Combining filter and embedded approaches to improve variable selection in land use change Cellular Automata models using Random Forests

Other Conference Item

Author(s):
Black, Benjamin

Publication date:
2022-07

Permanent link:
https://doi.org/10.3929/ethz-b-000593547

Combining filter and embedded approaches to improve variable selection in land use change Cellular Automata models using Random Forests

Benjamin Black & Prof. Dr. Adrienne Grêt-Regamey

iEMSs, 01.07.2022

Overview

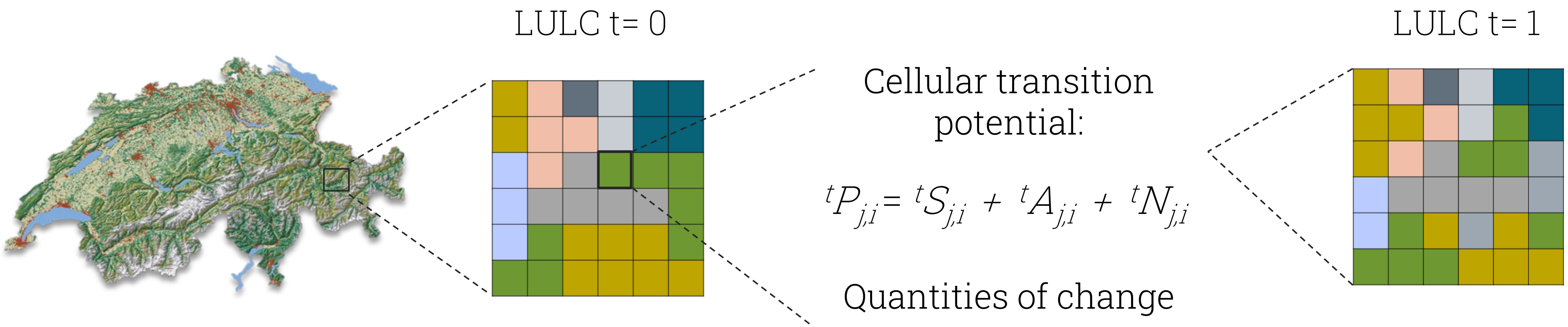
Combining **filter and embedded approaches** to improve variable selection in land use change Cellular Automata models using Random Forests

1. Introduction to Land use change Cellular Automata models
2. What are the benefits of variable(feature) selection
3. Applied example: Filter and embedded methods combined with Random Forests

Land Use Land Cover change Cellular Automata (LULCC-CA)

What:

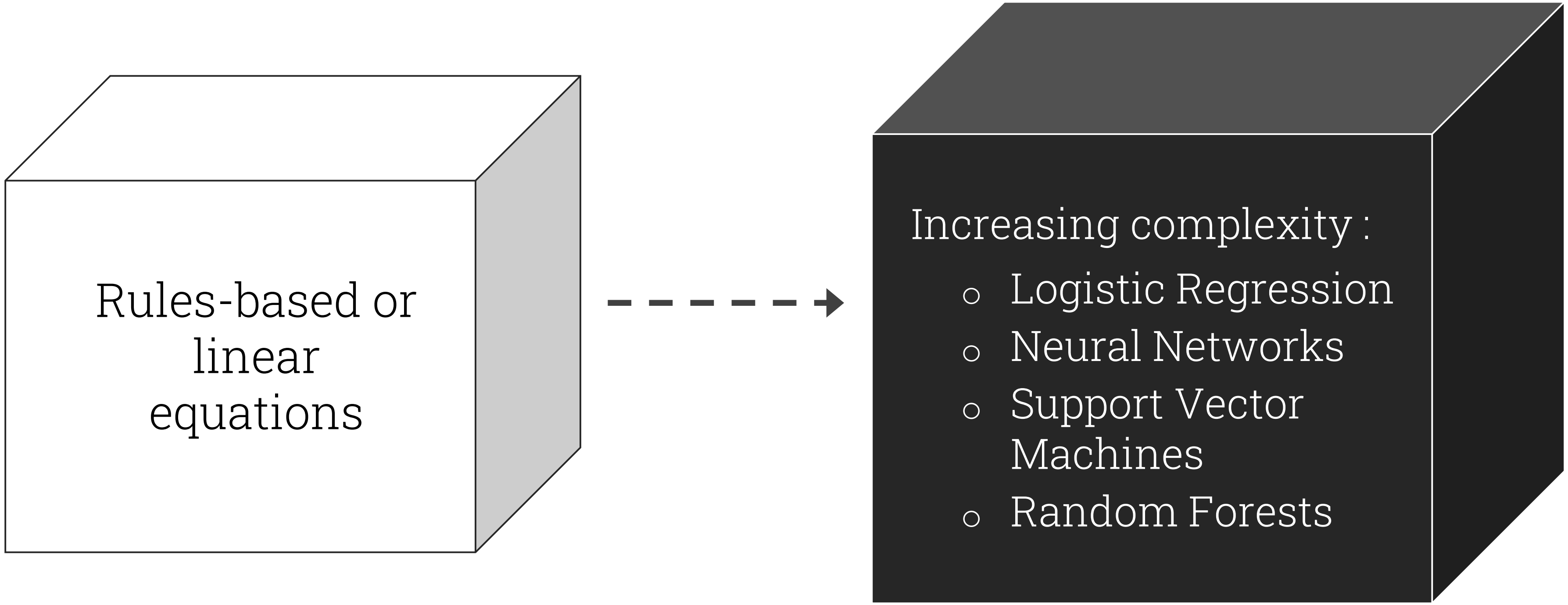
- Spatially explicit, 'patterns-based' approach to modelling LULCC
- Study area abstracted to cellular grid of LULC states
- LULCC simulated over discrete time with cells changing state on the basis of:
    - Previous state
    - Surrounding cells states: Neighborhood effect
    - Transition models encapsulating relationship between LULC transitions and driving variables (features)



Tobler 1979; R White and Engelen 1997; Roger White, Uljee, and Engelen 2012

Transition modelling in LULCC-CAs

- Driving variables represent an abstraction of the real-world processes of land use change
- Calibrated and validated on historical data and then used for future prediction
- Trend in field :



Kolb et al. 2013; Li and Yeh 2002; Yang et al. 2008; Rodrigues and Soares-Filho 2018; Kamusoko and Gamba 2015

Transition modelling in LULCC-CAs

- Driving variables represent an abstraction of the real-world processes of land use change
- Calibrated and validated on historical data and then used for future prediction
- Trend in field :

  - Lack of transparency in model behavior
  - Insufficient efforts to explore aspects of the model techniques:
    - Hyper-parameter tuning
    - Class imbalance
    - Feature selection

