Biased Resampling Strategies for Imbalanced Spatio-temporal Forecasting

MARIANA OLIVEIRA[1,2], NUNO MONIZ[1,2], LUÍS TORGO[1,2,3] AND VÍTOR SANTOS COSTA[1,2]

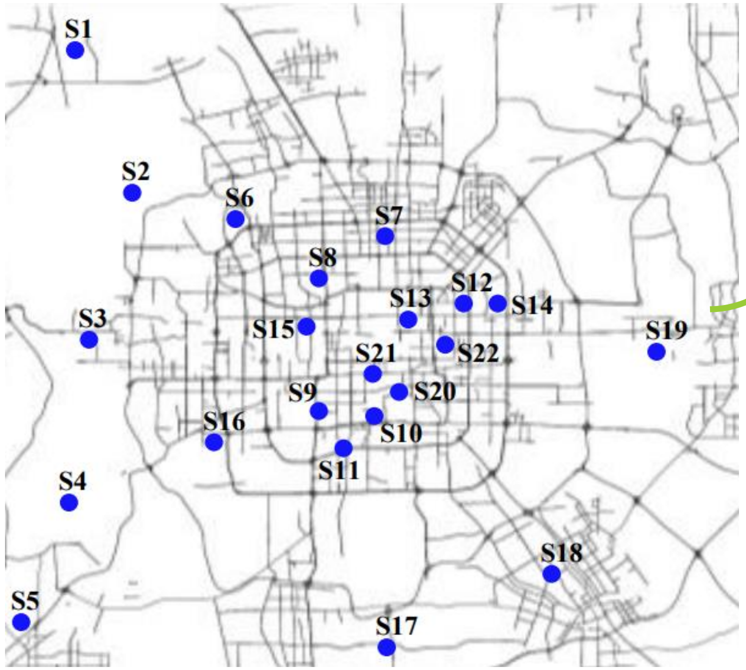
5-8 OCTOBER 2019

1 U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

2 INESCTEC TECHNOLOGY & SCIENCE ASSOCIATE LABORATORY PORTUGAL

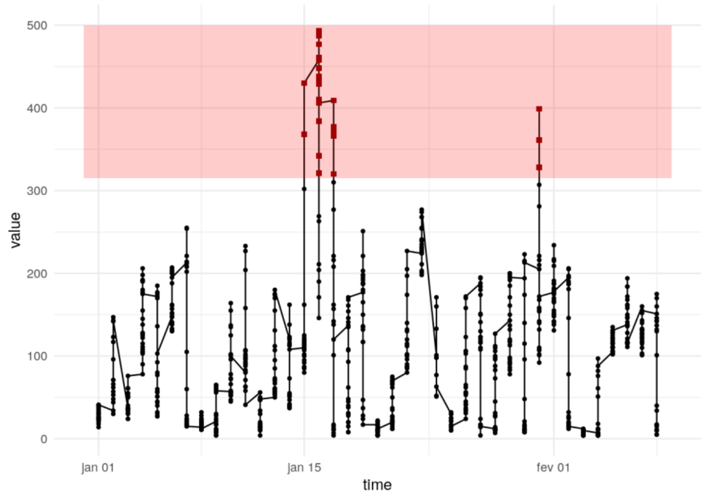
3 DALHOUSIE UNIVERSITY

Spatio-temporal Data



Air quality measurement station network (Source: Zheng et al., 2013)

