GHENT UNIVERSITY

FACULTY OF SCIENCES

Computing a branch's total added value from incomplete annual accounting data using machine learning techniques

Tom Marchal

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Stijn Vansteelandt
Tutor: François Coppens
Department of Applied Mathematics, Computer Science and Statistics

Academic year 2018 - 2019

GHENT UNIVERSITY

FACULTY OF SCIENCES

Computing a branch's total added value from incomplete annual accounting data using machine learning techniques

Tom Marchal

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Stijn Vansteelandt
Tutor: François Coppens
Department of Applied Mathematics, Computer Science and Statistics

Academic year 2018 - 2019

Foreword

This thesis was made as a final project for the study “master of science in statistical data analysis” at Ghent university. The dataset used in all analyses of this thesis was obtained from the National Bank of Belgium and contains annual accounting data of individual companies in the Belgian port sector. This data can be accessed publically though the National Bank’s central balance sheet office. Other information concerning previously researched topics and theory, handled in this this thesis, are properly referenced in the main text and reference list.

As several people have provided me with support to complete this thesis, I would like to thank them in this short introduction. First, I would like to thank my promotor Stijn Vansteelandt for his valuable input, useful feedback and overall guidance in the process of writing this thesis. Secondly, I would like to thank François Coppens from the National Bank of Belgium for concisely explaining the research previously performed on this topic, providing the R code of this previous research, his feedback and overall support.

Table of contents

Abstract

The National Bank of Belgium publishes (sectoral) studies on employment, economic value added and other important economic variables on a yearly basis. A main information source for these studies is the data collected from annuals accounts, filed yearly to the central balance sheet office of the bank by individual companies. In practice, however, annual accounts of different companies are not available at the same time or, in some cases, are filed with a large delay. This makes timely performance monitoring/reporting difficult and causes studies to be published with a large time lag. In order to publish studies in a more timely manner, the National Bank has been developing imputation techniques to estimate economic variables of interest in the presence of missing annual accounting data.

In this thesis, the performance of an alternative imputation framework to the previously researched estimation methods of Vansteelandt, Coppens, Vackier et al (2019) is investigated in the setting of predicting the total economic value added in the Belgian port sector in 2015. In the benchmark paper of Vansteelandt, Coppens, Vackier et al (2019), OLS under a missing at random assumption and size-based outlier correction method was determined as the best performing imputation strategy in terms of 95% prediction interval width for the total economic value added. The main goal of this thesis was to compare the performance of this benchmark model with several regression-tree based machine learning algorithms (bagged regression trees, random forest, boosted regression trees and Bayesian additive regression trees) as well as with a "Superlearner" ensemble, which combines all the aforementioned algorithms with the OLS method. In order to construct 95% prediction intervals for the machine learning algorithms, split conformal inference was used in combination with bootstrapping and all methods were tested in the setting of no outlier correction, size-based outlier correction and a leverage-based outlier handling mechanism. Under the assumptions of exchangeability, a symmetric distribution around the estimated values and homoscedasticity, reliable 95% prediction intervals were obtained.

In terms of performance, a distinction could be made between large companies and small companies in the Belgian port sector. For large companies, none of the researched methods performed better than the OLS benchmark. Possible reasons for this were superiority of the

OLS method, the fact that split conformal inference results in less available training data for the tested algorithms, or, depending on which outlier correction method was used, superiority of the size-based outlier correction method for large companies. For small companies, similar results were obtained and no method performed clearly better than OLS. The exception being that the use of a leverage-based outlier handling mechanism seemed to yield better results than the size-based outlier correction method, possibly due to a larger variability in other company characteristics than size alone for small companies. Overall, OLS can be seen as a viable, well-performing imputation strategy for estimating the economic value added in the Belgian port sector. Besides performing well in terms of prediction interval width, OLS is also an easy to understand, flexible, computationally efficient and easy automatable method, which could easily be adjusted to estimate other economic variables of interest in other sectors.

1. Introduction

The Central Balance Sheet Office (CBSO) of the National Bank of Belgium (NBB) is responsible for collecting the annual accounts of about 300,000 companies. In accordance with Belgian law, these annual accounts should be submitted to the CBSO of the NBB within thirty days after approval and within seven months after the end of the financial year (National Bank of Belgium, 2019).

Both the annual accounts published by large companies under a "full scheme" as well as the accounts filed by medium size and smaller companies under an "abbreviated" and "micro scheme", which contain subsets of the information submitted by large companies under the full scheme, contain important information for several studies and statistics published yearly by the NBB. In these studies, important (macro-)economic variables such as the economic value added and employment within a particular sector are analysed in more detail (Vansteelandt, Coppens, Vackier, Reynders & Van Belle, 2019).

One of the main issues encountered in these studies is the fact that companies of all sizes often publish their financial statements with a certain delay, hindering timely monitoring of the economic performance of a particular sector and causing the sectoral studies to be published with a large time lag. In practice, most annual accounts for a certain financial year y are filed with the CBSO between July $(y + 1)$ and August $(y + 1)$. For some companies however, it can take until February/March of the next year $(y + 2)$ to publish their annual accounts, making timely reporting on performance without imputation or estimation of the missing variables of interest for the missing annual accounts impossible. For example, the study of the economic performance of the Belgian port sector of the financial year 2015 (Mathys, 2017) was, due the unavailability of several annual accounts, published in June 2017. As port authorities demanded more timely publications, the NBB developed techniques to impute the missing values for the annual accounts that had yet to be submitted to the CBSO. The estimated company statistics resulting from these techniques are then added to the known values of the accounts already in the database, providing the Bank with a "flash estimate" of the variable of interest (Vansteelandt, Coppens, Vackier, Reynders & Van Belle, 2019).

The main issue with the NBB's previous approach was that no clear framework was provided on how this "flash estimate" was to be obtained and estimation methods were determined on an ad hoc basis. This entailed that the used methods were more labour-intensive and time

consuming, especially when a "flash estimate" needed to be produced for a different sector or variable of interest. Furthermore, the ad hoc nature of the methods implied that their accuracy and precision remained unclear. For this reason, a collaborative research project between the NBB and Ghent University (UGent) was set up in order to develop accurate, flexible and easily automatable methods for estimating aggregated economic variables of a particular sector in the presence of missing annual accounts data, alongside their uncertainty. This uncertainty assessment allows the NBB to decide when and at what level of detail the estimate of the economic variable can be made public or used in a sectoral study. This research is momentarily focused on the prediction of total economic value added of a sector by computing the total value added of individual companies in the Belgian port sector, but can easily be extended to other variables of interest or sectors (Vansteelandt, Coppens, Vackier, Reynders & Van Belle, 2019).

Currently several estimation methods have been developed to impute a branch's total value added from incomplete annual accounting data and the uncertainty accompanied with these predictions and its aggregate. All of the researched procedures relied on the assumption that companies that did not yet deposit their annual accounts to the CBSO were missing at random. The decision or necessity to delay the submission of a company's annual account should therefore not depend on the company's value added. This assumption can be tested retrospectively and should the assumption fail to hold, the proposed estimation methods can relatively easily be adjusted. For more detail and information concerning the performance of the different imputation methods, the reader is encouraged to study the paper of Vansteelandt, Coppens, Vackier et al (2019) on "Estimation methods for computing a branch's total value added from incomplete annual accounting data". In this research paper the assumptions and performance of all currently proposed estimation methods were analyzed in the setting of the Belgian port sector in 2014 and 2015 (Vansteelandt, Coppens, Vackier, Reynders & Van Belle, 2019).

The research paper concluded that of all studied estimation techniques, the method using ordinary least squares (OLS) should be preferred because of its flexibility in the use of auxiliary variables, weaker assumptions, smaller estimation errors (prediction intervals), ease to automate and relatively low complexity (Vansteelandt, Coppens, Vackier, Reynders & Van Belle, 2019). As the research in this thesis will be conducted on the same dataset of the Belgian port sector in 2015 as used in the paper cited, the results of the imputation method using OLS will be used as a benchmark to evaluate the performance of the imputation methods developed in this thesis. It should be noted that predictive performance is not the only measure to take into

account when comparing the resulting methods to the benchmark. Ease of use, implementation, complexity and generalizability to other sectors and variables are also aspects that should be carefully considered.

The main goal of this thesis is to evaluate the use of flexible statistical learning methods to model a branch's total value added from incomplete annual account's data with modern machine learning techniques. More specifically, the performance of several regression-tree based algorithms, together with an ensemble algorithm combing all of these methods with the OLS estimation method will be evaluated. The predictive performance of these flexible methods will be compared to the OLS benchmark in terms of prediction interval width of the aggregated total economic value added in the Belgian port sector (2015). The prediction intervals of the statistical learning methods will be constructed according to the theory of split conformal inference (Lei, G'Shell, Rinaldo, Tibshirani & Wasserman, 2018). It should be noted that the use of flexible statistical learning methods greatly increases the complexity, difficulty of automation and interpretability of the resulting imputation strategy and one should consider whether the possible increase in predictive capability is worth the increase in complexity compared to the current OLS method

Finally, a practical problem may arise for both the OLS as the flexible learning methods when several companies in the data set are much larger in terms of value added than others, leading them to be very influential on the fit (Vansteelandt, Coppens, Vackier, Reynders & Van Belle, 2019). Two possible solutions to dealing with this issue are proposed and implemented with the flexible learning methods of this paper. The first method takes a leverage based approach in dealing with possible outliers, while the second method takes into account solely the value added of companies in the previous year. This second approach was also suggested and implemented in the benchmark paper. Both approaches will be discussed in more detail in the methodology section of this thesis.

