NORTE2020 PROGRAMA OPERACIONAL REGIONAL DO NORTE
REPÚBLICA PORTUGUESA
PORTUGAL 2020
INESCTEC
FCT Fundação para a Ciência e a Tecnologia
PD/BD/128166/2016
U.PORTO
FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

PREDICTIVE ANALYTICS FOR SPATIO-TEMPORAL DATA

MARIANA OLIVEIRA

ADVISOR: LUÍS TORGO

CO-ADVISOR: VÍTOR SANTOS COSTA

MAP-I DOCTORAL PROGRAM

FACULTY OF SCIENCES, UNIVERSITY OF PORTO

13 DECEMBER 2021

universidade de aveiro
Universidade do Minho
U.PORTO

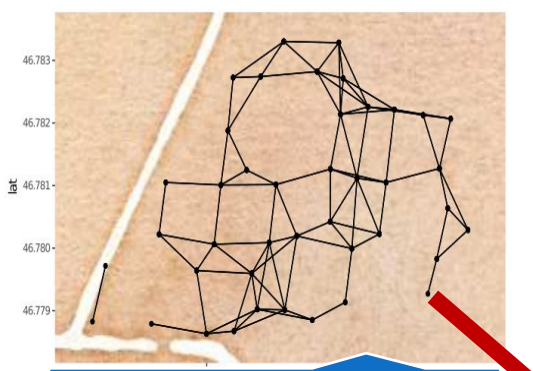
OUTLINE

- Context and Motivation
- Evaluating Spatio-Temporal Forecasting
- Extracting Spatio-Temporal Indicators
- Resampling Imbalanced Spatio-Temporal Data
- Final Remarks

CONTEXT AND MOTIVATION

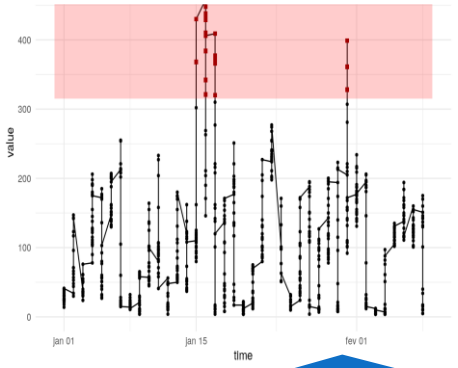
Predictive Analytics for Spatio-Temporal Data

GEO-REFERENCED TIME SERIES
CONTEXT AND MOTIVATION



Sensor Network

Sensor

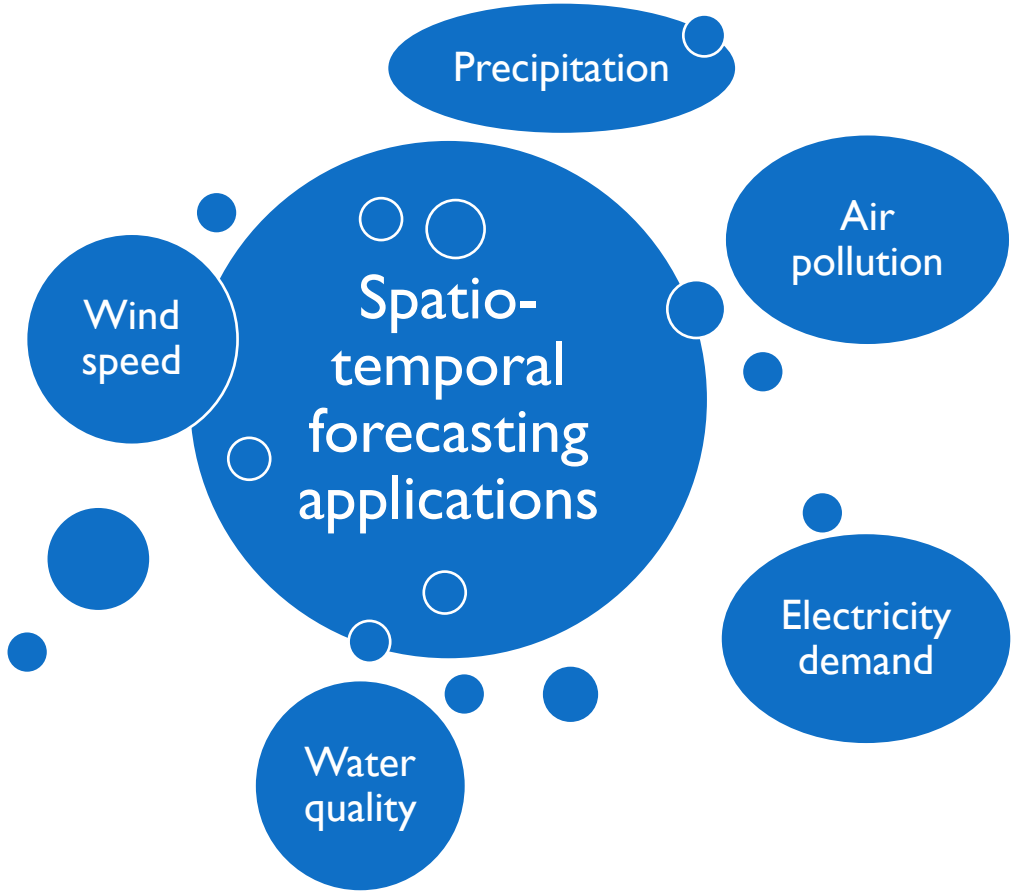


Geo-referenced time series

GEO-REFERENCED TIME SERIES FORECASTING
CONTEXT AND MOTIVATION

Spatio-temporal forecasting

- Given historical data, predict numeric values for a given location at a future time.



PROPERTIES AND CHALLENGES
CONTEXT AND MOTIVATION

Properties

- Autocorrelation
- Heterogeneity and non-stationarity
- Complex and implicit relationships
- Aggregation effects

PROPERTIES AND CHALLENGES
CONTEXT AND MOTIVATION

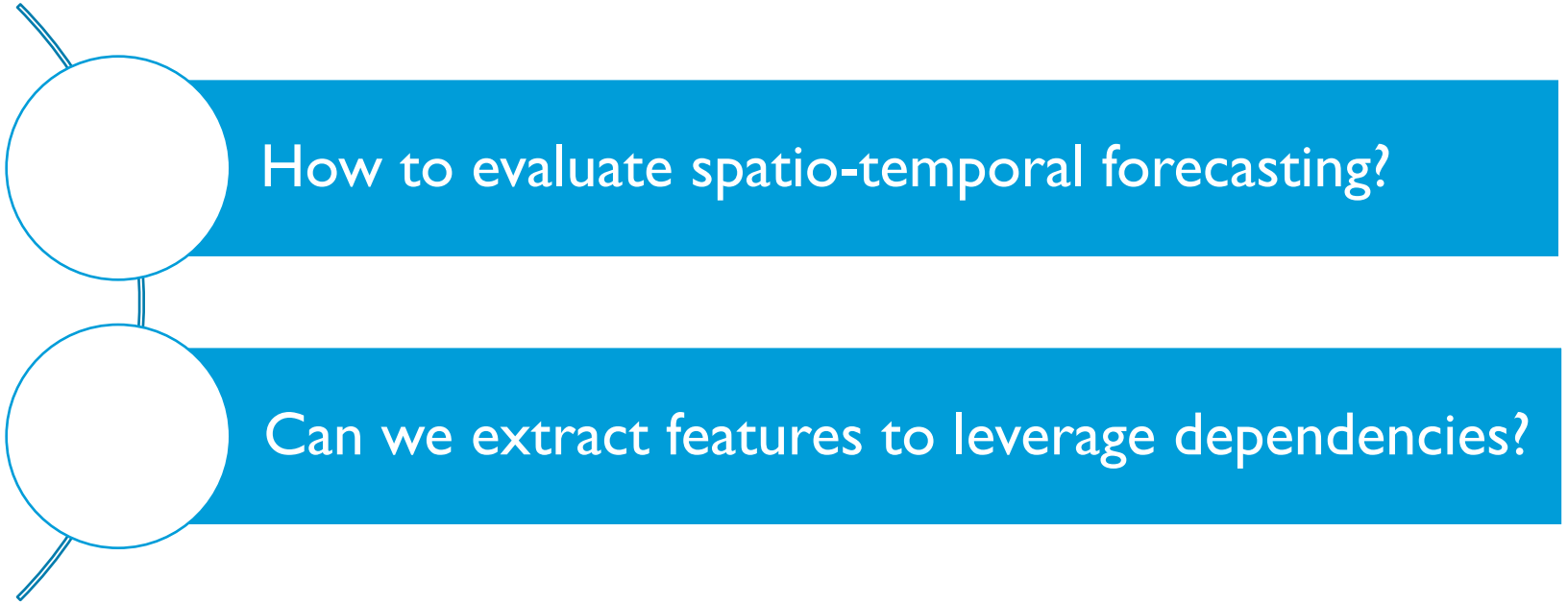
Properties

- Autocorrelation
- Heterogeneity and non-stationarity
- Complex and implicit relationships
- Aggregation effects

Challenges

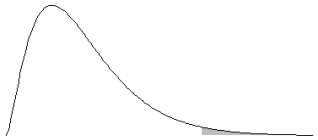
- Most learning and evaluation methods assume data to be i.i.d.
- Most methods do not leverage relationships between observations
- Results may vary depending on scale

RESEARCH QUESTIONS



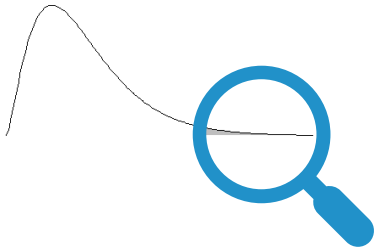
IMBALANCE PROBLEM
CONTEXT AND MOTIVATION

Properties



Imbalanced distribution

User interested in underrepresented values range

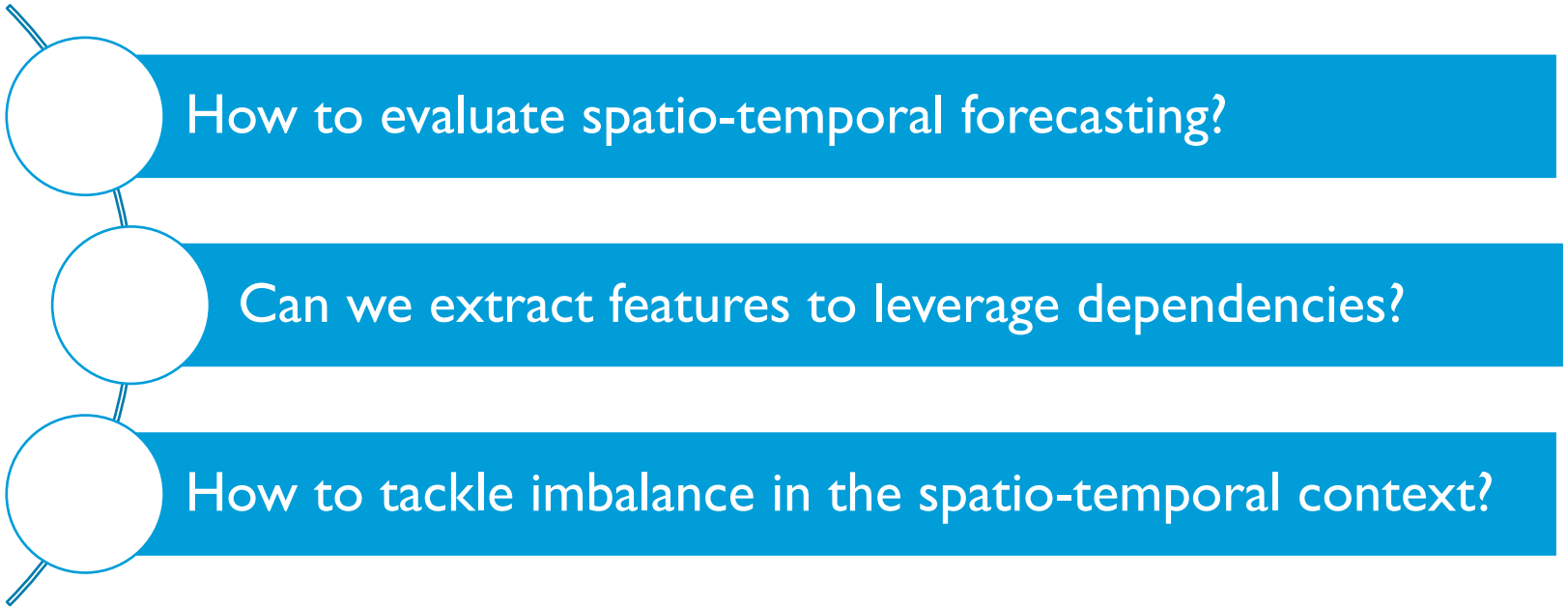


Challenges

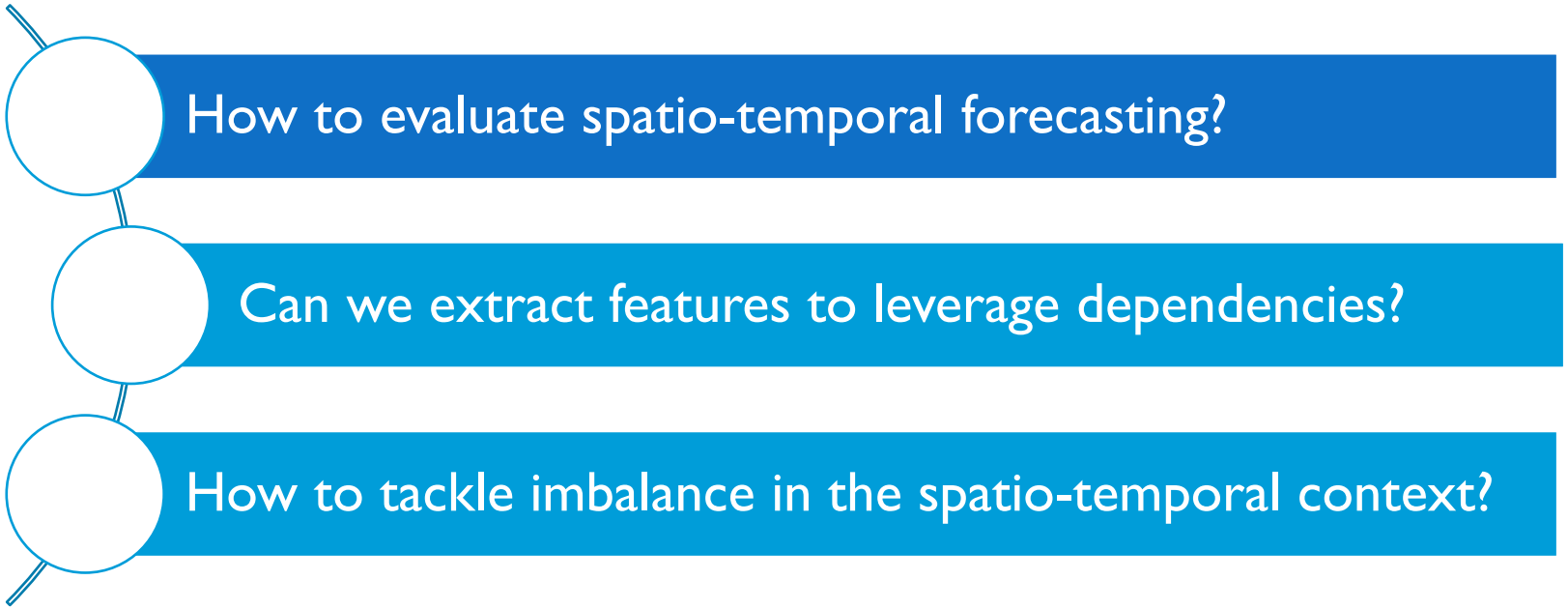
Most evaluation and learning methods assume balanced data

Most evaluation metrics assume equal benefits and costs for all numerical predictions

RESEARCH QUESTIONS



RESEARCH QUESTIONS

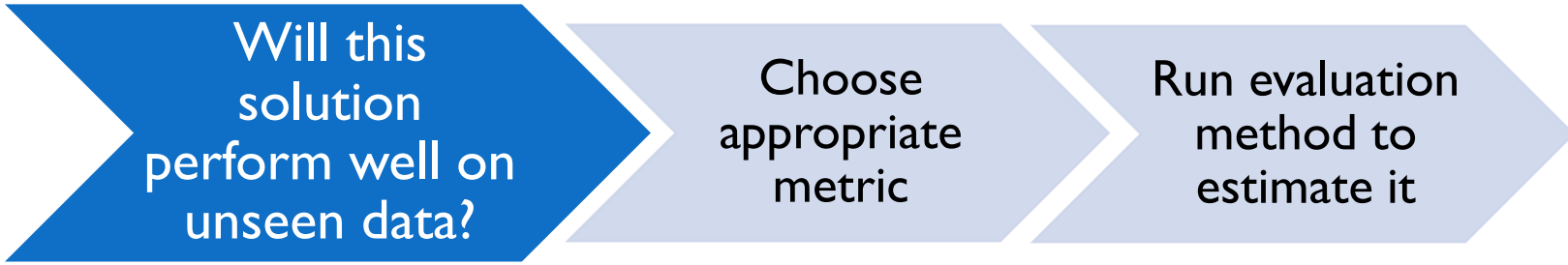


EVALUATING SPATIO-TEMPORAL FORECASTING

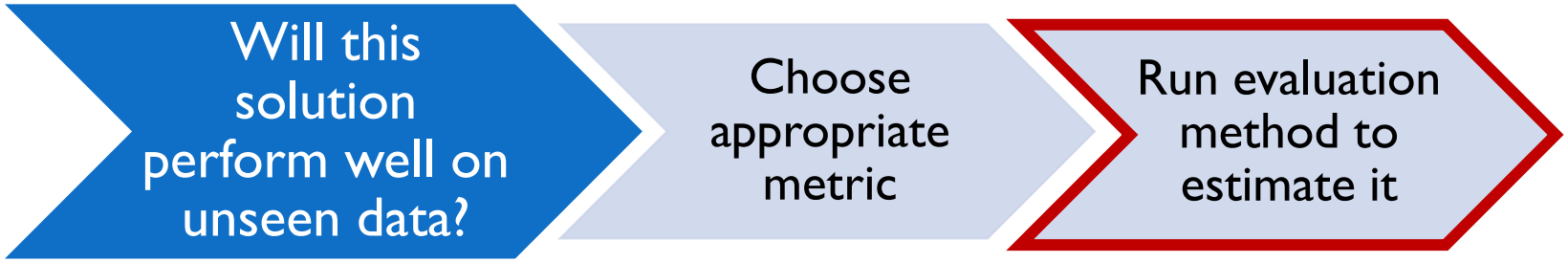
Mariana Oliveira, Luís Torgo, and Vítor Santos Costa. Evaluation Procedures for Forecasting with Spatio-Temporal Data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), volume 11051 LNAI, pages 703–718, 2018.

Mariana Oliveira, Luís Torgo, and Vítor Santos Costa. Evaluation Procedures for Forecasting with Spatiotemporal Data. Mathematics, 9(6), 2021.

