Conformal Prediction: An Alternative Approach to Uncertainty Quantification in Hierarchical Probabilistic Reconciliation

Azul Garza

October 28, 2023

azul@nixtla.io

Background

- Azul: CTO & Co-Founder of Nixtla

Background

- Azul: CTO & Co-Founder of Nixtla
- I'm a mexican woman, born in Salamanca, Mexico. (Annual GDP Per Capita: 8k USD)

Background

- Azul: CTO & Co-Founder of Nixtla
- I'm a mexican woman, born in Salamanca, Mexico. (Annual GDP Per Capita: 8k USD)
- Don't worry if this is new for you.

Background

- Azul: CTO & Co-Founder of Nixtla
- I'm a mexican woman, born in Salamanca, Mexico. (Annual GDP Per Capita: 8k USD)
- Don't worry if this is new for you.
- First generation student in my family.

Background

- Azul: CTO & Co-Founder of Nixtla
- I'm a mexican woman, born in Salamanca, Mexico. (Annual GDP Per Capita: 8k USD)
- Don't worry if this is new for you.
- First generation student in my family.
- This is my second in-person public talk in English (by myself).

Background

- Azul: CTO & Co-Founder of Nixtla
- I'm a mexican woman, born in Salamanca, Mexico.(Annual GDP Per Capita: 8k USD)
- Don't worry if this is new for you.
- First generation student in my family.
- This is my second in-person public talk in English (by myself).
- Let's take a moment to advocate for **diversity**.

Table of Contents

Table of Contents

Time series forecasting can be formally represented as a multivariate regression problem. For a given timestamp t and horizon H, we are interested in the following conditional probability

$$\mathbb{P}(y_{[t+1:t+H]} \mid X_{[:t+H]}) \quad \text{with} \quad X_{[:t+H]} = \{y_{[:t]},\ x^{(h)}_{[:t]},\ x^{(f)}_{[:t+H]},\ x^{(s)}\} \quad (1)$$

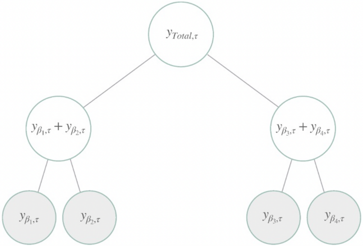
where $\mathbf{y}_{[:t]}$ are the historic values of the target variable, $x^{(h)}_{[:t]}$ the historic exogenous variables, $x^{(f)}_{[:t+H]}$ are future exogenous variables, and $x^{(s)}$ static variables.

Hierarchical time series are data organized at different levels of granularity, constrained by an aggregation rule.

Formally, a hierarchical time series can be denoted by the vector $\mathbf{y}_{[a,b],t}$ defined by the following aggregation constraint:

$$\mathbf{y}_{[a,b],t} = \mathbf{S}_{[a,b][b]}\mathbf{y}_{[b],t} \quad \Leftrightarrow \quad \begin{bmatrix} \mathbf{y}_{[a],t} \\ \mathbf{y}_{[b],t} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{[a][b]} \\ \mathbf{I}_{[b][b]} \end{bmatrix} \mathbf{y}_{[b],t}$$

where $\mathbf{y}_{[a],t}$ are the aggregate series, $\mathbf{y}_{[b],t}$ are the bottom level series and $\mathbf{S}_{[a,b][b]}$ are the hierarchical aggregation constraints.



The example's hierarchical, aggregated and base series are:

$$y_{\text{Total},\tau} = y_{\beta_1,\tau} + y_{\beta_2,\tau} + y_{\beta_3,\tau} + y_{\beta_4,\tau}$$

$$\mathbf{y}_{[a],\tau} = \left[y_{\text{Total},\tau},\ y_{\beta_1,\tau} + y_{\beta_2,\tau},\ y_{\beta_3,\tau} + y_{\beta_4,\tau}\right]^{\intercal}$$

$$\mathbf{y}_{[b],\tau} = \left[y_{\beta_1,\tau},\ y_{\beta_2,\tau},\ y_{\beta_3,\tau},\ y_{\beta_4,\tau}\right]^{\intercal}$$

The summing matrix of the example can be written as:

$$\mathbf{H} = \begin{bmatrix} \mathbf{S}_{[a][b]} \\ \\ \mathbf{I}_{[b][b]} \end{bmatrix} = \left[\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

Figure 1: Hierarchical Forecasting.

