Robust Gas Demand Forecasting With Conformal Prediction

Mouhcine Mendil MOUHCINE.MENDIL@IRT-SAINTEXUPERY.COM
IRT Saint Exupéry, Toulouse, France
IRT SystemX, Paris, France

Luca Mossina LUCA.MOSSINA@IRT-SAINTEXUPERY.COM
IRT Saint Exupéry, Toulouse, France
IRT SystemX, Paris, France

Marc Nabhan MARC.NABHAN@AIRLIQUIDE.COM
Air Liquide, Les-Loges-en-Josas, France
IRT SystemX, Paris, France

Kevin Pasini KEVIN.PASINI@IRT-SYSTEMX.FR
IRT SystemX, Paris, France

Editor: Henrik Boström, Lars Carlsson, Ulf Johansson, Zhiyuan Luo and Khuong An Nguyen

Abstract

Predicting the future trends of customer gas demand as precisely as possible is vital for securing the supply chain from production to distribution. The operations at Air Liquide require the predictions of a Machine Learning forecaster to be coupled with rigorous Uncertainty Quantification (UQ), building trustworthy and informative prediction intervals. To address these industrial needs, we propose to apply Conformal Prediction (CP), a framework that can provide probabilistic guarantees for any underlying predictive model.

The problem is formulated as time series forecasting, which may counter the CP hypothesis of data exchangeability. Nevertheless, our experiments show that CP methods enhance the predictive coverage of the tested UQ approaches. We also test EnbPI, a conformal method designed specifically for time series, and propose a locally adaptive variant. To carry out our experiments with prediction intervals using multiple regression models, we introduce our new python library PUNCC and a novel dataset (around 10k observations) provided by Air Liquide which leverages over 7 years of data of weekly gas consumption.

Keywords: Conformal Prediction, Gas Distribution, Prediction Intervals, Time Series, Uncertainty Quantification.

1. Introduction

In *Machine Learning* (ML), we build predictive models from experience, by choosing the right approach for the right problem, and from the accessible data, via algorithms. Despite our best efforts, we can encounter some underlying uncertainty that could stem from various sources or causes.

Typically, uncertainty in the ML process can be categorized into two types (Hüllermeier and Waegeman, 2021): aleatoric uncertainty, also known as statistical uncertainty, which is *irreducible* as due to the intrinsic randomness of the phenomenon being modeled; and epistemic uncertainty, also known as systematic uncertainty, which is *reducible* through additional information, e.g. via more data or better models.

In all its forms, uncertainty can have major impacts in industry. It would be very helpful to the on-site operators to quantify this uncertainty with a known guarantee, via *Conformal Prediction* (CP) methods for instance. To give a more practical viewpoint, we tackle a use case of industrial gas demand forecasting. An anonymized dataset was built by Air Liquide, a world leader in gases, technologies and services for industry, to be representative of the industrial process.

1.1. The Industrial Use Case

With several sources of production and multiple customers of industrial gases across France, it is the company's responsibility to guarantee the availability of supply, i.e. it should know when to deliver the right product to all its customers on time.

By taking into account geographical and logistical factors, managing the customers' products storage has an effect on the whole supply chain, from production to distribution. Therefore, the dispatchers, who are in charge of delivering the products, estimate the future trends of customers demands in advance based on their years of experience in the field. They have to take into account daily logistical constraints, e.g., limited transport resources, limited product availability, and accessibility issues.

In order to help the dispatchers in their estimations, multiple forecasting algorithms have been proposed to predict the needs in production as precisely as possible in the short term, and compare them with the dispatchers' estimations. Nonetheless, despite achieving remarkable performance, the comparison between the predictions and the estimations yields an uncertainty that is currently not measured. Such uncertainty can have a critical impact on production: if we produce more than intended, we face energy and product losses; conversely, if we produce less, we risk drying out the customers and failing to deliver the right amounts on time.

1.2. Challenges and Objectives

In light of this industrial context, even the best forecast model can suffer from uncertainty between the forecast predictions and the dispatchers estimations. The focus is then shifted toward **quantifying** the uncertainty associated with the forecast predictions. Since this forecasting use case is a regression problem, one effective way would be by building *Prediction Intervals* (PIs) via *Uncertainty Quantification* (UQ), which consists of upper and lower bounds that contain the forecast predictions with high probability. Hence, the dispatchers would have at their disposal minimum and maximum values that numerically quantify operational uncertainties.

Nonetheless, operational constraints have been expressed regarding the desired results. Obviously, these intervals should have a high coverage probability to be used in production with great trust, while also being as "narrow" as possible to be informative for the dispatchers. To approximate more effectively their predictions, dispatchers need a sort of statistical safety net in their decision-making process, that will increase the robustness of these operations, boost the trust in the forecast algorithm, and optimize the production and distribution of the products.

This is why we focused our work on ***Conformal Prediction* (CP) methods**, which constitute a promising solution to the operational constraints described before. These meth-

ods produce prediction intervals backed by theoretical guarantees while being able to be applied atop any regression algorithm. In its original form, CP (Papadopoulos et al., 2002; Vovk et al., 2005; Papadopoulos et al., 2011; Lei et al., 2018) assumes exchangeable or independent and identically distributed data. This assumption is usually not met in time series, where the data can be subject to distribution shifts and variable dependency, which can invalidate the guarantee of marginal coverage of CP.

Recent works (Tibshirani et al., 2019; Xu and Xie, 2021b; Barber et al., 2022) shed the light on approaches of CP that go beyond exchangeability. In particular, *Ensemble batch Prediction Interval* (EnbPI) (Xu and Xie, 2021b) is an ensemble-based wrapper around regression models that offers asymptotically valid coverage and efficient *Prediction Interval*s (PIs) for a class of time series. We propose in this paper to test EnbPI and other state-of-the-art CP methods on the problem of forecasting gas demand. We also propose a modification of EnbPI into a locally adaptive variant to enhance its conditional coverage.

The rest of the paper is structured as follows: PIs are presented in Section 2, along with the metrics used to evaluate their performances. In Section 3, CP methods are showcased for the benefit of the reader. Section 4 presents the industrial dataset, as well as our own developed library, before delving into the experiments that were conducted in this work. Finally, we compare and discuss the results obtained, before concluding in Section 5.

2. Uncertainty quantification with prediction intervals

Let $(X, Y) \sim \mathbb{P}_{XY}$, with X being the (random) vector of features and $Y \in \mathbb{R}$ the target in a learning task. Let $\alpha \in (0, 1)$ be a miscoverage level, interpretable as the proportion of predictive mistakes we are willing to accept in the long run. A PI (Hahn and Meeker, 1991) obtained via algorithm $C_\alpha(X)$ should contain the true value of the random variable Y, $(1-\alpha)100\%$ of the time. Ideally, $\mathbb{P}\{Y \in C_\alpha(X) \,|\, X = x\} = (1-\alpha)$ holds true conditionally on the value taken by X. An inferential procedure that satisfies this statement is said to produce conditionally **valid** intervals[1]. When the weaker condition $\mathbb{P}\{Y \in C_\alpha(X)\} = (1-\alpha)$ holds, the procedure is **marginally** valid (see Section 3).