PERFORMANCE EVALUATION
EVALUATING SPATIO-TEMPORAL FORECASTING



PERFORMANCE EVALUATION
EVALUATING SPATIO-TEMPORAL FORECASTING



Empirical study of performance estimation methods

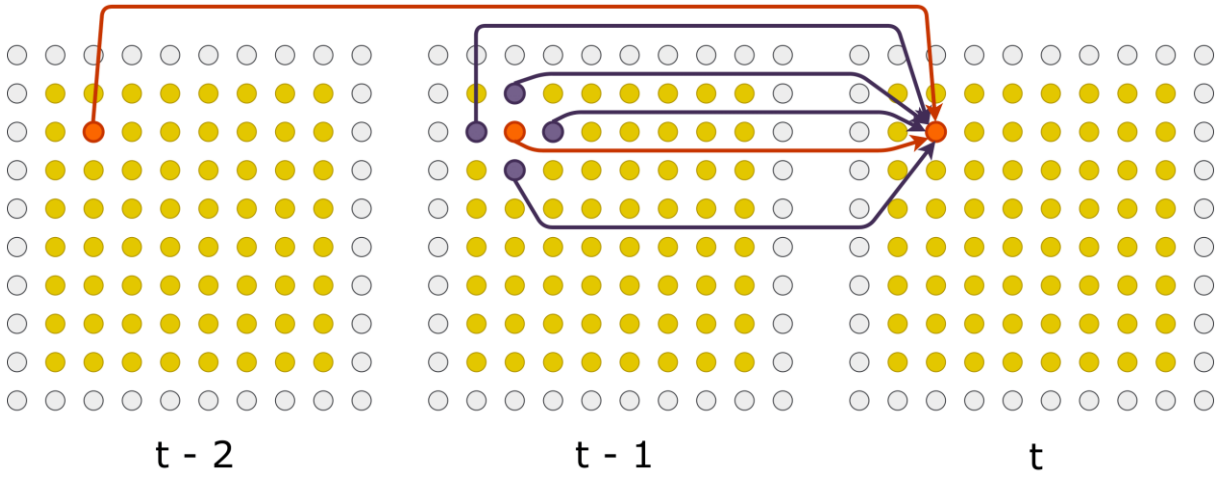
ARTIFICIAL DATA
EVALUATING SPATIO-TEMPORAL FORECASTING

- STARMA, STMA, STAR and NL-STAR
- Orders 2(10), 2(01) and 2(11)
- 8x8 and 20x20 regular grids
- 150 and 300 time points
- 4 sets of random coefficients each

---

192 datasets with embed 3(110)

STAR 2(10)



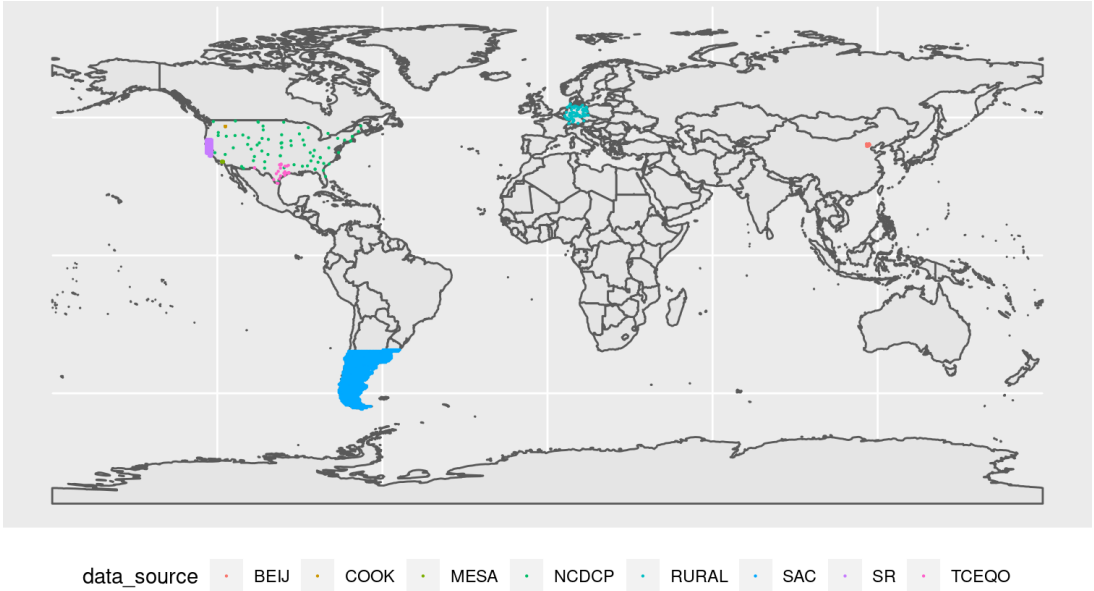
REAL-WORLD DATA

EVALUATING SPATIO-TEMPORAL FORECASTING

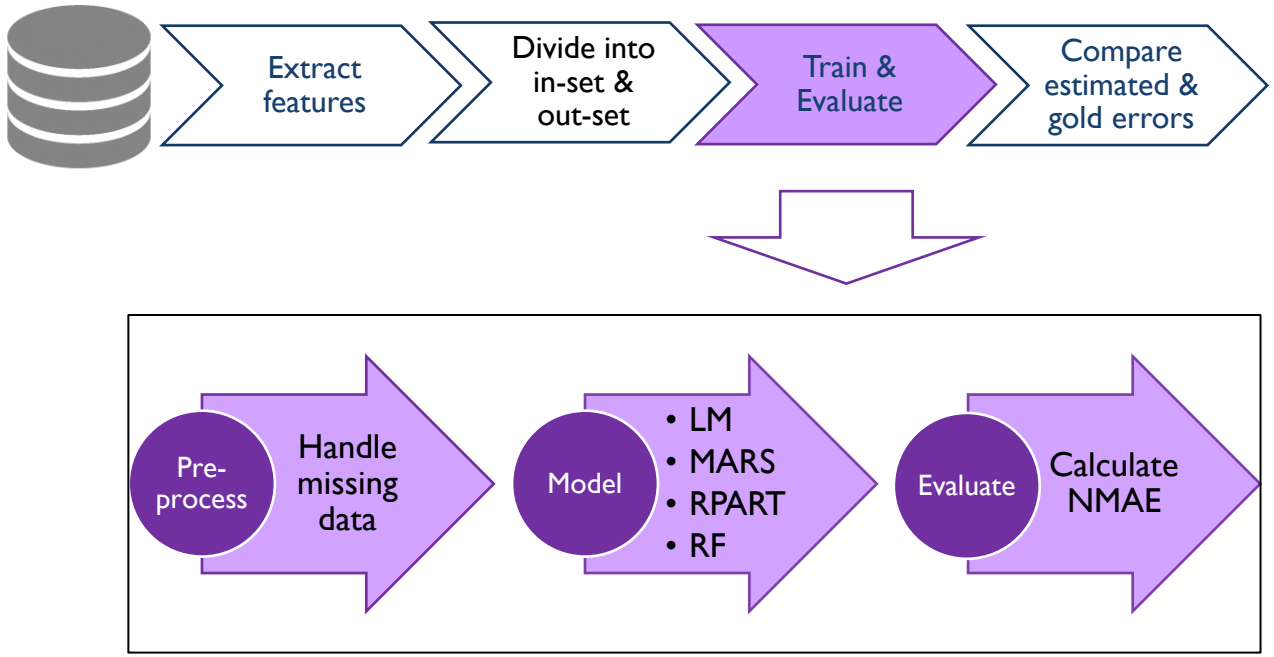
ID	Source	#vars	#times	#locs	% avail.
1	MESA	1	280	20	100
2	NCDC	2	105	72	100
3	TCE	3	360	26	100
4	Cook	3	729	40	74
5	SAC	1	144	900	100
6	airBase	1	4.4k	70	49
7	Beijing air	6	6.6k	36	67

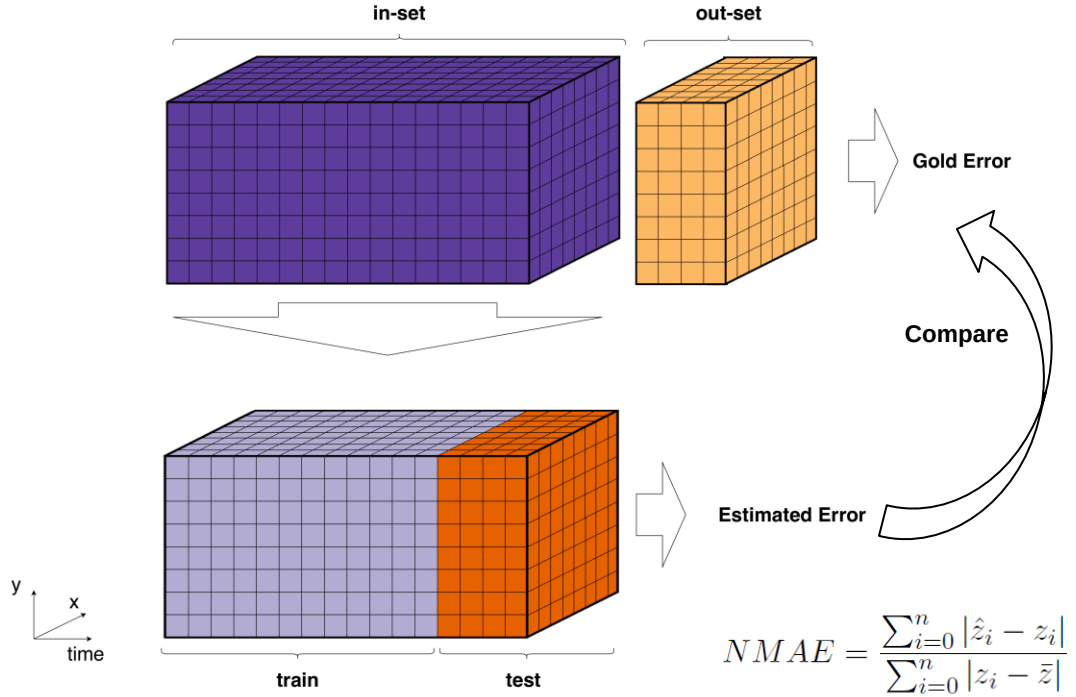
17 univariate datasets with spatio-temporal indicators

Data locations



EXPERIMENTAL SETUP
EVALUATING SPATIO-TEMPORAL FORECASTING



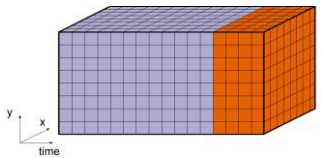


OOS or CV method

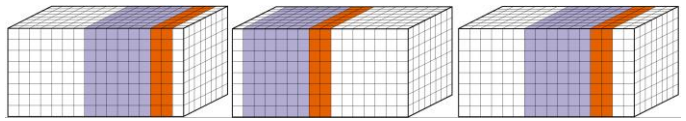
EVALUATION METHODS
EVALUATING SPATIO-TEMPORAL FORECASTING

Out-of-Sample (OOS)

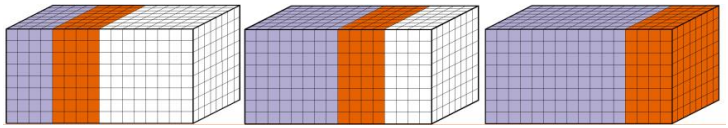
Holdout (HO)



Monte Carlo Holdout (MC)



Prequential temporal block evaluation (Preq-Tb)

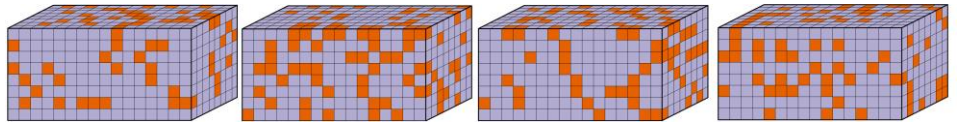


Cross-Validation (CV)

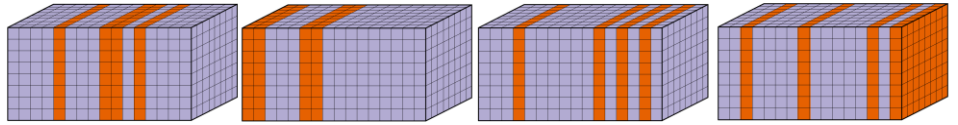
train

test

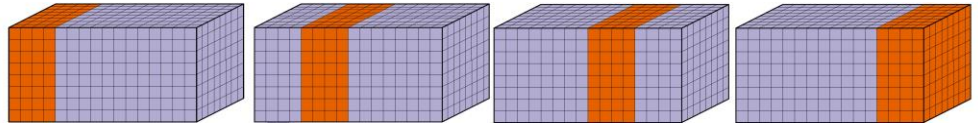
Standard CV (CV)



Time-sliced CV (CV-Tsl)

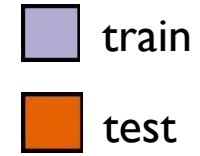


Temporal block CV (CV-Tb)

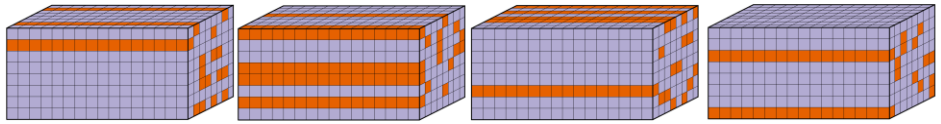


EVALUATION METHODS
EVALUATING SPATIO-TEMPORAL FORECASTING

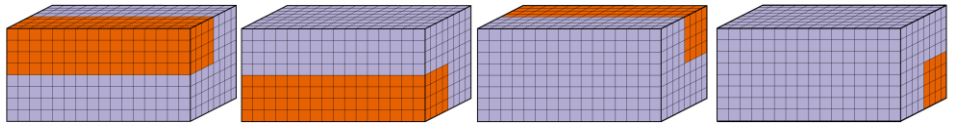
More Cross-Validation (CV)



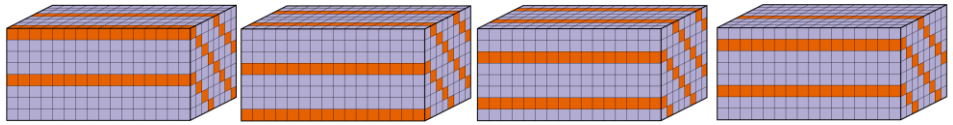
Spatial-block CV (CV-Sb)



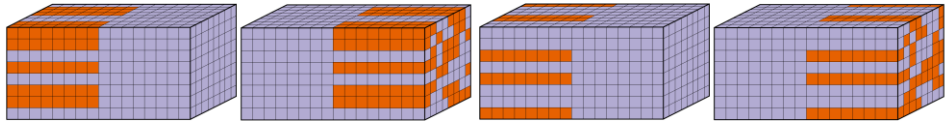
Contiguous spatial block CV (CV-Sb-cont)



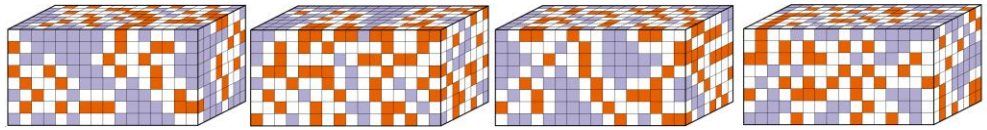
Systematic spatial block CV (CV-Sb-sys)



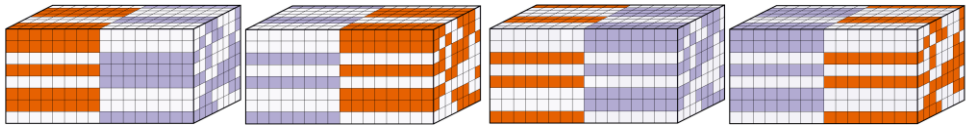
Spatio-temporal block CV (CV-STb)



Space-time buffered CV (CV-STbuf)

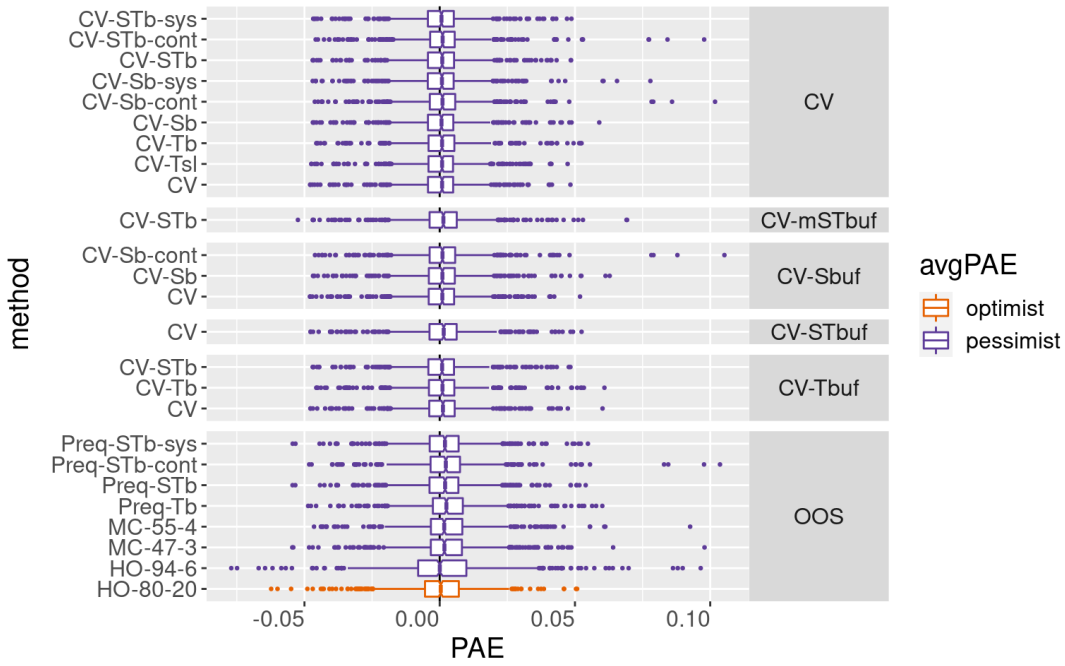


Leave location & time out CV (CV-mSTbuf)