Table of Contents

Current Limitations

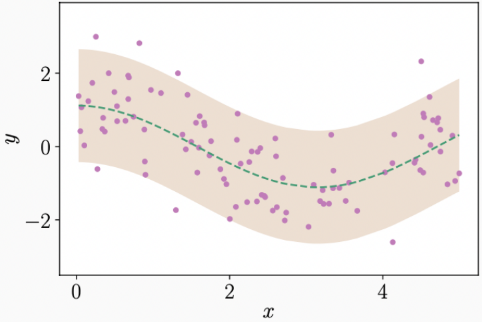
Traditional methods:

- Assume normally distributed forecast errors, which may not always hold in real-world scenarios.
- The interval can be narrow if the model doesn't fit well, leading to overconfidence.
- Does not adapt well to sudden changes or structural breaks in the data.
- Intervals can be sensitive to the choice of hyperparameters.

Conformal Prediction

- Converts point predictors into probabilistic predictors (calibration layer around any model, including black-box models).
- Based on non-conformity measures.
- Provides a layer of reliability to predictions.
- Relevant for time series: Ensures that forecasts respect certain prediction levels.

Main Intuition



$$\mathbb{P}\left\{Y_{n+1} \in \text{Interval}_{\alpha}\left(X_{n+1}\right)\right\} \geq 1-\alpha$$

Figure 2: Main Intuition of Conformal Prediction.

Conformal Prediction for Time Series [1]

- Let $\mathcal{D} = \{(X^{(i)}, y^{(i)})\}_{i=1}^{l}$ be a dataset of time series and forecasts following formulation (1), produced by the forecasting model M.
- CP operates by splitting the dataset into the *proper* training set of size n and the *calibration* set of size m: $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{cal}$
- The validation set is used to compute the *noncoformity score* $R_i \in \mathbb{R}^H$ based on the *residual error*, given for an observation as:

$$R_i = \Delta(M(X^{(i)}|\mathcal{D}, y^{(i)}) = |\hat{y}^{(i)} - y^{(i)}| \qquad (2)$$

[1] Stankeviciute, Kamile, Ahmed M Alaa, and Mihaela van der Schaar. "Conformal time-series forecasting." Advances in neural information processing systems 34 (2021): 6216-6228.

Conformal Prediction for Time Series

- The *empirical nonconformity score distribution* $\{R_i\}_{i=1}^{l}$ is used to compute the *critical nonconformity score* ε_τ, for each $\tau \in \{1, ..., H\}$, given by the $\lceil (m+1)(1-\alpha) \rceil$-th smallest residual.
- The prediction interval for a new observation on the test set and timestamp τ of the forecasting horizon, is given by:

$$\Gamma_\tau^\alpha(X) = [\hat{y}_\tau - \varepsilon_\tau, \hat{y}_\tau + \varepsilon_\tau] \tag{3}$$

Conformal Prediction for Time Series

Our Contribution: Leveraging the principles of conformal prediction for probabilistic hierarchical time series forecasting.

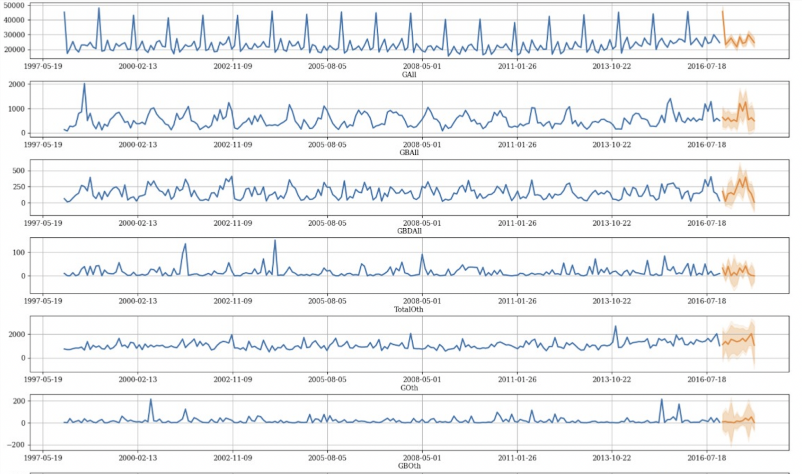


Table of Contents

Data and Preprocessing

- Data sourced from Tourism, Tourism Large, Labour, Traffic, Wiki.
- Preprocessing: Handling missing values.
- Transformation: Aggregating data at different hierarchical levels.

