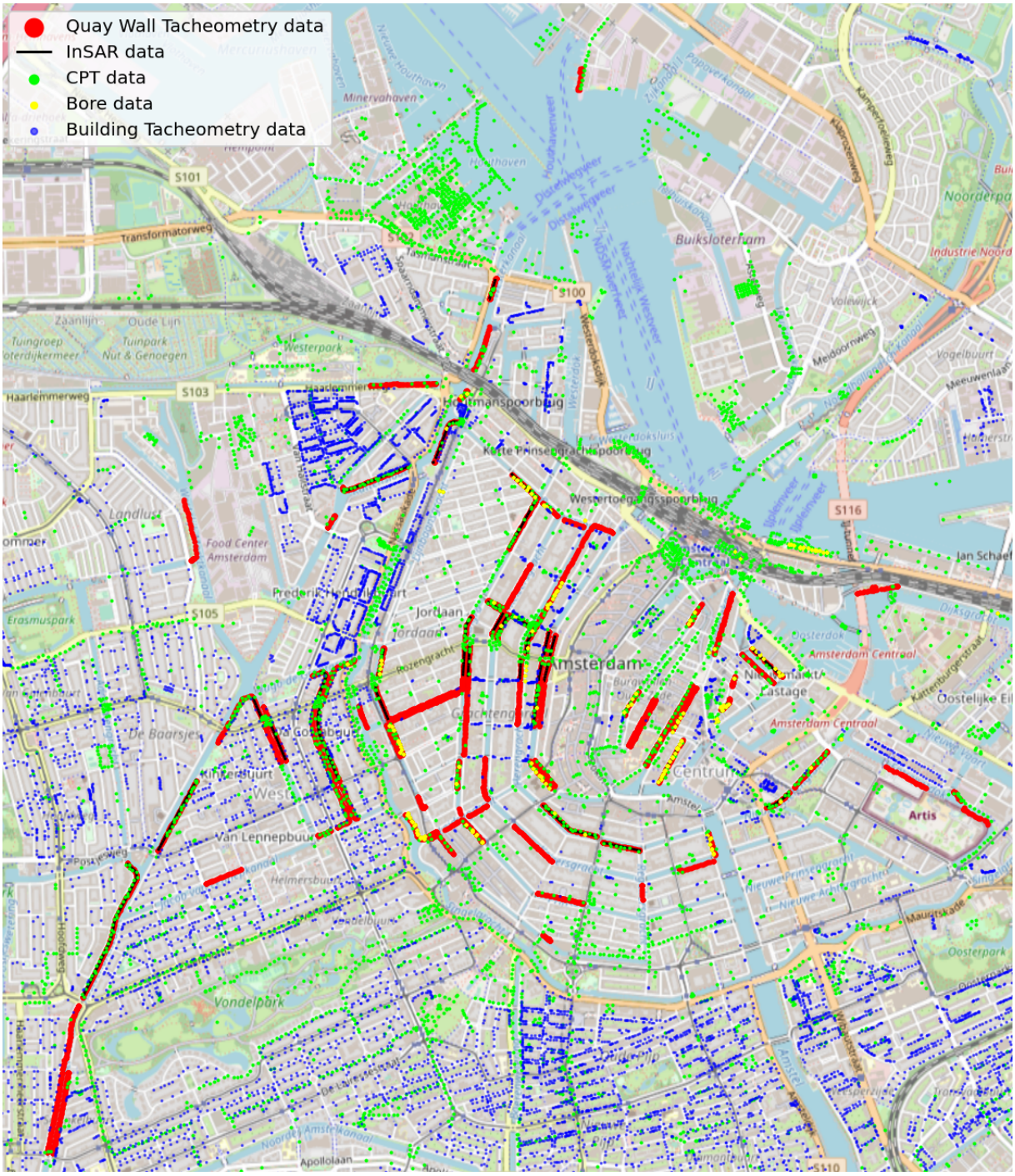


Figure 12: Map showing the spread of datapoints from all datasets available. The InSAR data was provided for the racks that had tacheometry data and CPT datapoints within 2m of their polygon. Consequently, the black lines indicate the racks assessed in this study.

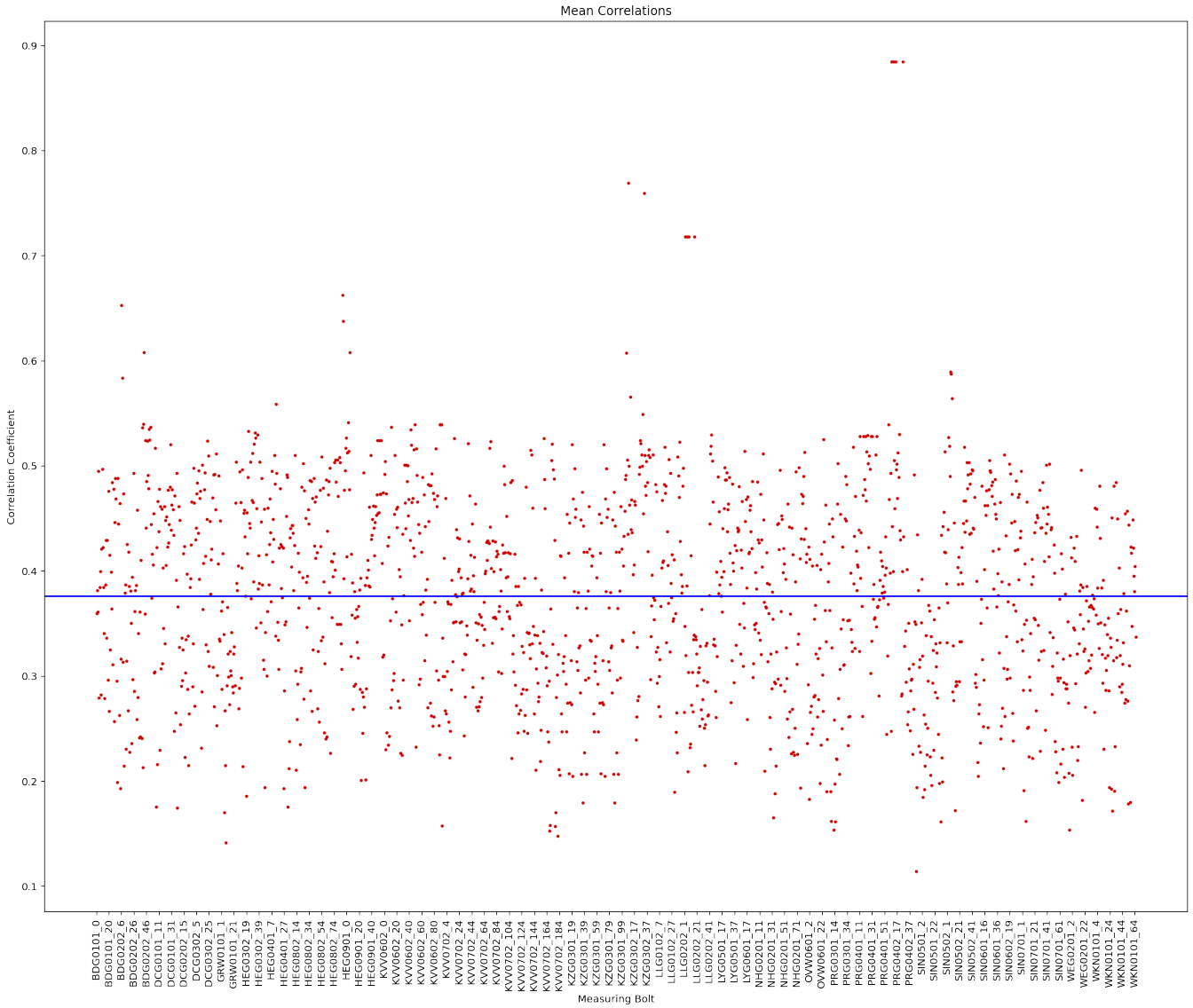


Figure 13: This figure shows for every measuring bolt the mean value of Pearson's R with every measuring bolt. The blue line indicates the overall mean.

Figure 14: Prediction scores of five months after the tacheometry data of the racks assessed in this study.

Figure 15: Deformation scores of the racks assessed in this study.

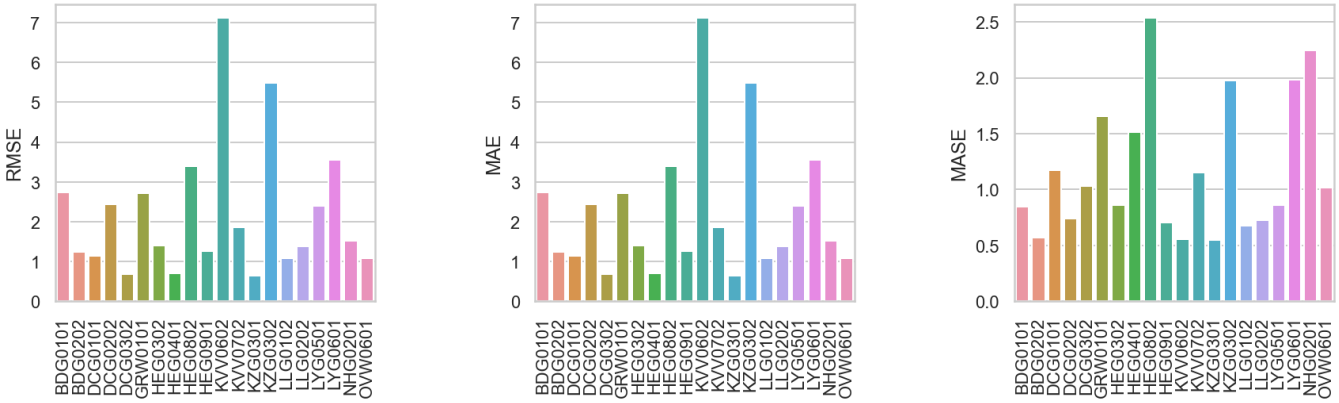


Figure 16: Metrics of BL1 in X-direction.

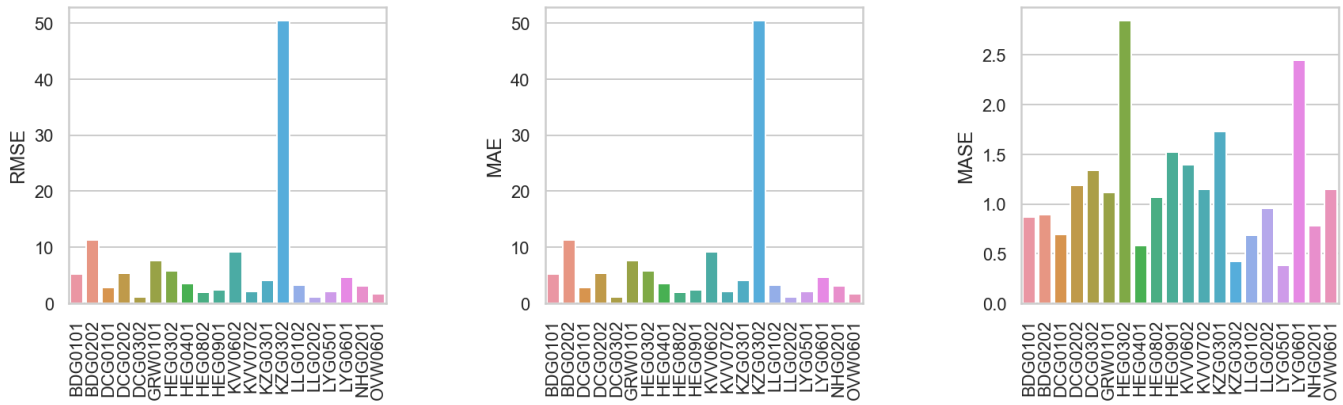


Figure 17: Metrics of BL1 in Y-direction.