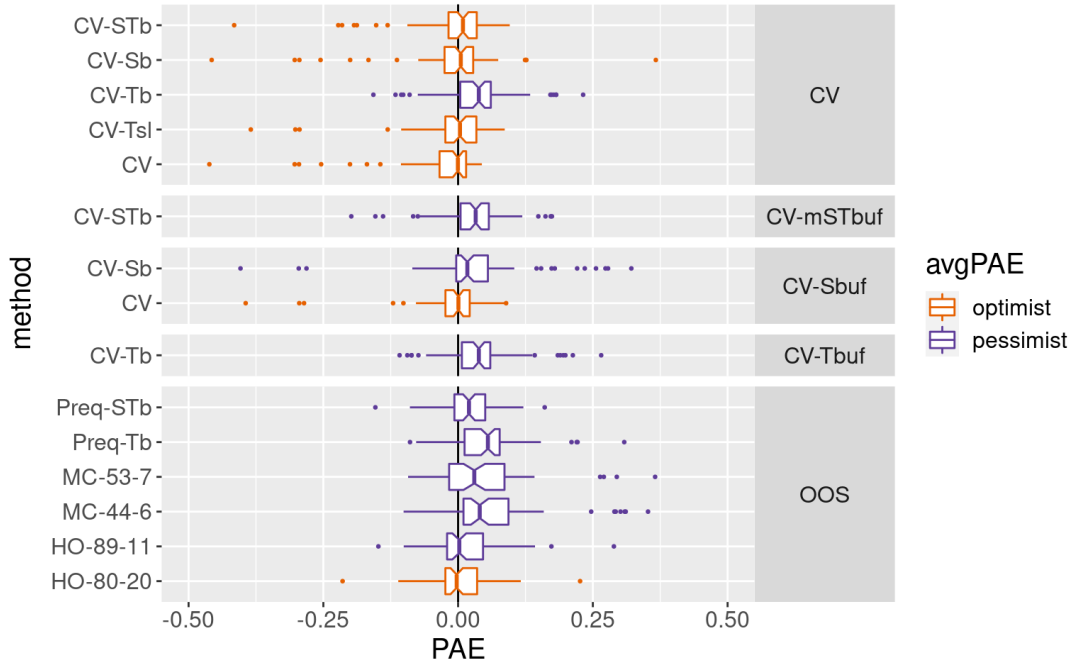


RESULTS: PREDICTIVE ACCURACY ERROR
PAE = ESTIMATED – GOLD
EVALUATING SPATIO-TEMPORAL FORECASTING

Artificial

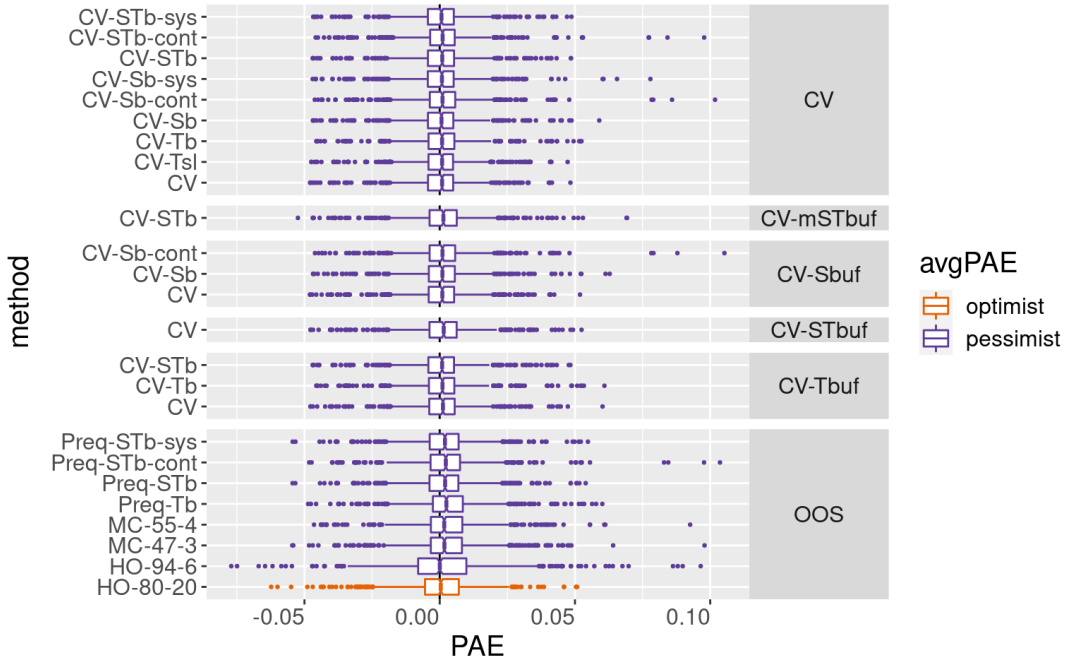


Real-world

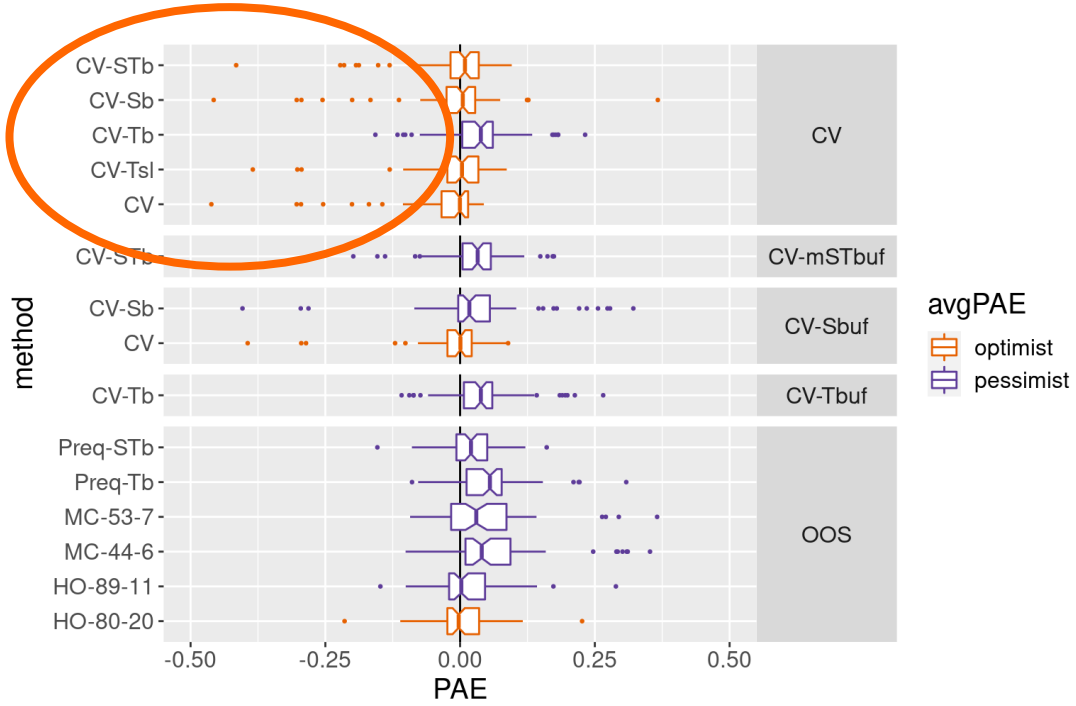


RESULTS: PREDICTIVE ACCURACY ERROR
PAE = ESTIMATED – GOLD
EVALUATING SPATIO-TEMPORAL FORECASTING

Artificial

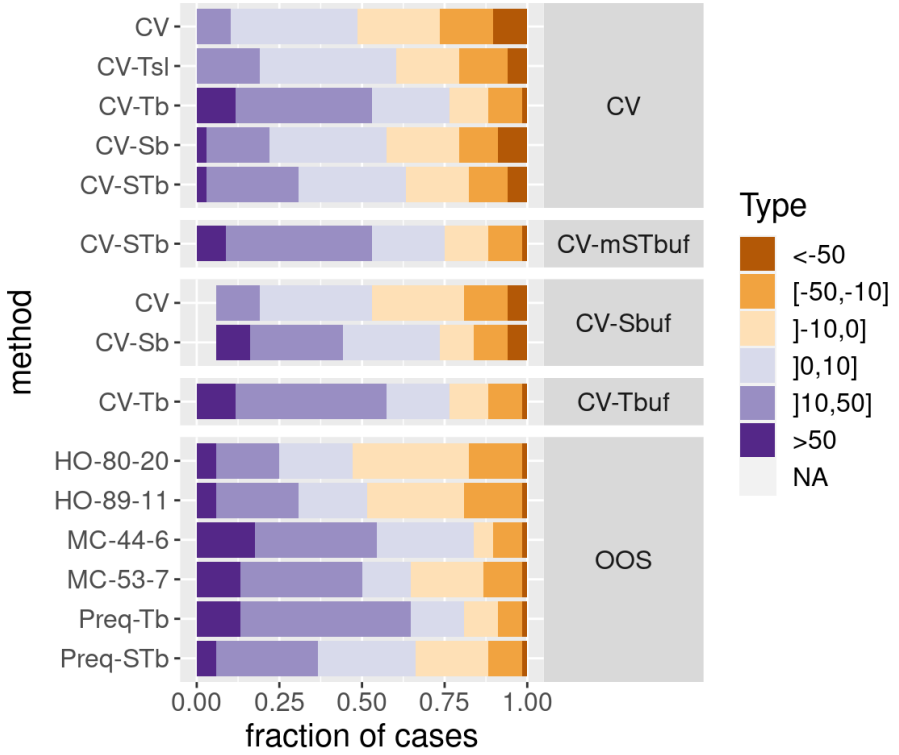


Real-world

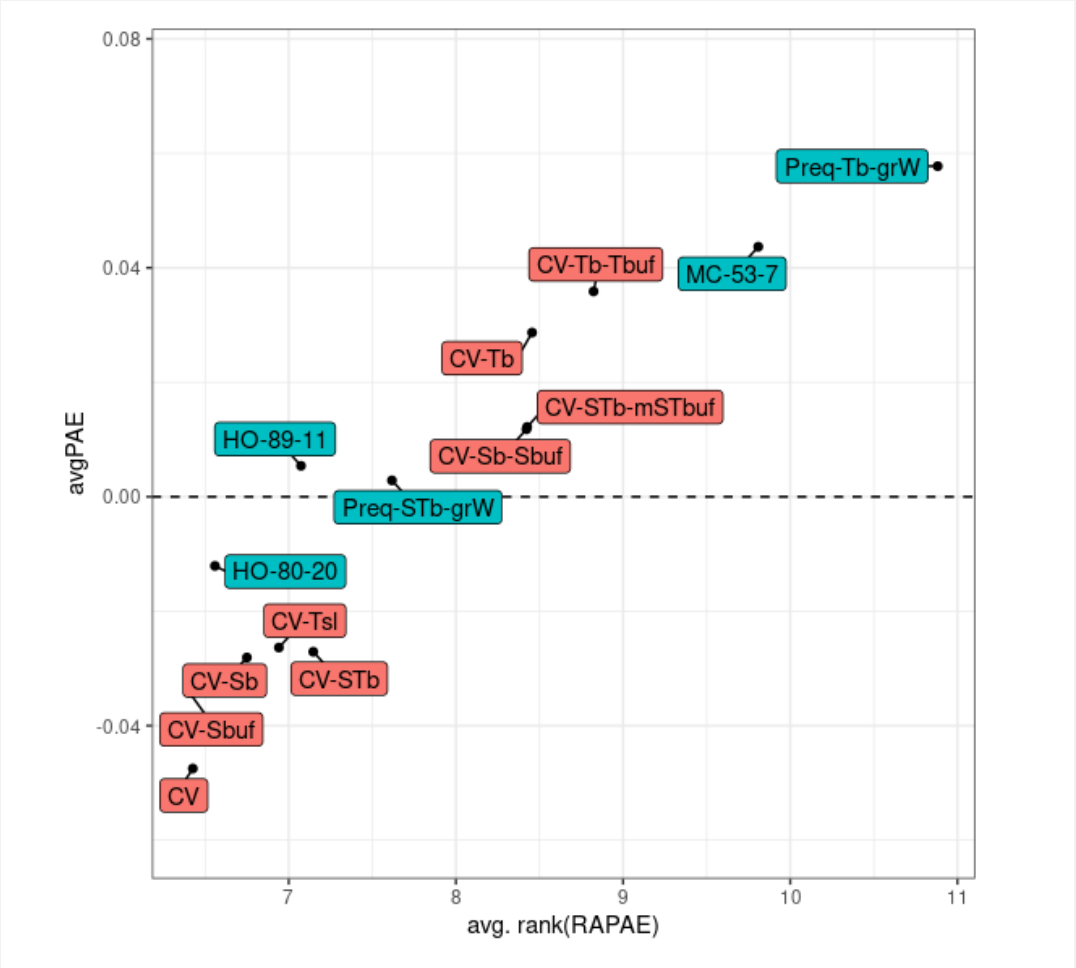


RESULTS: RELATIVE PREDICTIVE ACCURACY ERROR
RPAE = PAE/GOLD
EVALUATING SPATIO-TEMPORAL FORECASTING

Real-world



- **OOS** rarely highly optimistic, but more often severely over-estimate error
  - Only holdout tends to be optimistic more often
- **Standard CV** avoids highly pessimistic estimates, but has a higher rate of severe under-estimation
- **Temporal blocking** mitigates this



RESULTS: OPTIMISM VS ACCURACY

EVALUATING SPATIO-TEMPORAL FORECASTING

- Trade-off between:
  - average Absolute Predictive Accuracy Error
- and
  - average rank of Relative Absolute Predictive Accuracy Error

- Methods that produce **more accurate** estimates in absolute terms tend to suffer from severe error under-estimation, producing **overly optimistic** estimates

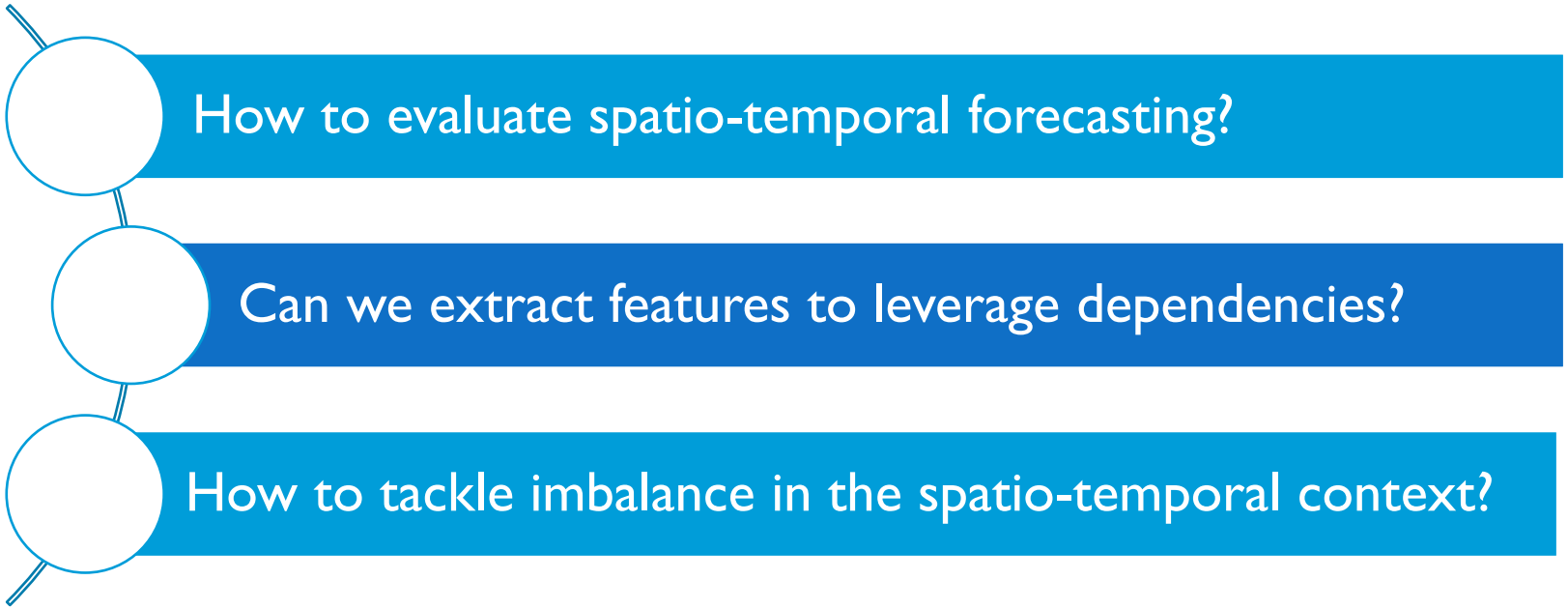
CONCLUSION
EVALUATING SPATIO-TEMPORAL FORECASTING

- **Spatio-temporal dependence** affects performance estimation
- **Empirical study** of 15+ estimation methods (artificial and real-world data)

- **Avoid** standard CV
  - Though reasonably accurate, it is often severely over-optimistic
  - Blocking data in time mitigates this
- OOS methods proved **competitive**
  - Less accurate and more pessimistic, but avoid severe error under-estimation
  - Inherently respect temporal order

- We will primarily use prequential temporal block evaluation (**Preq-Tb**) in the next sections

RESEARCH QUESTIONS



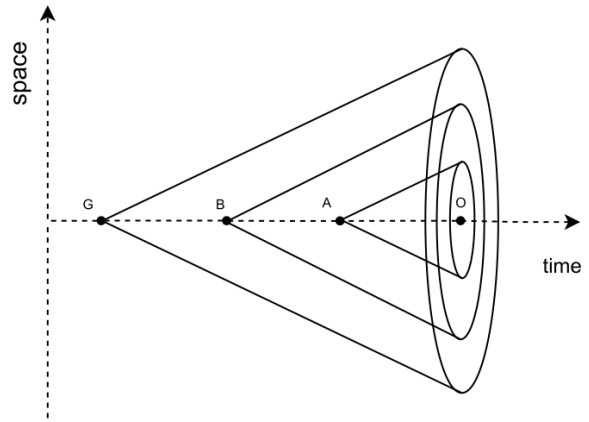
EXTRACTING SPATIO-TEMPORAL INDICATORS

SPATIO-TEMPORAL INDICATORS
EXTRACTING SPATIO-TEMPORAL INDICATORS

Ohashi & Torgo (2012)

Summary statistics on historical data from neighbourhoods within spatio-temporal distance

$$D_A = \alpha \cdot D_A^S + (1 - \alpha) \cdot D_A^T \leq \beta$$

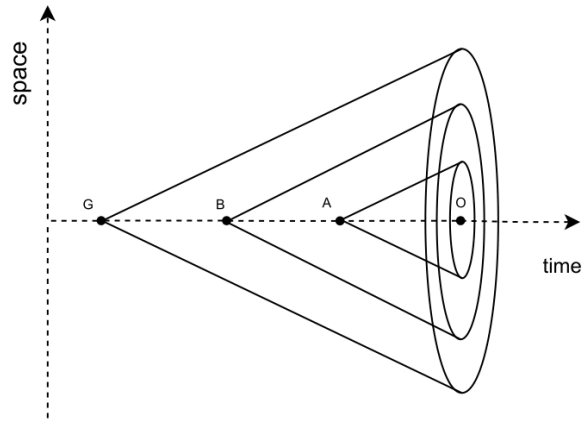


SPATIO-TEMPORAL INDICATORS
EXTRACTING SPATIO-TEMPORAL INDICATORS

Ohashi & Torgo (2012)

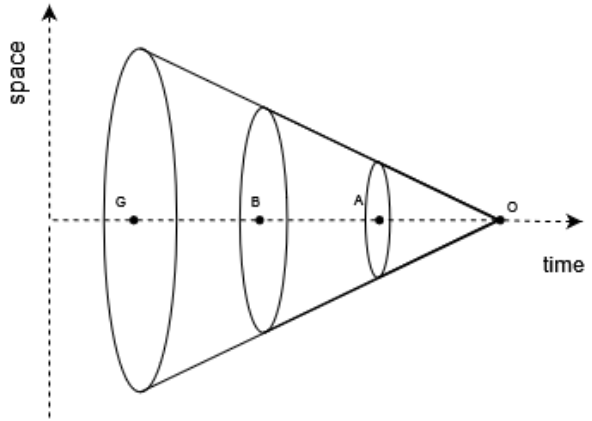
Summary statistics on historical data from neighbourhoods within spatio-temporal distance

$$D_A = \alpha \cdot D_A^S + (1 - \alpha) \cdot D_A^T \leq \beta$$



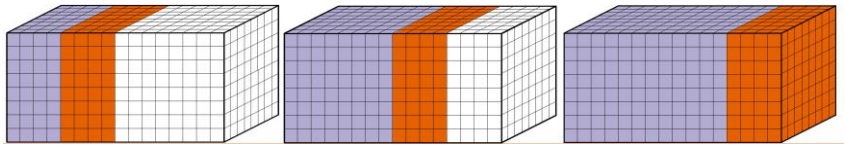
Our hypothesis

- Some processes take more time to "travel" between places
- **Can we improve performance by reversing the direction of the cone?**

