CBIC 2019
Belém . Pará
14th BRAZILIAN COMPUTATIONAL INTELLIGENCE MEETING
3-6 NOVEMBER 2019 . BELÉM . BRAZIL

Forecasting epidemiological time series based on decomposition and optimization approaches

PUCPR
GRUPO MARISTA

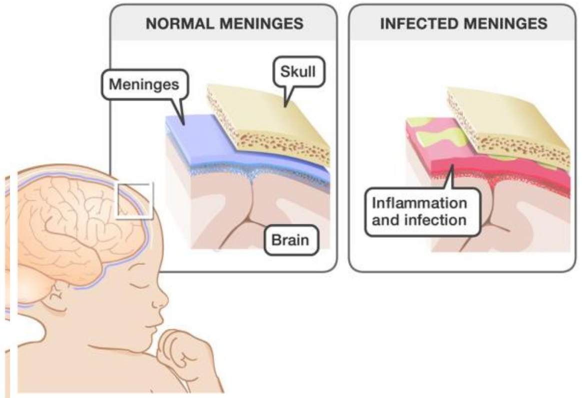
UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

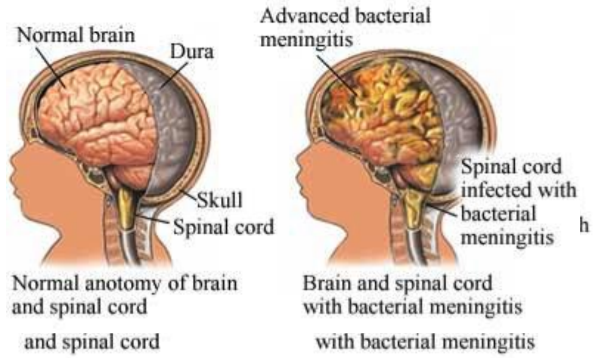
Matheus Henrique Dal Molin Ribeiro
Ramon Gomes da Silva
Naylene Fraccanabbia

Viviana Cocco Mariani
Leandro dos Santos Coelho

Presentation agenda

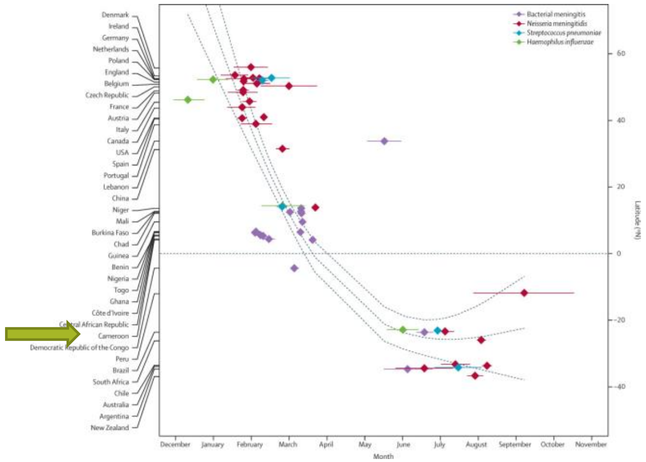
- Introduction
- Objectives
- Dataset
- Methodology
- Results
- Conclusion





Introduction

- **Meningitis** is an inflammation of the meninges, membranes that surround the brain and spinal cord;
- In **Brazil**, meningitis is considered an endemic disease.
- In **2018** have 15,000 cases of meningitis and 3,000 resulted in death;
- Forecasting meningitis cases allows to develop a **strategic planning**;
- Due to **chaotic behavior**, through the hybridization of ensemble, decomposition and optimization approaches is possible to build an efficient forecasting model.



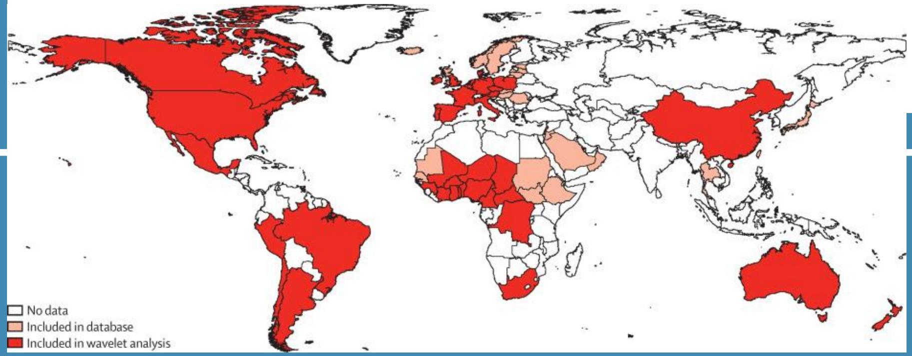
THE LANCET
Global Health

Volume 4, Issue 6, June 2016, Pages e370-e377

Seasonal dynamics of bacterial meningitis: a time-series analysis

Dr Juliette Paireau PhD [a, b], Angelica Chen [c], Helene Broutin PhD [d, e], Prof Bryan Grenfell PhD [f], Nicole E Basta PhD [f, g]

https://www.sciencedirect.com/science/article/pii/S2214109X1630064X#fig1



Objective

- This paper proposes a new hybrid approach that combines Ensemble Empirical Mode Decomposition, Quantile Random Forest based ensemble and Multi-Objective Optimization to forecast Meningitis Cases one-month ahead in
  - Para (PA)
  - Parana (PR)
  - Santa Catarina (SC)