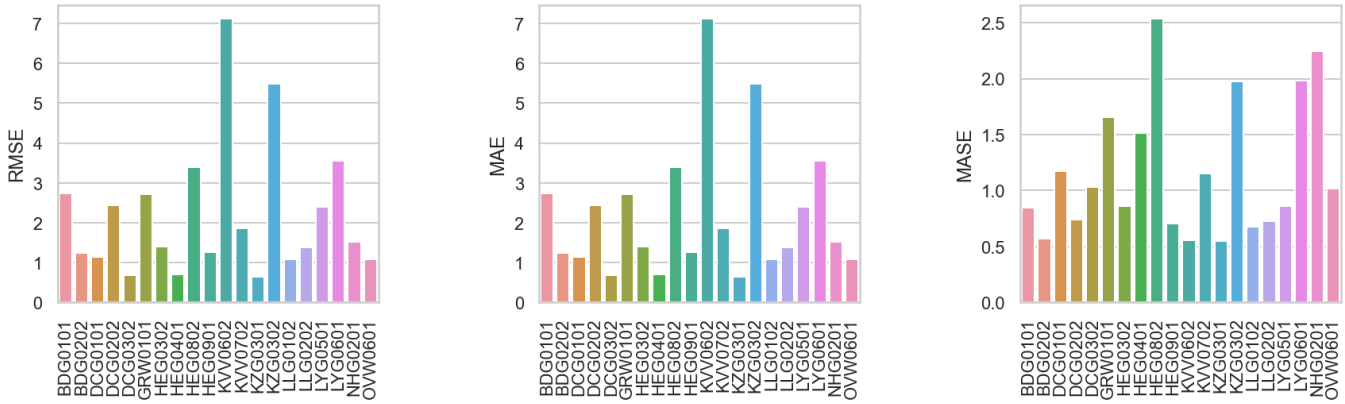


Figure 18: Metrics of BL1 in Z-direction.

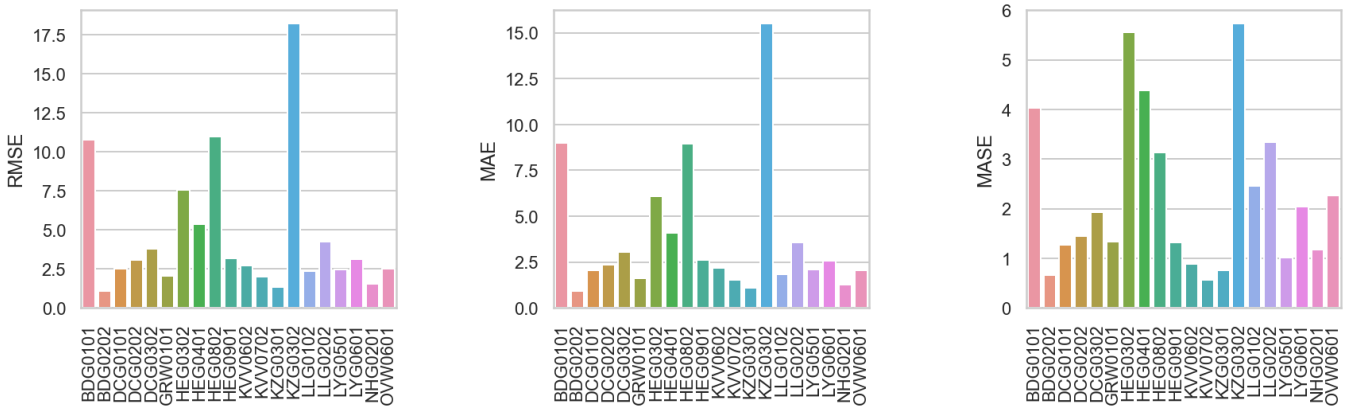


Figure 19: Metrics of BL2 in X-direction.

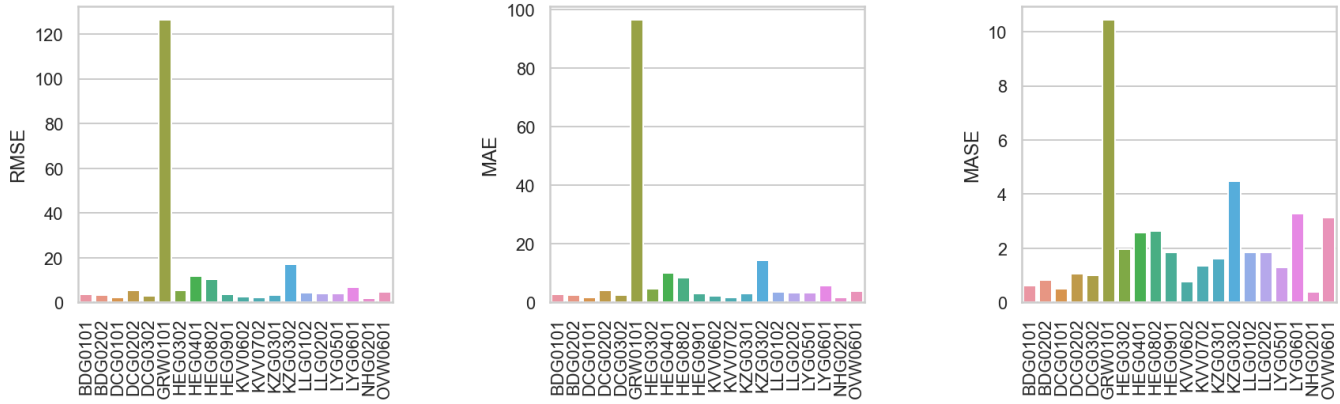


Figure 20: Metrics of BL2 in Y-direction.

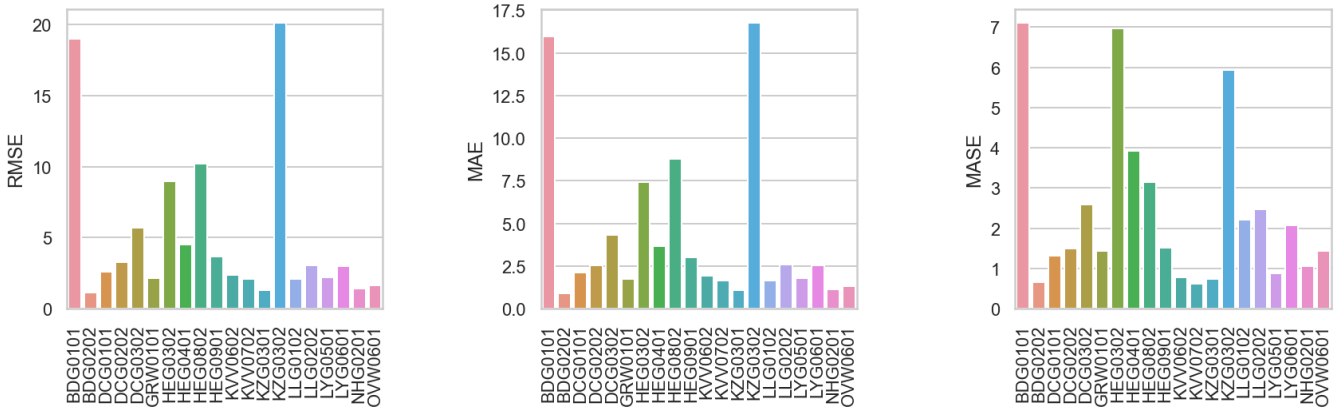


Figure 21: Metrics of BL2 in Z-direction.

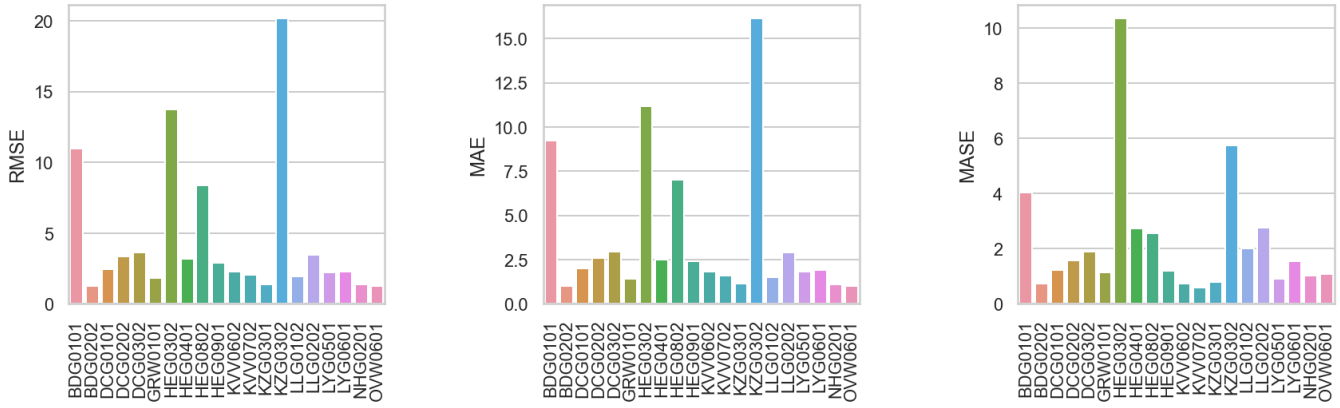


Figure 22: Metrics of BL3 with tacheometry as predictor in X-direction.

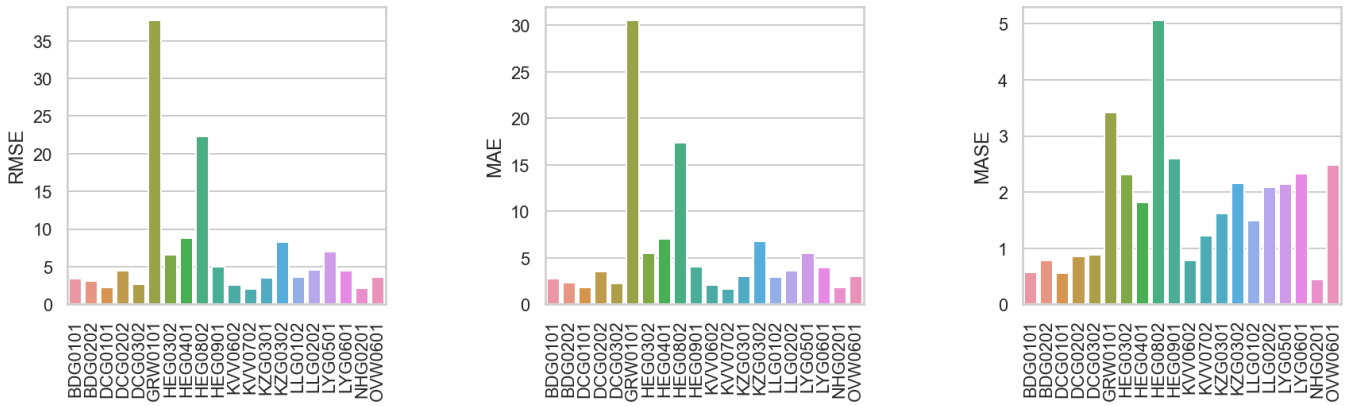


Figure 23: Metrics of BL3 with tacheometry as predictor in Y-direction.

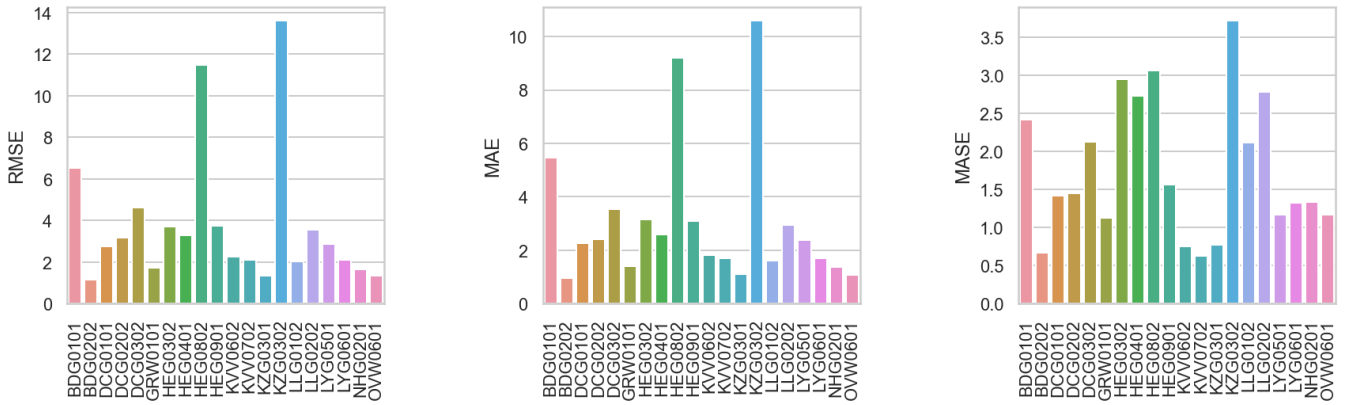


Figure 24: Metrics of BL3 with tacheometry as predictor in Z-direction.

