Forecasting The Cumulative Cases of COVID-19 In Four Large Brazilian Cities Using Machine Learning Approaches

Ramon Gomes da Silva[1]
Matheus Henrique Dal Molin Ribeiro[1,2]
José Henrique Kleinübing Larcher[3]
Viviana Cocco Mariani[3, 4]
Leandro dos Santos Coelho[1, 4]

gomes.ramon@pucpr.edu.br, mribeiro@utfpr.edu.br, jose.kleinubing@pucpr.edu.br, viviana.mariani@pucpr.br, leandro.coelho@pucpr.br

[1] Industrial and Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR), 1155, Rua Imaculada Conceicao. Curitiba, PR, Brazil.

[2] Department of Mathematics, Federal Technological University of Parana (UTFPR). Via do Conhecimento, KM 01 - Fraron, Pato Branco, PR, Brazil.

[3] Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR). 1155, Rua Imaculada Conceicao, Curitiba, PR, Brazil.

[4] Department of Electrical Engineering, Federal University of Parana (UFPR). 100, Avenida Coronel Francisco Heraclito dos Santos, Curitiba, PR, Brazil.

Overview

Introduction

REUTERS

World Business Markets Breakingviews Video More

HEALTHCARE & PHARMACEUTICALS OCTOBER 26, 2020 / 6:59 PM / UPDATED A DAY AGO

Brazil sees 15,726 new coronavirus cases, 263 new deaths: health minister

By Reuters Staff

1 MIN READ

BBC Sign in | Home | News | Sport | Reel | Worklife | Travel

NEWS

Home | US Election | Coronavirus | Video | World | US & Canada | UK | Business | Tech | Science | Stories

World | Africa | Asia | Australia | Europe | Latin America | Middle East

Covid: Brazil's coronavirus cases pass five million

8 October

Introduction

MENU

FOLHA DE S.PAULO

ENGLISH VERSION

LOGIN SEARCH

science & health

Country Exceeds Covid-19 157,000 Deaths

Daily average of deaths in the last seven days, however, was 468, according to the consortium

abc NEWS

Brazil reaches 150,000 deaths from COVID-19 milestone

Brazil's count of COVID-19 deaths has surpassed 150,000, despite signs the pandemic is slowly retreating in Latin America's largest nation

By **MARCELO DE SOUSA and TATIANA POLASTRI**

10 October 2020, 22:29 • 4 min read

Introduction

New coronavirus identified, currently named Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2), causing the coronavirus disease 2019 (COVID-19).

Brazil has more than 5 million cases and 150 thousand deaths.

Problems:

- Overcrowding of hospital beds;
- Lack of medical supplies to treat patients affected more severely;

Short-term forecast allows to:

- Reallocate patients;
- Reapportion resources and staff;
- Alert residents of a determined region in case of a spike;

Introduction

Models:

- Autoregressive Integrated Moving Average (ARIMA)
- *k*-Nearest Neighbor (kNN)
- Support Vector Regression (SVR)
- Cubist Regression (CUBIST)
- Stacked Generalization (STACK)

Objective

Present a **time series forecasting framework** based on the use of **machine learning** to predict cumulative cases of COVID-19 achieving accurate predictions, being possibly useful for decision making authorities.

Use forecasting models (CUBIST, kNN, and SVR) cooperatively in the form of a **Stacked Generalization** to bring the accuracy level up.