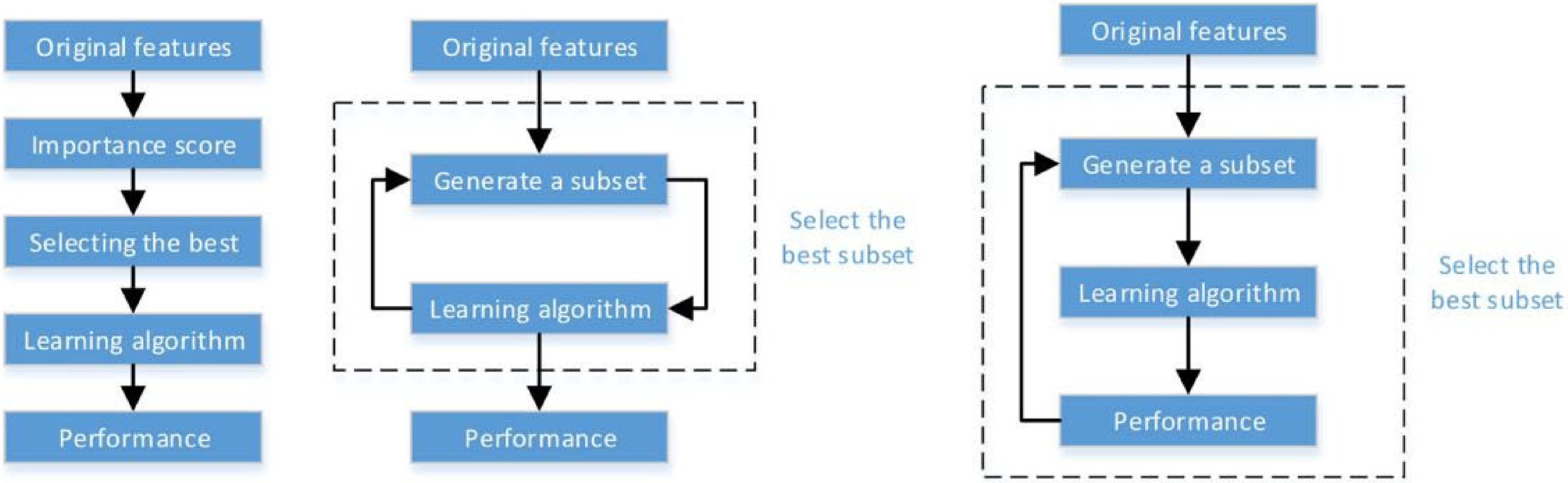
Increasing complexity :
- Logistic Regression
- Neural Networks
- Support Vector Machines
- Random Forests

Tong and Feng 2020

Transition modelling: Feature selection

- **What:** Selection of optimal set of features (variables/predictors) to give acceptable model performance whilst being representative, non-redundant and compact -> **parsimonious models**
- **3 approaches**:

1. Filter 2. Wrapper 3. Embedded

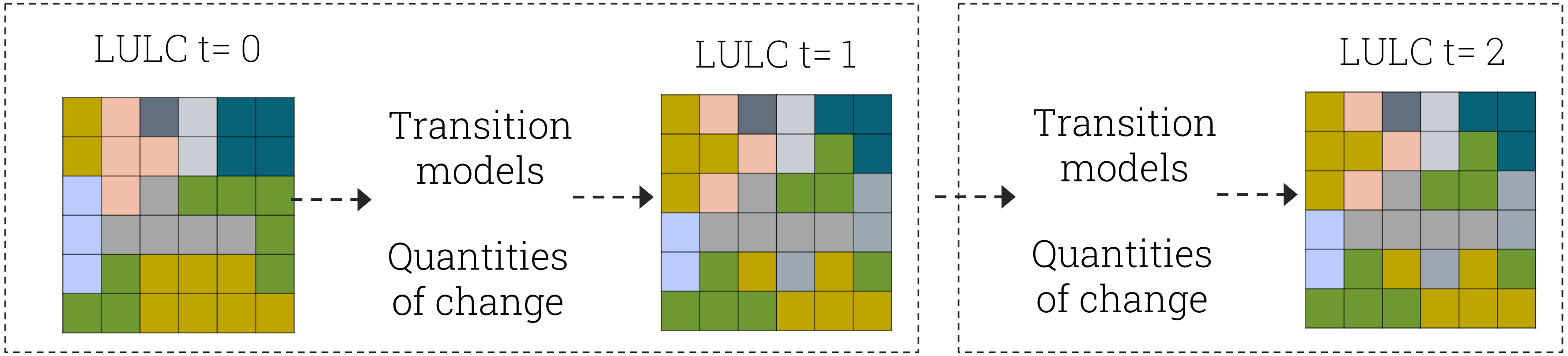


Wang et al. 2016

Feature selection rationale

Potential benefits for LULCC-CAs:

- **Model Generalizability:** improved performance on unseen data
- **Reduce burden for future simulations :** less variables must be extrapolated or assumed stationary



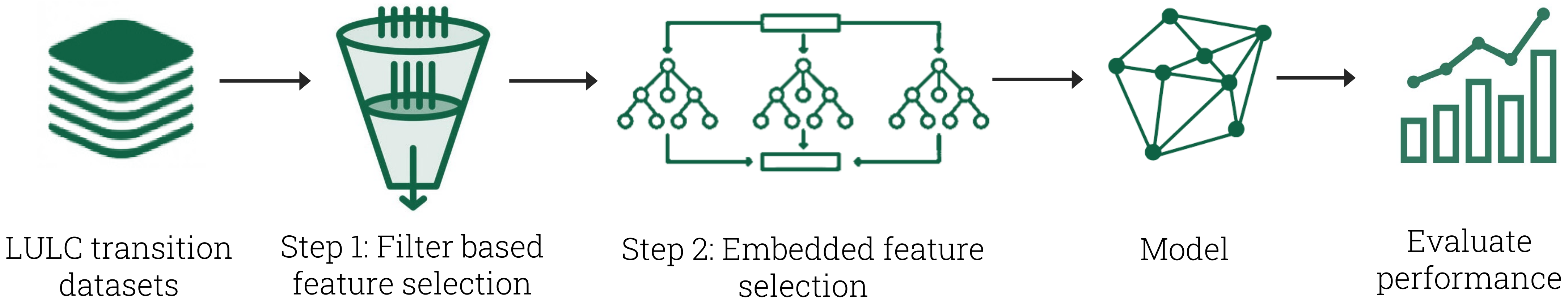
Within calibration interval historic data available for all variables

For future time points, transition models are stationary but require temporally dynamic variable data or assumption of stationarity: **Problematic**

