Conformal methods in official statistics: perspectives and challenges

Nina Deliu, Brunero Liseo [1]

Abstract

Survey sampling and, more generally, official statistics are experiencing an important renovation time. On one hand, there is the need to exploit the huge information potentiality that the digital revolution has made available in terms of data. On the other hand, this process occurred simultaneously with a progressive deterioration of the quality of classical sample surveys, due to a decreasing willingness to participate and an increasing rate of missing responses. The switch from survey-based inference to a hybrid system involving register-based information has made more stringent the debate and the possible resolution of the design-based versus model-based approaches controversy. In this new framework, the need of statistical techniques which provide exact coverage guarantees to model-based procedures is essential. Here we explore the use of conformal prediction methods, very popular in modern statistics, yet not very common in survey statistics. We argue that this approach can be beneficial both in design-based and model-based approaches to survey sampling.

Keywords: Finite sample coverage, Design-based coverage, Frequentist-Bayesian agreement, Prediction Intervals

1. Introduction

The new era of massive data availability has dramatically modified the role that National Statistical Offices are called to play. To contrast the widespread and increasing phenomenon of "refuse to respond" in surveys, most data collection activities are today based on the integration of administrative sources and ad hoc small surveys. Among the main consequences of this paradigm shift, there is a need for robust methods to quantify the uncertainty in the data production paradigm.

Today, the question of the correct quantification of uncertainty concerns all disciplines. Still, it plays a strategic role in official statistics, especially when it comes to forecasts for demographic or economic financial quantities.

There are many different sources of uncertainty. Some of them are related to measurement errors, caused by poor data quality, an excessive non-response rate, or gaps (where data are not available on some parts of the population)

Uncertainty quantification is crucial - for example - in small-area estimation where many "predictions" are necessary at a granular level and having area-specific small sample sizes is typical. This makes direct methods unreliable: they could achieve target frequentist coverage rates for all areas but don't take advantage of auxiliary information. On the other hand, model-based methods allow information to be shared across areas. This decreases variability but can introduce bias and alter area-level frequentist coverage rates from their nominal level (Bersson and Hoff 2024)

Two main approaches have been followed in the literature to quantify the accuracy, usually measured in terms of *mean-squared error* (or variance) of an estimator. The first approach involves resampling methods and mainly focuses on either parametric or non-parametric

[1]Nina Deliu (nina.deliu@uniroma1.it), Sapienza Università di Roma, Italy; Brunero Liseo (brunero.liseo@uniroma1.it), Sapienza Università di Roma, Italy.

bootstrap-based strategies (see e.g., Mashreghi *et al.* 2016; Shao 2003, for a survey). These heavily depend on the amount of data to be processed (which can be millions and millions in data from a register) and can lead to extreme computational burden. The second approach relies on approximations based on Taylor linearisation (see e.g., Graf and Tillé 2014; Vallée and Tillé 2019), and requires the effort to derive analytical formulas case by case.

However, the common bottleneck of all these methods is that their coverage guarantee is only restricted to large samples and reliable estimation of prediction confidence is a major challenge in all areas of statistics, but it is particularly crucial in the production of "official predictions".

In this paper, we explore and discuss a general method that guarantees the exact coverage of prediction intervals even for finite sample sizes, the so-called *Conformal Prediction* methodology (Vovk *et al.* 2005; Lei and Wasserman 2014). Conformal Prediction (CP, hereafter) is a relatively new method of making predictions with guaranteed confidence; it has been introduced and developed during the last two decades. CP allows the construction of prediction sets with the guaranteed error rate, exclusively under the simple *i.i.d.* assumption for the observations in the sample.

The rest of the paper is organised as follows: Section 2 introduces the terminology and outlines the framework within which we discuss CP. Section 3 describes in some detail the CP method, in its two main declinations, the **Full** CP and the **Split** CP. While the Full CP is the best approach to consider, its computational burden can be extremely heavy. This has suggested some practical modifications, such as Split CP, see below for details. Section 4 presents the results of a large simulation study where several competing approaches were tested on the data analysis of a publicly available dataset, `api`, which can be found in the **R** package `survey` (Lumley 2024).

Finally, Section 5 is a discussion on the limitations of CP, its role in more complex (non-exchangeable) sampling schemes, and the links with a Bayesian approach.

2. Problem formulation

Let U_N denote a finite population of N statistical units, with N fixed and known at a given moment of interest. Each unit j is characterised by a set of data $(\boldsymbol{X}_j, Y_j)$, where $\boldsymbol{X}_j \in \mathcal{X} \subseteq \mathbb{R}^p$, $p \geq 1$, represents the set of auxiliary variables and Y_j is a response variable of interest, for $j = 1, \ldots, N$. Typically, $\boldsymbol{X}_j$ is assumed to be fixed (and known) and the interest is in the response variable, which we assume to be continuous, i.e., $Y_j \in \mathbb{R}$, $j = 1, \ldots, N$. In official statistics, a common interest is in the population mean or total, often concerning a specific domain d. For example, in this work, we will focus on the academic performance index (API) score in a given district d in California. In these cases, the population mean will be the target parameter:

$$\theta^{(d)} = \frac{1}{N} \sum_{j=1}^{N} \gamma_j^{(d)} Y_j, \tag{1}$$

where $\gamma_j^{(d)} \in \{0, 1\}$ is the known domain membership indicator, with 1 indicating that unit j belongs to domain d.

It is important to note that $\theta^{(d)}$ is typically an unknown parameter as the values of $Y_j, j = 1, \ldots, N$, are only known for a subset of the N units at the given moment of interest and/or they are prone to some form of error (e.g., measurement error). In practice, surveys are conducted in order to observe the outcomes for a random sample $\mathcal{S}_n$ of size $n < N$, based on which inference is made on the full population U_N. We thus introduce a random indicator vector of size N to denote the sample membership: $\boldsymbol{I} = (I_1, \ldots, I_j, \ldots, I_N)^T$, with $I_j \in \{0, 1\}$ and $\sum_{j=1}^{N} I_j = n$.

In probabilistic surveys, the sample membership is defined according to a probabilistic sampling design, say P_I, that specifies the probability π_j for each unit $j = 1, \ldots, N$ to belong to the sample $\mathcal{S}_n$. In such a context, each unit j is thus characterised by the sample-covariate-outcome triplet with joint distribution:

$$(I_j, (\boldsymbol{X}_j, Y_j)) \sim P = P_I \times P_{(\boldsymbol{X},Y)|I}, \quad j = 1, \ldots, N. \tag{2}$$

An estimate of $\theta^{(d)}$ is performed on the basis of the sample $\mathcal{S}_n = \{j : I_j = 1\}$ and the associated (sample) data

$$\text{Data}_n^{\texttt{Sample}} = \{(\boldsymbol{X}_j, Y_j) : j \in \mathcal{S}_n\},$$

used for making inference and getting the predictions $\hat{Y}_j$ of the unobserved units $j \notin \mathcal{S}_n$. The generic estimator is thus given by:

$$\hat{\theta}^{(d)} = \frac{1}{n} \sum_{j \in \mathcal{S}_n} \gamma_j^{(d)} Y_j + \frac{1}{N-n} \sum_{j \notin \mathcal{S}_n} \gamma_j^{(d)} \hat{Y}_j, \tag{3}$$

and an illustrative sketch of the data format is presented in Table 2.1.

Table 2.1 – Example of the data format

Unit	Sample Membership I	Covariate X_1	...	Covariate X_p	Outcome Y_1
1	$i_1 = 1$	x_{11}	...	x_{1p}	y_1
2	$i_2 = 0$	x_{21}	...	x_{2p}	$\hat{y}_2$
...	...	...	...	...	...
j	$i_j = 1$	x_{j1}	...	x_{jp}	y_j
...	...	...	...	...	...
N	$i_N = 0$	x_{N1}	...	x_{Np}	$\hat{y}_N$

Source: authors' elaboration

Clearly, valid and accurate predictions for the unobserved units will lead to a valid and accurate estimator of the target population parameter (as well as its uncertainty). This work is concerned with this task and it aims to bring in the framework of conformal inference for valid uncertainty quantification into survey sampling.

