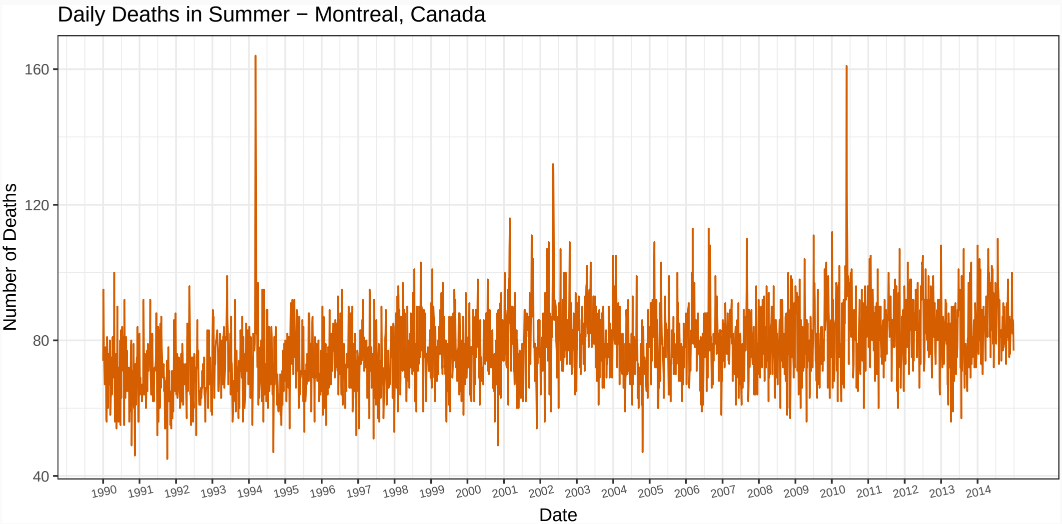
Optimal Predictor Selection for High-dimensional Nonparametric Forecasting

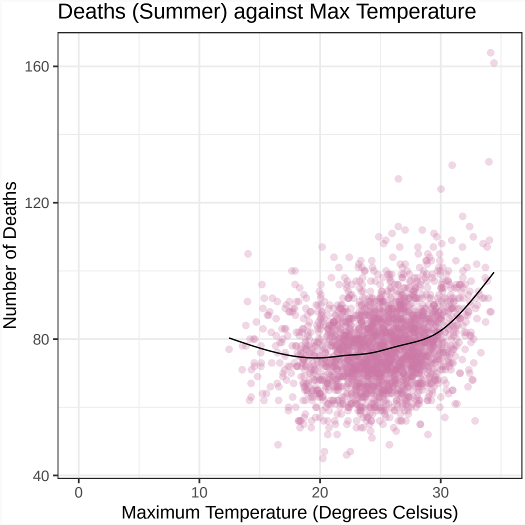
Nuwani Palihawadana

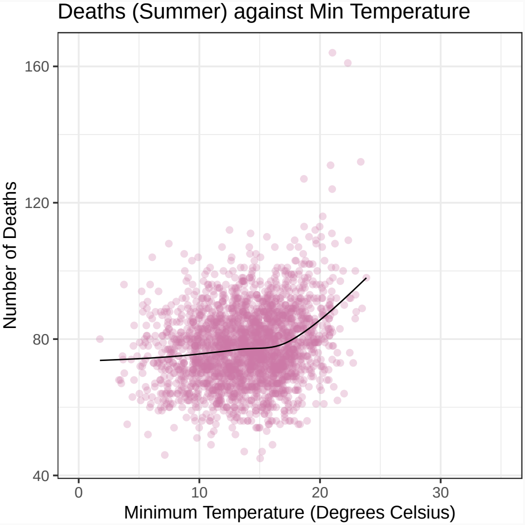
Supervisors:
Prof. Rob J Hyndman, Dr Xiaoqian Wang &
Prof. Louise M Ryan

