Model selection and the principle of parsimony in statistical modelling in agriculture

Zhanglong Cao

SAGI West, Curtin University

5 December, 2019, Adelaide

Overview

A contradiction in model selection

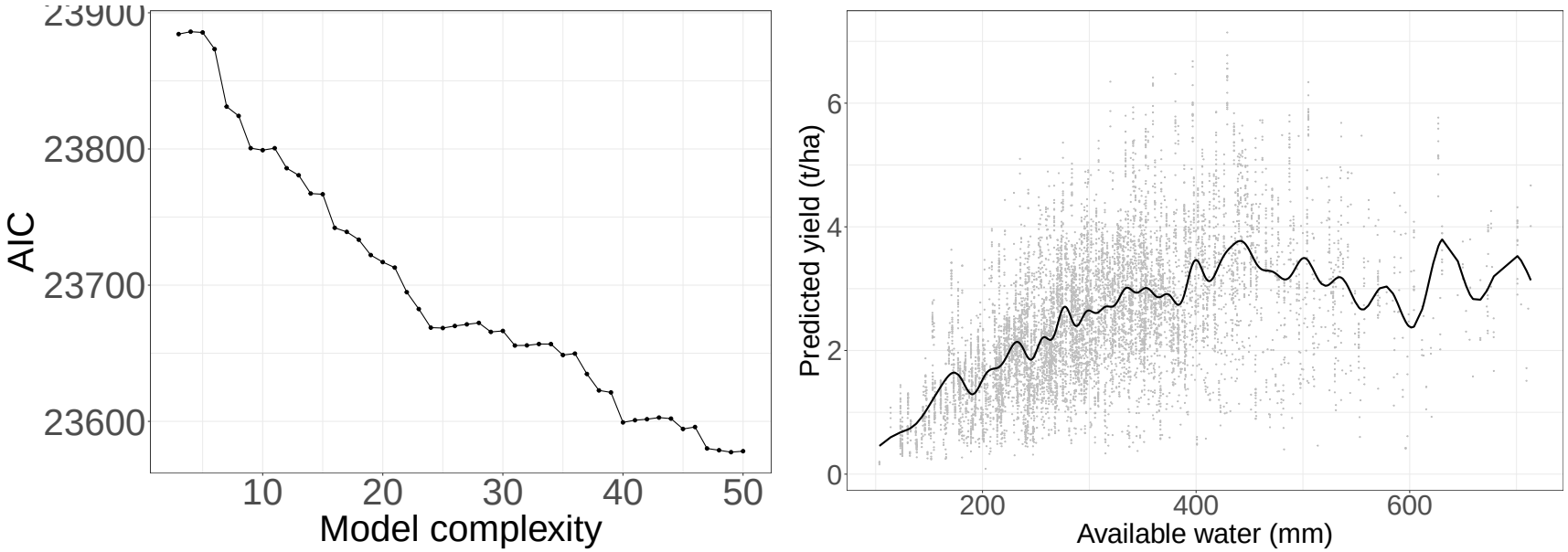
When we choose the "best model", it refers to the model which meets the nominated selection criteria,

- AIC [Akaike, H., 1973], BIC [Schwarz, G, 1978]
- Cross-validation score, including CV (leave-one-out, k-fold), GCV. [Wahba, G, 1985]
- Others: AIC_c, RIC (residual information criterion), MDL (minimum description length) etc.
  [Hurvich, C.M., et al, 1989, Shi, P, et al, 2002, Rissanen, J. 1983]

However, sometimes, the "best model" over-fits the data.

An example

Data is fitted by `gam` model.



(a) AIC against model complexity (number of knots)

(b) Data fitted by the "best model"

Figure 1: An example of over-fitting by the "best model".

Available water: the amount of water available for crop use in any growing season is defined to be one third of summer rainfall plus growing season rainfall.

Question

How can we find the best model that meets both our expectation and the statistical selection criteria?

Variety trial data

The data set used in [Chen, et al. 2019].