Guyon et al. 2006

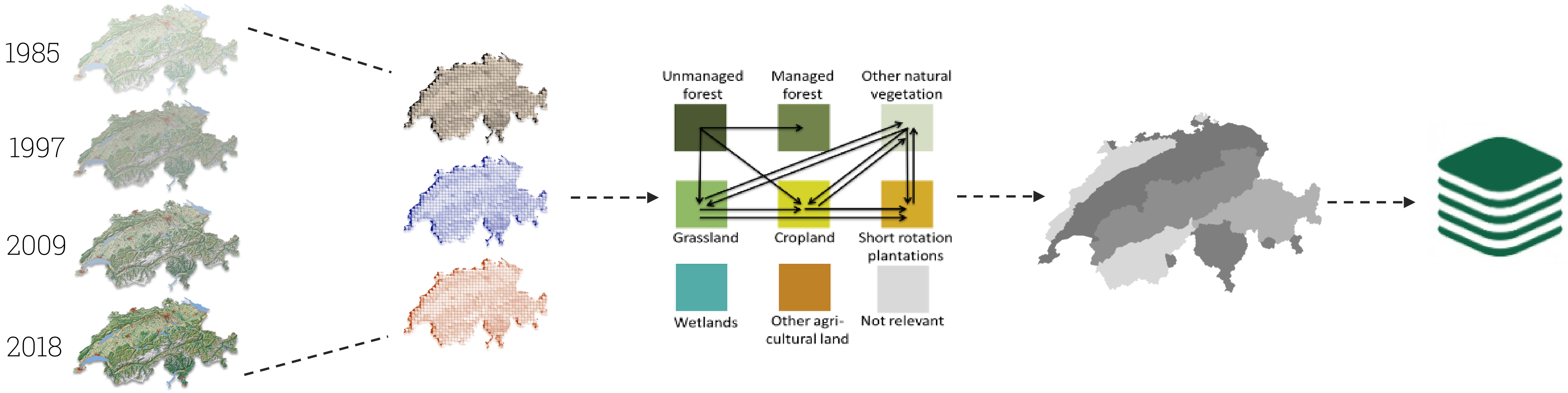
Applied example

- **What**: Two step feature selection approach for Random Forests transition models of LULC change in Switzerland
- **Aim**: Demonstrate benefits in terms of model generalisability and parsimony



Methods: Data preparation

LULCC transition datasets:



100m raster data for 4 time points, aggregated to 10 LULC classes

Identification of LULC transitions between time points: 3 change layers

Filtered to ~ 35 class-class transitions

Split by Swiss Bio-regions (6)

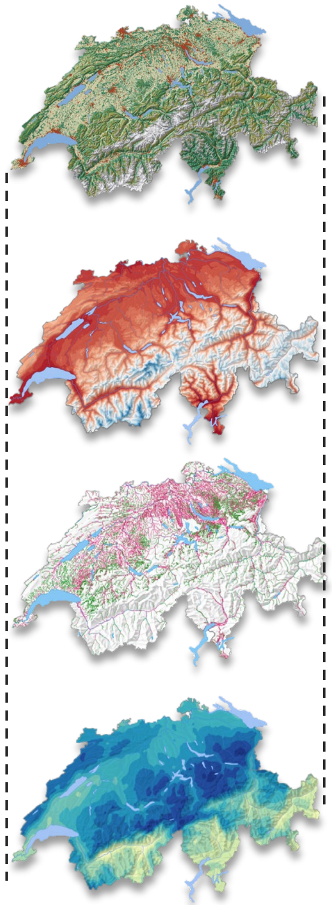
LULC transition datasets: 174-217 per period

Federal Office for statistics (FSO) 2021

Methods: Data preparation

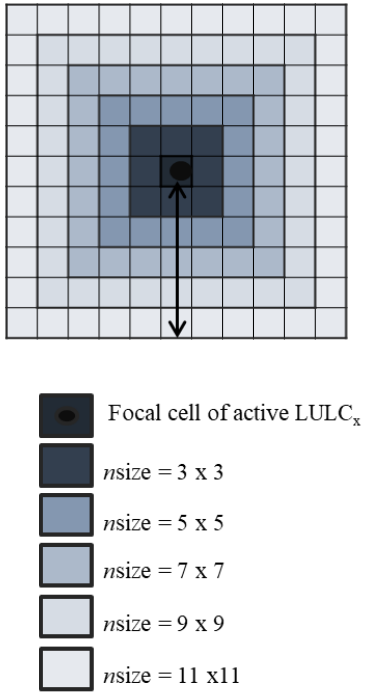
Predictors grouped by category: Accessibility and suitability vs. Neighbourhood

Suitability/Accessibility predictors
Distance to roads
Continentality
Light
Soil pH
Soil nutrients
Soil moisture
Soil moisture variability
Soil aeration
Soil humus
Change in average population per municipality
Change in no. of employees in primary sector per municipality
Change in no. of employees in secondary and tertiary sectors per municipality
Mean elevation
Aspect
Slope
Hillshade
Noise pollution index
Distance to lakes
Distance to rivers
Annual mean temperature
Average annual precipitation
Sum of growing days above 0 degrees
Sum of growing days above 3 degrees
Sum of growing days above 5 degrees



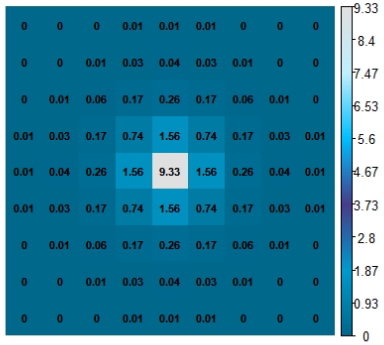
Methods: Neighborhood predictors

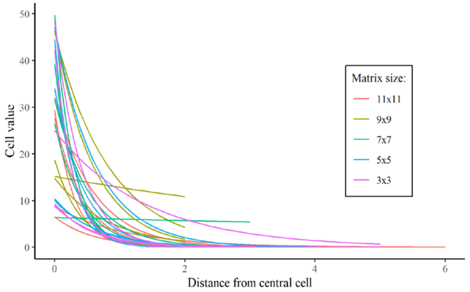
5x neighbourhood sizes



X

5 x Random matrices

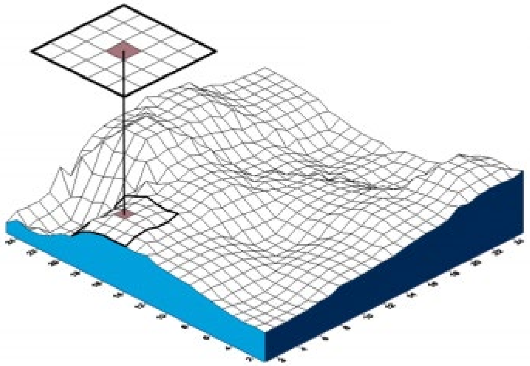




X

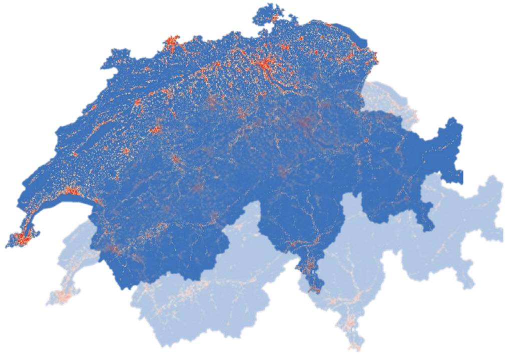
5 x Active LULC classes

Settlement/urban/amenities
Intensive agriculture
Alpine pastures
Grassland or meadows
Permanent crops