2.1. Construction of prediction intervals

Let $Y \in \mathbb{R}$ be a prediction target, such as the demand of product at any given time in our industrial use case. There are two main approaches to build the interval $\widehat{C}_\alpha(X) = [\,\widehat{Y}^L, \widehat{Y}^U\,]$, with lower bound $\widehat{Y}^L$ and upper bound $\widehat{Y}^U$. First, one can add an error margin δ_α to a point prediction: $\widehat{C}_\alpha(X) = [\widehat{f}(X) - \delta_\alpha(X)\,,\, \widehat{f}(X) + \delta_\alpha(X)]$. Here, $\widehat{f}(\cdot)$ usually estimates the conditional expected value $\mathbb{E}\{Y \,|\, X = x\}$, and $\delta_\alpha(X) \geq 0$ depends on the data and a *probabilistic statement* made within the PI procedure; in linear regression, for instance, we could assume the errors to be normally distributed (Wasserman, 2004). Second, one can estimate the bounds directly, e.g. by quantile regression: $\widehat{C}_\alpha(X) = [\widehat{q}_{\alpha_{lo}}(X)\,,\, \widehat{q}_{1-\alpha_{hi}}(X)]$, where $\widehat{q}_\beta(\cdot)$ is an estimation (Koenker and Bassett Jr, 1978; Meinshausen, 2006) of the conditional quantile $q_\beta(x) = \inf\{y : F(y \,|\, X = x) \geq \beta\}$, with $\alpha_{lo} + \alpha_{hi} = \alpha$. If valid, these intervals are *representative of the predictive error* due to the predictor (epistemic

1. This guarantee is on the *procedure* followed to sample the data and construct the PIs. The interval has a probabilistic interpretation only *before* observing the data. For a detailed explanation, see (Shafer and Vovk, 2008, Section 2.2) and (Hahn and Meeker, 1991, Section 2.3.6).

uncertainty) and the randomness in sampling (aleatoric uncertainty). Independently of the choice of $f(\cdot)$ or $q(\cdot)$, in Section 3 we show that conformal inference builds valid intervals for *any method* and can "robustify" existing PIs with rigorous probability guarantees.

2.2. Metrics for prediction intervals

In the literature (Pang et al., 2018; Bazionis and Georgilakis, 2021), we find **coverage metrics** to assess how often the predicted intervals $\widehat{C}_\alpha(X)$ *contain* the true value of Y, and **sharpness metrics** to measure the *width* of the intervals. Coverage and sharpness are connected, as a degradation in one usually leads to an improvement of the other.

Let n be the number of samples in a test dataset $D_{test} = \{(X_i, Y_i)\}_{i=1}^n$, $(X_i, Y_i) \sim \mathbb{P}_{XY}$. For a miscoverage level $\alpha \in (0,1)$, we get the corresponding nominal (target) coverage as $1-\alpha$. We measure how many times the PI contains the true value via the empirical coverage, or *Prediction Interval Coverage Probability*:

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\widehat{Y}_i^L \leq Y_i \leq \widehat{Y}_i^U\} \in [0,1]. \tag{1}$$

When PICP $\approx 1-\alpha$, the PIs are capturing, or "covering", the values of Y at the specified α. For a PICP $> 1-\alpha$ we have **overcoverage** and we are capturing too many points, for PICP $< 1-\alpha$, we have **undercoverage**.

In CP, the reference metric for sharpness is the *Average Width*, AW $= \frac{1}{n}\sum_{i=1}^n (\widehat{Y}_i^U - \widehat{Y}_i^L)$. The *Prediction Interval Normalized Average Width* allows for comparisons over different datasets, via the normalization $R = Y_{max} - Y_{min}$:

$$\text{PINAW} = \frac{1}{n\,R}\sum_{i=1}^{n} (\widehat{Y}_i^U - \widehat{Y}_i^L). \tag{2}$$

3. Conformal Prediction

Many popular ML algorithms do not offer UQ out of the box (Hastie et al., 2009; Goodfellow et al., 2016), and some UQ methods are valid only under hypotheses on $\mathbb{P}_{XY}$ or asymptotic properties of the algorithm (e.g. Section 2.1): they require enough data to get reliable PIs and can suffer from overfitting.

Conformal Prediction (CP) (Vovk et al., 2005) is a set of *distribution-free*, *model-agnostic* and *non-asymptotic* methods to do UQ by constructing valid *prediction sets* or *intervals*, whose probability coverage is backed by theoretical guarantees. Given a miscoverage probability $\alpha \in (0,1)$, a set of exchangeable[2] training data $\{(X_i, Y_i)\}_{i=1}^n$ and test point (X_{new}, Y_{new}) with common distribution $\mathbb{P}_{XY}$, a CP procedure $C_\alpha(\cdot)$ builds prediction sets so that:

$$\mathbb{P}\Big\{Y_{new} \in C_\alpha\left(X_{new}\right)\Big\} \geq 1-\alpha. \tag{3}$$

Over many calibration and test sets, $C_\alpha(X)$ will contain the observed values of Y with frequency of *at least* $100(1-\alpha)\%$. Within the CP framework, Equation 3 holds for any model, any data distribution $\mathbb{P}_{XY}$ and any finite sample sizes. Unlike the probability statement in

2. This includes independent and identically distributed data (iid) as a special case (Aldous, 1985).

Section 2, this holds for Y when averaging over all possible X, that is, **marginally** and not conditionally (Foygel Barber et al., 2020). We could overcover or undercover Y for some values of X.

CP can act as a *complementary tool* to attain rigorous probability coverages, as it can "conformalize" any existing predictor during or after training (black box predictors), yielding marginally valid PIs even under model misspecification (Chernozhukov et al., 2021). Furthermore, any theoretical property of the underlying predictor still holds[3].

3.1. Related Work

In this section, we present the most common CP methods of the literature. The reader is also referred to Angelopoulos and Bates (2021) for a hands-on introduction to CP and Shafer and Vovk (2008); Zeni et al. (2020) for an in-depth overview. Let $D_{train} = \{(X_i, Y_i)\}_{1=1}^{n_{train}} \sim \mathbb{P}_{XY}$ be the training data and $\alpha \in (0, 1)$ the miscoverage probability.