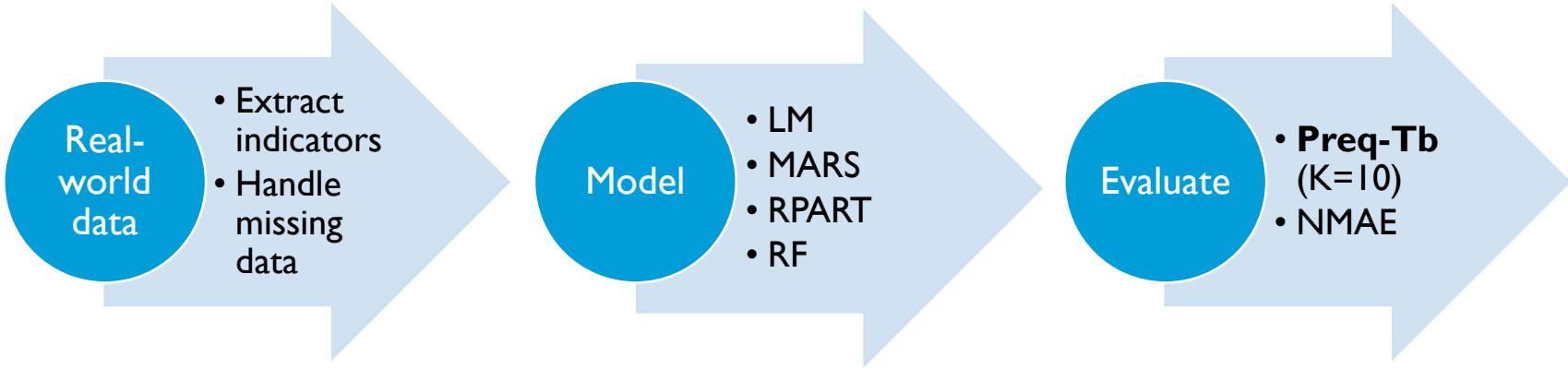


EXPERIMENTAL SETUP
EXTRACTING SPATIO-TEMPORAL INDICATORS

Parameter	Search space
Neighbourhood type	cone, reversed
Embed size	4, 8
Dimension weight α	0.1, 0.25, 0.5, 0.75, 0.9
Neighbourhood radii β	{0.01, 0.02, 0.03}, {0.02, 0.03, 0.04}



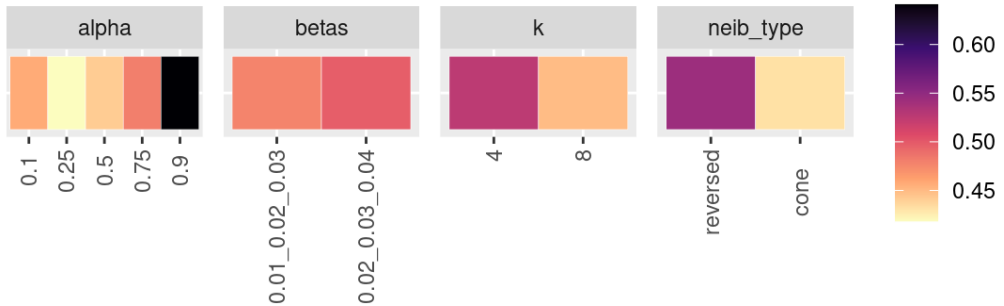
- Avoids severe error under-estimation
- Respects temporal order
- Prevents test-train spillover



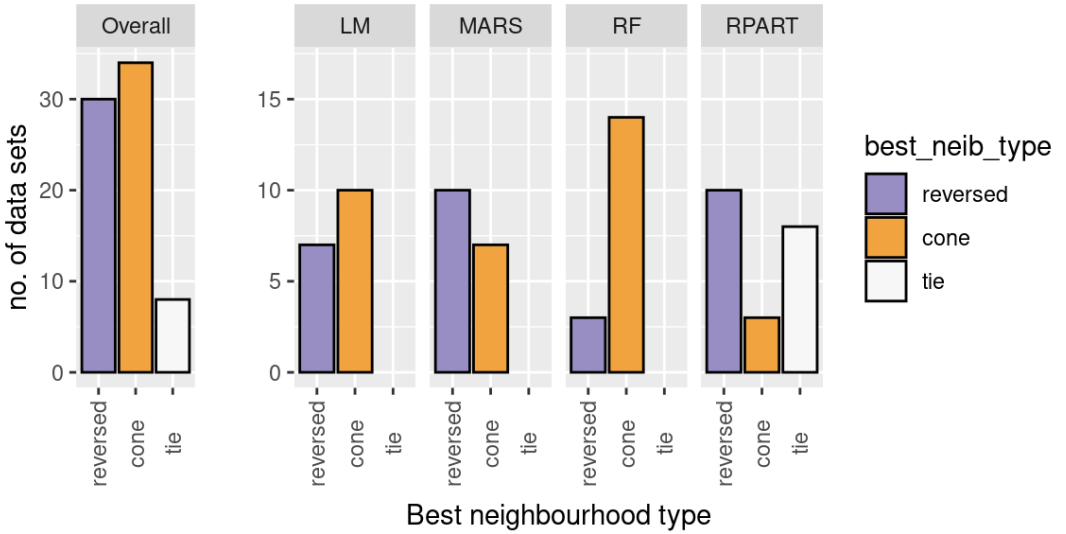
$$NMAE = \frac{\sum_{i=0}^{n} |\hat{z}_i - z_i|}{\sum_{i=0}^{n} |z_i - \bar{z}|}$$

RESULTS
EXTRACTING SPATIO-TEMPORAL INDICATORS

Average Rank

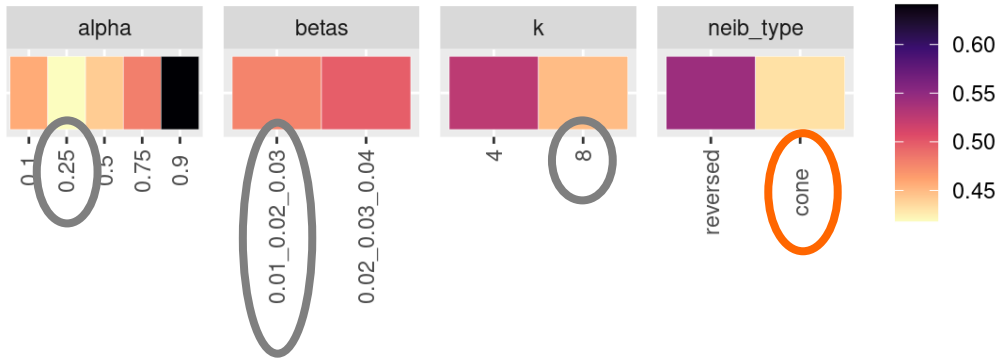


Best Performer

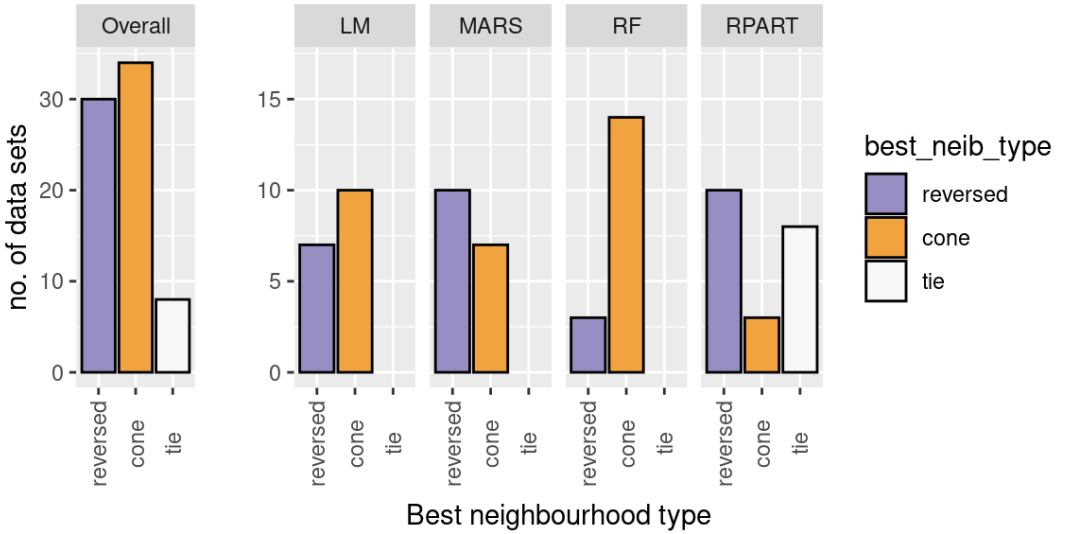


RESULTS
EXTRACTING SPATIO-TEMPORAL INDICATORS

Average Rank

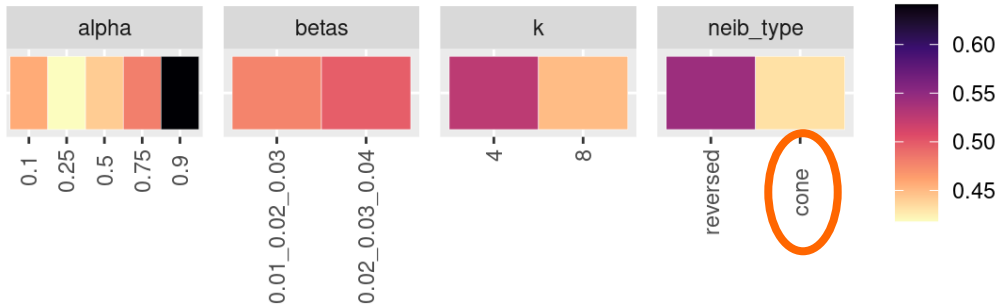


Best Performer

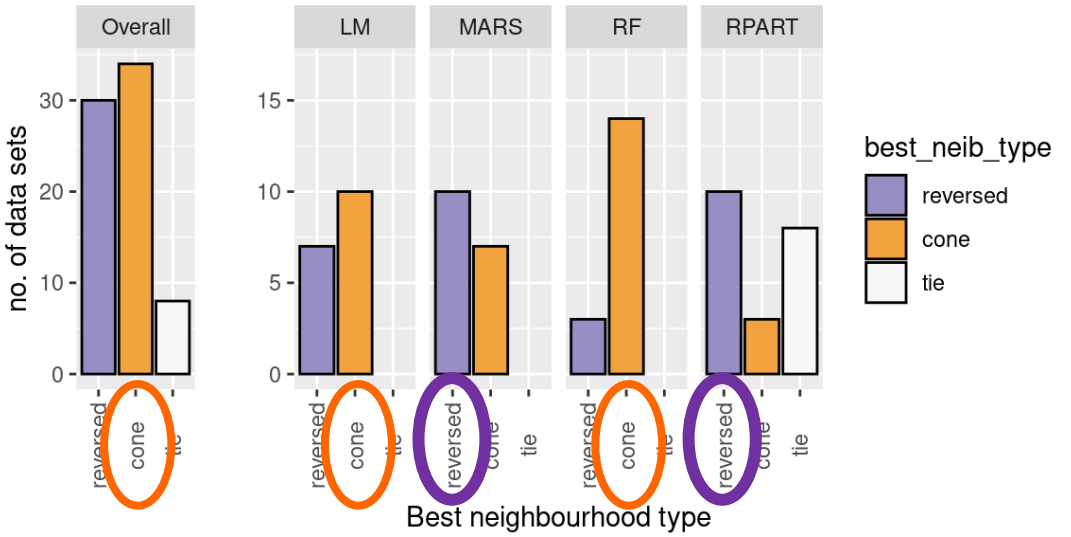


RESULTS
EXTRACTING SPATIO-TEMPORAL INDICATORS

Average Rank



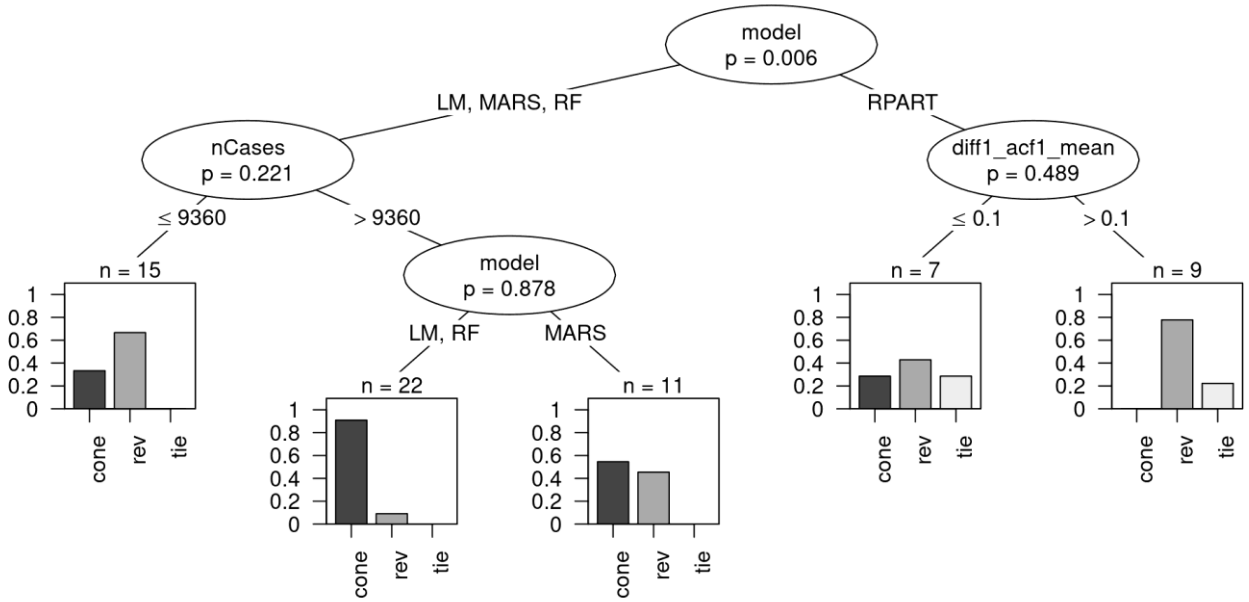
Best Performer



META-ANALYSIS: IMPACT OF DATA CHARACTERISTICS

EXTRACTING SPATIO-TEMPORAL INDICATORS

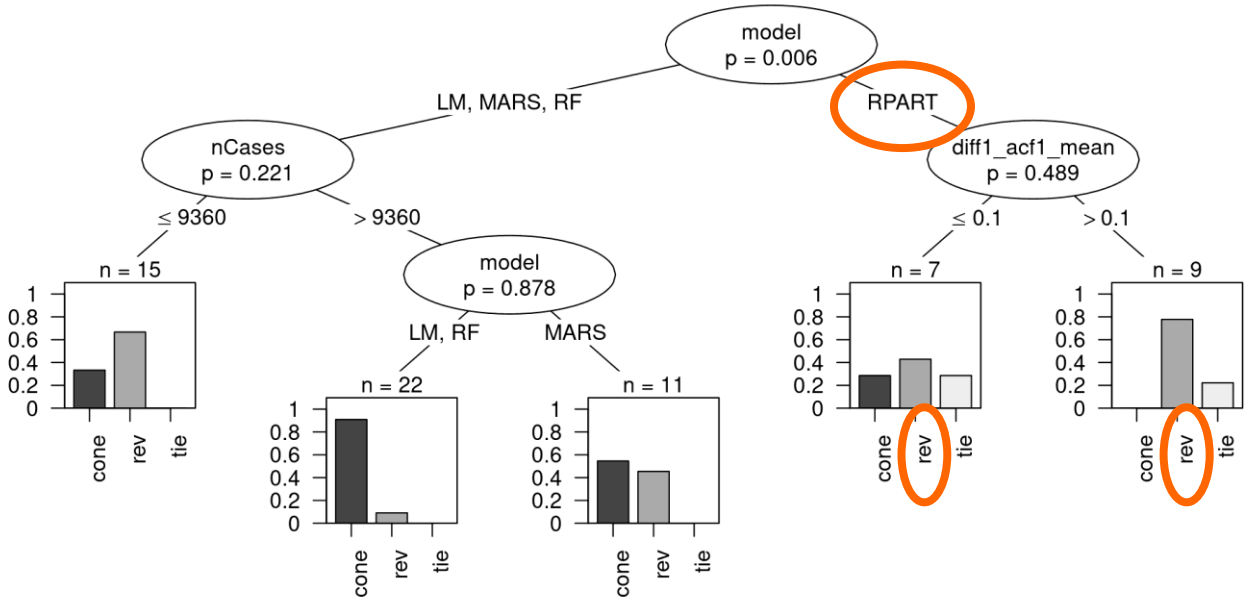
- Features describe data sets
  - Data size
  - Missing data
  - Target distribution characteristics
  - Spatial distribution of locations
  - Spatial/Temporal auto-correlation
- Meta-label: cone / reverse wins with "optimal" parameters
- Learn Conditional Inference Tree



META-ANALYSIS: IMPACT OF DATA CHARACTERISTICS

EXTRACTING SPATIO-TEMPORAL INDICATORS

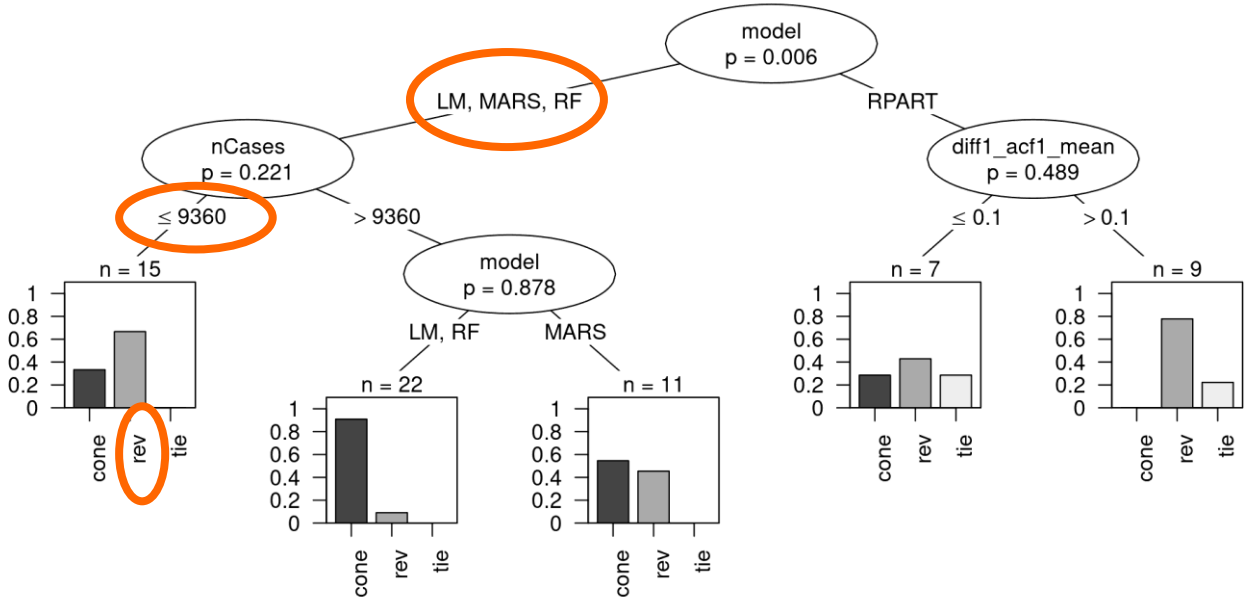
- Features describe data sets
  - Data size
  - Missing data
  - Target distribution characteristics
  - Spatial distribution of locations
  - Spatial/Temporal auto-correlation
- Meta-label: cone / reverse wins
- Learn Conditional Inference Tree



