Inhaltsverzeichnis

[Introduction 1](#_Toc174353948)

[Literature review 2](#_Toc174353949)

[Role of uncertainty in marketing 2](#_Toc174353950)

[Methodological literature overview 3](#_Toc174353951)

[The importance of prediction intervals 4](#_Toc174353952)

[Methods to derive prediction intervals 5](#_Toc174353953)

[Measures to assess reliability and sharpness 8](#_Toc174353954)

[Applied methods 10](#_Toc174353955)

[Bootstrap 10](#_Toc174353956)

[Mini Bootstrap / Ensemble 11](#_Toc174353957)

[Bayesian method 11](#_Toc174353958)

[Quantile regression 11](#_Toc174353959)

[Conformal prediction 12](#_Toc174353960)

[Conceptual comparison of the methods 14](#_Toc174353961)

[Results 14](#_Toc174353962)

[Methods’ Performance - Benchmarking Bootstrap 14](#_Toc174353963)

[General performance 14](#_Toc174353964)

[Deeper insights into and visualizations of methods’ performances 17](#_Toc174353965)

[Performance over varying training and prediction periods 20](#_Toc174353966)

[Application in marketing 21](#_Toc174353967)

[Managerial / research discussion 23](#_Toc174353968)

[Conclusion 23](#_Toc174353969)

[References 23](#_Toc174353970)

[Appendix 23](#_Toc174353971)

# Introduction

The objective of every company is to maximize revenue and profits by using the least input possible. Hence, departments inside companies must compete for scarce resources, and so does marketing. Marketing managers have to provide success and justify their budget and their expenditures, e.g. for campaigns or special discounts. A good resource allocation is therefore vital and to achieve this, customers, their potential and uncertainty must be evaluated as accurate as possible. The Customer Lifetime Value, CLV, is a well-known concept and gives information about the potential revenue (and profit) to expect from a customer. However, it is unknown by definition and must therefore be modeled, e.g. with the so-called pnbd (Pareto Negative Binomial Distribution) model. As it is usually the case for models, it is subject to uncertainty in its point predictions what puts marketing managers at risk when blindfold trusting these predictions.

This work will address this problem by introducing established methods to derive prediction intervals to the pnbd context and benchmarking them against the well-known Bootstrap method. The benchmarking will be done via several key metrics that are calculated for all methods for real-world data sets. The goal is to identify strengths and weaknesses of every method and assess their usefulness and reliability in practice. A special focus will be laid on Conformal prediction as a relatively young method that has gained a lot attention recently. As a final step, it will be assessed if the derived prediction intervals can help identify especially valuable customers, beyond assessing the uncertainty that is connected with their point forecasts.

This thesis can be seen in both, research and a managerial context.

# Literature review

## Role of uncertainty in marketing

To get a broader view of uncertainty more general in marketing, the notations of “uncertainty” and “marketing”, both shall be defined first. While there is no generally recognized definition of these notations, there are various attempts to provide a definition. Following Collin’s dictionary, then “Uncertainty is a state of doubt about the future or about what is the right thing to do.”[[1]](#footnote-1) Hubbard states in the context of business “The lack of complete certainty, that is, the existence of more than one possibility. The “true” outcome/state/result/value is not known.”[[2]](#footnote-2) Both definitions agree on the presence of unknown information with respect to a current or potential future state and connected actions (to be taken). On the other hand, there is a similar situation for the definition of marketing, the AMA (American Marketing Association) defines it as follows: “Marketing is the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large.” [[3]](#footnote-3) what makes it a very broad field but with a very rough focus on the placement of offerings to clients. The following paragraph shall introduce why uncertainty plays a vital role in marketing and why it needs to be taken into consideration.

As its name suggests, Marketing is concerned with the placement of offerings on a market and here is where uncertainty comes into play because an important aspect for marketing is the uncertain prospective demand. A market is a place which is heavily concerned with and driven by the actions of its agents, i.e. “customers, clients, partners, and society at large” (AMA), and their coordination.[[4]](#footnote-4) All of those agents come with uncertainty in their actions as they are ruled by human beings who make (ir-) rational or at least (un-) predictable (15RU) decisions, let it be a new product launch, the choice of a campaign, the location of new branch or simply a consumer’s unawareness of a competitor’s product which might be superior to their usual choice. Besides those human-driven uncertainties, estimating demand and placing offerings successfully in the market is affected by some additional dimensions of uncertainty, e.g. own product quality (16RU) that may vary by changing quality of delivered feedstocks. Competitors may bring unexpected technical advancements or the economic situation for the own product can change due to political conflicts and newly imposed taxes, to the good and bad, both. (17RU, 18RU) Also, if a campaign launched successfully in one region in the world does not imply that it would also work out in another (19RU, p. 2). This list could be continued nearly infinitely but should be sufficient to give motivation to why one wants to consider uncertainty in marketing decisions. It is obvious that from a marketing perspective, it is desirable to keep uncertainty as low as possible to make best decisions. Therefore, being able to understand the uncertainty in the specific case, i.e. identify its sources and quantify its amount is vital. An established approach in numerous contexts across science is to make predictions about the future by constructing models which depict a picture of reality, incorporating important aspects and leaving out unimportant ones for simplification. One example is the pnbd (Pareto Negative Binomial Distribution) which aims to predict future customers buying behavior and is the underlying model in this work.

## Methodological literature overview

To continue the thought of the previous paragraph and connect it with the context of this work, first it is to determine where uncertainty may arise in the CET prediction.

A diagram of a comparison between two different types of graphs

Description automatically generated with medium confidenceThe sources of uncertainty in models in general can be divided into aleatory and epistemic uncertainty (21RU) which also applies to the pnbd model. As there is, to my knowledge[[5]](#footnote-5), no clear, distinct definition of these two notations, the idea behind shall be briefly explained. Aleatory uncertainty refers to uncertainty coming from random events (20RU). It captures noise in the inherent observations and is therefore input dependent (21RU). The uncertainty that comes from inside the model, i.e. the parameters, is called epistemic uncertainty (21RU). In addition, it “[…] captures our ignorance about which model generated our collected data” (21RU, p. 2). Aleatory uncertainty cannot be reduced by e.g. collecting more data while this would be possible for epistemic uncertainty (21RU).

In the following, the sources will be discussed in more detail, starting with the aleatory side. Influences found in the literature that increase the uncertainty of customer behavior and hence the CLV do so because they are not fully considered in the model. Examples are campaigns of competitors, marketing contacts (in the past, presence and future) and state of the economy in a sense that people change their consumption behavior between recession and boom times. What is also a dominant issue that creates uncertainty is the possibility of a customer leaving the company forever either to switch to a competitor or stop consuming. The probability of being “alive” is included in the model but still, most customers won’t notify the company when they churn, so it stays a mere probability. The second part considers epistemic sources. N*ote that the papers quoted here are not necessarily concerned with CLV estimation but treat forecasting models in general or in other contexts, often time series or wind/energy forecasting. Nevertheless, since the pnbd model suffers from similar issues, these aspects are relevant here as well.* Especially often addressed in the literature is the parameter estimation which comes with uncertainty. This issue will be addressed later in this work. Also often mentioned is data uncertainty. This part is to locate on the intersection because errors can appear in the data collection and processing (aleatory) and there is a of lack of knowledge about data errors and potential biases (epistemic).

### The importance of prediction intervals

With these points raised, it is evident that mere point forecasts will be in most situations an insufficient indicator about the future values as they do not provide information about uncertainty (2PI,6PI,9PI,12PI). Hence, point estimates are often accompanied or even replaced by so-called confidence intervals (for e.g. parameter estimation) and prediction intervals in the context of forecasts. (5PI) The advantages are discussed in the following. An interval (forecast) is offering a range of possible values of future outcomes. (11PI) This means that the true value of the prediction will fall into this declared interval with a specific probability of p%. (4PI) point out 4 main points, why interval forecasts, and therefore PIs as well, are of such importance.

1. They “assess future uncertainty” (4PI, p. 476)
2. They enable the user to plan “different strategies for the range of possible outcomes” (4PI, p. 476). This means that one can prepare a strategy in case a high value inside the interval is realized and one for a low value, or the interval is so narrow and reliable that one can be sure with 95% that one specific strategy will be appropriate. In the context of CLV, it could help discover customers with high variability in their CLV and therefore target them especially. There are 2 rationales behind this approach: First, (16L) state that there is often a right tail distribution for CLV, and it hence makes sense to target a customer with high variability to realize that potential. Second, (33L) state one should focus on those as it offers the opportunity to learn and reduce uncertainty.
3. They “compare forecasts from different methods more thoroughly“ (4PI, p. 476). This means that PIs provide information about the reliability of each method what can be valuable when choosing methods for specific situations.
4. PIs” explore forecasts based on different assumptions more carefully“ (4PI, p. 476). When there is for example a method that assumes normal distribution, and one that is similar but does not make this assumption, and they produce different interval lengths, one might want to re-assess a model’s assumptions.