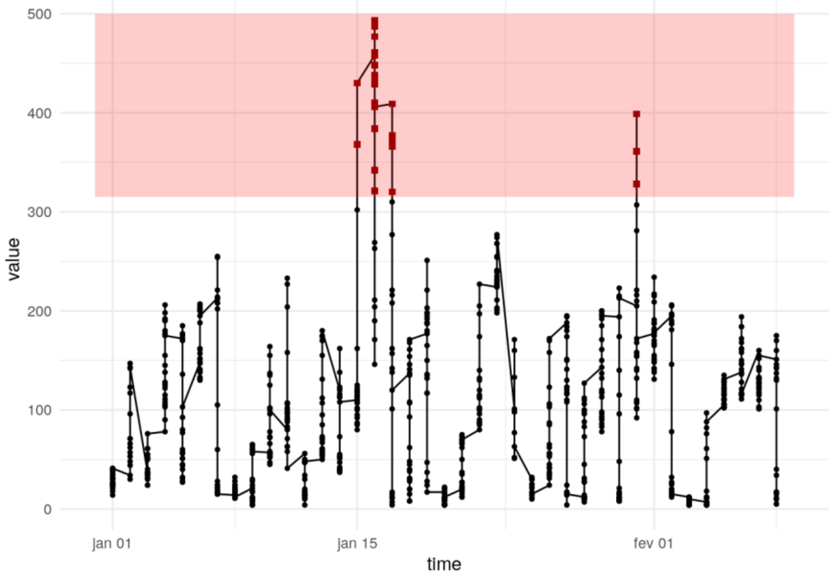
Remote monitoring equipment (Source: NDSU)



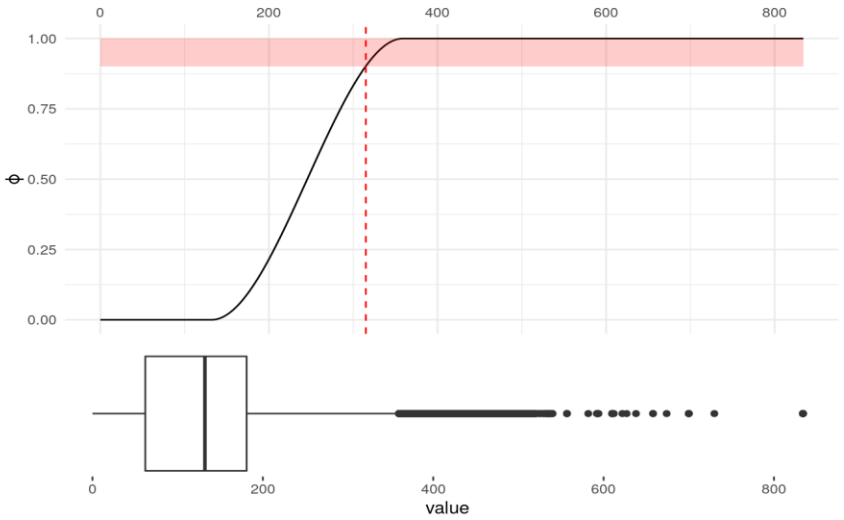
PM 2.5 pollution levels (time series)

Imbalanced Numeric Forecasting

IMBALANCED DOMAIN



RELEVANCE FUNCTION



Our Contribution

Motivation

- **Random resampling** approaches are often used to tackle this problem
- However, our **data is not i.i.d.** -- there are spatial and temporal dependencies

Research Questions

- Will introducing a **sampling bias** that takes into account spatio-temporal dependencies improve performance?
- Should we **weight the dimensions** differently?

