Inhaltsverzeichnis

[Introduction 1](#_Toc173586203)

[Literature review 1](#_Toc173586204)

[Role of uncertainty in marketing 1](#_Toc173586205)

[Methodological literature overview 2](#_Toc173586206)

[Types of uncertainty 2](#_Toc173586207)

[The importance of prediction intervals 4](#_Toc173586208)

[Methods to derive prediction intervals 4](#_Toc173586209)

[Measures to assess reliability and sharpness 8](#_Toc173586210)

[Applied methods 10](#_Toc173586211)

[Conformal prediction 10](#_Toc173586212)

[Ensemble (former Method 1) 12](#_Toc173586213)

[Bayesian method 12](#_Toc173586214)

[Quantile regression 12](#_Toc173586215)

[Conceptual comparison of the methods 13](#_Toc173586216)

[Results 14](#_Toc173586217)

[Methods’ Performance - Benchmarking Bootstrap 14](#_Toc173586218)

[General performance 14](#_Toc173586219)

[Deeper insights into and visualizations of methods’ performances 15](#_Toc173586220)

[Performance over varying training and prediction periods 18](#_Toc173586221)

[Managerial / research discussion 18](#_Toc173586222)

[Conclusion 18](#_Toc173586223)

[References 18](#_Toc173586224)

[Appendix 18](#_Toc173586225)

# Introduction

…

# 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 the 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 certainty, a state of limited knowledge where it is impossible to exactly describe the existing state, a future outcome, or more than one possible outcome.”[[2]](#footnote-2) Both definitions agree on the presence of unknown information with respect to a potential current or future state and connected actions (to be taken). On the other hand, there is a similar situation for marketing, the AMA 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. 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 that is where uncertainty comes into play as an important aspect here is the prospective demand. A market is a place which is heavily influenced and driven by the actions of its agents, i.e. “customers, clients, partners, and society at large”, and their connections.[[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, location of new branch or simply a consumer’s unawareness of a competitor’s product which is 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). 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 of the future by constructing models which depict a picture of reality, incorporating important aspects and leaving out unimportant ones for simplification.

## Methodological literature overview

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

A diagram of a comparison between two different types of graphs

Description automatically generated with medium confidenceThe sources of uncertainty in models can be divided into aleatory and epistemic uncertainty (21RU) which also applies to the pnbd model in the CLV context. As there is, to my knowledge, no clear, distinct definition of these notation, the idea behind these constructs shall be explained shortly. 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 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 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. *It is to note 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 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. One part, the generation of data is to locate in both areas because there can appear errors in the data collection and processing (aleatory) and it creates 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 shall be briefly discussed in the following. An interval (forecast) is offering a range of possible values of future outcomes with a specific level of confidence. (11PI) This means that the true value of the prediction will fall into this declared interval with a specific probability, e.g. 95%. (4PI) point out 4 main points, why PIs are of such importance.

1. They assess future uncertainty
2. hence enable the user to plan “different strategies for the range of possible outcomes”. This means that one can prepare a strategy in case a high value inside the interval is realized and one for a low value. 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 enable to assess different methods of forecasting more thoroughly
4. PIs allow to compare forecasts made under different assumptions more carefully.

Another point, made by (32PI) “forecasts cannot be expected to be perfect, and intervals emphasize this” which underlines maybe the most important characteristic 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 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 require different methods to derive PIs. In this work, the focus will be put on the work of 33PI, who suggest 4 big types of methods applicable in regression contexts, Bayesian approach, Ensembles, Direct interval estimation and Conformal prediction. 3 of 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 replaced by 2 different applications of bootstrap which picks up the idea of ensemble methods but is too far away to be called “ensemble”. A special emphasis will be put on conformal prediction due to its recent raise in attention in the statistical community.

#### 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. (1CP) It was first introduced by 8CP in 2005 and gained a lot of attention in recent years. Conformal prediction has two main ways 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). This comes due to 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 much more central, because Full CP is not applicable for the pnbd-model. This reason for this shall be roughly discussed in following 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 1. predictors Xi = 1:250 and 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 is lives in **Y.** The approach is to take n values as possible outcomes and reunite each record with the 249 unchanged records, ending up with n sets of records. For each set, a new model is being fitted which 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, the Ys 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 result in different models and would not allow to form PIs.