Another point, made by **(32PI)** “forecasts cannot be expected to be perfect, and intervals emphasize this” (alternatively: (2PI) 3.5.) which underlines maybe the most important power of PIs, namely pointing out to the user of forecasts that they are most probably wrong and hence treat them appropriately. Thinking one step beyond PIs, a more sophisticated option are density predictions, which are comparable with PIs, but they assign probabilities to each area in the interval and provide even more information about uncertainty. (6PI)

### Methods to derive prediction intervals

As the importance of PIs is justified now, it shall be introduced how they are obtained. First, it is important to note that different models and contexts can require different methods to derive PIs. In this work, the focus will be put on 33PI, who suggest 4 big classes of methods applicable in regression contexts: Bayesian approach, Ensembles, Direct interval estimation and Conformal prediction. These methods will be introduced in general in this chapter, explained with the concrete implementation in the CLV context in the next chapter and benchmarked against each other. The ensemble method will be implemented in connection with bootstrap which picks up the idea of ensemble methods but is too far away to be purely called “ensemble”. In addition, a pure Bootstrap method is used as benchmark. A special emphasis will be put on conformal prediction due to its recent raise in attention in the statistical community.

#### Bootstrap method

The Bootstrap method is the method that work as a benchmark in this work. In this chapter, it will be explained in general before going into the concrete application for the CLV context.

Bootstrap is a non-parametric and powerful approach to estimate statistics like a mean or quantiles of distribution and therefore as well PIs. The general approach to conduct a bootstrap is as follows.

1. From a sample of data of size n, draw n times with replacement
2. Repeat 1. A sufficiently often, e.g. 1000 times
3. For each of these new 1000 samples, calculate the desired metric.
4. From this distribution of the metric, take the central 90% of predictions. This is the desired interval.

The central assumptions for bootstrapping are the following: The initially sampled data, from which the new samples are created, must be representative for the whole population and independent from each other, i.e. they must be i.i.d.

#### Mini Bootstrap / Ensemble

Ensembles are in general a very straightforward method to derive prediction intervals. They can be described as follows: “An ensemble is a collection of a (finite) number of neural networks or other types of predictors that are trained for the same task. A combination of many different predictors can often improve predictions […]” (58PI, p.190) When one has enough fitted models and therefore enough point predictions, one can construct naïve prediction intervals (45PI) or calculate mean and variance, assume a (normal) distribution and derive PIs by calculating the respective z-values for desired quantile (33PI). 57PI suggest a special form of this approach, without referring to ensembles, which is conducted as follows.

1. Fit 1 model that has several parameters
2. Derive the covariance matrix of the parameters of this fitted model
3. Derive a large number of parameter combinations that have the characteristics described in the covariance matrix (one has to make assumptions about the distribution)
4. Treat these parameter combinations as independent models and make predictions with these models for each record
5. Take the naïve prediction intervals for each record

This approach has significant similarity with the previously described bootstrap approach and therefore, similar performance is to expect. In contrast, only 1 model is fitted and from there, all other models are derived by simply simulating parameters what makes it computationally more attractive.

#### Bayesian method

The roots of Bayesian statistics go back to the 18th century, to Thomas Bayes, as the name suggests (59PI). Who was the first one to make use of this approach to construct PIs is hard to tell but one of the pioneers in this field was Aitchison in 55PI who introduced the idea of using the Bayesian Approach’s strength in forming tolerance regions. The idea is to derive a probability distribution over the parameter(s) and based on this, derive a distribution of the outcomes and take the desired statistics. The process is described in the following and is based on 33PI, 51PI and 55PI.

1. Useful to have information about the parameter probability distribution (prior distribution) before the parameters are observed (or use an uninformative distribution) 51PI
2. Necessary to have a likelihood function that describes how likely it is to observe the data that are revealed step by step under the current parameter distribution 33PI
3. As more information (data) is revealed, update the prior parameter distribution with the new information to derive the posterior parameter distribution, using Bayes’ rule 33PI
4. Predict the outcomes based on the posterior parameter distribution to get a distribution of outcomes (predictive distribution)
5. Calculate the intervals based on this outcome distribution 33PI

Therefore, the Bayesian Approach is a method to estimate parameters and at the same time delivers a distribution of outcomes from where one can derive the prediction interval. Both, the posterior parameter distribution and the posterior prediction distribution, can be retrieved approximately by applying the Markov Chain Monte Carlo Method.

#### Quantile regression

Quantile regression is the last method that shall be used in the course of this work. It was first introduced by 54PI in 1978 and is a form of direct interval estimation what means a method that does not model a distribution of outcomes but is designed to directly output an interval. (33PI)

Following the process described in (33PI, 47PI, 54PI), Quantile regression does so by optimizing the parameters of a model with respect to a loss function that is employed while fitting. The idea is that one combination of parameters yields a specific number of overestimations and underestimations of outcomes. Aiming e.g. for a symmetric distribution would mean 50% overestimations and 50% underestimations. 47PI state that this is possible with any quantile other than the 50% quantile as well, as the same principle applies: The loss function penalizes deviation of the desired above-below ratio, which is 1:1 in the 50% case. Targeting intervals, e.g. a symmetric 90% interval, would require finding the 5%- and 95% quantile. Therefore, two parameter combinations must be found. Applying the same principle as above, the loss function would penalize according to the desired quantiles, yielding the parameter combinations that come closest to the objective.

This procedure requires interrupting and changing the model fitting procedure. As this would be out of scope of this work, a modified version will be implemented for the CLV context in the Applied Methods Chapter that keeps the core idea but simplifies the procedure. A step-by-step guide for the implementation will be provided in this later chapter.

#### Conformal prediction

Conformal prediction, or Conformal Inference is a relatively young method to derive prediction intervals with attractive empirical guarantees and few assumptions about the data and the model form to which it is applied. (1CP) It was first introduced by 23CP and 24CP in 1999 and gained a lot of attention in recent years. Conformal prediction has two main forms of implementation, Full conformal prediction (Transductive Conformal Prediction) and Split Conformal prediction (Inductive Conformal Prediction) where Full CP has been developed first and Split CP has emerged as an important special case (1CP) after being initially introduced by 26CP in 2002. The importance of the split version comes from the high computational costs associated with the full version but also sacrifices statistical efficiency (1CP). In this work, the focus will be exclusively put on the split version, because of the mentioned computational efficiency but also, which is crucial, because Full CP is not applicable for the pnbd-model. The reason for this shall be briefly outlined with a basic example outside of the pnbd-context, before going into detail with the applied split version.

Following 1CP and 9CP, full conformal prediction is implemented as follows: Assume there are 250 records of a) predictors Xi = 1:250 and b) observed outcomes Yi = 1:250. From these 250 records, 1 record is taken out. Assuming to not know what the true Y for this record is, one can only state that this outcome lives in the range of all possible future outcomes **Y.** The approach is to take n values as possible outcomes and reunite each of these invented records with the 249 unchanged records, ending up again with n sets of records. For each set, a new model is fitted what is computationally costly. Predicting with each of these different models the value that was left out, and applying a score function to this outcome, one ends up with n score values. From here one would go on and create prediction intervals. But in the context of the pnbd-model, it is not possible to continue because this model does not consider the true outcomes when fitting the model, as it is exclusively focused on the purchase history of a customer. Therefore, fitting n models by supplying n different outcomes for Y would not lead to different models and would not allow to form PIs.

Split conformal prediction

Even though CP is applicable to both, regression and classification problems, the focus of this work is by default exclusively on regression.

General procedure for of split conformal prediction in regression, following the (26CP, 1CP, 9CP)

1. Split the data in training, calibration and test set
2. Fit the prediction model on the training set
3. Make predictions with this model on the calibration
4. “Identify a heuristic notion of uncertainty […]” (1CP, p. 5), e.g. the absolute error of a prediction |y-f^(x)|
5. Define a score function (A score function can be chosen arbitrarily if it has the right orientation, i.e. lower values are better) 9CP and apply this function to the forecasting errors
6. Compute the quantile as see 1CP, p.5 of the calibration scores
7. Make predictions on the test set
8. Use the previously calculated quantile to form prediction intervals (add/subtract the quantile from the point prediction)

Regardless of the score function, these intervals have the coverage property, defined in (23CP).

 See 1CP, p.6

The only condition that must hold for this coverage guarantee is exchangeability (1CP, p. 50) in a sense that records, from training, validation and test (what is being predicted) are exchangeable which is weaker than i.i.d. data because exchangeability can be expressed with the following formula.

 9CP, p.3

The concrete implementation of CP for the pnbd case will be done in the next chapter.

### Measures to assess reliability and sharpness

The goal of this work is to assess different methods in their performance to derive prediction intervals and benchmark them against the Bootstrap approach. To fulfill this goal, several measures will be introduced in the following. They mainly address coverage, width and combined performance. The table below informs about recent works which employed these methods and how they are useful.

#### PICP (Prediction Interval Coverage Probability)

This measure holds the percentage of cases when the true value lays inside the constructed PI:

With n being the number of customers, yi the true observation for a customer’s CET and PIi the respective prediction interval.

#### ACE (Absolute Coverage Error)

This measure indicates how much on average the Prediction Interval Nominal Confidence, PINC, i.e. 90% and the true coverage, PICP, differ.