Biased Resampling

Proposed Resampling Strategies

Spatio-temporal Random Under-sampling *(STRUS)*

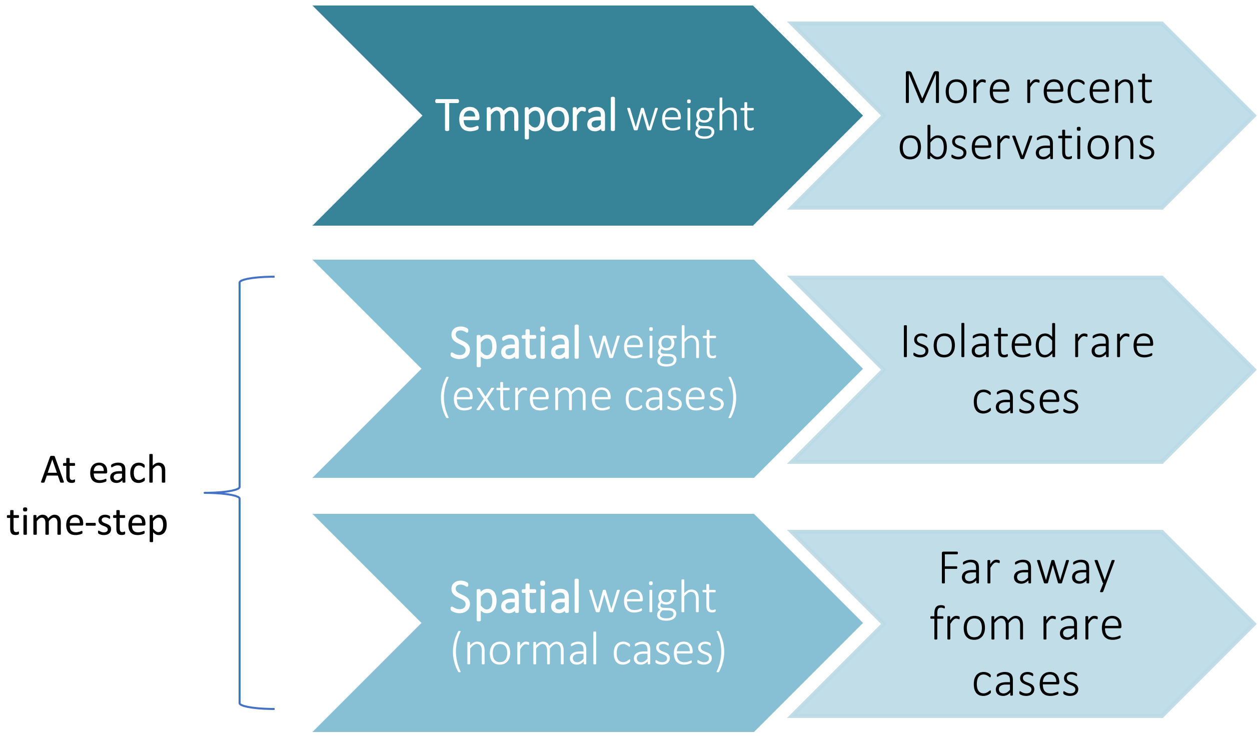
- Keep all extreme cases
- Keep only u% of normal cases, $0 < u < 100$ (with sampling bias)

Spatio-temporal Random Over-sampling *(STROS)*

- Keep all (normal and extreme) cases
- Add o% replicas of extreme cases, $o > 0$ (with sampling bias)

Spatio-temporal Sampling Bias

Which cases should have higher probability of being selected during resampling?



Spatio-temporal Sampling Bias

What if spatial and temporal dimensions have different impacts?

Add weighting parameter α

$$W_{i,j} = \alpha \times W^{T}_{i,j} + (1-\alpha) \times W^{L}_{i,j} + \epsilon$$

Experiments

Datasets

Data source	ID	# time IDs	# loc IDs	% available	% extreme
MESA	10	280	20	100	7.3
NCDC	20	105	72	100	6.0
	30				6.3
TCE	31	330	26	100	3.8
	32				2.4
Rural	40	4k	70	~49	7.5
	50				3.5
Beijing Air	51	11k	36	~40	5.5
	52				8.6
	53				3.8