Heat Exposure Related Daily Mortality



Heat Exposure Related Daily Mortality





Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-k}, y_1, \dots, y_{t-l}) + \varepsilon_t$$

y_t – variable to forecast

$\boldsymbol{x}_t$ – a vector of predictors

ε_t – random error

Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-k}, y_1, \dots, y_{t-l}) + \varepsilon_t$$

y_t – variable to forecast
$\boldsymbol{x}_t$ – a vector of predictors
ε_t – random error

- Impossible to estimate f for large k – ***curse of dimensionality***

Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-k}, y_1, \dots, y_{t-l}) + \varepsilon_t$$

y_t – variable to forecast
$\boldsymbol{x}_t$ – a vector of predictors
ε_t – random error

- Impossible to estimate f for large k – ***curse of dimensionality***
- Reasonable to impose additivity constraints

$$f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-k}) = \sum_{a=0}^{k} f_a(\boldsymbol{x}_{t-a})$$

Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-k}, y_1, \dots, y_{t-l}) + \varepsilon_t$$

y_t – variable to forecast
$\boldsymbol{x}_t$ – a vector of predictors
ε_t – random error

- Impossible to estimate f for large k – ***curse of dimensionality***
- Reasonable to impose additivity constraints

$$f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-k}) = \sum_{a=0}^{k} f_a(\boldsymbol{x}_{t-a}) \leftarrow$$ **Nonparametric Additive Model**

Background

Issues:

1. Challenging to estimate in a high-dimensional setting
2. Subjectivity in predictor selection, and predictor grouping to model interactions

Background

Issues:

1. Challenging to estimate in a high-dimensional setting
2. Subjectivity in predictor selection, and predictor grouping to model interactions

Index Models:

- **Mitigate difficulty of estimating a nonparametric component for each predictor**
- **Improve flexibility**

$$y_i = g\left(\boldsymbol{\alpha}^T \boldsymbol{x}_i\right) + \varepsilon_i$$

Sparse Multiple Index (SMI) Model

Semi-parametric model

$$y_i = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}) + \sum_{k=1}^{d} f_k(w_{ik}) + \boldsymbol{\theta}^T \boldsymbol{u}_i + \varepsilon_i, \quad i = 1, \dots, n,$$

- y_i – univariate response
- $\boldsymbol{x}_{ij} \in \mathbb{R}^{\ell_j}$, $j = 1, \dots, p$ – p subsets of predictors entering indices
- $\boldsymbol{\alpha}_j$ – ℓ_j-dimensional vectors of index coefficients
- g_j, f_k – smooth nonlinear functions
- Additional predictors :
  - w_{ik} – nonlinear
  - $\boldsymbol{u}_i$ – linear

Sparse Multiple Index (SMI) Model

Semi-parametric model

$$y_i = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}) + \sum_{k=1}^{d} f_k(w_{ik}) + \boldsymbol{\theta}^T \boldsymbol{u}_i + \varepsilon_i, \quad i = 1, \dots, n,$$

- y_i – univariate response
- $\boldsymbol{x}_{ij} \in \mathbb{R}^{\ell_j}$, $j = 1, \dots, p$ – p subsets of predictors entering indices
- $\boldsymbol{\alpha}_j$ – ℓ_j-dimensional vectors of index coefficients
- g_j, f_k – smooth nonlinear functions
- Additional predictors :
  - w_{ik} – nonlinear
  - $\boldsymbol{u}_i$ – linear

Allow elements equal to zero in $\boldsymbol{\alpha}_j$ – ”Sparse”

Sparse Multiple Index (SMI) Model

Semi-parametric model

$$y_i = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}) + \sum_{k=1}^{d} f_k(w_{ik}) + \boldsymbol{\theta}^T \boldsymbol{u}_i + \varepsilon_i, \quad i = 1, \dots, n,$$

- y_i – univariate response
- $\boldsymbol{x}_{ij} \in \mathbb{R}^{\ell_j}$, $j = 1, \dots, p$ – p subsets of predictors entering indices
- $\boldsymbol{\alpha}_j$ – ℓ_j-dimensional vectors of index coefficients
- g_j, f_k – smooth nonlinear functions
- Additional predictors :
  - w_{ik} – nonlinear
  - $\boldsymbol{u}_i$ – linear

Both "p" and the predictor grouping among indices are unknown.