Dataset

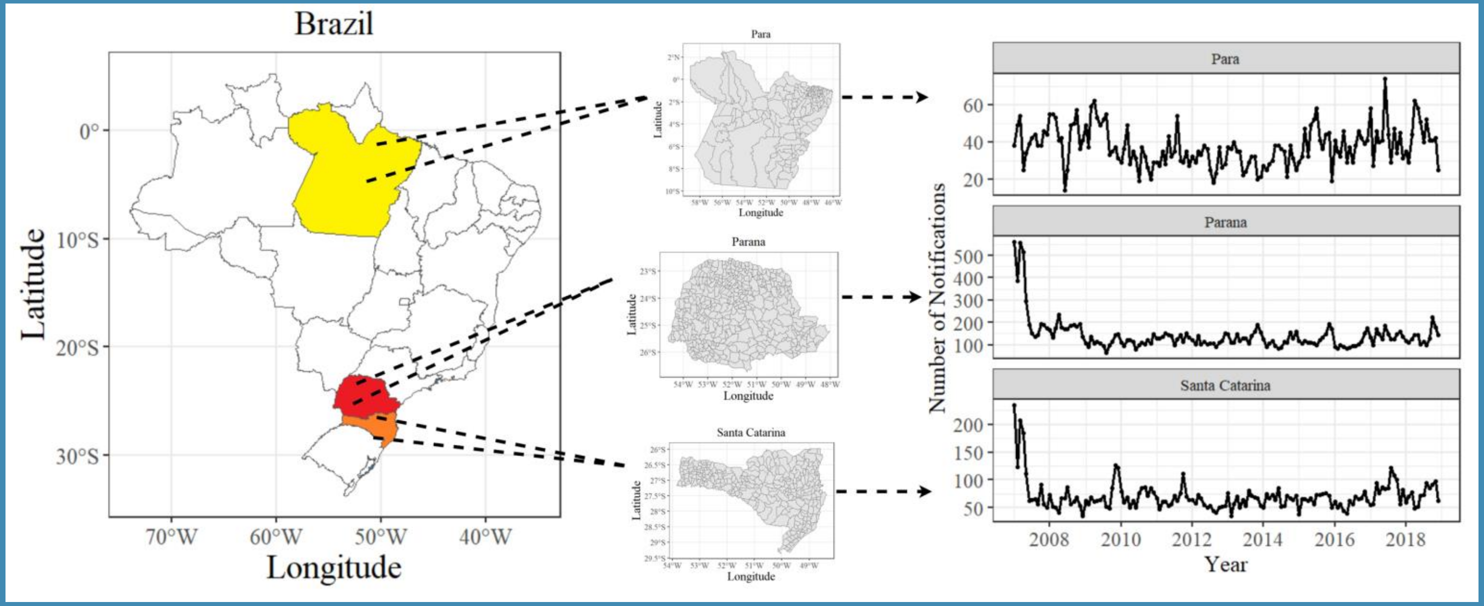
- The data set from DATASUS website;

  http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinannet/meningite/bases/meninbrnet.def

- Monthly meningitis cases from 2007 to 2018 recorded;

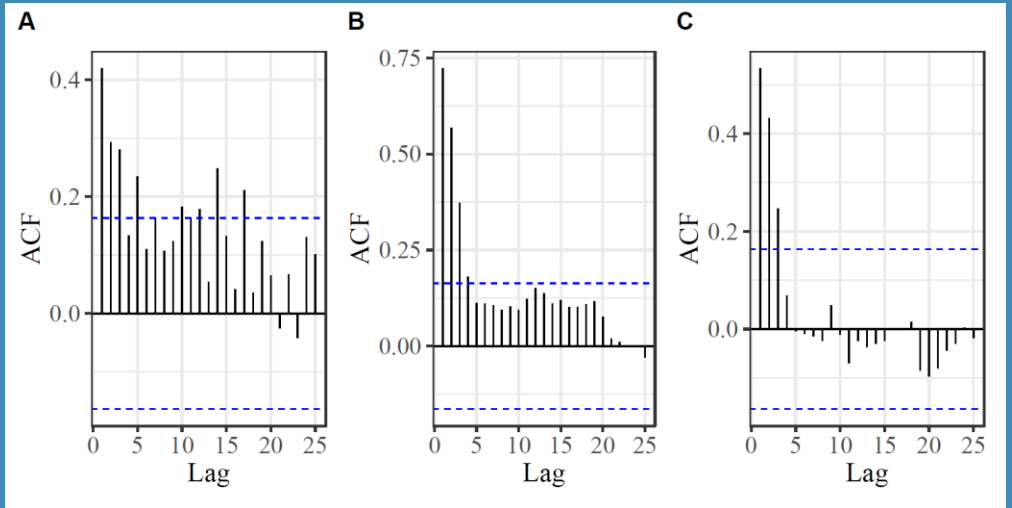
Dataset

12 years
(monthly measures)



STATISTICAL MEASURES FOR THE MENINGITIS NOTIFIED CASES NUMBER FOR ALL STATES.