2. Methods

2.1. Dataset

The estimation techniques described in the following section will be implemented using a dataset consisting of annual accounting data of the Belgian port sector in the year 2015. The reason for using the year 2015 is that currently all annual accounts of this year have been filed with the CBSO and the actual population value of the total economic value added in that year is therefore known. The dataset contains the dates at which companies' annual accounts of the year 2015 were published, making it easy to divide the dataset into two separate sets at a certain date. One dataset which contains the companies of which the added value in the year is known, to be used as the training set for the method, and another dataset for which the added values have to be imputed using the trained method. As the date increases, more and more companies are included in the training set while less values have to be imputed, resulting the estimated total economic value added of the sector to grow closer and closer to the population value. In this paper eight splits of the dataset are performed on the dates 15-07-2016, 30-07-2016, 15-08-2016, 31-08-2016, 15-09-2016, 30-09-2016, 15-10-2016 and 31-10-2016. On the first date, the 15th of July 2016, most of the annual accounts should already have been filed to the CBSO according to Belgian law. In practice, 1625 of the 4034 annual accounts had yet to be submitted for the port sector. Over the assessment period, the number of annual accounts to impute dropped continuously and at the final assessment date, the 31st of October 2016, only 137 missing accounts remained.

The main variables of interest for predicting the economic value added of companies in this thesis are the first digit of the NACE code, defining the branch of economic activities that the company undertakes, the value added of the company in the previous year, the number of people employed if available, the fiscal result (taxable profit), the existence/availability of added value of the company in the previous year, the availability of the fiscal result in the current year and the size of the company (small or large) defined by the required reporting scheme. As large and small companies differ significantly in required disclosure formalities and other characteristics, prediction algorithms are fitted separately for both the small and large companies at a certain date. This same strategy was implemented in the benchmark paper (Vansteelandt, Coppens, Vackier, Reynders & Van Belle, 2019).

2.2. Imputation techniques

In this section, the theory behind the employed estimation techniques will be summarized briefly. The estimation techniques used in this thesis are ordinary least squares (OLS), bagged regression trees, random forests, gradient boosting, Bayesian additive regression trees and an ensemble combining all of the methods listed above using the SuperLearner package in R. All the imputation techniques use the same predictor variables *P* and make the assumption that the missing annual accounts are missing at random. Missing at random in this setting means that for a given company *i* with characteristics P_i, the missingness does not depend on the economic value added (Y_i), or mathematically: $Y_i \perp\!\!\!\perp missing_i | P_i$ (Rubin, 1976). An assumption that was tested on the used dataset and can otherwise easily be adjusted for according to the paper of Vansteelandt S., Coppens F., Vackier M. et al (2019).

2.2.1. Ordinary least squares (OLS)

In the previous research paper of Vansteelandt, Coppens, Vackier et al (2019), OLS is suggested as the most suitable method for imputing the economic value added of a sector in the presence of missing annual accounting data and its underlying assumptions are checked. This well-known technique can be used to estimate the parameter coefficients (β's) of the following model:

$$Y_i = \beta_0 + \beta_1 \mathrm{Y}_{i,(t-1)} + \beta_2 \mathrm{Fiscal}_i + \beta_3 IFiscal_i + \beta_4 I\mathrm{Y}_{i,(t-1)} + \beta_5 \mathrm{PE}_i + \sum_{j=6}^{13} \beta_j \, NACE_{j,i} + \sum_{j=14}^{21} \beta_j \, NACE_{j,i} * \mathrm{Y}_{i,(t-1)} + \varepsilon_i$$

with,

Y_i: value added of company *i*

$\mathrm{Y}_{i,(t-1)}$: value added in the previous year, zero when missing

Fiscal_i: fiscal result of company *i*, zero when missing

$IFiscal_i$: indicator variable for the presence of fiscal (taxable profit) data

$I\mathrm{Y}_{i,(t-1)}$: indicator variable for the presence of added value data in the previous year

PE_i: number of people employed in company *i*

$NACE_{j,i}$: dummy variables for the first digit of the NACE code (8 dummies)

ε_i: error term with: cov($\varepsilon_i, \varepsilon_j | P_i, P_j$) = 0 ($\forall$ *i,j*) and $\varepsilon_i | P_i \sim N(0, \sigma^2)$

Note that interaction terms between the added value in the previous year and the first digit of the company's NACE code are included in the regression model. This is not unreasonable as different activities could lead to larger/smaller scale benefits and therefore larger/smaller increases or decreases in economic value added of a company of a certain size. Finally, the used regression model makes the assumptions of homoscedasticity, no serial correlation and normally distributed errors in order to yield unbiased estimates and prediction intervals (Gujarati & Porter, 2009). Using the regression model above, the total economic value added of a sector can be estimated by:

$$Y_{tot} = \sum_{j=1}^{n} f(P_j) + \sum_{i=1}^{m} Y_i$$

with,

Y_{tot}: total economic value added
n: the number of companies with missing annual accounting data
$f(P_j)$: estimated value added for company *j* with missing data and characteristics P_j
m: the number of companies for which the value added is known
Y_i: the known value added of company *i*

2.2.2. Bagged regression trees

Consider a training set consisting of *N* observations with predictor variables or features *P* and an outcome variable Y. A regression tree is then a prediction algorithm that in a step-by-step manner splits the predictor space into different regions in order to optimize a certain criterion depending on the outcomes Y of the training set. In the setting of this thesis, regression trees are built in order the minimize the sum of squared errors ($\sum_{i=1}^{N}(Y_i - f(P_i))^2$) in each split. With $f(P_i)$ the estimated value added of company *i* using predictors P_i. This is done until a predefined maximum number of splits or "nodes" are selected, after which a process known as "pruning" is applied in order to reduce tree size (Hastie, Tibshirani & Friedman, 2001). The predictors considered for building regression trees in this thesis are the same as the ones listed with the OLS method.

One of the main issues with the use of regression trees for prediction is their high variance due to their hierarchical nature. An erroneous split in the top of the tree is propagated to all splits below it, causing a small change in the observed data to potentially lead to very different regression trees. A possible solution to deal with this high variance in trees is bagging (Hastie, Tibshirani & Friedman, 2001).

Bagged regression trees can be seen as an ensemble learning method. The idea of ensemble learning is to create a better prediction method by combining the strengths of multiple, simpler base prediction algorithms. With bagged regression trees, the high variance of individual trees is lowered by constructing a prediction algorithm that averages the predictions of multiple, different regression trees. These different trees are constructed by bootstrapping or resampling the observations of the training set with equal probability and running the regression tree algorithm on each separate sample. More bootstrapped samples usually lead to a diminishing reduction in variance at the cost of higher computational time (Hastie, Tibshirani & Friedman, 2001). Bagged regression trees will be implemented in this thesis using the "ipred" package in R.

2.2.3. Random forests

While bagging greatly reduces the variance of regression tree-based prediction algorithms, the issue remains that the resulting trees from the bootstrapped samples are highly correlated. This is due to the fact that all trees calculated from the different samples still use the same set of P predictors to generate splits. A possible solution to this problem is the use of a random forest. A random forest is an ensemble prediction method that is a modification of the bagging algorithm. The goal of a random forest is to generate a large set of uncorrelated trees and finally, averaging the result. As the generated trees under a random forest are less correlated than under bagging, the variance of the resulting prediction method is expected to reduce (Hastie, Tibshirani & Friedman, 2001).

The uncorrelated trees in the random forest can be generated by the following procedure:

1. Generate a new training sample by bootstrapping from the training set.
2. From the predictors P in the bootstrapped sample, select m variables at random.
3. Use the m predictors to grow a regression tree on the bootstrapped sample.
4. Repeat steps 1-3 for the desired amount of trees.
5. Output the ensemble of trees and predict by averaging over all trees.

Just as with bagged regression trees, more trees in a random forest will generally lead to better predictions at a diminishing rate at the cost of higher computation time (Hastie, Tibshirani & Friedman, 2001). R's "randomForest" package will be used in this thesis to construct a random forest for the imputation of a branch's added value.

2.2.4. Boosted regression trees

Both bagged regression trees as well as random forests, introduced in the previous two sections, make predictions by resampling the observed data and averaging the result of a large number of simpler base methods (regression trees). In some manner, both methods can be regarded as "horizontal" ensemble algorithms, which average the results of similar base methods. Another class of ensemble algorithms takes a different approach. "Vertical" or additive ensemble learners combine the output of many "weak" learners in order to build a powerful prediction algorithm. Weak learners in the context of regression can be interpreted as prediction algorithms which error rates are only slightly better than an algorithm which just predicts the sample average. This leads to the idea behind boosting, a popular additive ensemble learning method, whose goal is to sequentially apply a weak learning algorithm such as regression trees to repeatedly modified versions of the data. This results in a sequence of weak learners which are added in a (weighted) sum to determine the final prediction of an observation (Hastie, Tibshirani & Friedman, 2001).

In the scenario of minimizing the total squared error loss ($\sum_{i=1}^{N}(Y_i - f(P_i))^2$) in the training set, a prediction method using boosting with regression trees can be built using the following simple procedure:

1. Set the number of estimators (trees) *m* as desired.
2. For the first tree, $k = 0$, construct a small regression tree using the predictors *P* and outcomes Y from the training set that minimizes the squared error loss $(Y_{i,0} - f_0(P_i)))^2$.
3. Calculate the current residuals $Y_{i,k} - f_k(P_i)$ for each observation where $f_k(P_i)$ represents the estimate of $Y_{i,k}$ using regression tree *k*.
4. Train a small regression tree $k + 1$ on the residuals from step 3 that minimizes the squared error loss and define $Y_{i,k+1} = Y_{i,k} - f_k(P_i)$.
5. Repeat steps 3 to 4 for each *k* ($k = 0, ..., m-1$) until the desired number of trees *m* are trained, training each tree on the residuals from the previous iteration.

6. Sum the predictions of each individual tree for a company in order to obtain its final prediction.

Boosting can easily be adjusted to optimize different performance criteria than the squared error loss and shrinkage/weighting of subsequent trees can be introduced to slow down learning and making the algorithm more sensitive to discovering local minima. Often the computational complexity increases with other performance measures and numerical optimization algorithms such as gradient descent are used to minimize the loss function/approach optimality (Hastie, Tibshirani & Friedman, 2001). Boosted regression trees can easily be implemented in R using the "gbm" package.

2.2.5. Bayesian additive regression trees

The final individual imputation technique employed in this thesis is another additive ensemble learner, namely: Bayesian additive regression trees (BART). BART can easily be implemented in R using the "BartMachine" package. The idea behind BART is to generate, just as with boosted regression trees, several different regression trees for which the individual predictions are added together in order to form a final prediction of Y or the value added of an individual company. The process of BART can be summed up in the following steps:

1. Set the number of trees *m* and the number of desired samples *t* from the posterior distribution in the algorithm as desired.
2. Initialize the *m* trees to single root nodes (no splitting variables, only a single value). As the total prediction is the sum over all *m* trees and all trees are single root nodes, the prediction associated which each tree is initialized to $\frac{\bar{Y}}{m}$, with $\bar{Y}$ being the average of the dependent variable (value added) in the training set.
3. Using a Metropolis-Hastings (MH) algorithm, generate a tree structure T_1^* for the first regression tree. The values to be modelled here are $R_{1,i} = Y_i - \sum_{j \neq 1} f_j(P_i)$, where the second term is the sum of current predictions of the other trees. The basic idea of the MH algorithm used here is create a Markov chain that converges to the posterior distribution by generating a new regression tree structure T_1^* from the previous regression tree T_1. This generated tree T_1^* is then accepted with a certain probability depending on the probability of observing R_1 under T_1 versus T_1^*, the probability of observing T_1, the probability of observing T_1^*, the probability of moving from T_1 to T_1^* and the probability of moving from T_1^* to T_1. This process is

repeated for all *m* trees in order to obtain a single draw of the posterior distribution (Yoayuan & Roy, 2019).

4. Repeat step 3 for the number of draws *t.*
5. Average the *t* predictions for an individual company in order to obtain its final prediction.

Just as other algorithms making use of Markov Chain Monte Carlo (MCMC), BART often makes use of a burn-in period, meaning that the procedure described in step 3 above is ran for several iterations before the outcomes are saved. This is done to make the outcome of BART independent of the starting position (initialized structures) of the trees or in other words, to obtain convergence to the posterior distribution. For extra information regarding the technical and mathematical details behind BART, the reader is encouraged to review the article "Bayesian additive regression trees and the General BART model" written by Yoayuan and Roy (2019). Other parameters to consider when building BART is the maximum depth/penalization of tree size.

2.2.6. Superlearner ensemble

In the context of this thesis all imputation techniques mentioned in the previous sections will be applied to estimate the value added of a branch and their aggregate in the presence of missing annual accounting data. However, it is possible to combine all these imputation techniques in a single imputation strategy. This can be accomplished using the "SuperLearner" package in R. This algorithm uses cross-validation to estimate the performance of several different individual modeling techniques. It then creates an ensemble algorithm consisting of a weighted average of the individual methods, giving high performance methods a higher weight in the overall ensemble (Kennedy, 2017). The Superlearner method can be summarized in the following steps:

1. Set the number of folds *V* to be used for cross-validation.
2. On each combination of *V-1* folds train all the imputation methods to be taken into consideration (OLS, bagged regression trees, random forest, boosted regression trees, BART). In this thesis individual methods were trained to minimize the total squared error loss.
3. Predict outcomes for each remaining validation fold using each candidate learner.
4. Use the sets of predicted outcomes for each learner as independent variables in a regression of the actual outcome under the condition that the coefficients sum to one

and are non-negative. This is done using non-negative least squares (Polley, LeDell, Kennedy, Lendle & van der Laan, 2018).

5. Train all the individual learners on the entire training set and use the coefficients found in step 4 as weights for each method in the total prediction. The total prediction can be seen as a weighted sum of the predictions of the individual learners (van der Laan, Polley & Hubbard, 2007).

The individual algorithms to be used as components of the Superlearner in this thesis are OLS, bagged regression trees, random forest, boosted regression trees and BART. Ten folds were used in the cross-validation.

2.3. Inductive Conformal Inference

Currently, the regression tree-based algorithms proposed in the previous section 2.2. as well as the Superlearner ensemble only yield point predictions of a branch's added value at a certain date. In order to get a better view on the prediction uncertainty of the different imputation techniques, prediction intervals are needed. In this section, the theory behind inductive or split conformal inference is explained briefly, together with an extension to create prediction intervals of the aggregate of individual predictions (total value added in an economic sector). The choice of using split conformal inference is made in this thesis as it yields reliable prediction intervals at relatively low computational cost in a similar manner for all imputation methods (Lei, G'Shell, Rinaldo, Tibshirani & Wasserman, 2018).

2.3.1. Individual prediction intervals

Consider data (Y_i, P_i), $i = 1, 2, \dots, N$, for which a regression algorithm is run to predict the mean of Y_i ($f(P_i)$), depending on a set of predictor variables (P_i) and for which a prediction interval needs to be constructed at a certain marginal coverage level $1 - \alpha$ ($\alpha \in (0,1)$). The method of split conformal inference accomplishes this under the assumptions of exchangeable samples, a symmetric distribution of Y_i around $f(P_i)$ and homoscedasticity. Samples (Y_1, P_1), …, (Y_N, P_N), are exchangeable if every permutation of the sequence of samples has the same joint probability distribution, a weaker assumption than i.i.d. samples (Shafer & Vovk, 2008). The theory behind split conformal inference is based on the idea that that the ranks of the absolute observed residuals $R_i = |Y_i - f(P_i)|$ are uniformly distributed over the set {1, 2, …, N} under the assumptions mentioned above. Having an empirical set of absolute residuals R,

one could easily rank the residuals and, knowing that these ranks are uniformly distributed, find the residual value *d* which corresponds to the $1 - \alpha$ percent highest rank so that:

$$P(R \leq d) \geq 1 - \alpha$$

Solving for *d* and relying on the symmetry assumption of $f(P_i)$, a prediction interval with a marginal coverage of 1 - α for a new observation *j* is then given by:

$$PI_{j,1-\alpha} = [f(P_j) - d; f(P_j) + d]$$

In order to obtain a set of absolute errors independent from the data used to train the regression algorithm, split conformal inference splits the training data set into two separate sets. One set for training the regression algorithm (I_1) and another set to calculate the absolute residuals (I_2). The split conformal inference algorithm can be summarized in the following steps:

1. Determine the marginal coverage level $1 - \alpha$ ($\alpha \in (0,1)$) of the prediction intervals.
2. Randomly split the *N* samples into two (equal-sized) subsets I_1, I_2.
3. Train the imputation method on set I_1.
4. Predict the outcomes $f(P_i)$ in set I_2 and calculate the corresponding absolute residuals $R_{i,I2} = |Y_i - f(P_i)|$.
5. Calculate *d* = the k^{th} smallest value in the set R_{I2} with k = $\lceil (N/2 + 1) * (1 - \alpha) \rceil$ in the case of an equal-sized split in step 2.
6. For each prediction *j*, the corresponding $1 - \alpha$ prediction interval is given by: $PI_{j,1-\alpha} = [f(P_j) - d; f(P_j) + d]$.

In the setting of this thesis, the split between I_1 and I_2 was performed to obtain sets of (approximately) equal size. Note that split conformal inference leads to prediction intervals of equal width for each prediction. This is due to the homoscedasticity assumption made in the beginning of this section. One could easily adjust the algorithm in the case of heteroscedasticity by modelling the absolute error size on the considered predictors P_i and implementing the theory of locally-weighted conformal inference (Lei, G'Shell, Rinaldo, Tibshirani & Wasserman, 2018). The implementation of this method is, however, beyond the scope of this thesis. This is not problematic as the benchmark OLS method of Vansteelandt, Coppens, Vackier et al (2019) used for comparison is developed in the setting of homoscedasticity as well. Finally, it should be mentioned that the use of split conformal inference could be detrimental to the predictive capabilities of an imputation method. Splitting the dataset in two

and training a regression algorithm on only half of the data likely results in less accurate predictions and higher errors. These higher errors can easily lead to wider prediction intervals than the analytically calculated prediction intervals resulting from a method trained on the full dataset. Split conformal inference can be implemented in R using the "conformalInference" library.

2.3.2. Aggregate prediction interval

While the method of split conformal inference could easily be implemented to generate marginal prediction intervals for the added value of individual companies in the presence of missing annual accounting data using any of the imputation methods described in this thesis, the problem of constructing a prediction interval for the sum of individual branches' added values or total value added remains. In order to generate a prediction interval of the total value added of missing companies, the following novel procedure, based on split conformal inference and bootstrapping, is proposed in this thesis:

1. Run step 1-4 of the split conformal inference algorithm for individual predictions from section 2.3.1. for the prediction method, evaluation date and company size (small/large) under consideration. This results in a set of absolute residuals R_{I2}.
2. Determine the number of individual branches *m* for which the added value has to be imputed.
3. Randomly sample *Z* times *m* residuals from R_{I2}, each time summing the *m* residuals $E_i = |\sum_{j=1}^{m} R_{j,I2}|$.
4. Rank the set of *Z* residuals *E* and calculate *d* = the k^{th} smallest value in the set *E* with k = $\lceil (Z+1) * (1-\alpha) \rceil$ and $1-\alpha$ the desired coverage level.
5. The prediction interval for the total added value of missing branches is then given by: $PI_{total,1-\alpha} = [\sum_{j=1}^{m} f(P_j) - d; \sum_{j=1}^{m} f(P_j) + d]$ with $\sum_{j=1}^{m} f(P_j)$ the sum of predictions.

In this thesis, the number of samples *Z* was 10000 and all prediction intervals were constructed at the 95% level. Note that the procedure outlined above needs to be implemented for each prediction method, at each evaluation date, for each company size (small or large) and suffers from the same drawbacks/assumptions as the split conformal inference method.

2.4. Outlier handling

When imputing the economic value added of a branch, issues could arise under any imputation method when several companies in the training set are much larger in terms of value added or possess extreme characteristics in terms of other predictor variables, therefore possibly influencing the method's fit. In this section, two possible mechanisms for dealing with these outliers are discussed. Both outlier handling mechanisms as well as the imputation methods that do not account for outliers were implemented in R in combination with all previously mentioned prediction techniques.

2.4.1. Size-based approach

The first outlier handling approach considered in the case of imputing a branch's economic value added in the presence of missing annual accounting data relies solely on the size of the companies that have to be imputed. More specifically, companies in the training set with a value added in the previous year ($Y_{(t-1)}$) larger than the maximum or lower than the minimum added value in the previous year of the companies in the set of companies to impute, are removed when training the imputation algorithm. As separate imputation methods are trained for large and small companies, determined by the corresponding reporting scheme, the outlier removal is done separately on both sets at each evaluation date. This size-based approach for removing outliers was also implemented for the OLS method of Vansteelandt, Coppens, Vackier et al (2019) which is used as the benchmark in this thesis.

2.4.2. Leverage-based approach

A possible drawback of the size-based approach is that it only considers size and no other characteristics such as business activities (NACE), number of employees (PE), etc., which could greatly differ between companies. The leverage-based approach tries to identify outliers as companies that are extreme in terms of all predictors. The leverage of a branch *i* in this thesis was calculated as the i^{th} diagonal element of the hat matrix $H = X(X^T X)^{-1} X^T$ with X the predictor matrix of the combined training and imputation set in the OLS scenario (Kutner, Nachtsheim, Neter & Li, 2004). A branch was considered to be an outlier if its leverage value h_{ii} exceeded $2 * \frac{p}{n}$, with p the number of predictors and n the total amount of

observations, a common used rule-of-thumb to determine high-leverage observations (Kutner, Nachtsheim, Neter & Li, 2004)

After high-leverage observations were identified, the training and imputation data could easily be grouped into 4 different sets. For the training data, sets T_1 and T_2 were created, with T_1 the set containing all companies of the training dataset and T_2 the training set with high leverage observations removed. The imputation set was divided into the sets S_1 and S_2 with S_1 the companies with missing annual accounting data and high leverage values and S_2 the branches with low leverage values ($h_{ii} \leq 2 * \frac{p}{n}$,). Set T_1 was used to train the prediction algorithm for S_1 and set T_2 for imputing the added values of branches from S_2.

3. Results

In the results section of this thesis, the performance of all imputation methods will be analyzed in the three different outlier scenarios described in the methodology section. The performance or uncertainty accompanying a prediction can be a assessed in multiple ways. In the context of this thesis, the performance of an imputation method for estimating the total economic value added in the Belgian port sector at certain evaluation date *t*, will be evaluated by the relative width *w* of its 95% prediction interval or:

$$w_t = \frac{\widehat{ub_t} - \widehat{lb_t}}{\widehat{Y_t}}$$

With,

$\widehat{ub_t}$ = upper bound of the 95% prediction interval for time *t*

$\widehat{lb_t}$ = lower bound of the 95% prediction interval for time *t*

$\widehat{Y_t}$ = Estimated total value added for time *t*

This width of the 95% prediction interval is expected to decrease at later evaluation dates. Small companies represented only a small percentage (2,81%) of the total value added in the Belgian port sector in the year 2015. Deviations in the predictions for small companies can therefore be seen as relatively less important than the ones for large companies.

All methods will be compared to the benchmark method from the paper of Vansteelandt, Coppens, Vackier et al (2019). This benchmark uses the OLS method with size-based outlier

correction for which the prediction intervals can calculated analytically or with bootstrapping (not using inductive conformal inference).

3.1. No outlier correction

For predicting the total value added in the Belgian port sector in the setting of no correction for outliers, figures 1, 2, 3 and 4 give the relative width of the 95% prediction interval for each imputation method developed in this thesis as well as for the benchmark, both for large (figure 1 & 2) and small companies (figure 3 & 4) in 2015.

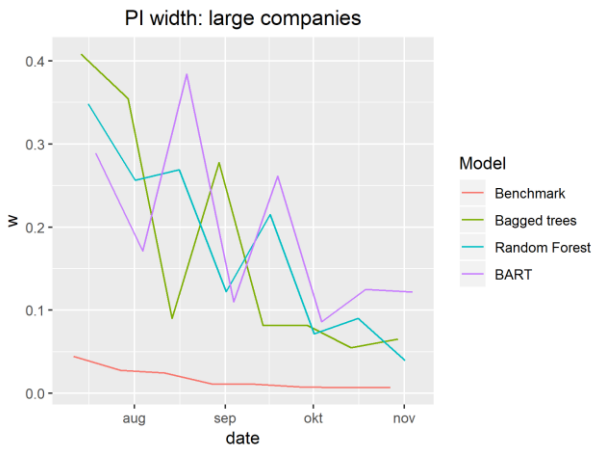

Figure 1: PI width large companies 2015 (Bagging, Random Forest & BART)

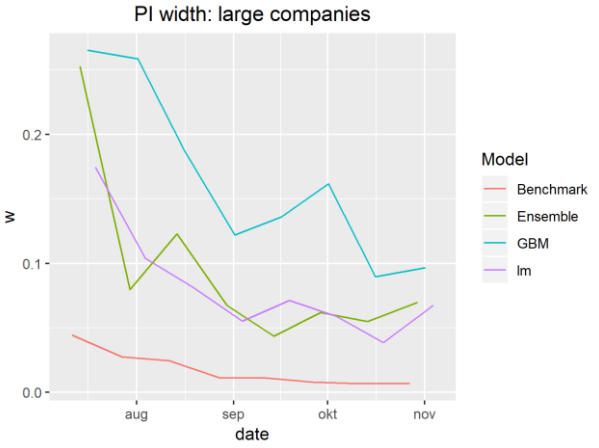

Figure 2: PI width large companies 2015 (Ensemble, GBM & lm)

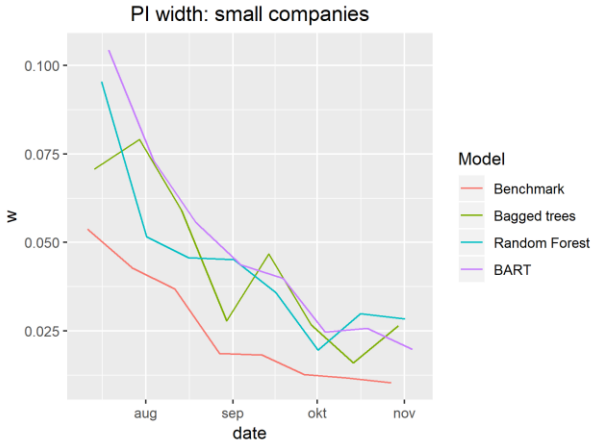

Figure 3: PI width small companies 2015 (Bagging, Random Forest & BART)

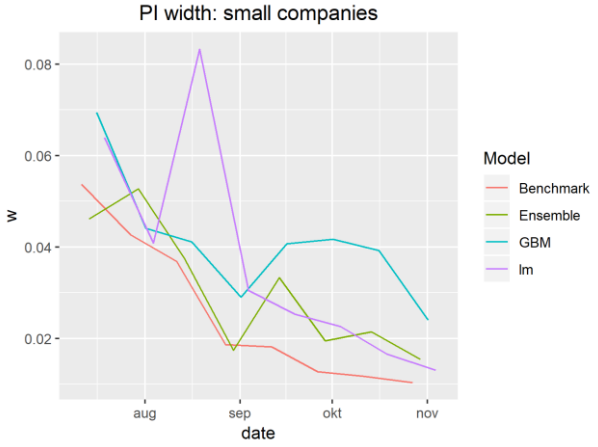

Figure 4: PI width small companies 2015 (Ensemble, GBM & lm)

Note that the benchmark model, OLS with size-based outlier adjustment, performs better than all other imputation methods in terms of 95% prediction interval width for nearly all evaluation dates. This result is observed for both large and small companies. Possible reasons for this

worse performance are not taking into account outliers in this setting for the non-benchmark methods, the superiority of the OLS method, or the fact that induced conformal inference requires the imputation algorithms to be trained on a smaller dataset than the benchmark model. In figure 5 and 6, the performance of the lm (OLS with split conformal inference) and ensemble (Superlearner) method are disclosed in more detail for respectively large and small companies in the no outlier correction setting.

In these figures, the 95% prediction intervals of each method are plotted as error bars alongside the corresponding point predictions at the different evaluation dates. The black line in each plot represents the actual added value of the companies in the Belgian port sector derived from the CBSO when all annual accounts were submitted. Note that the prediction intervals almost always contain the actual value, as expected in 95% of the cases.

Figures 5 and 6 show, just like the previous prediction interval width plots, smaller prediction intervals for the benchmark model with decreasing interval size at later evaluation dates. Note that while prediction intervals are larger for the non-benchmark methods, that the point predictions are all relatively close to the actual value and in the case of the ensemble method for small companies, even closer than the benchmark model.

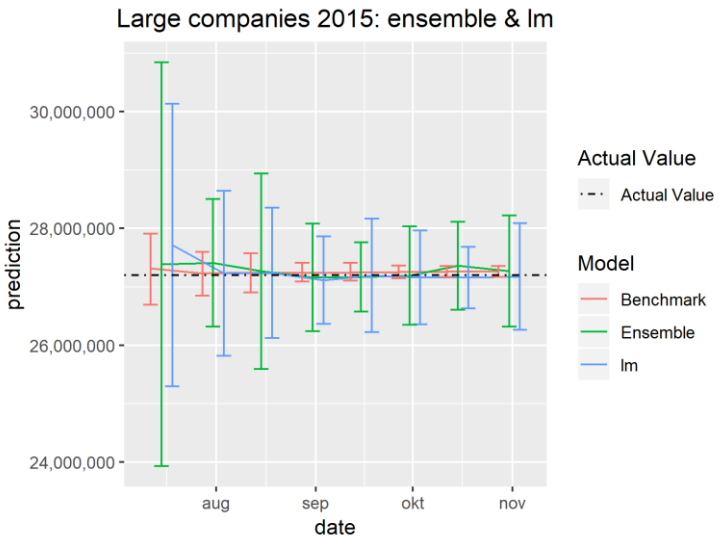

Figure 5: Error bar plot large companies 2015 (lm & Ensemble)

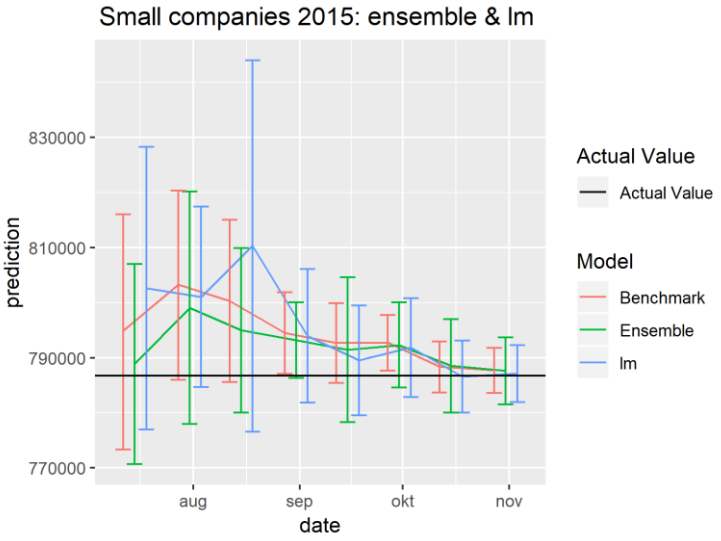

Figure 6: Error bar plot small companies 2015 (lm & Ensemble)

3.2. Size-based outlier correction

In this section, the performance of the prediction methods for which the size-based outlier correction approach of section 2.4.1 is applied, is analyzed in comparison to the benchmark. Remember in the previous results section, where the performance of the different imputation methods without outlier correction were compared to the benchmark, three different possible reasons were given for the underperformance of the different imputation methods. These reasons were: not taking into account outliers, the superiority of the OLS method or the smaller training sets used in split conformal inference. In this section, the same outlier handling technique as implemented for the benchmark is used. Underperformance can therefore only be attributed to an inferior imputation method or smaller training set. Furthermore, for the "lm" model, a wider prediction interval can only be caused by a smaller training set as the same OLS method and outlier correction method are used as the benchmark. Figures 7, 8, 9 and 10 plot the relative 95% prediction interval widths *w* for respectively small (figure 7 & 8) and large (figure 9 & 10) companies.

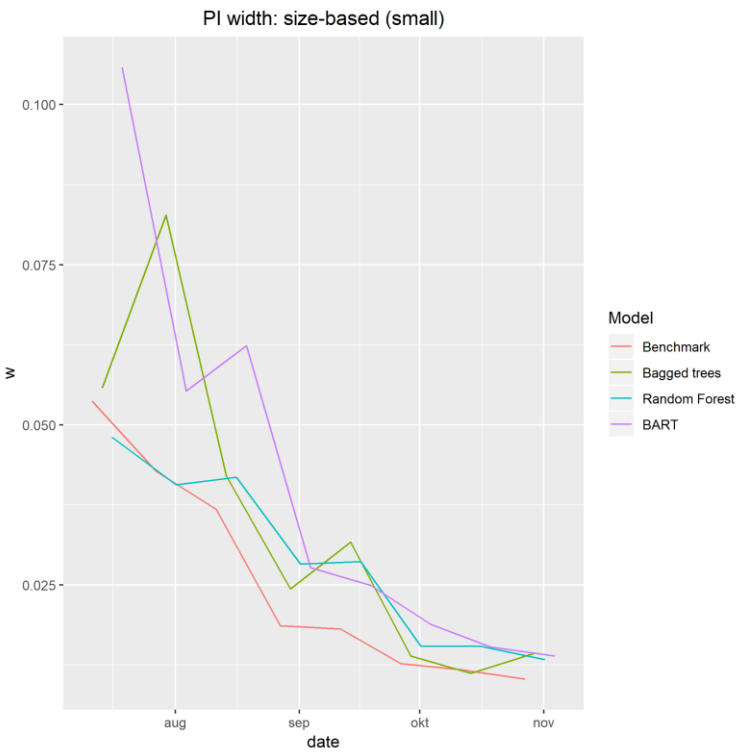

Figure 7: PI width small companies 2015 size-based (Bagging, Random Forest & BART)

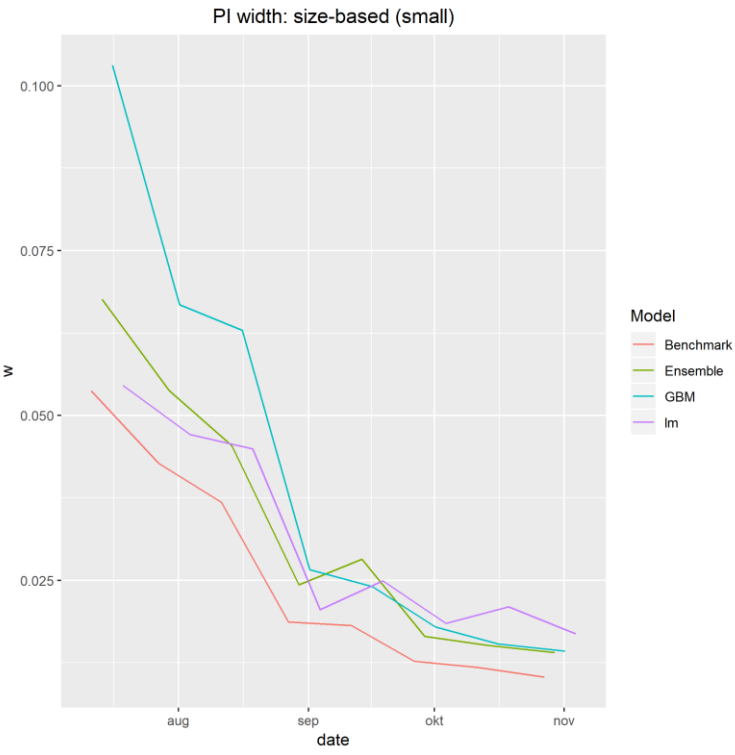

Figure 8: PI width small companies 2015 size-based (Ensemble, GBM & lm)