Even though CP is applicable to both, regression and classification, the focus of this work will be put exclusively on regression. The procedure in general of split conformal prediction for regression works as follows (1CP, 9CP)

1. Split the data in training and calibration set and fit the prediction model on the calibration set
2. Define a heuristic notion of uncertainty, e.g. |y-f^(x)|
3. Define a score function (A score function can be chosen arbitrarily if it has the right orientation, i.e. lower values are better) 9CP
4. Compute the quantile as see 1CP, p.5 of the calibration scores
5. Use this quantile to form prediction intervals (in this work, add/subtract the quantile from the point prediction)

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

See 1CP, p.6

The only condition that must hold for this coverage guarantee is exchangeability 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.

(Y1 Yn+1) d = (Y (1) Y (n+1)) for all permutations, directly taken from 9CP, p.3

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

#### 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 centre 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) When one has enough models fitted 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. 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 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, while only fitting 1 model and from this derive all other models 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 cinstruct PIs is hard tell. As of my knowledge, Aitchison was the first one to introduce the idea of using the Bayesian Approach’s strength in forming tolerance regions in 55PI. 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. Necessary to have information about the parameter probability distribution before the parameters are observed (or use an uninformative prior 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 derive 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 (33PI, 47PI, 54PI) Quantile regression does so by optimizing the parameters with respect to a loss function that is employed while model fitting. The idea is that one combination of parameters yields x overestimations and y underestimations of outcomes where it is the objective to balance outcomes above and below when this symmetric distribution is the objective. 47PI state that this is possible with any other quantile than the 50% quantile as well as the same principle applies: The loss function penalizes deviation of the desired above-below ratio, which 1:1 in the 50% (median case). Targeting e.g. the symmetric 90% intervals would require finding the 5%- and 95% quantile, so now, two parameter combinations must be found. Applying the same principle as above, the loss function would penalize according to the desired quantiles differently, 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.

### 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 address coverage, width and combined performance.

#### Scaling

Some measures require appropriate scaling which is usually done, dividing some parts of the formula by either the true outcome or the prediction. Potential candidates for this are the true value and the prediction itself. To be consistent, in this work will be use the prediction. Nevertheless, in the case of CLV or more precisely CET, it occurs that both options happen to be 0 in special occasions. Considering different alternatives, it appears to be the most reasonable solution to scale by the 2nd smallest observation, in case they are 0. This equivalent is defined as:

#### (True) coverage / PICP

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

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

#### ACE

This measure indicates how much on average the intended nominal prediction coverage, PINC, i.e. 90% and the true coverage, PICP, differ.

#### 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.

#### MIS

The Mean Interval Score is based on the interval score, proposed by (35) and averaged over all customers instead of timesteps as it originates from time series forecasting. Due to the scale difference between small and large values for CLV, the differences in the initial formula were scaled by the estimated value (scaling by the true value would result in division by 0 if the customer would not buy again). Also, it is not to be confounded with MSIS which would in addition be scaled by mean absolute seasonal difference. (31) This measure assesses reliability and sharpness at the same time because it penalized interval width and true values outside of the CI at the same time.

#### MSIW (Mean Scaled Interval Width)

This measure assesses the interval width with respect to the level of the true value. This takes account for the special CLV situation where an interval having a width of $10 is of different value for a CLV of $20 or $200 what is a realistic case for different customers. The actual MSIW is the mean over all customers.

#### MSIWW (Mean Scaled Interval Width Weighted)

Lower CLVs tend to have a higher (worse) MSIW for all methods (see appendix). At the same time, firms may want to put more emphasize on higher CLV customers. This measure will therefore weigh customers with higher CLV more, i.e. the weight for a customer will be its CLV relative to the sum of all CLVs.