META-ANALYSIS: IMPACT OF DATA CHARACTERISTICS

EXTRACTING SPATIO-TEMPORAL INDICATORS

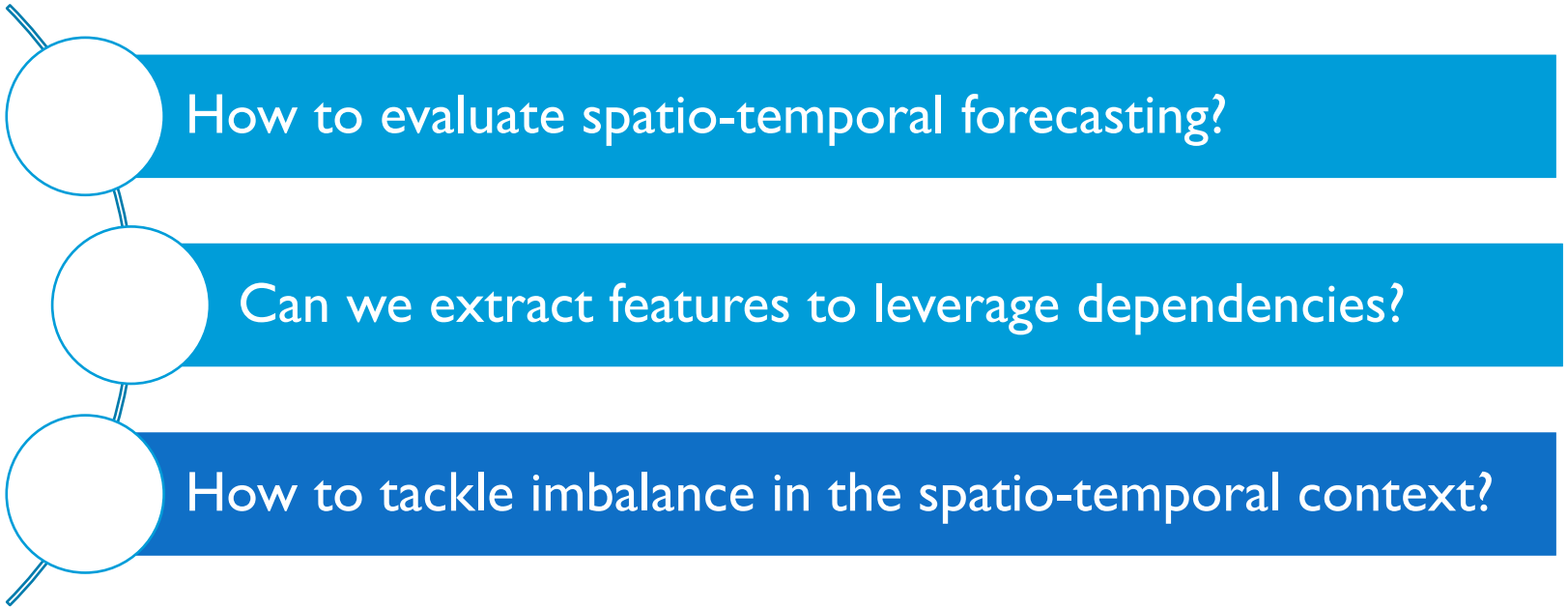
- Features describe data sets
  - Data size
  - Missing data
  - Target distribution characteristics
  - Spatial distribution of locations
  - Spatial/Temporal auto-correlation
- Meta-label: cone / reverse wins
- Learn Conditional Inference Tree



CONCLUSION
EXTRACTING SPATIO-TEMPORAL INDICATORS

- Extracted features from a conic spatio-temporal neighbourhood, reversing the direction of Ohashi & Torgo's 2012 proposal
- On average, features using the original proposal ranked better
- Depending on the parametrization, our proposal achieved the best results
  - Most often when using decision trees or MARS
  - When using linear regression on smaller data sets

RESEARCH QUESTIONS



RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

Mariana Oliveira, Nuno Moniz, Luís Torgo, and Vítor Santos Costa. Biased Resampling Strategies for Imbalanced Spatio-Temporal Forecasting. In IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 100–109. IEEE, 2019. doi:10.1109/dsaa.2019.00024

Mariana Oliveira, Nuno Moniz, Luís Torgo, and Vítor Santos Costa. Biased resampling strategies for imbalanced spatio-temporal forecasting. International Journal of Data Science and Analytics, 12(3):205–228, 2021. doi:10.1007/s41060-021-00256-2

SPATIO-TEMPORAL BIASED RESAMPLING

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

Motivation

- **Random resampling**: common approach to imbalance
- Spatial and temporal dependencies
- Strenghth of dependence along the dimensions may differ

Research Questions

- Can we improve performance by introducing a spatio-temporal **sampling bias**?
- Should we **weight the dimensions** differently?

PROPOSED RESAMPLING STRATEGIES
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

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

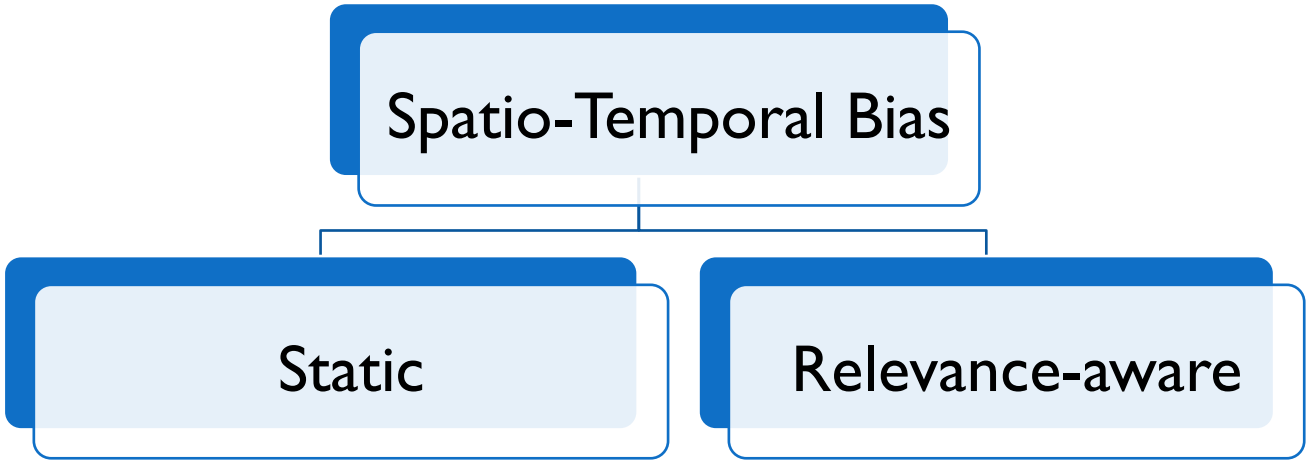
- Keep all extreme cases
- Keep only u% of normal cases, $0 < u < 100$ (with sampling bias)

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 BIAS

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

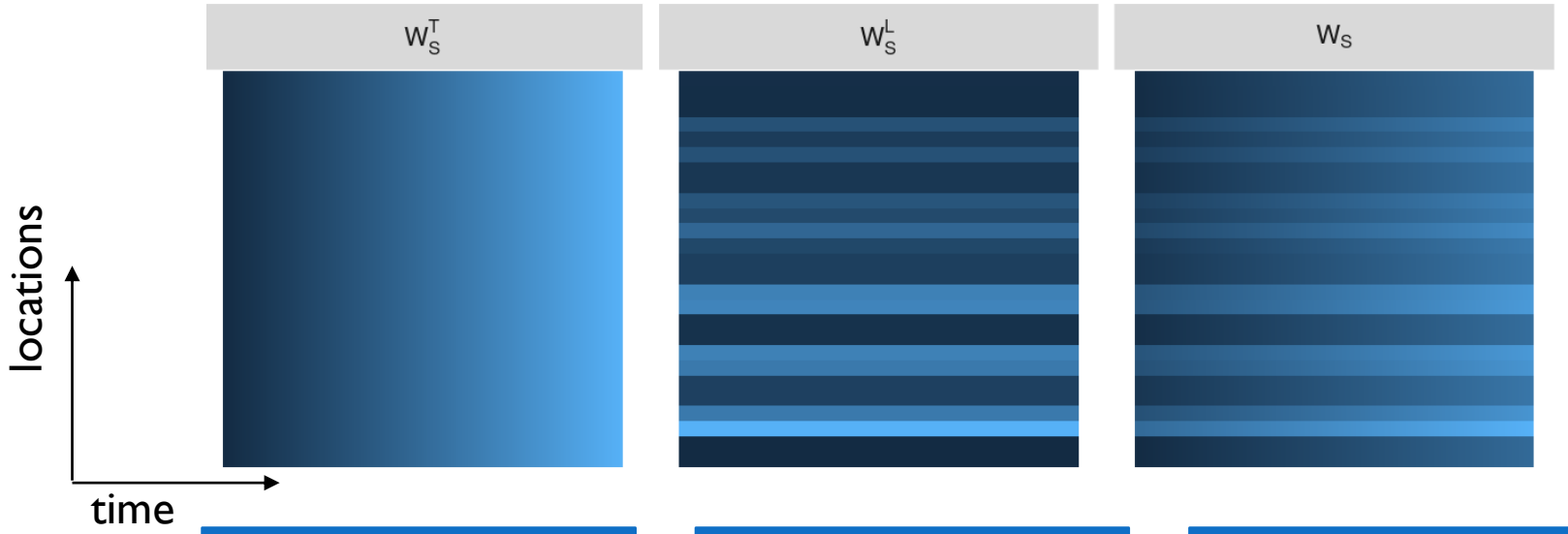


$$W^{S}_{i,j} = \alpha \times W^{T_S}_{i,j} + (1-\alpha) \times W^{L_S}_{i,j} + \epsilon$$

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

STATIC SPATIO-TEMPORAL BIAS

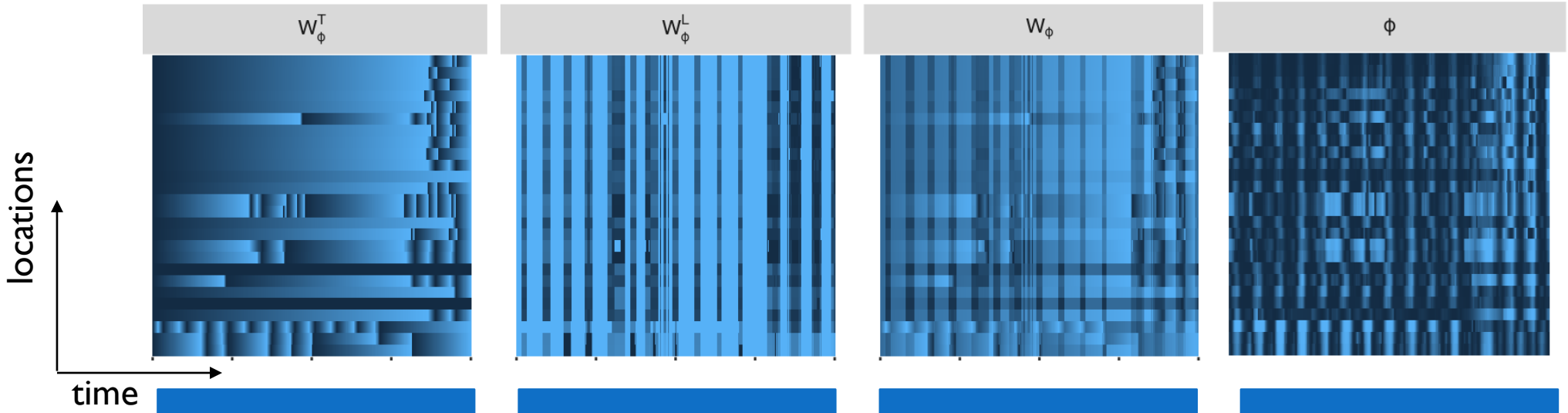
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



Temporal	Spatial	Spatio-temporal
• Keep/replicate more recent observations	• Keep/replicate spatially isolated observations	• Linear combination ($\alpha = 0.5$)

RELEVANCE-AWARE SPATIO-TEMPORAL BIAS

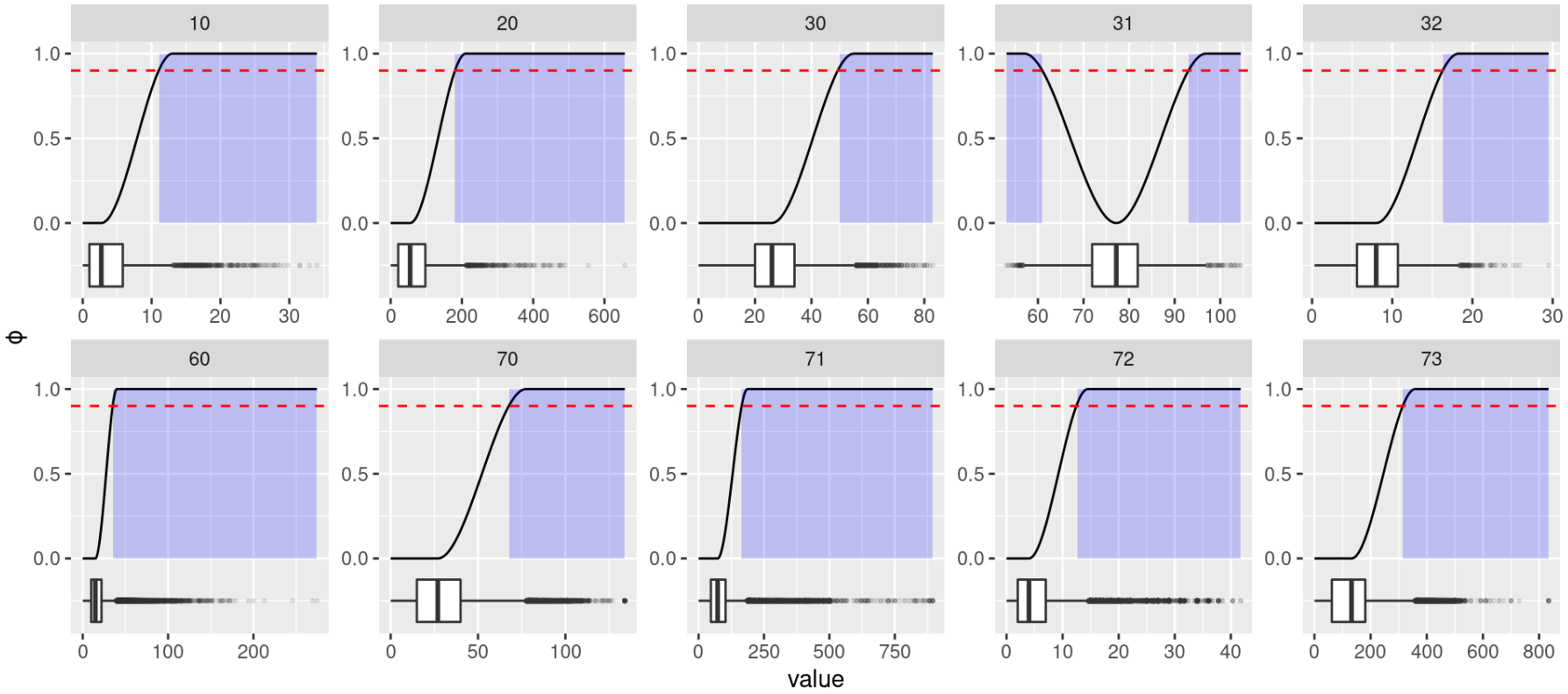
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



Temporal	Spatial	Spatio-temporal	Relevance
• Keep/replicate observations at the end/start of a normal/extreme sequence	• Replicate isolated extreme cases • Remove normal neighbours of extremes	• Linear combination ($\alpha = 0.5$)	• Derived from boxplot statistics

IMBALANCED REAL-WORLD DATA

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



Boxplots and derived relevance functions (ϕ)

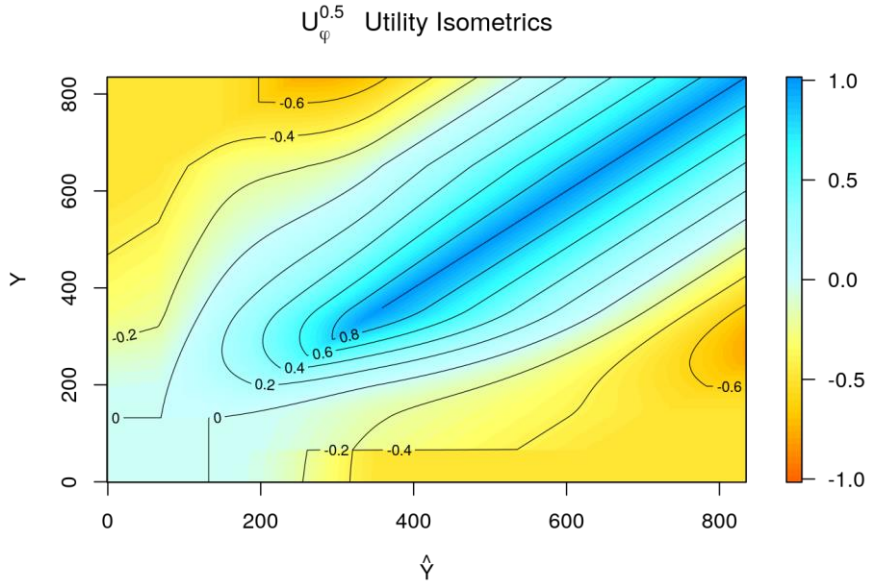
EXPERIMENTAL SETUP: EVALUATION METRICS
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

- **Utility-based** precision and recall for numeric prediction [Moniz et al. 2019]:

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

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

$$F_1^u = 2 \cdot \frac{prec_\phi^u \cdot rec_\phi^u}{prec_\phi^u + rec_\phi^u}$$



[Torgo & Ribeiro, 2007]

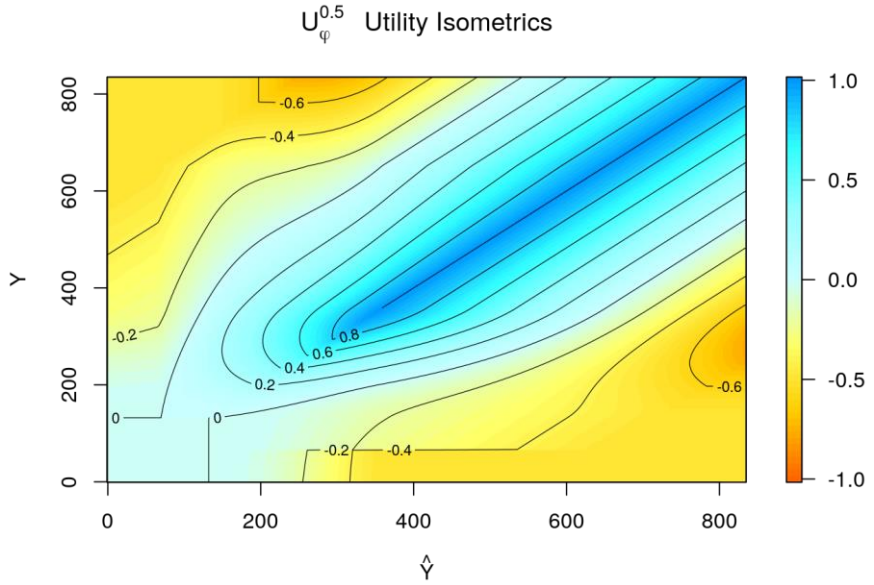
EXPERIMENTAL SETUP: EVALUATION METRICS
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