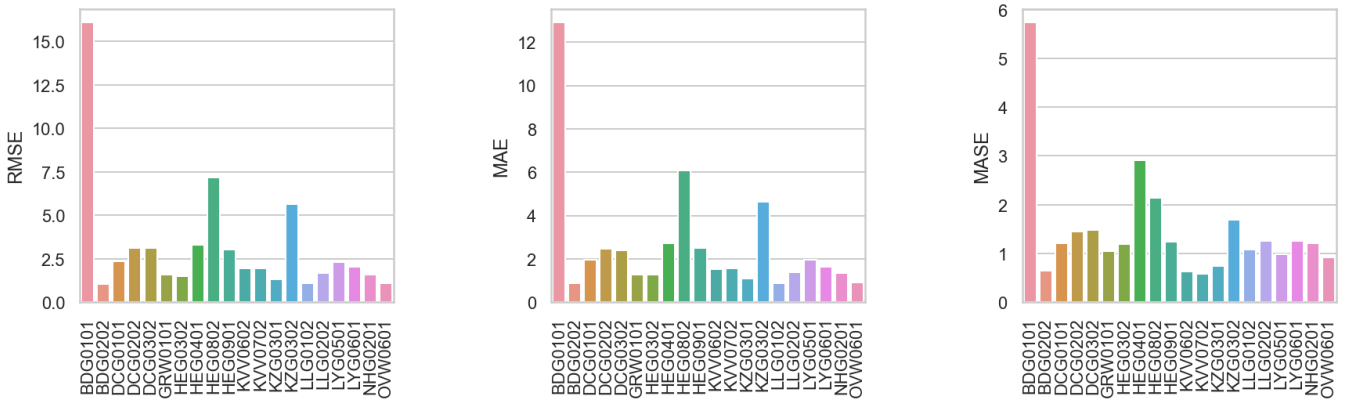


Figure 25: Metrics of BL4 with tacheometry as predictor in X-direction.

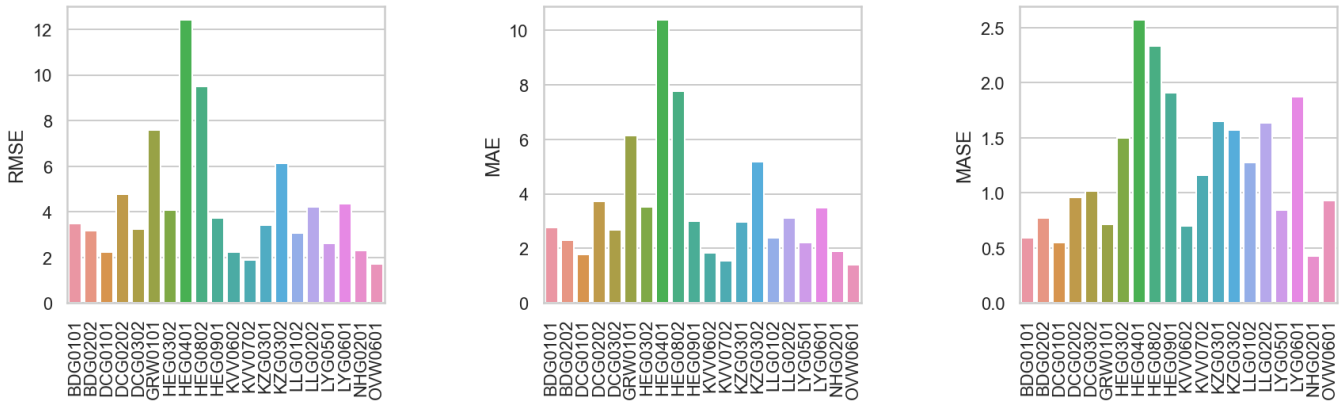


Figure 26: Metrics of BL4 with tacheometry as predictor in Y-direction.

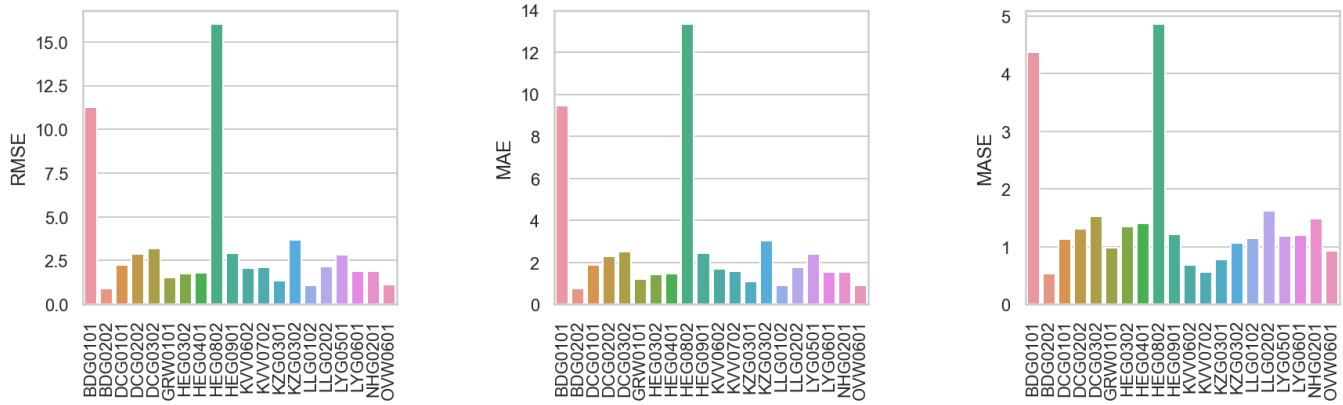


Figure 27: Metrics of BL4 with tacheometry as predictor in Z-direction.

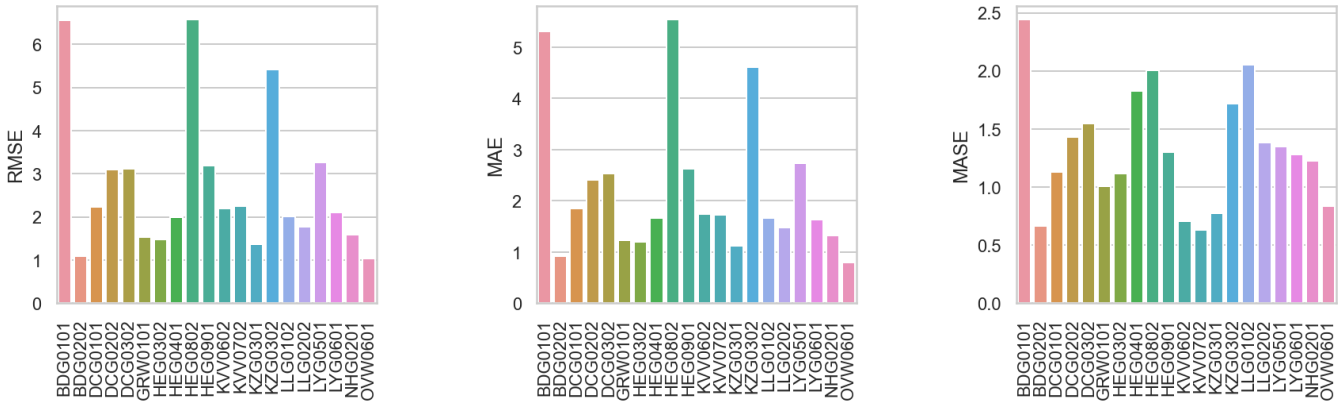


Figure 28: Metrics of EF1 with tacheometry as predictor in X-direction.

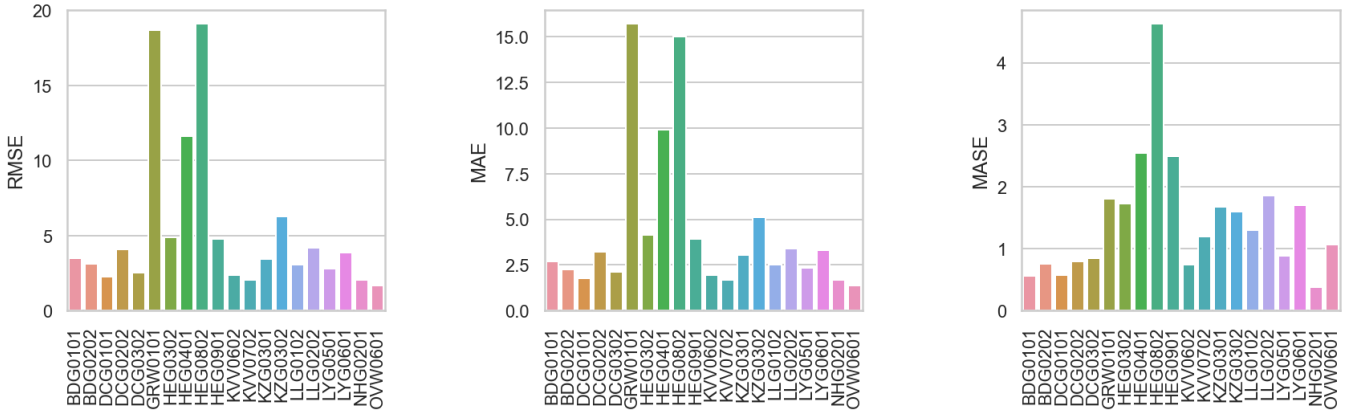


Figure 29: Metrics of EF1 with tacheometry as predictor in Y-direction.

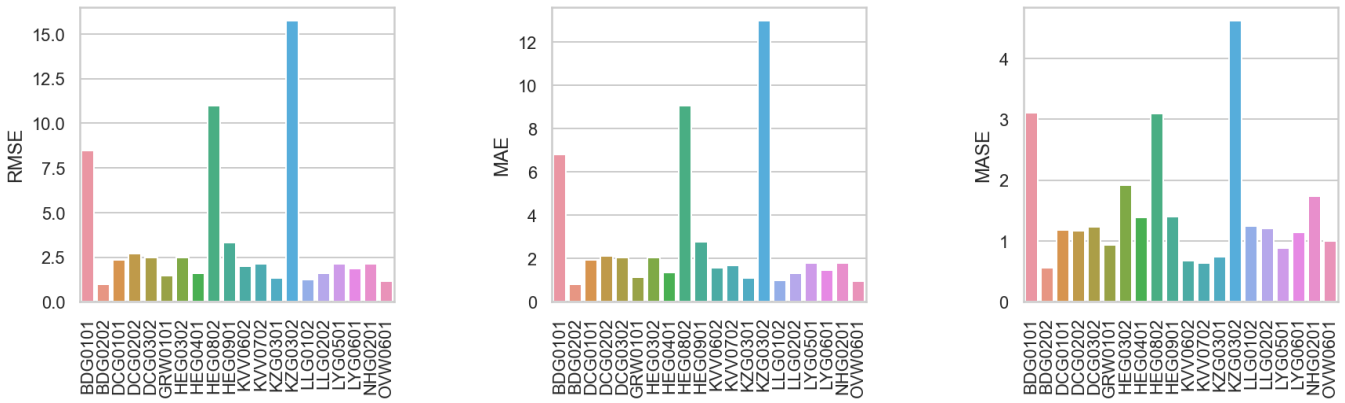


Figure 30: Metrics of EF1 with tacheometry as predictor in Z-direction.