#### SWR (Sharpness Width Ratio)

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

With euqi being the estimate. The formula was taken from (2L) and was slightly adapted.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2L | 15 | 16 | 20 | 31PI | 35 | 34 | 37 | 38 | 39 | This work | Reliability and Sharpness | Reliability | Sharpness | Downside risk | Upside potential | Context independent/Generalizability | Simplicity |
| MSIS[[5]](#footnote-5) | x |  |  |  | x |  |  |  |  | x | x | x |  |  |  |  |  |  |
| (True) coverage / PICP / ECP | 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 |  |  | x | x | x |  |  |  | x |  | x |  | x |  |  |  | x | x |
| MSIW |  |  |  |  |  | (x) | (x) | (x) | (x) |  | x |  |  | x |  |  |  |  |
| MSIWW |  |  |  |  |  |  |  |  |  |  | x |  |  | x |  |  |  |  |
| SWR |  |  |  |  |  |  |  |  |  |  | x | x |  |  |  |  | x | x |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

To eventually come to the purpose of this work, the derived PIs shall be evaluated. The following table gives an overview about possible measures found in the literature and measures that will be deployed in this work.

This 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 CLV context as well. The measures listed above are also employed in this work. Additional measures like MSIS (2L, 31, 39), NMPIL (34,35), PINAW (37) or Bias (2L) were not directly translatable to the CLV context[[6]](#footnote-6), but their idea is used in a slightly changed measure, hence they are noted in brackets. All the measures used will be explained in the section below. The bias has a different position as it does not evaluate PIs but the point forecasts.

# Applied methods

* Concrete implementation of methods to derive PIs
* Concrete implementation of measures to assess PI performance

## Conformal prediction

* Full conformal prediction is computationally intense and does not work because the model does not take true values into account while fitting the parameters, see e.g. 9CP, p.10 red marked
* -> Only split conformal prediction

Split conformal prediction

* Approach follows the one from the literature review but with adaptations
* 2 major problems in advance:
* **1. True CET/PTS is unknown** for the respective customers (CP relies on having the true values at some point)
  + Potential solution in reality: Get the quantile and standard deviations from an old cohort/panel, fit the model for the new cohort/panel and use the old quantile and standard deviations for the PIs of the new data
  + Strong assumption: New customers behave similar to the old ones
  + Not yet implemented and I don’t know if it will work
* **2. Heteroskedasticity:** “Standard deviations” (how strong true and predicted values are differing) are necessary because in CLV context, you cannot use absolute residuals:
  + customer with true total spending of $1, PTS of 21$ -> delta = $20 is bad
  + customer with true total spending of $1000, PTS of 1021$ -> delta = $20 is good
  + high values would suffer under-coverage, low values over-coverage
  + Solution (1CP, 9CP): Take sd for each customer (or any useful measure of uncertainty) and scale residuals with it
  + Procedure for getting sds
    - For (x in 1:80) do (x chosen arbitrarily, 80 is ok)
      * Split the data[[7]](#footnote-7) randomly (train and test)
      * Train the pnbd model incl. spending
      * Predict for test customers the CET and PTS
      * Take the differences to the true values

end

* + - average over the difference for each customer
    - average over the CET and PTS predictions for each customer
    - it is useful to make sds dependent on the predicted values and thus make it applicable to any customer
      * -> sd\_CET(customer\_i) = f(CET\_i)
    - f is not generally defined and can be chosen more or less arbitrarily (1CP, p.9)
    - linear coherence of CET\_i seems reasonable
      * -> fit linear model: avg(sd\_i) ~ avg(pred\_i)
* Another issue: Bias: When doing split conformal prediction then the choice of customers for training and calibration changes the result a lot, thus do it many times and average over the results, again 80 seems reasonable

**Complete Procedure:**

1. Get sd(CET\_i)

For (x in 1:80) do

* Random data split (training and validation)
* Train the model with training
* Predict CET and PTS for validation
* Collect the differences between true value and CET and the predictions

End

Calculate the mean of the collected sds, PTS and CET predictions for each customer

Fit linear model: Regress mean(sd\_i) on mean(CET\_i)

1. Get quantile q (and PIs)

* q = ceiling(((n + 1) \* (1 - alpha)))/n (formula often seen, e.g. 1CP, p.5)

For (x in 1:80) do

* Random data split (training, validation, test)
* Transfer model parameters from the trained model to the calibration and test models (to imitate the normal machine learning procedure[[8]](#footnote-8))
* Predict CET (PTS) on the calibrate set
* Collect the absolute residuals (1.)
* Scale them (residual/sd\_i) by the respective sd\_i (2.)
* Take the q-quantile of the scaled residuals (3.)
* Make point predictions on the test data set
* Get test data CET intervals by: CET point prediction\_i +- (quantile\_x \* sd\_i) (4.)
* Collect quantile\_x