=

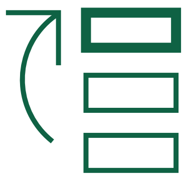
125 x Neighborhood realisations (raster layers)

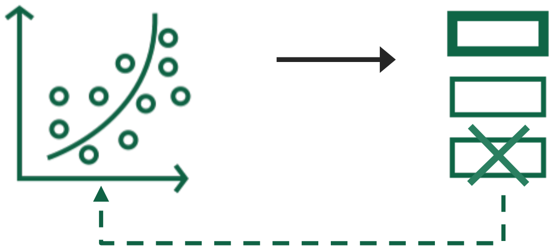


Roodposhti et al. 2020

Step 1: Filter based feature selection

$$Y = a + bX$$





Features predictors grouped categorically (Suitability and accessibility; neighbourhood split according to the active LULC class)

Univariate GLMs produced for each predictor in each group

p-values from these models used to produce ranked lists of predictors (lowest to highest)

Pairwise Pearson’s correlation coefficients calculated iteratively with the lowest (list-wise) variable removed until the subsequent correlation value was below a threshold of 0.7

Output: Datasets with different number of remaining predictors: max of 1 neighbourhood predictor for each active LULC class and as many suitability and accessibility predictors that passed the <0.7 correlation cut-off.

Step 2: Embedded feature selection

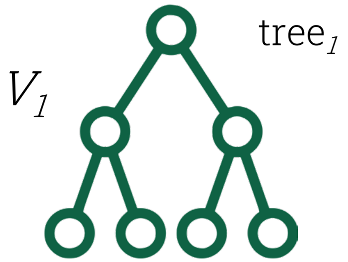
- Guided Regularized Random Forests (GRRF)
- Purpose: Select "compact" (non-redundant) subsets of predictors directly utilising the RF algorithm

Prior to GRRF:

Fit standard RF model and calculate normalized feature importance (NFI) scores

Ensemble decision tree construction in GRRF:

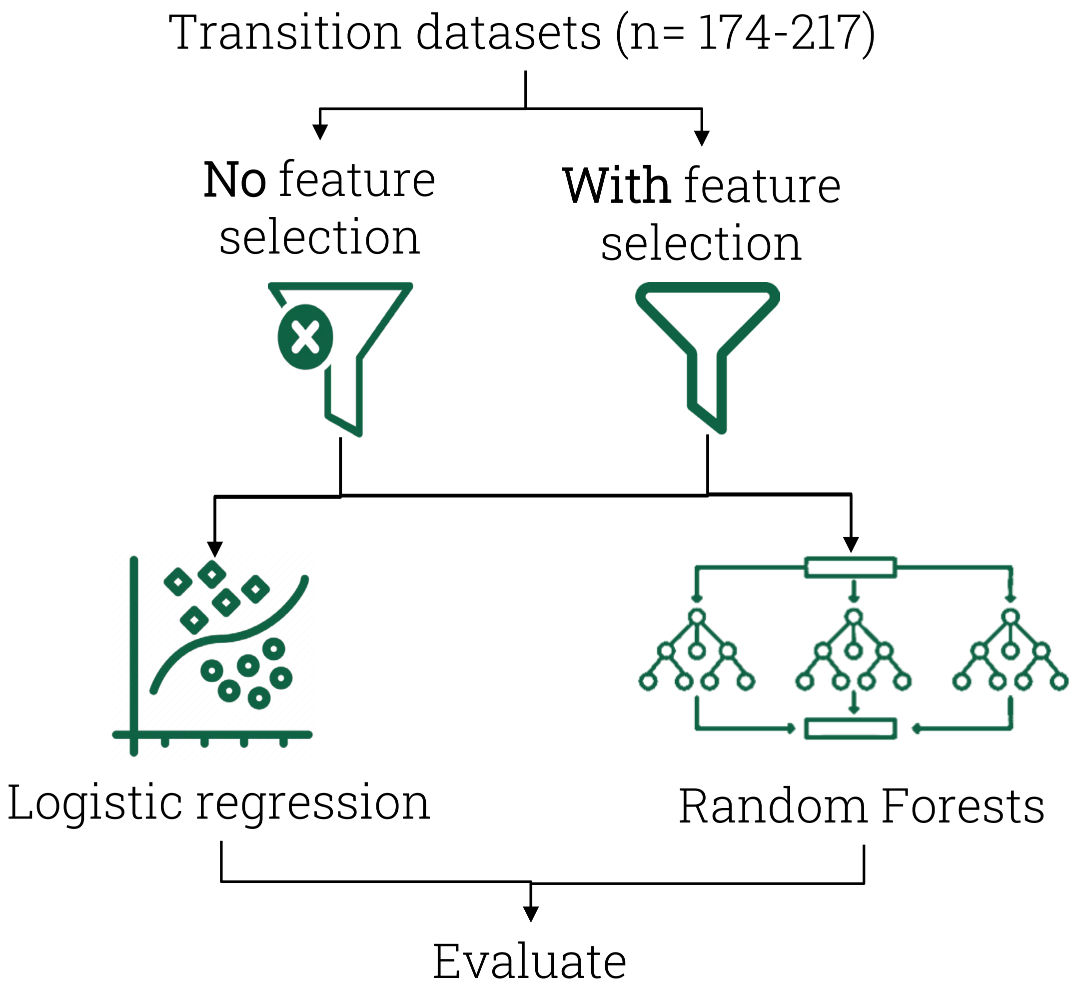
- Instantiate empty subset (F) of features
- Features used at tree nodes are added to F



- At each node (V) splitting occurs as per RF with Gini Information gain (GI) calculated for each feature (X_j)
- GI values modified by a penalty factor if X_j is present in F, with penalty scaled by NFI
- X not included in F must have high importance to overcome penalization