State	Description	n	Min	Mean	Max	SD
Para	All data	144	14	37.54	74	10.66
	Training	132	14	37.02	74	10.51
	Test	12	25	43.33	62	11.01
Parana	All data	144	63	141.63	561	71.46
	Training	132	63	142.15	561	73.95
	Test	12	97	135.92	222	35.06
Santa Catarina	All data	144	35	70.68	235	26.78
	Training	132	35	70.39	235	27.56
	Test	12	49	73.92	97	16.30



Autocorrelations (ACF) for PA (A), PR (B) and SC (C) states

Augmented Dickey-Fuller test

PA, PR, SC series **are non-stationary**
(DF = -5.35 - -3.41, p-value > 0.05).

Seasonality in the data (Kruskal-Wallis test)

PA and SC series: there is **no** evidence of seasonality
PR state series : there is evidence of seasonality.

Methodology

Ensemble Empirical Mode Decomposition (EEMD)

Norden E. Huang et al. (**NASA**)
HHT (1998, cited by 19636), EEMD (2009, cited by 4857)

Hilbert-Huang Transform (HHT)

Empirical Mode Decomposition (EMD)

Hilbert Spectrum (HS)

Ensemble Empirical Mode Decomposition **(EEMD)**

Studies on its properties:
decomposing white noise

This approach consists of sifting an **ensemble of white noise-added signal (data)** and treats the mean as the final true result.

In this sense, it is performed the decomposition of time series signal with objective to **extract the coexisting oscillatory functions, named IMF (intrinsic mode functions) and residual component, from original data.**

N. E. Huang et al., "The Empirical Mode Decomposition Method and the Hilbert Spectrum for Non-stationary Time Series Analysis," *Proc. Roy. Soc. London*, 454A, pp. 903-995, 1998.
Z. Wu and N. E. Huang, "Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method," *Advances in Adaptive Data Analysis*, Volume 1, No. 1, pp. 1-41, 2009.

Quantile Random Forest (QRF)

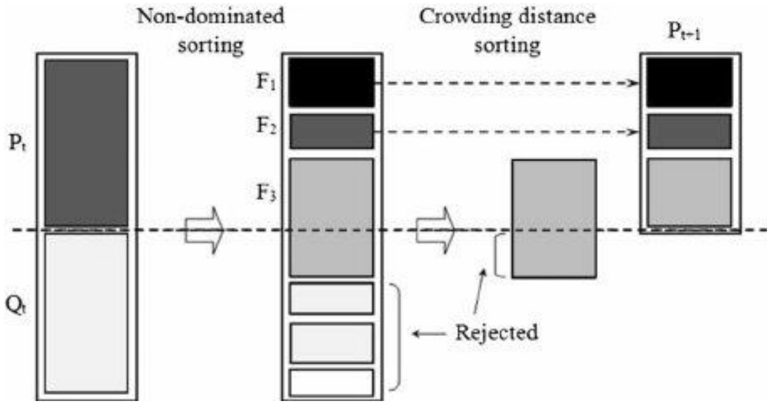
It provides information about the full conditional distribution of the response variable, not only about the conditional mean.

Quantile random forests give a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables.

Non-Dominated Sorting Genetic Algorithm (NSGA-II)

NSGA-II procedure has three features:

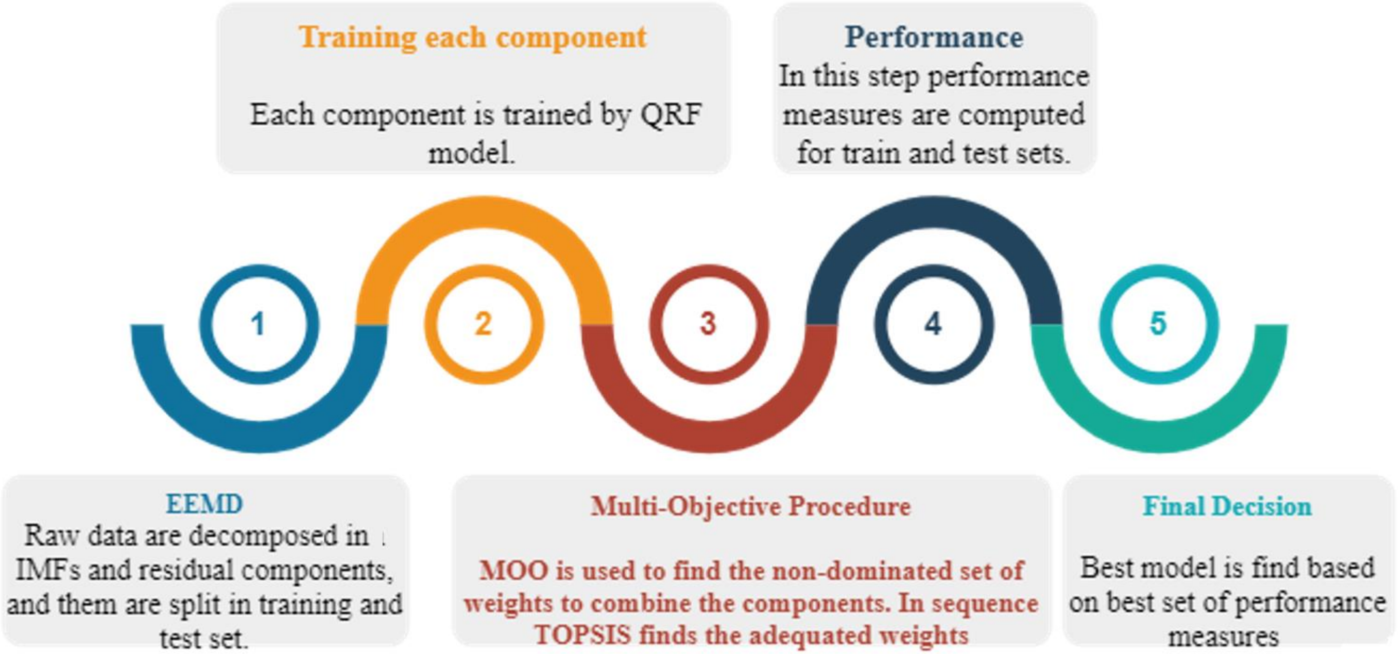
- It uses an elitist principle
- It emphasizes non-dominated solutions.
- It uses an explicit diversity preserving mechanism