- **Utility-based** precision and recall for numeric prediction [Moniz et al. 2019]:

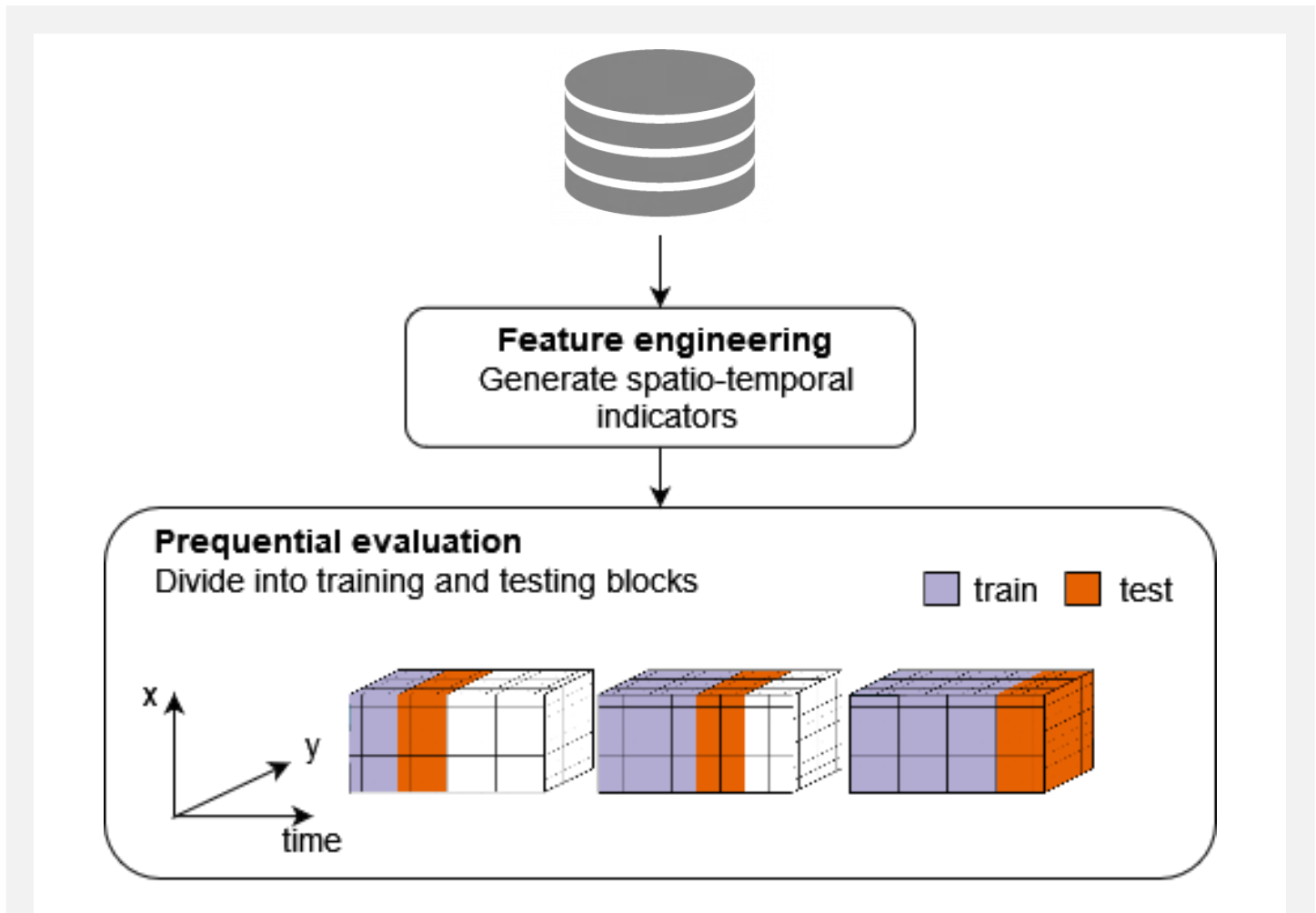
$$prec_{\phi}^{u} = \frac{\sum_{\phi(\hat{y}_i) \geq t_R, \phi(y_i) \geq t_R} (1 + u(\hat{y}_i, y_i))}{\sum_{\phi(\hat{y}_i) \geq t_R} (1 + \phi(\hat{y}_i))}$$

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

$$F_1^u = 2 \cdot \frac{prec_{\phi}^{u} \cdot rec_{\phi}^{u}}{prec_{\phi}^{u} + rec_{\phi}^{u}}$$



[Torgo & Ribeiro, 2007]



EXPERIMENTAL SETUP

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

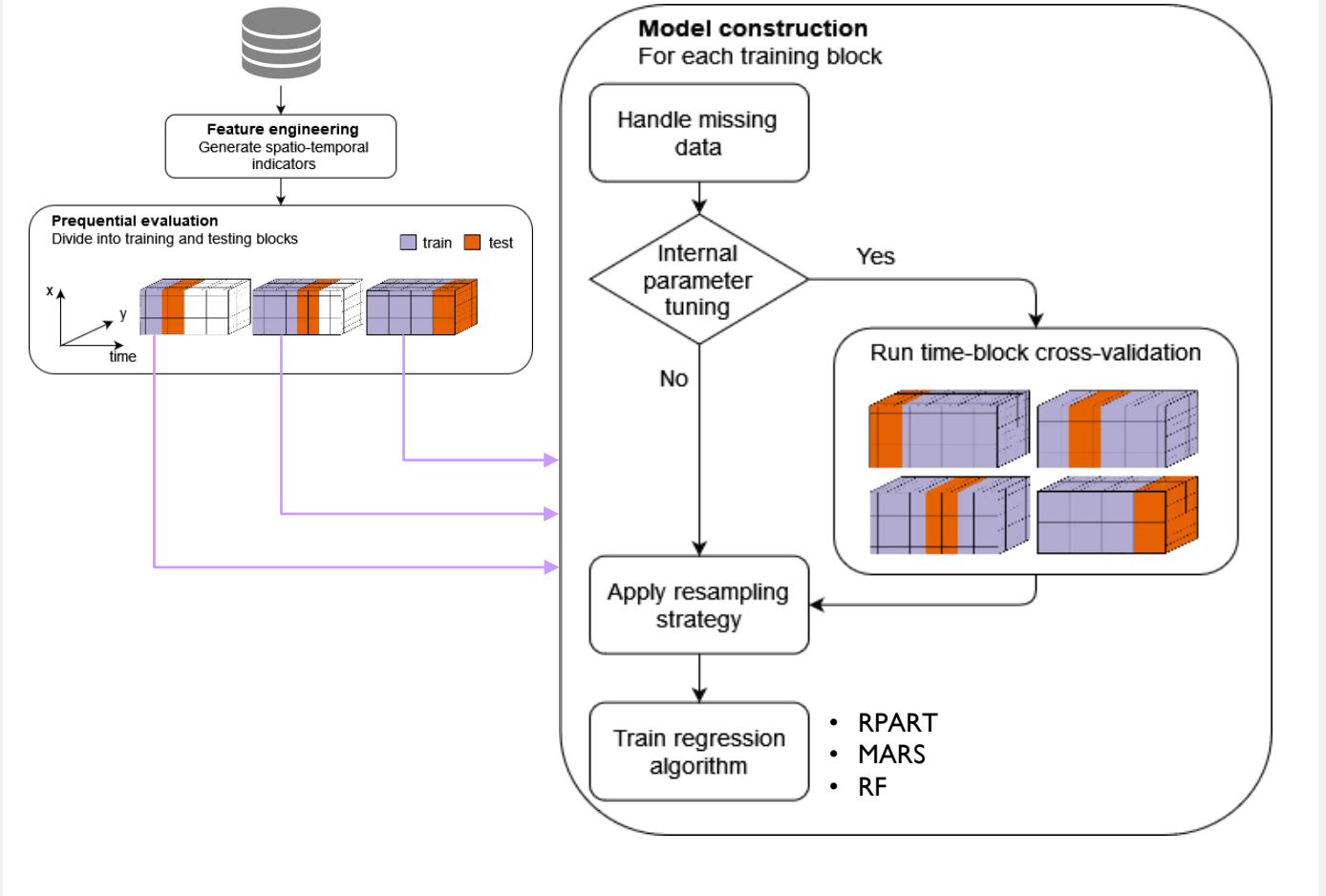
Preq-Tb

- Avoids severe error under-estimation
- Respects temporal order
- Prevents test-train spillover
- Includes data from every location in all training sets (we are interested in global extremes)

Process **repeated 10x** for reliability
(except computationally heavy RF)

EXPERIMENTAL SETUP

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



CV-Tb

- Similar to Preq-Tb, but uses the whole data set (potentially important on small training sets)
- Mitigates severe error under-estimation of standard CV
- Does not completely ignore temporal order

EXPERIMENTAL SETUP: PARAMETRIZATION
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

Internal tuning	Fixed	Optimal
• Internal CV-Tb to select best parameters per training block	• Same set of parameters selected **a priori** for whole data set $\alpha = 0.5$, $u = 0.6$, $o = 2$	• Same set of parameters selected **a posteriori** for whole data set • Parameters with best F_1^u per data-model pair

Parameter	Search space
Dimension weight α	0, 0.25, 0.5, 0.75, 1
Under-sampling *u*%	0.2, 0.4, 0.6, 0.8, 0.95
Over-sampling *o*%	0.5, 1, 2, 3, 4

EXPERIMENTAL SETUP: PARAMETRIZATION
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

Internal tuning

- Internal CV-Tb to select best parameters per training block

Most reliable

Fixed

- Same set of parameters selected **a priori** for whole data set

  $\alpha = 0.5$,
  $u = 0.6$,
  $o = 2$

Show effects of using "defaults"

Optimal

- Same set of parameters selected **a posteriori** for whole data set
- Parameters with best F_1^u per data-model pair

Show full potential

Parameter	Search space
Dimension weight α	0, 0.25, 0.5, 0.75, 1
Under-sampling *u*%	0.2, 0.4, 0.6, 0.8, 0.95
Over-sampling *o*%	0.5, 1, 2, 3, 4

RESULTS: F_1^U–SCORE AVERAGE RANK
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

Average Rank

Method	tuning	optimal	fixed
None	**6.4**	**6.9**	**6.2**
ROS	4.4	4.7	3.7
$STROS_S$	3.8	3.4	3.5
STROSφ	4.8	4.2	4.7
RUS	3.4	4.3	3.7
$STRUS_S$	3.1	2.7	3.8
STRUSφ	2.0	1.7	2.4

- **All forms of resampling improve results**
- One type of bias (or both) always able to improve random resampling
- Relevance-aware under-sampling worked best for every parametrization

RESULTS: F_1^U–SCORE AVERAGE RANK
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

Average Rank

Method	tuning	optimal	fixed
None	6.4	6.9	6.2
ROS	4.4	4.7	3.7
STROS$_S$	**3.8**	**3.4**	**3.5**
STROS$_\varphi$	4.8	**4.2**	4.7
RUS	3.4	4.3	3.7
STRUS$_S$	**3.1**	**2.7**	3.8
STRUS$_\varphi$	**2.0**	**1.7**	**2.4**

- All forms of resampling improve results
- **One type of bias (or both) always able to improve random resampling**
- Relevance-aware under-sampling worked best for every parametrization

RESULTS: F_1^U–SCORE AVERAGE RANK
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

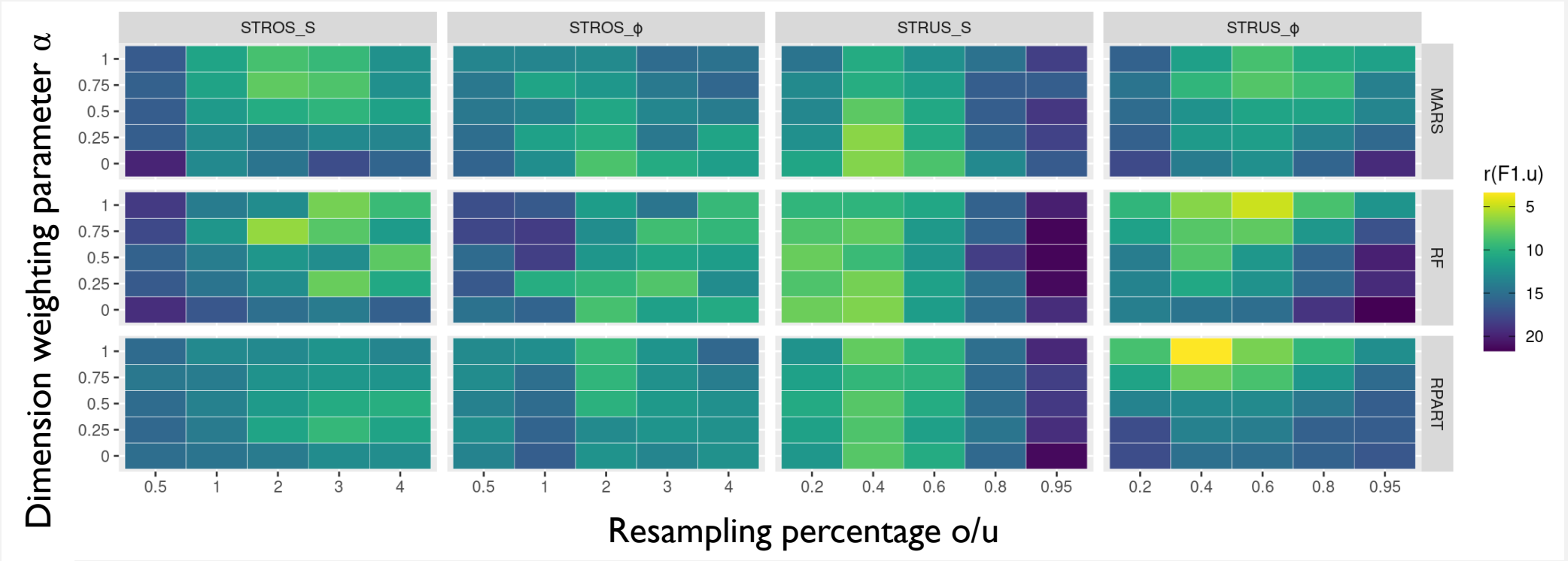
Average Rank

Method	tuning	optimal	fixed
None	6.4	6.9	6.2
ROS	4.4	4.7	3.7
$STROS_S$	3.8	3.4	3.5
STROSφ	4.8	4.2	4.7
RUS	3.4	4.3	3.7
$STRUS_S$	3.1	2.7	3.8
STRUSφ	**2.0**	**1.7**	**2.4**

- All forms of resampling improve results
- One type of bias (or both) always able to improve random resampling
- **Relevance-aware under-sampling worked best for every parametrization**

RESULTS: PARAMETER SENSITIVITY

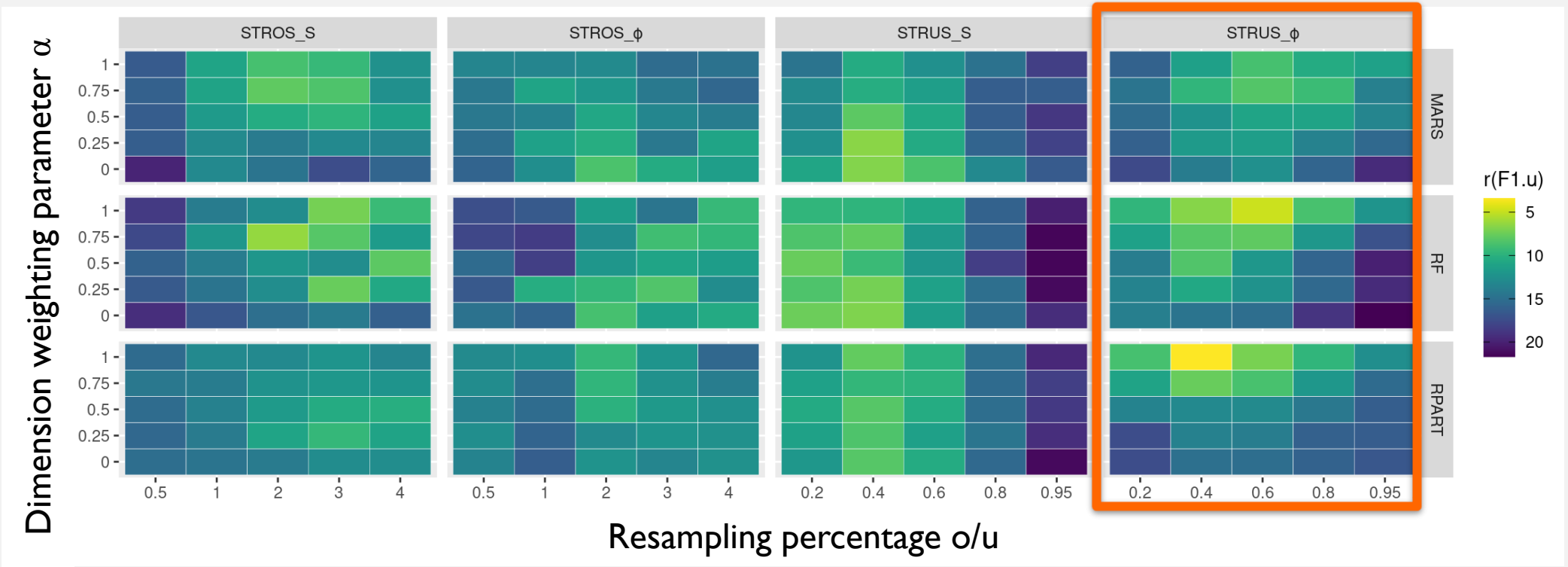
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- Most strategies more sensitive to changes in **resampling percentage** (↔) than in **α** (↕)
  - Exception is STRUS φ
- Relevance-aware and over-sampling strategies more stable than their counterparts
- Fairly consistent across models

RESULTS: PARAMETER SENSITIVITY

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- Most strategies more sensitive to changes in **resampling percentage** (↔) than in **α** (↕)
  - **Exception is STRUS φ**
- Relevance-aware and over-sampling strategies more stable than their counterparts
- Fairly consistent across models