Deng and Runger 2013

Modelling

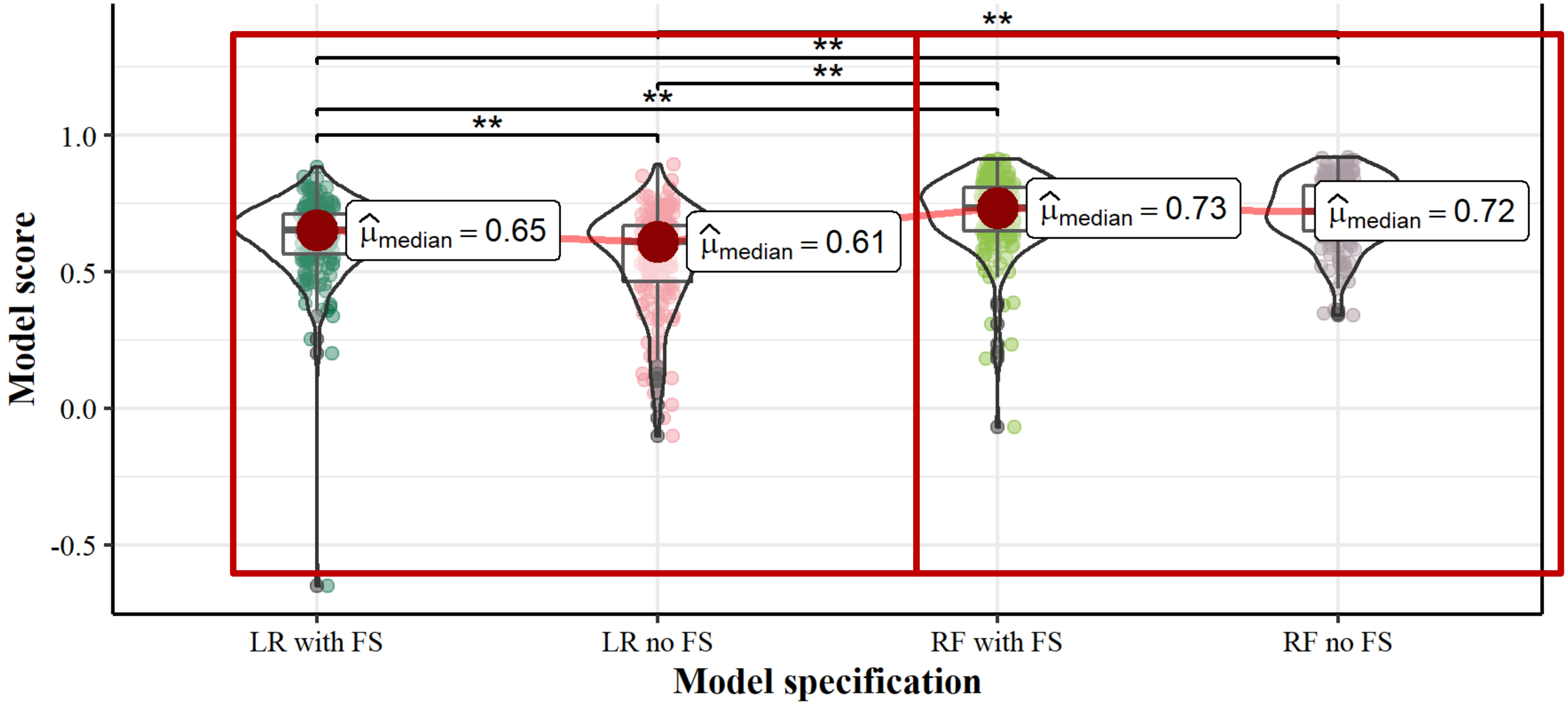


- All models fitted on 5 replicates using a **70:30 split of training/test data** to allow for independent validation
- Models evaluated using threshold and non-threshold metrics averaged over replicates:

Model score [-1, 1] = $\bar{x}$(norm(AUC ROC), Boyce index)

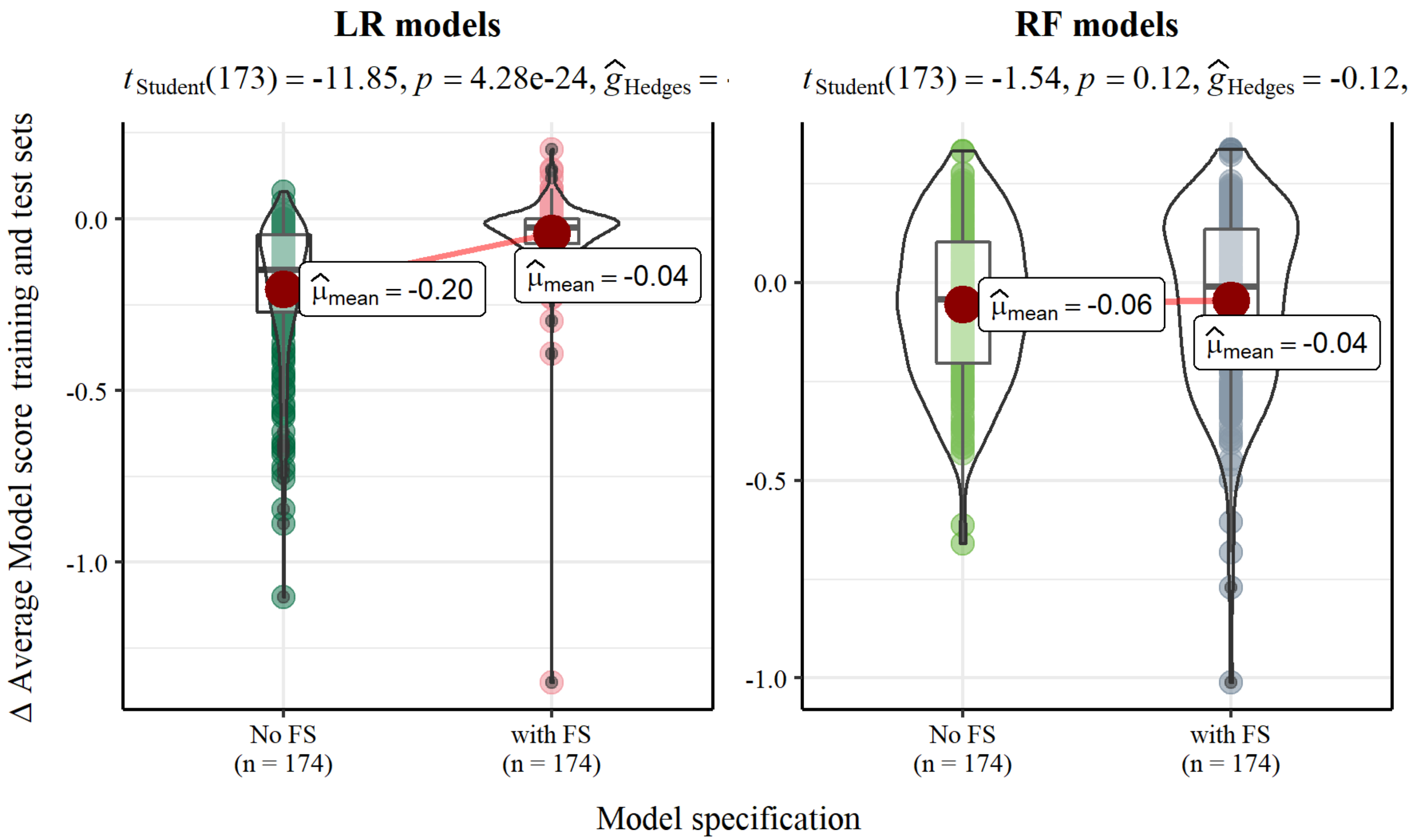
Boyce et al. 2002; Hirzel et al. 2006; Paegelow et al. 2018