- It consists of 9,116 yield estimates taken from 109 variety trials.
- It covers a 40-year period of time from 1975 to 2014.
- Figure 2 shows the distribution of 775 locations in variety trials conducted in the WA grain belt, overlaid on 30-year average rainfall [Garlinge, J., 2005]

Variety trial data

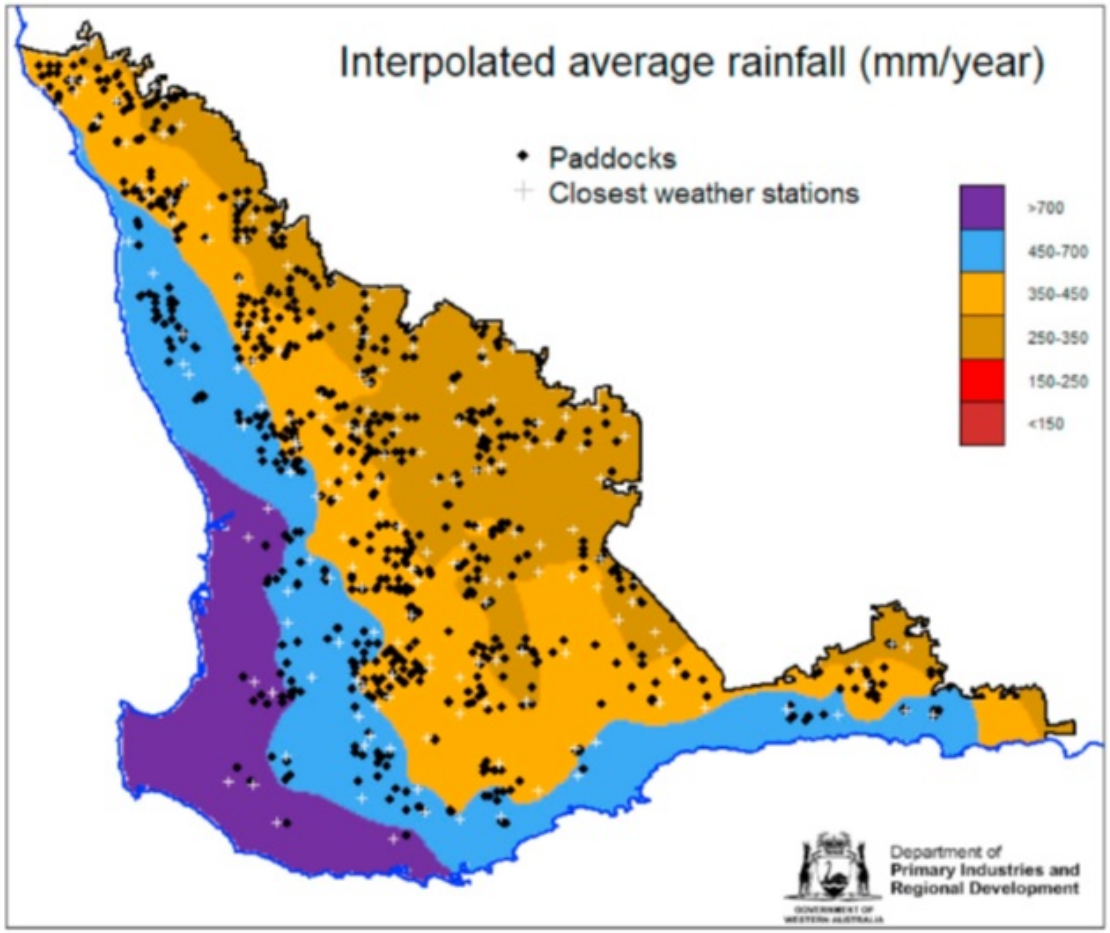


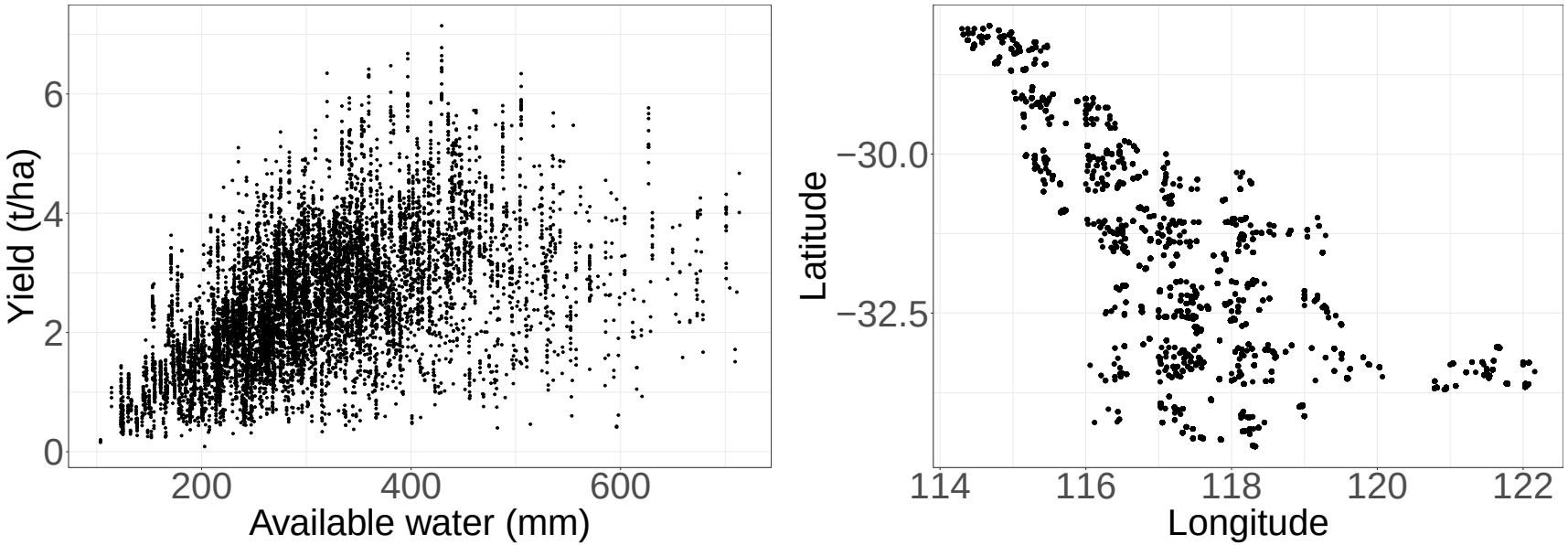
Figure 2: Map of the paddocks and patched point weather stations in variety trials with interpolation of 30-years average rain fall. [Chen, et al. 2019]