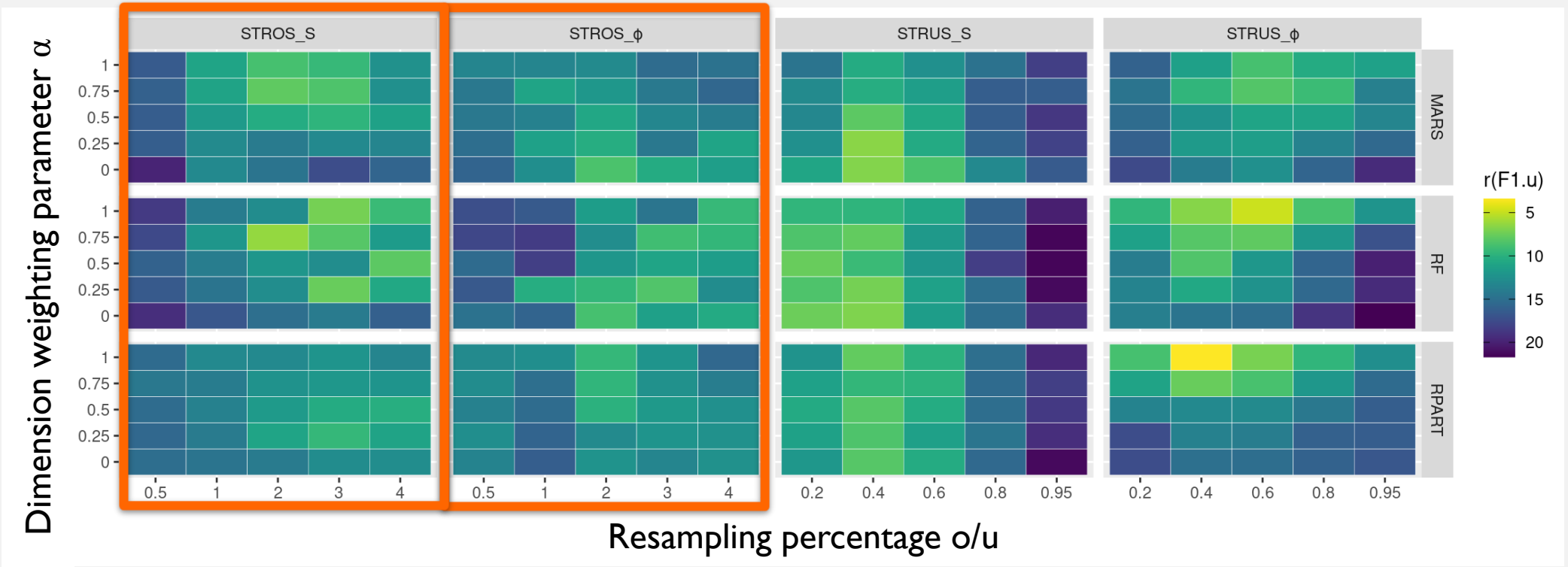
RESULTS: PARAMETER SENSITIVITY

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

- Most strategies more sensitive to changes in resampling percentage (↔) than in α (↕)
  - Exception is STRUS φ
- **Relevance-aware** and over-sampling strategies **more stable** than their counterparts
- Fairly consistent across models

RESULTS: PARAMETER SENSITIVITY

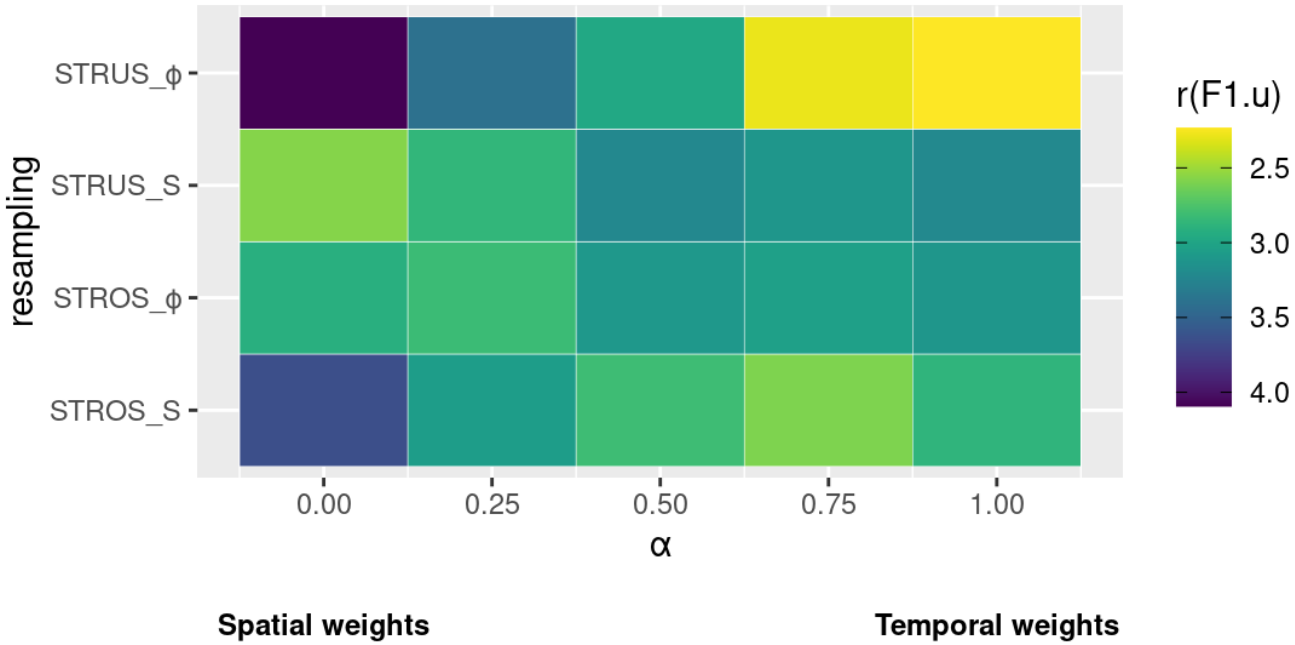
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- Most strategies more sensitive to changes in resampling percentage (↔) than in α (↕)
  - Exception is STRUS φ
- Relevance-aware and **over-sampling** strategies **more stable** than their counterparts
- Fairly consistent across models

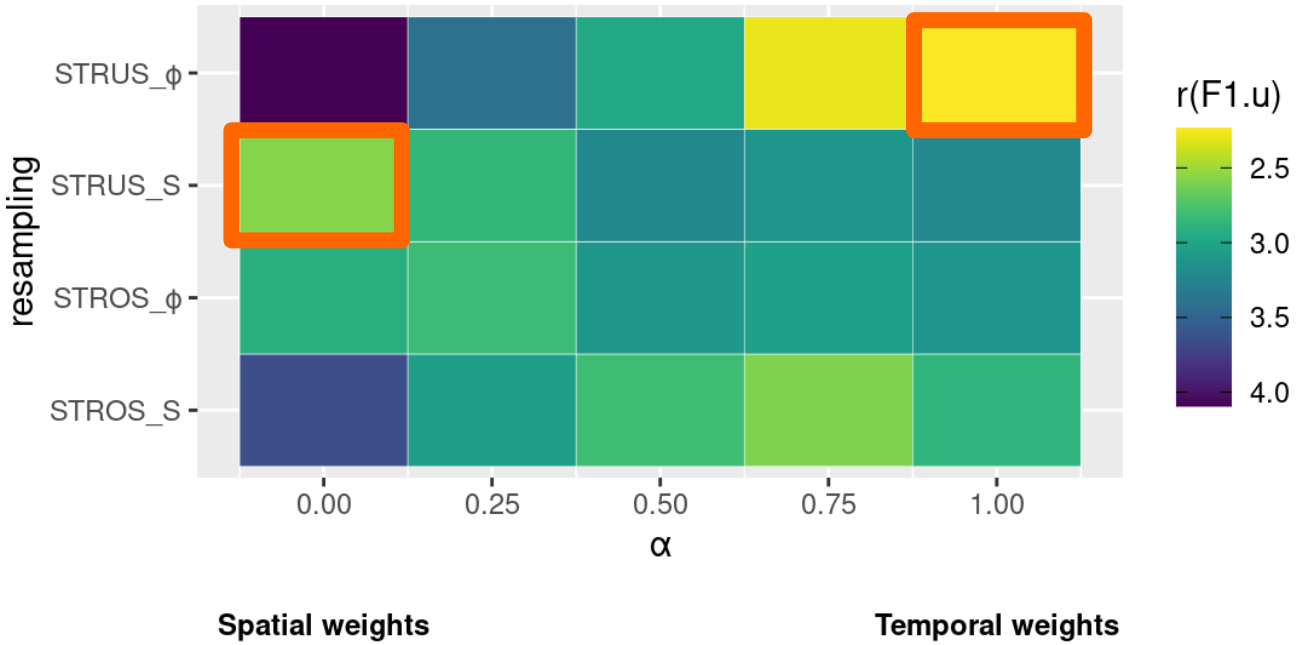
RESULTS: SENSITIVITY TO α

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



RESULTS: SENSITIVITY TO α

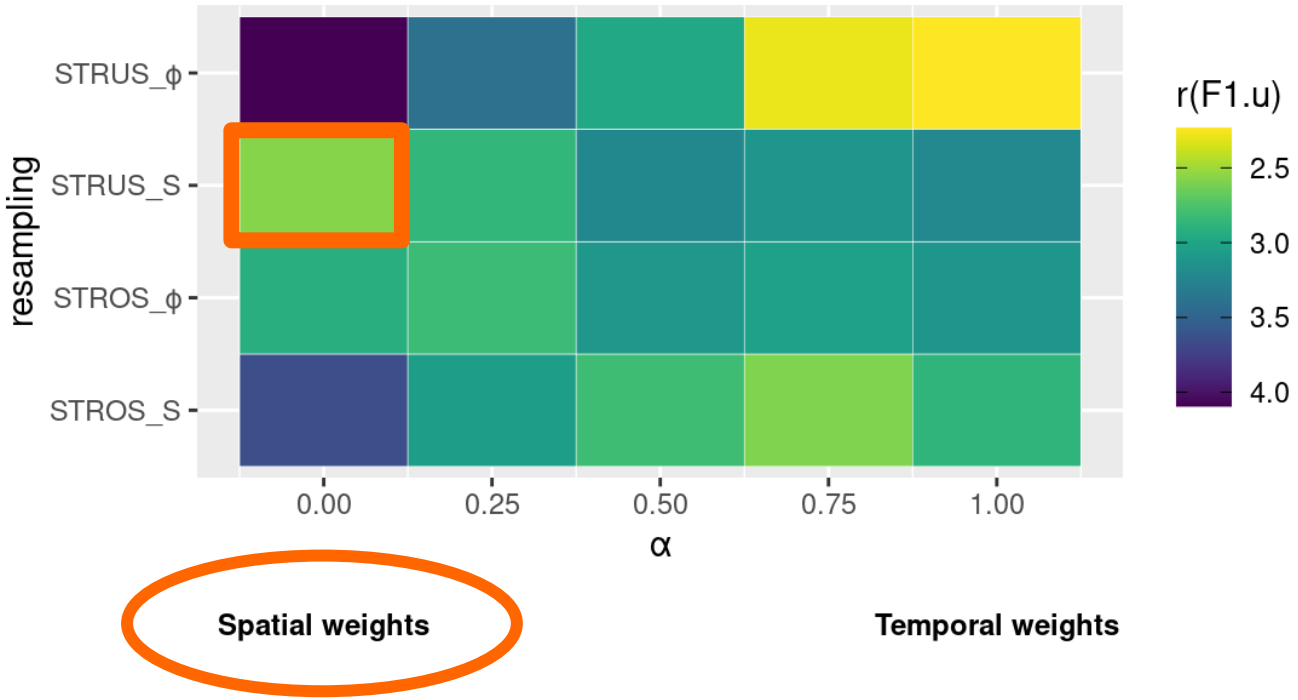
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- **Under-sampling** strategies worked best by using **"pure" weights**

RESULTS: SENSITIVITY TO α

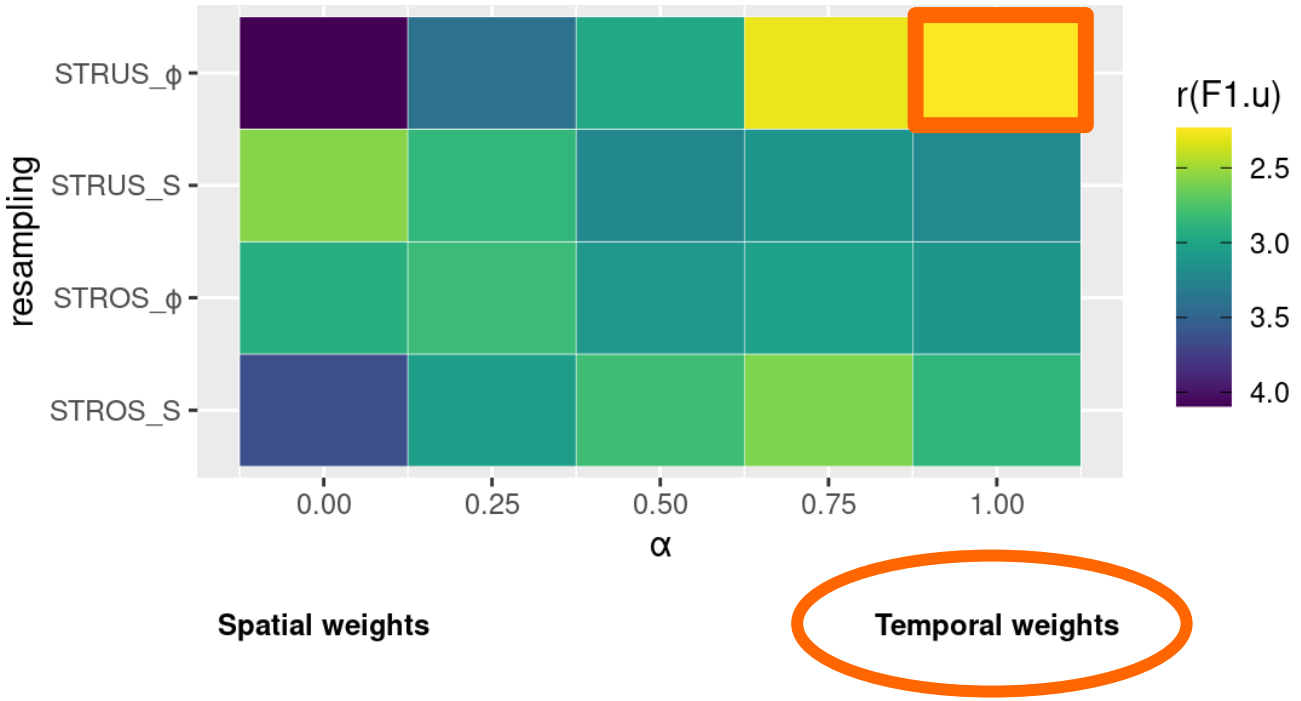
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- Under-sampling strategies worked best by using "pure" weights
  - **static bias using spatial weights**

RESULTS: SENSITIVITY TO α

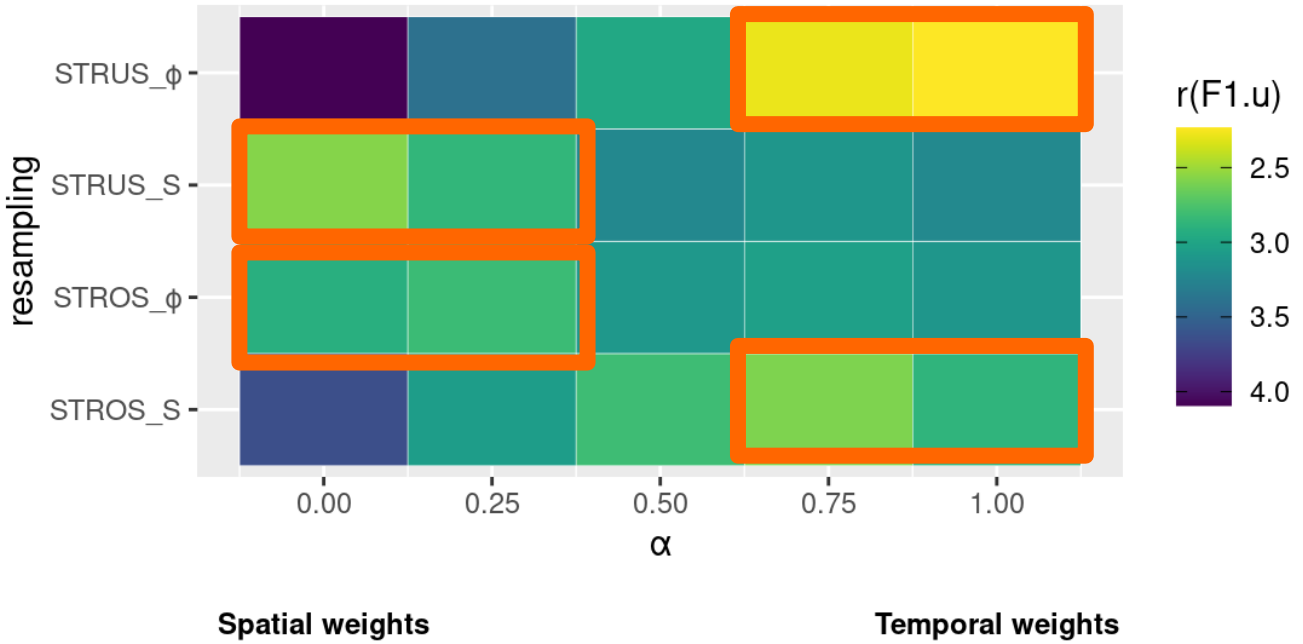
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- Under-sampling strategies worked best by using "pure" weights
  - static bias using spatial weights
  - **relevance-aware using temporal weights**

RESULTS: SENSITIVITY TO α

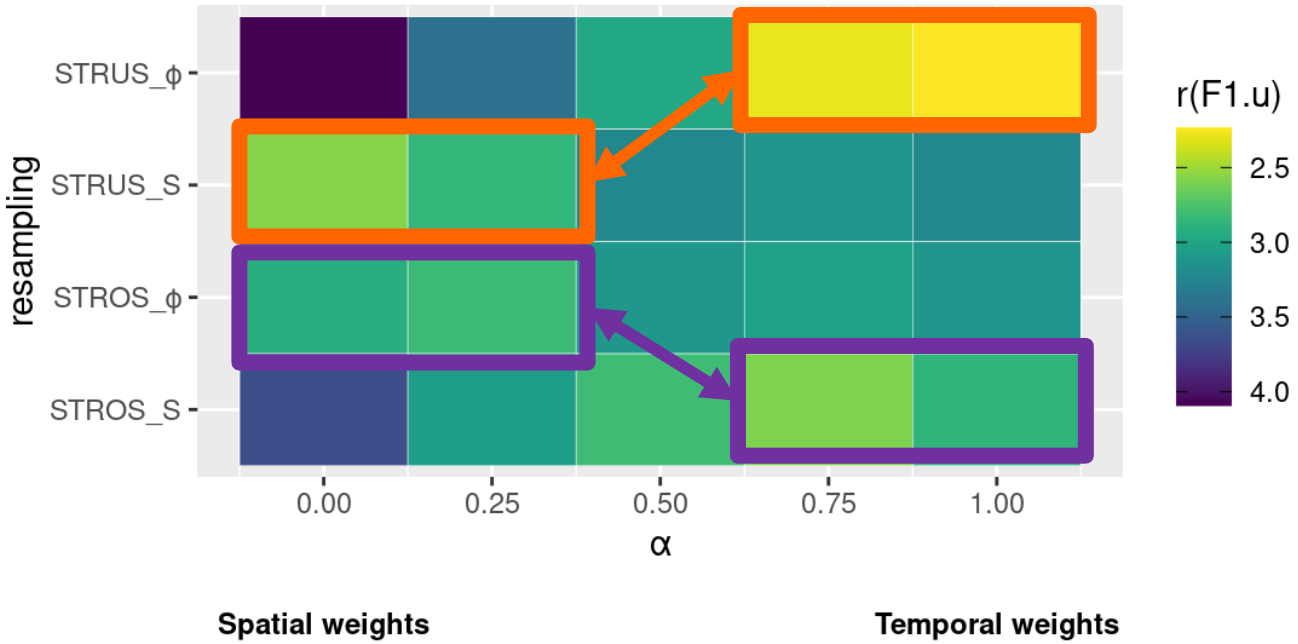
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- Under-sampling strategies worked best by using "pure" weights
  - static bias using spatial weights
  - relevance-aware using temporal weights
- **Combining weights also works well**