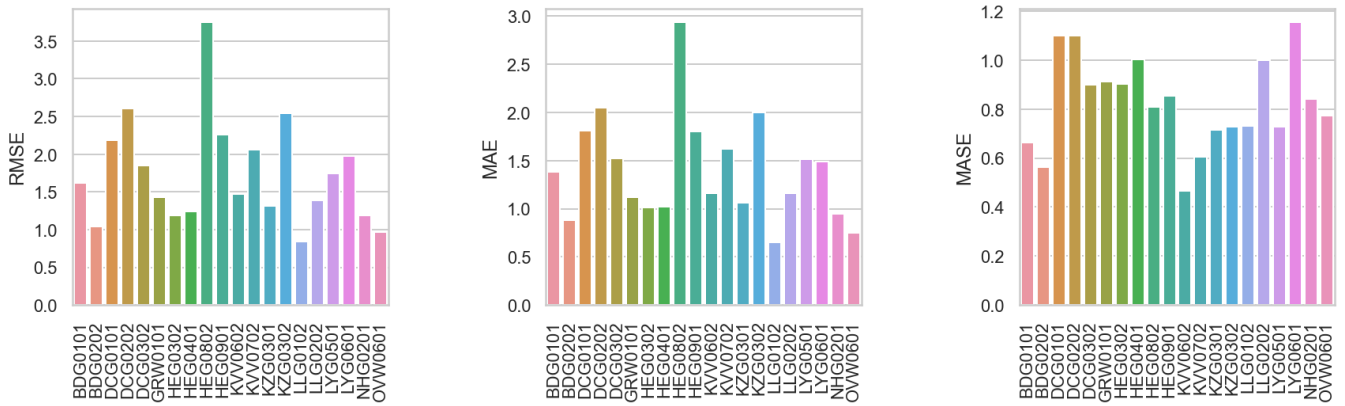


Figure 31: Metrics of IF1 with tacheometry as predictor in X-direction.

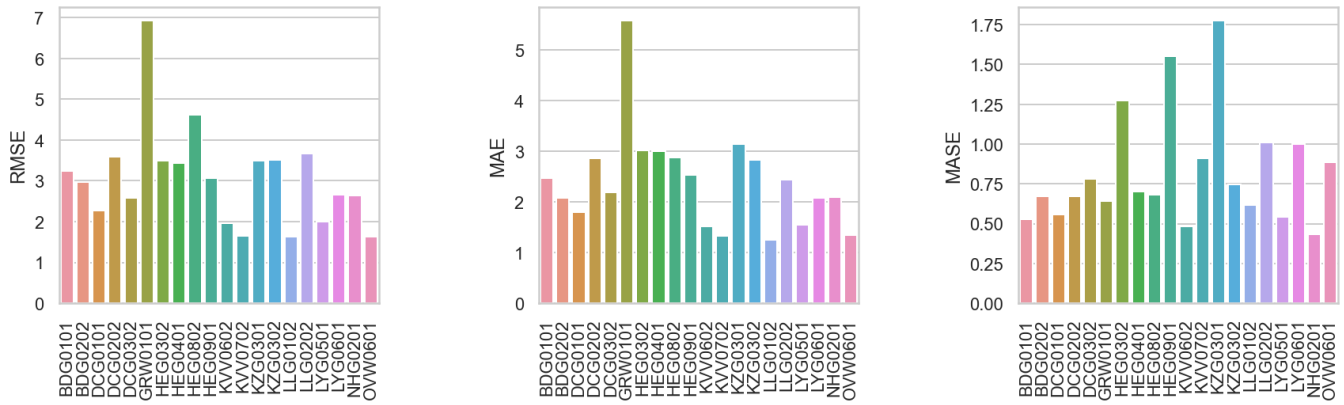


Figure 32: Metrics of IF1 with tacheometry as predictor in Y-direction.

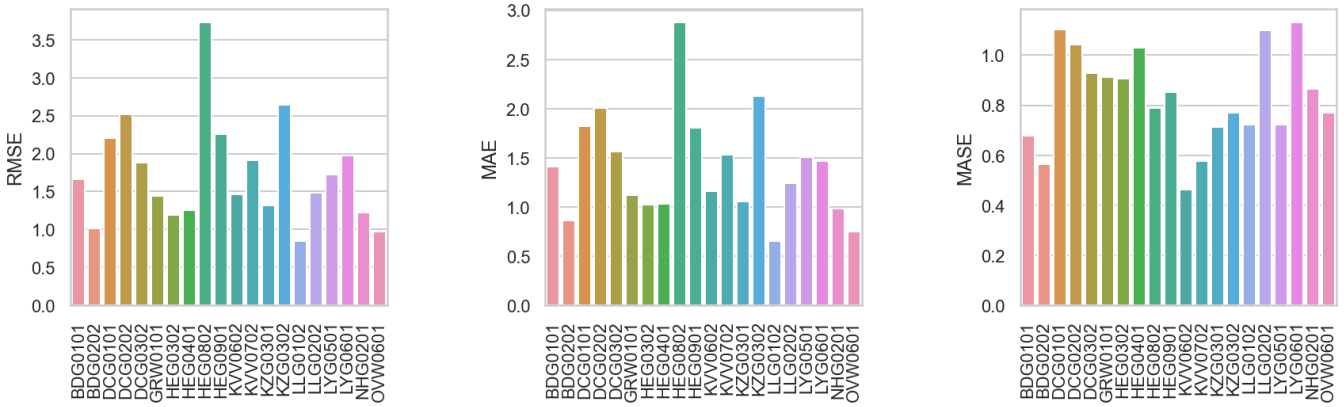


Figure 33: Metrics of IF1 with tacheometry as predictor in Z-direction.

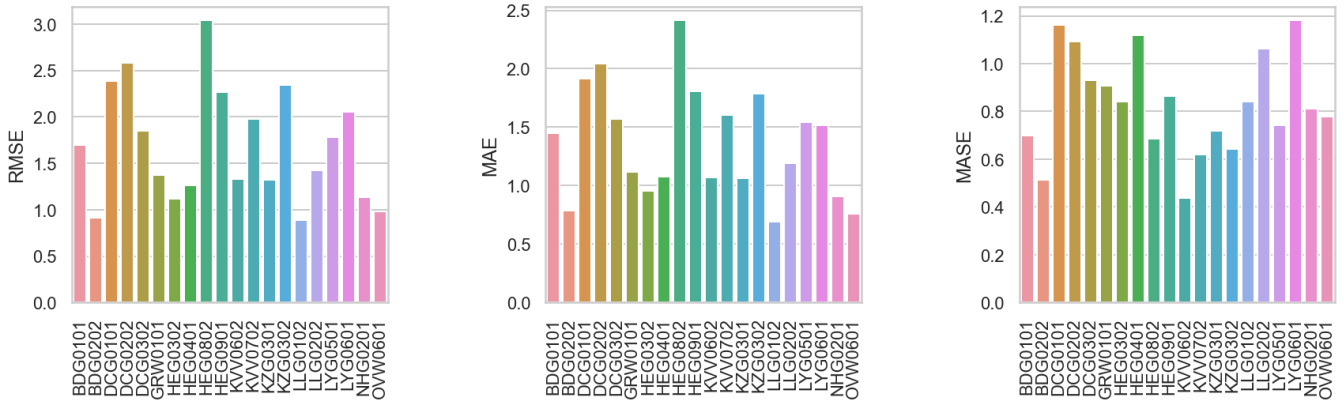


Figure 34: Metrics of IF2 with tacheometry as predictor in X-direction.

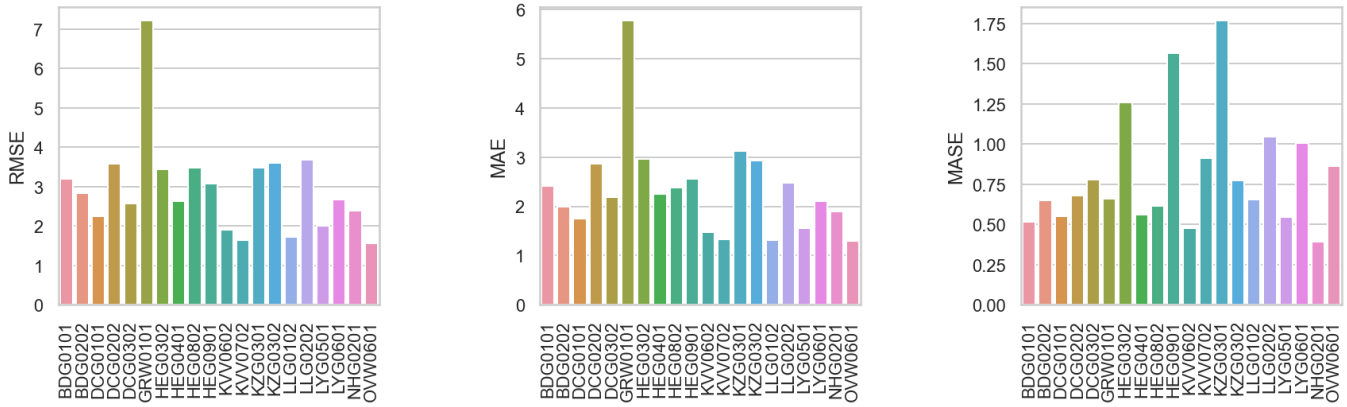


Figure 35: Metrics of IF2 with tacheometry as predictor in Y-direction.

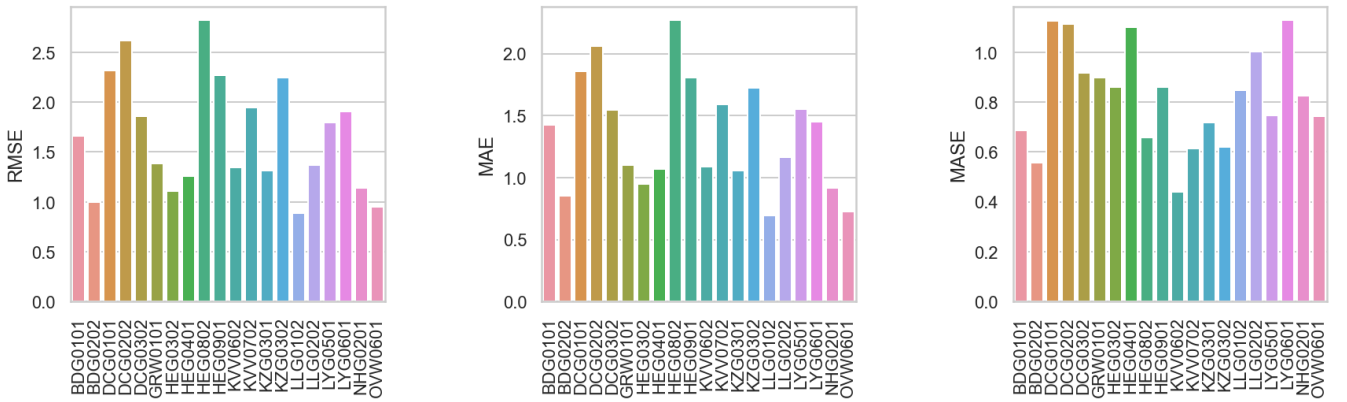


Figure 36: Metrics of IF2 with tacheometry as predictor in Z-direction.

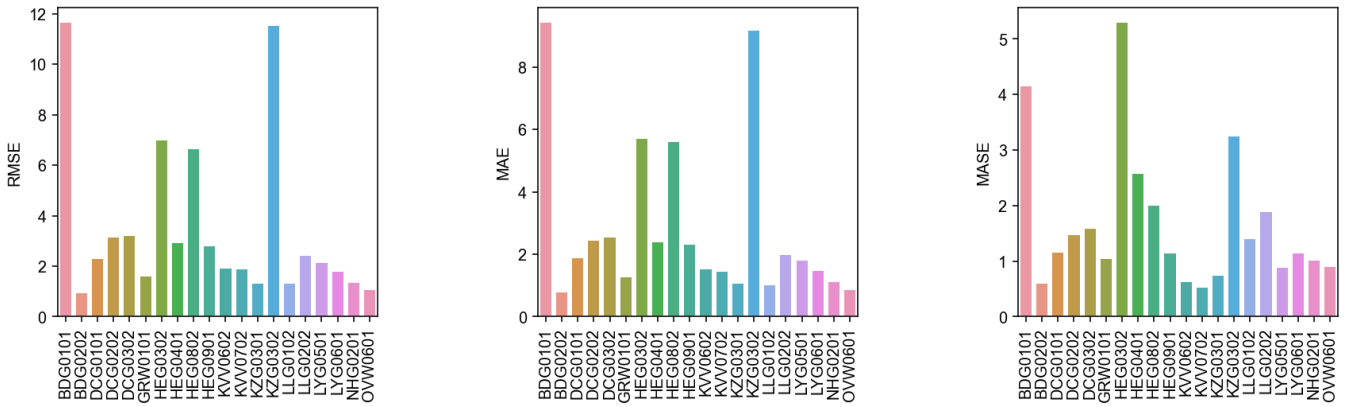


Figure 37: Metrics of LF1 with tacheometry as predictor in X-direction.