Variety trial data

Following [Chen, et al. 2019], we only focus on the following variables which contribute most to yield estimation and prediction:

- Growing season available water.
- Location information (longitude and latitude).

Data visualization



(a) Yield (t/ha) against available water.

(b) Location information

Figure 3: Data visualization

GAM

A generalised additive model (GAM) is a generalised linear model in which the predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions [Wood, S. N. 2017, Chen, et al. 2019, Hastie, T., et al, 2009].

$$\mathrm{E}(Y \mid X_1, \ldots, X_p) = \alpha + f_1(X_1) + \ldots + f_p(X_p). \tag{1}$$

GAM

A preliminary `gam` model from package `mgcv` [Wood, S. N. 2017] for yield prediction is:

$$\text{yield} \sim s(\text{water}, k_1) + s(\text{longitude}, \text{latitude}, k_2), \tag{2}$$

where k_1 and k_2 are the number of knots that control the complexity of the model.

Balance in model selection

In the process of model selection, there is a *tradeoff* between bias and variance [Yu, L, et al, 2006]. Given a new data x_0 and observed y_0,

$$\mathrm{E}[(y_0 - \hat{f}(x_0))^2] = \mathrm{Var}(\hat{f}_0) + \mathrm{Bias}^2(\hat{f}_0) + \sigma^2, \tag{3}$$

where $\mathrm{Var}(\hat{f}_0) = \mathrm{E}[(\hat{f}_0 - \mathrm{E}[\hat{f}_0])^2]$, $\mathrm{Bias} = \mathrm{E}[\hat{f}_0] - y_0$ and σ^2 is the irreducible error (beyond our control).