3. Conformal prediction

Conformal inference represents a flexible yet robust framework that is attracting considerable attention in modern statistics. It is, in principle, used for quantifying the uncertainty in the predictions made by arbitrary prediction algorithms, and its underlying idea is quite simple. The basic theory behind conformal inference stems from the relationship between exchangeable random variables, rank statistics, and sample quantiles. Specifically, if $U_1, \ldots, U_n$ and U_{n+1} are exchangeable (e.g., i.i.d.) realisations of a scalar random variable, then the rank of e.g., U_{n+1} (among $U_1, \ldots, U_n, U_{n+1}$) is uniformly distributed over $\{1, \ldots, n+1\}$. Furthermore, for a given miscoverage level $\alpha \in (0, 1)$, we have that:

$$\mathbb{P}(U_{n+1} \leq q_{n,1-\alpha}) \geq 1 - \alpha, \tag{4}$$

where $q_{n,1-\alpha}$ is the sample quantile of $U_1, \ldots, U_n$ and is given by

$$q_{n,1-\alpha} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)}, & \text{if } \lceil (n+1)(1-\alpha) \rceil \leq n, \\ \infty, & \text{if } \lceil (n+1)(1-\alpha) \rceil = n+1, \end{cases}$$

where $U_{(j)}$ denotes the j-th ordered sample element and $\lceil \cdot \rceil$ is the ceiling function.

In our specific problem setup, if we assume exchangeable samples $Z_j = (I_j, (\boldsymbol{X}_j, Y_j)) \sim P$, in all the population units $j = 1, \ldots, N$, then, we have the following guarantee:

$$\mathbb{P}(Y_j \in \mathcal{C}_{n,1-\alpha}(\boldsymbol{X}_j)) \geq 1 - \alpha, \quad \forall j \notin \mathcal{S}_n, \tag{5}$$

if $\mathcal{C}_{n,1-\alpha}(\cdot)$ is a conformal prediction interval. Here, $\mathcal{C}_{n,1-\alpha}(\boldsymbol{X}_j)$ denotes the $(1-\alpha)$-level conformal prediction interval for the response variable Y_j of a non-observed unit $j \notin \mathcal{S}_n$ with covariate information $\boldsymbol{X}_j$. It is indexed by n to emphasise that its computation is based on the n sample response data $Y_j, j \in \mathcal{S}_n$.

One of the main features of conformal prediction, compared to e.g., asymptotic prediction intervals, resampling methods, or a mere adoption of the sample quantile property in Eq. (4), is the introduction of a *conformity measure* to quantify the similarity or conformity of a certain (new) given point to the observed (sample) data. Assuming a symmetric conformity measure, say r, the conformity scores $R_1 = r(Z_1), \ldots, R_N = r(Z_N)$ are the main working quantities invoked in the process. Exchangeability of $Z_1, \ldots, Z_N$ and the symmetry of r ensure exchangeability in the set $\{R_1, \ldots, R_N\}$ and makes it possible to extend the result in Eq. (4) for the construction of prediction intervals. Two approaches are now outlined for this purpose: split conformal prediction and full conformal prediction. In both cases, assuming - for the sake of clarity - a regression framework with continuous Y, we take as the conformity measure the absolute fitted residuals, that is, $R_j = |Y_j - \hat{f}_n(\boldsymbol{X}_j)|$, where $\hat{f}_n \colon \mathcal{X} \to \mathbb{R}$ is a point estimator of the underlying regression model fitted on the n data points. A plethora of alternative conformity measures have been introduced to address specific problems; their main goal is to adapt the size of the prediction interval to the available local information, see e.g., Romano *et al.* (2019). We briefly discuss this issue in Section 5.

Full Conformal Prediction. The original idea of conformal prediction is rooted in what is today referred to as full conformal prediction. Although this approach is less commonly adopted in practice due to its computational complexity, it nonetheless remains an elegant and robust methodology, often leading to more efficient results.

In essence, the construction of a full conformal prediction interval for a unit $j \notin \mathcal{S}_n$, denoted by j^* for simplicity, involves evaluating the inclusion or acceptance of a set of *candidates* $y \in \mathcal{Y} \subseteq \mathrm{R}$, where $\mathcal{Y}$ is a reasonable grid for the response variable of interest Y. Fixing the covariate information $\boldsymbol{X}_{j^*}$ for unit $j^* \notin \mathcal{S}_n$, the acceptance of the candidate y is evaluated according to its conformity score $R((\boldsymbol{X}_{j^*}, y))$ compared with the conformity scores of the full set of sample data $(\boldsymbol{X}_j, Y_j), j \in \mathcal{S}_n$. To ensure exchangeability among the $n+1$ conformity scores, the predictor $\hat{f}_{j^*}$ is trained on the *augmented train* data $\{(\boldsymbol{X}_j, Y_j)_{j \in \mathcal{S}_n}, (\boldsymbol{X}_{j^*}, y)\}$, leading to:

$$\text{Calibration scores:} \qquad R_j = |Y_j - \hat{f}_{n+1}(\boldsymbol{X}_j)|, \qquad j \in \mathcal{S}_n, \tag{6}$$

$$\text{Candidate score:} \qquad R_{j^*} = |y - \hat{f}_{j^*}(\boldsymbol{X}_{j^*})|, \qquad j^* \notin \mathcal{S}_n, \tag{7}$$

We term the residuals in Eq. (6) *calibration scores* to enhance resemblance and comparability with the split conformal prediction that will be shortly introduced. Their role is in the derivation of a valid sample quantile (i.e., which leads to coverage guarantees) as in Eq. (5).

Given α, the inclusion of the candidate y in the full conformal prediction interval for unit j^*, say $\mathcal{C}^{\text{full}}_{n,1-\alpha}(\boldsymbol{X}_{j^*})$, is determined according to its score rank when compared to the calibration scores:

$$y \in \mathcal{C}^{\text{full}}_{n,1-\alpha}(\boldsymbol{X}_{j^*}) \quad \iff \quad R_{j^*} \leq q_{n,1-\alpha} = R_{(\lceil (n+1)(1-\alpha) \rceil)} \text{ among } \{R_j\}_{j \in \mathcal{S}_n},$$

and the $1-\alpha$ full conformal prediction interval for unit j^* is therefore given by:

$$\mathcal{C}^{\text{full}}_{n,1-\alpha}(\boldsymbol{X}_{j^*}) = \{y\colon R_{j^*} \leq R_{(\lceil (n+1)(1-\alpha) \rceil)} \text{ among } \{R_j\}_{j\in\mathcal{S}_n}\}. \tag{8}$$

By construction, the conformal prediction interval in Eq. (8) guarantees valid *finite-sample* coverage. Moreover, full conformal prediction generally produces accurate intervals with coverage levels closely aligning with the nominal target $1-\alpha$, while avoiding significant over-coverage.

Split Conformal Prediction. For a given non-sample unit $j^* \notin \mathcal{S}_n$ with covariate $\boldsymbol{X}_{j^*}$, full conformal prediction requires refitting the predictor $\hat{f}_{n+1}(\boldsymbol{X}_{j^*})$ on each candidate $y \in \mathcal{Y}$. Depending on the predictive algorithm used, the dimensionality of the grid, and the number of non-sample units, this process can easily become computationally prohibitive. Under these premises, split conformal prediction has been introduced as a more feasible yet valid alternative to the original conformal procedure.

The concrete idea of split conformal prediction can be summarised as follows. First, the (labelled) sample set $\text{Data}_n^{\texttt{Sample}}$ is partitioned into two subsets of approximately same sizes n_1 and n_2, with $n_1 + n_2 = n$, called *train set* and *calibration set*, respectively; that is,

$$\text{Data}_n^{\texttt{Sample}} = \text{Data}_{n_1}^{\texttt{Train}} \cup \text{Data}_{n_2}^{\texttt{Cal}}, \qquad \text{with } \text{Data}_{n_1}^{\texttt{Train}} \cap \text{Data}_{n_2}^{\texttt{Cal}} = \emptyset.$$

The first set $\text{Data}_{n_1}^{\texttt{Train}}$ is then used to fit the predictive strategy $\hat{f}_{n_1}$, while the second set $\text{Data}_{n_2}^{\texttt{Cal}}$ has the key role of determining the calibration scores as in Eq. (6):

$$\text{Calibration scores}\colon \qquad R_j = |Y_j - \hat{f}_{n_1}(\boldsymbol{X}_j)|, \qquad j \in \text{Data}_{n_2}^{\texttt{Cal}}, \tag{9}$$

and allowing for the computation of a valid quantile $q_{n,1-\alpha}$. The validity of the quantile is guaranteed by the sample-split procedure, which, conditioned on $\text{Data}_{n_1}^{\texttt{Train}}$, treats all the points in the calibration set equally, as none of these is used in the fitting procedure. In this way, the exchangeability of the calibration scores is preserved. The split conformal prediction interval for a new observation $j^* \notin \mathcal{S}$ is then obtained as:

$$\mathcal{C}^{\text{split}}_{n,1-\alpha}(\boldsymbol{X}_{j^*}) = \left[\hat{f}_{n_1}(\boldsymbol{X}_{j^*}) - q_{n,1-\alpha},\ \hat{f}_{n_1}(\boldsymbol{X}_{j^*}) + q_{n,1-\alpha}\right], \tag{10}$$

where $q_{n,1-\alpha} = R_{(\lceil (n_2+1)(1-\alpha) \rceil)}$ amongst the set $\{R_j;\, j \in \text{Data}_{n_2}^{\texttt{Cal}}\}$.

Interestingly and crucially from a pragmatic perspective, split conformal prediction does not require repeated refitting of the predictor, which is anchored to the train set $\text{Data}_{n_1}^{\texttt{Train}}$, with substantial gains in memory and computation expense. Furthermore, it also provides a preliminary "in-sample" evaluation of the coverage, which is approximately guaranteed. In fact, as data in $\text{Data}_{n_2}^{\texttt{Cal}}$ are exchangeable with those in $\text{Data}_{n_1}^{\texttt{Train}}$, the former can be used as an "in-sample" test set to evaluate coverage without overfitting issues. In contrast, in full conformal prediction, first, all units in $\text{Data}_n^{\texttt{Sample}}$ are used in the fitting process, and secondly, and most importantly, the candidate points $y \in \mathcal{Y}$ do not adhere to exchangeability (they are generally sampled from a uniform grid). This makes full conformal prediction unsuitable for coverage evaluation using the sample data, necessitating additional (test) data points for assessing such a guarantee.

3.1 Conformal inference under the design- vs. model-based frameworks

Design-based conformal inference The role of conformal prediction in design-based sampling has been addressed by (Wieczorek 2023), where the most common sampling designs

are discussed in detail and the problem of the lack of exchangeability among the units in sampling schemes different from the standard simple random sampling with replacement (SRS-WR) is addressed through the notion of *covariate shift*, which we discuss in Section 5. However, design-based inference is already tailored to produce unbiased point estimates and, when possible, well-calibrated confidence intervals. The potential gain of using CP in a design-based framework is that CP provides a *finite-sample size* exact coverage rather than invoking asymptotic results. Using the notation introduced in Section 2, the design-based CP ensures the following coverage guarantees

$$\mathbb{P}_I(Y_{j^*} \in \mathcal{C}(\boldsymbol{X}_{j^*})) \geq 1 - \alpha, \quad j^* \notin \mathcal{S}_n.$$

where the probability is computed over the randomness in the sampling design, and related sample-membership variable I, while conceiving the unit characteristics (both $\boldsymbol{X}$ and Y) as fixed quantities.

Model-based conformal inference The model-based approach, on the other hand, has not been addressed in the survey sampling literature. The main shift from the former is in the conceived randomness, which is now attributed to $(\boldsymbol{X}, Y)$, assumed to be generated according to a super-population model. In this case, conformal inference comes with the guarantee:

$$\mathbb{P}_{(\boldsymbol{X},Y)}(Y_{j^*} \in \mathcal{C}(\boldsymbol{X}_{j^*})) \geq 1 - \alpha, \quad j^* \notin \mathcal{S}_n.$$

Under this framework, we argue that the use of a CP approach can be greatly beneficial for at least two reasons outlined below:

- The use of a model in finite population inference is often criticised because of the risk of misspecification and - more generally - because of the use of untestable assumptions: a model-based inference *corrected* by a conformal step greatly reduces this kind of risk and, at the same time, guarantees the correct coverage.
- Given the exact coverage, one can simply choose *among* alternative approaches, either design-based or model-based, in terms of the average length of the resulting prediction intervals. As we will illustrate in the empirical study (Section 4), a model-based approach comes with remarkable advantages in terms of length efficiency.

4. Empirical evaluation

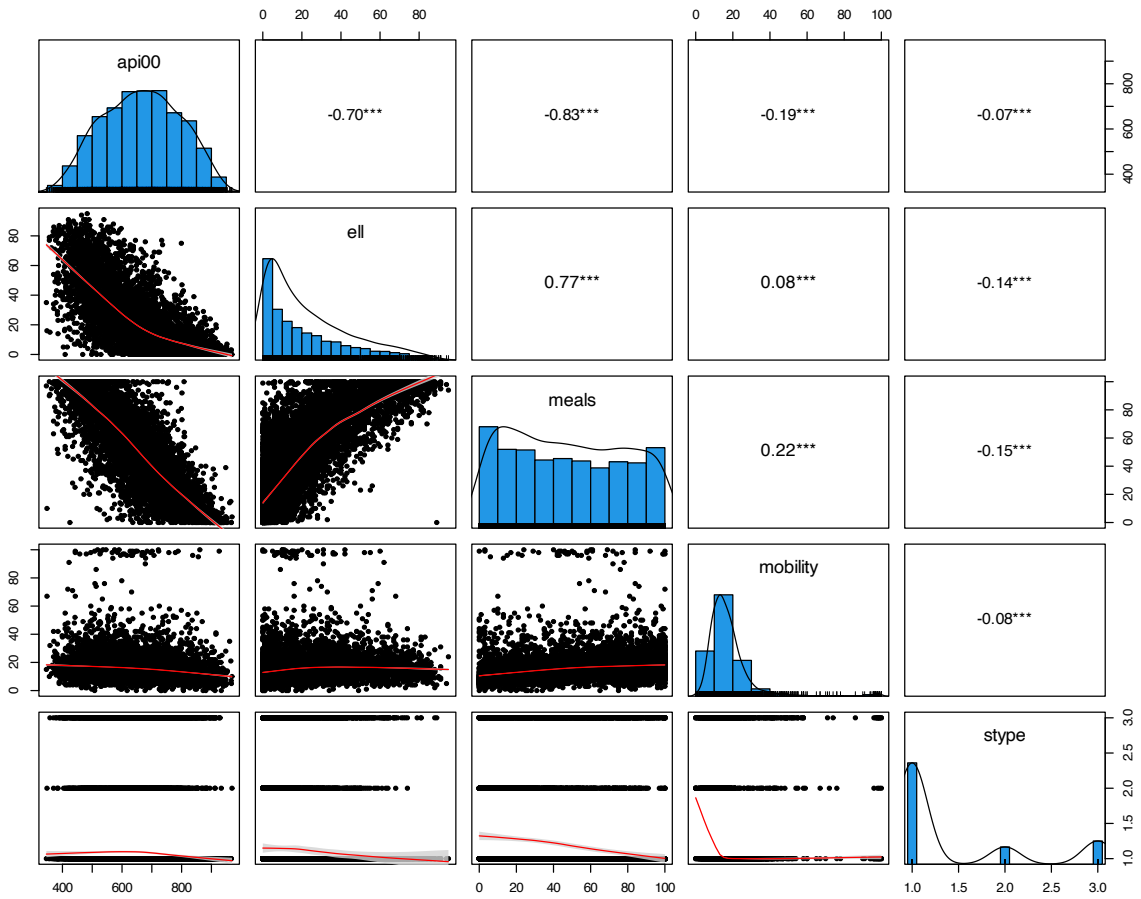
This section illustrates the application of conformal inference methods – in comparison to more traditional techniques – to the `apipop` dataset. These data have been widely used in education and research as a tool for evaluating statistical methodologies in survey sampling, and are available in the `survey` package in `R`. They provide information on the Academic Performance Index (API), which serves as a measure of academic performance designed to track the progress in educational outcomes in California over time.

More in detail, the dataset contains info on the API score for the year 2000, covering all California schools with at least 100 students, for a total of $N = 6194$ units. It also includes a set of auxiliary variables such as school type, and enrollment size, as well as information on students' characteristics (e.g., English learners, disabilities, and socioeconomic disadvantages). In this work, we will take into account the following variables, whose graphical representation is provided in Figure 4.1:

Y :: `api00` Numeric response variabile representing the API score (range: 200 to 1000)

X_1 :: `stype` Categorical variable representing the school type (elementary, middle, high)

Figure 4.1 - Summary of the set of `apipop` variables analysed in this work.



Source: Own computation

X_2 :: `ell` Numeric variable given by the percentage of English Language Learners

X_3 :: `meals` Numeric variable being the percentage of students eligible for subsidised meals

X_4 :: `mobility` Numeric variable for the percentage of first-year students at the school

The objective of the application study is to construct prediction intervals for the outcome variable Y for the set of units (schools) not included in the sample. In this work, we consider two sample sizes $n \in \{200, 500\}$, which are extracted according to a simple random sampling design with replacement (SRS-WR) from the full dataset of size $N = 6194$. To assess the properties of different methods for interval construction, population but non-sample units ($U_N \setminus \mathcal{S}_n$) will be used with the role of a test set:

$$\text{Data}_{N-n}^{\texttt{Test}} = \{(\boldsymbol{X}_{j^*}, Y_{j^*}) : j^* \in U_N \setminus \mathcal{S}_n\}.$$

Specifically, two measures will be evaluated to assess the performance of the different statistical approaches: coverage (in the test set) and length of the prediction interval. These can

be regarded as a measure of validity (coverage) and efficiency (length) and are defined as follows.

(Test-set) coverage Let $\mathcal{C}^{\texttt{rule}}_{n,1-\alpha}(\boldsymbol{X}_{j^*})$ denote the $(1-\alpha)$ prediction interval for Y_{j^*} obtained with method `rule` using $\text{Data}^{\texttt{Sample}}_n$. Then, the test-set coverage is defined as:

$$\text{Cov}^{\texttt{rule}} = \frac{1}{N-n} \sum_{j^* \notin \mathcal{S}_n} \mathbb{1}(Y_{j^*} \in \mathcal{C}^{\texttt{rule}}_{n,1-\alpha}(\boldsymbol{X}_{j^*})), \quad (\boldsymbol{X}_{j^*}, Y_{j^*}) \in \text{Data}^{\texttt{Test}}_{N-n},$$

where $\mathbb{1}(\cdot)$ represents the indicator function.

Length Let $\mathcal{C}^{\texttt{rule}}_{n,1-\alpha}(\boldsymbol{X}) = [L^{\texttt{rule}}_{n,1-\alpha}(\boldsymbol{X}), U^{\texttt{rule}}_{n,1-\alpha}(\boldsymbol{X})]$ denote the lower and upper bounds, respectively, of the prediction interval obtained with the method `rule` for a unit with covariate $\boldsymbol{X}$. Then, the associated length is defined as:

$$\mathcal{L}^{\texttt{rule}}(\boldsymbol{X}) = U^{\texttt{rule}}_{n,1-\alpha}(\boldsymbol{X}) - L^{\texttt{rule}}_{n,1-\alpha}(\boldsymbol{X}).$$

If one has access to a test set $\text{Data}^{\texttt{Test}}_{N-n}$, as an overall length performance, one may consider the "marginal" length (e.g., averaging over the individual prediction intervals) or some "conditional" alternative (e.g., for a given domain d such as school type).

The different performance measures will be evaluated according to a number of $M = 1000$ independent replications $\mathcal{S}^{(1)}_n, \ldots, \mathcal{S}^{(M)}_n$ of the considered sampling design. We will let both n and α vary, with $n \in \{200, 500\}$ and $\alpha \in \{0.02, 0.05\}$.

Compared approaches The focus of this evaluation is on conformal prediction methods, with a greater emphasis on split-conformal prediction, given its simpler structure, more efficient computation, and wide adoption. The conformal approach will be compared to traditional approaches, which include the design-based approach, the model-based approach, and the bootstrap approach. Model-based approaches (both traditional and conformal) will account for the following three different linear models:

Model M1 The response variable Y is regressed on the categorical predictor X_1, with poor predictive abilities (correlation $\rho = -0.07$; Figure 4.1).

Model M2 The response variable Y is regressed on the numerical predictor X_2, with good predictive abilities (correlation $\rho = -0.70$; Figure 4.1).

Model M3 The response variable Y is regressed on the full set of four covariates $X_1, \ldots, X_4$.

An additional extension of conformal prediction aimed to improve conditional (i.e., domain-specific) performance will be presented in Section 5.1.

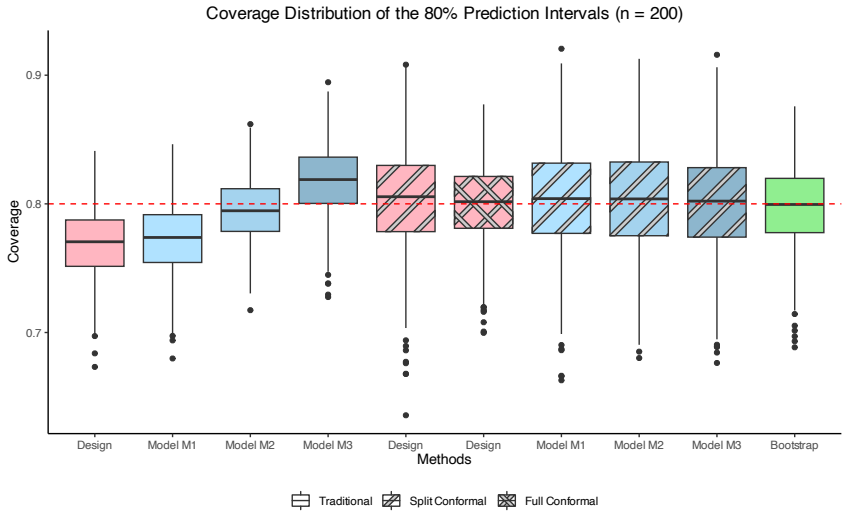
4.1 Results

Empirical results on coverage performance when $n = 200$ are reported in Figures 4.2 and 4.3, for all methods and for the two miscoverage levels, respectively.

When $\alpha = 20\%$, the traditional design-based and model-based (with M1) approaches do not ensure proper 80% coverage, with a substantial undercoverage in the case of M1, a slight undercoverage for M2, and over-coverage for M3. In all these cases, conformal prediction ensures nearly exact coverage, with the most accurate performance achieved for the full conformal prediction. Proper coverage is achieved also by the traditional non-parametric

bootstrap approach, but these become less optimal when $\alpha = 0.05$. As illustrated in Figure 4.3, when $\alpha = 0.05$, the bootstrap approach undercovers, while other traditional approaches are characterised by substantial (design-based, model-based M1) or mild (model-based M2) overcoverage. This pattern can be explained by the distributional characteristics of the response variable Y, which shows a platykurtic behaviour, with an empirical kurtosis much lower than that of the normal distribution (note that normality serves as a basis for the underlying theory of traditional approaches). When compared to the standard normal distribution, the empirical quantile of the standardised outcome distribution at $\alpha/2$ is much higher when $\alpha = 5\%$ (that is, -1.84 vs. -1.96 of a standard normal) and this explains the overcoverage in this case (Figure 4.2). On the contrary, for $\alpha = 20\%$, its empirical quantile is much lower than the one of a standard normal (-1.36 vs. -1.28, respectively), explaining the over-coverage in the former case (Figure 4.3). Interestingly, in all these cases, conformal prediction approaches ensure proper coverage with a near-optimal (exact) coverage in the case of full conformal prediction.

Figure 4.2 - Expected coverage with the compared methods, for a target coverage $1 - \alpha = 80\%$ (red dashed line). Results are based on $M = 1000$ independent generations of an SRS-WR with sample size $n = 200$ from the `apipop` dataset with population size $N = 6194$.



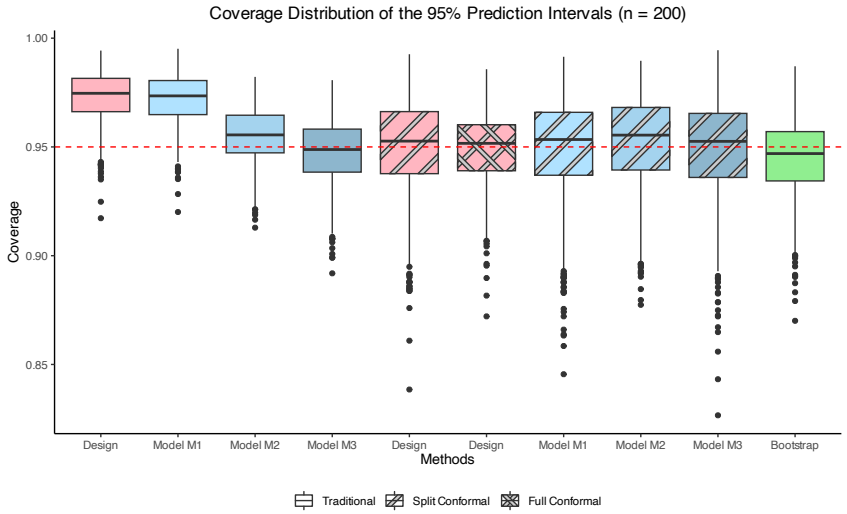
Source: Own computation

Moving now to efficiency evaluation, Figures 4.4 and 4.5 provide an illustrative summary of the size of the different prediction intervals along with their average length.

As expected, an increase in the length is observed for $\alpha = 5\%$, however, the pattern remains quite similar for the two α levels. In the specific, the non-parametric bootstrap and all design-based methods are characterised by wider intervals with an increased expected length. Independently of the traditional vs. conformal approach, the availability and the use of good predictive covariates (e.g., M2 and M3) lead to enhanced efficiency (and lower prediction uncertainty). This suggests model-based approaches as a more suitable methodology for the construction of prediction intervals. While concerning efficiency, the standard conformal framework shows comparable performance to traditional model-based approaches, there exist simple (or more sophisticated) variations of conformal prediction that can greatly outperform their comparators. This will be illustrated in Section 5.1.

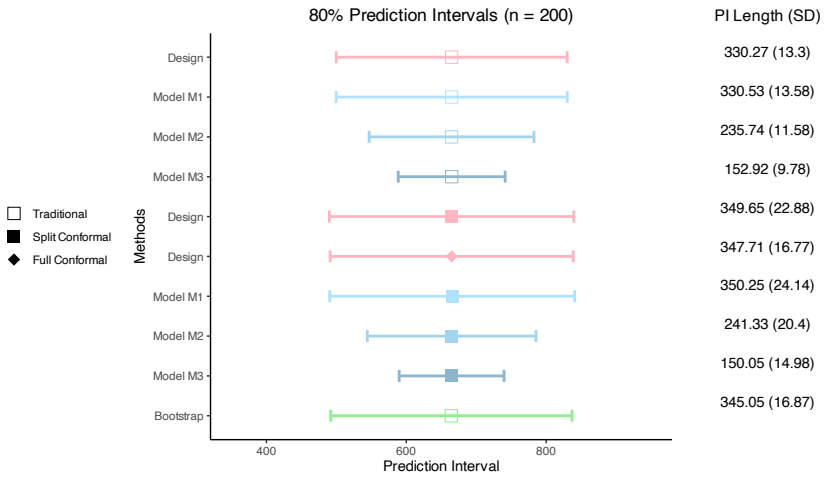
Results for $n = 500$ show a similar pattern in terms of both validity and efficiency (with a much lower degree of uncertainty in their estimates; compare e.g., the prediction length's

Figure 4.3 - Expected coverage with the compared methods, for a target coverage $1 - \alpha = 95\%$ (red dashed line). Results are based on $M = 1000$ independent generations of an SRS-WR with sample size $n = 200$ from the `apipop` dataset with population size $N = 6194$.



Source: Own computation

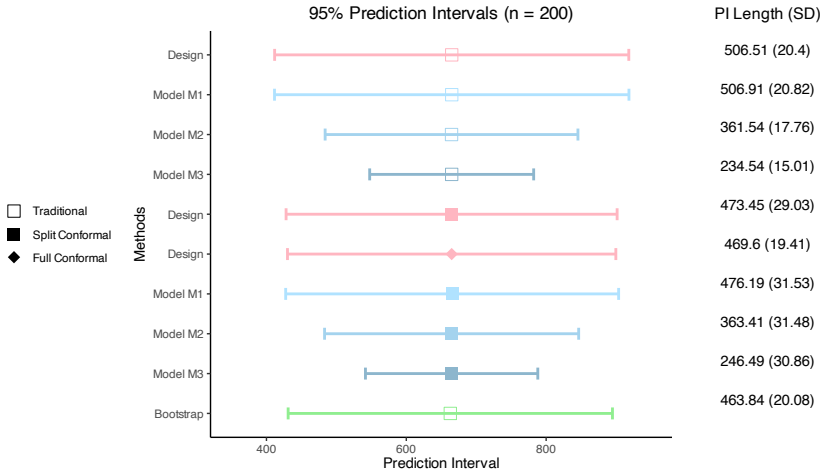
Figure 4.4 - Expected prediction interval, along with their length (standard deviation; SD) with the compared methods, for a target level $\alpha = 20\%$. Results are averaged across $M = 1000$ independent generations of an SRS-WR with sample size $n = 200$ from the `apipop` dataset with population size $N = 6194$.



Source: Own computation

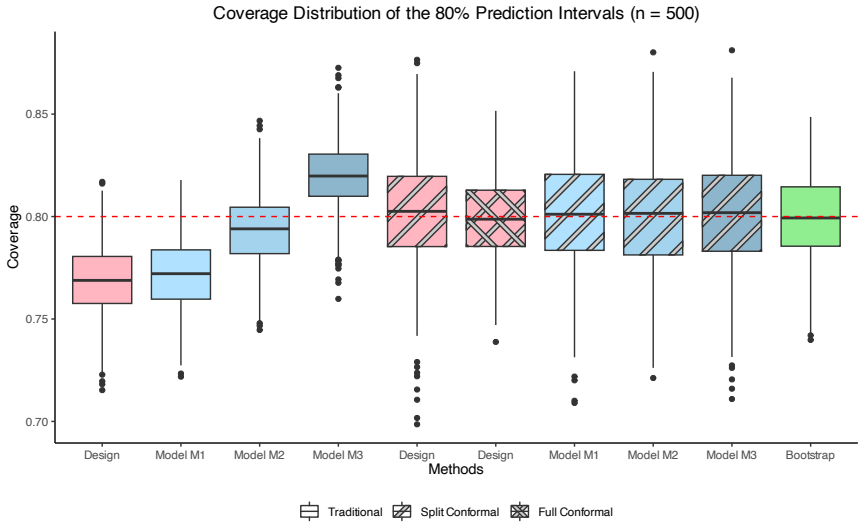
SD in Figure 4.4 and Figure 4.8). Overall, conformal prediction represent a valid inferential framework with finite-sample coverage guarantees, which are not always met in other traditional methods. Furthermore, they can be coupled with an underlying model to ensure increased efficiency in terms of the derived prediction interval's length.

Figure 4.5 - Expected prediction interval, along with their length (standard deviation; SD) with the compared methods, for a target coverage $1 - \alpha = 95\%$ (red dashed line). Results are averaged across $M = 1000$ independent generations of an SRS-WR with sample size $n = 200$ from the `apipop` dataset with population size $N = 6194$.



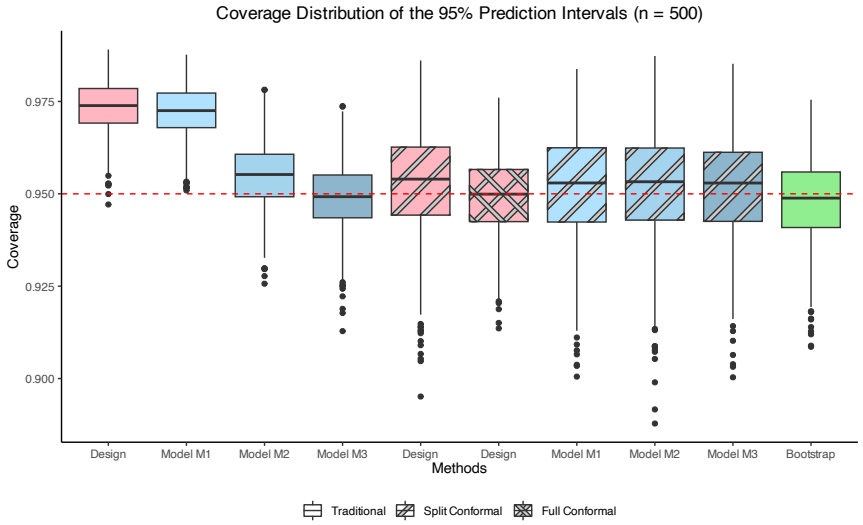
Source: Own computation

Figure 4.6 - Expected coverage with the compared methods, for a target coverage $1 - \alpha = 80\%$ (red dashed line). Results are based on $M = 1000$ independent generations of an SRS-WR with sample size $n = 500$ from the `apipop` dataset with population size $N = 6194$.



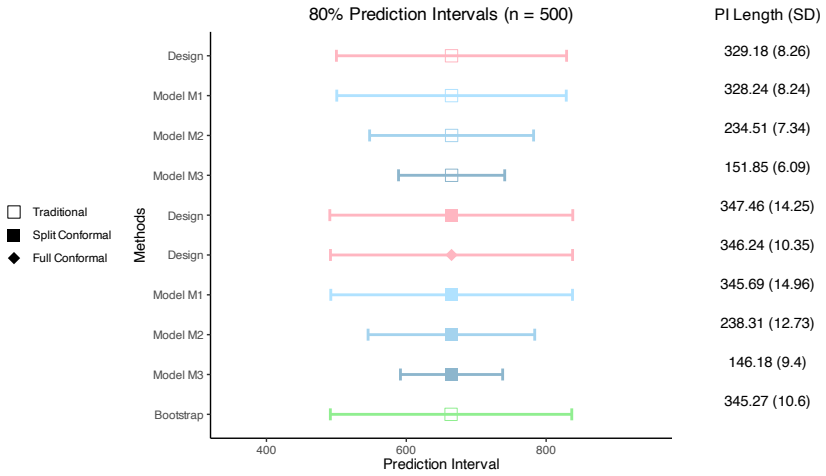
Source: Own computation

Figure 4.7 - Expected coverage with the compared methods, for a target coverage $1 - \alpha = 95\%$ (red dashed line). Results are based on $M = 1000$ independent generations of an SRS-WR with sample size $n = 500$ from the `apipop` dataset with population size $N = 6194$.



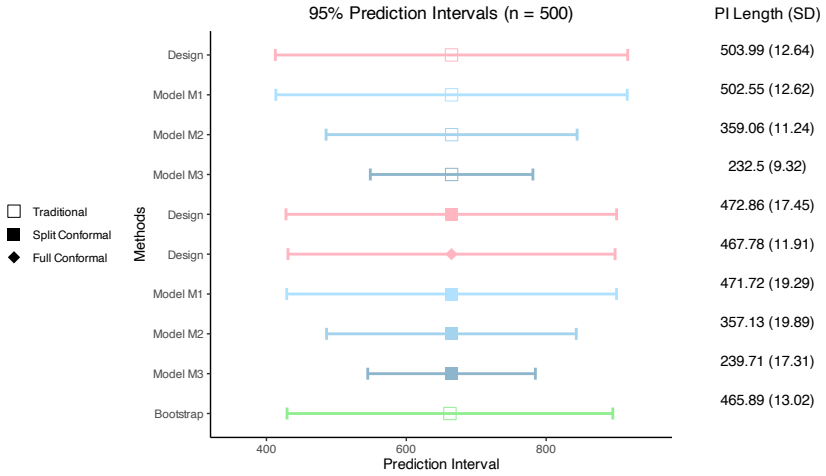
Source: Own computation

Figure 4.8 - Expected prediction interval, along with their length (standard deviation; SD) with the compared methods, for a target level $\alpha = 20\%$. Results are averaged across $M = 1000$ independent generations of an SRS-WR with sample size $n = 500$ from the `apipop` dataset with population size $N = 6194$.



Source: Own computation

Figure 4.9 - Expected prediction interval, along with their length (standard deviation; SD) with the compared methods, for a target level $\alpha = 5\%$. Results are averaged across $M = 1000$ independent generations of an SRS-WR with sample size $n = 500$ from the `apipop` dataset with population size $N = 6194$.



Source: Own computation

5. Conformal prediction in survey sampling: challenges and directions

5.1 Conditional Coverage

A closer look at the theoretical results regarding conformal prediction indicates that the frequency guarantee of exact coverage is only *marginal*, i.e., on the sample space $\mathcal{X} \times \mathcal{Y}$. In other words, there is no guarantee of obtaining exact conditional coverage for a specific value of $x \in \mathcal{X}$. However, this is not a peculiar problem of the conformal methods. Lei and Wasserman (2014) show that no method can provide conditional coverage, in a meaningful way, in a distribution-free setting; analogous considerations are given in McCullagh *et al.* (2009) for linear regression. More precisely, Lei and Wasserman (2014) show that, if $\mathcal{C}_{n,1-\alpha}(x)$ is a prediction interval for Y_{n+1} such that

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{n,1-\alpha}(x) | \boldsymbol{X}_{n+1} = x) \geq 1 - \alpha, \quad \forall P,\ \forall P_X \text{ for almost every } x \in \mathcal{X},$$

then, $\mathbb{P}\left(\lim_{\delta \to 0} \sup_{x \in B_\delta(x_0)} \mu\left(\mathcal{C}_{n,1-\alpha}(x)\right) = \infty\right) = 1$, $\forall P$, and any non-atom point x_0 of the distribution P_X. Nonetheless, notice that, although conditional coverage is impossible in a distribution-free setting, this does not mean that different methods cannot show significantly different behaviours in practice, in terms of approximate conditional coverage.

A simple yet efficient extension of conformal prediction to address this issue is the so-called Locally-Weighted Conformal Inference (Lei, G'Sell, *et al.* 2018). Consider, for the sake of simplicity a standard regression framework. A very natural correction to the residuals R_j can be done using their studentised version, namely

$$\tilde{R}_j = \frac{|Y_j - \hat{f}_{n_1}(\boldsymbol{X}_j)|}{\sigma_{n_1}(\boldsymbol{X}_j)}, \qquad j \in \text{Data}_{n_2}^{\text{Cal}}, \tag{11}$$

where $\sigma_{n_1}(\boldsymbol{X}_j)$ has the role of a "spread predictor" at $\boldsymbol{X}_j = x$. In Lei, G'Sell, *et al.* (2018), this is obtained by fitting the conditional mean absolute deviation $|Y_j - \hat{f}_{n_1}(\boldsymbol{x})|$ as a function

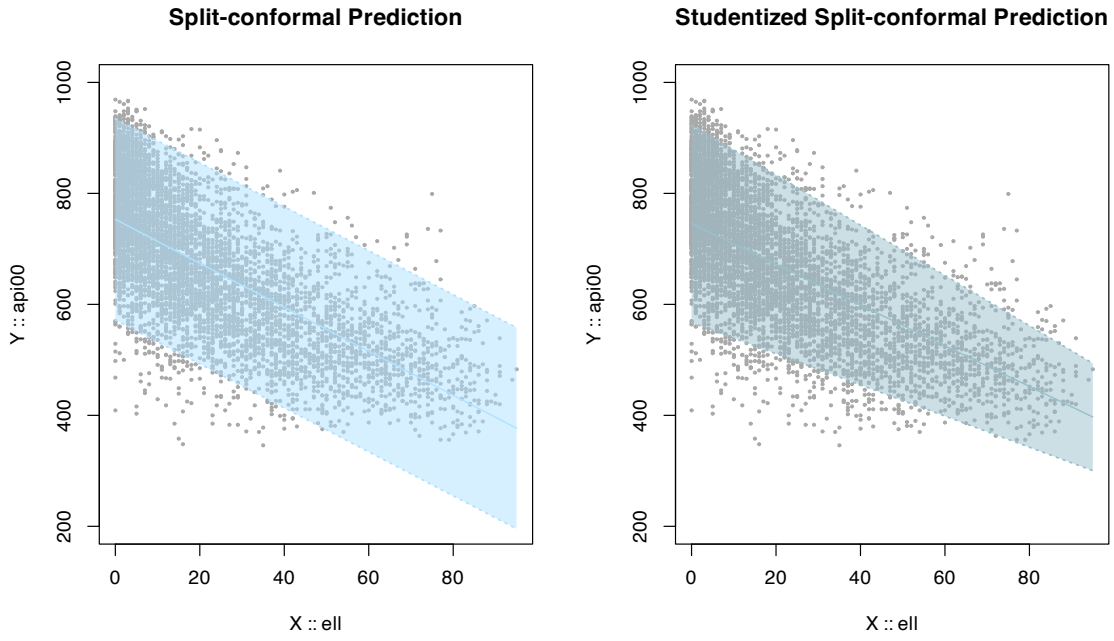
of $\boldsymbol{x} \in \mathcal{X}$, using $\text{Data}_{n_1}^{\text{Train}}$, as for the main predictor $\hat{f}_{n_1}$. With this correction, an estimated conformal prediction interval (in the split case) is obtained as:

$$\mathcal{C}_{n,1-\alpha}^{\text{ST-split}}(\boldsymbol{X}_{n+1}) = \left[\hat{f}_{n_1}(\boldsymbol{X}_{n+1}) - \sigma_{n_1}(\boldsymbol{X}_{n+1})q_{n,1-\alpha},\ \hat{f}_{n_1}(\boldsymbol{X}_{n+1}) + \sigma_{n_1}(\boldsymbol{X}_{n+1})q_{n,1-\alpha}\right], \quad (12)$$

where $q_{n,1-\alpha} = \tilde{R}_{(\lceil (n_2+1)(1-\alpha) \rceil)}$ amongst the $\{\tilde{R}_j\}_{j \in \text{Data}_{n_2}^{\texttt{Cal}}}$.

Interestingly, this simple modification greatly enhances the prediction interval estimation. In particular, it allows for adaptive prediction bounds, with larger intervals expected in correspondence to $\boldsymbol{x} \in \mathcal{X}$ that are harder to predict (e.g., small sample sizes or greater variability). This can be depicted in Figure 5.1 where we reconsider the simulation studies described in Section 4, focusing on the model-based split-conformal approach with M2, where the covariate has a numerical nature. Furthermore, when a group/domain variable is

Figure 5.1 - A comparison between the standard split-conformal prediction interval $\mathcal{C}_{n,1-\alpha}^{\text{split}}(\boldsymbol{x})$ and its studentised counterpart $\mathcal{C}_{n,1-\alpha}^{\text{ST-split}}(\boldsymbol{x})$ for different values of $\boldsymbol{x}$ (in relation to model M2). Here, $n = 200$ and $\alpha = 5\%$. Light blue areas represent the 95% prediction bands and gray points are the observed values.



Source: Own computation

considered (e.g., M1), it also improves the conditional efficiency (i.e., length) of the conformal procedure. Specifically, when comparing the traditional model-based and the standard model-based conformal prediction with the studentised version, we can notice narrower intervals, while still achieving good guarantees on the coverage (see Table 5.1).

An alternative and more general approach is discussed in Romano *et al.* (2019), where a quantile approach is considered and prediction intervals are no longer constructed around a point estimate.

Table 5.1 – Coverage and length with the different model-based methods, for a target level $\alpha = 5\%$ and a sample size $n = 200$ from the `apipop` dataset with population size $N = 6.194$. Results are reported for the model-based approaches with Model M1.

	$\text{Cov}^{\texttt{trad}}$	$\text{Cov}^{\texttt{split}}$	$\text{Cov}^{\texttt{ST-split}}$	$\mathcal{L}^{\texttt{trad}}$	$\mathcal{L}^{\texttt{split}}$	$\mathcal{L}^{\texttt{ST-split}}$
$X_1 = \texttt{Elementary}$	96.08%	94.92%	94.95%	488	476	477
$X_1 = \texttt{Middle}$	96.65%	95.78%	95.24%	490	476	465
$X_1 = \texttt{High}$	97.81%	97.66%	94.16%	491	476	399

Source: Own computation

5.2 Covariate shift

The main requirement for using CP is that the observations must be exchangeable so that the conformal measure computed on the test unit can be considered exchangeable with the others. This may not be the case in survey sampling where observed values in the sample $\mathcal{S}_n$ may be the result of a complex sampling design, while units for which we need to make a prediction might be generated by a different system. This problem has been considered in Tibshirani *et al.* (2019), by adopting a weighted version of the conformal scores. More in detail, assume that while the original sample data were generated by a model

$$(\boldsymbol{X}_j, Y_j) \overset{\text{i.i.d}}{\sim} P = P_I \times P_{Y|\boldsymbol{X}} \times P_{\boldsymbol{X}}, \quad j \in \mathcal{S}_n$$

the new observation comes from a different marginal distribution of $\boldsymbol{X}$, say

$$(\boldsymbol{X}_{j^*}, Y_{j^*}) \overset{\text{i.i.d}}{\sim} P^* = P_I \times P_{Y|\boldsymbol{X}} \times P^*_{\boldsymbol{X}}, \quad j^* \notin \mathcal{S}_n. \tag{13}$$

This problem can arise in survey sampling, for example, when the n sample data are generated by a stratified sample and the "next" one is selected at random among the unsampled ones. This implies a different meaning of the conformal scores, whose relevance now depends on the original distribution P or P^*. The problem is solved by weighting the original conformal scores of the observations $(x_1, x_2, \ldots, x_n)$ using the likelihood ratio $w(x_j) = \mathrm{d}P^*(x_j)/\mathrm{d}P(x_j)$, which has a "weight" role.