K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, 2002.

https://ieeexplore.ieee.org/document/996017

Roadmap of hybrid framework

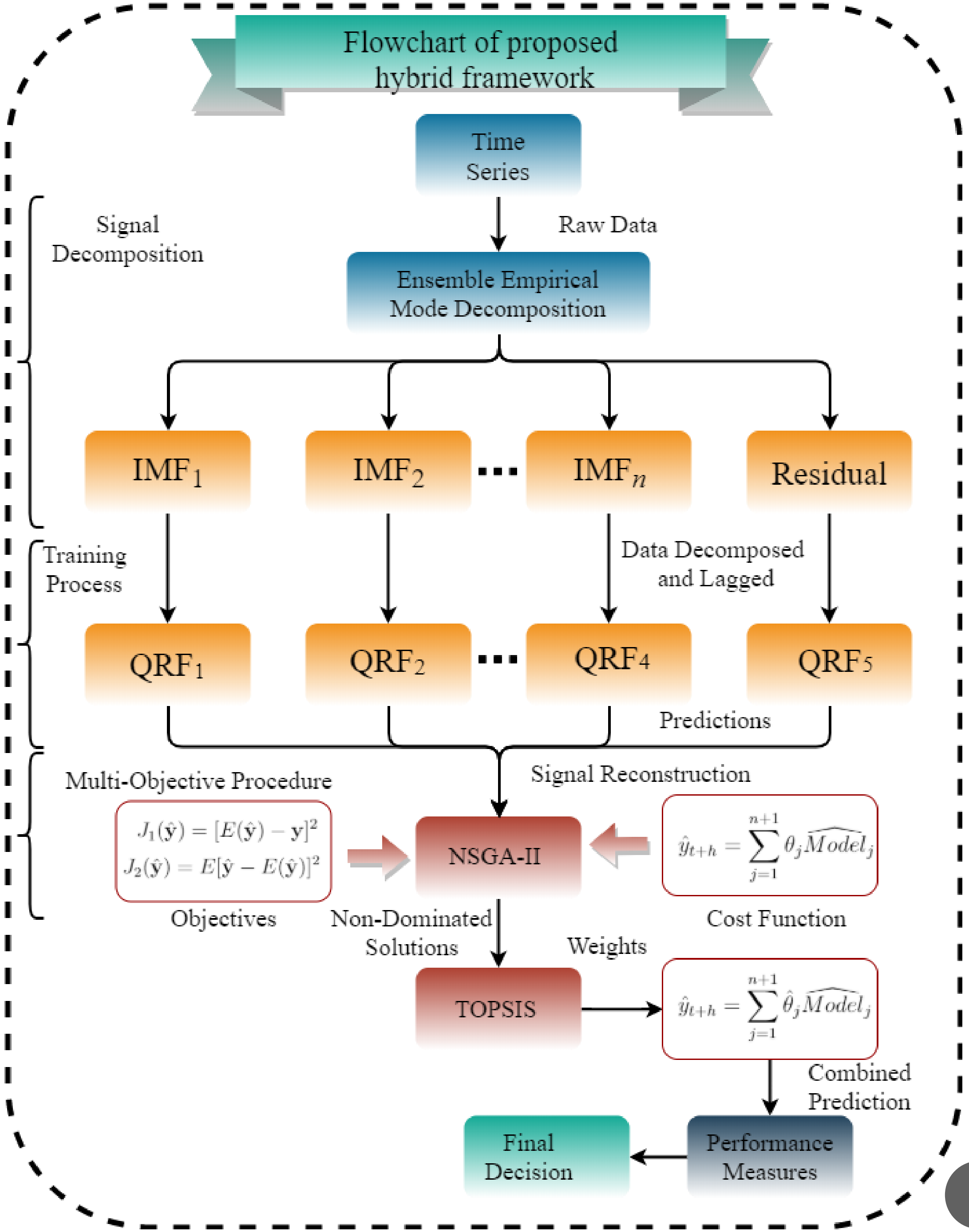


IMFs: intrinsic mode functions

TOPSIS: Technique for Order Preference by Similarity to Ideal Solution

Methodology

$$J_1(\hat{\mathbf{y}}) = \overbrace{[E(\hat{\mathbf{y}}) - \mathbf{y}]^2}^{Bias-Error} \quad \text{and} \quad J_2(\hat{\mathbf{y}}) = \overbrace{E[\hat{\mathbf{y}} - E(\hat{\mathbf{y}})]^2}^{Variance-Stability}$$