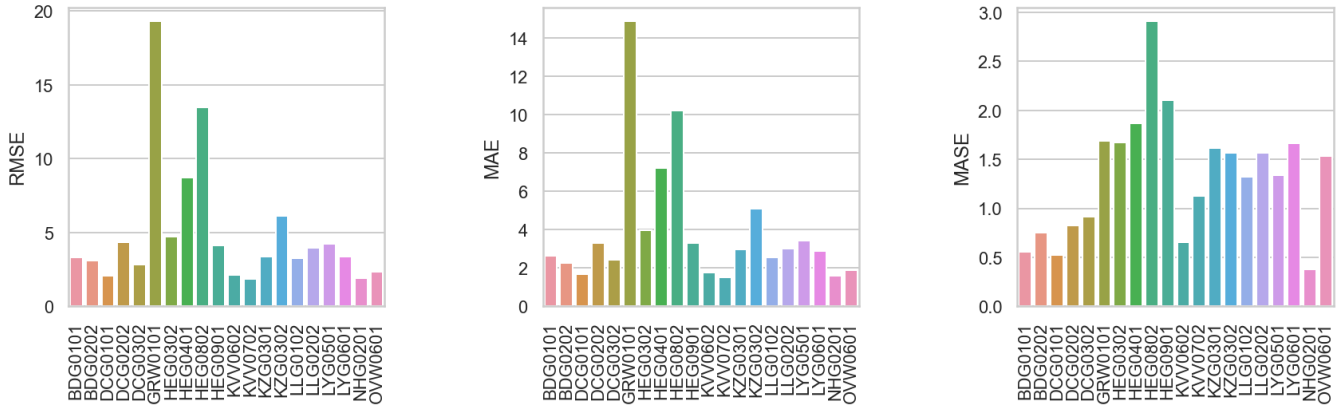


Figure 38: Metrics of LF1 with tacheometry as predictor in Y-direction.

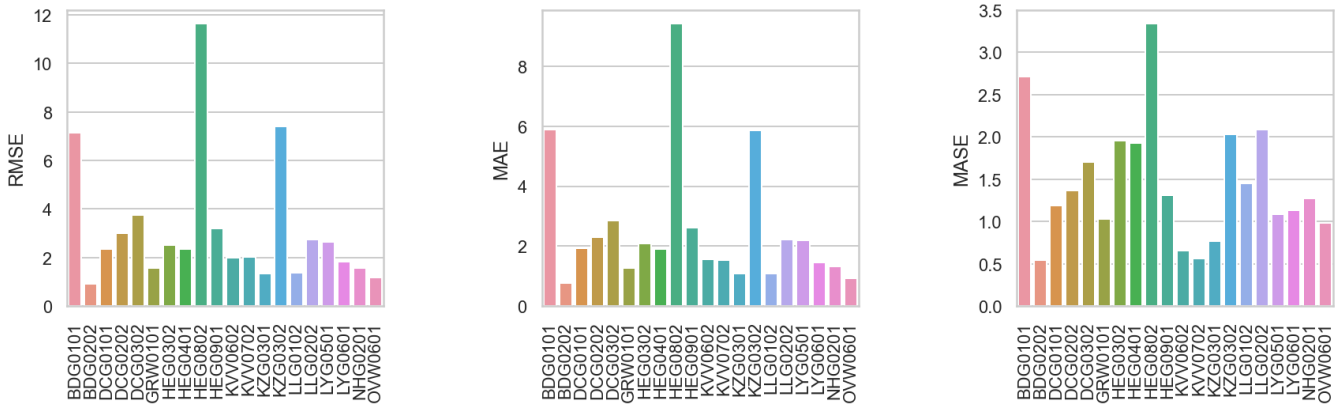


Figure 39: Metrics of LF1 with tacheometry as predictor in Z-direction.

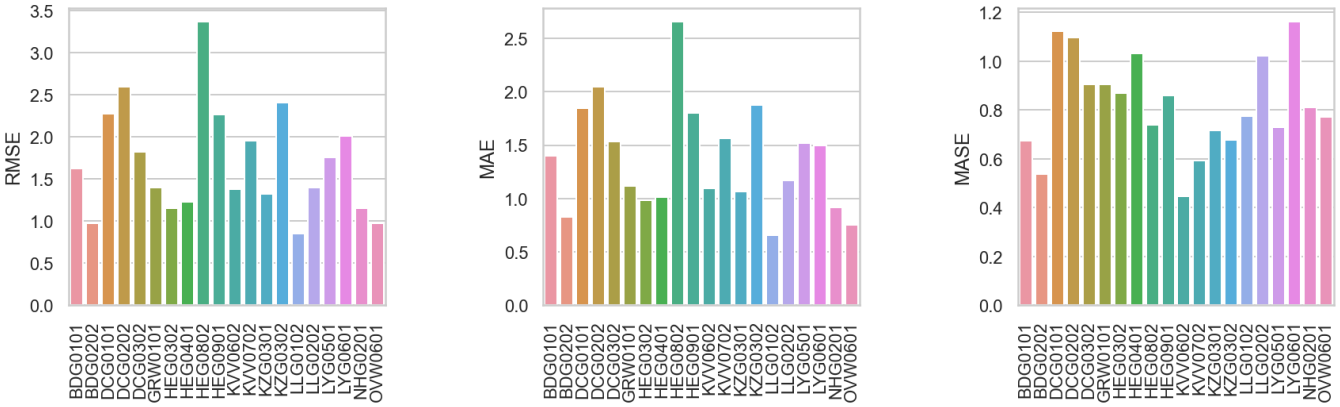


Figure 40: Metrics of LF2 with tacheometry as predictor in X-direction.

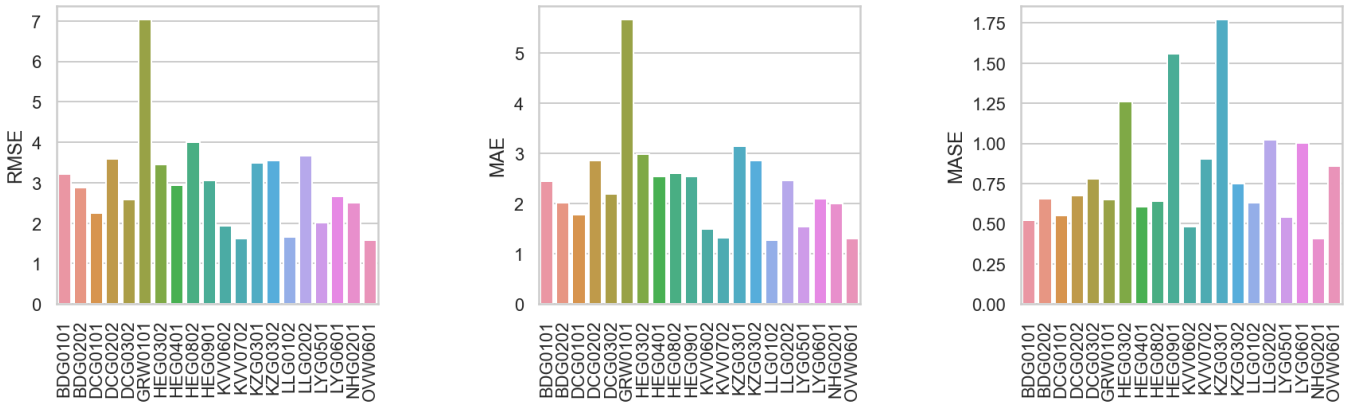


Figure 41: Metrics of LF2 with tacheometry as predictor in Y-direction.

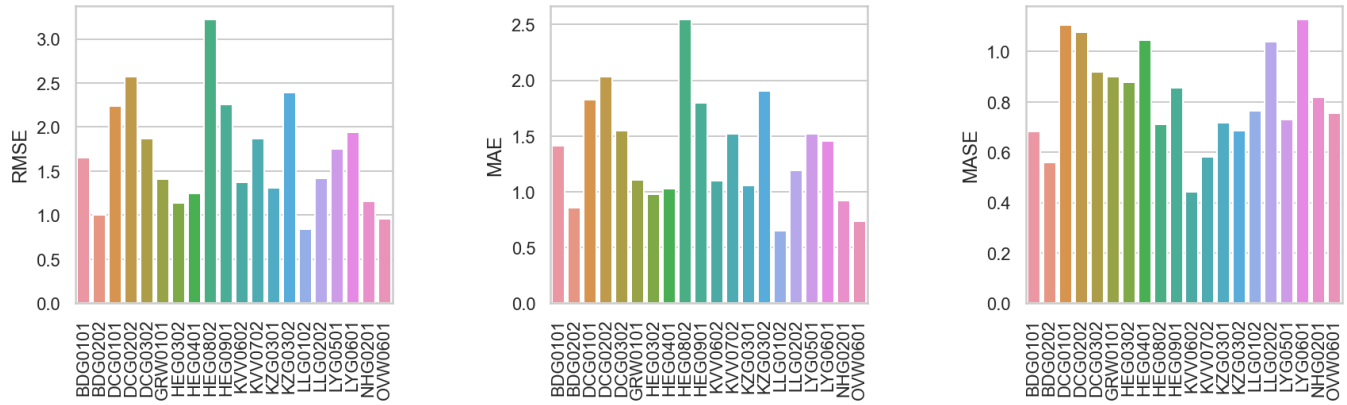


Figure 42: Metrics of LF2 with tacheometry as predictor in Z-direction.

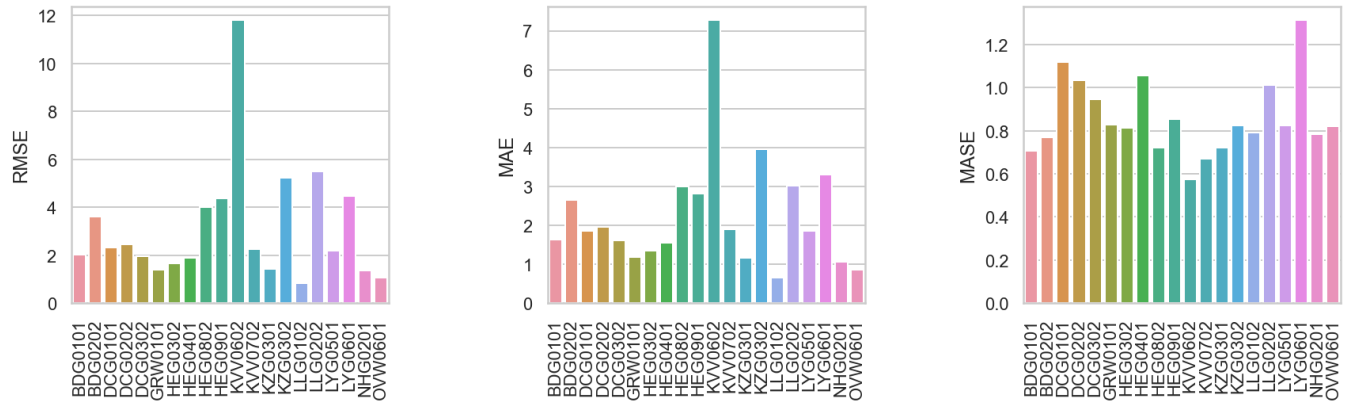


Figure 43: Metrics of BL3 without tacheometry as predictor in X-direction.