Model performance

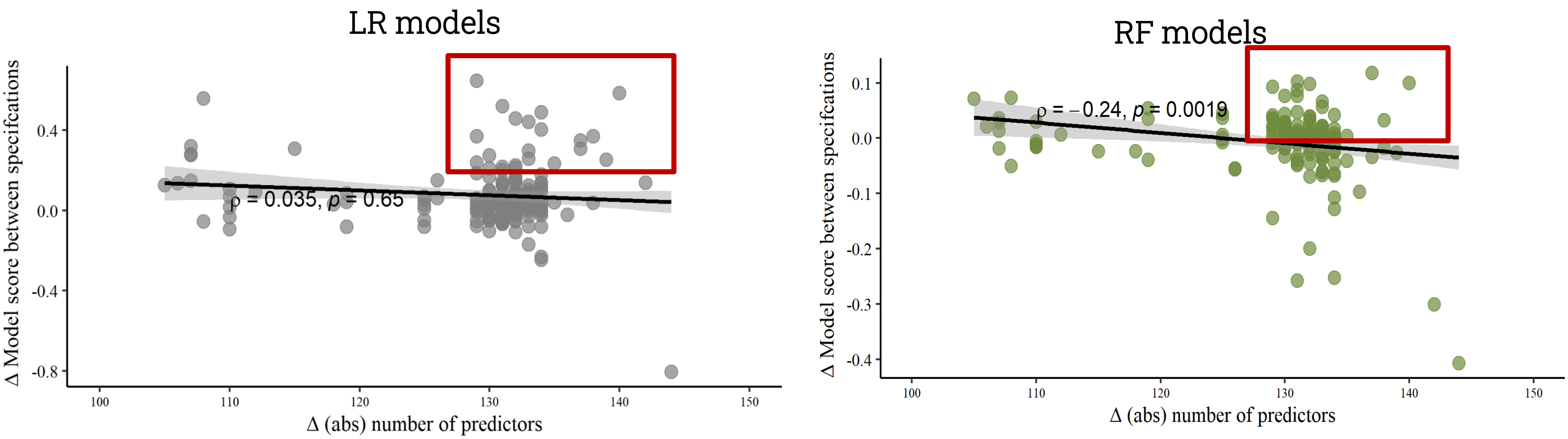


Bars between violins indicate significant pairwise differences between groups under the Conover's all-pairs comparisons test (*p<0.05, **p<0.01).

Model generalisability



Model parsimony



Scatter plots of the differences (Δ) in the model score metric against absolute (abs) differences in the number of predictors between the models with feature selection and without feature selection models with linear trend line and correlation (Spearman's) coefficient.

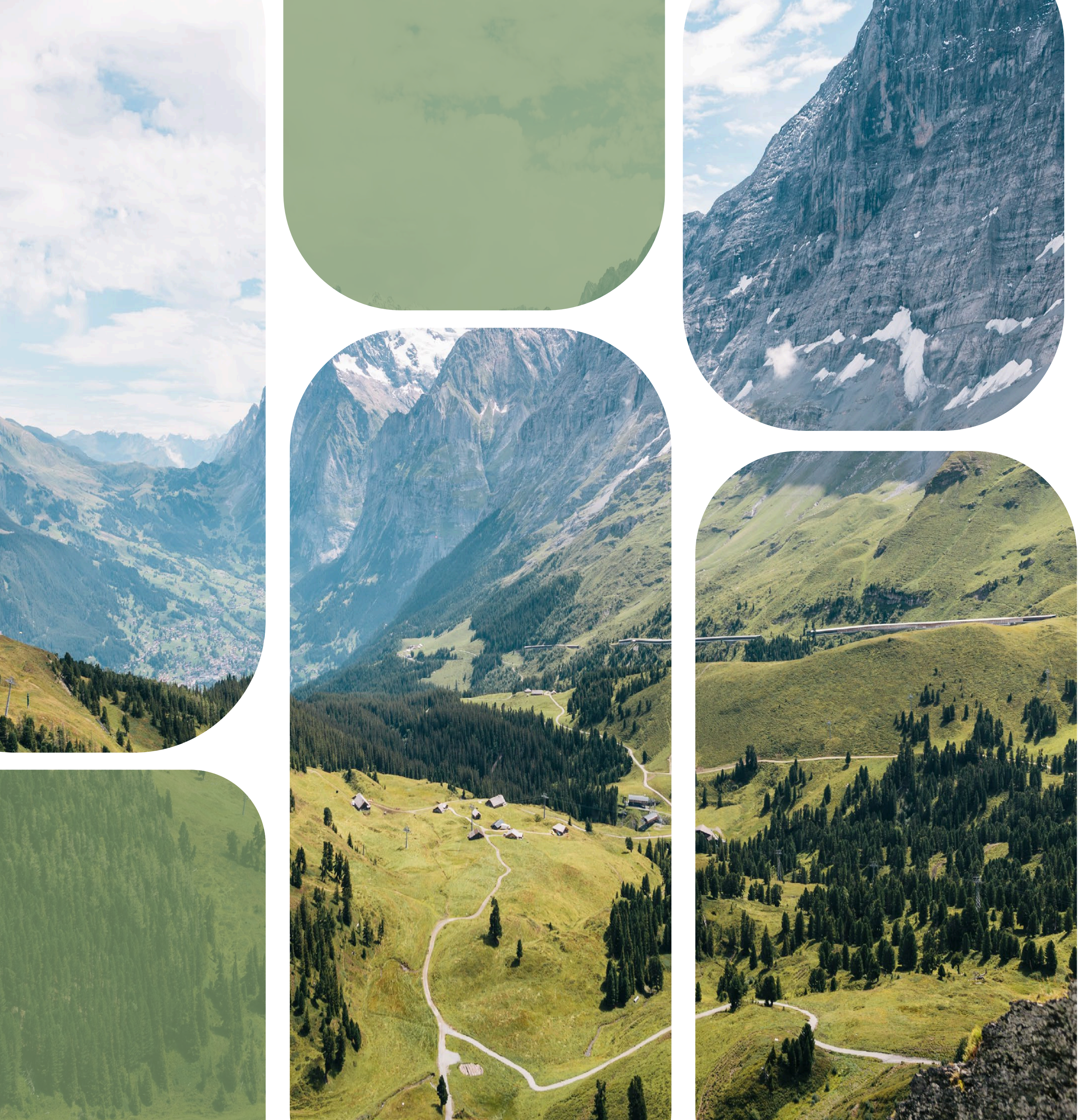
Conclusion

Takeaway: LULCC-CAs should include feature selection as a process of training transition models because it offers two benefits:

- Improved model generalisability
- Moderate reduction in number of predictors for only small decreases in performance

Publication:

Black, B. Grêt-Regamey, A. Van Strien, M. 2022. Improving calibration of land use change models through proactive validation of transition potential predictions. *In preparation*.