3.1.1. The split conformal prediction procedure

Following the *Split* CP method (Papadopoulos et al., 2002; Lei et al., 2018), also known as *Inductive* CP, here are the fundamental steps of conformalization:

- **(Step 1)** Choose (or receive) a base prediction model: $f(\cdot)$
- **(Step 2)** Choose a nonconformity score: $R = s(\widehat{f}(X), Y)$
- **(Step 3)** Choose a data scheme: split training data D_{train} into two partitions, a *fit* subset D_{fit} and a *calibration* subset $D_{calibration}$
- **(Step 4)** Fit & calibrate: compute $\widehat{f}(\cdot)$ on D_{fit} and scores $\bar{R} = \{R_i\}$, $i = 1, \ldots, |D_{calibration}|$ on $D_{calibration}$
- **(Step 5)** Get error margin $\delta_\alpha^f = (1-\alpha)(1 + \frac{1}{|D_{calibration}|})$-th empirical quantile of $\bar{R}$
- **(Step 6)** Inference: build CP interval $\widehat{C}(X_{new})$, for observation X_{new}

The **base prediction model** $f(\cdot)$ (Step 1) is either a pre-fitted black-box $\widehat{f}$, or is selected through empirical experimentation or for operational needs. The selection of **nonconformity score** (Step 2) is at the heart of CP: here, the default score is the absolute deviation $R_i = |Y_i - \widehat{f}(X_i)|$ (Section 3.1.2 for other possible scores). The split **data scheme** (Step 3) used to fit and conformalize a predictor works well with large datasets. We partition D_{train} in two disjoint sets: D_{fit}, to fit the predictor, and $D_{calibration}$ for calibration. For smaller datasets, better options exist (Section 3.1.2). For **calibration** (Step 4), the data scheme determines the out-of-sample evaluation of $\widehat{f}$: here, we compute the scores $\bar{R} = \{R_i\}_{i=1}^{|D_{calibration}|}$ on $D_{calibration}$. In (Step 5), we compute the error margin δ_α^f as a quantile of the sorted scores; δ_α^f is a constant and does not depend on the test point X_{new}. At **inference** (Step 6), we get the PI as:

$$\widehat{C}_\alpha(X_{new}) = \left[\widehat{f}(X_{new}) - \delta_\alpha^f \,,\, \widehat{f}(X_{new}) + \delta_\alpha^f\right]. \tag{4}$$

When conformalizing a pre-trained predictor $\widehat{f}$, one only needs additional data $D_{calibration} \sim \mathbb{P}_{XY}$, skipping (Step 3) and the fit of (Step 4).

3. e.g., the asymptotic conditional coverage of Quantile Regression Forests (Meinshausen, 2006).

3.1.2. Conformal regression methods

In the CP literature, different methods stem from the choice of nonconformity score and data scheme. Split CP turns into *Locally Adaptative Conformal Prediction* (LACP) when using scaled scores $R_i = \frac{|Y_i - \widehat{f}(X_i)|}{\widehat{\sigma}(X_i)}$, where $\widehat{\sigma}(X)$ is a predictor of dispersion learned on D_{train}. This was proposed by Papadopoulos et al. (2002) and further studied by Papadopoulos et al. (2011); Papadopoulos and Haralambous (2011); Johansson et al. (2014); Boström et al. (2017), for different types of predictors. Here, we follow Lei et al. (2018) and $\widehat{\sigma}(X_i)$ predicts the conditional *Mean Absolute Deviation* (MAD) of $\left(Y - \widehat{f}(X)\right)$, conditioned on $X = x$. LACP is "local" in the sense that, for a point $(X_i, Y_i) \sim P_{XY}$, the prediction interval size will be corrected to represent the variability at $Y_i|X_i = x_i$. The PI is given by:

$$\widehat{C}_\alpha(X_{new}) = \left[\widehat{f}(X_{new}) - \widehat{\sigma}(X_{new})\,\delta_\alpha^{f,\sigma}\,,\ \widehat{f}(X_{new}) + \widehat{\sigma}(X_{new})\,\delta_\alpha^{f,\sigma}\right]. \tag{5}$$

where the normalized error margin $\delta_\alpha^{f,\sigma}$ is scaled up by $\widehat{\sigma}(X_{new})$.

Split CP can be extended to quantile (and interval) base predictors $q(\cdot)$ via the nonconformity score $R_i = \max\{\widehat{q}_{\alpha_{lo}}(X_i) - Y_i, Y_i - \widehat{q}_{1-\alpha_{hi}}(X_i)\}$, $i = 1, \ldots, |D_{calibration}|$. This gives the *Conformalized Quantile Regression* (CQR) method (Romano et al., 2019), with PI:

$$\widehat{C}_\alpha(X_{new}) = \left[\widehat{q}_{\alpha_{lo}}(X_{new}) - \delta_\alpha^q\,,\ \widehat{q}_{1-\alpha_{hi}}(X_{new}) + \delta_\alpha^q\right], \tag{6}$$

where δ_α^q is a **correction** margin that guarantees finite-sample coverage for the quantile predictions. If $\delta_\alpha^q \approx 0$, we get a confirmation of its predictive marginal validity backed by theory. Although designed for exchangeable data, in Section 4.4 we show its reasonable effectiveness on time series.

For small datasets, the *jackknife+* (Barber et al., 2021) uses either a *Leave-One-Out* (LOO) or *K–fold* data schemes for better statistical efficiency, at the cost of fitting n_{train} and K models. When the base predictor is an ensemble learner using *bagging* (Breiman, 1996a), one can compute the nonconformity scores via the *Out-of-Bag* (OOB) "trick" (Breiman, 1996b) with negligible computational overhead[4]. Notably, the *Jackknife+–after–Bootstrap* (Kim et al., 2020) and the *Quantile Out–of–Bag* (Gupta et al., 2019) CP methods are, respectively, compatible with any point $f(\cdot)$ and interval $q(\cdot)$ predictor. For the above methods, the inference (Step 6) is modified to account for the multiple fitted predictors.

3.1.3. Conformal prediction for time series

As previously mentioned in the introduction, the community is moving towards extensions of CP for non-exchangeable data (Barber et al., 2022; Tibshirani et al., 2019; Xu and Xie, 2021b). For time series in particular, the literature introduces additional assumptions on the problem to get close to the properties of CP. The method in (Chernozhukov et al., 2018) aims at recovering data exchangeability via permutations of sequences of data samples. Furthermore, Gibbs and Candès (2021) and Zaffran et al. (2022) worked on obtaining

4. When *bagging*, one fits B predictors on B bootstrap datasets S_b, sampled *with replacement* from D_{train}. Following a probabilistic fact (see citations), on average, about $\sim 30\%$ of the samples are left out of S_b, or *out-of-bag*, and we can compute the scores with out-of-sample data, approximating a *Leave-One-Out* (LOO) scheme.

adaptive miscoverage probabilities α for online CP, while in (Stankevičiūtė et al., 2021) CP is applied to multiple time series assumed to be exchangeable. Finally, Diquigiovanni et al. (2021) worked with multivariate functional time series applied to demand prediction in the Italian gas market, and in (Wisniewski et al., 2020) we find a modified, empirically tested, version of split CP applied to financial data.

Among the most promising results is the *Ensemble batch Prediction Interval* (EnbPI) algorithm (Xu and Xie, 2021b), a modification of the Jackknife+-after-Bootstrap that also uses OOB estimation of nonconformity scores (see Footnote 4). By doing so, EnbPI avoids overfitting without recourse to data-splitting, which improves the computational efficiency. By construction, EnbPI yields constant-size intervals. We implemented a locally adaptive extension referred to as *Adaptive Ensemble batch Prediction Interval* (aEnbPI), by integrating an ensemble estimation of the conditional MAD $\widehat{\sigma}$ to the original algorithm. This is a straightforward application of LACP to EnbPI, yielding scaled nonconformity scores.