#### PICPW (Prediction Interval Coverage Probability Weighted)

This measure assesses how the coverage develops for more valuable customers. It weighs the coverage with the number of true transactions and therefore overweighs important customers compared to the neutral PICP. E.g. if there were 1000 repeat purchases across all customers of a dataset, a customer has 15 purchases and the interval covers these 15, this “1” for “value covered” would be weighted with 15/1000. If in contrast a customer makes 0 transactions, it would not increase PICPW, regardless of the interval covering this value because its weight is 0/1000. Hence PICPW only measures if high value customers are identified and not if low value customers are identified. Obviously, it must be evaluated together with the width. Another interpretation of this measure is essentially the coverage of repurchases across the customer base.

#### PIARW (Prediction Interval Average Relative Width)

This measure assesses the interval width with respect to the level of the estimation. It takes account of the special CET situation where the predictions have significantly different values and hence, intervals must be assessed accordingly. E.g. an interval having the width of 4 is of different value for a prediction of 1.2 and a prediction of 50. The actual PIARW is the mean over all customers.

#### PIARWW (Prediction Interval Average Relative Width Weighted)

This measure is the equivalent for PICPW but for PIARW.

#### MSIS (Mean Scaled Interval Score)

The MSIS is based on the interval score, proposed by (36PI) but scaled by the estimation and averaged over all customers. In other time-series contexts, see the table below, the scaling is done with the seasonal differences which is not applicable here. This measure assesses reliability and sharpness at the same time because it penalized interval width and true values outside of the CI. A low value is therefore better than a higher one.

#### SWR (Sharpness Width Ratio)

This measure evaluates coverage per width achieved by the intervals. A higher value is better than a lower.

#### Upper coverage

This measure indicates the percentage of times the upper prediction limit was not exceeded by the true value.

With ULi being the upper limit of the prediction interval.

#### Lower coverage

This measure indicates the percentage of times the lower prediction limit was smaller than the true value.

With LLi being the lower limit of the prediction interval.

#### Computational time

This measure holds the time which was needed by a method to calculate the prediction intervals.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2L | 15 | 16 | 20 | 31PI | 35 | 36 | 60 | 61 | 38 | 59 | 39 | This work | Reliability and Sharpness | Reliability | Sharpness | Downside risk | Upside potential | Context independent/Generalizability | Simplicity |
| MSIS[[6]](#footnote-6) | x |  |  |  | x |  | x |  |  |  |  | x | x | x |  |  |  |  |  |  |
| PICP[[7]](#footnote-7) | x | x | x | x | x | x |  | x | x | x | x |  | x |  | x |  |  |  | x | x |
| Upper coverage | x |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  |  | x | x | x |
| Lower coverage |  |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  | x |  | x | x |
| ACE[[8]](#footnote-8) |  |  | x | x | x |  |  |  |  | x |  |  | x |  | x |  |  |  | x | x |
| PIARW |  |  |  |  |  |  |  | x | x |  | x |  | x |  |  | x |  |  |  |  |
| PIARWW |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |
| SWR |  |  |  |  |  |  |  |  |  |  |  |  | x | x |  |  |  |  | x | x |
| PICPW |  |  |  |  |  |  |  |  |  |  |  |  | x |  | x |  |  | x | x |  |
| Comp. time |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  | x | x |

The following table is not exhaustive in a sense that it does not contain all papers that dealt with evaluating their PIs but contains a collection of recent studies that employed inter alia methods that are applicable for the CET context as well.

# Applied methods

In this chapter, it will be explained how the previously introduced methods to derive PIs are concretely implemented in the pnbd context.

## Bootstrap

The bootstrap approach is used as benchmark for all other methods to be introduced later, as it is the most established method, spread in all scientific domains and straight-forward to implement and understand. The procedure goes as follows.

1. Replicate x bootstrap samples of the customers and their transactions
2. For each bootstrap sample, fit a new pnbd-model
3. With each model, predict (all?) customers
4. Take the central x% confidence interval of all predictions for each customer

## Mini Bootstrap / Ensemble

As the name suggests, this method is a combination of several methods. Its idea has been introduced by 57PI and the implementation follows their approach.

1. Fit one pnbd model on the purchasing history of the customers for whom the prediction shall be made
2. Receive parameter estimates and the respective covariance matrix
3. From the estimates and the covariance matrix, simulate n draws of the parameters, assuming the parameters to be multivariate normal distributed
4. Each draw of this parameter simulation is the basis for a new model
5. Run the prediction with each of these new models
6. Receive n values for the CET
7. Take the central x% interval of the predictions for each customer

Its great advantage compared to the bootstrap is that requires only a single model fit.

## Bayesian method

This approach requires to fit the whole pnbd model with the Bayesian approach and then take the intervals from the posterior predictive distribution. As it would be out of scope for this work to re-estimate the model with the Bayesian approach, the existing implementation from the BTYD package in R was used. Therefore, only a rough explanation how the Bayesian model fitting works will be given.

1. Estimate the pnbd model with the Bayesian approach: (BTYD package)
   * No previous information about the parameter distribution is given
   * Use the mcmc method to get the parameter posterior distribution
   * Use the mcmc method to get the draws of the posterior predictive distribution
2. Take the desired quantiles of the posterior predictive CET distribution for each customer

## Quantile regression

As indicated before, the implementation of Quantile regression will be a modified from the “original” approach but keeps the idea of direct interval estimation. To avoid introducing an optimization in the model fitting process, the optimization is 1. Conducted after the model fitting and 2. Broken down into to a grid search. More precisely,

1. Build a grid of parameter combinations
2. Each parameter combination is the basis for a new model
3. Predict the CET for each customer with every new model
4. Introduce a distance measure what serves as a loss function for
   1. The parameters for the Upper bound parameters
   2. The parameters for the Lower bound parameters

When the central (1-a)-quantile is requested, a/2 of the observations should be above and a/2 below. The first part of the equations measures exactly this coverage and then a/2 is deducted.

1. Collect these calculated differences
2. Select the parameter combinations that yield the lowest absolute differences, one combination for the upper, one combination for the lower bound.

It is apparent that the true values are required in this method in order to calculate out the coverage of each combination. In reality, one does either not have these values (because they lay in future, and they shall be predicted) or they are known because they lay in the past. Then, there would be no point in predicting them. To overcome this issue, one could assume that customer behavior for one firm will not change a lot. In this case, there is no reason to assume that the optimal parameter combinations that yield the desired quantiles would change. So, the first part could be run on old data of the company to figure out the optimal parameter combination for each bound and then use those on the new, interesting data to construct prediction intervals.

A few notes on the first part of the procedure:

* For setting up the grid, it is helpful to have some prior knowledge where parameters should be located approximately. This step was done manually, running several attempts by hand for a rough orientation and then giving several alternative values around. It turns out that, regardless of the dataset (and in the next chapter, regardless of the learning and holdout period lengths), nearly the same parameters are selected. This is a very convenient situation for the application in practice as the combinations do not need to be identified anymore and one can run the approach on a smaller grid.
* Many customers don’t buy again after their initial purchase. Regardless, which parameter combination is selected, the model will never predict exactly 0 but something very close to 0. It makes sense to introduce a small tolerance and set those predictions to 0 which are reasonably close to 0 to give method a chance to perform well. This seems arbitrary but in practice, one could argue that a managerial decision for a customer will barely differ if CET = 0 or CET = 0.1 (what is the used tolerance). Also, one could argue that it is “unfair” against the other methods. Quantile regression is the only method that suffers from this issue and adding this tolerance to increases at the same time the QR-interval’s widths, so it comes at a cost. Doing this trade-off for other methods would not increase their performance.

## Conformal prediction

As indicated before, only Split Conformal prediction will be implemented. The implementation follows in principle the steps from the general description but there are several modifications to be made.

1. Heteroskedasticity of the outcomes: It appears that customers have a very different repurchasing behavior and might buy again 0 or 50 times. When the model is off by 3, say for the first customer (y\_1 = 0), it predicts 3 and for the second customer (y\_2 = 50) it predicts 53, the absolute delta would be equal, but the model would have done a bad job for the 1st customer and good job for the 2nd customer. Assuming that model is equally good at all levels, it is reasonable to employ adaptive confidence intervals. Otherwise, the method would suffer from an over coverage for small and under coverage for large values. A solution to that issue is given in (1CP, 9CP), as they suggest scaling the residuals by their standard deviation, “studentization”. The approximate standard deviation for customers of one company can be approximated as a linear function of their CET level, retrieving sd(CET\_i). This process can be summarized as
   1. Fit pnbd-model on the old and known data
   2. Make predictions
   3. Get the absolute differences for each prediction to their true value
   4. Fit a linear model absolute\_difference(CET\_i) ~ CET\_i

For every CET, there is now a reasonable scale.

1. As it was before the problem with Quantile regression, Conformal prediction needs the true data, not only for the standard deviation but the actual method functionality. Again, one could make the assumption that customers’ behavior for one firm is approximately constant over time. The whole process of model fitting, and derivation of the quantile and standard deviation could therefore be conducted on old, known data and then the quantile and standard deviation forwarded to the current period.

Assuming that the linear model for the standard deviation has been derived already, the whole process can be summarized:

For the old data set

1. Split the old data set into a train and a test set (split customer wise)
2. Train the pnbd-model on the training set
3. Make point predictions on this training and test set
4. Take the deviations from the training set and scale them (divide by their estimated standard deviation)
5. Take the desired x%-quantile of the scaled residuals (formula)
6. Rescale the quantile with the CETs of the test data set
7. Add and subtract these individually scaled residuals to the point predictions on the test set
8. Evaluate the average coverage on the test set

Theoretically, it is not necessary to split the old data set, but it is reasonable to check if the quantile works at least with old data, without having it transferred to the new data.