Overlapping of predictors among indices is not allowed.

Optimisation Problem

Let q be the *total number of predictors* entering indices.

$$\min_{\beta_0, p, \boldsymbol{\alpha}, \boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}} \quad \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) - \sum_{k=1}^{d} f_k(w_{ik}) - \boldsymbol{\theta}^T \boldsymbol{u}_i \right]^2$$

$$+ \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} \mathbb{1}(\alpha_{jm} \neq 0) + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\alpha}_j\|_2^2$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \mathbb{1}(\alpha_{jm} \neq 0) \in \{0, 1\} \quad \forall m$$

- $\lambda_0 > 0$ – controls the number of selected predictors
- $\lambda_2 \geq 0$ – controls the strength of the additional shrinkage

MIQP Formulation

$$\min_{\beta_0, p, \boldsymbol{\alpha}, \boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{z}} \quad \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) - \sum_{k=1}^{d} f_k(w_{ik}) - \boldsymbol{\theta}^T \boldsymbol{u}_i \right]^2 + \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} z_{jm} + \lambda_2 \sum_{j=1}^{p} \sum_{m=1}^{q} \alpha_{jm}^2$$

$$\text{s.t.} \quad |\alpha_{jm}| \leq M z_{jm} \quad \forall j, \forall m,$$

$$\sum_{j=1}^{p} z_{jm} \leq 1 \quad \forall m,$$

$$z_{jm} \in \{0, 1\} \quad \leftarrow \quad z_{jm} = \mathbb{1}(\alpha_{jm} \neq 0)$$

- $M < \infty$: If $\boldsymbol{\alpha}^*$ is an optimal solution, then $\max\left(\{|\alpha_{jm}^*|\}_{j \in [p], m \in [q]}\right) \leq M$

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

- **PPR:** Projection Pursuit Regression Based Initialisation
- **Additive:** Nonparametric Additive Model Based Initialisation
- **Linear:** Linear Regression Based Initialisation
- **Multiple:** Pick One From Multiple Initialisations

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Step 2 : Estimate nonlinear functions

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Step 2 : Estimate nonlinear functions

Step 3 : Update index coefficients

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Step 2 : Estimate nonlinear functions

Step 3 : Update index coefficients

Step 4 : Iterate steps 2 and 3 – until:

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Step 2 : Estimate nonlinear functions

Step 3 : Update index coefficients

Step 4 : Iterate steps 2 and 3 – until:

- convergence
- loss increases for 3 consecutive iterations OR
- max iterations

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Step 2 : Estimate nonlinear functions

Step 3 : Update index coefficients

Step 4 : Iterate steps 2 and 3 until stopping criteria are reached

Step 5 : Add a new index with dropped predictors, and repeat step 4

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Step 2 : Estimate nonlinear functions

Step 3 : Update index coefficients

Step 4 : Iterate steps 2 and 3 until stopping criteria are reached

Step 5 : Add a new index with dropped predictors, and repeat step 4

Step 6 : Increase p by 1 in each iteration of step 5 – until:

Estimation Algorithm

Step 1 : Initialise index structure and index coefficients

Step 2 : Estimate nonlinear functions

Step 3 : Update index coefficients

Step 4 : Iterate steps 2 and 3 until stopping criteria are reached

Step 5 : Add a new index with dropped predictors, and repeat step 4

Step 6 : Increase p by 1 in each iteration of step 5 – until:

- no.of indices reaches q
- loss increases after the increment model OR
- solution maintains same no.of indices as previous iteration, and abs(difference of index coefficients between two successive iterations) $<=$ tolerance

Forecasting Heat Exposure Related Daily Mortality

Variables

- **Response: Daily deaths in Summer** – 1990 to 2014 – Montreal, Canada
- **Index Variables:**
  - Death lags
  - Max temperature lags
  - Min temperature lags
  - Vapor pressure lags
- **Nonlinear:** DOS (day of the season), Year

Forecasting Heat Exposure Related Daily Mortality

Variables

- **Response:** **Daily deaths in Summer** – 1990 to 2014 – Montreal, Canada
- **Index Variables:**
  - Death lags
  - Max temperature lags
  - Min temperature lags
  - Vapor pressure lags
- **Nonlinear:** DOS (day of the season), Year

$$\textbf{Deaths} = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{X}\boldsymbol{\alpha}_j) + f_1(\textbf{DOS}) + f_2(\textbf{Year}) + \varepsilon,$$

Forecasting Heat Exposure Related Daily Mortality

Variables

- **Response:** **Daily deaths in Summer** – 1990 to 2014 – Montreal, Canada
- **Index Variables:**
  - Death lags
  - Max temperature lags
  - Min temperature lags
  - Vapor pressure lags
- **Nonlinear:** DOS (day of the season), Year

Data Split

- **Training Set:** 1990 to 2012
- **Validation Set:** 2013
- **Test Set:** 2014

$$\textbf{Deaths} = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{X}\boldsymbol{\alpha}_j) + f_1(\textbf{DOS}) + f_2(\textbf{Year}) + \varepsilon,$$

Results

Model	Predictors	Indices	Test Set 1		Test Set 2	
			MSE	MAE	MSE	MAE
SMI Model (5, 12) - PPR	61	7	**85.233**	**7.140**	**97.353**	**7.772**
SMI Model (1, 0) - Additive	61	59	96.398	7.481	112.199	8.156
SMI Model (6, 11) - Linear	61	2	100.231	7.719	120.542	8.598
Backward Elimination	40	—	136.204	9.319	140.867	9.385
Group-wise Additive Index Model	61	4	90.763	7.247	106.251	7.928
Projection Pursuit Regression	61	4	90.698	7.343	110.497	8.057

SMI Model (a, b) $\rightarrow$ $\boldsymbol{\lambda}_0 = a, \boldsymbol{\lambda}_2 = b$

- **Test Set 1:** Three months (June, July and August 2014)
- **Test Set 2:** One month (June 2014)

Summary

Key features:

- Automatic selection of number of indices and predictor grouping
- Automatic predictor selection
- A wide spectrum: from single index models to additive models

Limitations:

- Initialisation: we encourage trial-and-error
- Computational cost: increases with number of predictors and indices

Paper : github.com/nuwani-palihawadana/smimodel_paper

R Package - smimodel

- Open source implementation of **SMI Modelling Algorithm**
  - **model_smimodel()**
- Penalty parameter tuning with greedy search
  - **greedy_smimodel()**
- Functions to fit benchmark models
  - **model_backward()**
  - **model_gaim()**
  - **model_ppr() etc.**

github.com/nuwani-palihawadana/smimodel

Uncertainty Estimation

- **"Uncertainty"** of a forecast $\rightarrow$ **Prediction Interval (PI)**

Uncertainty Estimation

- **"Uncertainty"** of a forecast $\rightarrow$ **Prediction Interval (PI)**
- Theoretical $100(1-\alpha)\%$ prediction interval:

$$\hat{y}_{t+h|t} \pm z_{\alpha/2} * \hat{\sigma}_h,$$

where

- y – time series $y_1, \dots, y_T$
- $\hat{y}_{t+h|t}$ – h steps ahead point forecast for y_{t+h}
- $z_{\alpha/2}$ – $\alpha/2$ quantile of standard normal distribution
- $\hat{\sigma}_h$ – an estimate of std. deviation of h-step forecast distribution