3.2. Adaptive Ensemble Batch Prediction Interval

The *Ensemble batch Prediction Interval* (EnbPI) algorithm (Xu and Xie, 2021b) is a modification of the Jackknife+-after-Bootstrap (Kim et al., 2020), using the *Out-of-Bag* (OOB) trick of Breiman (1996b) to estimate Leave-One-out (LOO) nonconformity scores and aggregated predictors. EnbPI allows to build marginally valid PIs for time series with few assumptions in addition to the CP framework. From now on, we use the original notation of the paper, which will be immediately clear to the reader.

The data is supposed to be generated by the model $Y_t = f(X_t) + \epsilon_t$, where the function f is unknown and the errors ϵ_t are *identically distributed* but not necessarily independent. The d-dimensional feature vector X_t can include the history $\{Y_{t-1}, Y_{t-2}, \dots\}$ of Y_t, such that both traditional time series models (e.g., ARIMA) and complex ML techniques are eligible.

We propose the *Adaptive Ensemble batch Prediction Interval* (aEnbPI) method as an extension of the LACP to EnbPI. In Algorithm 1, the highlighted lines represent the additions or modifications brought in aEnbPI with respect to EnbPI.

From Line 1 to Line 11, the procedure is derived from the Jackknife+-after-Bootstrap of Kim et al. (2020): 1) the b-th bootstrap estimators ($b = 1, \dots, B$) of the conditional mean $\widehat{f}^b$ and the conditional MAD $\widehat{\sigma}^b$ are trained on bootstrap samples of the training data 2) the i-th LOO estimates ($i = 1, \dots, T$) of the conditional mean $\widehat{f}^{\phi}_{-i}$ and the conditional MAD $\widehat{\sigma}^{\phi}_{-i}$ are approximated by aggregation of a subset of bootstrap models for which the sample (x_i, y_i) is OOB and 3) the OOB nonconformity scores $\boldsymbol{\epsilon}^{\phi}_i$ are computed. Note that we replace the absolute residuals $|y_i - \hat{f}^{\phi}_{-i}(x_i)|$ by the scaled residuals $\frac{|y_i - \hat{f}^{\phi}_{-i}(x_i)|}{\hat{\sigma}^{\phi}_{-i}(x_i)}$.

Line 13 marks the start of the inference loop to determine the PIs on a time horizon T_1 with respect to the batch size s. At the end of every batch of sequential predictions (see Line 19), we update the nonconformity scores $\widehat{\boldsymbol{\epsilon}}$ by replacing the s oldest values with the s most recent. The batch update allows for dynamic calibration of PIs: under potential data drift, discarding old errors and considering the most recent ones can improve the accuracy of PI widths even without refitting the models on the newly available data. Finally, on the basis of the $(1 - \alpha)$-th quantile of the nonconformity scores, the OOB estimation of

Algorithm 1: *Adaptive Ensemble batch Prediction Interval* (aEnbPI). Additions or modifications of EnbPI (Xu and Xie, 2021b) are highlighted.

Input: Training data $\{(x_i, y_i)\}_{i=1}^{T}$, point prediction algorithm A^f, variability prediction algorithm A^σ, miscoverage level α, aggregation function ϕ, number of bootstrap models B, batch size s, and test data $\{(x_t, y_t)\}_{t=T+1}^{T+T_1}$; y_t is revealed only after the batch of s PIs with t in the batch are constructed.

for $b = 1, \ldots, B$ **do**
  Sample with replacement an index set $S_b = (i_1, \ldots, i_T)$ from indices $(1, \ldots, T)$;
  Compute $\widehat{f}^b \leftarrow A^f(\{(x_i, y_i) | i \in S_b\})$;
  Compute $\widehat{\sigma}^b \leftarrow A^\sigma(\{(x_i, y_i) | i \in S_b\})$;
end
$\widehat{\boldsymbol{\epsilon}} \leftarrow \{\}$;
for $i = 1, \ldots, T$ **do**
  $\hat{f}^\phi_{-i}(x_i) \leftarrow \phi(\{\hat{f}^b(x_i) | i \notin S_b\})$;
  $\hat{\sigma}^\phi_{-i}(x_i) \leftarrow \phi(\{\hat{\sigma}^b(x_i) | i \notin S_b\})$;
  $\hat{\boldsymbol{\epsilon}}^\phi_i \leftarrow \frac{|y_i - \hat{f}^\phi_{-i}(x_i)|}{\hat{\sigma}^\phi_{-i}(x_i)}$;
  $\widehat{\boldsymbol{\epsilon}} \leftarrow \widehat{\boldsymbol{\epsilon}} \cup \{\hat{\boldsymbol{\epsilon}}^\phi_i\}$;
end
$\widehat{\boldsymbol{C}} \leftarrow \{\}$;
for $t = T+1, \ldots, T+T_1$ **do**
  $\hat{f}^\phi_{-t}(x_t) \leftarrow (1-\alpha)$ quantile of $\{\hat{f}^\phi_{-i}(x_t)\}_{i=1}^{T}$;
  $\hat{\sigma}^\phi_{-t}(x_t) \leftarrow \phi\{\hat{\sigma}^\phi_{-i}(x_t)\}_{i=1}^{T}$;
  $w^\phi_{T,t} \leftarrow (1-\alpha)$ quantile of $\boldsymbol{\epsilon}$;
  $C^{\phi,\alpha}_{T,t}(x_t) \leftarrow [\hat{f}^\phi_{-t}(x_t) \pm w^\phi_t \, \hat{\sigma}^\phi_{-t}(x_t)]$;
  $\widehat{\boldsymbol{C}} \leftarrow \widehat{\boldsymbol{C}} \cup C^{\phi,\alpha}_{T,t}(x_t)$;
  if $t - T = 0 \ mod \ s$ **then**
    for $j = t-s, \ldots, t-1$ **do**
      $\hat{\boldsymbol{\epsilon}}^\phi_j \leftarrow \frac{|y_j - \hat{f}^\phi_{-j}(x_t)|}{\hat{\sigma}^\phi_{-j}(x_t)}$;
      $\widehat{\boldsymbol{\epsilon}} \leftarrow (\widehat{\boldsymbol{\epsilon}} - \{\hat{\epsilon}^\phi_1\}) \cup \{\hat{\epsilon}^\phi_j\}$ and reset index of $\widehat{\boldsymbol{\epsilon}}$;
    end
  end
end
return *Ensemble prediction intervals* $\widehat{\boldsymbol{C}} = \{C^{\phi,\alpha}_t(x_t)\}_{t=T+1}^{T+T_1}$

the local dispersion $\widehat{\sigma}^\phi_{-t}$ enables to construct rescaled prediction intervals $C^{\phi,\alpha}_{T,t} = [\hat{f}^\phi_{-t}(x_t) \pm w^\phi_t \, \hat{\sigma}^\phi_{-t}(x_t)]$ that account for data heteroskedasticity.