Consider, for simplicity, a full CP setup where the calibration scores are computed for the full sample dataset $\text{Data}_n^{\texttt{Sample}}$ and the augmented candidate y. Under a weighted version, the new set of empirical conformal scores will then be $(R_1 p_1(x), \ldots, R_n p_n(x), R_{j^*} p_{j^*})$, where

$$p_j(x) = \frac{w(\boldsymbol{X}_j)}{\sum_{i=1}^n w(\boldsymbol{X}_i) + w(x)}, \quad j \in \mathcal{S}_n; \quad p_{j^*}(x) = \frac{w(x)}{\sum_{i=1}^n w(\boldsymbol{X}_i) + w(x)}, \quad j^* \notin \mathcal{S}_n.$$

5.3 Calibrated Bayes

In recent years, we have witnessed a lively debate around the pros and cons of Bayesian ideas contamination in survey sampling, in a mood to maintain the characteristics of non-distortion and robustness of design-based and - at the same time - use the potential of the model-based approach. Recent contributions in this direction include Little (2011, 2022) and Di Zio *et al.* (2024).

Conformal methods operate as a *correction* factor to achieve the correct frequentist coverage and it can be applied to any parametric or nonparametric methods. On the other hand, it is well known that if any conformalised prediction algorithm leads to valid coverage, better algorithms (for the prediction problem at hand) lead to smaller prediction sets (Lei and Wasserman 2014).

One of the main criticisms regarding the adoption of model-based techniques in survey sampling is the potential dependence on the assumed model, which could be misleading in the case of misspecification. In addition, the frequentist performance of Bayesian methods can be jeopardised by the introduction of strong extra-experimental information through the prior distribution. The adoption of a conformal *modification* of the estimates produced via a full model-based Bayesian approach is then a promising way to proceed in order to obtain what Little (2011) calls *Calibrated Bayes*. In fact, one could exploit and combine all the information sources via a Bayesian hierarchical, model-based approach and take, as a natural conformity measure, the posterior predictive distribution, both in a Full or in a Split CP scenario. This road has been explored, in the field of small area estimation, by Michal *et al.* (2024) and Bersson and Hoff (2024) who propose a conformal method to obtain a prediction region for an unsampled unit in a small area. Computation details are described in Fong and Holmes (2021).

5.4 Combining prediction intervals

Conformal prediction produces uncertainty intervals at the unit level. However, the interest in official statistics may be generally on an aggregated quantity such as population totals or means as in Eq. (1). The process of aggregating intervals is not always straightforward.

Consider, for example, the case where we need to estimate the size of a population and, for each individual not sampled, we have a prediction interval. In this case, it may be natural to "sum" the individual intervals to obtain an overall prediction interval for the population size. However, there is no guarantee of exact coverage for this overall interval. This is a direction of further study, which we aim to pursue in the next future.

References

Bersson, E., and P. Hoff. 2024. "Optimal Conformal Prediction for Small Areas". *Journal of Survey Statistics and Methodology*, Volume 12, N. 5: 1464–1488. https://doi.org/10.1093/jssam/smae010.

Di Zio, M., B. Liseo, and M. Ranalli. 2024. "Bayesian ideas in survey sampling: The legacy of Basu". *Sankhya A*, Volume 86, N. Suppl 1: 71–94. https://doi.org/10.1007/s13171-023-00327-5.

Fong, E., and C. Holmes. 2021. "Conformal Bayesian Computation". In Ranzato, M., A. Beygelzimer, Y. Dauphin, P. Liang, and J.W. Vaughan (*eds.*). *Advances in Neural Information Processing Systems*, Volume 34: 18268–18279. Red Hook, NY, U.S.A: Curran Associates, Inc.

Graf, E., and Y. Tillé. 2014. "Variance estimation using linearization for poverty and social exclusion indicators". *Survey Methodology*, Volume 40, N. 1: 61–80. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14000-eng.pdf.

Lei, J., M. G'Sell, A. Rinaldo, R. Tibshirani, and L. Wasserman. 2018. "Distribution-free predictive inference for regression". *Journal of the American Statistical Association*, Volume 113, N. 523: 1094–1111. https://doi.org/10.1080/01621459.2017.1307116.

Lei, J., and L. Wasserman. 2014. "Distribution-free prediction bands for non-parametric regression". *Journal of the Royal Statistical Society: Series B*, Volume 76, N. 1: 71–96. https://doi.org/10.1111/rssb.12021.

Little, R. 2022. "Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference". *Survey Methodology*, Volume 48, N. 2: 257–281. https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00001-eng.htm.

Little, R. 2011. "Calibrated Bayes, for Statistics in General, and Missing Data in Particular". *Statistical Science*, Volume 26: 162–174. https://api.semanticscholar.org/CorpusID:2207054.

Lumley, T. 2024. "`survey`: analysis of complex survey samples. R package version 4.4". https://cran.r-project.org/web/packages/survey/index.html.

Mashreghi, Z., D. Haziza, and C. Léger. 2016. "A survey of bootstrap methods in finite population sampling". *Statistics Surveys*, Volume 10: 1–52. https://doi.org/10.1214/16-SS113.

McCullagh, P., V. Vovk, I. Nouretdinov, D. Devetyarov, and A. Gammerman. 2009. "Conditional prediction intervals for linear regression". In *2009 International Conference on Machine Learning and Applications*: 131–138.

Michal, V., J. Wakefield, A. Schmidt, A. Cavanaugh, B. Robinson, and J. Baumgartner. 2024. "Model-Based Prediction for Small Domains Using Covariates: A Comparison of Four Methods". *Journal of Survey Statistics and Methodology*, Volume 12, N. 5: 1489–1514. https://doi.org/10.1093/jssam/smae032.

Romano, Y., E. Patterson, and E. Candes. 2019. "Conformalized quantile regression". In Wallach, H., H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (*eds.*). *Advances in neural information processing systems*, Volume 32. Red Hook, NY, U.S.: Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf.

Shao, J. 2003. "Impact of the bootstrap on sample surveys". *Statistical Science*, Volume 18, N. 2: 191–198. https://doi.org/10.1214/ss/1063994974.

Tibshirani, R., R. Foygel Barber, E. Candes, and A. Ramdas. 2019. "Conformal Prediction Under Covariate Shift". In Wallach, H., H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (*eds.*). *Advances in Neural Information Processing Systems*, Volume 32. Red Hook, NY, U.S.: Curran Associates, Inc.

Vallée, A., and Y. Tillé. 2019. "Linearisation for Variance Estimation by Means of Sampling Indicators: Application to Non-response". *International Statistical Review*, Volume 87, N. 2: 347–367. https://doi.org/10.1111/insr.12313.

Vovk, V., A. Gammerman, and G. Shafer. 2005. *Algorithmic Learning in a Random World*. New York City, NY, U.S.: Springer.

Wieczorek, J. 2023. "Design-based conformal prediction". *Survey Methodology*, Volume 49, N. 2: 443–473. http://www.statcan.gc.ca/pub/12-001-x/2023002/article/00007-eng.htm.