Learning Process

Feature engineering

Calculate spatio-temporal indicators

Resampling

- None
- RUS
- ROS
- STRUS
- STROS

Model

- MARS
- Random Forest
- RPART

PRE-PROCESSING

Experimental Evaluation

Evaluation metrics

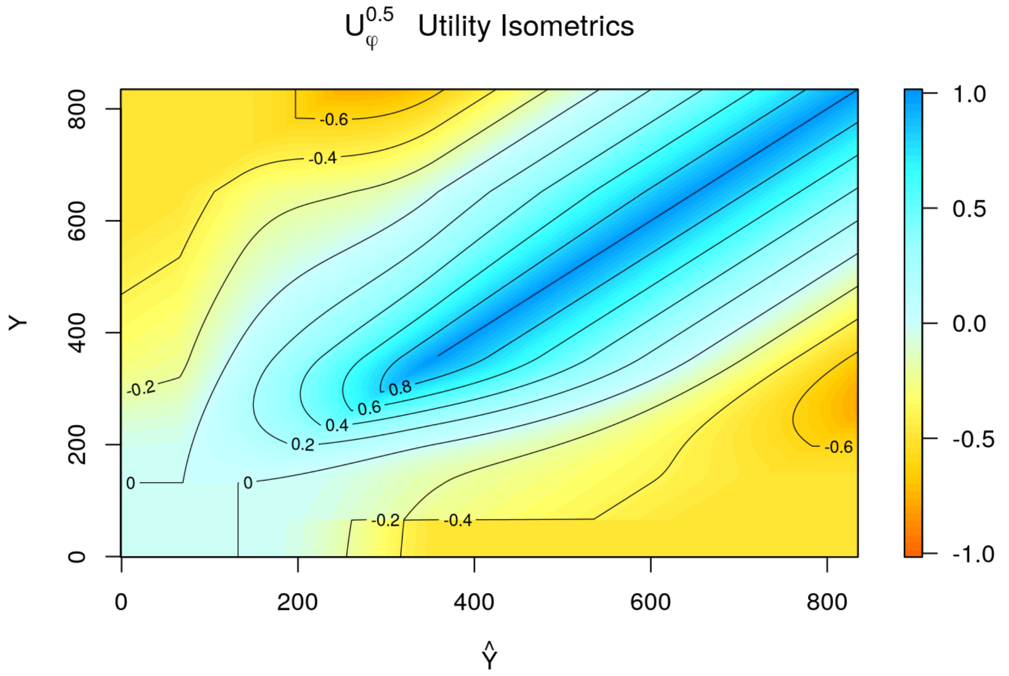
Performance estimation procedure

Evaluation Metrics

- **Utility-based** precision and recall for numeric prediction:

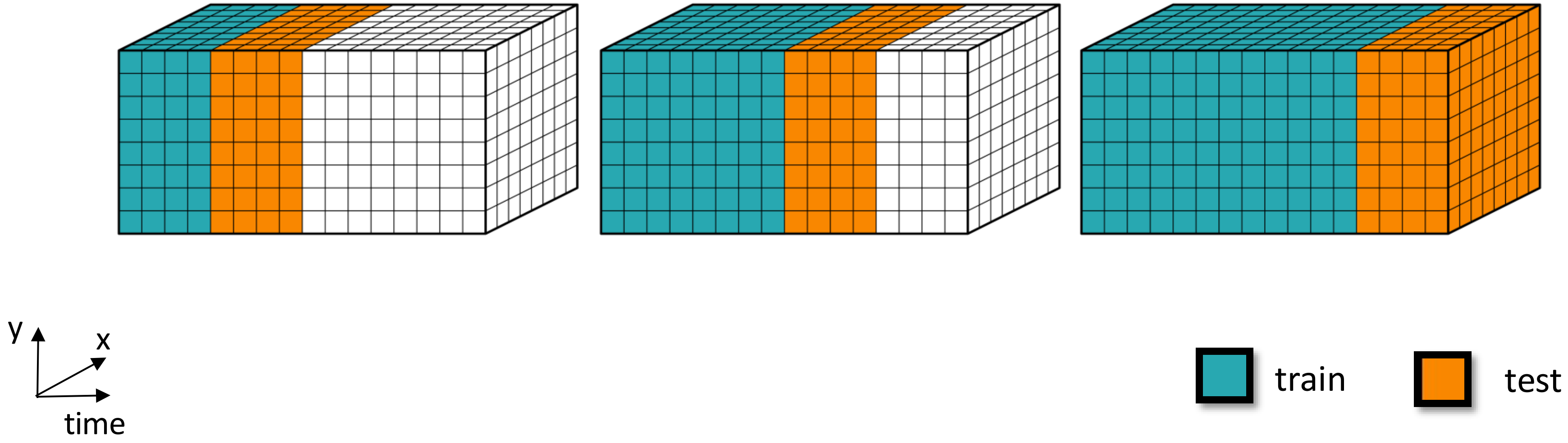
$$prec_{\phi}^{u} = \frac{\sum_{\phi(\hat{y}_i)\geq t_R, \phi(y_i)\geq t_R} (1 + u(\hat{y}_i, y_i))}{\sum_{\phi(\hat{y}_i)\geq t_R} (1 + \phi(\hat{y}_i))}$$

$$rec_{\phi}^{u} = \frac{\sum_{\phi(\hat{y}_i)\geq t_R, \phi(y)\geq t_R} (1 + u(\hat{y}_i, y_i))}{\sum_{\phi(y_i)\geq t_R} (1 + \phi(y_i))}$$