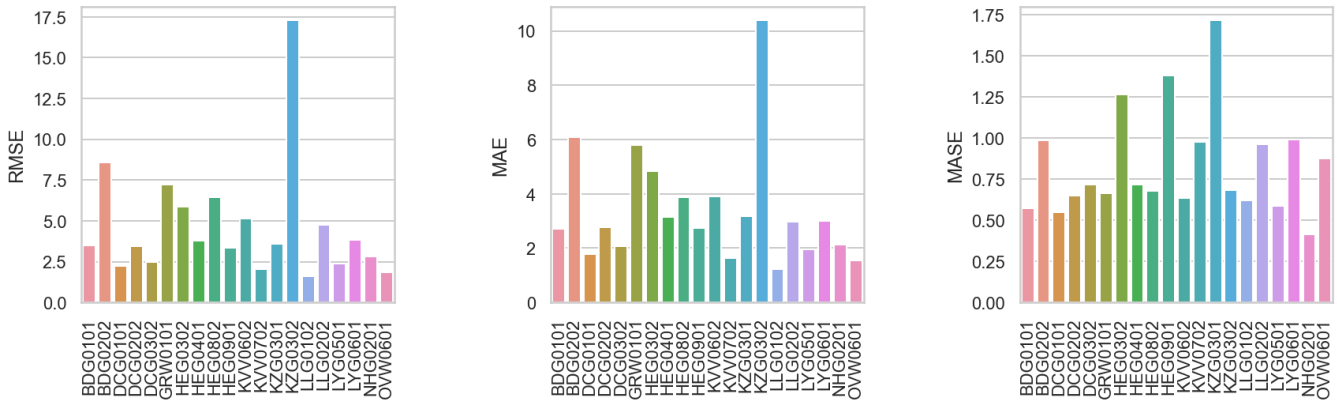


Figure 44: Metrics of BL3 without tacheometry as predictor in Y-direction.

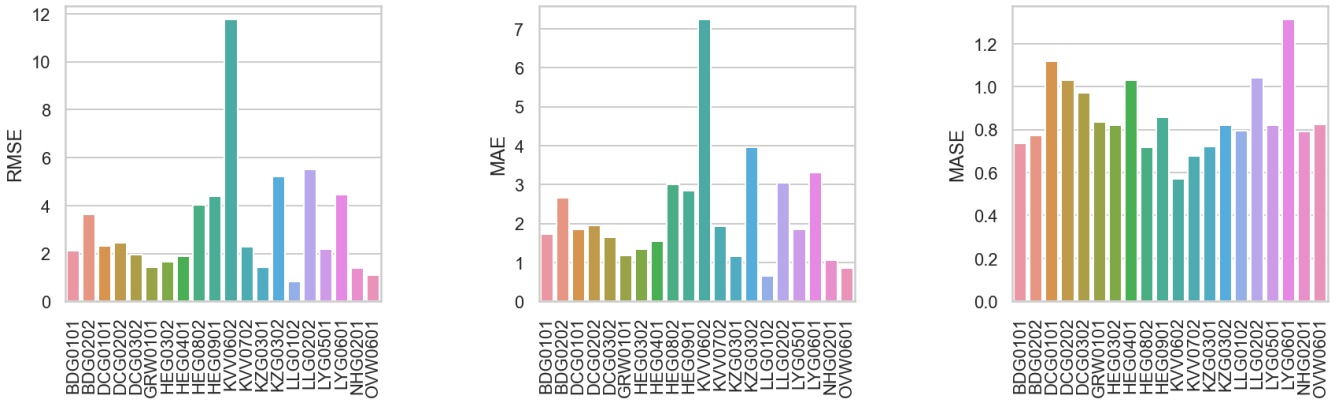


Figure 45: Metrics of BL3 without tacheometry as predictor in Z-direction.

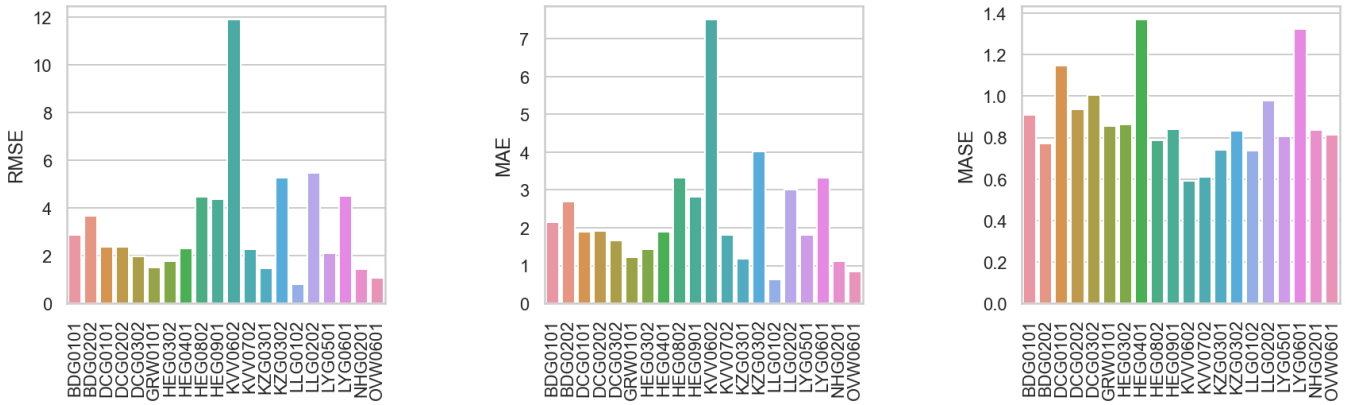


Figure 46: Metrics of BL4 without tacheometry as predictor in X-direction.

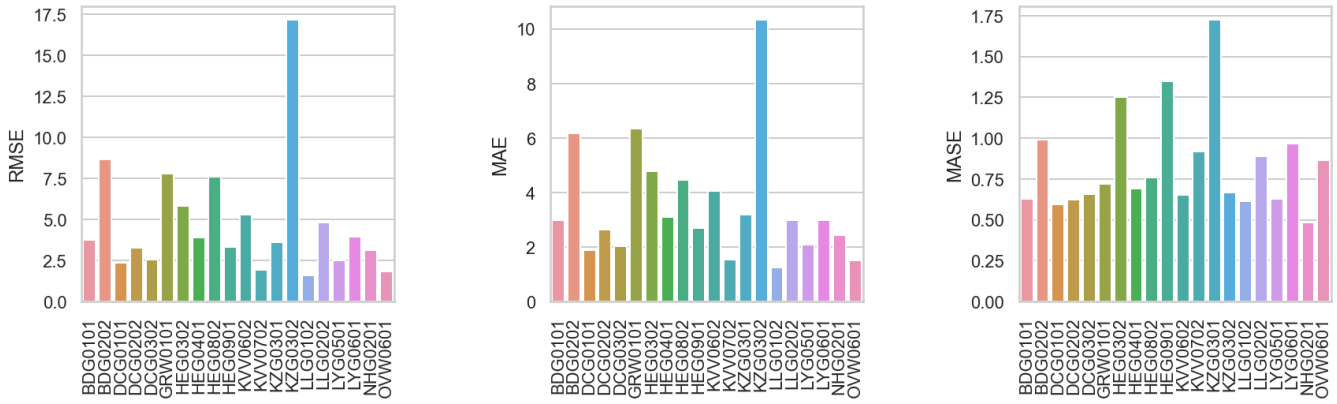


Figure 47: Metrics of BL4 without tacheometry as predictor in Y-direction.

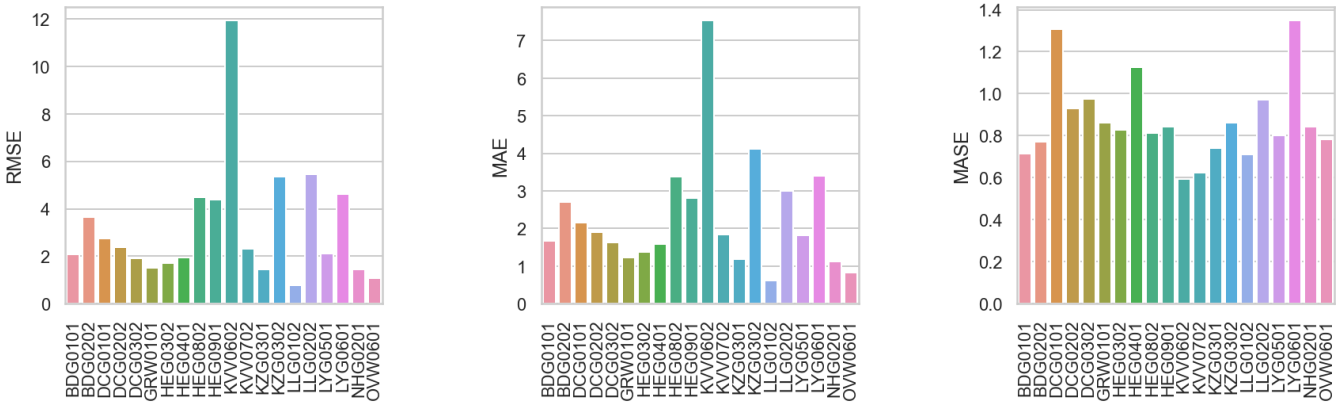


Figure 48: Metrics of BL4 without tacheometry as predictor in Z-direction.

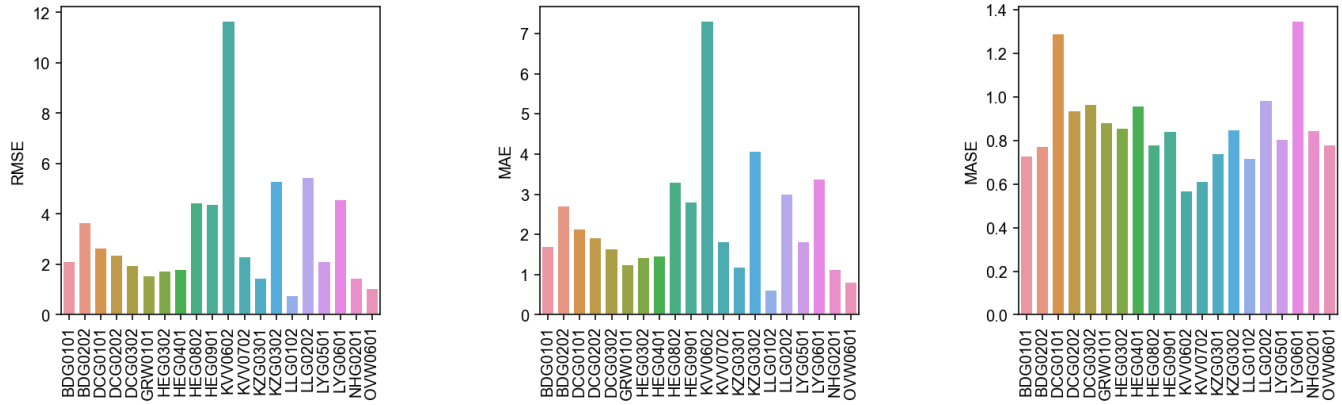


Figure 49: Metrics of EF1 without tacheometry as predictor in X-direction.

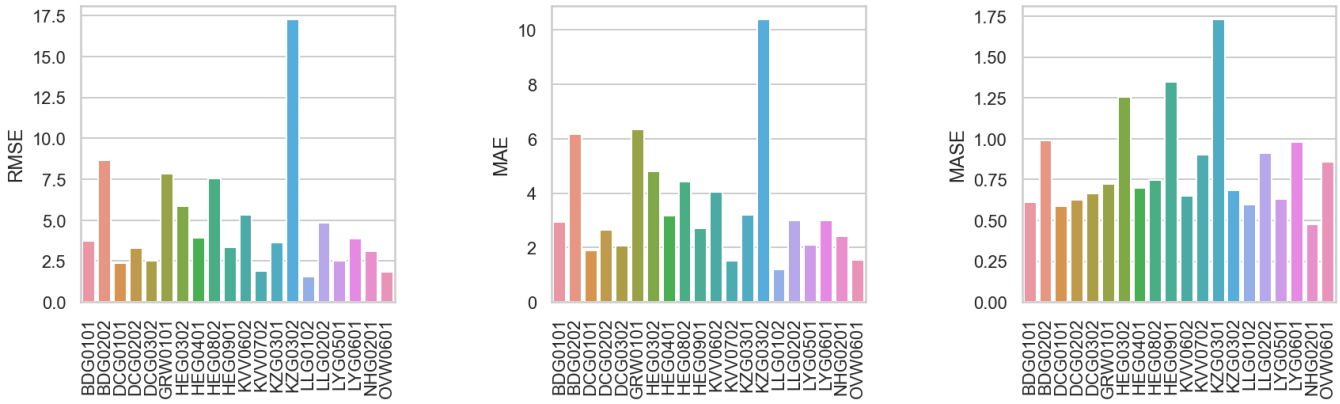


Figure 50: Metrics of EF1 without tacheometry as predictor in Y-direction.