Thank you for listening

I will now take any questions.

References

Boyce, Mark S, Pierre R Vernier, Scott E Nielsen, and Fiona K. A Schmiegelow. 2002. 'Evaluating Resource Selection Functions'. *Ecological Modelling* 157 (2): 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4.

Deng, Houtao, and George Runger. 2013. 'Gene Selection with Guided Regularized Random Forest'. *Pattern Recognition* 46 (12): 3483–89. https://doi.org/10.1016/j.patcog.2013.05.018.

Federal Office for statistics (FSO). 2021. 'Areal Statistics According to Nomenclature 2004, Surveys 1979-1985, 1992-1997, 2004-2009, 2013-2018'. https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/geostat/geodaten-bundesstatistik/boden-nutzung-bedeckung-eignung/arealstatistik-schweiz.assetdetail.20104753.html.

Guyon, Isabelle, Masoud Nikravesh, Steve Gunn, and Lotfi A. Zadeh, eds. 2006. *Feature Extraction: Foundations and Applications*. Vol. 207. Studies in Fuzziness and Soft Computing. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35488-8.

Hirzel, Alexandre H., Gwenaëlle Le Lay, Véronique Helfer, Christophe Randin, and Antoine Guisan. 2006. 'Evaluating the Ability of Habitat Suitability Models to Predict Species Presences'. *Ecological Modelling* 199 (2): 142–52. https://doi.org/10.1016/j.ecolmodel.2006.05.017.

Kamusoko, Courage, and Jonah Gamba. 2015. 'Simulating Urban Growth Using a Random Forest-Cellular Automata (RF-CA) Model'. *ISPRS International Journal of Geo-Information* 4 (2): 447–70. https://doi.org/10.3390/ijgi4020447.

Kolb, Melanie, Jean-François Mas, and Leopoldo Galicia. 2013. 'Evaluating Drivers of Land-Use Change and Transition Potential Models in a Complex Landscape in Southern Mexico'. *International Journal of Geographical Information Science* 27 (9): 1804–27. https://doi.org/10.1080/13658816.2013.770517.

Li, Xia, and Anthony Gar-On Yeh. 2002. 'Neural-Network-Based Cellular Automata for Simulating Multiple Land Use Changes Using GIS'. *International Journal of Geographical Information Science* 16 (4): 323–43. https://doi.org/10.1080/13658810210137004.

Paegelow, M., M. T. Camacho Olmedo, and J. F. Mas. 2018. 'Techniques for the Validation of LUCC Modeling Outputs'. In *Geomatic Approaches for Modeling Land Change Scenarios*, edited by María Teresa Camacho Olmedo, Martin Paegelow, Jean-François Mas, and Francisco Escobar, 53–80. Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-60801-3_4.

References

Rodrigues, H., and B. Soares-Filho. 2018. 'A Short Presentation of Dinamica EGO'. In *Geomatic Approaches for Modeling Land Change Scenarios*, edited by María Teresa Camacho Olmedo, Martin Paegelow, Jean-François Mas, and Francisco Escobar, 493–98. Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-60801-3_35.

Roodposhti, Majid Shadman, Richard J. Hewitt, and Brett A. Bryan. 2020. 'Towards Automatic Calibration of Neighbourhood Influence in Cellular Automata Land-Use Models'. *Computers, Environment and Urban Systems* 79 (January): 101416. https://doi.org/10.1016/j.compenvurbsys.2019.101416.

Tobler, W. R. 1979. 'Cellular Geography'. In *Philosophy in Geography*, edited by Stephen Gale and Gunnar Olsson, 379–86. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-9394-5_18.

Tong, Xiaohua, and Yongjiu Feng. 2020. 'A Review of Assessment Methods for Cellular Automata Models of Land-Use Change and Urban Growth'. *International Journal of Geographical Information Science* 34 (5): 866–98. https://doi.org/10.1080/13658816.2019.1684499.

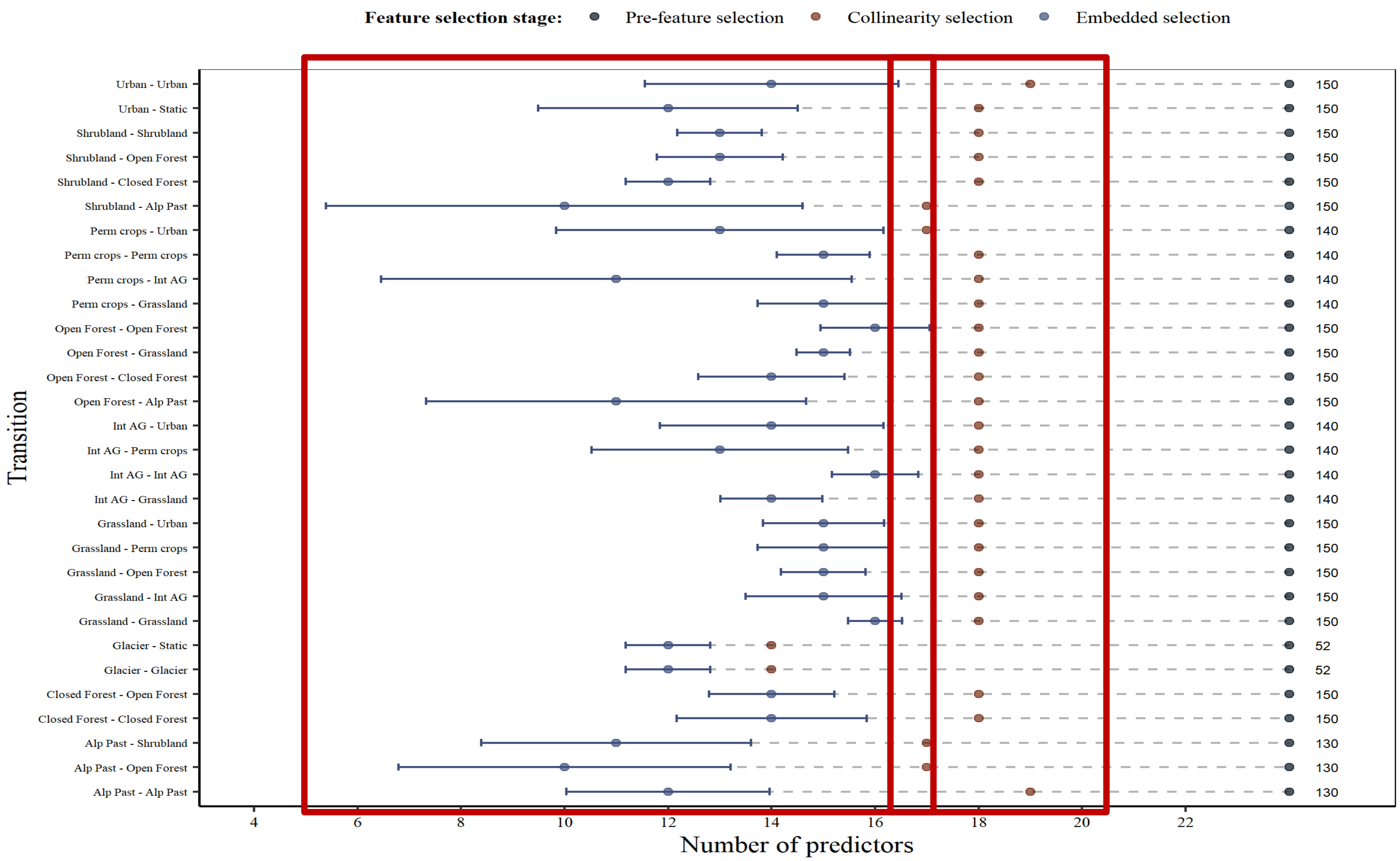
Wang, Shuihua, Yin Zhang, Tianmin Zhan, Preetha Phillips, Yu-Dong Zhang, Ge Liu, Siyuan Lu, and Xueyan Wu. 2016. 'Pathological Brain Detection by Artificial Intelligence in Magnetic Resonance Imaging Scanning'. *Progress In Electromagnetics Research* 156 (August): 105–33.