Implementing Conformal Prediction

- **Models**:
  - Statistical Models: AutoARIMA, AutoETS, MSTL, DynamicOptimizedTheta, Theta.
  - Sparse Models: ADIDA, CrostonClassic, CrostonOptimized, IMAPA.
- All the models were reconciled using MinTrace (ols).
- **Conformal layer**: Using non-conformity measure based on prediction errors.
- **Calibration**: Using a hold-out dataset to calibrate the conformal predictor.

Evaluation Metric

$$\mathrm{sCRPS}(\mathbb{P}, \mathbf{y}_{[i],\tau}) = \frac{2}{N} \sum_{i} \frac{\int_0^1 \mathrm{QL}(\mathbb{P}_{i,\tau}, y_{i,\tau})_q dq}{\sum_i |y_{i,\tau}|}$$

where N is the number of time series, $\mathrm{QL}(\hat{\mathbb{P}}_{i,\tau}, y_{i,\tau})_q$ stands for the quantile loss at the q level, between the estimated forecast probability $\hat{\mathbb{P}}i, \tau$ and the observation $y_{i,\tau}$. We use a Riemann approximation to the sCRPS with dq quantile intervals of 1 percent.

Approach

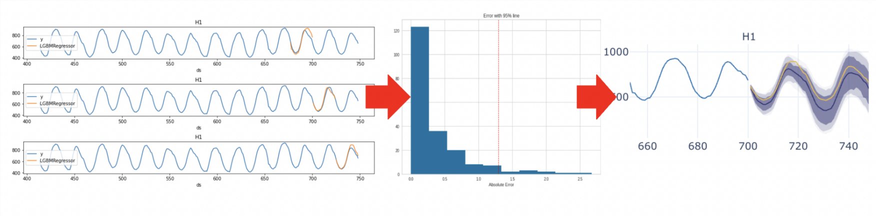


Figure 3: Conformal Prediction for Time Series.

Results Sparse Models

Dataset	TourismLarge	Traffic	Wiki2	TourismSmall	Labour
ADIDA	0.4239	0.1317	0.5616	0.1946	0.0201
CrostonClassic	0.4212	0.1325	0.4727	0.1882	0.0265
CrostonOptimized	0.4237	0.1317	0.5616	0.1946	0.0201
CrostonSBA	0.4190	0.1398	0.4746	0.1792	0.0656
IMAPA	0.4239	0.1317	0.5616	0.1946	0.0201

Figure 4: Scaled CRPS for Sparse Models.

Comparison Against Normality

Dataset	TourismLarge		Traffic		Wiki2		TourismSmall		Labour	
Uncertainty	**Normality**	**Conformal**	**Normality**	**Conformal**	**Normality**	**Conformal**	**Normality**	**Conformal**	**Normality**	**Conformal**
Naive	0.9930	0.5263	0.2459	0.1848	2.1552	0.4050	0.2995	0.2342	0.0184	0.0177
SeasonalNaive	0.4090	0.4660	0.1259	0.1012	1.4894	0.4178	0.1590	0.1712	0.0242	0.0278
AutoETS	0.3317	0.3648	0.1022	0.1014	0.8613	0.8546	0.1648	0.1867	0.0122	0.0162
DynamicOptimizedTheta	0.9246	0.3510	0.5037	0.1343	5.0259	0.6502	0.7259	0.1611	0.0315	0.0160
Theta	0.9547	0.3507	0.5069	0.1347	5.2842	0.6190	1.2139	0.1695	0.0716	0.0156
AutoARIMA	0.3462	0.3934	0.1080	0.1047	1.1527	0.9674	0.1647	0.1897	0.0157	0.0214
MSTL	0.3724	0.4307	0.0976	0.1105	0.7042	0.7973	0.1850	0.2176	0.0102	0.0147
ADIDA	nan	0.4239	nan	0.1317	nan	0.5616	nan	0.1946	nan	0.0201
CrostonClassic	nan	0.4212	nan	0.1325	nan	0.4727	nan	0.1882	nan	0.0265
CrostonOptimized	nan	0.4237	nan	0.1317	nan	0.5616	nan	0.1946	nan	0.0201
CrostonSBA	nan	0.4190	nan	0.1398	nan	0.4746	nan	0.1792	nan	0.0656
IMAPA	nan	0.4239	nan	0.1317	nan	0.5616	nan	0.1946	nan	0.0201

Figure 5: Scaled CRPS for Models.