End

1. Average for each customer their retrieved PI values (from when they were in the test data)
2. Managerial version:
   1. Average over all collected quantile\_x
   2. Fit a pnbd model on the new data and retrieve PIs:

PI\_i = CET point prediction\_i +- (quantile \* sd\_i)

#### Validity

The measure validity assesses how exact the prediction intervals perform. They must cover on average the true value in (1-α)\*100% of the cases to be called valid. The current approach delivers in this regard (and with the simulated test data) satisfying results when repeated often enough (what is typically not the case when only as many runs are conducted as are necessary to sample each customer at least once as described in the approach above).

#### Exchangeability

## Ensemble (former Method 1)

**Procedure** (based on [[9]](#footnote-9), online book, section 16.5: https://www.openforecast.org/adam/adamRefitted.html)

1. Fit pnbd and gg model and receive parameter estimates and respective covariance matrices
2. From the estimates and covariance matrices, simulate n draws of the parameters
3. Run the prediction with each draws
4. Receive n values for CLV prediction
5. Take the central 90% interval of the predictions

## Bayesian method

1. Estimate the pnbd model with the Bayesian approach:
   * Use the mcmc method to get the parameter posterior distribution
   * Use the mcmc method to get the draws of the posterior predictive distribution (outcomes, only CET because gamma model for spending not implemented)
2. Until here everything done by BTYDplus
3. Take the desired quantiles of the posterior predictive CET distribution for each customer (Predicted Total Spending PTS is not implemented)

## Quantile regression

* Actually: Use an optimization problem to find the parameters that minimize loss function and yield the number of predictions above/below the desired quantile
* Issue: Complex problem, solving algorithms fail
* Alternative approach:
  + Do not: Implement the loss function optimization during model estimation
  + Do: Use grid of possible parameter combinations and calculate the necessary measures (distances to the desired quantile) for each parameter combination

**Procedure:**

1. Build a grid of parameters (use prior knowledge of where you expect the parameters to be)
2. For each parameter combination
   1. Predict the CET and PTS with this parameter combination
   2. Calculate for CET (PTS) 2 “distance measures”, see 54PI, p.7 (alternatively, one can take differences to the desired quantile directly and achieve similar performance)
      1. CET Upper:
      2. CET Lower[[10]](#footnote-10):
3. Select for each desired quantile (CET/PTS Lower/Upper) the model (parameter combination) that reaches the distance closest to 0
4. Predict with the 4 selected models for all customers their intervals

* Managerial implementation: “Recycle” old models (their parameters) and apply them to the new data to get PIs (not implemented yet, I don’t know if it works)

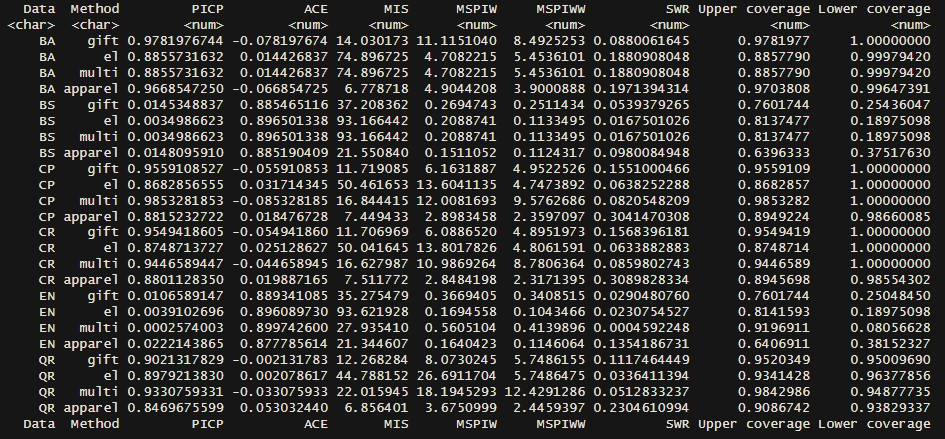
## Conceptual comparison of the methods

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **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 | Non 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** | **Academic** | Model uncertainty | Exchangeability | Yes | High | Medium | (Yes) |
| **Managerial** | Model uncertainty | Exchangeability | (Yes) | Low | Low | (Yes) |