A final word on Line 15, where the point estimate is given by the empirical quantile of the T OOB predictions. In the latest, revised version of the paper currently under review Xu and Xie (2021a), this inferential step is modified by taking an aggregation of the OOB

predictions: $\hat{f}^{\phi}_{-t}(x_t) = \{\hat{f}^{\phi}_{-i}(x_t)\}_{i=1}^{T}$. More insights on this point are also given in Zaffran et al. (2022).

The validity of aEnbPI derives from EnbPI. Roughly speaking (interested reader can find the details in Xu and Xie (2021b, Section 4.1))[5], the prediction errors must be reasonably "well-behaved", in the sense that there is a number of time steps after which current errors are not affected by past errors (independence). Second, the predicted residuals must be close to the true residuals. This hypothesis could be invalidated, for instance, if we were overfitting the training data. As the authors point out, "[...] the series can exhibit arbitrary dependence and be highly non-stationary, but still have strongly-mixing (or even i.i.d.) errors". For more details on the implications of this assumptions, see Xu and Xie (2021b, Section 4.2).

4. Experiments

4.1. The customers demands forecasting dataset

The customers demands forecasting dataset was to be shared among the contributors of the *confiance.ai* consortium while containing sensitive customers data. A necessary step of data anonymization and numerical transformations was then inevitable to ensure customers data confidentiality. Having that in mind, a brief description of the dataset will follow to grasp this use case without compromising possible critical information.

The dataset includes data related to historical consumption of industrial gas across seven years, customers and orders specifics, geographical distribution, and seasonality parameters. In short, there are several sources of production, a handful of products concerned, and multiple customers to deliver to. Geographical data is retrieved for the sources of production to locate the origin of the quantity that has been produced. The product type is also essential to be involved in the dataset, since quantities can obviously differ among various products needed. Customers-specific data will give more information into the clients involved and how they are dispatched to the production sources according to the quantities and product type needed. For each combination of **source**, **product**, and **customer**, we obtain a unique time series of gas delivery.

After aggregating product quantities weekly, the objective is to predict accurate predictions for the following week. Having 29 combinations of source, product, and customer overall, the dataset is a concatenation of all these time series over seven years, as illustrated in Figure 1. With approximately 350 weekly quantities per time series, the dataset is composed of around 10 000 observations in total. The forecast algorithm's sliding window leverages the last four weeks of data to predict the quantities for the next one. Given that some time series are dependent, it is operationally efficient to learn a single joint model for predicting future demands of all combinations. Additional features related to the registered orders and holidays information were also added to provide a maximum amount of knowledge. Further details are provided in Appendix A.

5. See Assumption 1 & 2 in Section 4.1 of the paper. See also the additional material: http://proceedings.mlr.press/v139/xu21h.html

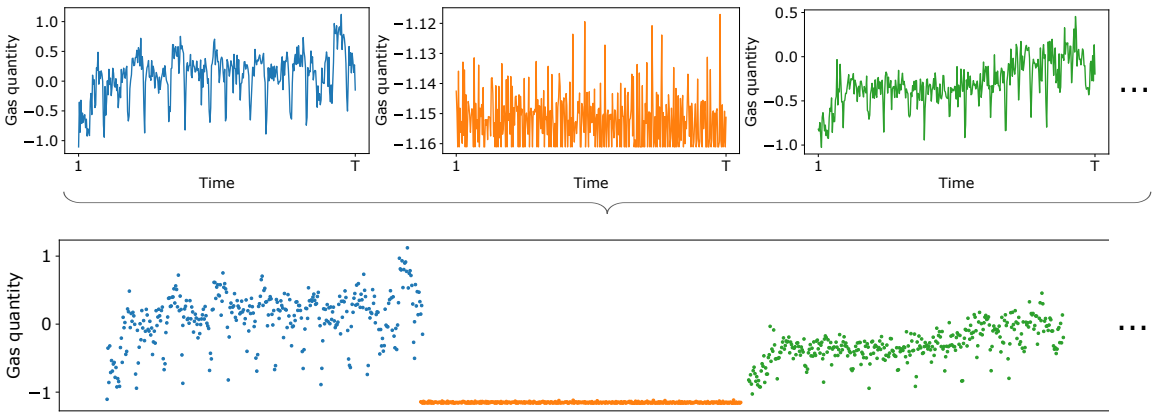

Figure 1: Partial visual representation of the dataset: individual time series each corresponding to a combination of source, product and customer (top); concatenation of all time series to form the dataset (bottom). (Gas Quantities are numerically transformed for confidentiality, see Section 4.1)

4.2. PUNCC

Predictive UNcertainty Calibration and Conformalization (PUNCC) is an open-source library[6] that enables ad-hoc integration of *Artificial Intelligence* (AI) models into a theoretically sound UQ framework based on CP. From a wrapper of point or interval predictors, PIs are constructed with guaranteed coverage probability according to a selected level of error α. PUNCC exposes two API levels following the user's need:

- High-level API: a low-barrier-to-entry tool for fast prototyping and deployment. The calibration and conformalization are performed by a ready-to-use wrapper around regression models. PUNCC provides various state-of-the-art CP methods.

- Low-level API: offers extra control on each stage of the calibration and conformalization. It provides generic structures to design custom nonconformity scores, data partition schemes, and in-depth user-defined algorithms for building PIs.

The design of PUNCC is characterized by its simplicity and the modularity of its components, enabling a flexible and broad range of uses. The core interface of the library, namely `ConformalPredictor`, is in charge of the calibration and conformalization of ML models. The high-level API provides a collection of ready-to-use `ConformalPredictor` subclasses that includes split CP (Papadopoulos et al., 2002), locally adaptive CP (Lei et al., 2018), *Conformalized Quantile Regression* (CQR) (Romano et al., 2019) and EnbPI (Xu and Xie, 2021b). More generally, the `ConformalPredictor` object is constructed following the Predictor-Splitter-Calibrator paradigm —exposed by the low-level API—. It boils down to putting together three modular components:

6. https://github.com/deel-ai/puncc

1. Predictor: interface standardizer for AI models, making PUNCC interoperable with various ML libraries (such as tensorflow, pytorch and scikit-learn).
2. Splitter: a strategy to assign data into fit and calibration sets, such as K-fold or a split based on seasons for time-series.
3. Calibrator: an estimator of nonconformity scores used for building and calibrating PIs.

Dedicated visualization and metrics modules are also provided to help build and interpret the CP results. At the time of writing, PUNCC covers CP for regression. In the future, we aim to include more other ML tasks such as classification and risk control.