Dataset

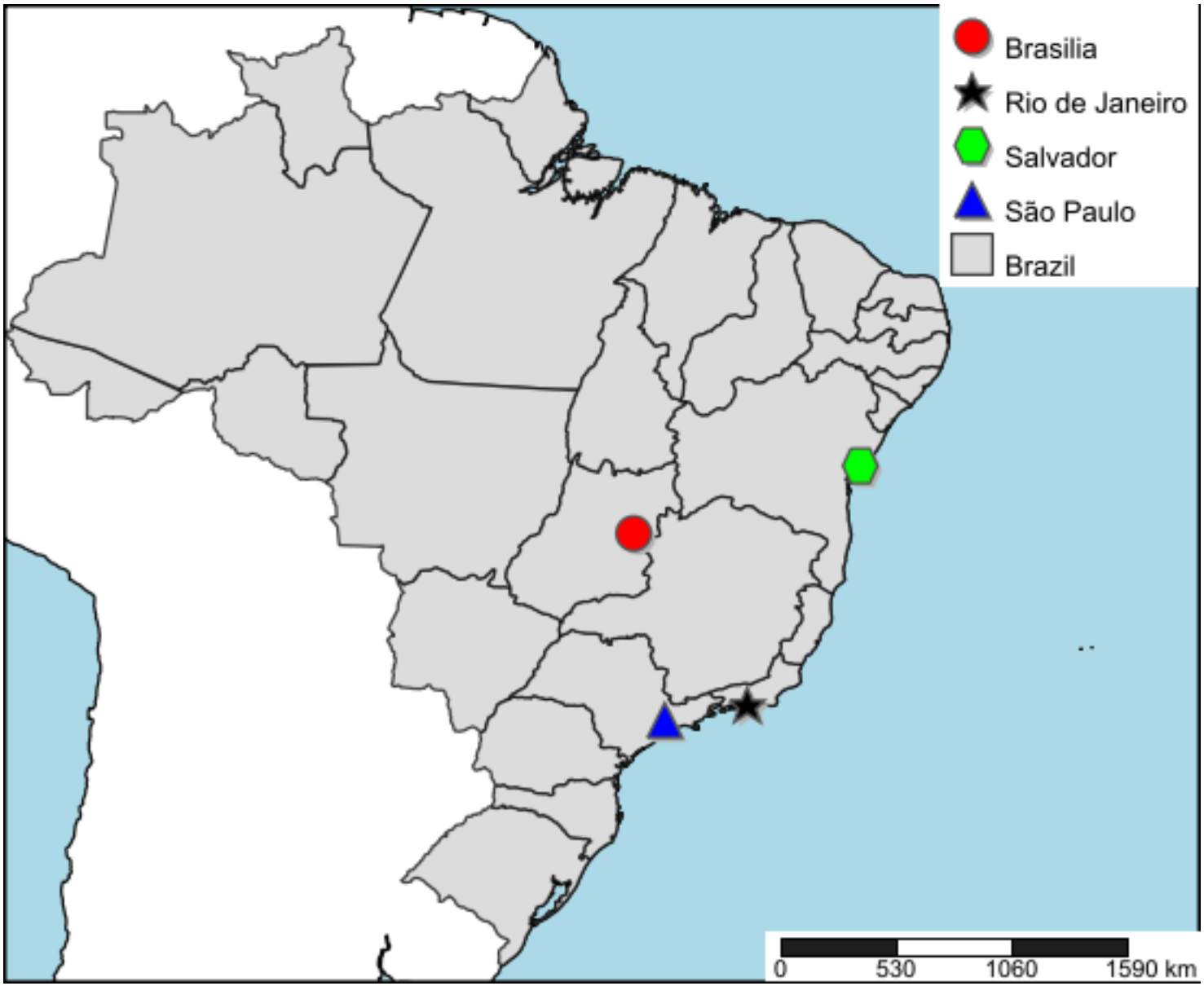
Aggregated data from Brazilian State Health Offices from March 13th until August 14th, 2020.

The data is gathered in a collaborative project with various contributors and made available with an API or direct download.