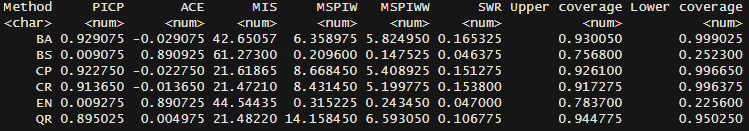
# Results

## Methods’ Performance - Benchmarking Bootstrap

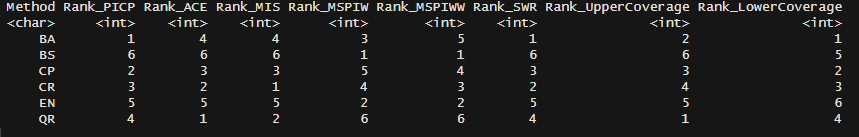
### General performance



Averaged over datasets:



Ranked:



* BS and EN have very low coverage and small interval width
* BA, CP, CR and QR have a good coverage and high interval width
* Homogenous performance across datasets for all methods for PICP
* More heterogenous for width (should be due to the model itself when PICP is constant but needs different widths for this?)
* Note on heterogeneity in the appendix
* Trade-off:
  + High coverage or small width
  + No middle way, just either or in this case
  + MIS: The lower the better because it includes penalties, QR, CR and CP are good here
  + SWR: The most sharpness per width offers BA, followed by CR and CP
* Including weights for more worthy customers decreases the width for all methods, least for BA, most for QR, see later why

A diagram of a graph

Description automatically generated

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

#### Coverage

A graph with blue and red dots

Description automatically generated

* No middle way
* No big performance differences over the data sets
* BS, EN always under coverage
* QR, CR, CP, BA appropriate values, sometimes over coverage, sometimes under coverage

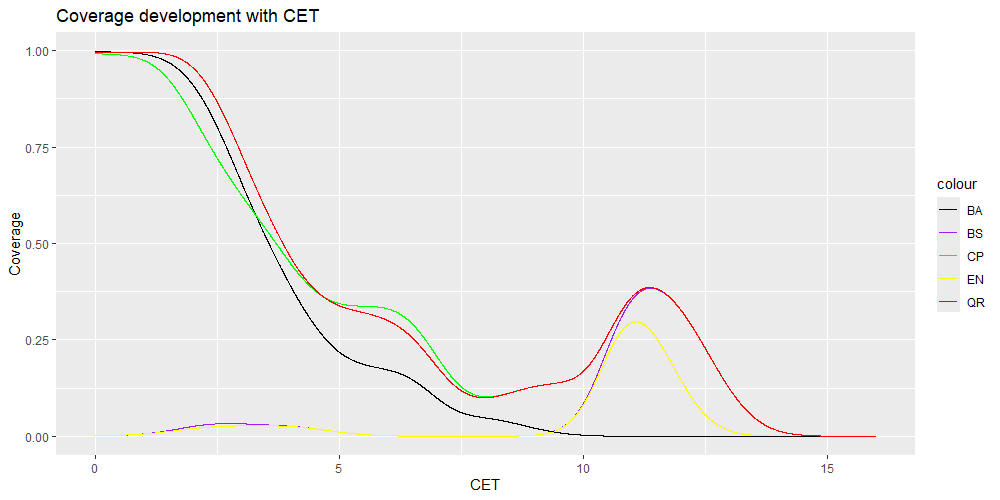
Problem of BS and EN can be visualized in the following plot