For the new data set

1. Train the pnbd-model on the new data set
2. Make point predictions on this data set
3. Take the quantile and linear function of the standard deviation that were previously derived
4. Rescale the quantile with the predicted standard deviation for each customer i (quantile \* sd(CET\_i))
5. Subtract and add the individually scaled quantile for each point prediction

Three additional remarks on the procedure:

1. If the old data set is not very large and there is a potential bias when splitting the old dataset, one may want to repeat this procedure several times and average over the retrieved quantiles. This is done in this work as well.
2. Typically, one would use a separate validation set in Conformal prediction to get realistic quantiles from new data which the model has not yet seen. Since the model is fitted on the new data anyway, the use of a calibration set becomes redundant.
3. The intervals would usually reach below 0, so they are cut off at 0 to incorporate obvious knowledge and make them comparable to other methods in terms of interval width.

## Conceptual comparison of the methods

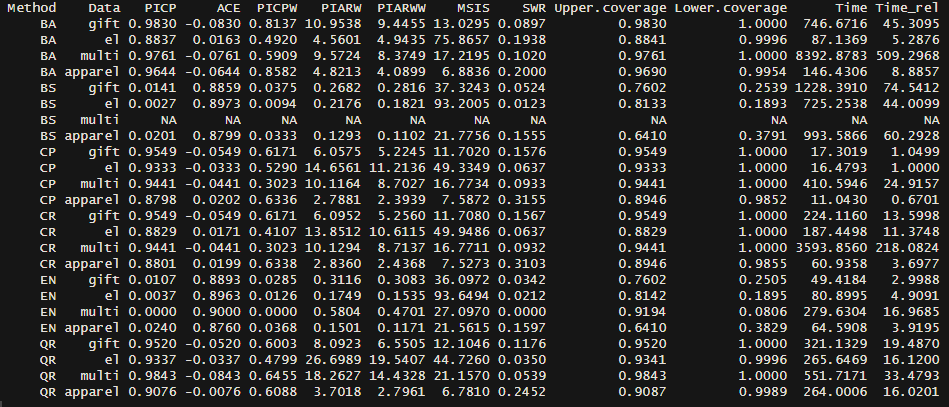
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Focused uncertainty** | **Assumptions** | **True values needed** | **Computational  effort** | **Approach complexity** | **Frequentist approach** |
| **Bootstrap** | | Data uncertainty/Bias | None | No | High | Low | Yes |
| **Ensemble** | | Model and parameter uncertainty | Normal distribution of re-sampled parameters | No | High | Medium | Yes |
| **Bayesian** | | Model and parameter uncertainty | None assumptions but set priors for parameters | No | Medium | High | No |
| **Quantile regression** | | (Distributional uncertainty of the response variable), model misspecification | None | Yes | Depends on prior knowledge on parameters | Medium | Yes |
| **Conformal prediction** | **CP** | Model uncertainty | Exchangeability | Yes | Medium | Medium | (Yes) |
| **CR** | Model uncertainty | Exchangeability | Yes | High | High | (Yes) |

# Results

## Methods’ Performance - Benchmarking Bootstrap

In the course of this work, all[[9]](#footnote-9) methods are applied to all datasets and evaluated with the measures that were introduced above. The following table contains these results, which shall be subject to some deeper analyses.

### General performance



A black screen with numbers

Description automatically generatedA black screen with white text

Description automatically generatedThe results table from the previous page have been summarized by every method by averaging over the respective results and a ranking of these averages has been produced. The following insights in the method performance are to be noted.

Major findings

1. The 2 bootstrap-based methods deliver consistently under coverage across data sets (BS, EN) while the 4 other methods deliver roughly the desired or even over coverage (BA, CP, CR, QR).
2. The widths of the 2 underperforming methods are significantly lower than the widths of the other 4 methods.
3. There is not trade-off method that combines both strengths and finds a compromise. This is visualized in graphics…
4. CP and EN are significantly faster than all other methods, BA is the most time consuming.

Other findings

1. When over weighing more important customers, generally decreases the the coverage, meaning that all methods perform better with important customers. The 4 satisfyingly covering methods range roughly between 50% and 70% with BA leading with 68.9% over the next closest, QR with 58.4%. The 2 bootstrap-based methods give relatively signifantly better results than without the weighing but are both still below 3%.
2. Regarding the width, EN and BS have the by far lowest width with the respect to the prediction, followed by BA, CR and CP, serving a middle way. QR has in 2 data sets comparable values to these 3 methods but in the electronics and the multichannel case, it develops very far intervals, resulting in a very high average. One can infer for that the performance of QR has higher variability and dependence of the data set.
3. Intervals for all methods shrink in relation to CET. This can be mostly explained by not scaling with excessively small numbers.
4. The combined assessment of sharpness and width sees BS and EN on the lower rank, due to the penalization of non-coverages. The rest of the methods have roughly the same performance regarding MSIS with a slight disadvantage for BA.
5. The same picture can be seen in SWR. The 4 reliable methods have around 2-3x more coverage per width than BS and EN.
6. CP and CR hit the desired 90% the most accurate. (least over or under covering)
7. BS and EN have, scaled to their actual performance, a high variability (coefficient of variation) in reliability measures across data sets. For shaprness measures, all methds have comparable variability across data sets, see variation table in the appendix.

A graph of data set

Description automatically generated with medium confidenceThe insights can be observed in the following visualizations.

A graph with green and red lines

Description automatically generatedOne can clearly observe the two different types of methods, the bootstrap-based on the left with low coverage and small intervals on the left and the other methods on the right, delivering the desired coverage at the cost of wider intervals. What is also interesting to note is that data sets cause either wider or shorter intervals across CR, CP and QR. This means that the methods realize model uncertainty or inaccuracy and react to it, as the widen the intervals to consistently deliver the desired coverage (what is what they are optimizing for). This can also be observed for EN and BS but in another order, surprisingly.

This graph shows the concrete lengths of prediction intervals for concrete customers and the true number of transactions exemplary for the apparel data set (see in the appendix the graphs for the other data sets). Again, one can observe the small ranges that are covered by EN and BS and the wide spread of true observations that show how unlikely it is for these methods to cover a true point. The other methods capture the uncertainty appropriately and (in this case at least) cover all true points. For this data set and selected range of CET, BA has the longest intervals, but this is not representative for the rest of the data sets and ranges of CET as the overview table shows and as can be seen below.

Summary

There are two types of methods, those width wide intervals and appropriate coverage and those with narrow intervals and low coverage. In combined measures of reliability and sharpness, usually the methods with high coverage outperform the other two. These domains of method expertise are visualized in the radar chart below.

A diagram of a star

Description automatically generated with medium confidence

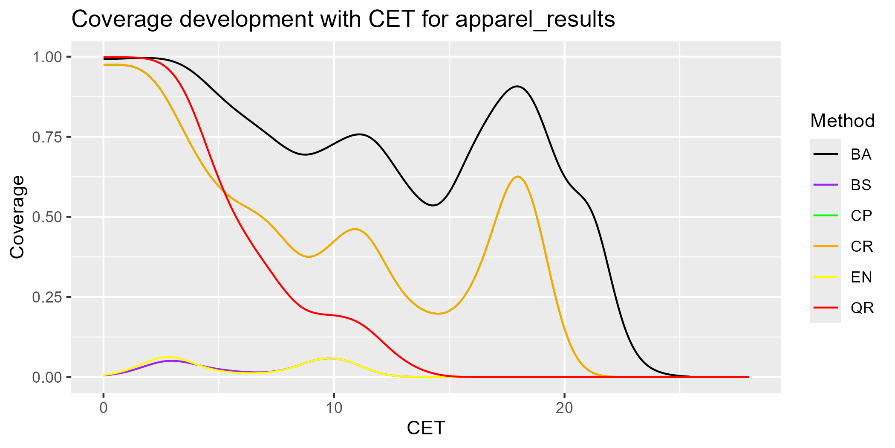
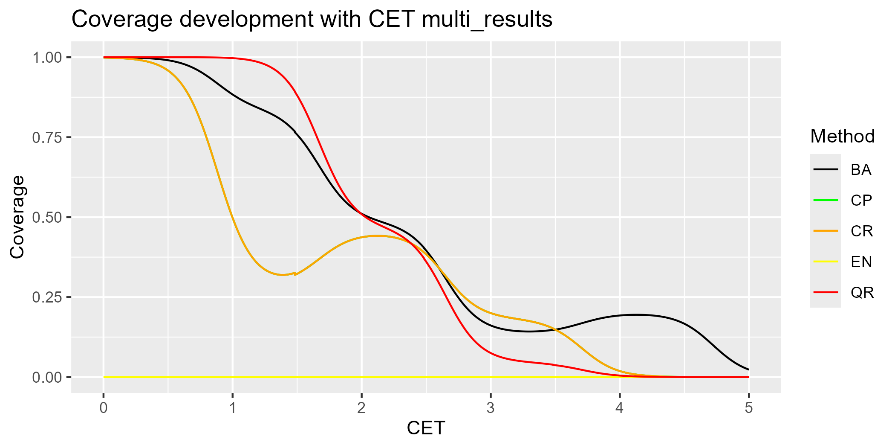
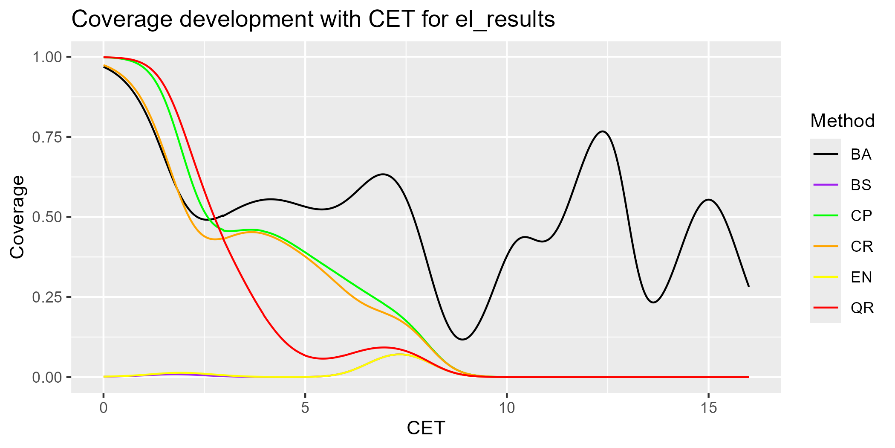
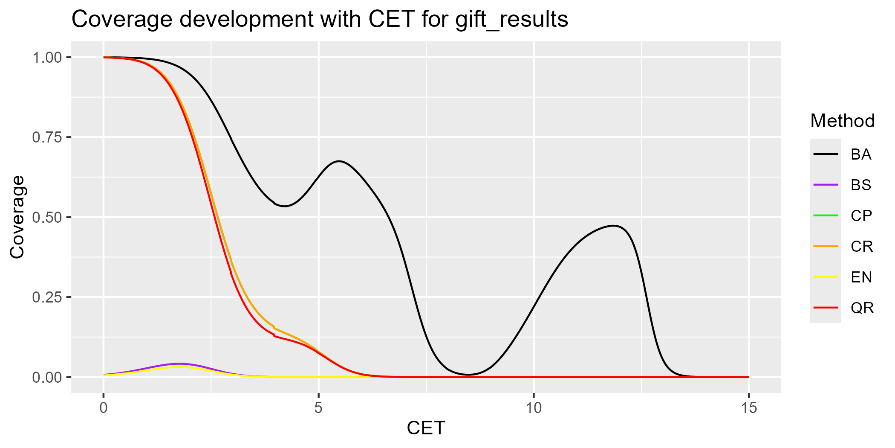
### Deeper insights into and visualizations of methods’ performances

In the general insights section, it was already mentioned that methods might perform differently at different levels of CET. This shall be examined more in-depths.