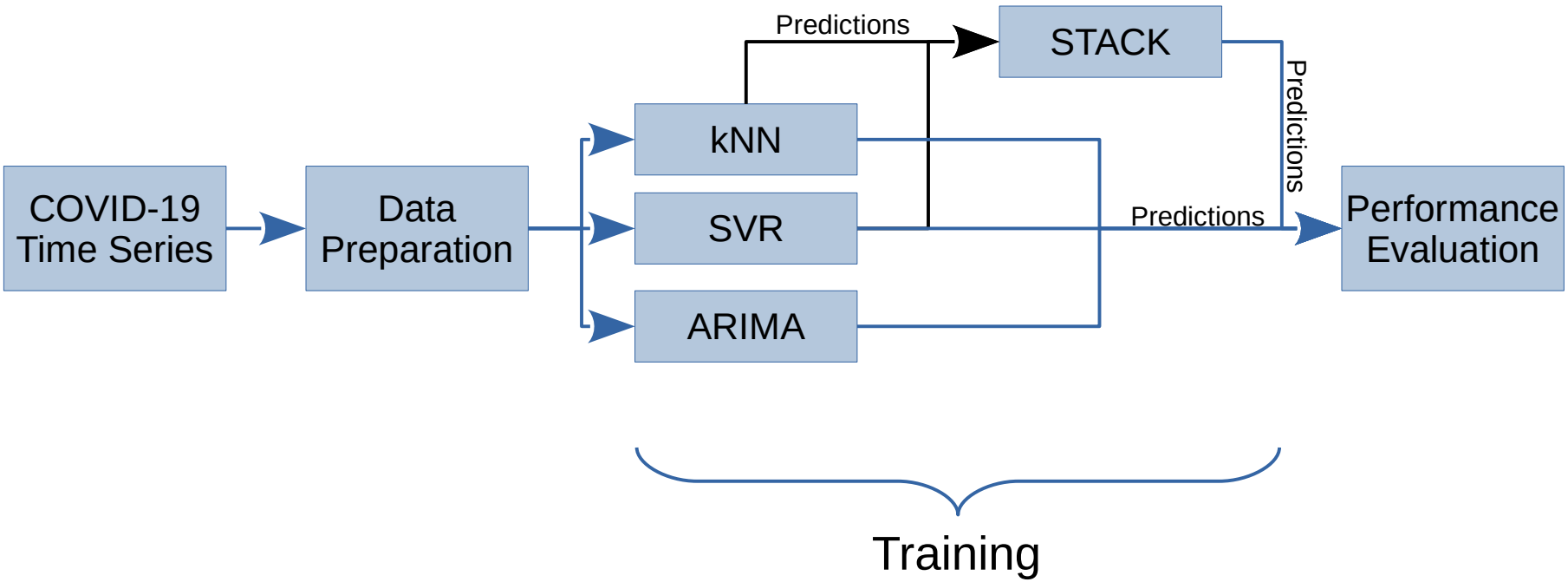
- Brasil.IO

Cities: Brasília (DF); Rio de Janeiro (RJ); Salvador (BA); and São Paulo (SP).

Dataset


Brasilia
Rio de Janeiro
Salvador
São Paulo
Brazil
0
530
1060
1590 km

Proposed forecasting framework



Proposed forecasting framework

Step 1

- Train-test split
- Normalization
- Past 5 days as features
- Recursive one-day-ahead model

Proposed forecasting framework

Step 2

- Base-learners (kNN and SVR) are trained and its predictions used as input for CUBIST meta-learner.
- Leave-one-out cross-validation is used for training.
- Machine learning models use caret R package.
- ARIMA uses auto.arima function from forecast R package.
- All computations are made in R software environment.

Proposed forecasting framework

Step 3

- Out-of-sample metrics are computed:
  - Improved percentage index (IP)
  - Mean absolute error (MAE)
  - Root-mean-square error (RMSE)
  - Symmetric mean absolute percentage error (sMAPE)