Balance in model selection

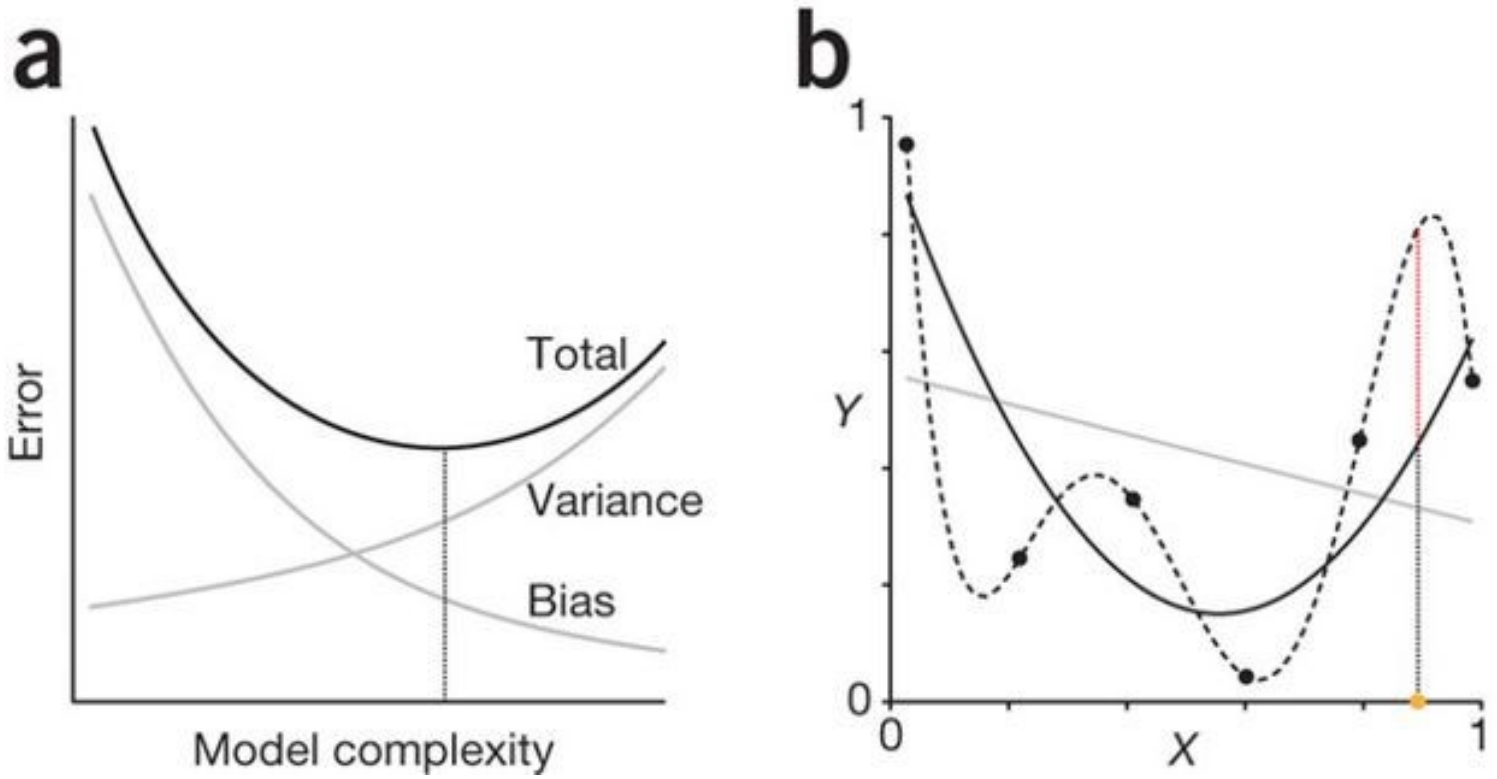


Figure 4: Over-fitting is a challenge for regression and classification problems [Lever,J.K. et al 2016].

When model complexity increases, generally bias decreases and variance increases. The choice of the optimal model is informed by the goal of minimizing the total error (dotted vertical line).