Methodology

Performance measures

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{\sum_{i=1}^{n}[y_i - \overline{y}_i]^2},$$

$$sMAPE = 100 \times \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{(|y_i| + |\hat{y}_i|/2)} \right|,$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

Statistical tests

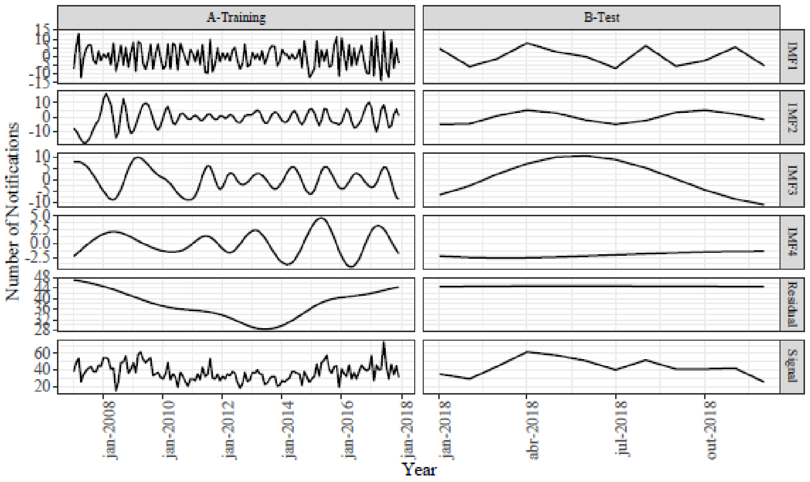
- **Friedman** test and **Nemenyi** test post-hoc

critical difference

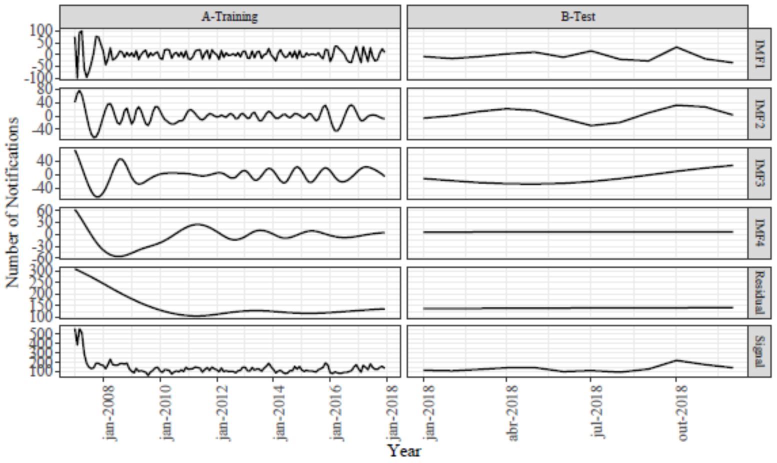
$$FD = \frac{12n}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right],$$

$$CD = \frac{q_{\infty,k,\alpha}}{\sqrt{2}}\sqrt{\frac{k(k+1)}{6}},$$

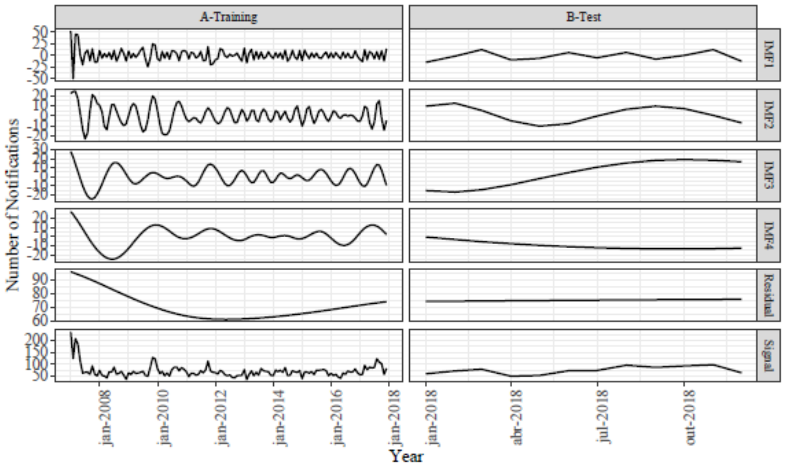
Results



(a) Decomposed series for PA state.



(b) Decomposed series for PR state.