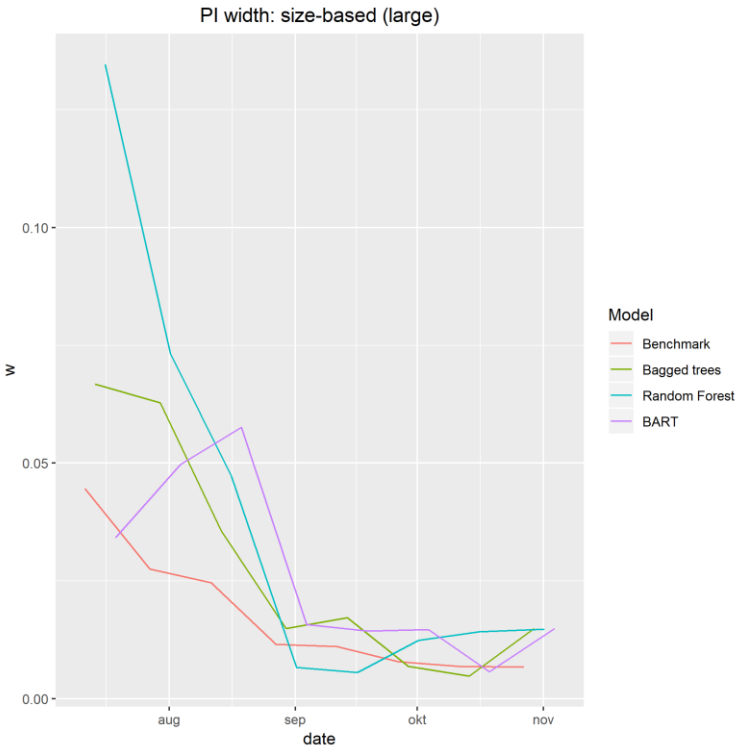

Figure 9: PI width large companies 2015 size-based (Bagging, Random Forest & BART)

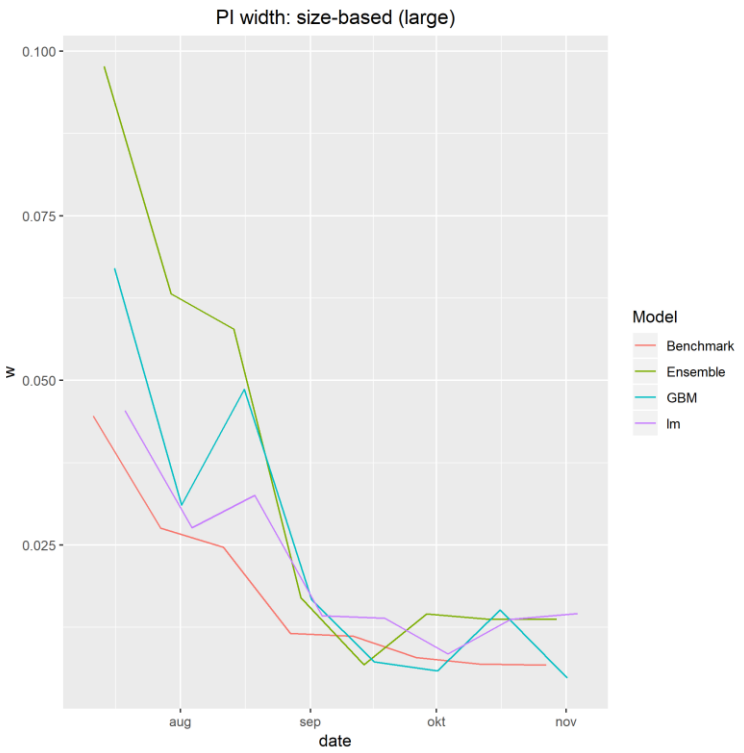

Figure 10: PI width large companies 2015 size-based (Ensemble, GBM & lm)

For small companies figures 7 and 8 show that the benchmark model still outperforms all other imputation methods based on relative 95% prediction interval width overall, especially at earlier evaluation dates with the exception of the random forest algorithm. The lm model performs worse than the benchmark at all evaluation dates, visualizing the effect on prediction interval width of the use of a smaller training set for small companies. Another interesting trend is the higher instability of relative prediction interval width over the evaluation period for other imputation methods. Non-benchmark methods often seem to experience sharper increases and decreases in prediction interval width at subsequent evaluation periods compared to the benchmark, which experiences a smoother drop in width. The non-benchmark methods that perform reasonably well in the size-based outlier setting from start to finish of the evaluation period are: the random forest, lm model and ensemble (Superlearner) method. Note that the interval width for these three imputation strategies are relatively close at all evaluation dates, suggesting no clear superiority of the random forest and Superleaner method over OLS for small companies.

While some methods perform better and others worse at certain evaluation dates for the small companies in the size-based versus no-outlier correction setting. Figures 9 and 10 suggest that for large companies, the companies that almost entirely determine the total value added in the

Belgian port sector, all imputation strategies perform clearly better with size-based outlier correction in terms of prediction interval width compared to the no-correction setting. This was not the case for small companies and can possibly be attributed to the larger variability in company size for the large companies.

The OLS benchmark still has better overall performance than the other methods, especially at earlier evaluation dates. At some evaluation dates, certain methods yield smaller prediction intervals but, note that these other methods also show high instability in prediction interval width at subsequent evaluation dates in contrast with the benchmark, which shows a strict drop in width.

A final finding is that while a random forest and the Superlearner method performed well for small companies, both methods perform worse compared to the other algorithms for large companies, especially at earlier evaluation dates. The methods that performed better for large companies overall were BART, lm, GBM and bagged regression trees. Figure 11 shows the error bar plot of these four methods together with the OLS benchmark. Point predictions all seem relatively close to the actual population value and intervals always cover the true population value. The OLS benchmark outperforms the other methods at almost all evaluation dates and shows a steady decrease in interval width. For the four other methods, all trained on a smaller training set under conformal inference, not a single method is clearly superior. At earlier evaluation dates, the lm model seems to be stable with relatively low prediction interval width, while at later evaluation dates, other methods sometimes yield smaller intervals.

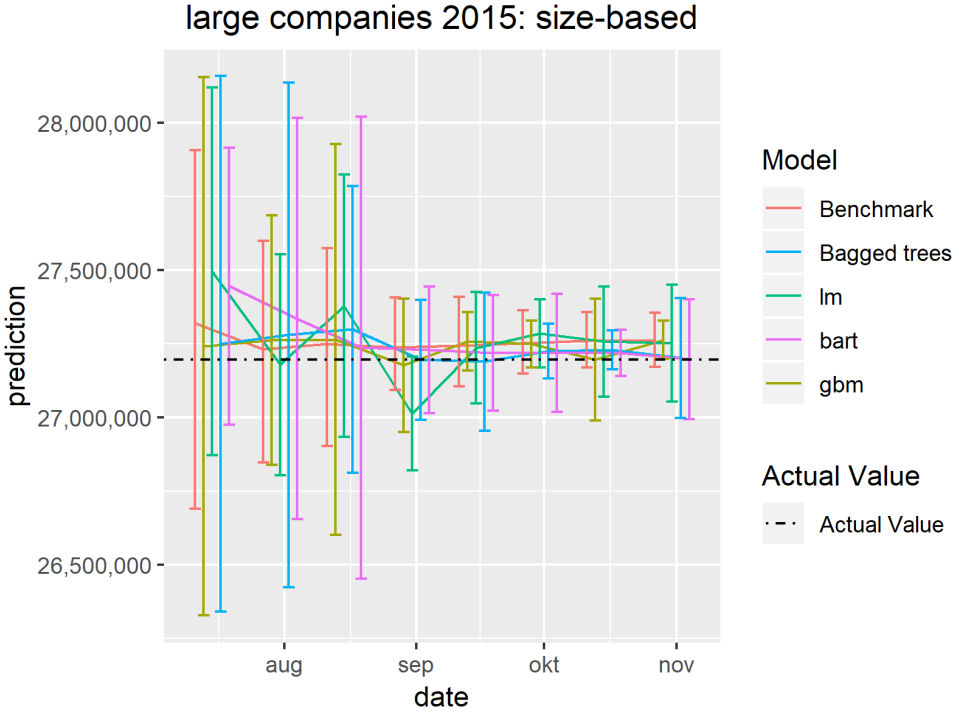

Figure 11: Error bar plot large companies 2015 (lm, bagging, BART & GBM)

As there is no clear difference between the performance of the lm model and other well-performing imputation methods for both small and large companies, one could argue that the lm model is preferable to predict the total economic value added in the Belgian port sector for the year 2015 if outlier correction is done using the size-based approach. This because prediction intervals for OLS, which is used for the lm model, can easily be calculated without the use of (split) conformal inference. This allows the lm model to be trained on the entire dataset, yielding more accurate predictions and the same results as the benchmark OLS method, which has been shown to be superior to all other analyzed methods in the setting of size-based outlier correction. Furthermore, OLS can be considered as a fairly simple, highly flexible and easy automatable estimation technique.