Moving-window CV

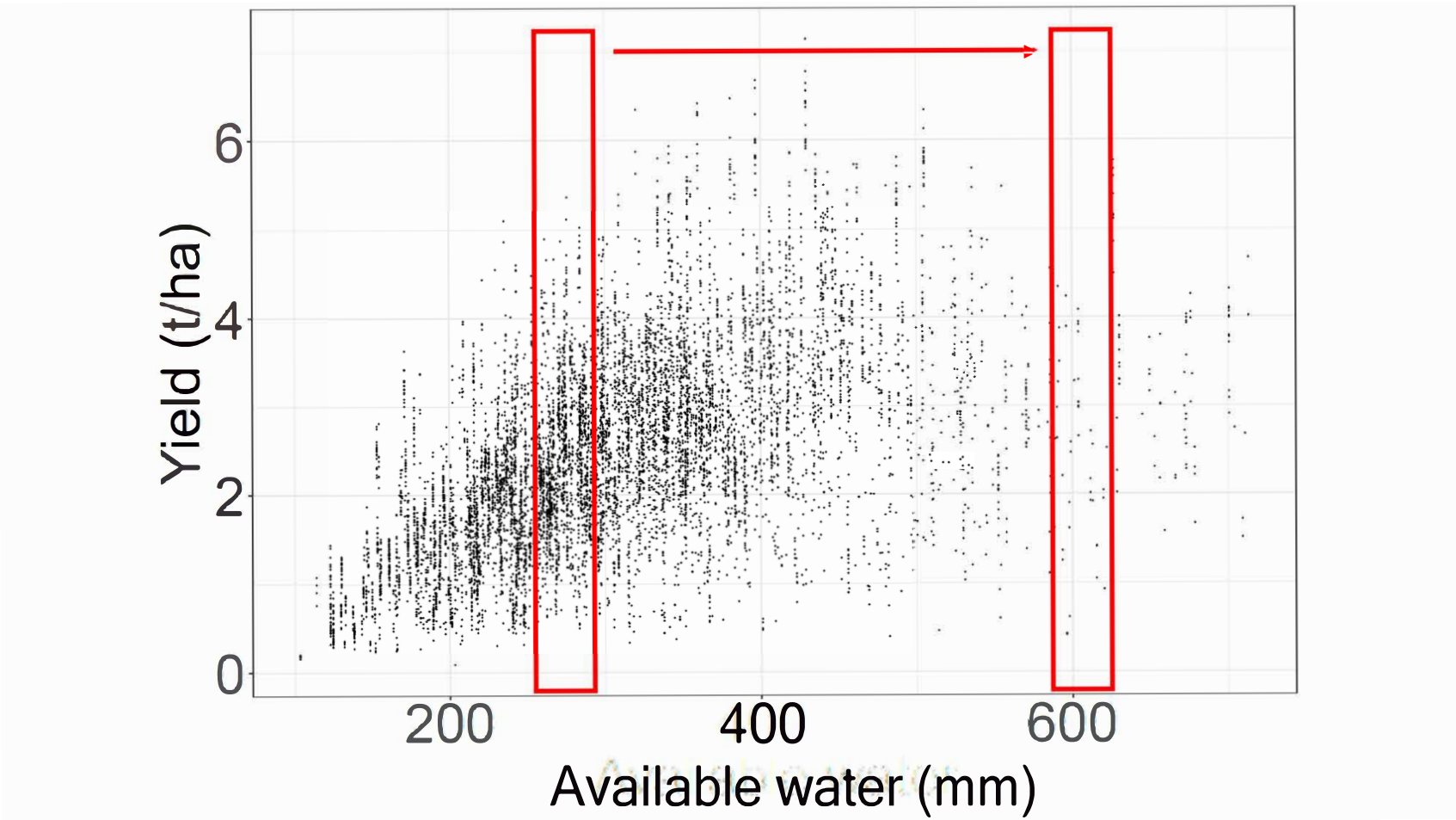


Figure 5: Data in the red frame is removed for testing. The rest of the data is used for training. The window moves forward and the process repeats.