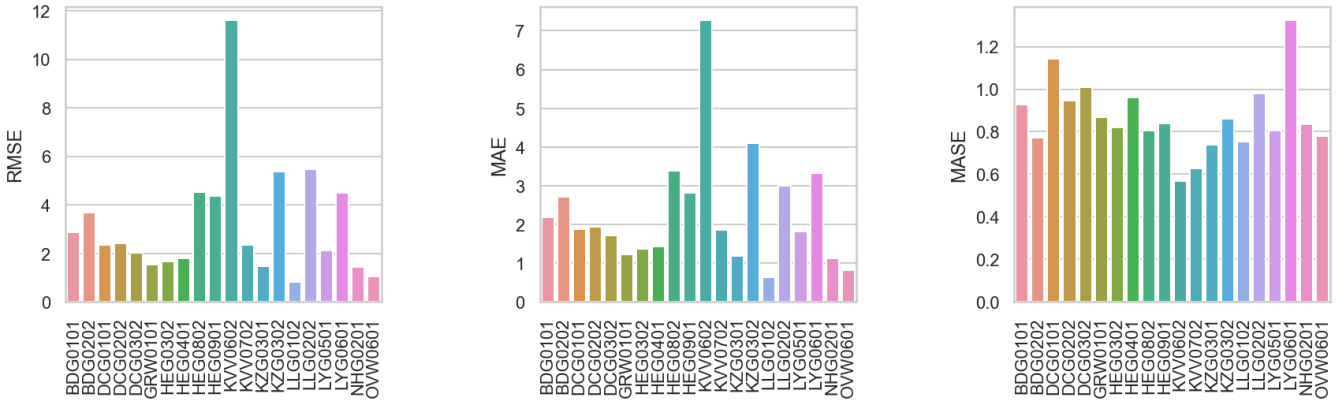


Figure 51: Metrics of EF1 without tacheometry as predictor in Z-direction.

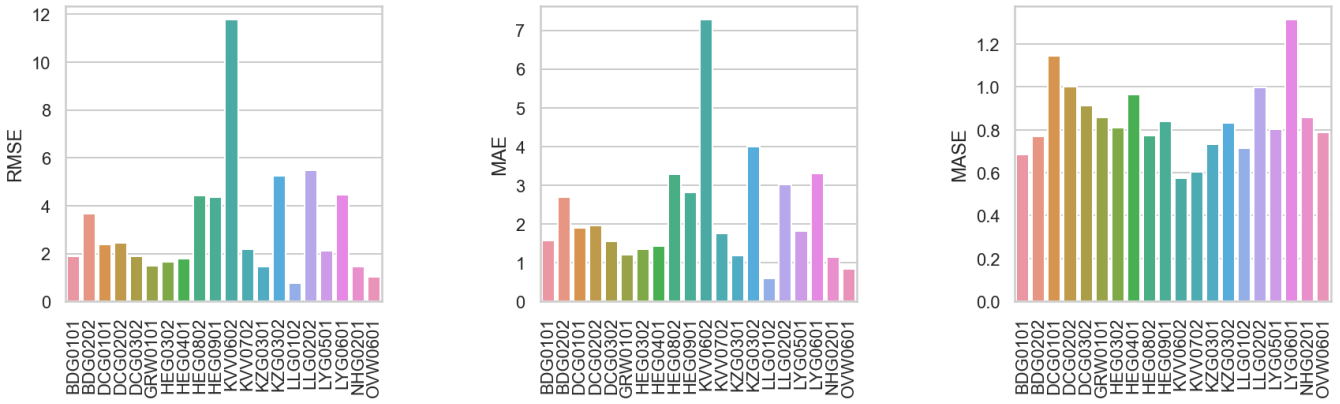


Figure 52: Metrics of IF1 without tacheometry as predictor in X-direction.

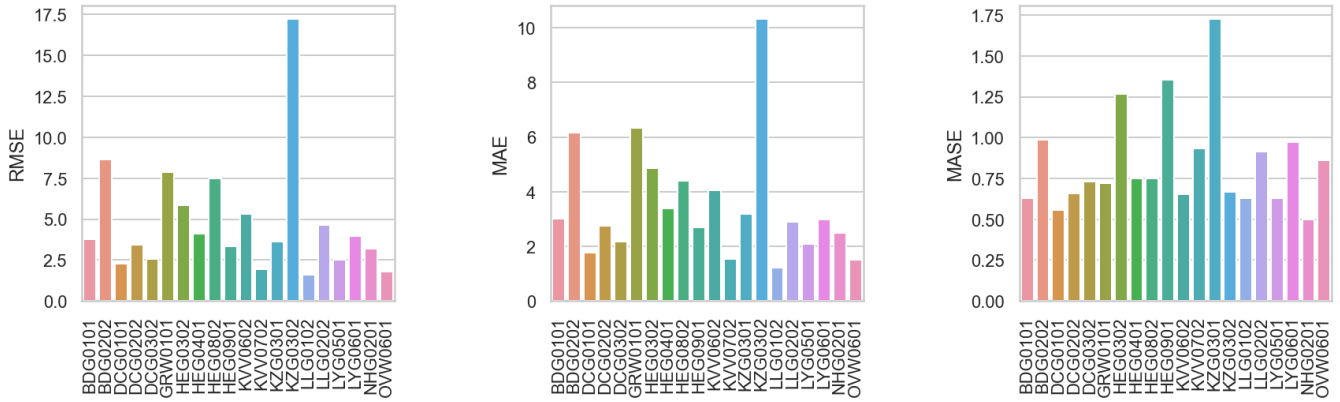


Figure 53: Metrics of IF1 without tacheometry as predictor in Y-direction.

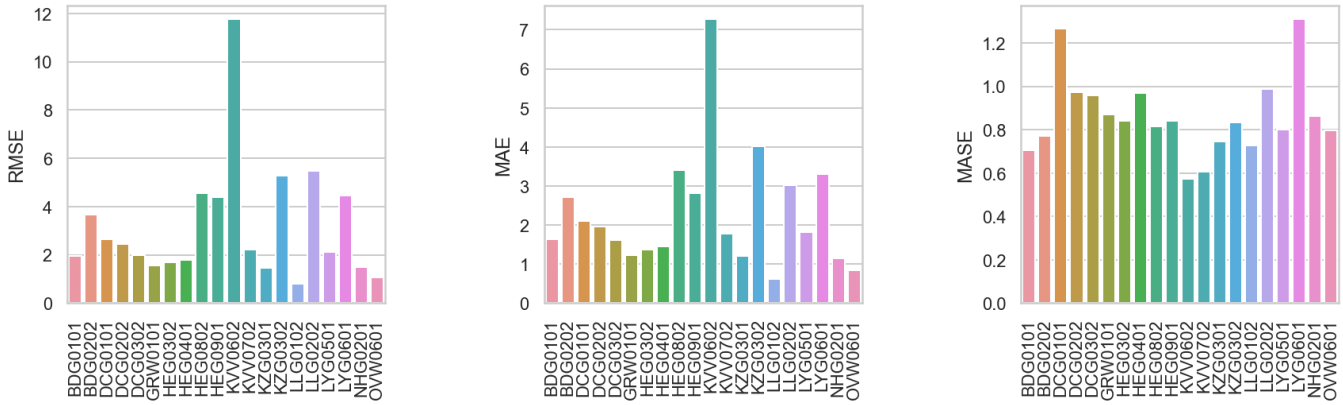


Figure 54: Metrics of IF1 without tacheometry as predictor in Z-direction.

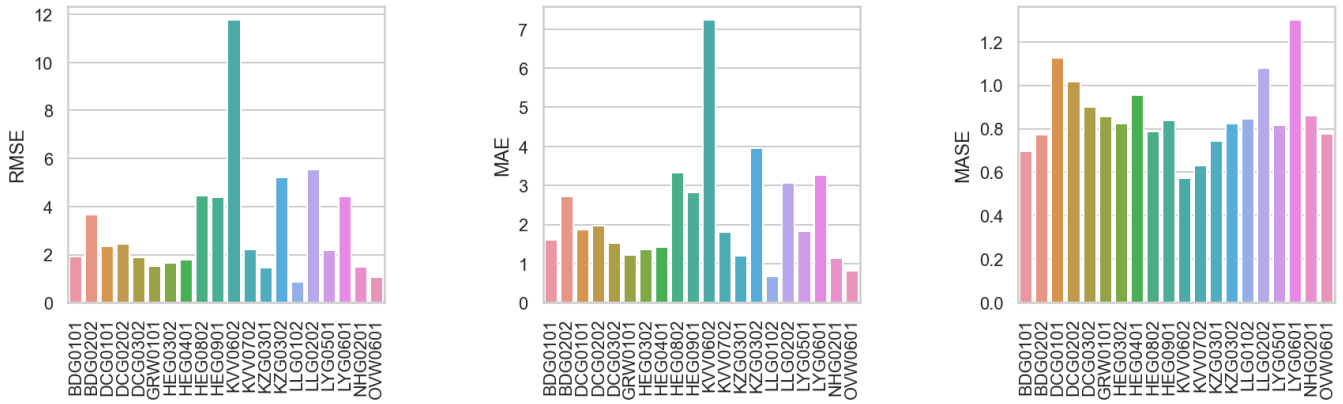


Figure 55: Metrics of IF2 without tacheometry as predictor in X-direction.

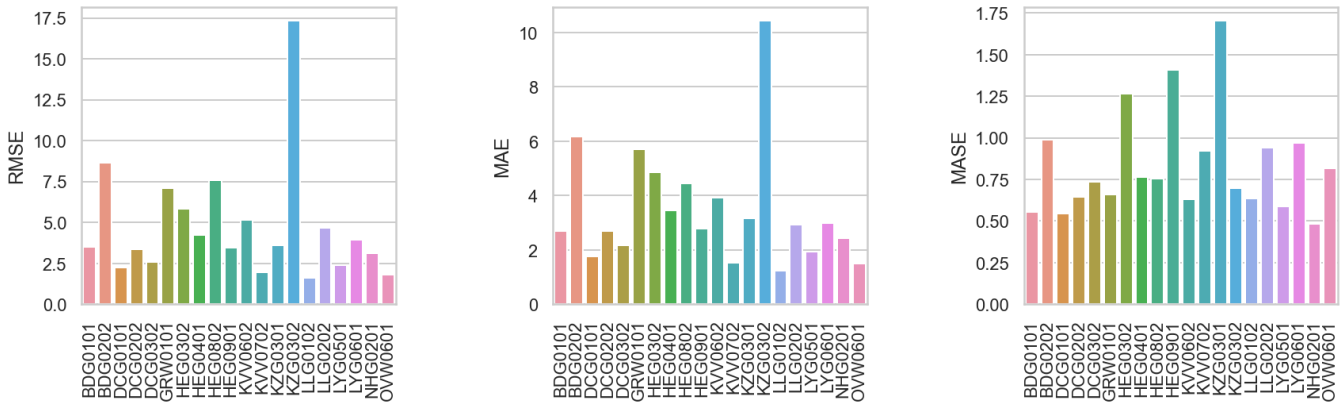


Figure 56: Metrics of IF2 without tacheometry as predictor in Y-direction.