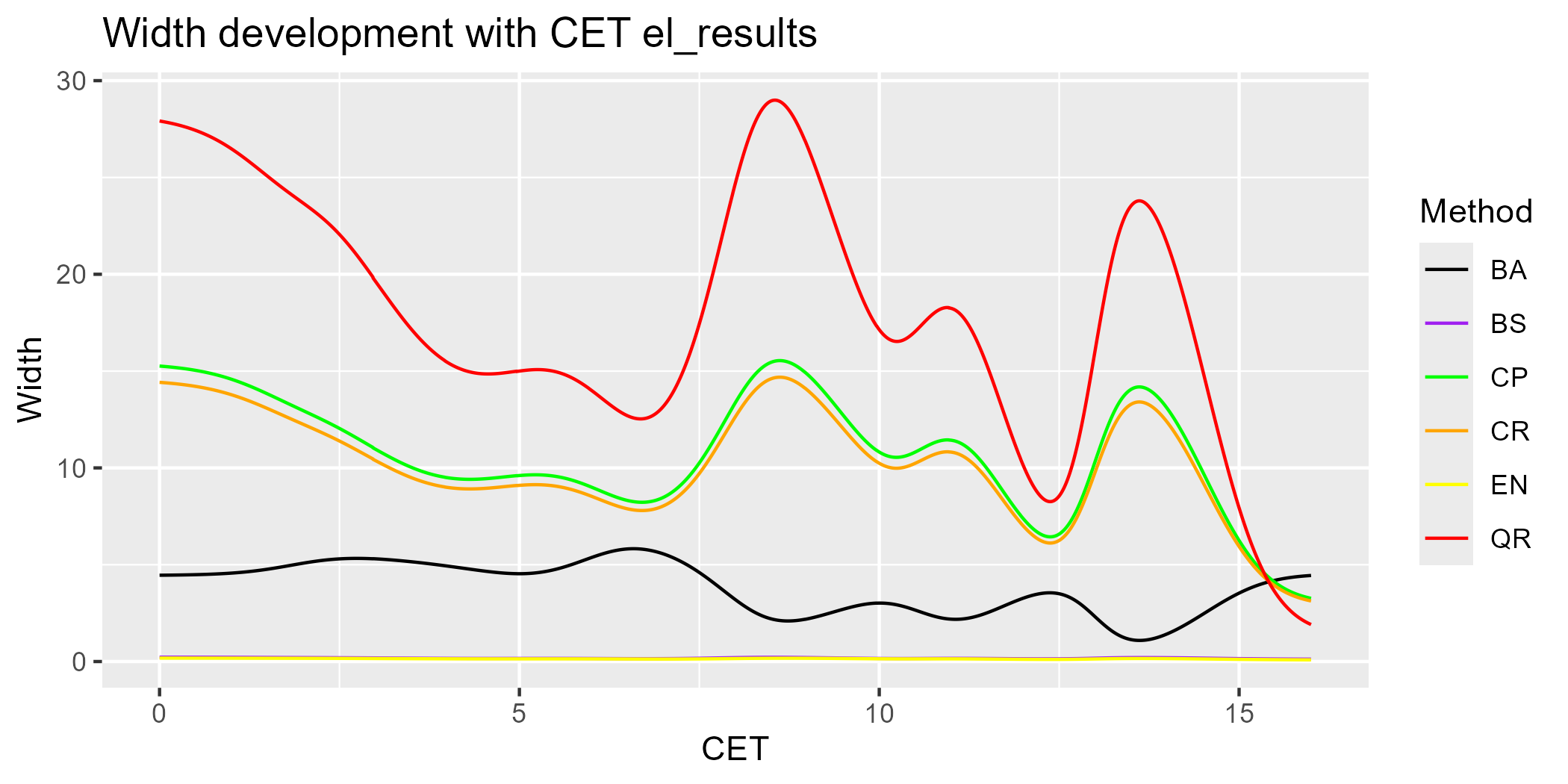
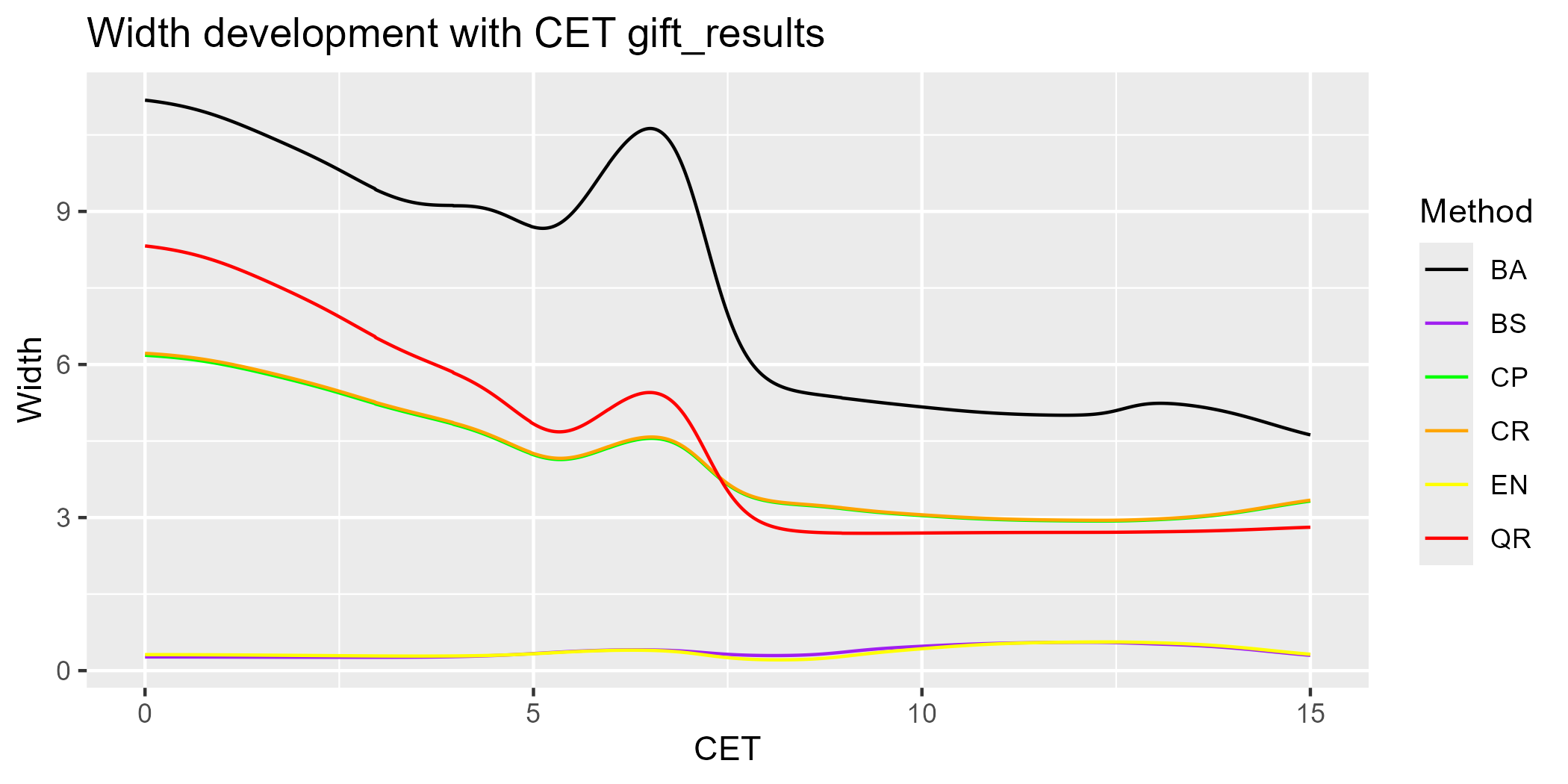
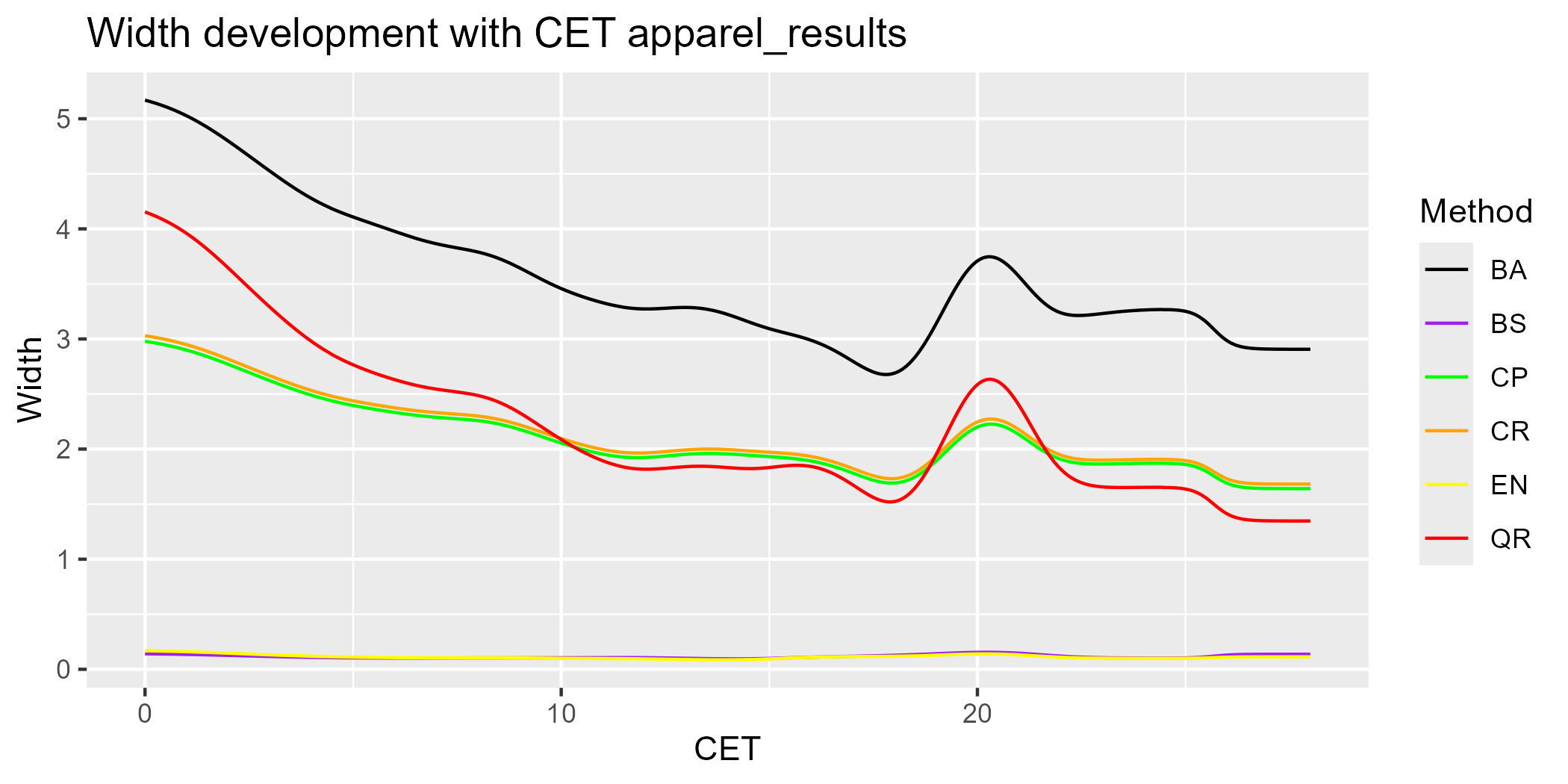
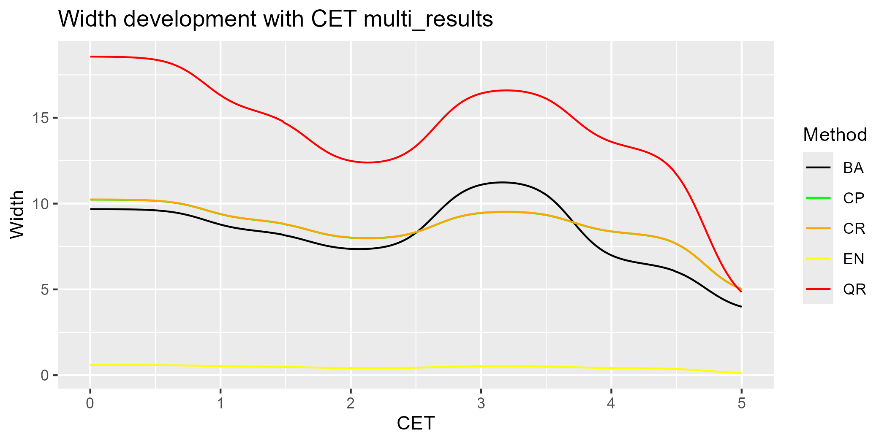
#### PICP across CET levels

As the coverage develops across different CET levels, this development is not quite the same for different data sets, therefore see below this development for all data sets. The following graphs smooth the coverage across CET by the help of a kernel with normal distribution.

Across all data sets, BA has the highest coverage what is not surprising because is has the highest over coverage and EN and BS have the lowest coverage across all levels, being nearly 0. The second note to make is generally a downward tendency that is approximately true for all methods on all data sets. This means that the intervals are less likely to include the true value if it is high. After the analysis of PICPW, this comes not surprising but poses an issue as it is exactly this segment that marketers are interested in. The high coverage rate at lower values comes from many true values being 0 and usually intervals are covering 0 either by nature or by the small tolerance that was given in case of QR. The lower coverage on high values is visible across methods what leads to the conclusion that the model might be less accurate in this area. Indeed, the relative error in for higher true values is a lot higher than for smaller values (see appendix). In the methods, this behavior is not covered and hence explains as well the downward trend.



#### Width across CET levels



Considering PICP across CET levels, it must be seen in connection with the respective width development. The situation across data sets is very heterogeneous but also here, there is generally a downward trend visible, while either BA or QR delivering the widest intervals and BS and EN delivering across all levels intervals with a length of nearly 0. The downward trend is not as dominant as for PICP but still visible. A lot of this phenomenon can be explained by dividing “normal”-sized intervals at lower levels by very small predictions, delivering relatively wider intervals. The key insight therefore is that a decreasing PICP with increasing customer value is first and foremost not the responsibility of relatively narrowing PIs. Nevertheless, it is arguable that more sophisticated methods should realize the differentiated model performance and adapt their PIs accordingly.

#### Summary

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Adequate coverage | Consistent coverage across outcomes (CET, PTS) | Consistent coverage across outcome levels | Consistent performance across data sets | Works in all examined datasets | Usefulness | Computational intensity |  |
| Bootstrap | No | - | No | Yes | No | Low | High |  |
| Ensemble | No | Yes | No | Yes | Yes | Low | Low |  |
| Bayesian | Yes | - | No | Yes | Yes | Medium | High |  |
| CP | Yes | Yes | No | Yes | Yes | Medium | Low |  |
| CR | Yes | Yes | No | Yes | Yes | Medium | Medium |  |
| Quantile regression | Yes | Yes | No | Coverage yes, but not width | Yes | Medium | Medium |  |

## Performance over varying training and prediction periods

Motivation

The analysis so far has held constant the learning and prediction period for each data set and examined how different performance measures vary across methods and CET levels. This chapter will deal with the cases when the periods for learning and prediction are varied and assess how the overall coverage is influenced. This scrutiny is mainly motivated by the methods QR, CP and CR which use old data from previous periods. Those data exist in the case of this work and the used data sets, but in reality, a lack of past data is not unrealistic and might cause problems. I.e. the data could be biased so that inappropriate quantiles (CR and CP) are learnt, or the wrong parameters are selected (QR). Another issue is that the model has difficulty to be fitted due to the lack of data and inherently loses predictive power. As this topic is not the core of this work, it will be held concise and concentrate on the most central insights.