Moving-window spatially CV

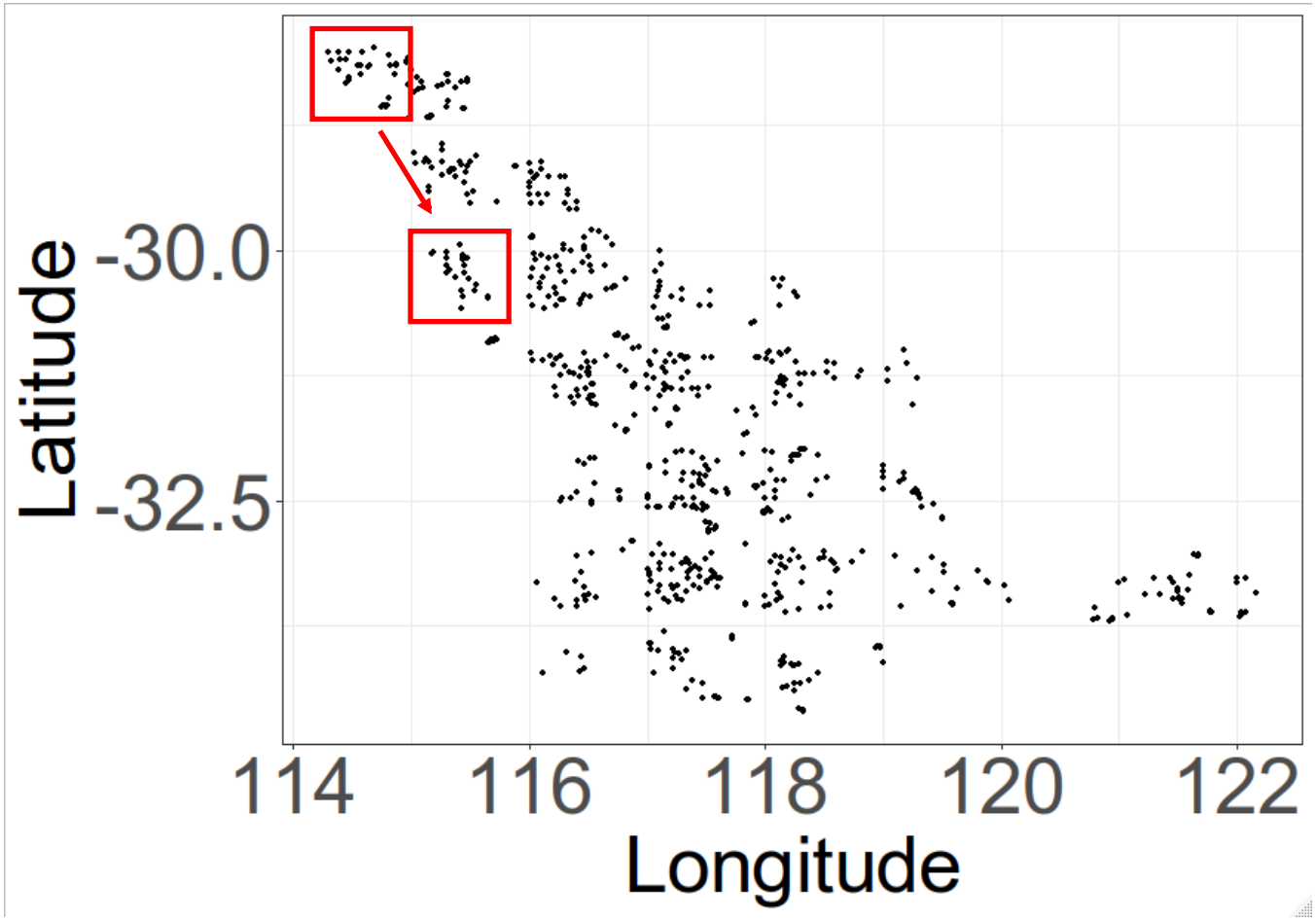


Figure 6: Spatial cross-validation removes spatially autocorrelated data from the data set for testing [Brenning, A. 2012].

Adaptive MW cross-validation

In adaptive moving-window cross-validation, the bandwidth of the window depends on the density of the data.

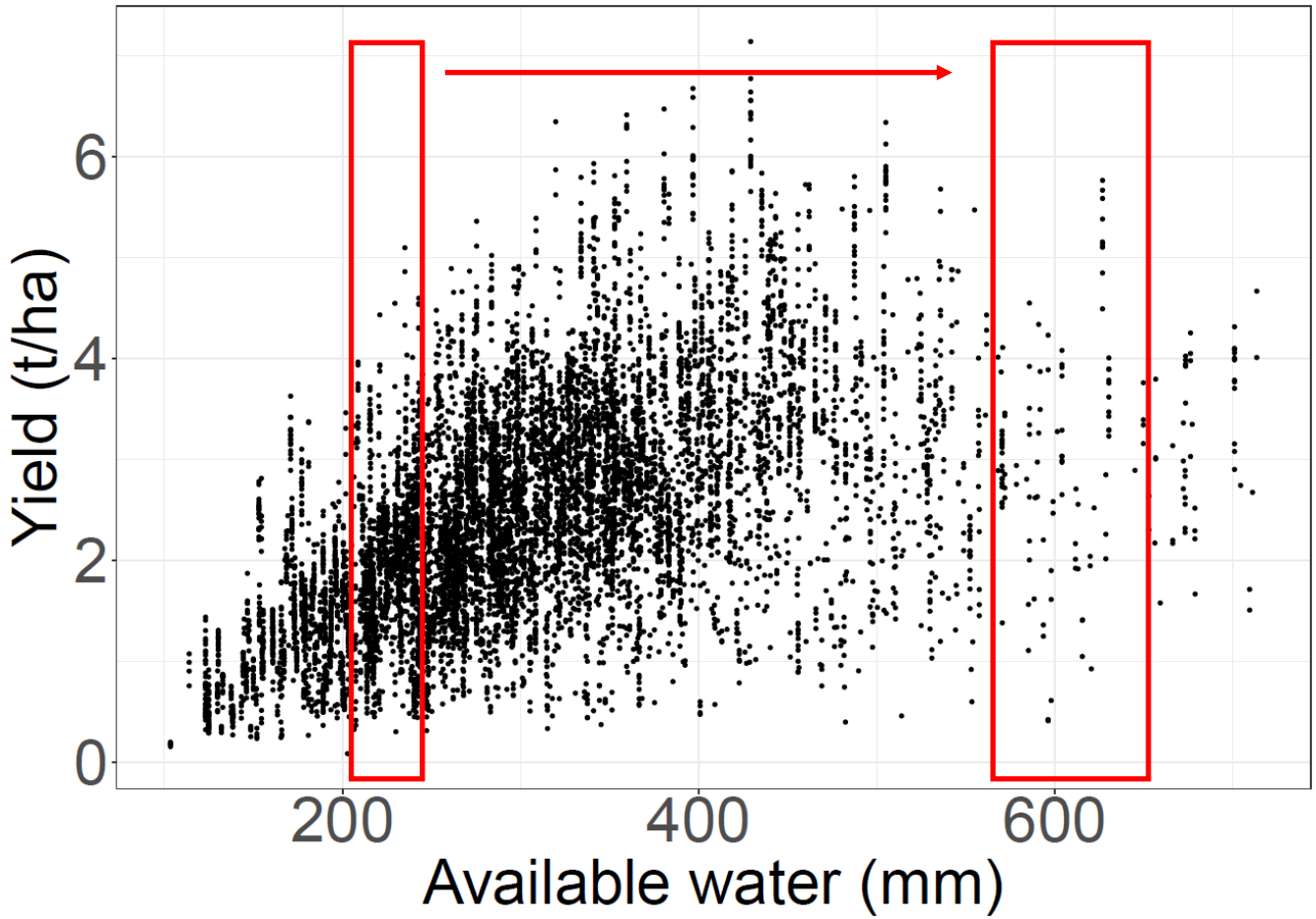


Figure 7: Narrower window in high density area and wider window in sparse area.