Distribution of Errors

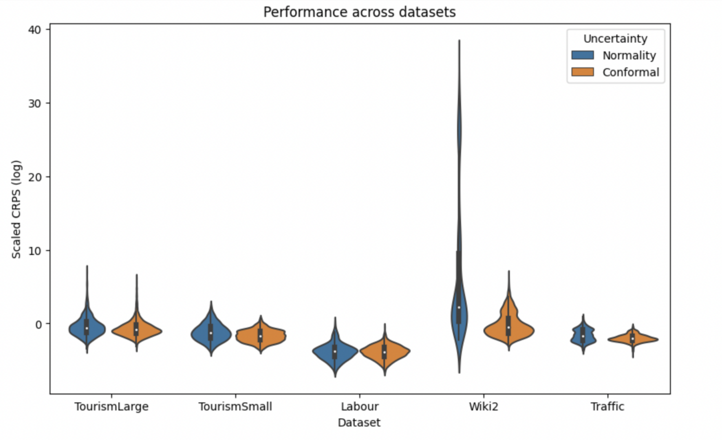


Figure 6: Distribution of errors across datasets.

Distribution of Errors Across Models

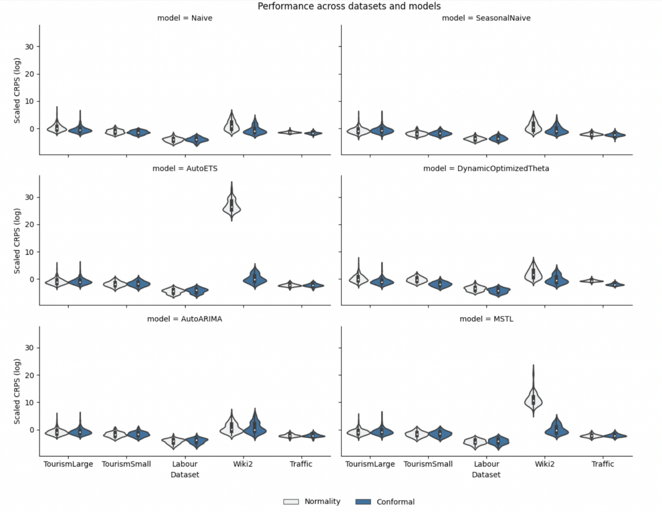


Figure 7: Distribution of errors across datasets and models.

Intuition of Results

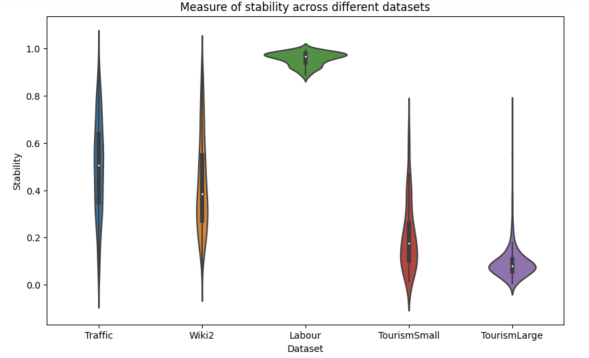


Figure 8: Stability measure across datasets.

Table of Contents

Conclusion

- Conformal Prediction offers accurate uncertainty quantification for sparse methods.
- Is more accurate than traditional approaches measured by sCRPS.
- Offers an interesting alternative to uncertainty quantification for new hierarchical time series forecasting methods.

Future Work

- Explore other non-conformity measures for better calibration.
- Expand to multivariate hierarchical time series.
- Integration with other forecasting models.

Questions?

Thank you for your attention!