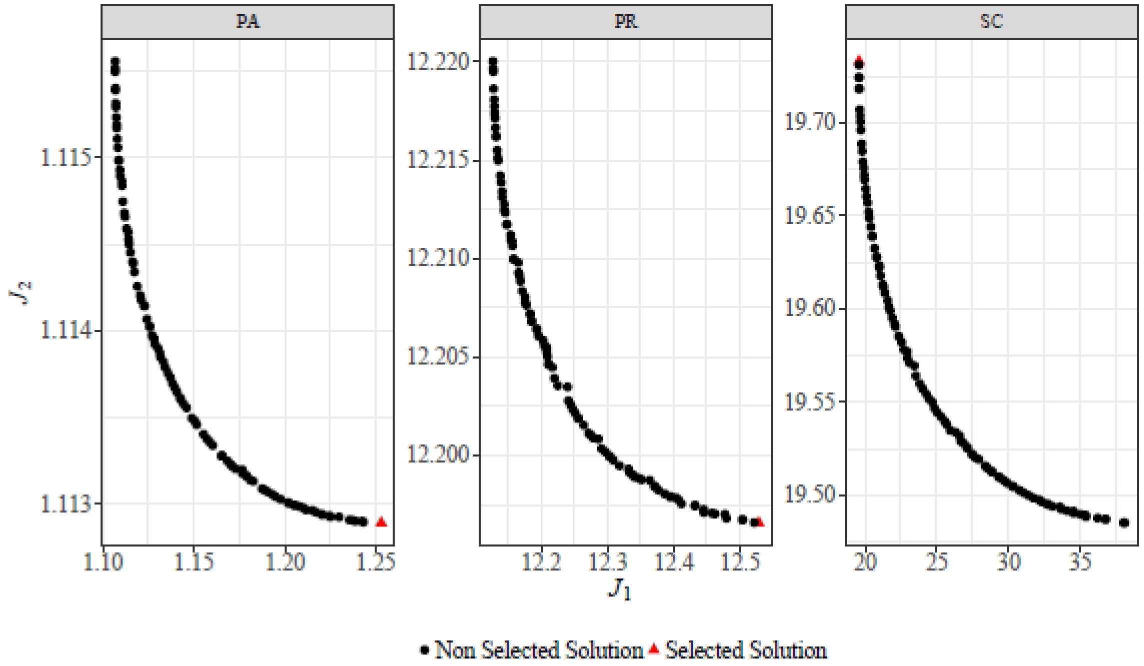
(c) Decomposed series for SC state.

Results

CONTROL HYPERPARAMETERS (*mtry*) EMPLOYED IN EACH MODEL.

Component/Model	PA	PR	SC
IMF_1	2	4	3
IMF_2	4	4	4
IMF_3	4	2	3
IMF_4	4	2	4
Residual	4	3	4
QRF	4	4	2

Pareto front



WEIGHTS ADOPTED FOR EACH IMF AND RESIDUAL IN THE SIGNAL RECONSTRUCTION.

State	θ_1	θ_2	θ_3	θ_4	θ_5
PA	1.1020	1.0409	1.0151	1.0126	1.0146
PR	1.0904	1.0214	1.0143	0.9798	0.9938
SC	1.0691	1.0162	0.9118	1.1994	0.9892

Results

STANDARD DEVIATION FOR MODELS SE FOR TEST SET.

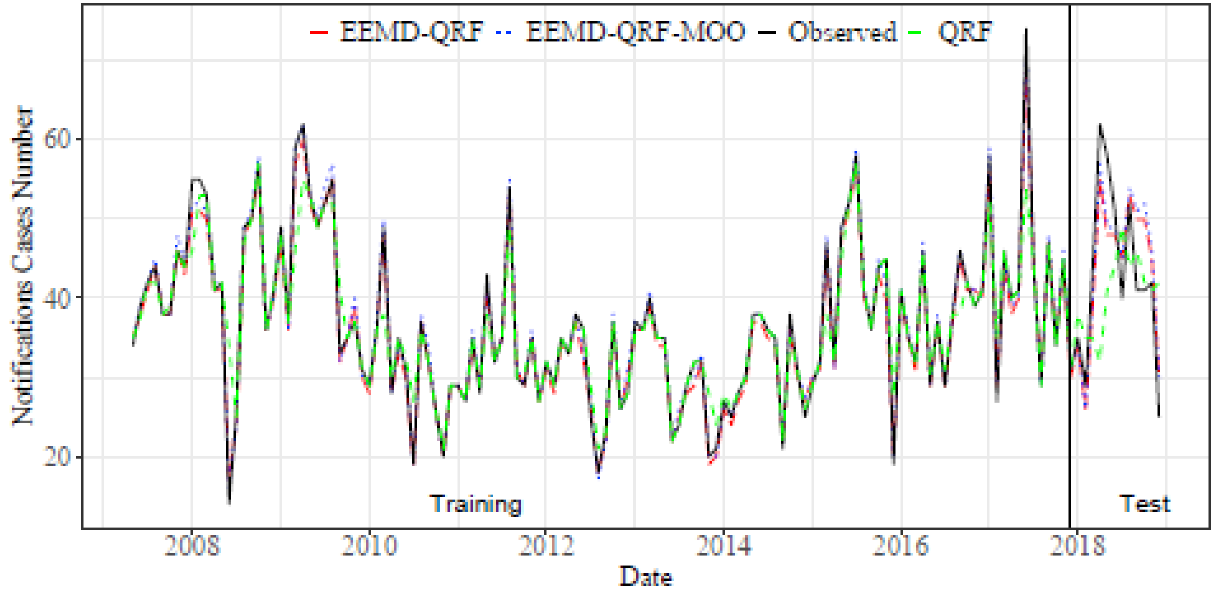
State	EEMD-QRF	EEMD-QRF-MOO	QRF
PA	1.0500	**0.9902**	6.0696
PR	2.1544	**2.1542**	7.2229
SC	1.6042	**1.4264**	4.3714