RESULTS: SENSITIVITY TO α

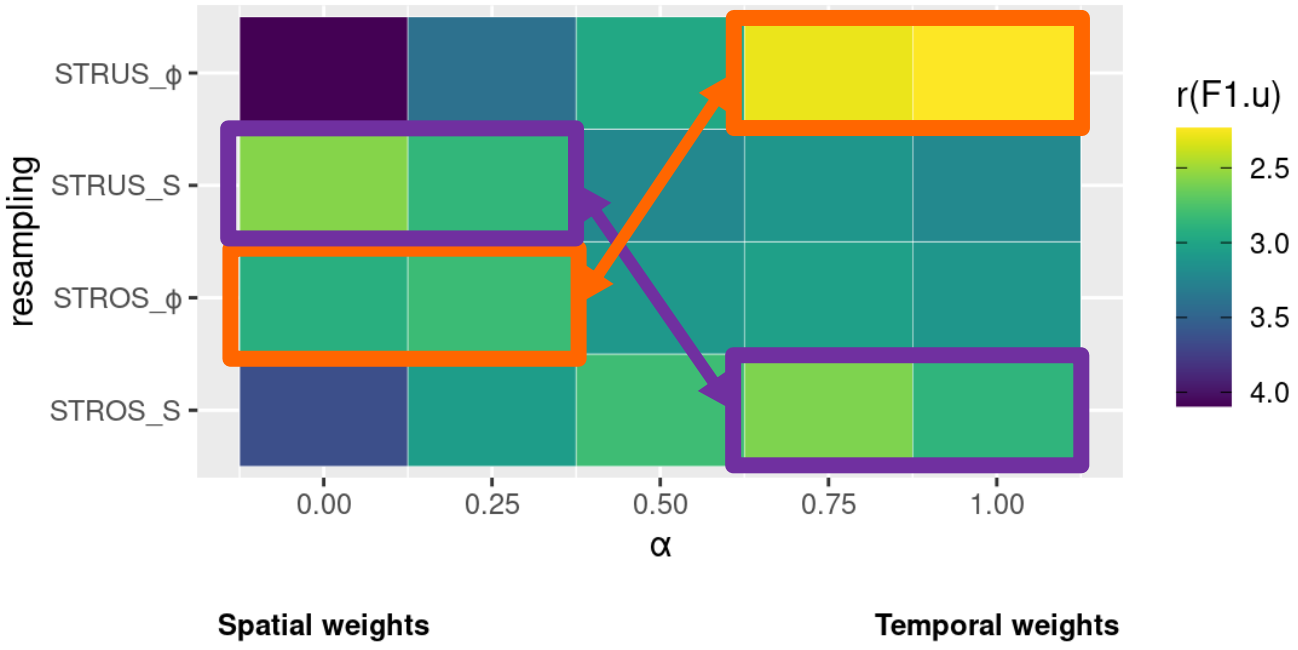
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA



- Under-sampling strategies worked best by using "pure" weights
  - static bias using spatial weights
  - relevance-aware using temporal weights
- Combining weights also works well
- **Under- and over-sampling show opposite tendencies when using the same bias**

RESULTS: SENSITIVITY TO α

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

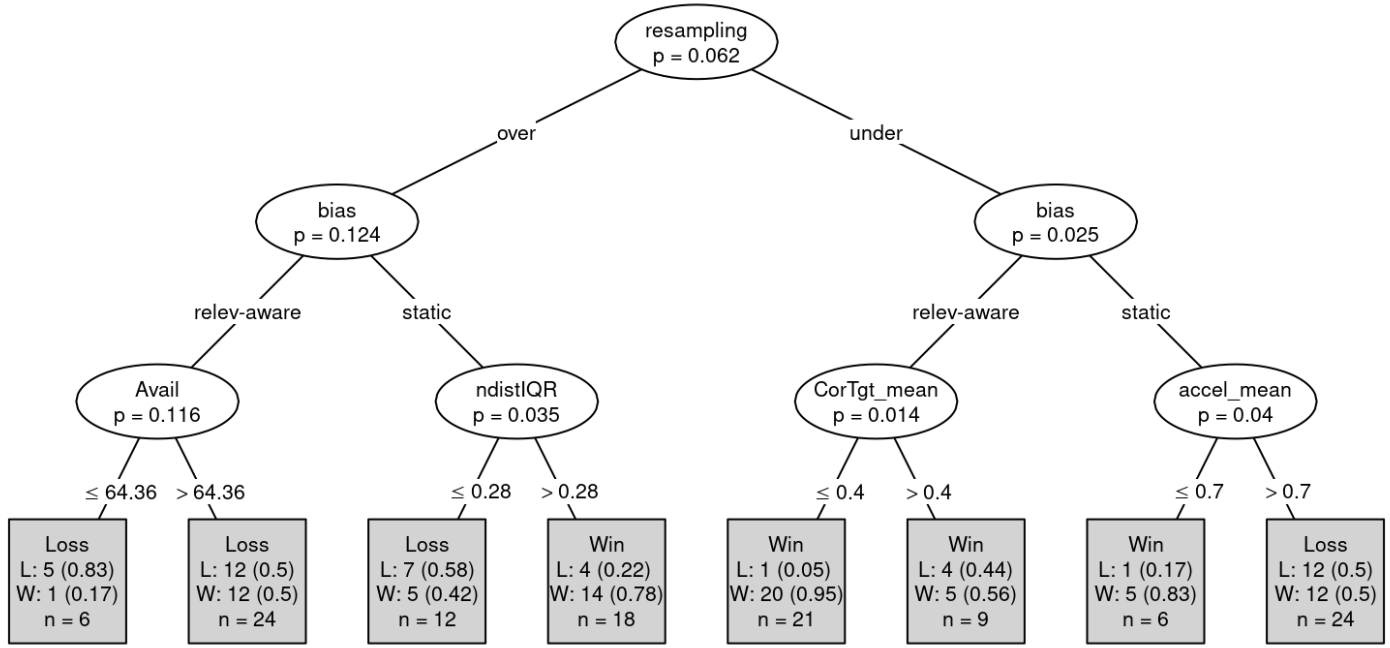


- Under-sampling strategies worked best by using "pure" weights
  - static bias using spatial weights
  - relevance-aware using temporal weights
- Combining weights also works well
- Under- and over-sampling show opposite tendencies when using the same bias
  - **Same applies to relevance-aware and static resampling**

META-ANALYSIS: IMPACT OF DATA CHARACTERISTICS

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

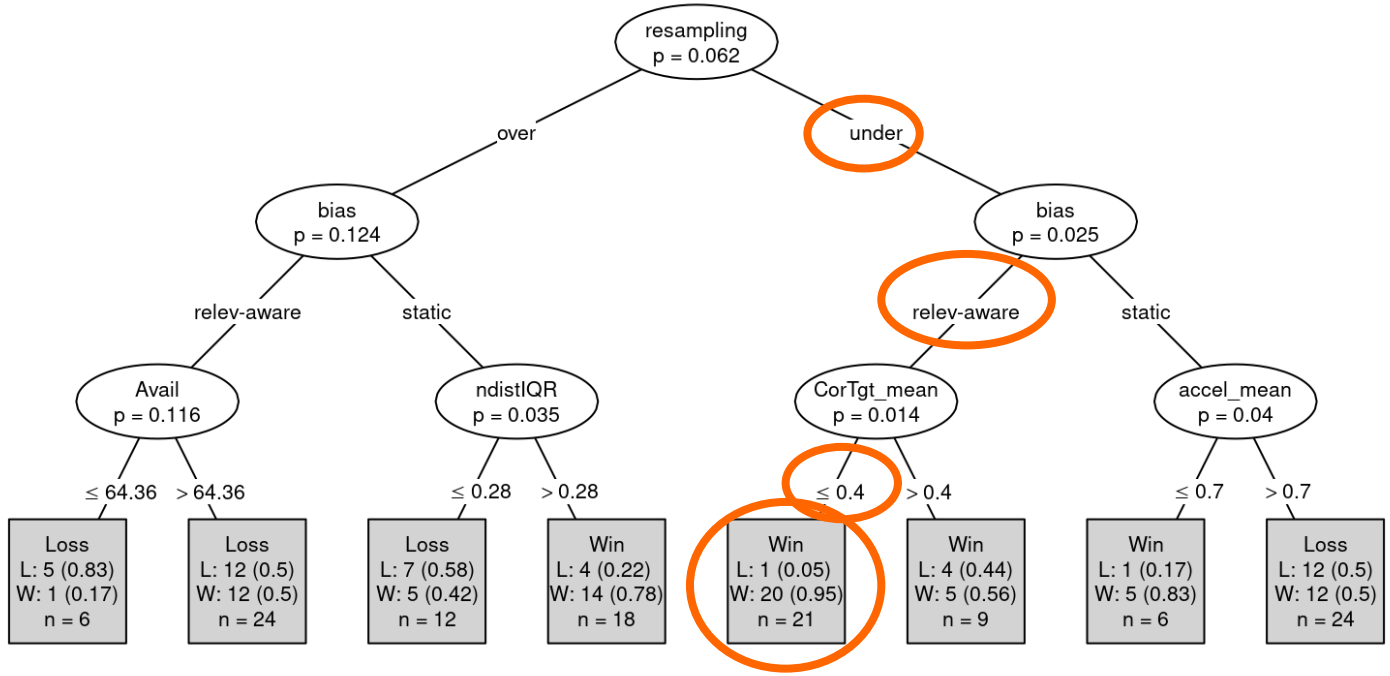
- Features describe data sets
  - Data size
  - Missing data
  - Target distribution characteristics
  - Spatial distribution of locations
  - Spatial/Temporal auto-correlation
- Meta-label: biased resampling wins/loses with **internally tuned parameters**
- Learn Conditional Inference Tree



META-ANALYSIS: IMPACT OF DATA CHARACTERISTICS

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

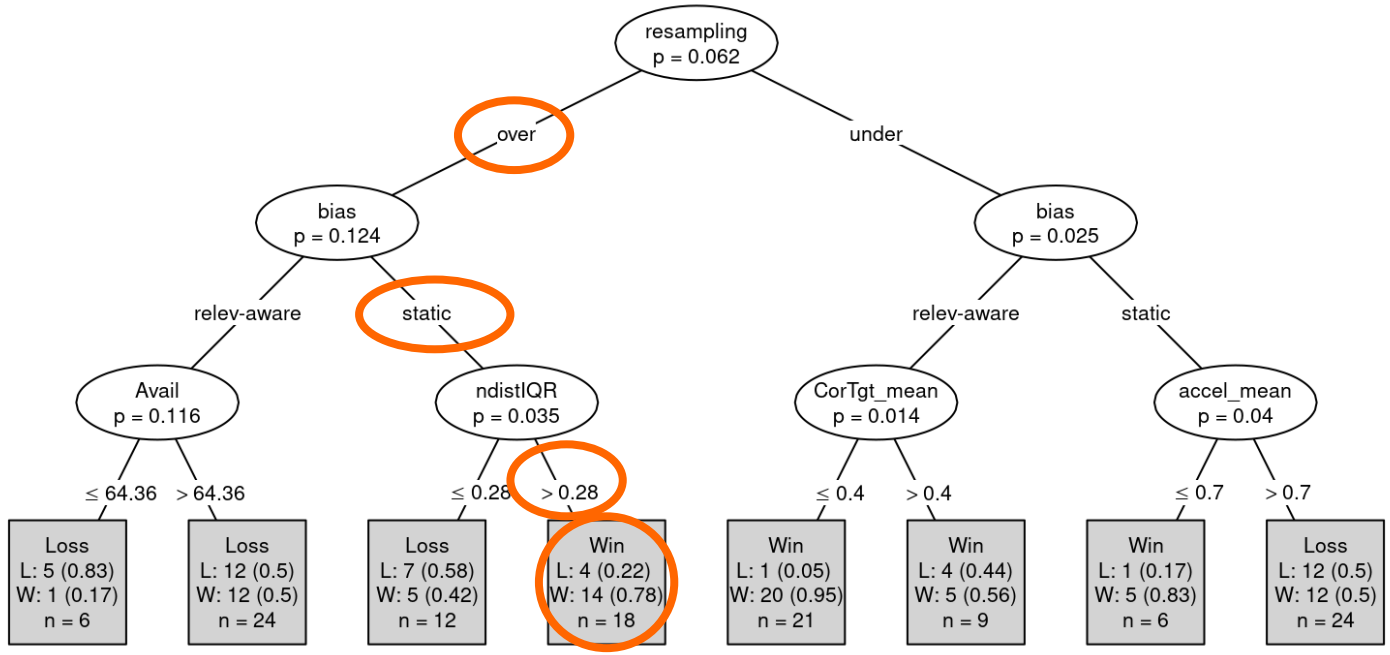
- Features describe data sets
  - Data size
  - Missing data
  - Target distribution characteristics
  - Spatial distribution of locations
  - Spatial/Temporal auto-correlation
- Meta-label: biased resampling wins/loses with **internally tuned parameters**
- Learn Conditional Inference Tree



META-ANALYSIS: IMPACT OF DATA CHARACTERISTICS

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

- Features describe data sets
  - Data size
  - Missing data
  - Target distribution characteristics
  - Spatial distribution of locations
  - Spatial/Temporal auto-correlation
- Meta-label: biased resampling wins/loses with **internally tuned parameters**
- Learn Conditional Inference Tree



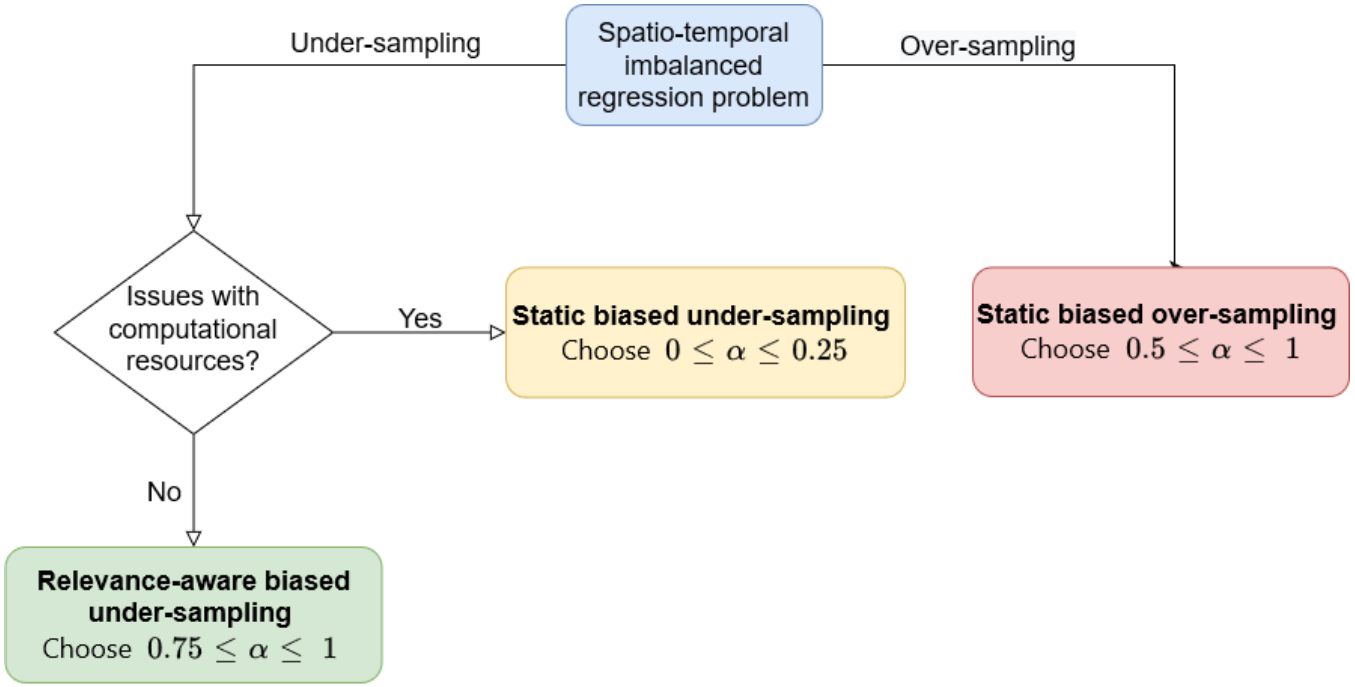
META-ANALYSIS: IMPACT OF DATA CHARACTERISTICS

RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

- We calculated **meta-feature importance** from CIFs:
  - **Type of bias** one of the most important meta-features when comparing strategies using the same resampling percentage
  - Much **less impactful after internal parameter tuning**
- Suggests our proposals for bias may have their merits under different circumstances **when properly tuned**

CONCLUSION
RESAMPLING IMBALANCED SPATIO-TEMPORAL DATA

- Our proposed spatio-temporal resampling strategies **improve prediction of extremes**
- **Relevance-aware biased under-sampling** worked best
  - Especially if **favouring the temporal dimension**



FINAL REMARKS

Predictive Analytics for Spatio-Temporal Data

MAIN CONTRIBUTIONS
FINAL REMARKS

Guidelines for performance estimation	Incentivize adequate evaluation of future spatio-temporal forecasting solutions
Proposal for extracting spatio-temporal indicators	Shows potential to capture some phenomena better than a previous proposal
New biased resampling strategies for imbalanced domains	Improve prediction of extreme numeric values in real-world applications
Study of impact of domain-specific data properties on proposals' efficacy	Inform future practitioners using our proposals
Freely available open-source software	R packages STEvaluation and STResampling

CONCLUSION & FUTURE DIRECTIONS

FINAL REMARKS

Spatio-temporal **data dependencies** present challenges when forecasting future values, but **provide opportunities** to improve predictive analytics and evaluation by leveraging contextual information

Future directions:

- Evaluating under irregularity and imbalance
- Explicitly addressing seasonality, lack of stationarity, and other heterogeneities
- Theoretical analysis of performance estimation methods
- Global vs. Local relevance

THANK YOU!

Publications:

- Mariana Oliveira, Luís Torgo, and Vítor Santos Costa. Evaluation Procedures for Forecasting with Spatio-Temporal Data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), volume 11051 LNAI, pages 703–718, 2018. doi:10.1007/978-3-030-10925-7_43
- Mariana Oliveira, Nuno Moniz, Luís Torgo, and Vítor Santos Costa. Biased Resampling Strategies for Imbalanced Spatio-Temporal Forecasting. In IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 100–109. IEEE, 2019. doi:10.1109/dsaa.2019.00024
- Mariana Oliveira, Luís Torgo, and Vítor Santos Costa. Evaluation Procedures for Forecasting with Spatiotemporal Data. Mathematics, 9(6), 2021b. doi:10.3390/math9060691
- Mariana Oliveira, Nuno Moniz, Luís Torgo, and Vítor Santos Costa. Biased resampling strategies for imbalanced spatio-temporal forecasting. International Journal of Data Science and Analytics, 12(3):205–228, 2021. doi:10.1007/s41060-021-00256-2

R packages:

- https://github.com/mrfoliveira/STEvaluation-MDPI2021
- https://github.com/mrfoliveira/STResampling-JDSA2020