Uncertainty Estimation

- **"Uncertainty"** of a forecast $\rightarrow$ **Prediction Interval (PI)**
- Theoretical $100(1-\alpha)\%$ prediction interval:

$$\hat{y}_{t+h|t} \pm z_{\alpha/2} * \hat{\sigma}_h,$$

where

  - y – time series $y_1, \dots, y_T$
  - $\hat{y}_{t+h|t}$ – h steps ahead point forecast for y_{t+h}
  - $z_{\alpha/2}$ – $\alpha/2$ quantile of standard normal distribution
  - $\hat{\sigma}_h$ – an estimate of std. deviation of h-step forecast distribution
- Nonparametric Additive Models:
  - No distributional assumptions
  - Serially correlated errors $\rightarrow$ **Impossible to estimate theoretical PIs**

Methodology

Block Bootstrap

- Resampling from empirical distribution of historical model residuals
  $\rightarrow$ Bootstrapping

Methodology

Block Bootstrap

- Resampling from empirical distribution of historical model residuals
  $\rightarrow$ Bootstrapping
- Randomly resample blocks from the historical model residuals, and join together $\rightarrow$ **Block Bootstrapping**
- Retains serial correlation in the data

Methodology

Block Bootstrap

- Resampling from empirical distribution of historical model residuals $\rightarrow$ Bootstrapping
- Randomly resample blocks from the historical model residuals, and join together $\rightarrow$ **Block Bootstrapping**
- Retains serial correlation in the data
- block length:
  - Long enough to capture autocorrelation patterns
  - Short enough to construct sufficient number of blocks

Methodology

Conformal Prediction (CP) – Vovk et al. (2005)

- A distribution-free approach
- Relies only on the assumption of **exchangeability of data**
- Provides theoretical coverage guarantees

Methodology

Conformal Prediction (CP) – Vovk et al. (2005)

- A distribution-free approach
- Relies only on the assumption of **exchangeability of data**
- Provides theoretical coverage guarantees

Split Conformal Prediction (SCP)

- A holdout method for generating prediction intervals
  - **Training set** – forecasting model is trained
  - **Calibration set** – forecasting errors (*nonconformity scores*) are calculated
  - **Test set** – prediction intervals are obtained

Methodology

- CP methods for **non-exchangeable data**:

Methodology

- CP methods for **non-exchangeable data**:

Weighted Conformal Prediction (WCP) Methods

- Tibshirani et al. (2019):
  - Depends on **”covariate shift”** assumption
  - *Nonconformity scores* are weighted using ratio of likelihoods of training and test covariate distributions
  - Likelihood ratio is assumed to be known or accurately estimated

Methodology

- CP methods for **non-exchangeable data**:

Weighted Conformal Prediction (WCP) Methods

- Tibshirani et al. (2019):
  - Depends on **"covariate shift"** assumption
  - *Nonconformity scores* are weighted using ratio of likelihoods of training and test covariate distributions
  - Likelihood ratio is assumed to be known or accurately estimated
- Barber et al. (2023):
  - Weighting *quantiles* to avoid assumption of exchangeability
  - Weights are "fixed" rather than being data dependent

Methodology

- CP methods for **non-exchangeable data**:

Adaptive Conformal Prediction (ACP) – Gibbs & Candès (2021)

- Update nominal α based on achieved coverage
  - If achieved coverage is larger – increase α
  - If achieved coverage is smaller – decrease α

Methodology

- Prediction interval construction methods
  - Block Bootstrapping (BB)
  - Conformal Prediction (CP) methods: SCP, WSCP, ACP

- Applied using *online learning framework* proposed by *Wang and Hyndman (2024)*

Forecasting Heat Exposure Related Daily Mortality

Data Recap

- **Response:** **Daily deaths in Summer** – 1990 to 2014 – Montreal, Canada
- **Index Variables:**
  - Death lags
  - Max temperature lags
  - Min temperature lags
  - Vapor pressure lags
- **Nonlinear:** DOS (day of the season), Year