4.3. Methodology of benchmark experiments

We experiment with different ways of conformalizing several candidate models to forecast time series. First, four models are used for estimating the mean, dispersion, and quantiles of the customers demand (Model hyperparameters are available in materials):

1. ***eXtreme Gradient Boosting* (XGB)**: for mean regression (Chen and Guestrin, 2016).
2. ***Gradient Boosting Quantile Regression* (GBQR)**: for quantile regression with the pinball loss (Hastie et al., 2009).
3. ***Quantile Random Forests* (QRF)**: for quantile regression using RF resampling (Meinshausen, 2006).
4. ***Random Forest Mean Variance* (RFMV)**: for mean and variance estimation using mean squared error loss (Breiman, 2001).

During the learning phase, the parameters θ of each model are obtained through minimization of the loss function $\mathcal{L}$:

$$\theta := \min_{\theta} \sum_{t \in fit} \mathcal{L}(y_t, f_\theta(x_t))$$

For XGB, RFMV and QRF, $\mathcal{L}$ corresponds to the squared error: $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$.
For GBQR, $\mathcal{L}$ is the pinball loss: $\mathcal{L}_\alpha(y, \hat{y}) = \begin{cases} (y - \hat{y}) \cdot \alpha & \text{if } y \leq \hat{y} \\ (\hat{y} - y) \cdot (1 - \alpha) & \text{if } y \geq \hat{y} \end{cases}$

Then, we rely on five CP methods to build and calibrate the PIs (see Section 3):

1. ***Split Conformal Prediction* (SCP)**: for constant PIs from a conditional mean estimator.
2. ***Locally Adaptative Conformal Prediction* (LACP)**: for adaptive PIs from both conditional mean and dispersion estimators.
3. ***Conformalized Quantile Regression* (CQR)**: for adaptive PIs from upper and lower quantile estimators.

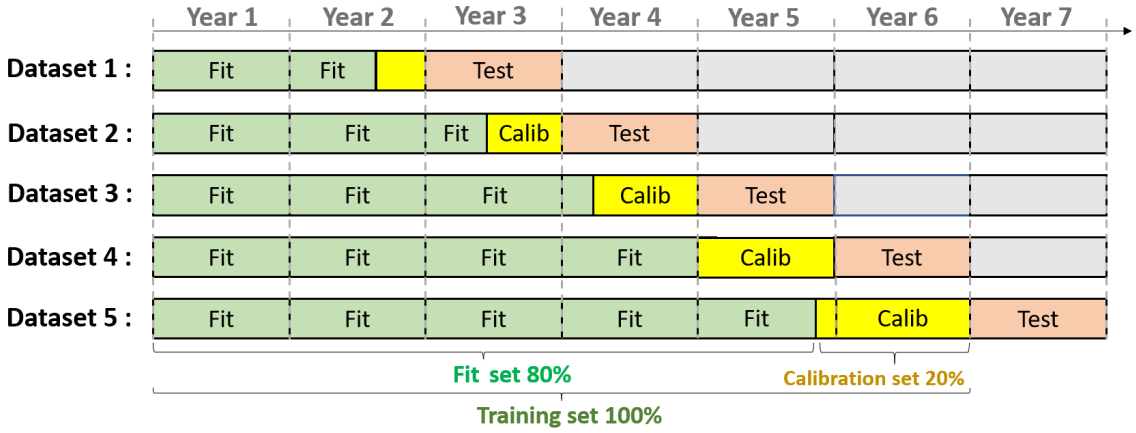

Figure 2: Sequential cross-validation scheme: testing conformal prediction on a cumulative time window.

4. ***Ensemble batch Prediction Interval* (EnbPI)**: for constant PIs from a mean estimator.
5. ***Adaptive Ensemble batch Prediction Interval* (aEnbPI)**: our *variant*, which builds adaptive PIs from both mean and dispersion estimators.

Finally, to ensure the robustness of the results obtained, we set up a sequential cross-validation scheme (Hastie et al., 2009; Cerqueira et al., 2020) with five datasets, as depicted in Figure 2. Each dataset covers the data corresponding to all 29 combinations of source, product, and customer for a number of years. The first dataset covers the first three years of data, the second one includes the first four years of data, and so on. The final year of each dataset is used as the test set. Each training set is then split into fit and calibration sets according to an 80%-20% split ratio following Sesia and Candès (2020) and preliminary tests.

Table 1 summarizes the nine configuration evaluated empirically by combining predictors and conformalization methods[7]. For the baseline configurations without conformalization, as well as EnbPI approaches, the models are fitted on the whole training dataset. As for all the other CP-enabled, we use the sequential cross-validation scheme as described in Figure 2.

4.4. Results

Table 2 and Figure 3 report the averages of the metrics and their standard deviations deriving from the 5-fold cross-validation scheme (see Figure 2).

For the Conformal Predictor approaches based on a mean predictor (XGB–SCP and XGB–EnbPI), we observe coverage (87.8%) close to the target coverage of 90% with a sharpness gain in PINAW for the EnbPI approach. For the UQ approaches without conformalization (QRF, RFMV and GBQR), we observe a rather large deviation from the target

7. The code and data supporting our findings can be made available by the authors upon request.

Predictor		Conformalization				
	without[1]	SCP[2]	EnbPI[1]	CQR[2]	LACP[2]	aEnbPI[1]
XGB		✓	✓			
GBQR	✓			✓		
QRF	✓			✓		
RFMV	✓				✓	✓

Table 1: Summary of the methods evaluated in the experiments.
[1]Approaches without conformalization are trained on 100% of the training set. [2]Sequential cross-validation CP procedure; 80% (resp. 20%) of the train data are assigned to the fit (resp. calibration) set (see Figure 2).

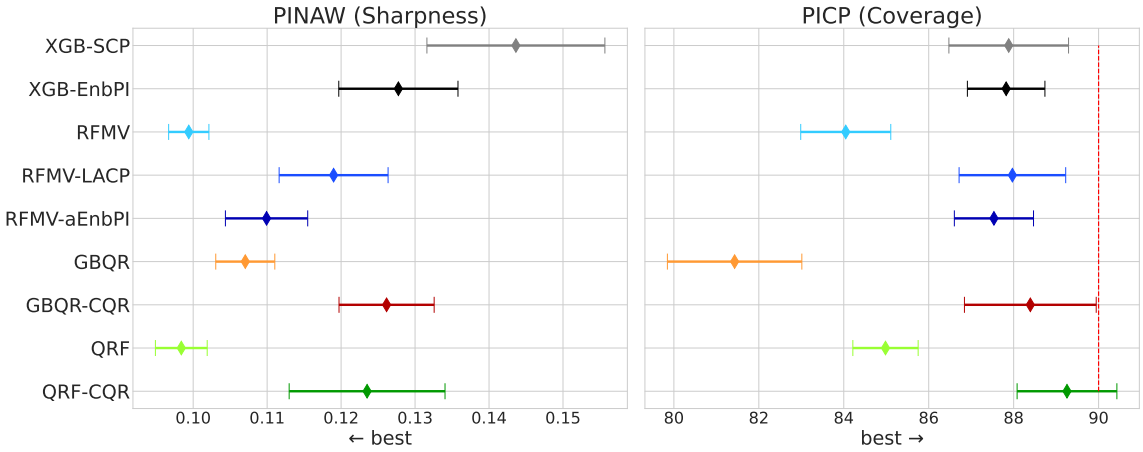

Figure 3: PINAW and PICP (%) (average metric ± standard deviation) averaged over the five datasets (target coverage: 90%).