Performance measures

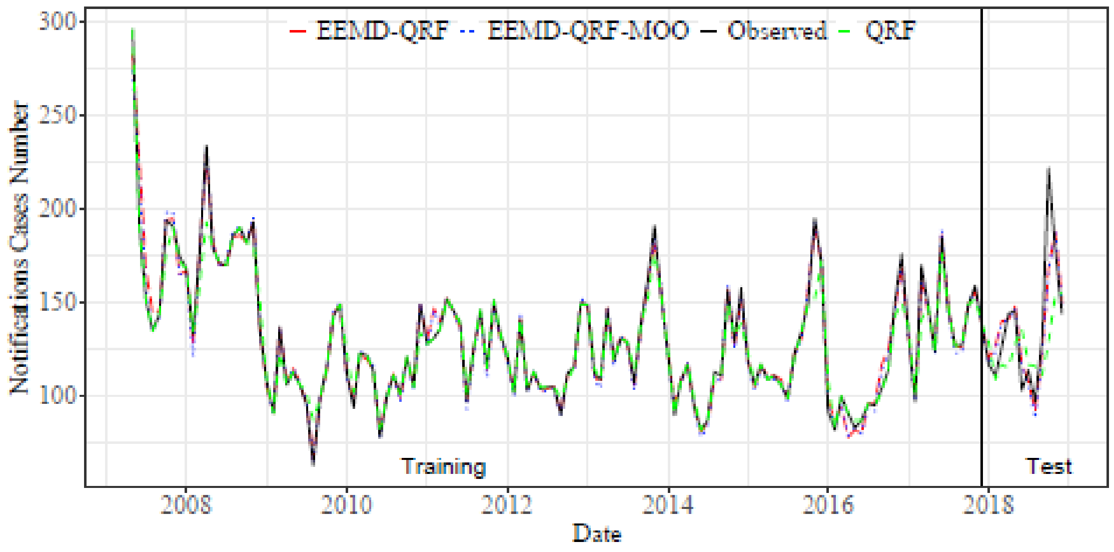
PERFORMANCE MEASURES ADOPTED IN THE MODELS EVALUATION.

State	Measure	EEMD-QRF	EEMD-QRF-MOO	QRF
Training set				
PA	RMSE	1.3550	**1.0825**	3.6596
	R^2	0.9882	**0.9893**	0.8919
	sMAPE	2.57%	**2.08%**	4.19%
PR	RMSE	4.0243	**3.6518**	7.6555
	R^2	0.9867	**0.9901**	0.8552
	sMAPE	**1.93%**	2.03%	2.47%
SC	RMSE	4.9434	**4.4362**	7.0872
	R^2	0.9148	**0.9317**	0.8552
	sMAPE	6.50%	**5.51%**	4.04%
Test set				
PA	RMSE	6.0690	**6.0346**	12.7777
	R^2	0.6693	**0.6751**	0.0369
	sMAPE	11.73%	**11.57%**	23.54%
PR	RMSE	17.5855	**17.3518**	31.8764
	R^2	0.7323	**0.7378**	0.1869
	sMAPE	8.24%	**7.97%**	15.28%
SC	RMSE	**8.1240**	9.6220	15.4407
	R^2	0.8472	**0.8628**	0.1355
	sMAPE	**9.82%**	12.03%	19.73%

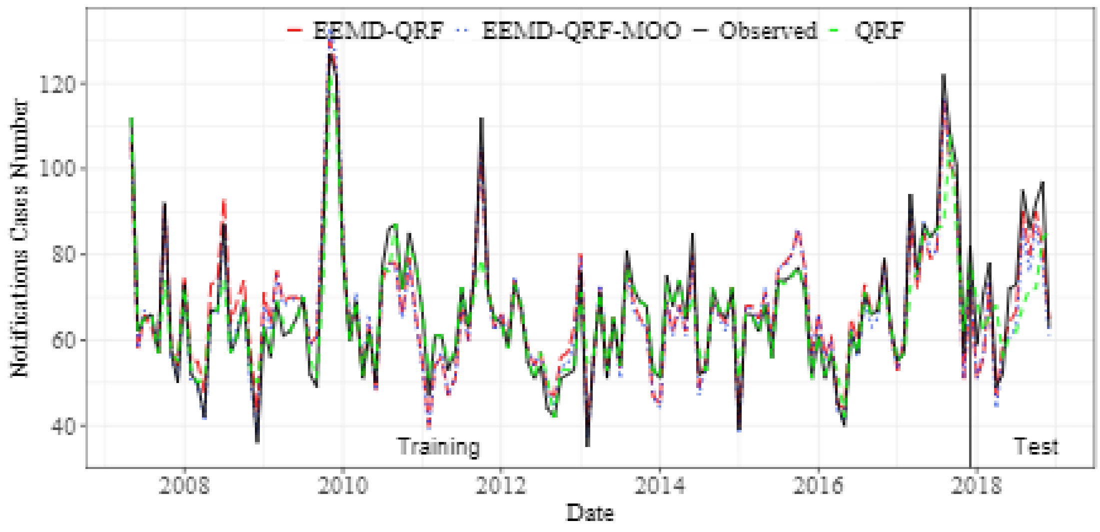
Results
observed x predicted values



(a) Observed and Predictions for PA state.