Forecasting Heat Exposure Related Daily Mortality

Data Recap

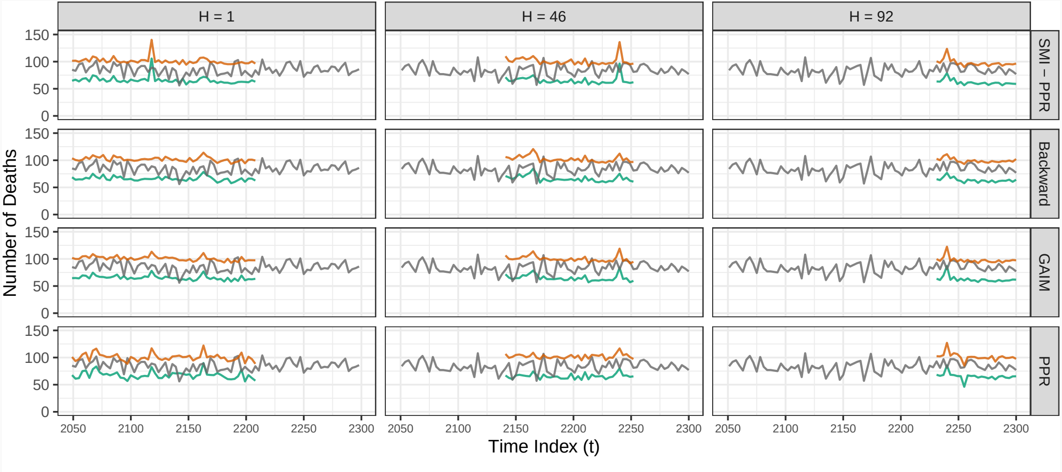
- **Response:** **Daily deaths in Summer** – 1990 to 2014 – Montreal, Canada
- **Index Variables:**
  - Death lags
  - Max temperature lags
  - Min temperature lags
  - Vapor pressure lags
- **Nonlinear:** DOS (day of the season), Year

New Data Split

- **Training Set:** 1990 to 2007
- **Validation Set:** 2008
- **Test Set:** 2009 to 2014

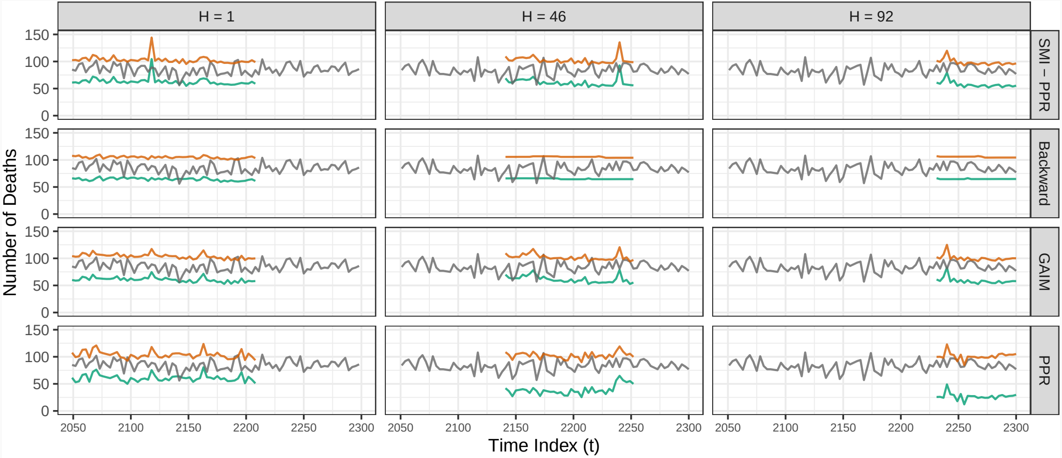
Results

Block Bootstrap 95% Prediction Intervals (block length = 59):