Predictor	CP approach	PINAW (*sharpness*)	PICP [%] (*coverage*)
XGB	SCP	0.144 ± 0.024	87.88 ± 2.86
XGB	EnbPI	0.128 ± 0.016	87.82 ± 1.82
RFMV	-	0.099 ± 0.005	84.05 ± 2.12
RFMV	LACP	0.119 ± 0.015	87.97 ± 2.51
RFMV	aEnbPI	0.110 ± 0.011	87.54 ± 1.86
GBQR	-	0.107 ± 0.008	81.43 ± 3.17
GBQR	CQR	0.126 ± 0.013	88.39 ± 3.10
QRF	-	0.098 ± 0.007	84.98 ± 1.53
QRF	CQR	0.124 ± 0.021	89.26 ± 2.35

Table 2: PINAW and PICP (%) (average metric ± standard deviation) averaged over the five datasets (target coverage: 90%).

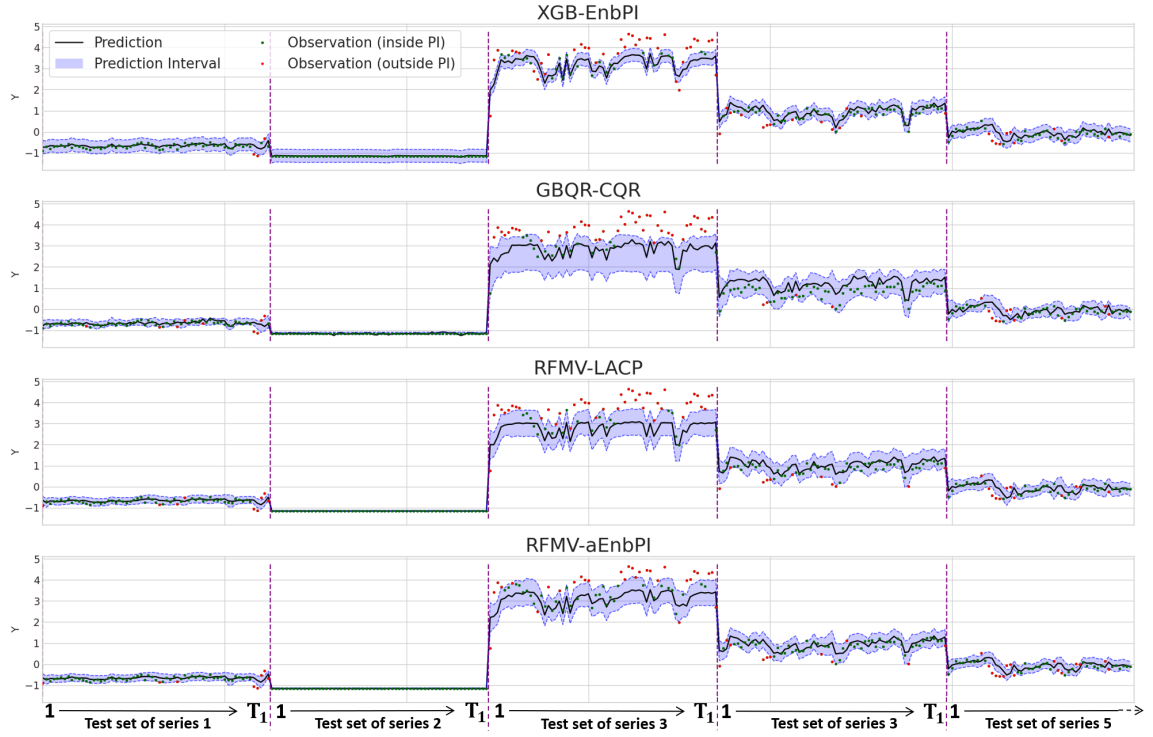

Figure 4: PIs on a test example for four different CP approaches.

coverage for the non-conformalizing approaches (in the range of $81\% - 85\%$). When we apply a conformalization on these (QRF–CQR, RFMV–LACP, GBQR–CQR and RFMV–aEnbPI), we obtain better coverages in the range of $87.5\% - 89.2\%$, closer to the target coverage.

If we analyze the PINAW metrics of conformalized approaches that have a comparable coverage performance, we notice that:

- XGB–SCP produces constant PIs with larger (worse) PINAW performance (0.144).
- XGB–EnbPI produces a narrower constant interval (PINAW performance of 0.128) for the same marginal coverage.
- UQ conformalized approaches (XGB–CQR, QRF–CQR, RFMV–LACP & RFMV–aEnbPI) produce adaptive PIs with a similar PINAW performance (1.24 - 1.28)
- RFMV–aEnbPI produces the narrowest adaptive PIs (1.10) for a slightly lower marginal coverage than other CP approaches.

By analyzing the PIs in Figure 4 produced by 4 approaches XGB–EnbPI, GBQR–CQR, QRF–CQR and RFMV–aEnbPI, we observe that:

- XGB–EnbPI shows constant-size PIs, which seems not ideal given the different variances of the series.
- GBQR–CQR, QRF–CQR and RFMV–EnbPI, on the other hand, show adaptability with narrower PIs for the series with low variability.

- The GBQR–CQR shows larger PIs for some series with high variability than the 2 others approaches.

4.4.1. Coverage Analysis for Operational Profiles

In industrial operations, one can be interested in the coverage related to contextual information that have an operational interest. For example, we can analyze the performances related to each of the 29 series which bear an operational meaning. To simplify the visualization, we cluster the 29 series in three sets: *low-var* for the 10 series whose standard deviation between 0.00001 and 0.2, *mid-var* for the 9 series with standard deviation between 0.2 and 0.38, and *high-var* for the 10 series with standard deviation between 0.38 and 1.0.

In Figure 5, sharpness (on the left) and coverage (on the right) are color-coded following these partition for the 5 conformalized approaches XGB–EnbPI, RFMV–LACP, RFMV–aEnbPI, QRF–CQR and GBQR–CQR. For XGB–EnbPI we observe, as expected, constant PINAW on the three subsets. Contextual coverage shows undercoverage ($\sim 75\%$) for the series with high variance and overcoverage ($\sim 98\%$) on the low variance series. The two compensate each other in the marginal coverage (in black) on the whole dataset.

Conformalized UQ approaches can produce PIs of adaptive size: smaller on series with low variability (in blue) and larger on series with high variability (in red). This adaptability allows balancing the contextual coverage, which are all slightly closer to the target coverage. We find the same observation about XGB–CQR, with larger intervals for series with high variances (in red) than others approaches. Such in-depth analyses can be relevant to industrial considerations, especially in the presence of operational contexts of varying relevance (such as more or less critical systems or periods).