Performance Estimation Procedure

- **Prequential** temporal block evaluation



Parametrization

Internal tuning

Fixed a priori

Optimal a posteriori

Parametrization

Internal tuning

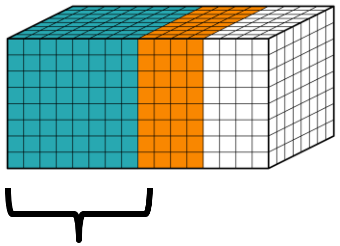
Fixed a priori

Optimal a posteriori

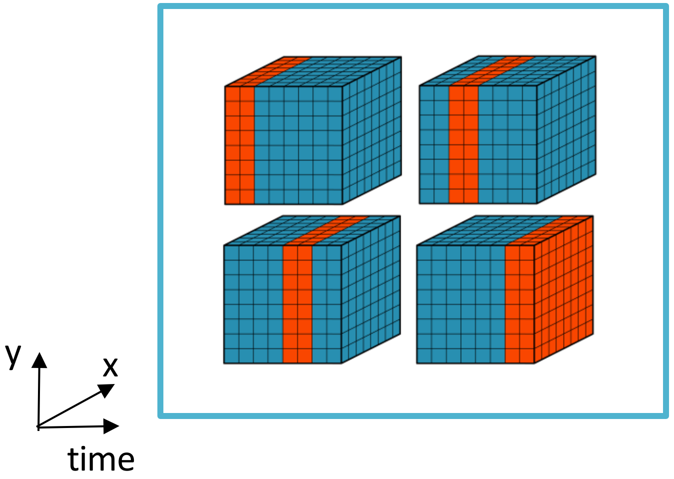
Internal Tuning

INTERNAL ESTIMATION PROCEDURE

For each training set:



Temporal-block CV



PARAMETER GRID SEARCH

Parameter	Values
u	0.2; 0.4; 0.6; 0.8; 0.95
o	0.5; 1; 2; 3; 4
α	0; 0.25; 0.5; 0.75; 1

Parametrization

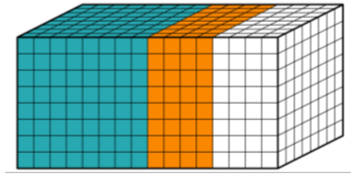
Internal tuning

Fixed a priori

Optimal a posteriori

Fixed a priori

For all training sets:



Fixed parameters at middle of the grid.

FIXED PARAMETERS

Parameter	Values
u	~~0.2; 0.4;~~ **0.6**; ~~0.8; 0.95~~
o	~~0.5; 1;~~ **2**; ~~3; 4~~
α	~~0; 0.25;~~ **0.5**; ~~0.75; 1~~

Parametrization

Internal tuning

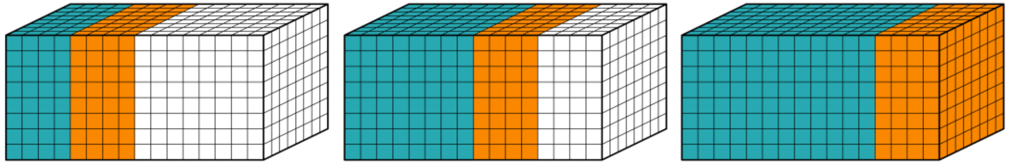
Fixed a priori

Optimal a posteriori

Optimal a posteriori

EXTERNAL ESTIMATION PROCEDURE

For each data set:



Choose parameters with **best** results on the external (prequential) procedure.