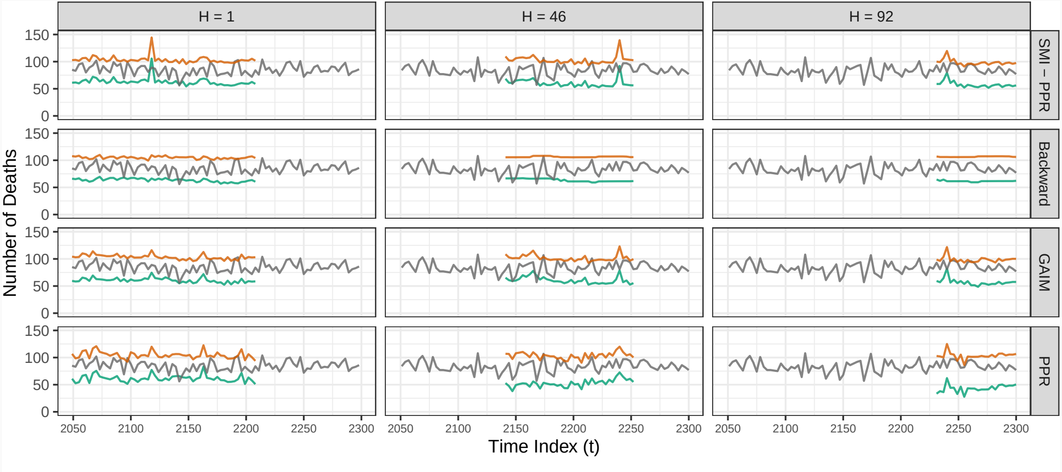


Results

Split Conformal Prediction 95% Prediction Intervals:

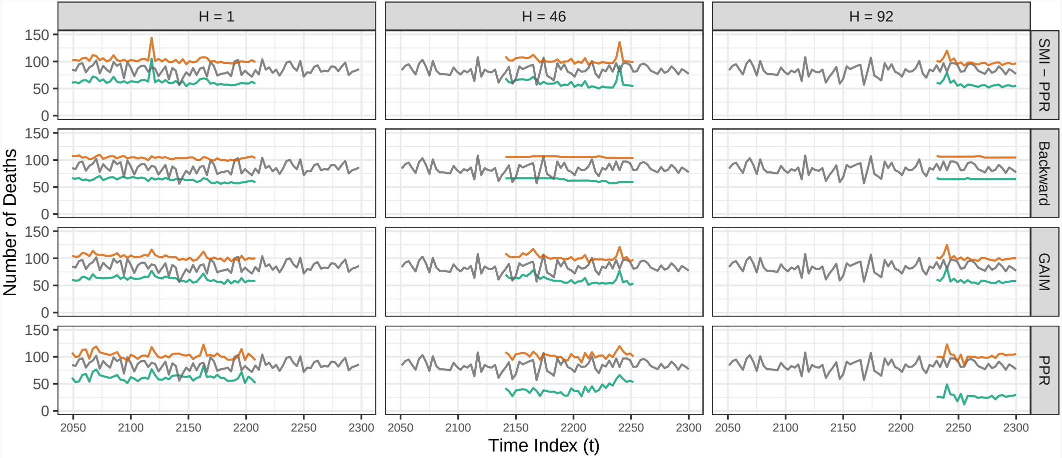


Weighted Split Conformal Prediction 95% Prediction Intervals:



