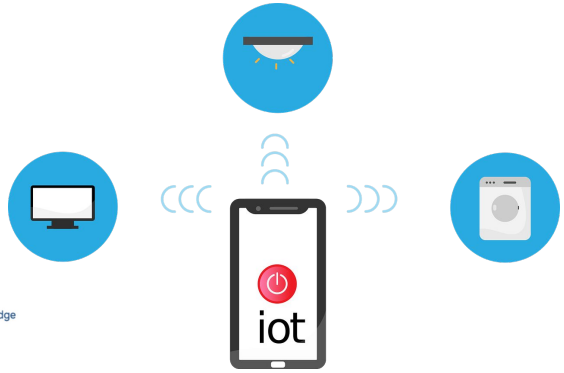
DALHOUSIE UNIVERSITY
Inspiring Minds

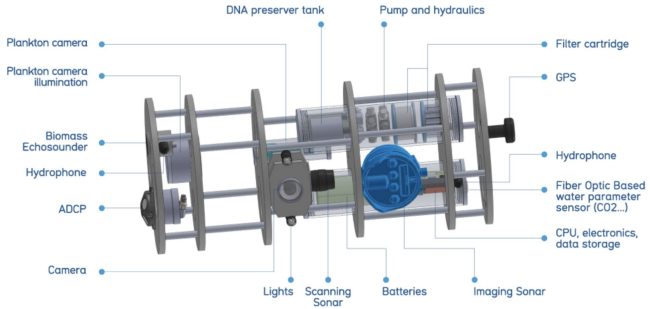
Time Series Forecasting

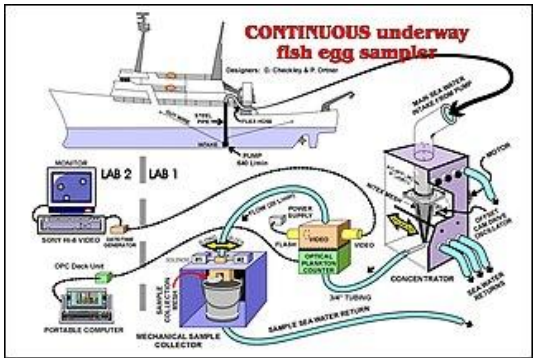
Some Challenges and Possible Solutions

Luis Torgo
ltorgo@dal.ca web.cs.dal.ca/~ltorgo

The Increasing Importance of Time Series



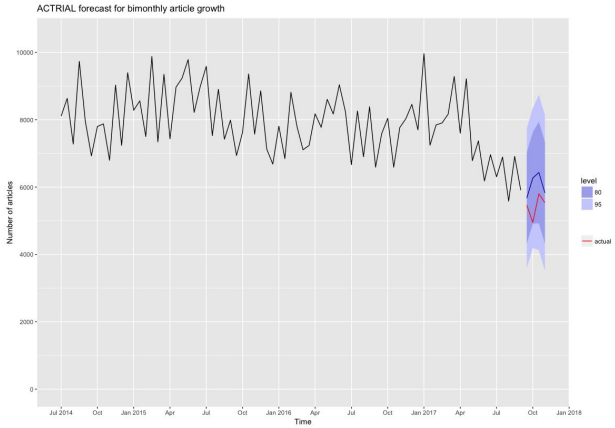


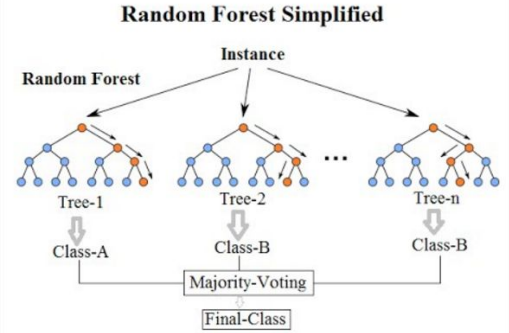


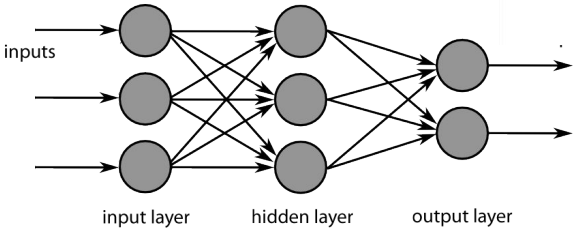
DALHOUSIE UNIVERSITY
Inspiring Minds

Many Available Approaches

- Many research areas
  - Econometrics
  - Statistics
  - Machine learning
  - ….
- Many different types of models
  - Different assumptions
  - Different pros and cons
  - Different results on different benchmarks







Some Questions We Will Address

Q1: Which model/approach should I use for this dataset?

- Model evaluation and comparison

Q2: Are there models/approaches that are clearly better than others?

- Discussion of some existing benchmark results

Q3: Are ensembles a good answer to face the diversity of problems?

- Diversity of models; aggregation; adaptation to different regimes

Q1: Which model/approach should I use for this dataset?

Performance Estimation Methods for Time Series Forecasting Models

Cerqueira V., Torgo L. and Mozetic I. (2020): **Evaluating Time Series Forecasting Models: an empirical study on performance estimation methods**. In *Machine Learning*, 109 (11), 1997-2028.

Why?

- Crucial Step of Predictive Analytics
  - Deliver not only a model but what you can expect from it in terms of predictive performance

- It is the basis of proper parameter selection / model tuning
  - Complex models have far too many parameters to set

- Allows the analyst to select the “best” model for an application
  - Too many models to choose from

Performance Estimation

- *Goal*: **using the available data, obtain a reliable estimate of the performance of any model on unseen data**
  - Performance on seen data is unreliable due to overfitting

- Main Classes of Approaches for Performance Estimation for Time Series:
  - Cross validation
  - Out of Sample
  - Prequential
- Different forms of using the available data for estimating performance
  - What is the impact on the quality of the estimates given the properties of time series data?
- Time series observations are not independent
  - **Ignoring these dependencies may introduce biases in the performance estimates**

Cross Validation Approaches

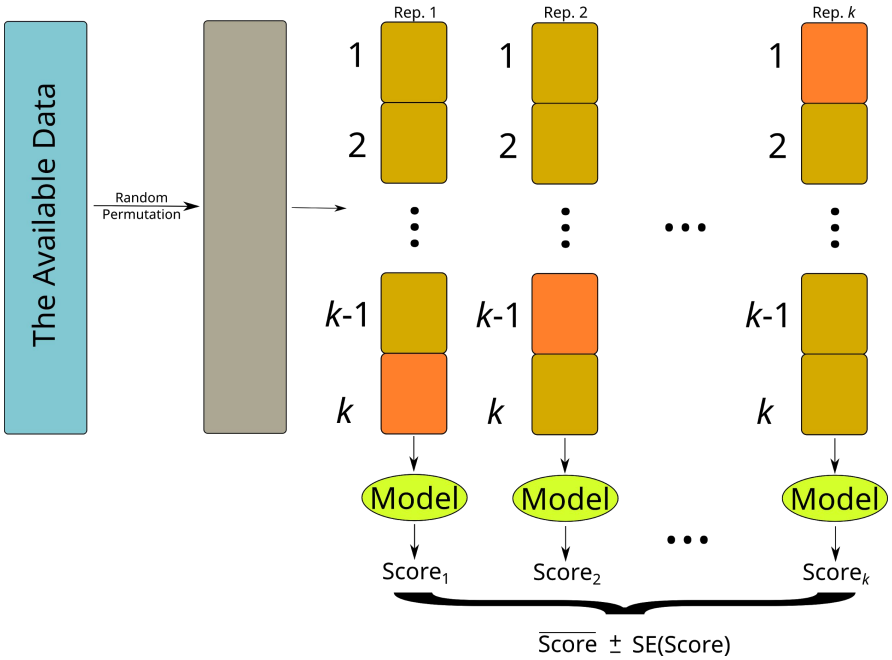
- Iterative process that uses the available data in a very efficient way
  - All observations are used for both training and testing across the iterations
  - This efficiency is particularly relevant for small data sets

- Key Potential Problem: order of the observations is not preserved
  - Bergmeir et al (2018) show that there is no problem for stationary time series
  - Real world time series are frequently non-stationary
  - Several variants of CV have been proposed to overcome this potential drawback

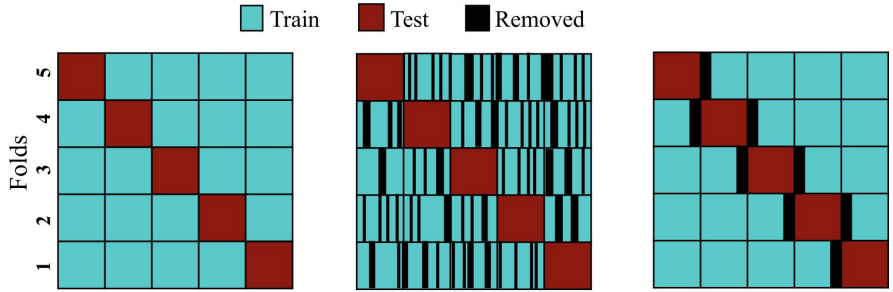
Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83.

Cross Validation Approaches (cont.)

- Standard CV

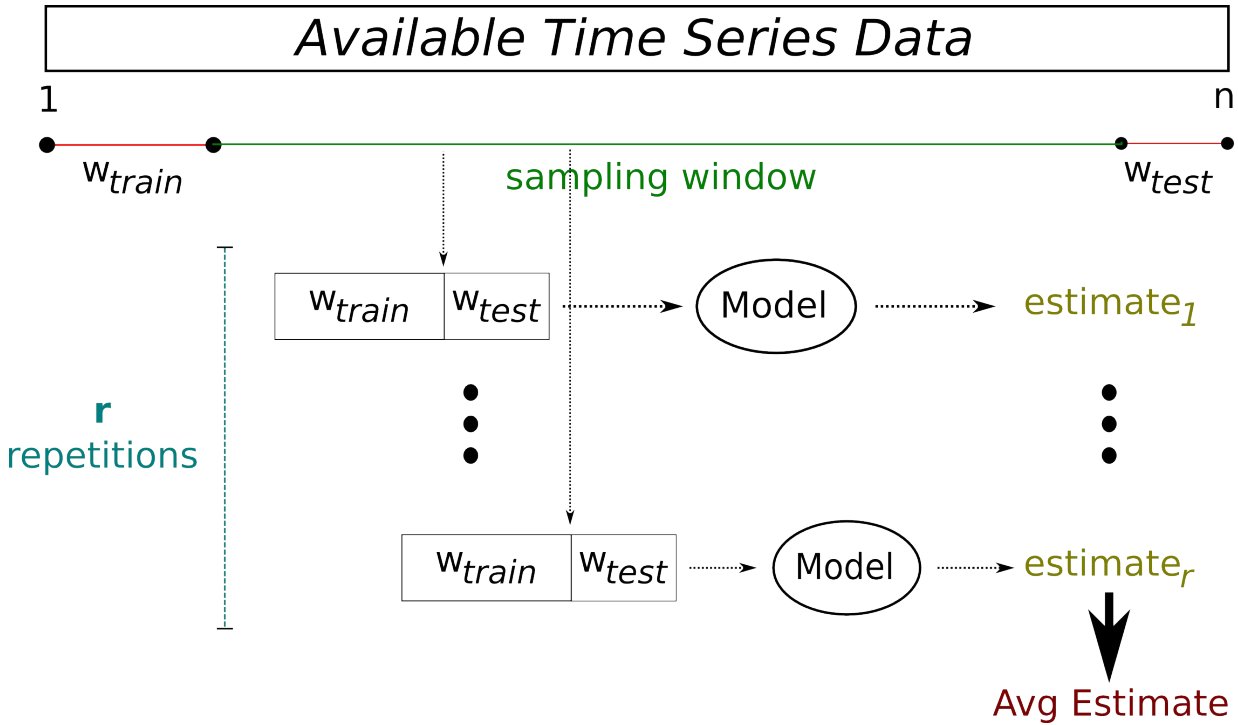


- Some variants
  - Blocked K-fold CV (Snijders, 1988)
    - No shuffling
    - Order within blocks maintained
  - Modified K-fold CV (McQuarrie & Tsai, 1998)
    - Remove from training set obs correlated with the test samples
  - hv-blocked K-fold CV (Racine, 2000)
    - Similar to blocked K-fold CV, but obs adjacent to train and test are removed



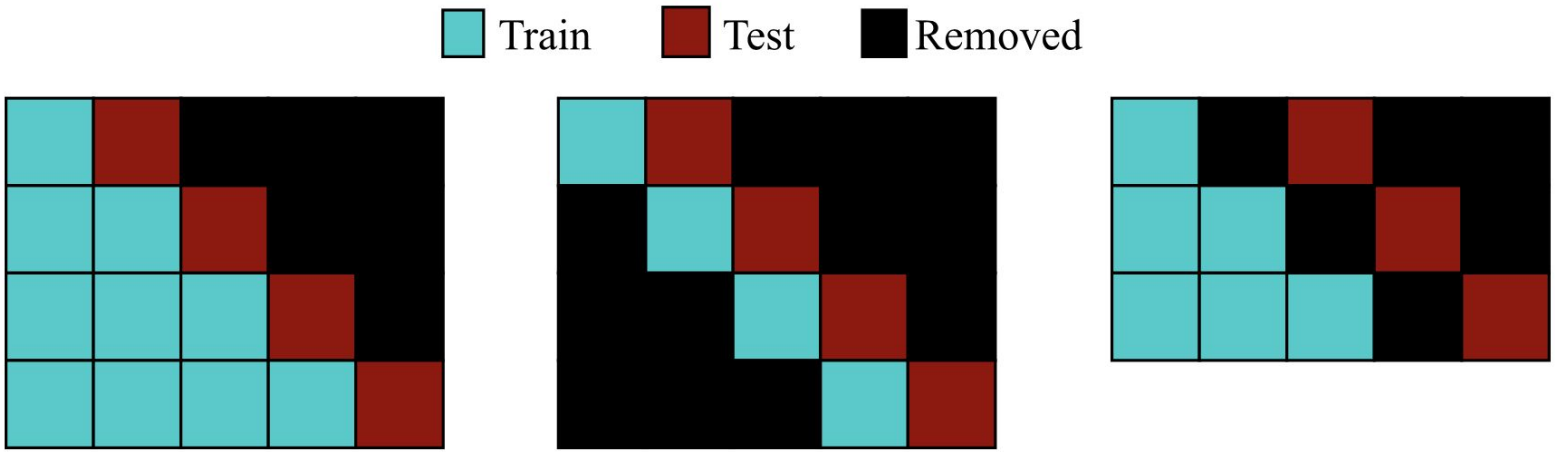
Out of Sample Approaches

- Respect the order of the observations
- Train models on a window and test them on the subsequent window
- Repeat, for different random dates



Prequential Approaches

- Each observation is first used for testing and then for training
  - Implemented in different ways
    - Growing, sliding, or even adding gaps



Comparing Estimators

Used Procedure

- Check which estimation method g_i produces the best estimate $\hat{g}_i$ of the true loss L^m of the model *m*

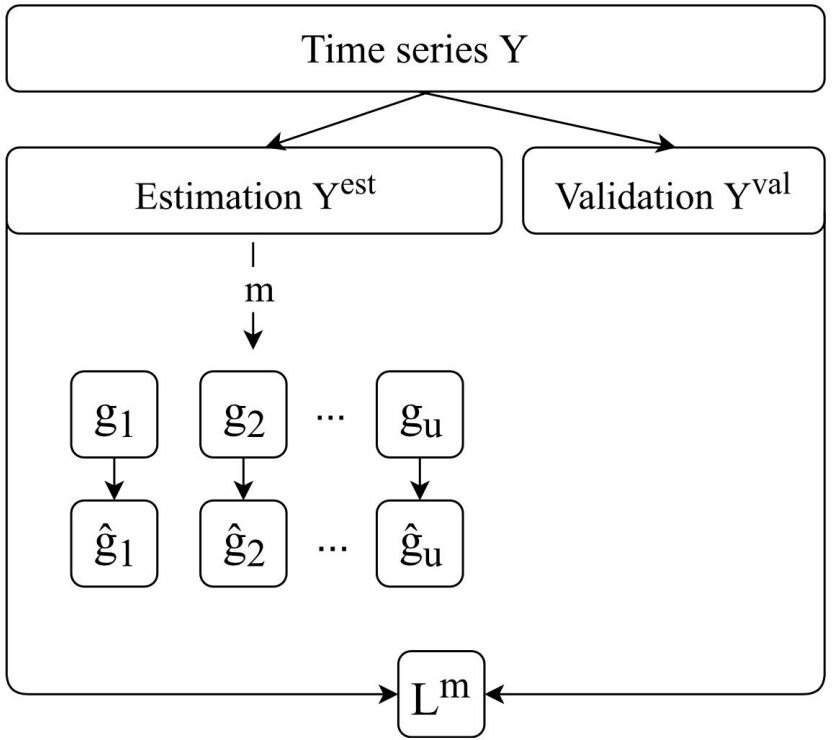
Quality of the estimation methods evaluated by:

Absolute Predictive Accuracy Error

$$APAE = |\hat{g}_i^m - L^m|$$

Predictive Accuracy Error

$$PAE = \hat{g}_i^m - L^m$$

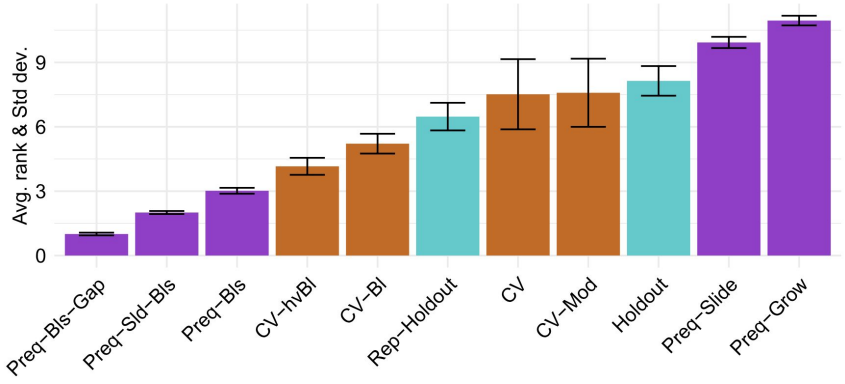


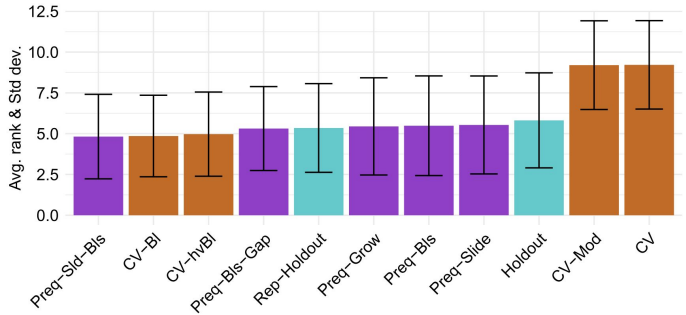
The Experiments

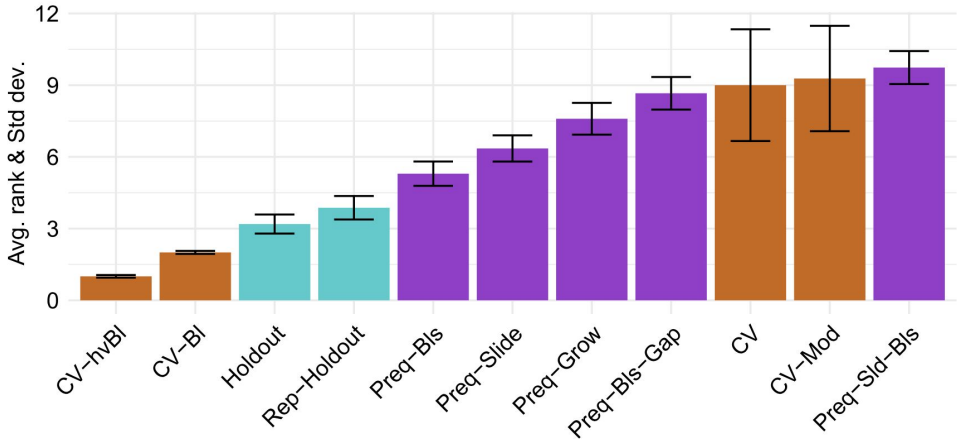
- Main estimation methods
  - **CV** : standard k-fold CV
  - **CV-Bl** : blocked K-fold CV
  - **CV-Mod** : modified K-fold CV
  - **CV-hvBl** : hv-blocked K-fold CV
  - **Holdout** : standard holdout (70% train)
  - **Rep-Holdout** : Out of Sample
  - **Preq-Bls** : prequential with growing window
  - **Preq-Sld-Bls** : prequential with sliding window
  - **Preq-Bls-Gap** : prequential with growing+gap
- Learning algorithms
  - **RBR:** Cubist (rule based model)
  - **RF** : random forest
  - **GLM** : generalized linear model

- Time series data
  - Synthetic time series
    - 3 time series from previous studies
  - Real world time series
    - 174 time series from different domains (finance, physics, meteorology, etc.)
      - 97 stationary and 77 non-stationary
- Prediction tasks from the time series
  - Forecasting *t+1* using an embed of the previous *p* values of the series
    - Embed size estimated using the False Nearest Neighbours method

Results with Synthetic Time Series

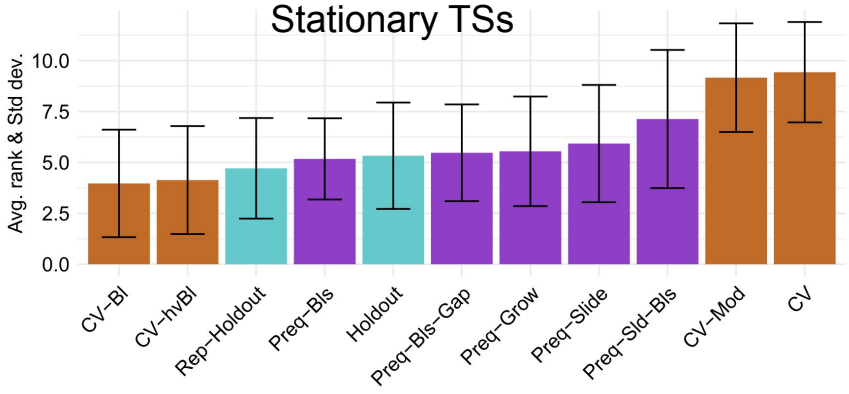




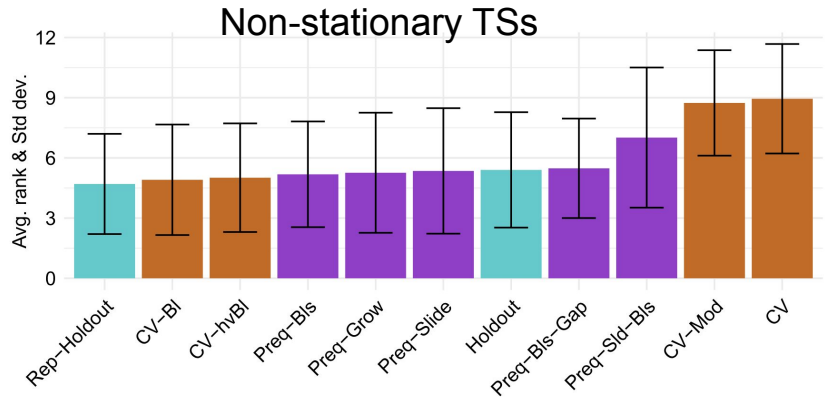


- Results from Bergmeir et. al (2018) confirmed
  - **CV approaches, particularly blocked, outperform OOS approaches**
- Our experiments show that **prequential methods also show good results**

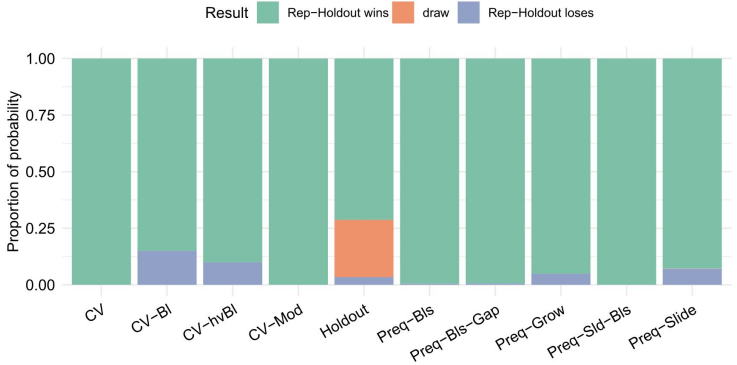
Results with Real World Time Series



- OSS approaches more competitive



OSS repeated holdout achieving the best results



Main Outcomes and Recommendations from our study

- For stationary time series **blocked K-fold CV** is the best option
- For non-stationary time series the best estimates are obtained with the OSS approach **repeated holdout**

- Other observations
  - Prequential applied in blocks (**Preq-Bls**) is the best of the prequential alternatives
  - Results were similar across the different learning algorithms

More details/information:
Cerqueira V., Torgo L. and Mozetic I. (2020): **Evaluating Time Series Forecasting Models: an empirical study on performance estimation methods**. In *Machine Learning*, 109 (11), 1997-2028.

DALHOUSIE UNIVERSITY
Inspiring Minds

Q2: Are there models/approaches that are clearly better than others?

Benchmarks of Time Series Forecasting Models

Cerqueira V., Torgo L. and Soares C. (2019): **Machine Learning vs Statistical Methods for Time Series Forecasting: size matters**. In *arXiv*, 1909.13316.

Motivation

- Machine learning (ML) models have witnessed noticeable success in many predictive tasks
- Forecasting literature is still dominated by statistical methods like ARIMA or exponential smoothing. Why?
- Several experimental studies (e.g. Makridakis et al., 2018) have shown that these methods outperform ML methods in forecasting univariate time series

- **Our Hypothesis**: these studies are biased regarding one particular characteristic of the data - sample size

S. Makridakis, E. Spiliotis, and V. Assimakopoulos (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889.

Our Experimental Study

Goal: empirical analysis of the impact of size on the relative performance of different forecasting methods

- Two categories : ML vs Statistical approaches (to match previous studies)
- 90 univariate time series from different domains

Machine Learning Approaches

RBR: Cubist rule learning system
RF: random forests
GP: Gaussian process regression
MARS: multivariate adaptive regression splines
Glm: generalized linear model regression with a Gaussian distribution and a different penalty mixing

Statistical Approaches

ARIMA: auto-regressive integrated moving average
Naive2: seasonal random walk forecasting
Theta: theta method, exponential smoothing with drift
ETS: exponential smoothing state-space model
Tbats: exponential smoothing state-space model with Box-Cox transformation, ARMA errors, trend and seasonal components

Some methodological details

- ML methods were applied to a dataset using an embed of the 10 previous values of the series
- Statistical models used the value automatically determined by R package *forecast* (Hyndman et al, 2014)
- In terms of time series pre-processing we follow Makridakis et al., 2018
  - Apply Box-Cox transformation to stabilize variance
  - For all models not copying directly with seasonality we used several tests to decompose and remove it
  - Use Cox-Stuart test to detect and then remove trend

How we tested the size impact?

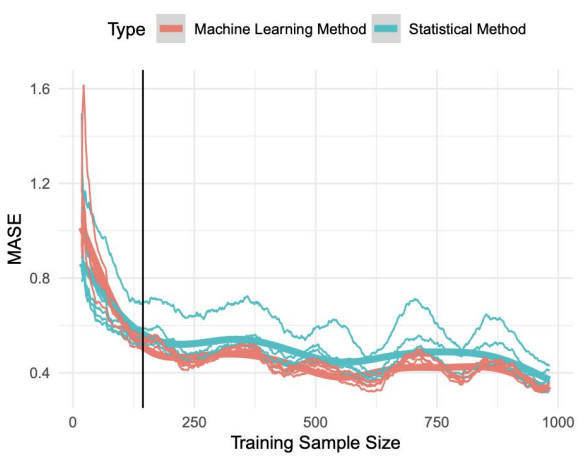
Following Makridakis et al., 2018 the first 18 observations are used to fit the models that are then used to obtain forecasts for the next 18 (multi step ahead) or for the next (one step ahead)

To test the influence of size we have used a prequential procedure with a growing training window to obtain a learning curve

Contrary to Makridakis et al., 2018, that considered a maximum of 144 points, we continued until 1000 observations to check for the impact of size

Models evaluate using mean absolute scaled error (MASE) and also by the average rank across all time series

Results for one step ahead forecasts



Results of Makridakis et al., 2018 are confirmed but as size grows the conclusions are different!

Results for multi step ahead forecasts

Size advantage of ML methods is not so evident

Summary of these experiments

Size of the time series seems to be an important factor on the performance of ML models

ML models tend to need larger data sets to achieve their best performance

For small time series we confirmed a clear advantage of more traditional approaches to time series forecasting

For larger time series the edge seems to be on the side of ML methods

> The difference may eventually be larger if recent models like Deep Neural Networks (that require very large data sets) are considered

Q3: Are ensembles a good answer to face the diversity of problems?

Ensemble Approaches to Time Series Forecasting

- Oliveira M., Torgo L., Costa V.S. (2015). Ensembles for Time Series Forecasting. In Proc. ACML'2015
- Cerqueira V., Torgo L., Pinto F., Soares C. (2019). Arbitrage of Forecasting Experts. In *Machine Learning*, 108(6), 913-944.

Motivation

Different and varying results of different approaches across different problems

Non-stationary time series are frequent in the real world, frequently showing rather different dynamic regimes

Ensembles are formed by sets of models that are used together to model a problem

> Known as efficient methods to fight complex problems by injecting diversity

Our Hypothesis: ensembles can help in coping with the diversity of regimes an non-stationarities often found in real world time series

Ensembles basics

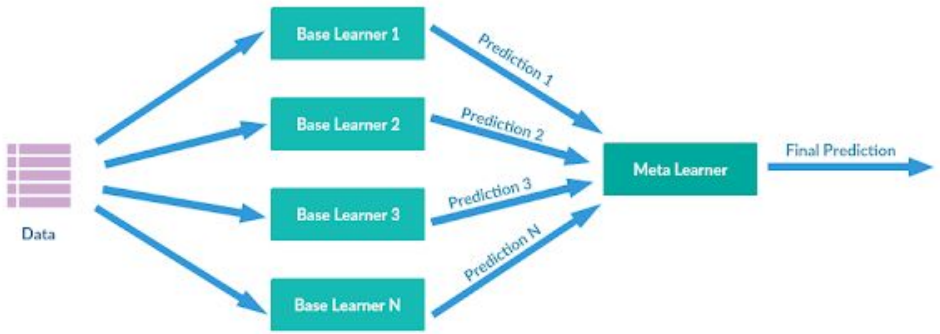
Several models solving the problem

Some key ideas:

1) How to generate the base learners?

Diversity known to be crucial

2) How to aggregate the predictions of the base models?



1) Generating diversity among base learners

Frequent methods

- Varying the training data
- Varying the variables

Ensembles for Time Series (Oliveira et al, 2015)

- Variant of bagging with diversity generated based on characteristics of time series
  - Different ways of handling diverse dynamic regimes
  - Different ways of handling non-stationarities

Oliveira M., Torgo L., Costa V.S. (2015). Ensembles for Time Series Forecasting. In Proc. ACML'2015

Main characteristics of the tried ensemble variants

	Embed size	Extra predictors
E	All models use k_{max}.	None.
E+S	All models use k_{max}.	All models use μ_Y and σ_Y^2 calculated with the respective embed.
DE	One third of the models use k_{max}, another third uses $k_{max}/2$, and the last third uses $k_{max}/4$.	None.
DE+S	One third of the models use k_{max}, another third uses $k_{max}/2$, and the last third uses $k_{max}/4$.	All models use μ_Y and σ_Y^2 calculated with the respective embed.
DE±S	One third of the models use k_{max}, another third uses $k_{max}/2$, and the last third uses $k_{max}/4$.	Half of the models using a certain embed size use μ_Y and σ_Y^2 calculated with the respective embed.

Experimental Validation

Hypothesis: The new forms of generating ensembles are able to outperform a normal ensemble (E - bagging) and are competitive with state of the art standard forecasting methods (ARIMA)

Comparison of different variants of our proposal on 14 real world time series

Performance measure with mean squared error over 10 repetitions of holdout (OSS)

Several variants of the number of models in the ensemble and the size of the embed (k_{max}) were tried

Results and discussion

Variants with more diversity (E+S) and (DE±S) have the best results

Results confirm the validity of the proposal

Open questions:

- Would diverse base models help?
- Is it possible to select the best configuration automatically?

M	k_{max}		E	E+S	DE	DE+S	DE±S	ARIMA
1020	20	mean	4.36	**2.00**	4.21	2.29	2.14	3.50
		sd	*0.84*	1.18	0.89	1.07	0.86	2.59
	30	mean	3.93	**2.29**	3.64	2.57	2.57	3.86
		sd	1.44	1.27	1.34	*1.16*	1.28	2.57
1500	20	mean	4.43	**2.00**	4.14	2.29	2.14	3.50
		sd	*0.65*	1.18	1.03	1.07	0.86	2.59
	30	mean	3.86	**2.36**	3.79	2.64	**2.36**	3.86
		sd	1.41	1.28	1.42	1.28	*1.01*	2.57

2) How to aggregate the predictions of the base models?

Base method consist of simply averaging the predictions

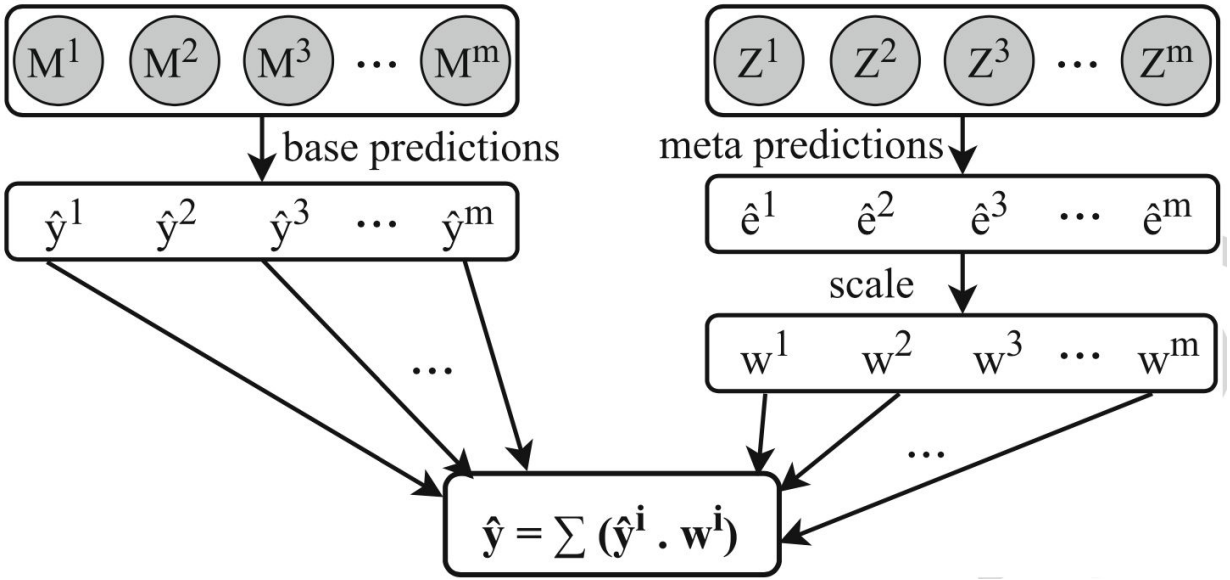
More sophisticated models track the recent performance of the models and use this information to dynamically weight the combination

Metalearning has been used (e.g. stacking) to learn the inter-dependencies among base learners and dynamically decide the form of better aggregating them.

Our Proposal: use metalearning to learn the individual capabilities of each model and make sure we have a diverse set of models with different specializations. Use the results of metalearning to decide which models to use in the aggregation at each time step

Cerqueira V., Torgo L., Pinto F., Soares C. (2019). Arbitrage of Forecasting Experts. In *Machine Learning*, 108(6), 913-944.

Arbitrated Dynamic Ensemble (ADE)



M's are trained to forecast the time series

Z's are trained to forecast the error of a certain M model

Diversity in ADE

- Implicit
  - Using different learning algorithms for the base learners (M's)

- Explicit
  - During the aggregation we take into account not only the predicted error but the correlation between the models calculated over a recent time window

How ADE makes predictions?

Predicting Y_{t+1}

Models M_i are asked for their predictions

Models Z_i are asked for their estimate of the error of each M_i for this test case

A committee is formed with the %k models with the lowest error in the past x cases

The weights of the models in the committee are determined by a scaling transformation of their error estimated by the respective meta model Z

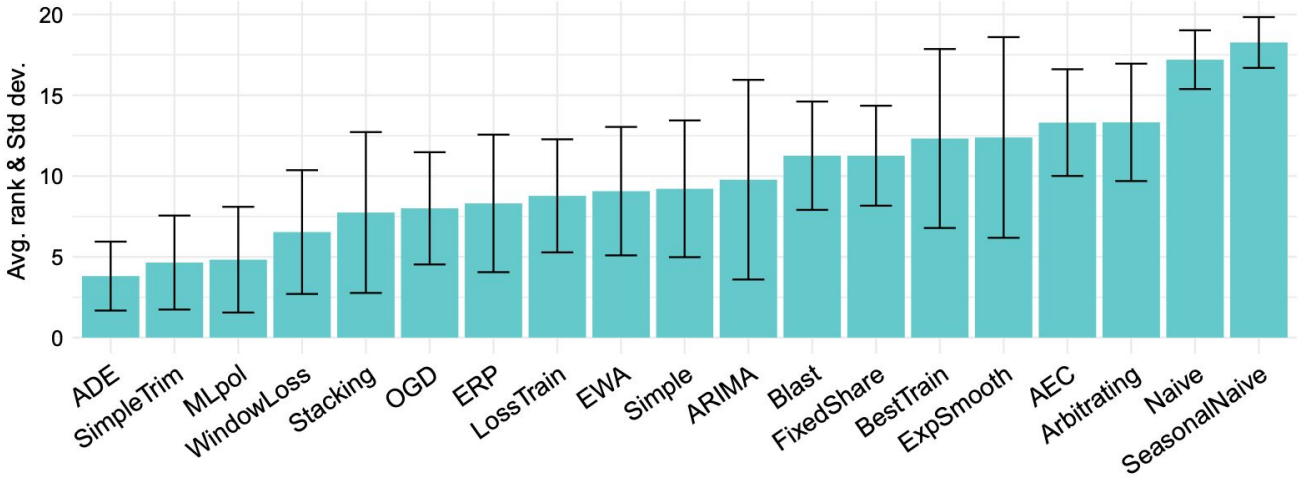
$$w_{t+1}^{j} = \frac{scale\left(-\hat{e}_{t+1}^{j}\right)}{\sum_{j \in \Omega M} scale\left(-\hat{e}_{t+1}^{j}\right)}$$

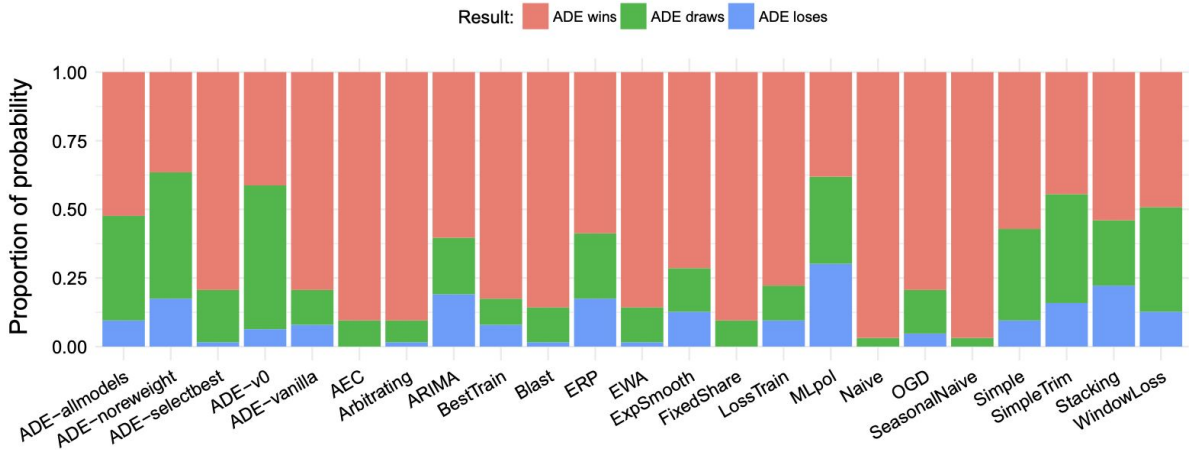
The final weights are adjusted to penalise models that are correlated to each other

The final prediction is give by $\hat{y}_{t+1} = \sum_{j \in \Omega M} \hat{y}_{t+1}^{j} \cdot w_{t+1}^{\prime j}$

Some results

ADE beats most methods with statistical significance





DALHOUSIE UNIVERSITY
Inspiring Minds

Concluding Remarks

Summary

- Model evaluation is a key step for comparing, selecting and tuning forecasting models
  - Time series data raises some challenges that should be taken into account
- Different models have different characteristics and pros&cons
  - Benchmarks are important but they should consider carefully the characteristics of models
  - ML models seem to be rather competitive with classical forecasting approaches when data abounds
- Ensembles are interesting approaches to cope with the diversity of challenges of real world time series data
  - ADE is a very competitie ensemble incorporating novel forms of diversity and aggregation methodology

Some open challenges

- Automatic forecasting
  - Improve ADE to be able to automatically adjust several of its components
    - Sets of models
    - Sets of features
    - Etc.
- Extension of some of these ideas to other data dependencies
  - Spatial
  - Spatiotemporal
  - Network data
- Explore other interesting and important problems related with time series
  - Hierarchical time series
  - Activity monitoring (from forecasting to actions)

Acknowledgements

Vitor Cerqueira

Mariana Oliveira

Chaires de recherche du Canada / Canada Research Chairs

Canada

Natural Sciences and Engineering Research Council of Canada / Conseil de recherches en sciences naturelles et en génie du Canada

Canada

FCT

Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

ltorgo@dal.ca

web.cs.dal.ca/~ltorgo