Adaptive MW cross-validation

In adaptive moving-window cross-validation, the bandwidth of the window depends on the density of the data.

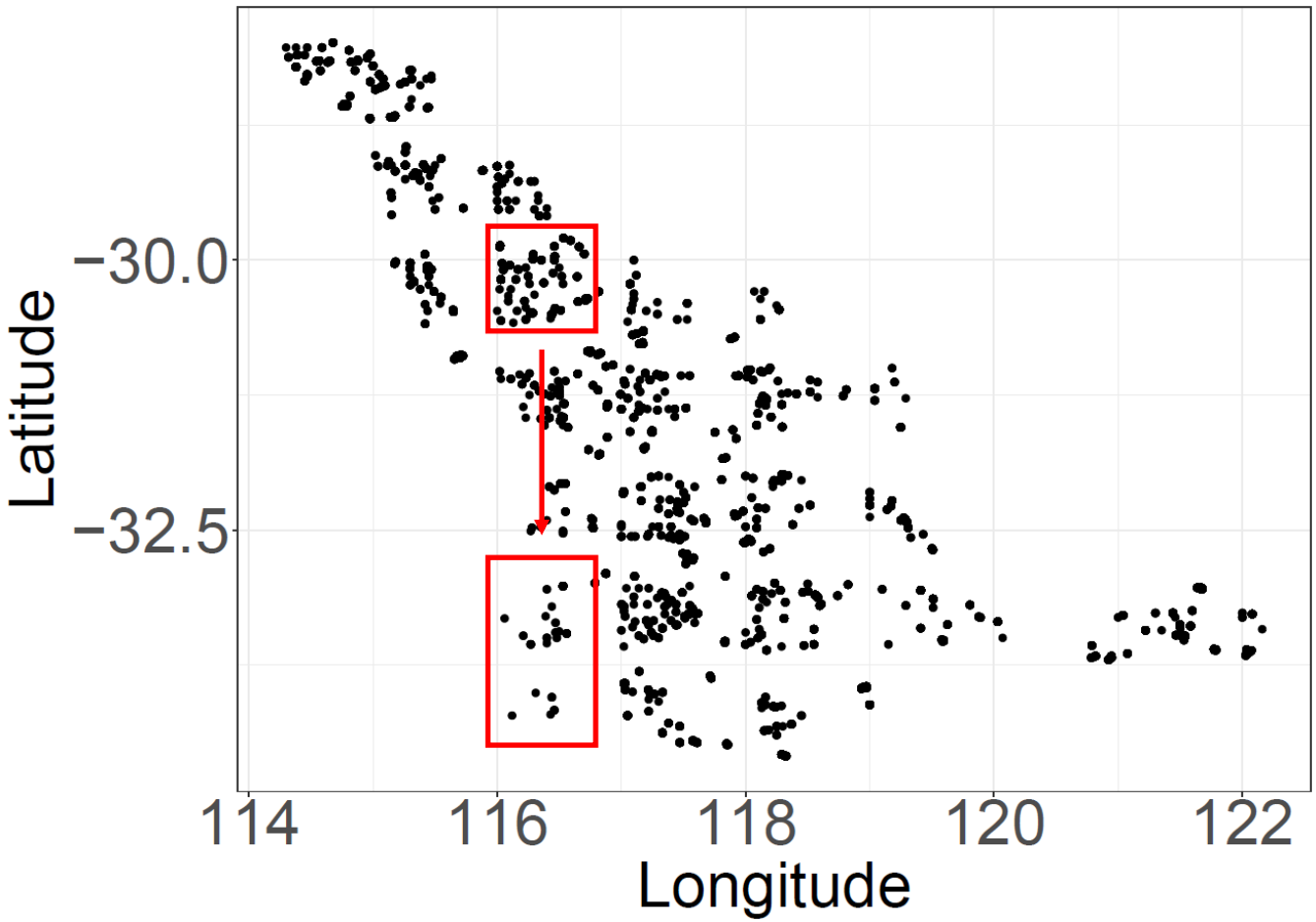


Figure 8: Narrower window in high density area and wider window in sparse area.