3.3. Leverage-based outlier correction

In this section, the performance of the imputation methods using the final outlier correction method covered in this thesis, the leveraged based approach, are shown. Figures 12, 13, 14 and 15 give the relative widths of the 95% prediction intervals for all imputation methods in both the setting of large (figure 12 & 13) and small companies (figure 14 & 15) for the total added value of the Belgian port sector in 2015.

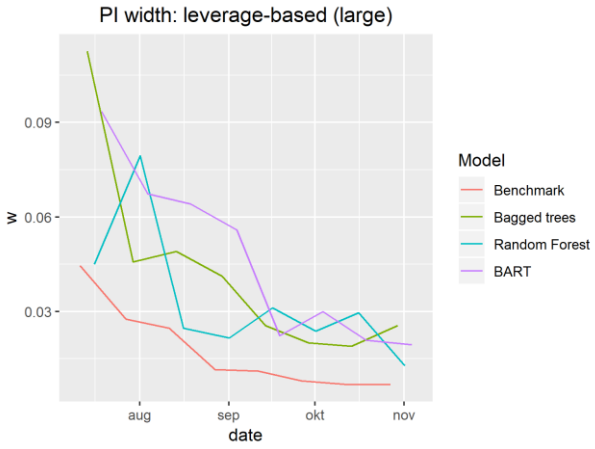

Figure 12: PI width large companies leverage-based 2015 (Bagging, Random Forest & BART)

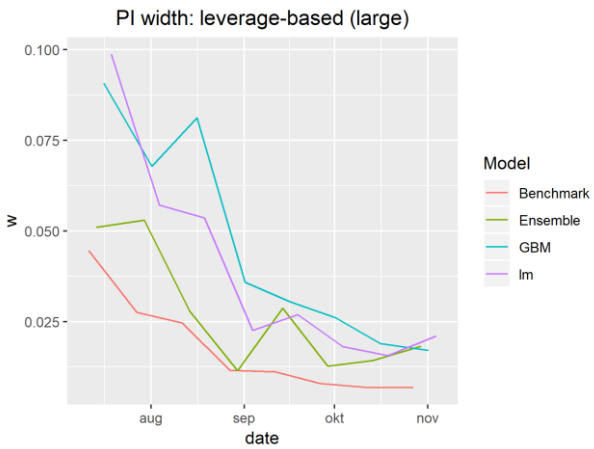

Figure 13: PI width large companies 2015 leverage-based (Ensemble, GBM & lm)

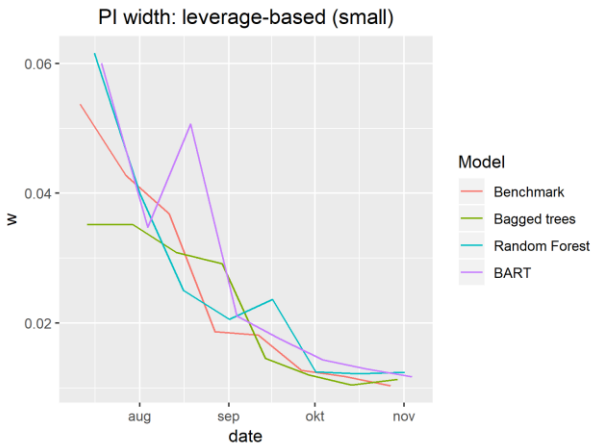

Figure 14: PI width small companies leverage-based 2015 (Bagging, Random Forest & BART)

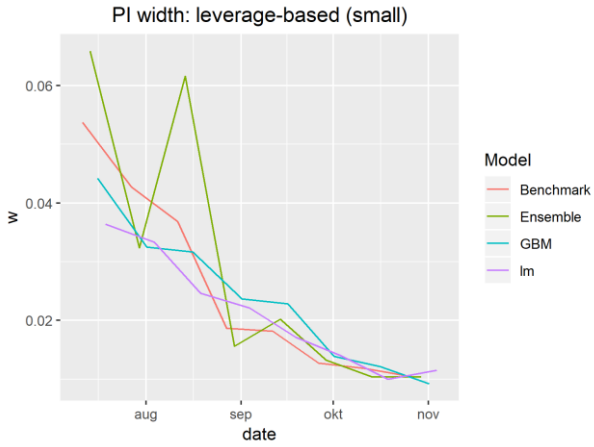

Figure 15: PI width small companies 2015 leverage-based (Ensemble, GBM & lm)

For the large companies, the benchmark OLS with size-based outlier adjustment still outperforms all other imputation methods, followed by the leverage-based Superlearner ensemble which seems to be more unstable with several increases in interval width for subsequent evaluation dates. Possible reasons why this method still performs worse than the benchmark are the use of a smaller training set in the split conformal inference framework or simply inferiority of the method compared to the OLS method. Another method that seems to perform better under the leverage-based approach is the random forest while BART, bagged regression trees, GBM and the lm model seem to perform worse than under the size-based approach.

Figures 14 and 15 depict the relative 95% prediction intervals for the imputation strategies combined with the leverage-based approach for small companies in 2015. At earlier evaluation dates GBM, bagged regression trees as well as the lm model perform better than the benchmark, after which, the relative 95% prediction interval widths of all methods converge. The lm model performing better than the benchmark, even with a smaller training set, is quite informative as both use the same underlying modelling technique. This means that for small companies, a leverage-based outlier handling method could improve benchmark results. A possible reason for the better performance of the leverage-based approach for small companies could be that small companies were more different in terms of other company characteristics than size alone as compared to the set of large companies.

3.4. Results comparison

Previous sections support the use of the benchmark OLS. No single other combination of imputation strategy with a particular outlier handling mechanism yielded better results for large companies. Reasons for inferior results were the use of a smaller training set needed to construct prediction intervals with split conformal inference, inferior imputation methods or, in the case of no/leverage-based outlier correction, the use of an inferior outlier correction method. Furthermore, for large companies no clear difference can be observed between BART, bagged regression trees, GBM, lm model (size-based adjustment) and the Superlearner ensemble (leverage-based adjustment), which were the methods that performed best not taking into account the benchmark. In other words, the complex methods introduced in this thesis did not clearly outperform the simple linear model for large companies.

For small companies, the benchmark model performed relatively well, only being outperformed at earlier evaluation dates by bagged trees, GBM and the lm model combined with the leverage-based approach. This signifies superiority of the leverage-based approach over the size-based approach for estimation the total economic value added for small companies as the lm model (OLS), even with a smaller training set, outperforms the OLS benchmark using the size-based approach. This outcome could be due to the larger diversity in company characteristics for small companies. Figures 16 and 17 visualize the results for the lm model under the different outlier scenarios for large and small companies respectively.

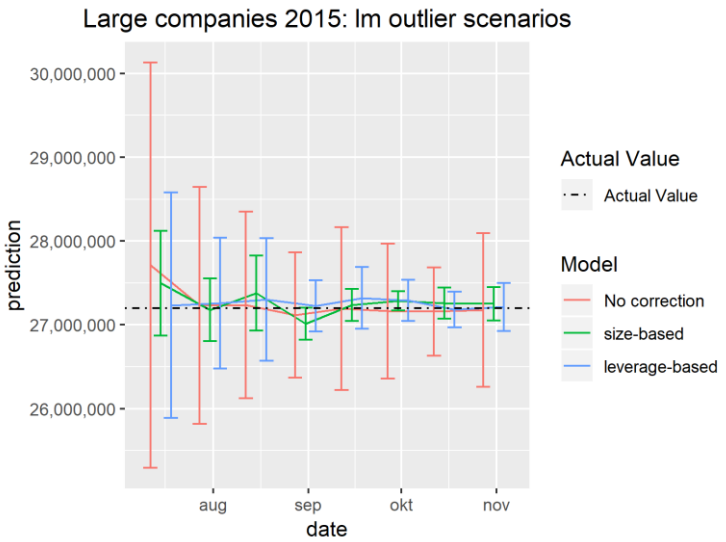

Figure 16: Error bar plot large companies 2015 under different outlier scenarios (lm model)

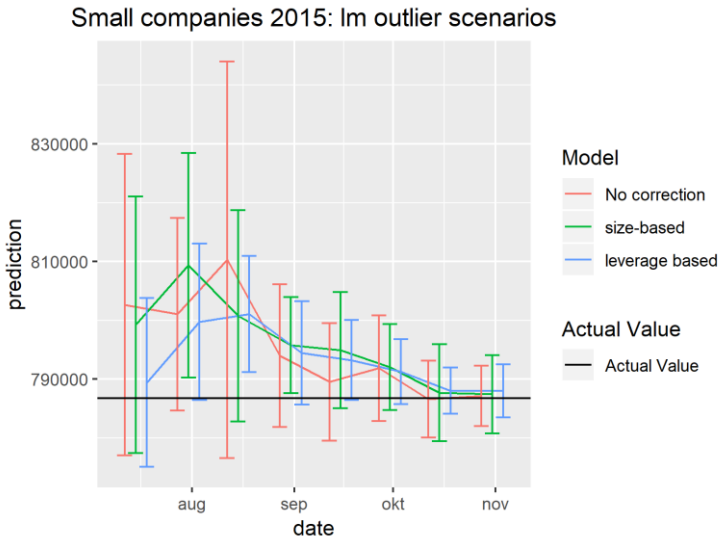

Figure 17: Error bar plot small companies 2015 under different outlier scenarios (lm model)

From figure 16 it can easily be deduced that the size-based approach leads to smaller prediction intervals for large companies, while figure 17 confirms the superiority of the leverage-based approach for small companies under OLS. Applying this method to the full dataset, in other words, without split conformal inference, could therefore be beneficial for the prediction of total economic value added for small companies.