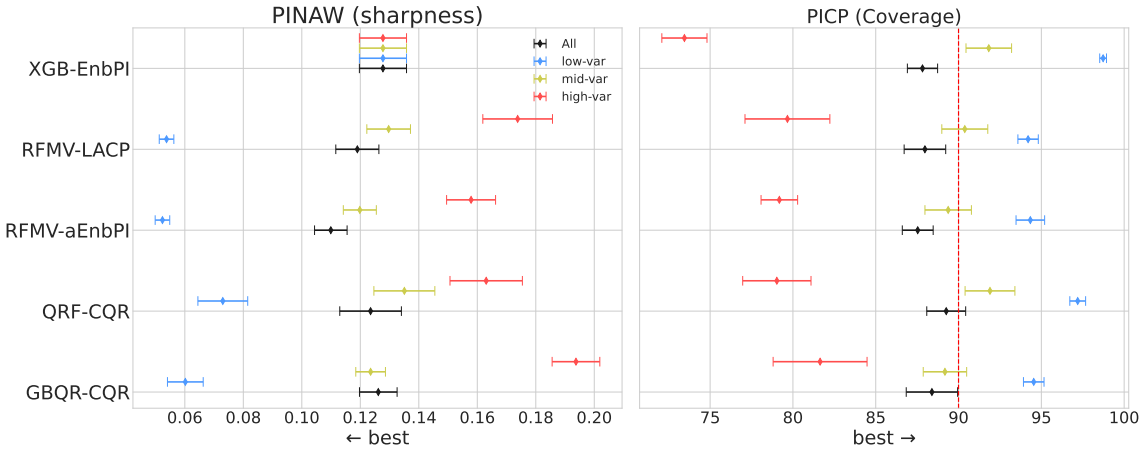

Figure 5: Contextual PINAW and PICP (%) (average metric ± standard deviation) averaged over the five datasets (target coverage 90%). Three time series clusters are considered: *low-var*, *mid-var* and *high-var*.

5. Conclusion

In this paper, we addressed the problem of quantifying the uncertainty in gas demand forecasting for Air Liquide. We opted for several CP methods to provide the probabilistic guarantees needed by the on-site operators. The prediction of gas demand, formalized as a time series regression task, is a priori inconsistent with the assumption of data exchangeability required by some CP methods. Therefore, we designed different experiments to assess the reliability and efficiency of state of the art CP approaches, namely the Split CP, CQR, LACP and EnbPI. Also, a novel version of EnbPI, which we call aEnbPI, was proposed to enhance its local adaptability. The presented results were obtained with PUNCC, a new open-source and user-friendly library that offers ready-to-use wrappers of several CP methods referenced in the literature.

Based on our experience, we recommend CQR on time series as a simple yet effective starting point, to be validated empirically on in-house data. Quantile predictors are accessible to ML engineers in multiple open source implementations, and conformalization is a lightweight post-processing step. If compatible with operational constraints, CP can be executed and updated online following, for instance, the Online Sequential Split CP scheme tested by Zaffran et al. (2022). Finally, we note the relevance of EnbPI and our locally adaptive version for their ability to offer valid coverage when traditional methods fail to attain it.

Following our results, a first on-site testing phase of several weeks is being launched, which could see extended exploitation according to the field results. The versatility of the CP offers many opportunities to increase the trustworthiness in AI and strengthen the connection between state-of-the-art UQ research and reliable industrial applications.

Acknowledgments

This work received French government aid under the Investments for the Future program (PIA) within the framework of SystemX Technological Research Institute. We acknowledge the support of the *confiance.ai* research program, a French consortium with industrial and academic partners that aims to design and industrialize trustworthy AI-based critical systems: `www.confiance.ai`.

References

David J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198. Springer, 1985.

Anastasios Nikolas Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511v3*, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021.

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability, 2022.

Ioannis K Bazionis and Pavlos S Georgilakis. Review of deterministic and probabilistic wind power forecasting: Models, methods, and future research. *Electricity*, 2(1):13–47, 2021.

Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial Intelligence*, 81(1–2):125–144, oct 2017.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996a.

Leo Breiman. Out-of-bag estimation. Technical report, Technical report, Statistics Department, University of California Berkeley, 1996b.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11): 1997–2028, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016.

Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference On Learning Theory*, 2018.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48), 2021.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Distribution-free prediction bands for multivariate functional time series: an application to the italian gas market, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2020.

Isaac Gibbs and Emmanuel Candès. Adaptive Conformal Inference Under Distribution Shift. *arXiv:2106.00170*, June 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Chirag Gupta, Arun K Kuchibhotla, and Aaditya K Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *arXiv preprint arXiv:1910.10562*, 2019.

Gerald J. Hahn and William Q. Meeker. *Statistical intervals: a guide for practitioners*. John Wiley & Sons, Inc., 1991.

T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition edition, 2009.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine learning*, 97(1-2):155–176, 2014.

Byol Kim, Chen Xu, and Rina Barber. Predictive inference is free with the jackknife+-after-bootstrap. In *Advances in Neural Information Processing Systems*, 2020.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006.

Jingyue Pang, Datong Liu, Yu Peng, and Xiyuan Peng. Optimize the coverage probability of prediction interval for anomaly detection of sensor-based monitoring series. *Sensors*, 18(4):967, 2018.

Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, 2002.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, 2019.

Matteo Sesia and Emmanuel J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9:e261, 2020. doi: https://doi.org/10.1002/sta4.261.

Glenn Shafer and Vladimir Vovk. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Kamilė Stankevičiūtė, Ahmed Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Larry Wasserman. *All of Statistics*. Springer, 2004.

Wojciech Wisniewski, David Lindsay, and Sian Lindsay. Application of conformal prediction interval estimations to market makers' net positions. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, 2020.

Chen Xu and Yao Xie. Conformal prediction for dynamic time-series. *arXiv preprint arXiv:2010.09107v12*, 2021a.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, 2021b.

Margaux Zaffran, Aymeric Dieuleveut, Olivier Féron, Yannig Goude, and Julie Josse. Adaptive conformal predictions for time series. *arXiv preprint arXiv:2202.07282*, 2022.

Gianluca Zeni, Matteo Fontana, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *arXiv preprint arXiv:2005.07972*, 2020.

Appendix A. Data details

The dataset has a set of 29 features in total for each observation. There are four ordered categorical features, three categorical ones and finally, three numerical features and a numerical target (gas demand).

- **Years (Ordered-categorical-1):** Year as categorical feature (7 values)
- **Week number (Ordered-categorical-2):** Week as categorical feature (53 values)
- **Categorical-0:** 9 One-hot feature (one for each value)
- **Categorical-1:** 4 One-hot feature (one for each value)
- **Categorical-2:** 2 One-hot feature (one for each value)
- **Ordered-categorical-2:** Categorical feature (3 values)
- **Ordered-categorical-3:** Categorical feature (8 values)
- **Numerical-0**: 4 Latest values of the target (numerical lag)
- **Numerical-1**: 4 Latest values of a correlated signal (numerical lag)
- **Numerical-2**: Latest values of a correlated signal (numerical lag)
- **Numerical-3**: Latest values of a correlated signal (numerical lag)