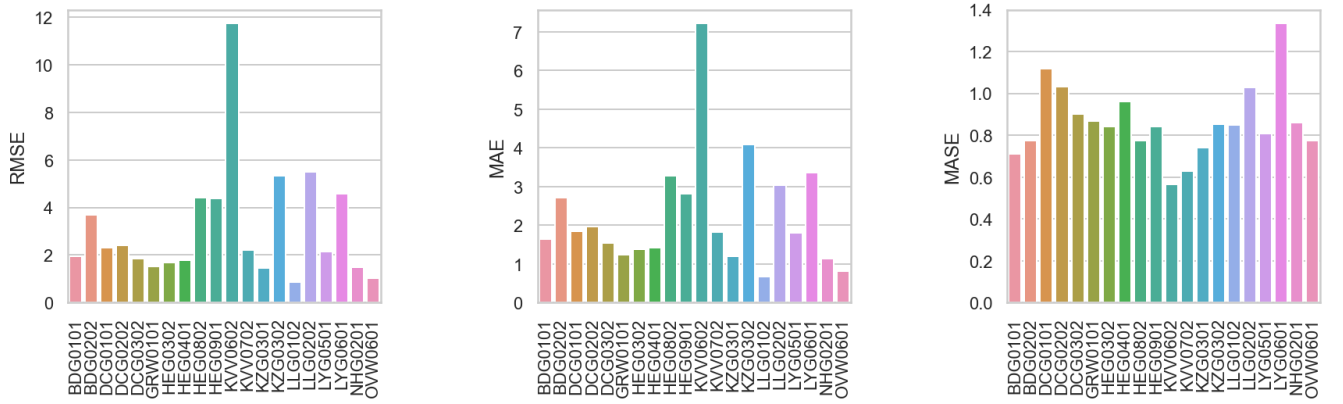


Figure 57: Metrics of IF2 without tacheometry as predictor in Z-direction.

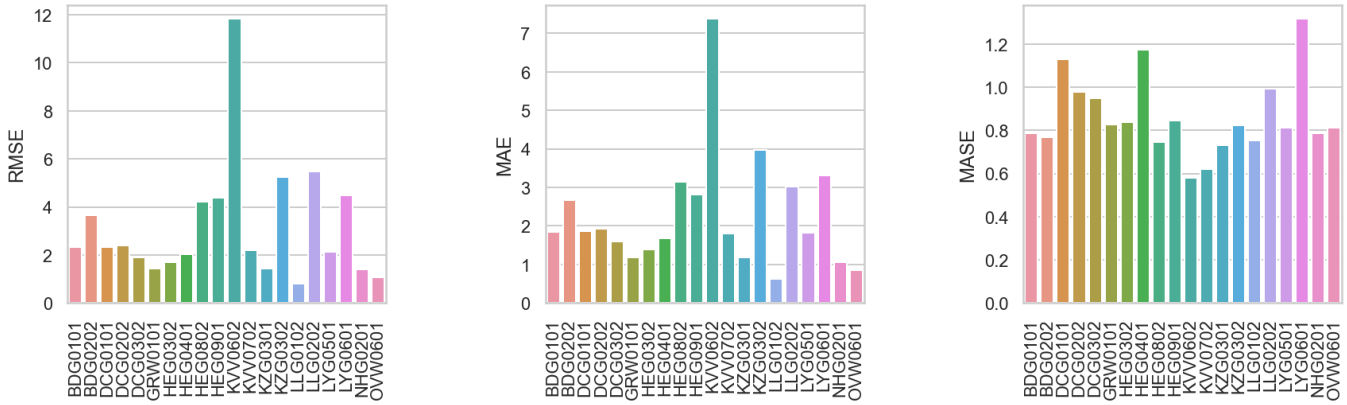


Figure 58: Metrics of LF1 without tacheometry as predictor in X-direction.

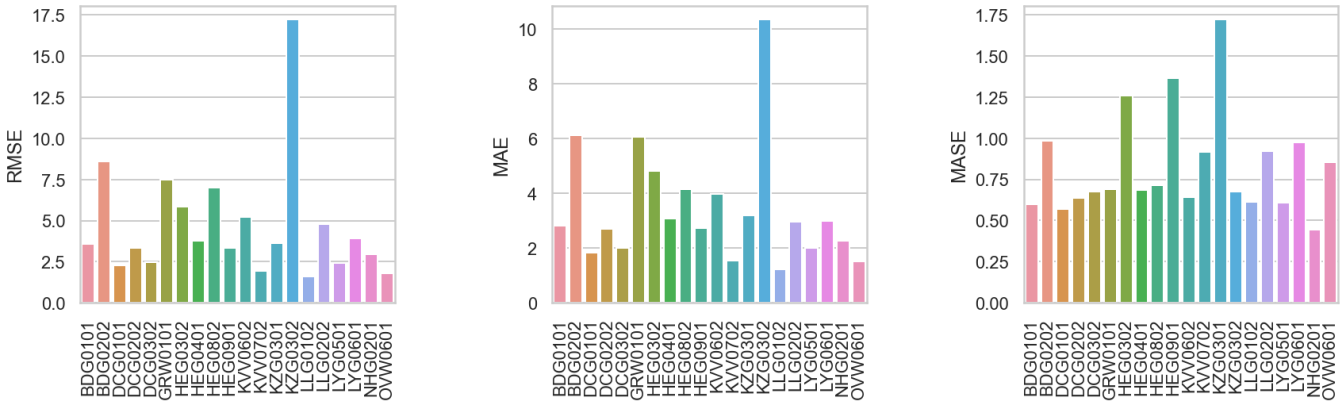


Figure 59: Metrics of LF1 without tacheometry as predictor in Y-direction.

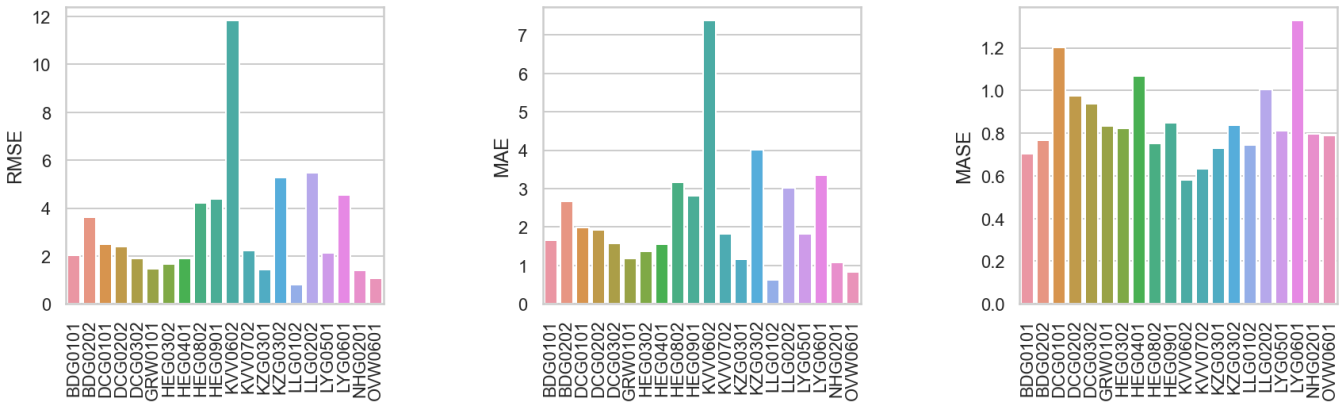


Figure 60: Metrics of LF1 without tacheometry as predictor in Z-direction.

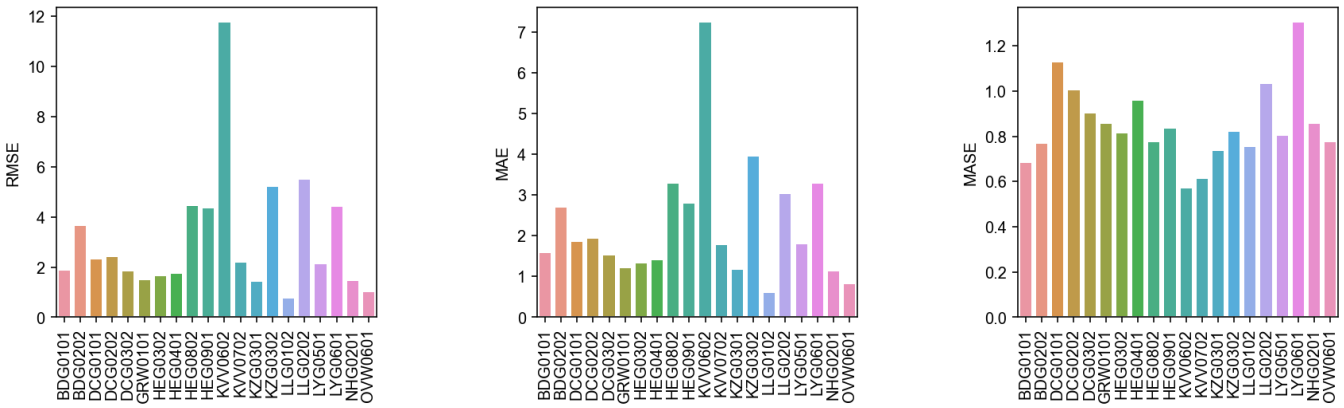


Figure 61: Metrics of LF2 without tacheometry as predictor in X-direction.

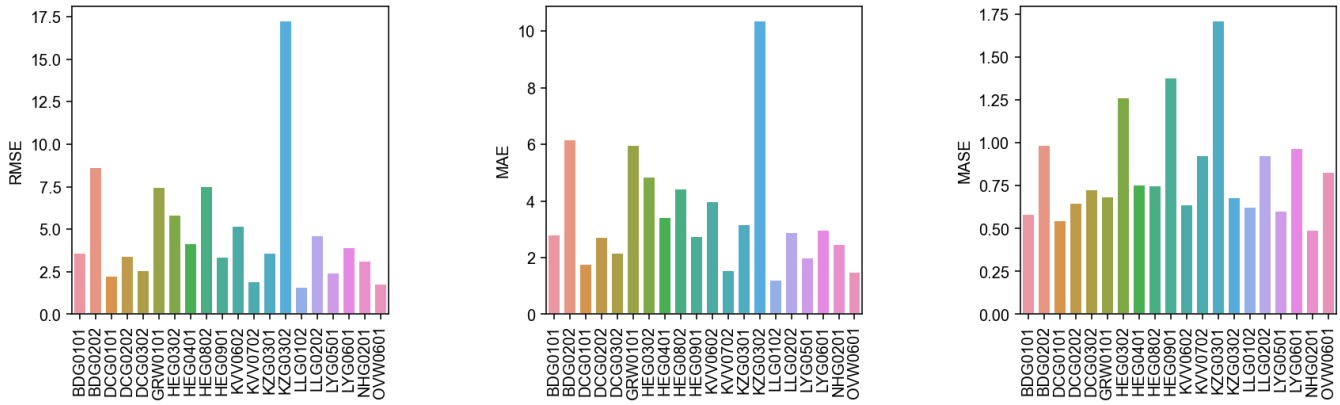


Figure 62: Metrics of LF2 without tacheometry as predictor in Y-direction.

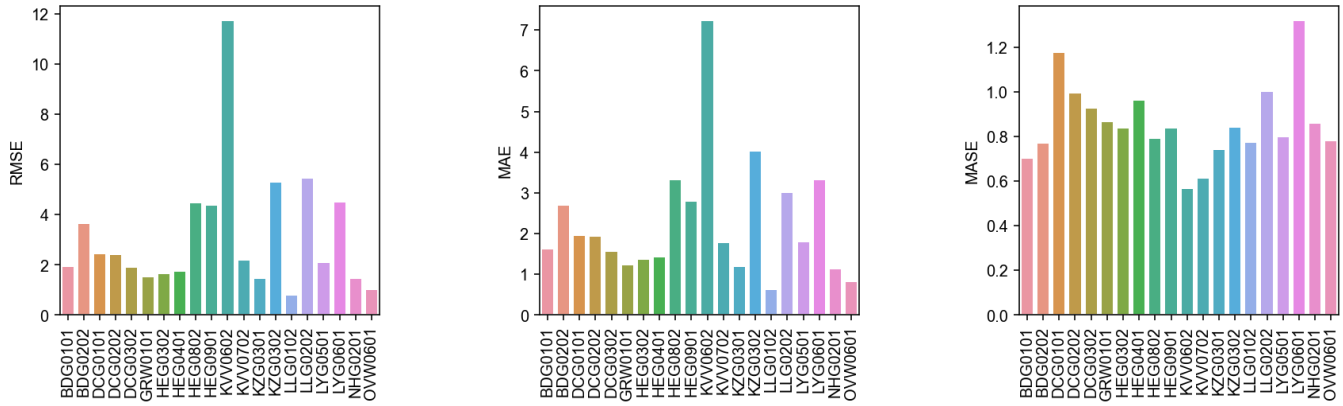


Figure 63: Metrics of LF2 without tacheometry as predictor in Z-direction.