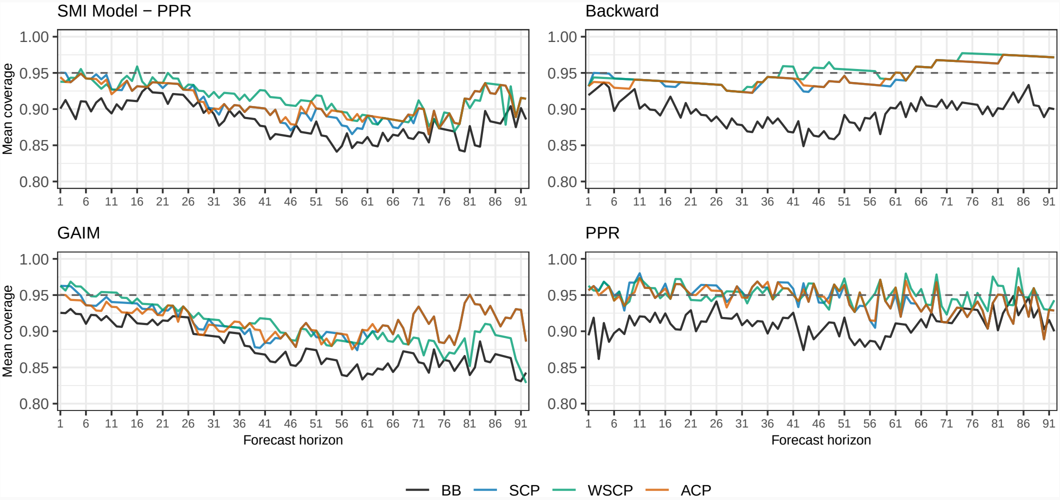
Results

Adaptive Conformal Prediction 95% Prediction Intervals:



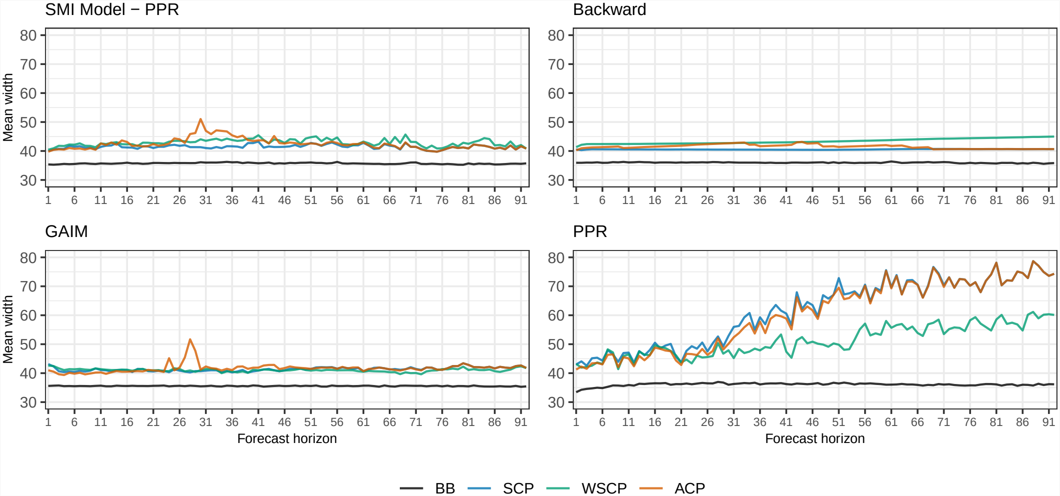
Results

Mean Coverage:



Results

Mean Width:



Summary

Summary of Results (work-in-progress):

- **Block Bootstrap** – Under-coverage; too narrow
- **Conformal Prediction** – Better achieves a target coverage, with acceptable sharpness

Limitations:

- Test set is not long enough for larger forecast horizons
- Hyper-parameter choices

Future Work

- We propose a novel methodology: **Conformal Block Bootstrap (CBB)**
  - **A natural integration of BB and SCP**
  - **Exploits the strengths of both the methods**

Future Work

- We propose a novel methodology: **Conformal Block Bootstrap (CBB)**
  - **A natural integration of BB and SCP**
  - **Exploits the strengths of both the methods**

Find me :

nuwanipalihawadana.netlify.app
in/nuwani-palihawadana
@nuwani-palihawadana
nuwani.kodikarapalihawadana@monash.edu

References

- Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic learning in a random world*, New York, NY: Springer.
- Tibshirani, R., Barber, R., Candès, E., and Ramdas, A. (2019), "Conformal prediction under covariate shift", *Advances in neural information processing systems*, 2526–2536.
- Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2023), "Conformal prediction beyond exchangeability", *The Annals of Statistics*, 51, 816–845.
- Gibbs, I., and Candès, E. (2021), "Adaptive conformal inference under distribution shift", *Advances in neural information processing systems*, 1660–1672.
- Wang, X., and Hyndman, R. J. (2024), "Online conformal inference for multi-step time series forecasting".