Implementation

Due to computational reasons, the analysis will be limited to the electronics and gift data sets and the bootstrap based methods will be left out because only coverage is assessed here and a loss in this area is more relevant to BA[[10]](#footnote-10), QR, CP and CR. For both data sets, around 50 combinations of learning and prediction times[[11]](#footnote-11) for the old and the new cohorts are used. For each period combination, a model is fitted, predictions are made, and prediction intervals are derived and assessed regarding coverage. The resulting table can be found in the appendix. The following graph resulted from the table.

A graph with blue and red squares

Description automatically generated

BA and QR were, regarding coverage, not affected by changing period lengths or data set and deliver constantly around the desired 90% quantile. In contrast, the performance of the conformal prediction implementations suffer for the gift data set a lot while delivering decent results for the electronics data set. Looking directly gives several insights but no definite explanation for the underperformance and especially not for the difference between the two data sets. Typically, low values resulted from short learning or prediction periods for the old cohort, what makes intuitively sense as it is here where the quantiles are derived. On the other hand, one can also observe cases in which all periods, for both cohorts have been low which performed decently. It suggests that when both cohorts have been treated equally in terms of model fitting and quantile forming conditions, the system works properly.

The motivation for this chapter was to give an intuition how stable the methods are and if managers can apply them safely. QA and QR seem to work stable, regardless of data set or learning or prediction time. The Conformal prediction-based methods seem to suffer on some datasets but work well on other ones. The reasoning behind it cannot be clarified absolutely but tendencies are observable. It will be necessary to use more runs and data sets to get deeper results for all methods but one should be especially careful to use CP or CR with short periods.

## Application in marketing

In this final chapter, the usefulness of the results beyond the assessment of the model uncertainty shall be discussed. In particular, it shall be examined if and how prediction intervals can help to identify valuable customers in terms of number of future transactions. There have been numerous studies about how to identify valuable customers, amongst others the so-called RFM analysis where Recency, Frequency and the Monetary value of transactions are assessed. As in this work, the focus is on the number of transactions, only recency and frequency will be taken as a benchmark. A third benchmark is the point prediction CET.

Implementation

1. Use only 1 combination of times per data set
2. Train the model and make predictions for each data set
3. Calculate prediction intervals with all methods for all data sets
4. Apply different strategies or metrics to the outcomes and rate customers with them
5. Order the customers according to these metrics
6. Pick the top 10% of the customers and calculate their performance (how many transactions did the selected customers, compared to the theoretically maximum possible transactions by selecting the 10% customers, based on their true transactions)

The point where uncertainty comes into play is the calculation of the metrics which involves information from the intervals, in order to beat the benchmarks.

**Metrics**

**Benchmarks**

* Highest Point predictor (hpp)
* Highest Frequency of previous purchases (hfq)
* Recency of previous purchases (hrc)

**Challengers**

* Highest upper limit of the intervals (hul)
* Highest interval width[[12]](#footnote-12) (hiw)
* Highest upwards uncertainty (Difference between the point prediction and the upper interval limit) (huu)
* Highest CET/width (hcw)
* Lowest CET/width (lcw)
* Highest frequency/width (hfw)
* Lowest frequency/width (lfw)
* Highest recency/width (hrw)
* Lowest recency/width (lrw)
* Highest frequency\*width (hfxw)
* Lowest frequency\*width (lfxw)
* Highest recency\*width (hrxw)
* Lowest frequency\*width (lfxw)

The results from this implementation can be found in the following table

A screen shot of a computer screen

Description automatically generated

It must be acknowledged that the metrics based on the uncertainty quantification were not able to beat the benchmarks, especially frequency. On the contrary, most performed significantly worse across all data sets. As the only methods that did not incorporate one of the benchmarks, huu and hcw were occasionally able to keep up with or beat the predictive power of recency or the point predictor. None of the methods was able to beat frequency and one must recognize that the intervals in this form are not useful to identify especially valuable customers. Using the absolute interval width instead of the relative one has produced similar results.

An intuitive explanation at least for BA, QR, CP and CR for that behavior is that these methods do generally not hold any information about single customers. Much more, they hold general uncertainty information about the sum of customers and the overall model performance.

* EN:
* BA: The intervals from BA come from the estimated posterior distribution of model parameters from where a posterior predictive distribution of the actual outcomes is derived. The posterior parameter distribution is for sure based on the data that came from the customers, but as soon as the posterior parameter distribution exists, it does not hold any individual information about a single customer anymore.
* QR: A similar situation as for BA can be found for QR as well. The whole information about uncertainty of customers is summarized by 8 parameters, 4 for the lower boundary and 4 for the upper boundary. It is not possible to retrieve individual information from this point.
* CP/CR: All the information about uncertainty from an old cohort is summarized by a single quantile which is then applied to new customers and scaled by their individual point prediction which does by definition not include any information about uncertainty.