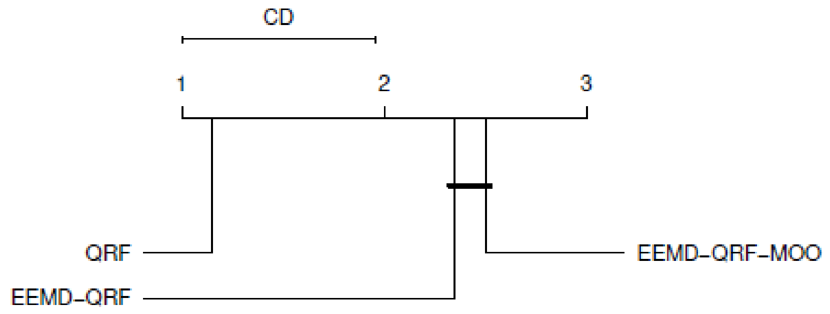


(b) Observed and Predictions for PR state.

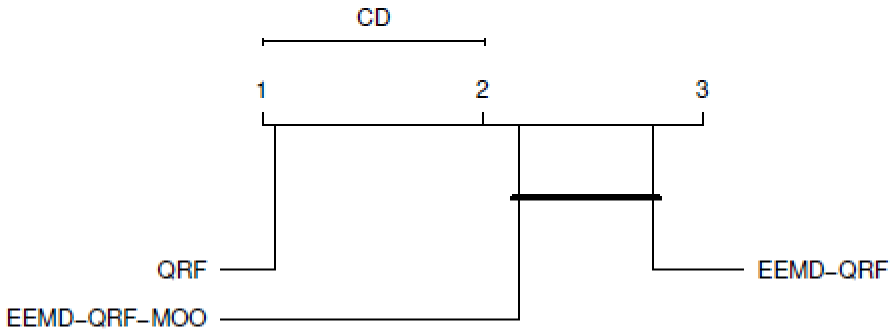


(c) Observed and Predictions for SC state

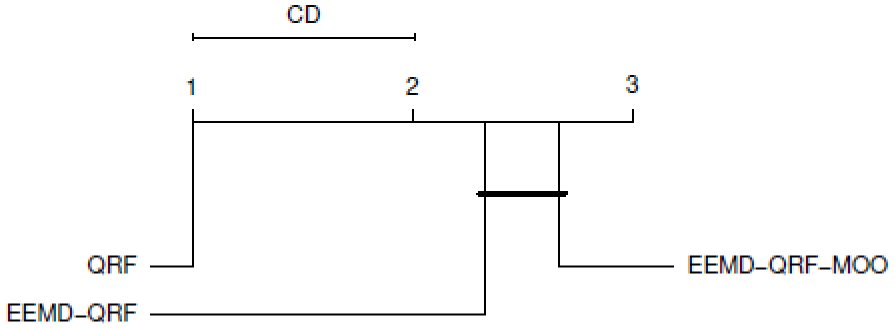
Results



(a) CD diagram for PA state.



(c) CD diagram for SC state



(b) CD diagram for PR state.

Friedmann Test = $\chi_2^2 = 12.51 - 14.97, p - value < 0.05$

CD = 1.1088, 1.1722 and 1.1722 for PA, PR and SC.

PR and PA states EEMD-QRF-MOO shows lower errors than EEMD-QRF but **no** statisical significant.

Conclusion

- A hybrid framework combining EEMD, QRF and MOO was proposed;
- EEMD was employed to decompose the series, QRF to forecast each obtained component and MOO to find weights for these components;
- One-month ahead forecasting the meningitis cases in PA, PR and SC states was studied;
- EEMD-QRF-MOO is competitive with 2 cases better than EEMD-QRF and all cases better than QRF model.
- Decomposition and optimization allow to enhance models performance;
- For future works is intend
  - Adopt different combinations of models for EEMD components.
  - Increasing the number of steps ahead to forecasting.

Acknowledgments

CAPES

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

FUNDAÇÃO ARAUCÁRIA
Apoio ao Desenvolvimento Científico e Tecnológico do Paraná

SCIENTIA, VITA ET FIDES
PUCPR

CBIC 2019
Belém . Pará
14th BRAZILIAN COMPUTATIONAL INTELLIGENCE MEETING
3-6 NOVEMBER 2019 . BELÉM . BRAZIL

Thank you

Comments: HHT, EEMD

	Fourier	STFT	Wavelet	HHT
Basis	A priori	A priori	A priori	Adaptive
Frequency	Convolution: global, uncertainty	Convolution: regional, uncertainty	Convolution: regional, uncertainty	Differentiation: local, certainty
Presentation	Energy-frequency	Energy-time-frequency	Energy-time-frequency	Energy-time-frequency
Nonlinear	No	No	No	Yes
Nonstationary	No	Yes	Yes	Yes
Feature Extraction	No	Yes	Discrete: No Continuous: Yes	Yes
Theoretical Base	Theory complete	Theory complete	Theory complete	Empirical