A graph of a number of numbers

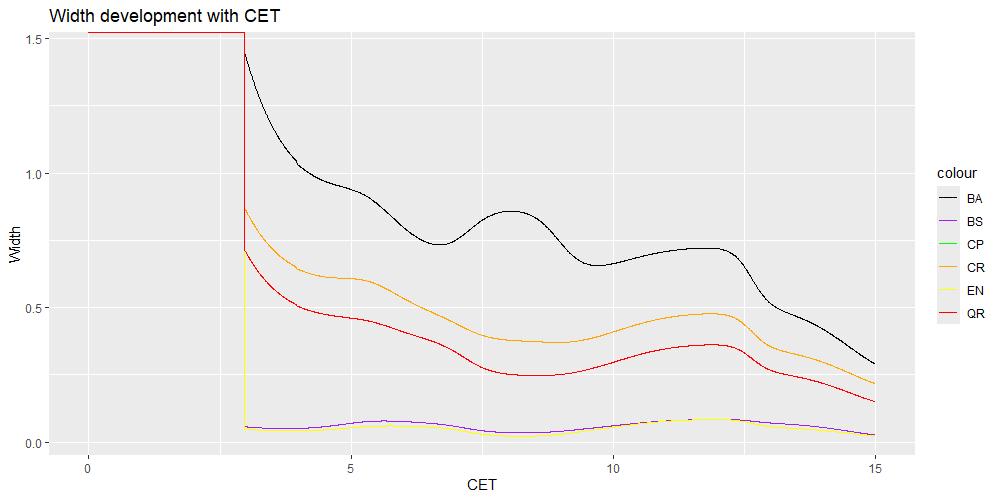
Description automatically generated with medium confidence

* Too short intervals for BS and EN, coverage only when the model is performing very well
* The other methods “behave” more like PIs, that they are large is due to the model’s performance
* Note: Bayesian intervals come from another model estimation and are therefore not centred around the actual model’s prediction

#### Adaptability and stability across CET levels



* Coverage probability depends on CET level
* BA, CP and QR perform well on lower levels and decrease rapidly for medium levels
* CP and QR have another peak 70-80% of the maximum level
* BS and EN have basically no coverage at low levels and have a peak like CP and QR at 70-80% of the maximum level

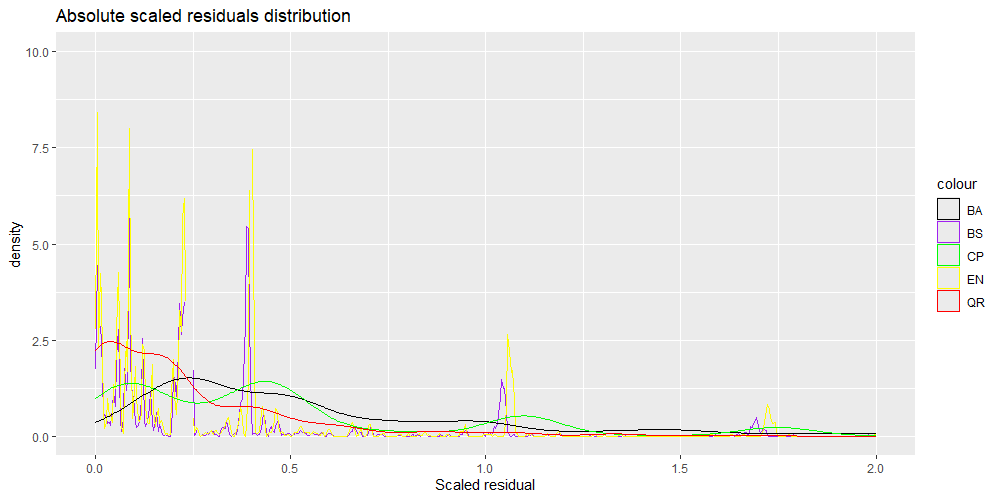


* Downward trend, the higher the CET level, the shorter the interval (relatively)/ the “surer” the methods are
* Very different across datasets but generally downward tendency
* For gift: Same across BA, CP, CR, QR

#### Distribution of residuals and predictions

The following figure shows how residuals (distance of the uncovered points to their closest interval boundary) are distributed.

* very unsteady for BS and EN, with many peaks and outliers (y and x), some even needed to be cut to keep dimensions properly
* BA, CP and QR have a rather steady distribution



How do datasets differ?