Findings - Brasília

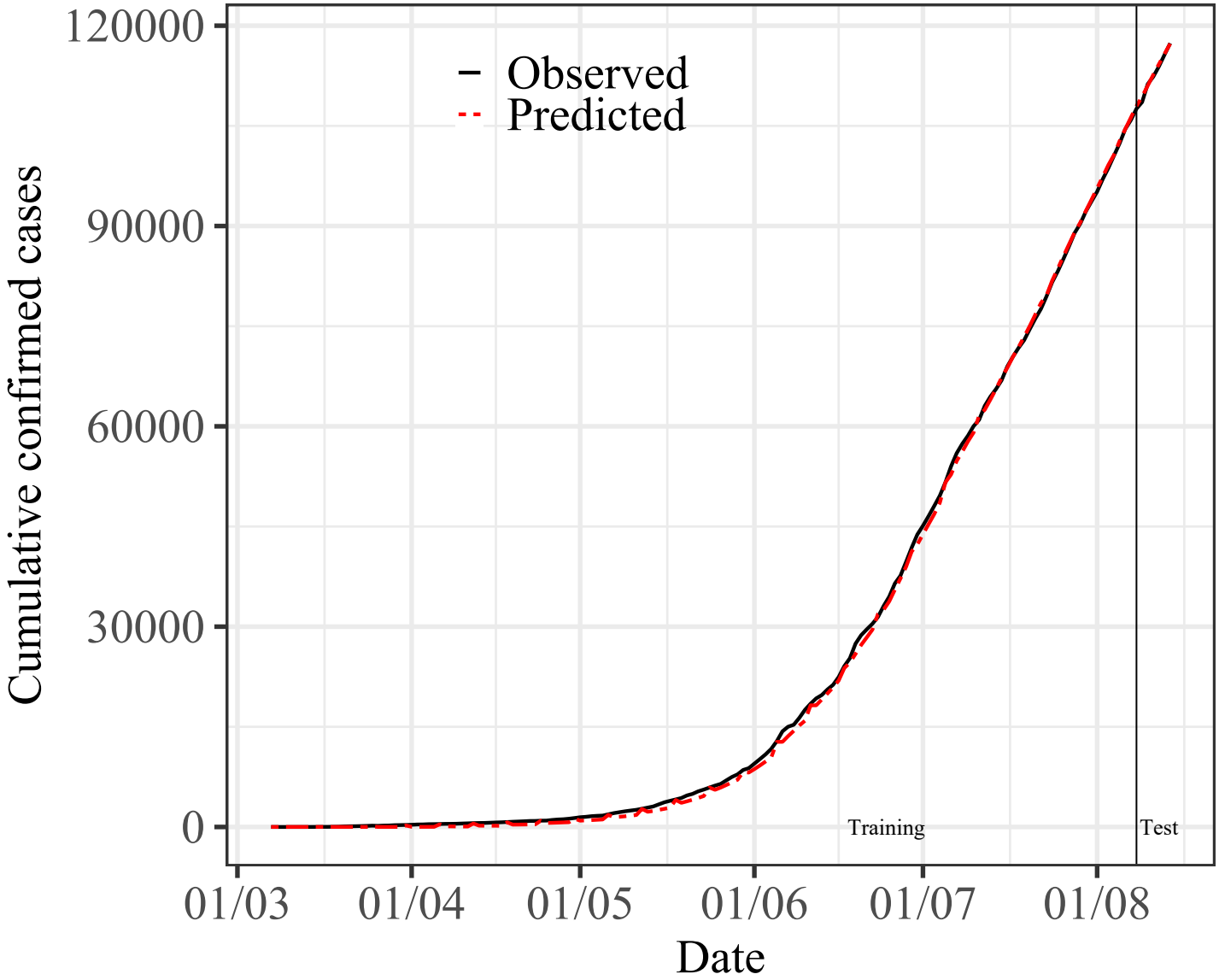
STACK approach could be considered to forecast COVID-19 cases, once the model outperformed others in all performance criteria.

Regarding MAE the performance improved in the ranges of 38.59% – 96.58%.