White, R, and G Engelen. 1997. 'Cellular Automata as the Basis of Integrated Dynamic Regional Modelling'. *Environment and Planning B: Planning and Design* 24 (2): 235–46. https://doi.org/10.1068/b240235.

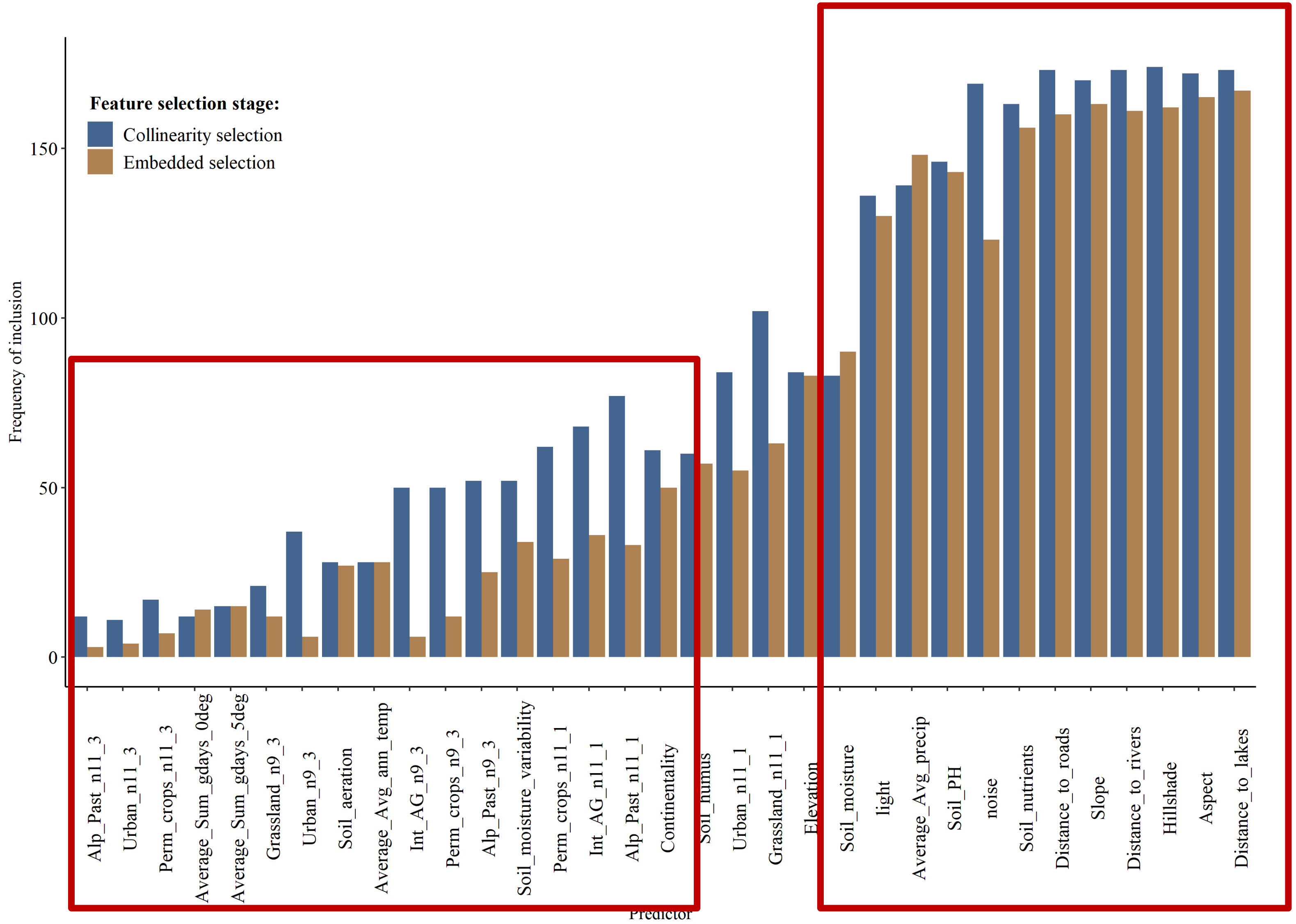
White, Roger, Inge Uljee, and Guy Engelen. 2012. 'Integrated Modelling of Population, Employment and Land-Use Change with a Multiple Activity-Based Variable Grid Cellular Automaton'. *International Journal of Geographical Information Science* 26 (7): 1251–80. https://doi.org/10.1080/13658816.2011.635146.

Yang, Qingsheng, Xia Li, and Xun Shi. 2008. 'Cellular Automata for Simulating Land Use Changes Based on Support Vector Machines'. *Computers & Geosciences* 34 (6): 592–602. https://doi.org/10.1016/j.cageo.2007.08.003.

Results: Feature reductions



Results: Feature retention



Model performance

– RF is generally invariant to redundant predictors so the approach is unlikely to produce better performing models: