Biased Resampling Strategies for Imbalanced Spatio-temporal Forecasting

Mariana Oliveira[1,2] · Nuno Moniz[1,2] · Luís Torgo[1,2,3] · Vítor Santos Costa[1,2]
(mariana.r.oliveira@inesctec.pt) (nmoniz@inesctec.pt)

1 INESCTEC
2 U.PORTO FC Faculdade de Ciências Universidade do Porto
3 Dalhousie University

Motivation

- Extreme and rare events like spikes in air pollution can have serious repercussions, and many of these events arise from spatio-temporal processes;
- Learning approaches usually assume that:
  - Users have uniform domain preferences when, in reality, the accurate forecasting of extreme values may be of more importance;
  - Data is i.i.d. which is often false for spatio-temporal data;
- When working with imbalanced domains:
  - Relevance functions and utility-based metrics should be used for evaluation;
  - Random resampling is usually used to improve prediction of extreme values;
- We investigate the following research questions:
  - Will Introducing a sampling bias that takes into account the implicit spatio-temporal dependencies in the data improve performance?
  - Should we weight the spatial and temporal dimensions differently?

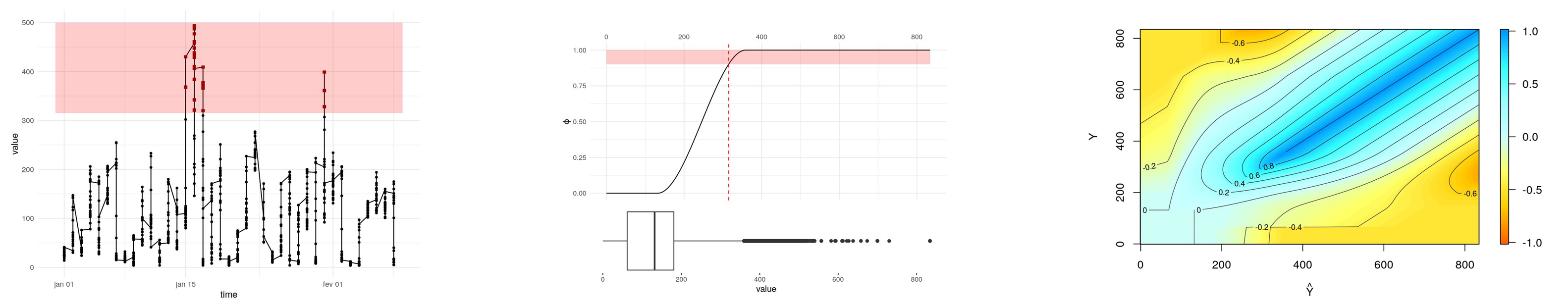

Figure 1. Time series of air pollution measured at a Beijing station (left); Relevance function automatically calculated for the domain based on boxplot of values (middle); Utility (u) isometrics of predictions (Ŷ) of real values (Y) (right).

Biased Resampling Strategies

Spatio-Temporal bias Random Under-Sampling (STRUS)

1. Keep all (normal and extreme) cases;
2. Add o% replicas of extreme cases with sampling bias, o>0

Spatio-Temporal bias Random Over-Sampling (STROS)

1. Keep all extreme cases;
2. Add u% of normal cases with sampling bias, 0<u<100

Resampling bias weight Which cases should be prioritized during resampling?

- **Temporal weight**: Keep/add more recent observations *with higher probability*
- **Spatial weight**: At each time-step, add more isolated extreme cases / keep normal cases that are farther away from extreme cases *with higher probability*

What if temporal and spatial dimensions have different impacts? Add parameter α.

$$W_{i,j} = \alpha \times W^T_{i,j} + (1-\alpha) \times W^L_{i,j} + \epsilon$$

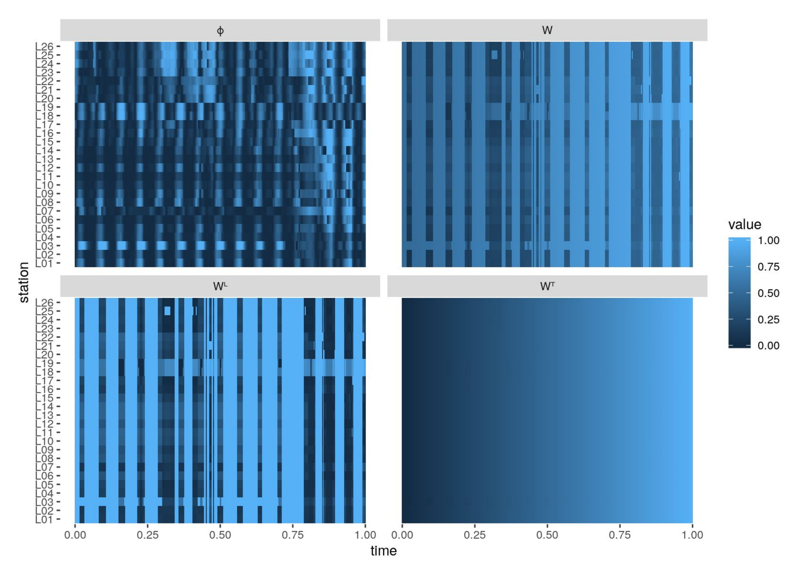

Figure 2. Heatmap showing relevance (ϕ), spatial bias weight (W^L), temporal bias weight (W^T), and spatio-temporal bias weight (W) for an example data set. Each cell corresponds to an observation at a given location and time point.

Experimental setup

Goal: Compare random under- and over-sampling (RUS and ROS) against proposed spatio-temporal bias under- and over-sampling (STRUS and STROS) and a baseline.

Datasets & Learning models

- 10 variables from 5 real-world data sources including climate and air pollution of different sizes, levels of missing data, and proportions of normal and extreme values (2.4 - 8.6 % extreme values);
- 3 different off-the-shelf learning models: MARS, RPART and Random Forests (RF).

Evaluation metrics & Estimation procedures

We calculate utility-based precision and recall:

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

To estimate evaluation metrics, we use prequential temporal-block evaluation.

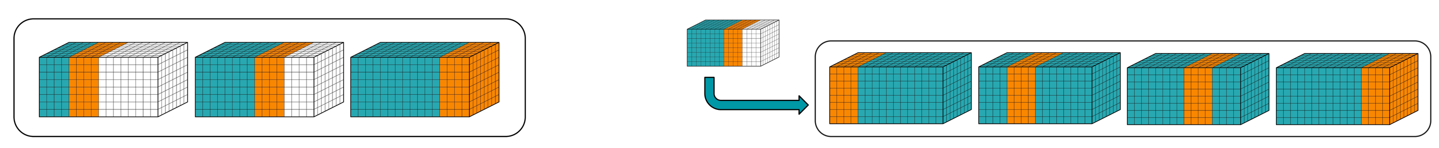

Figure 3. Prequential temporal block evaluation method (left) and internal temporal-block validation (right).

Parametrization scenarios

- **Internal validation**: For each training set, use temporal block cross-validation to estimate best parameters;
- **Fixed a priori**: For all training sets, use the same parameters;
- **Optimal a posteriori**: For each data set, choose parameters that achieved best average results over the training sets.

Results

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	2.77	2.57	**2.40**
Optimal a posteriori	5.00	3.07	3.07	2.93	**1.73**

Table 1. Average ranks of F1u results.. Ranks were always calculated separately for each learning model and data set pair before averaging. Lower ranks correspond to better results.

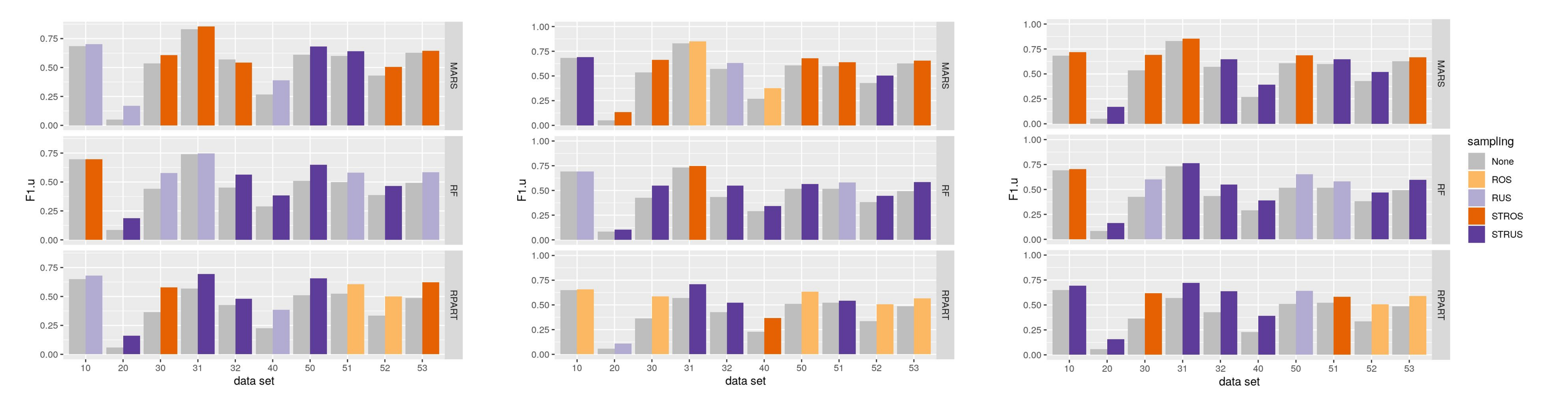

Figure 4. Baseline and best F1u achieved for each data set and learning model. Parameters internally tuned (left), fixed a priori (middle) and optimal found a posterior (left).

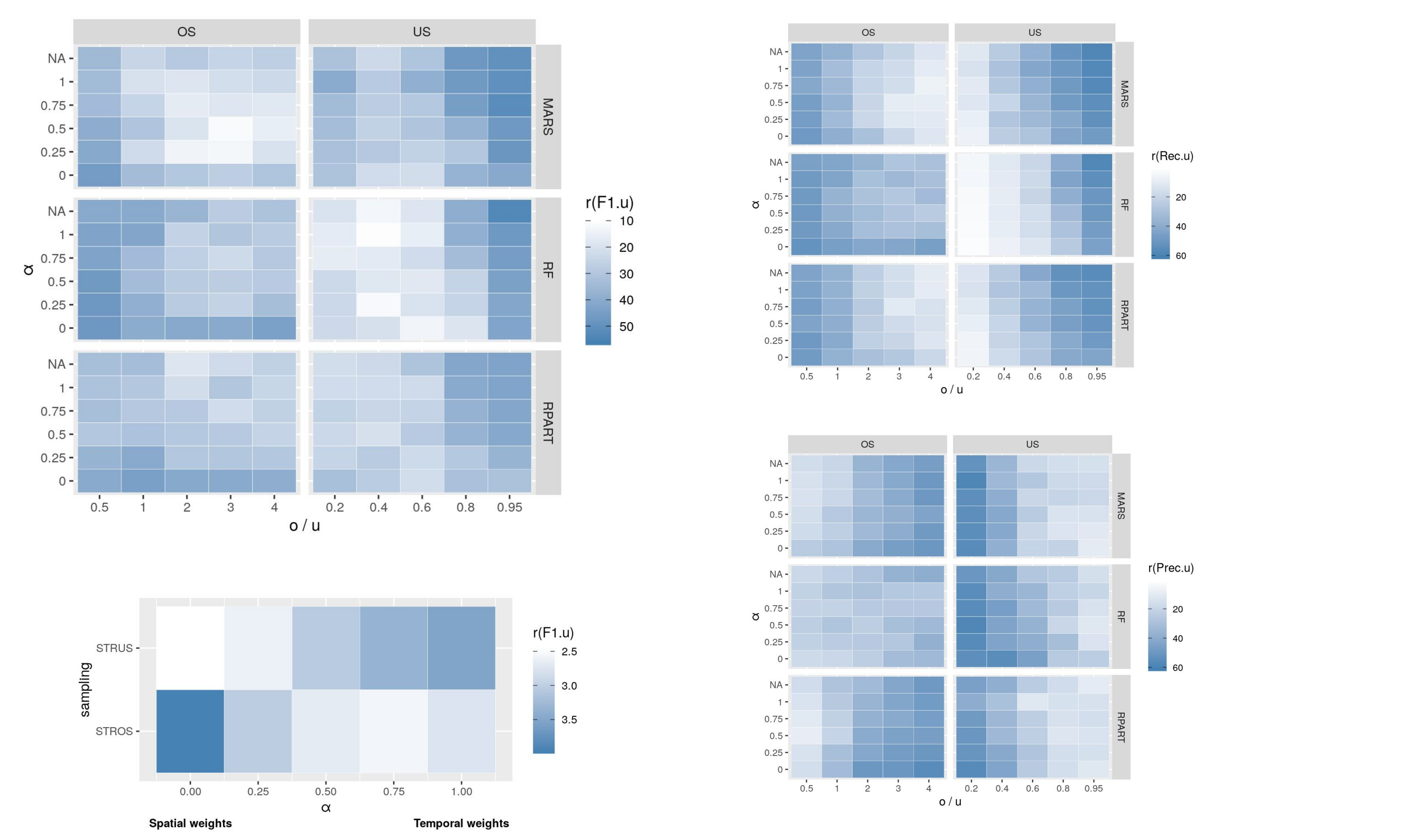

Figure 5. Average rank according to F1u (left), precision (top right) and recall (bottom right) for 60(+1) different parametrizations (aggregated by α on bottom left). The baseline was included in rank calculation, but excluded from the graphs. Non-biased resampling is denoted by α=NA.

Conclusions

- Including spatio-temporal bias when resampling improves performance;
- The contributions of each dimension should be weighted:
  - When over-sampling, favour temporal weight and prioritize more recent cases;
  - 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

Acknowledgements. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within the project: UID/EEA/50014/2019. Mariana Oliveira is supported by FCT/MAP-i PhD research grant (PD/BD/128166/2016). The work of L. Torgo was undertaken, in part, thanks to funding from the Canada Research Chairs program. V. Santos Costa gratefully acknowledges the project POCI-01-0145-FEDER-031356 -(PTDC/CCI-BIO/31356/2017).

FCT Fundação para a Ciência e a Tecnologia