Ranking

- STACK
- ARIMA
- SVR
- kNN

Findings - Brasília



Findings - Rio de Janeiro

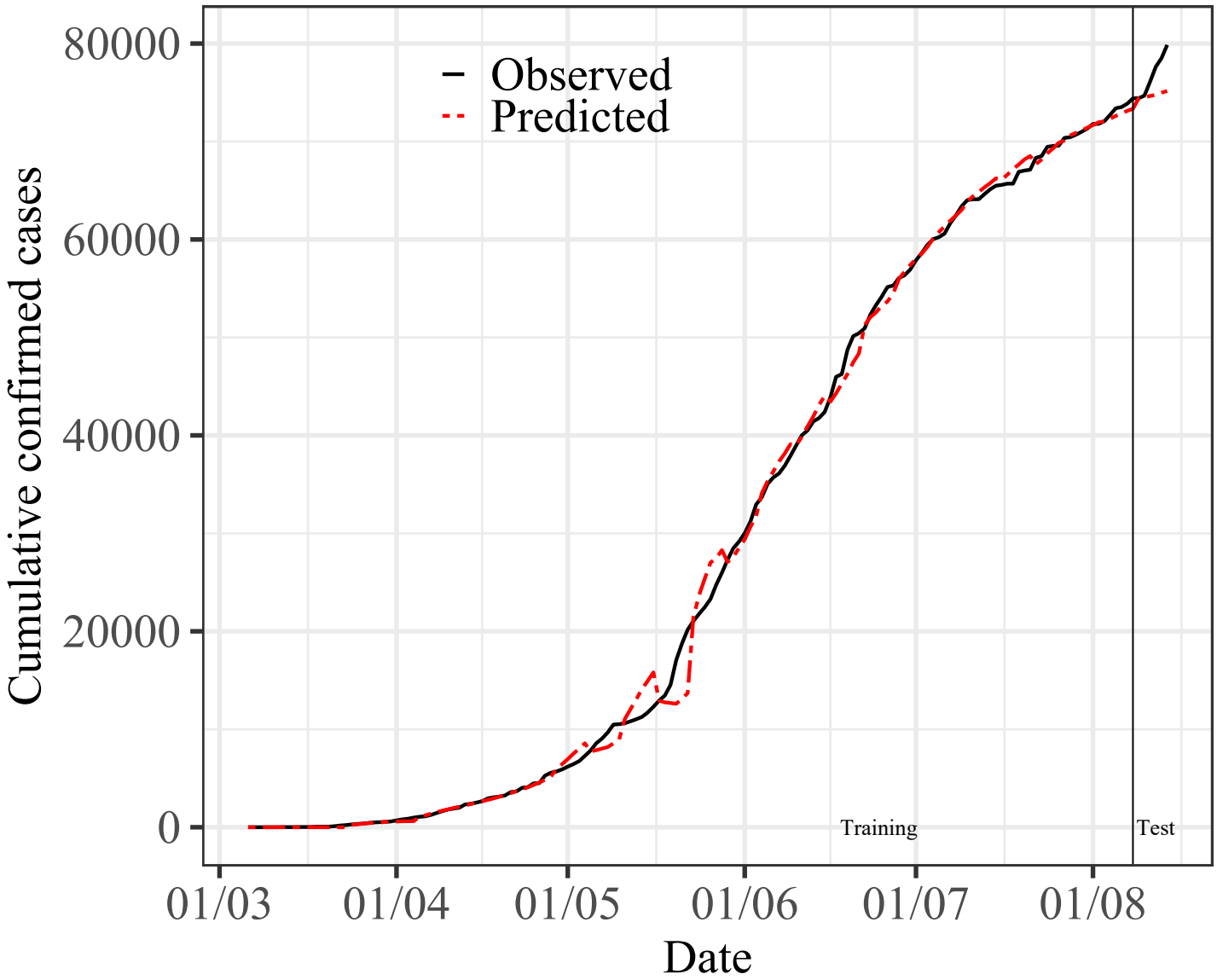
For Rio de Janeiro's cases, the SVR model achieved better accuracy than other models

The improvement in MAE ranged in 9.14% – 61.98%.