Finding the best method (for which purpose?)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Adequate coverage | Coverage across outcomes (CET, PTS) | Coverage across levels | Performance across data sets | Upper and lower coverage | Works in all examined datasets | Usefulness |
| Bootstrap | No |  | No |  |  | No |  |
| Ensemble | No |  | No |  |  | Yes |  |
| Bayesian | Yes |  | Not really |  |  | Yes |  |
| Conformal prediction | Yes |  | Acceptable |  |  | Yes |  |
| Quantile regression | Yes |  | Acceptable |  |  | Yes |  |

## Performance over varying training and prediction periods

* Comparison between methods shifted from different measures to only PICP and from one period-combination to several combinations
* Address managerial question: Performance when only limited data available especially relevant for CP and QR
* Are the methods equally good across combinations? How do they vary?
* Which combinations are good, which aren’t? (reorder table and put in appendix)

# Managerial / research discussion

Strengths, weaknesses, shortcomings, compromises, limitations

Impact on strategical planning, applicability in reality

# Conclusion

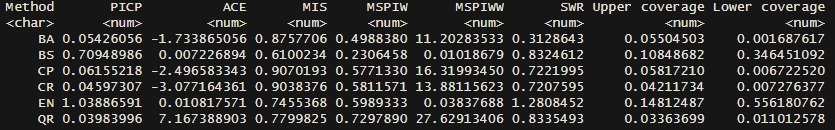
…

# References

…

# Appendix

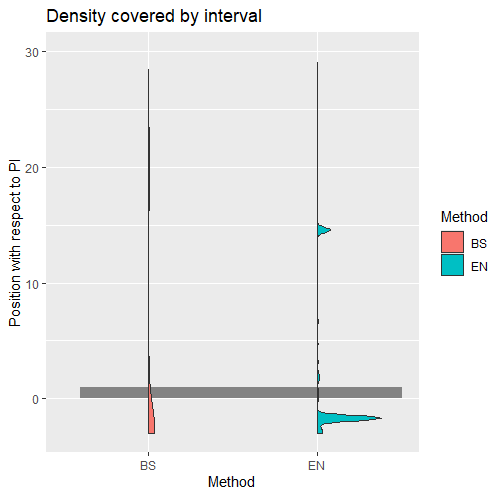
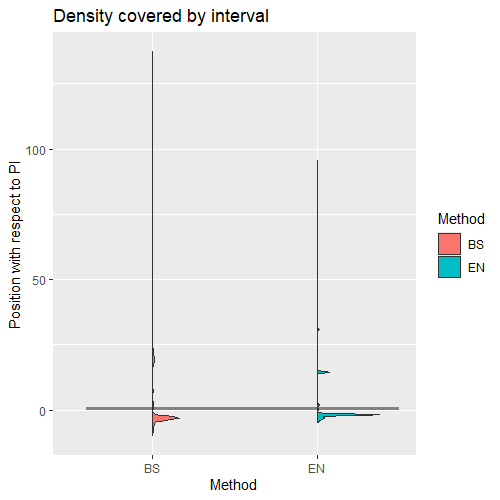
Coefficient of variation for different methods and measures across datasets



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

A graph of a graph

Description automatically generated



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) [↑](#footnote-ref-4)
5. 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-5)
6. Since the literature on benchmarking PIs in the CLV context is very limited, it was necessary to conduct research in other fields where it is more common to calculate and assess PIs. Typical examples are time series forecast competitions (31) or the prediction of wind (16, 20, 37) or electricity-related forecasts (35, 38) [↑](#footnote-ref-6)
7. A customer and its transactions can only be in either of the sets [↑](#footnote-ref-7)
8. Train the model on training data, use the model to make predictions on the test or calibration data. In CLVTools, I can to my knowledge not just predict a dataset with a model that has not been trained on these data, so I need to train on each data set the model first and then change it accordingly) [↑](#footnote-ref-8)
9. - Svetunkov, I., Pritularga, K., 2023. Incorporating Parameters Uncertainty in ETS. Department of Management Science Working Paper Series. 1–19.)

   - Svetunkov, I. (2023). Forecasting and Analytics with the Augmented Dynamic Adaptive Model (ADAM) (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781003452652> [↑](#footnote-ref-9)
10. Often the true value is 0 for both, CET and PTS, but predicting 0 is not achievable by any parameter combination but something like 0.0000214. To not let these cases destroy the approach, a small tolerance *CET\_tol* is added to the true value. It makes sense from a managerial point of view because one would barely treat a customer differently if the prediction is 0.0000214 or exactly 0. [↑](#footnote-ref-10)