PARAMETER GRID SEARCH

Parameter	Values
u	0.2; 0.4; 0.6; 0.8; 0.95
o	0.5; 1; 2; 3; 4
α	0; 0.25; 0.5; 0.75; 1

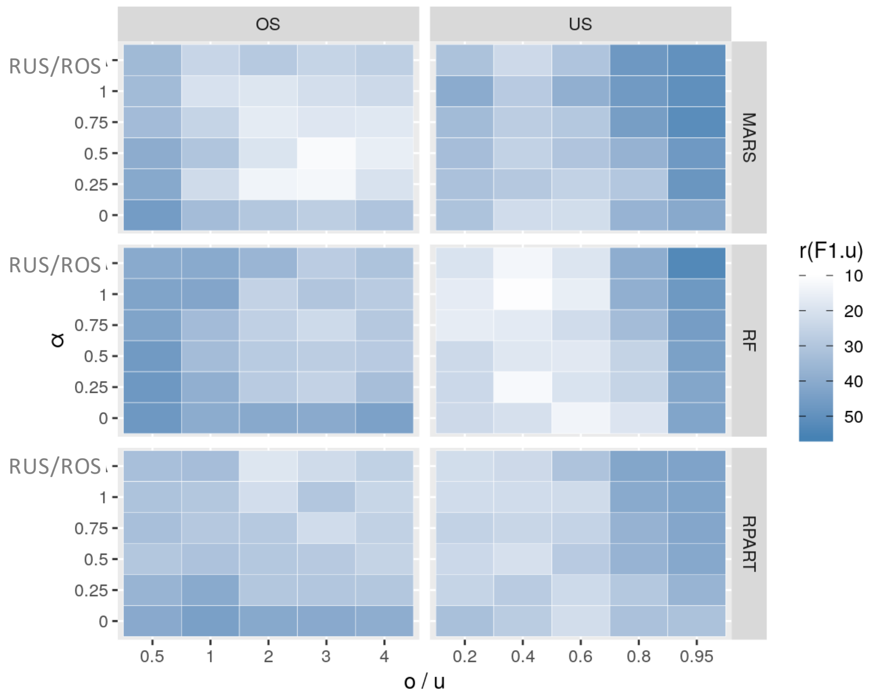
Results

Average Rank of F_1^u

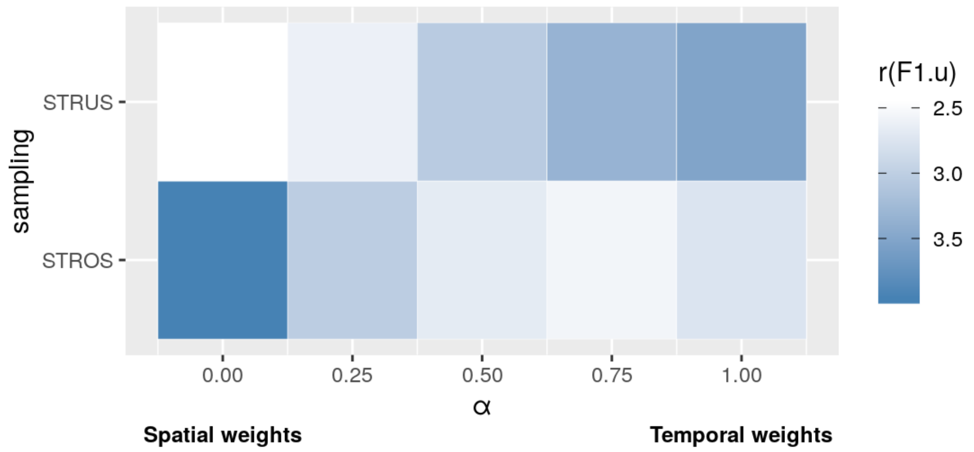
Parametrization	None	ROS	STROS	RUS	STRUS
Internal tuning	4.60	3.07	2.37	2.67	2.30
Fixed a priori	4.53	2.77	2.73	2.57	2.40
Optimal a posteriori	5.00	3.07	2.27	2.93	1.73