Ranking

- SVR
- ARIMA
- STACK
- kNN

Findings - Rio de Janeiro



Findings - Salvador

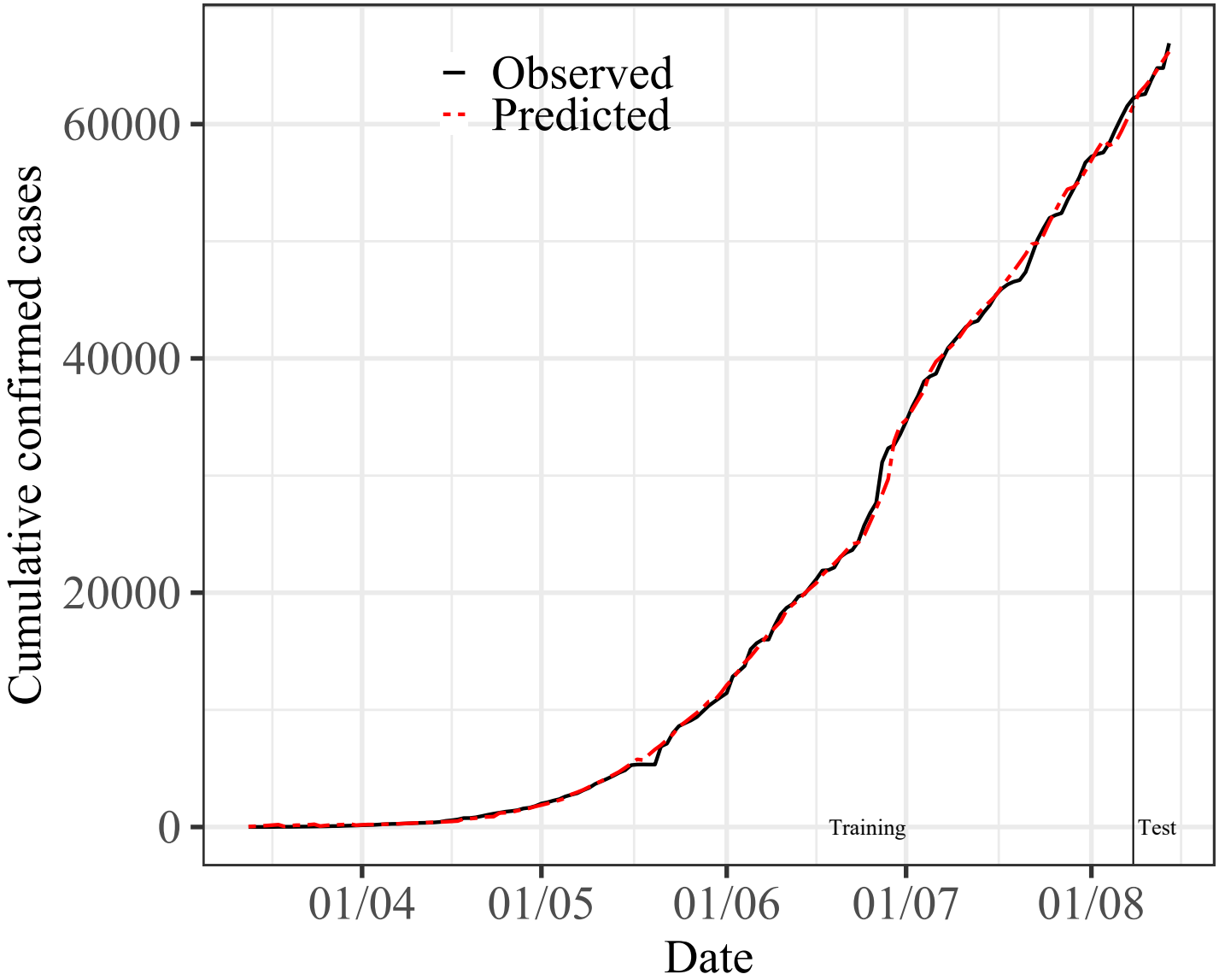
STACK approach outperformed the other models in Salvador's cases.

MAE improvement ranged between 36.78% – 88.71%.