In summary, the methods developed in this thesis did not outperform the benchmark for large companies simply due to the OLS method yielding better results or the necessity to split the training set in order to build the prediction intervals with split conformal inference. For small companies, some methods performed slightly better at earlier evaluation dates, possibly due to the use of the leverage-based instead of the size-based outlier correction method. The OLS method can therefore be identified as an efficient method for estimating the total value added for companies in the Belgian port sector in the presence of missing annual accounting data. This method had comparable or even better performance than the regression-tree based methods introduced in this thesis while maintaining interpretability, easy automatability, low computational requirements and still being easy to adapt to other sectors/economic variables.

4. Discussion

In this thesis, an alternative framework to the OLS method with size-based outlier adjustment, proposed by Vansteelandt, Coppens, Vackier et al (2019), for estimating the added economic value of an individual branch and its aggregate in the presence of missing annual accounting data was developed and analyzed. Research was carried out using a dataset of the Belgian port sector in 2015, for which predictions were made at several different evaluation dates. The main goal being to obtain an estimation framework that is easy to implement, easy to automate, relatively simple and generalizable to other sectors/economic variables while yielding accurate predictions. The developed methods all relied on the same testable missing at random assumption as the earlier proposed OLS method as well as exchangeability, a symmetric distribution around the estimated values and homoscedasticity.

The regression algorithms considered for estimating the added value in this thesis were OLS, bagged regression trees, random forests, boosted regression trees, Bayesian additive regression trees (BART) and an ensemble "Superlearner" method which combines all the aforementioned algorithms. The performance, in terms of relative 95% prediction interval width of the estimated total economic value added, of all these methods was investigated in a setting with no correction for outliers, a size-based approach to deal with outliers, similar as the one implemented by Vansteelandt, Coppens, Vackier et al (2019), and a leverage-based outlier handling mechanism. Prediction intervals were obtained using a combination of split conformal inference and bootstrapping. The main drawback of this method being that it requires a split of the training set, yielding less available observations for training the regression algorithms. This results in less accurate predictions/wider prediction intervals compared to an estimation method that can be trained on the entire dataset, obtaining prediction intervals analytically/without split conformal inference.

In terms of performance, slightly different results were obtained for the added value of large and small companies in the Belgian port sector for the year 2015. For large companies none of the developed methods in this thesis outperformed the OLS with size-based outlier adjustment benchmark, trained on the entire dataset. Furthermore, no clear difference was observed between BART, bagged trees, GBM, lm model (size-based adjustment) and the Superlearner ensemble (leverage-based adjustment), which were the methods that performed best after the

benchmark. The OLS method with split conformal inference (lm model) performed as well as the other, more complex methods for large companies. For this reason, OLS with a size-based outlier correction can be seen as a viable, well performing imputation strategy for estimating the total value added of a sector, especially when knowing that OLS could easily be trained on the entire dataset, calculating prediction intervals analytically (without split conformal inference), resulting in the clearly superior benchmark results.

Small companies only account for a small percentage of the total value added in the port sector. In 2015 this was approximately 2,81 percent. Deviations in predictions for small companies can therefore be seen as relatively less important in the setting of the Belgian port sector as these only account for a small part of the total economic value added. For the small companies, the benchmark model performed worse at earlier evaluation dates than bagged regression trees, GBM and the lm model, all three using the leverage-based outlier adjustment. This supports the use of the leverage-based instead of the size-based approach to estimate the total economic value added for small companies. This because the lm model (OLS) performed better than the benchmark, even with a smaller training set. Applying the leverage-based outlier adjustment to the benchmark dataset could therefore lead to improvements in added value estimation. A possible reason for the leverage-based approach yielding better results in this setting could be the larger diversity of small companies in terms of characteristics other than size.

In summary, no clear evidence was found that any of the developed methods performed better than OLS for estimating the total added value of the Belgian port sector in terms of relative 95% prediction interval width. In contrast, the OLS benchmark with size-based outlier adjustment outperformed all other methods when estimating the total value added for large companies in the Belgian port sector. For small companies, the benchmark model performed well but improvements could be made by combining OLS estimation with leverage-based outlier adjustment. Finally, the OLS method is less complex than the other developed methods, easy to generalize to other sectors and predictor/outcome variables and easy to automate and implement as a general estimation framework. In other words, OLS is a viable, well-performing imputation strategy for estimating the economic value added in the Belgian port sector.

Just as any study, the research performed in this thesis has several limitations. First of all, the performance of the different methods was only analyzed in the setting of the Belgian port sector in 2015. There is no guarantee that the OLS benchmark performed better than the other methods

for different years/sectors/other economic variables of interest. This could however be analyzed in future research with the code provided in the R appendix and a different dataset. Second, except for the OLS and Superlearner method, only regression-tree based machine learning algorithms were considered. It could be that other methods such as spline or support vector regression (SVR) perform just as well or even better than the benchmark. Third, better performance of the analyzed algorithms could possibly be obtained by tuning their hyperparameters using i.e. grid searches. The performance of these algorithms with different sets of hyperparameters was thus not investigated. One could argue however, that this would greatly increase the complexity of the estimation methods, requiring different optimization for each different sector, year and economic variable of interest and is not really necessary as the OLS method performs well while still being easy to implement and understand. Fourth, in this thesis split conformal inference was used to determine the 95% prediction intervals which requires splitting the training set. Regular conformal inference, which does not require this split, could have been implemented. This was impossible in this thesis for computational reasons. Furthermore, in this thesis a 50-50 split was used for split conformal inference. One could experiment with different proportions for better results. Finally, for the leverage-based outlier adjustment, only one leverage cut-off level was considered. Other cut-off levels could possibly lead to better results.

Possible avenues for future research are the evaluation of the performance of the OLS method with size-based outlier correction for other sectors or economic variables, possibly comparing the performance of OLS with the methods developed in this thesis in these other settings. Furthermore, as better results were obtained for small companies with the leverage-based approach, further development of this approach, testing it in other settings and for other cut-off levels could form an interesting topic for further research. Finally, investigating the performance of methods such as SVR or spline regression could form an interesting research question. These methods could easily be implemented in the split conformal inference framework of thesis.

5. Reference list

Gujarati, D., & Porter, D. (2009). *Basic econometrics* (Fifth edition ed.). New York: McGraw Hill.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning* (Second Edition ed.). New York: Springer.

Kennedy, C. (2017). Guide to SuperLearner. Retrieved from https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models* (Fifth ed.). New York: McGraw-Hill.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association, 113*(523), 1094-1111.

Mathys, C. (2017). *Economic importance of Belgian Ports: Flemish maritime ports, Lige port complex and the port of Brussels,* Report 2015, NBB Working paper Series, No 321, June 2017.

NBB Central Balance Sheet Office (2019). Central Balance Sheet Office: filing. Retrieved from https://www.nbb.be/en/central-balance-sheet-office/filing

Polley E, LeDell E, Kennedy C, Lendle S, van der Laan M. *Package 'SuperLearner'*. 2018. https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf.

Rubin, D.B. (1976). *Inference and Missing Data. Biometrika*, 63, 581-592.

Shafer, G., & Vovk, V. (2008). *A tutorial on conformal prediction*. Journal of Machine Learning Research, 9, 371-421.

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1).

Vansteelandt, S., Coppens, F., Vackier, M., Reynders, D., & Van Belle, L. (2019). *Estimation methods for computing a branch's total value added from incomplete annual accounting data.* Ghent University.

Vincent Tan, Y., & Roy, J. (2019). *Bayesian additive regression trees and the General BART model*.

GHENT UNIVERSITY

FACULTY OF SCIENCES

Computing a branch's total added value from incomplete annual accounting data using machine learning techniques

Tom Marchal

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Stijn Vansteelandt
Tutor: François Coppens
Department of Applied Mathematics, Computer Science and Statistics

Academic year 2018 - 2019