Parameter Sensitivity Analysis

TWO PARAMETERS

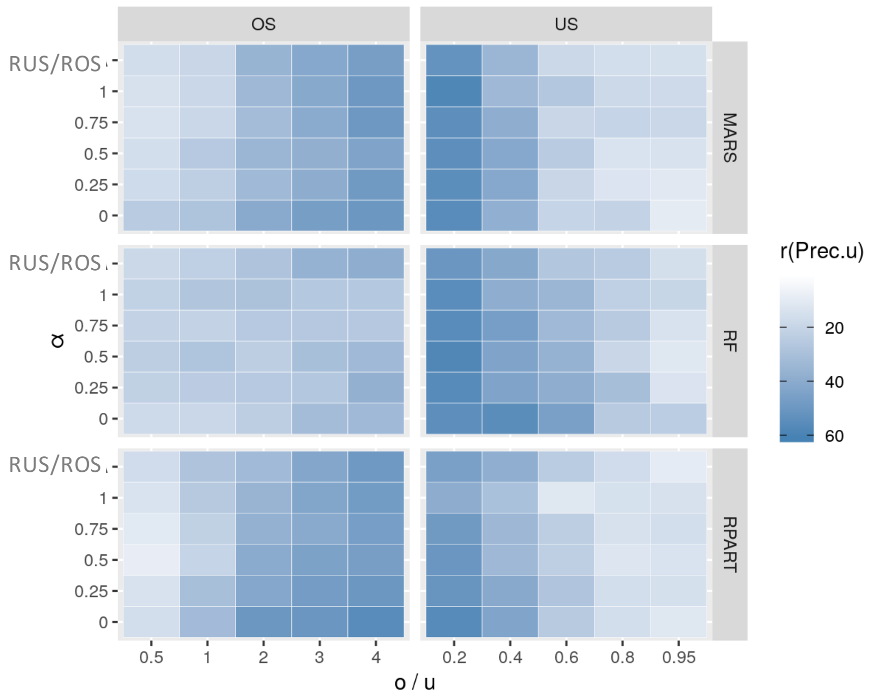


DIMENSION WEIGHTING

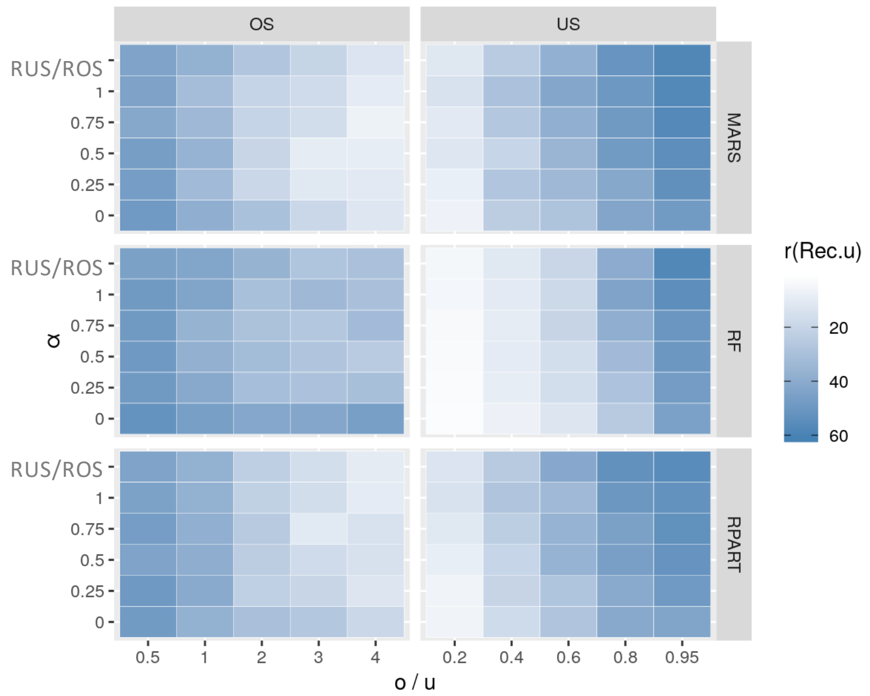


Precision and Recall Trade-off

PRECISION



RECALL



Conclusion

Conclusion

- Including **spatio-temporal bias** when resampling improves performance
- The contributions of each dimension should be **weigthed**:
  - When **over-sampling**: favour temporal weight and prioritize **more recent** observations
  - When **under-sampling**: favour spatial weight and prioritize **isolated** rare cases and normal cases that are spatially distant from extreme cases
- Future work:
  - Study the impact of data characteristics on performance
  - Consider local instead of global definitions of extreme values

Thank you!

Code available at https://github.com/mrfoliveira/STResampling-DSAA2019