Ranking

- STACK
- ARIMA
- SVR
- kNN

Findings - Salvador



Findings - São Paulo

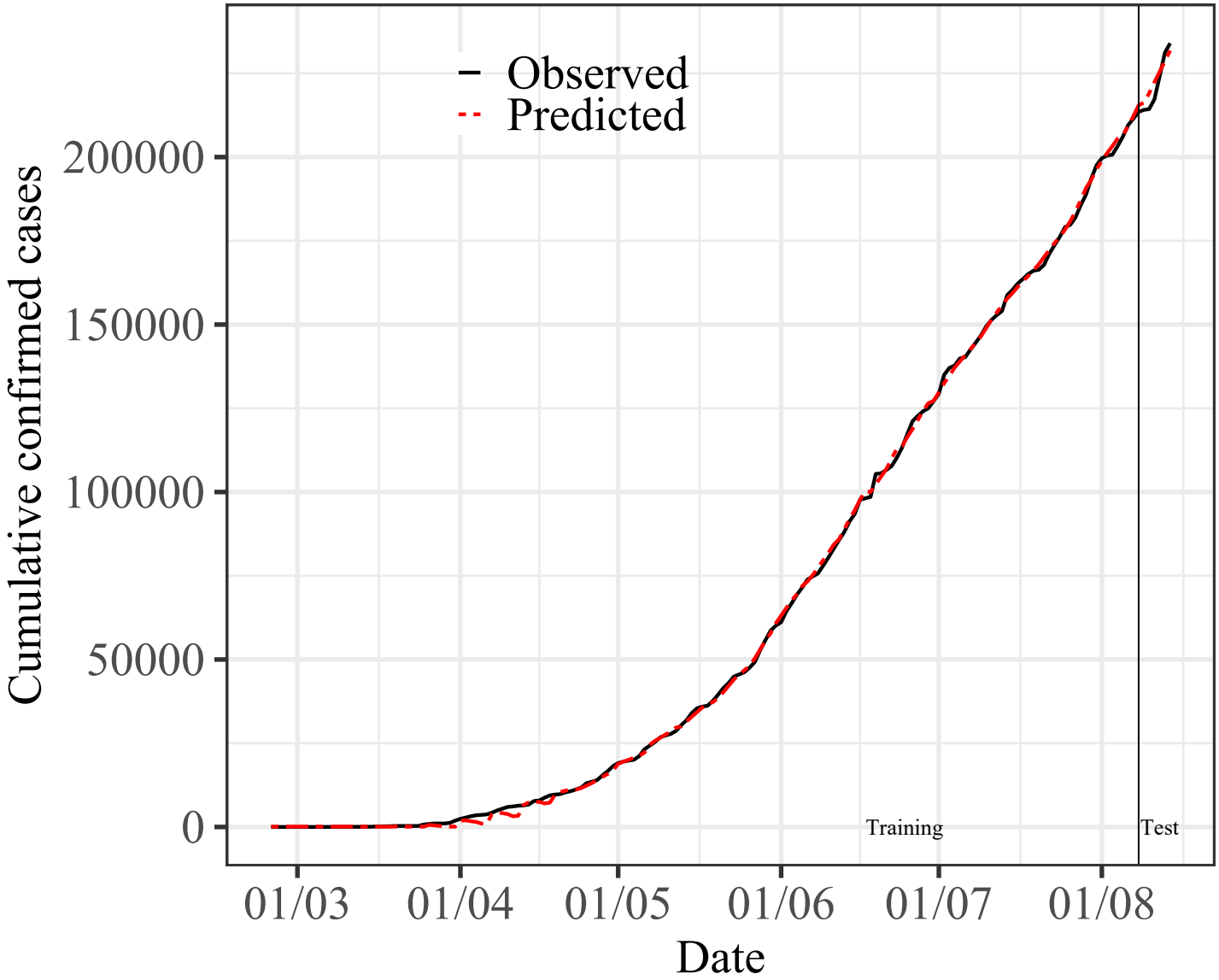
STACK approach achieved better accuracy than other models.

The improvement in MAE ranged in 20.30% – 78.01%.

Ranking

- STACK
- ARIMA
- SVR
- kNN

Findings - São Paulo



Findings - Overview

Table: Performance measures of the models

City	Criteria	ARIMA	kNN	STACK	SVR
Brasília	MAE	501.83	9020.17	**308.17**	3982.00
	RMSE	528.16	9471.26	**394.21**	4196.15
	sMAPE	0.44%	8.27%	**0.28%**	3.56%
Rio de Janeiro	MAE	1374.83	3285.83	2120.83	**1249.17**
	RRMSE	1700.76	3838.18	2734.95	**1502.23**
	sMAPE	1.78%	4.33%	2.76%	**1.62%**
São Paulo	MAE	673.33	3769.67	**425.67**	2346.83
	RMSE	766.89	4064.64	**500.17**	2429.35
	sMAPE	1.05%	6.02%	**0.66%**	3.57%
Salvador	MAE	3784.83	13716.50	**3016.50**	5277.33
	RRMSE	4242.16	15833.98	**3324.96**	7355.84
	sMAPE	1.69%	6.30%	**1.36%**	2.35%

Conclusion

Stacking-ensemble learning and **SVR** models are suitable tools to forecast COVID-19 cases for most of the adopted cities.

These approaches were able to learn the **nonlinearities** inherent to the evaluated **epidemiological time series**.

ARIMA also showed good consistent results for a six-days-ahead time window.

Conclusion

The ranking of models in **Rio de Janeiro**'s scenarios is:

1. SVR
2. ARIMA
3. STACK
4. KNN

For all the **other cities**:

1. STACK
2. ARIMA
3. SVR
4. KNN

Conclusion

Remark

- Even though the models discussed in this paper presented accurate forecasting capabilities, they should be used cautiously.
- The **dynamics** of the analyzed data, as well as the diversity of **exogenous factors** that can affect the daily notifications of COVID-19

Acknowledgments

Thank you!