# Managerial / research discussion

Strengths, weaknesses, shortcomings, compromises, limitations

Impact on strategical planning, applicability in reality

# Conclusion

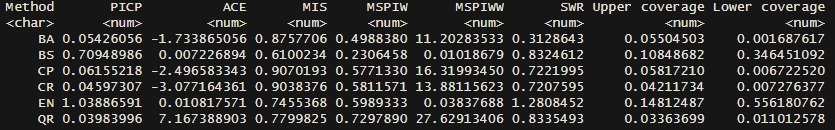
…

# References

…

# Appendix

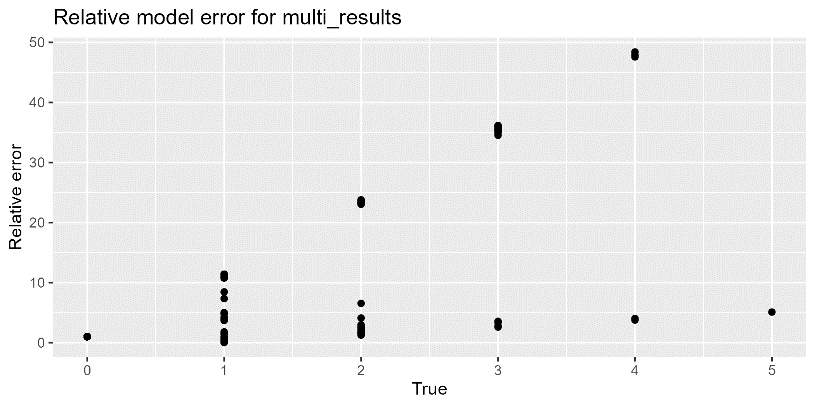
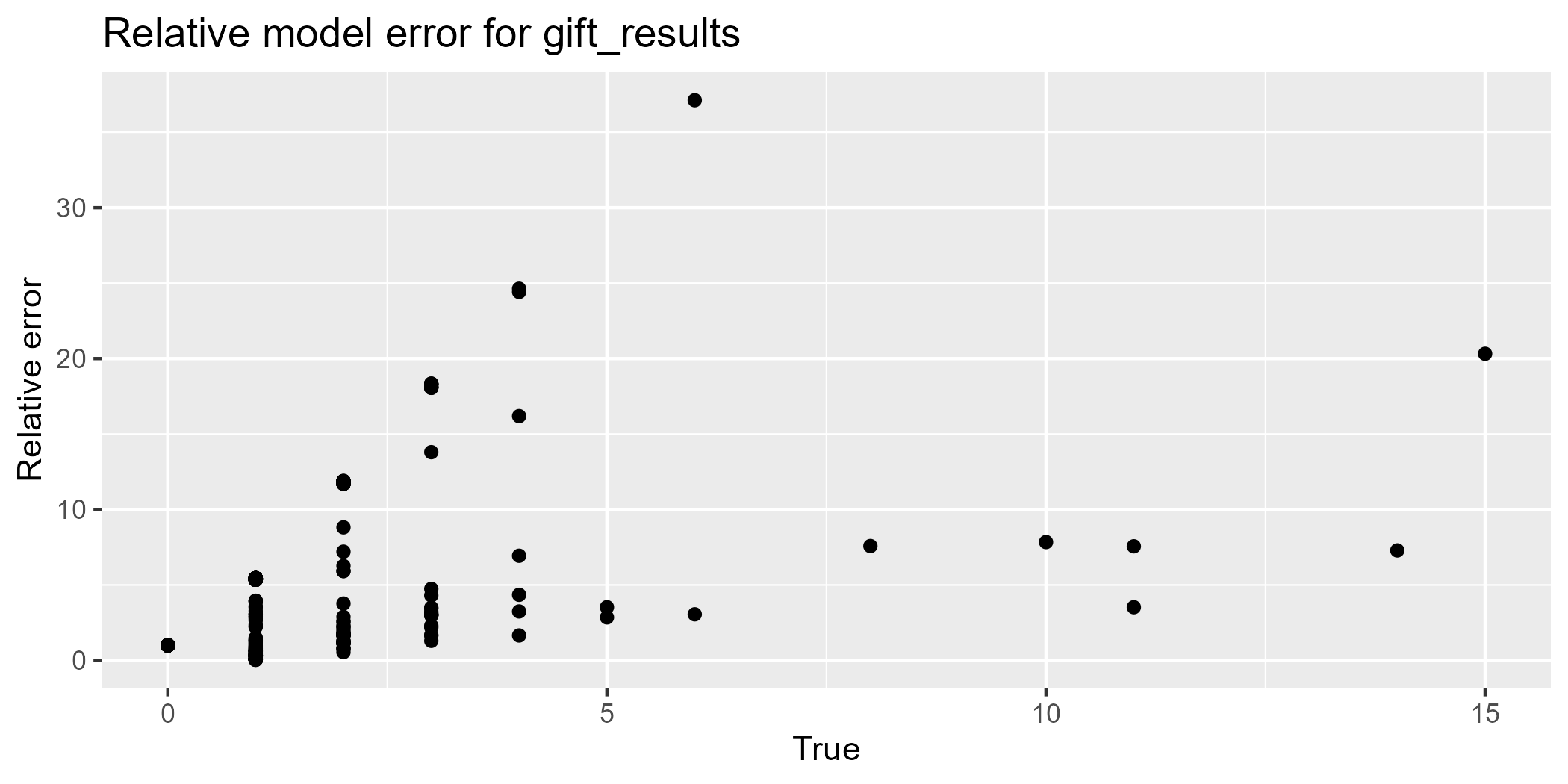
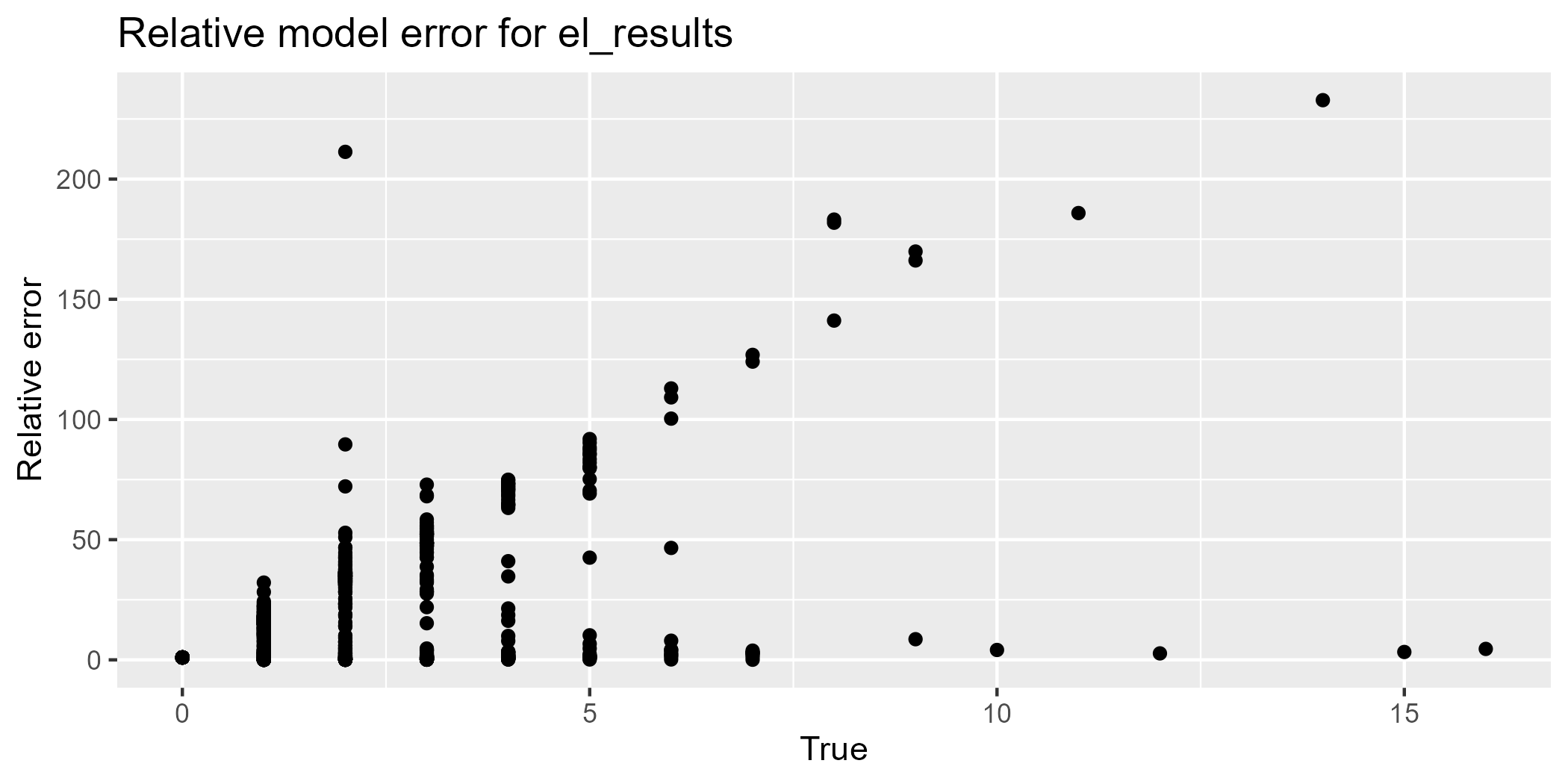
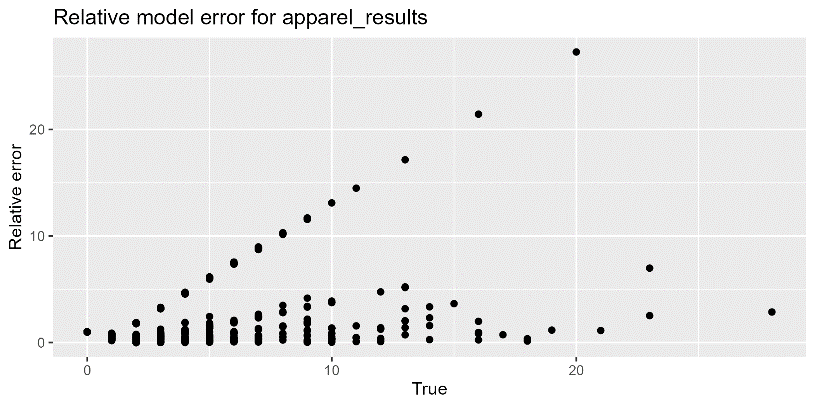
Coefficient of variation for different methods and measures across datasets



* Standard deviation scaled by the mean of the respective measure
* Lower variation for PICP, higher variation for MSPIW and especially MSPIWW

A screen shot of a computer screen

Description automatically generatedPerformance with absolute width instead of relative width



1. https://www.collinsdictionary.com/de/worterbuch/englisch/uncertainty#google\_vignette [↑](#footnote-ref-1)
2. https://en.wikipedia.org/wiki/Uncertainty [↑](#footnote-ref-2)
3. https://www.ama.org/the-definition-of-marketing-what-is-marketing/ [↑](#footnote-ref-3)
4. [The market system : what it is, how it works, and what to make of it](https://swisscovery.slsp.ch/discovery/fulldisplay?docid=alma991142075309705501&context=L&vid=41SLSP_NETWORK:VU1_UNION&lang=de&search_scope=DN_and_CI&adaptor=Local%20Search%20Engine&isFrbr=true&tab=41SLSP_NETWORK&query=any%2Ccontains%2CThe%20Market%20System%3A%20What%20It%20Is%2C%20How%20It%20Works%2C%20and%20What%20to%20Make%20of%20It&sortby=date_d&facet=frbrgroupid%2Cinclude%2C9048714116547914341&offset=0), p. 35 [↑](#footnote-ref-4)
5. https://link.springer.com/chapter/10.1007/0-387-35429-8\_5 [↑](#footnote-ref-5)
6. In the works that used this measure before, it was in a time-series context and scaled by seasonal differences. In the CLV context, this scaling is not applicable and is replaced with a scaling by equ. [↑](#footnote-ref-6)
7. Different names have been found in the literature for this measure: 2L: Coverage; 15: True coverage; 16, 20, 35: PICP; 31: Coverage rate; 38: ECP (Empirical coverage probability) [↑](#footnote-ref-7)
8. Different names have been found in the literature for this measure: 16, 20. 38: ACE; 31: ACD (absolute coverage difference) [↑](#footnote-ref-8)
9. Except the bootstrap approach which repeatedly failed with the multichannel retailer dataset. [↑](#footnote-ref-9)
10. Even though BA does not need old data, it might still be affected by learning and prediction periods of the current cohort, therefore it was included. [↑](#footnote-ref-10)
11. Note that for the electronics data set, it was not possible to go much below 50 weeks for learning for the old cohort. For comparability reasons, the records with a short learning period in the gift data set were also left out. The results have merely changed. [↑](#footnote-ref-11)
12. Note that with „width“ is meant “relative width” [↑](#footnote-ref-12)