Comparison

$$\text{yield} \sim s(\text{water}, k_1) + s(\text{longitude}, \text{latitude}, k_2) \tag{4}$$

	MW (fixed)	AIC	BIC	5-fold	GCV
k_1	3	50	24	50	50
k_2	7	10	9	10	10

Plots

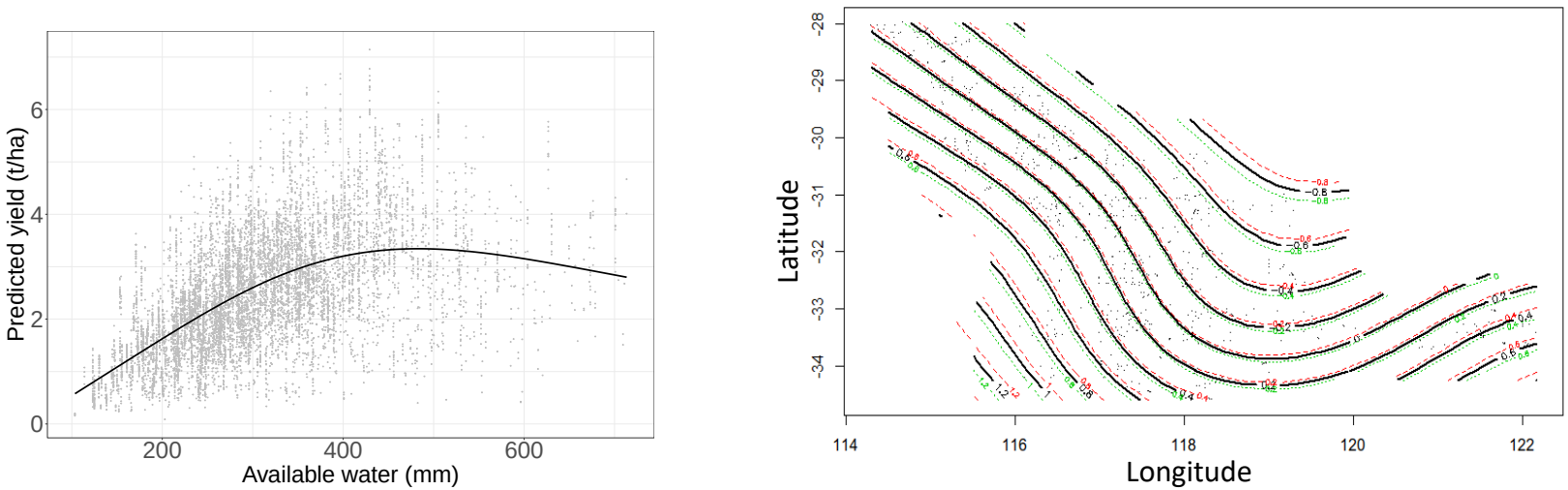


Figure 9: Smooth plots which meet both our expectation and the balance in model selection.

The principle of parsimony

- A parsimonious model has the minimum number of parameters and maximum predictive power, in which every parameter reflected a known effect on yield predictive system to allow mechanistic interpretation [Landau, S., et al, 2000, Oweis, T., et al, 2006].
- As the model in our study is built for the purpose of prediction, a simple and parsimonious model with few inputs was selected [Chen, et al. 2019].

Conclusion

- *No Free Lunch Theorems for Optimization* [Wolpert, D.H, et al, 1997].
- If data is autocorrelated, moving window/spatially cross-validation is recommended.
- Constrained Generalized Additive Model (CGAM), and Shape Constrained Additive Models (SCAM)

Discussion

A further question: how to determine the bandwidth or the window size for the moving-window cross-validation?

- Kernel density estimation
- Clustering methods (such as K-means)

Refer to the paper: Rakshit, S. et al. (2019)

Novel Approach to the Analysis of Spatially-varying Treatment Effects in On-farm Experiments

submitted to *Field Crops Research*.

Acknowledgement

Many thanks to my lovely team members: Katia Stefanova, Karyn Reeves, Kefei Chen, Suman Rakshit, Smaila Sanni, Angelika Pilkington and Andrew Grose for their support and comments.

References

Chen, Kefei and O'Leary, Rebecca A. and Evans, Fiona H. (2019).

A simple and parsimonious generalised additive model for predicting wheat yield in a decision support tool.
Agricultural systems: 173, 140-150.

Garlinge, Jenny (2005).

2005 Crop variety sowing guide for Western Australia.
Department of Agriculture and Food, Western Australia, Perth. Bulletin 4655.

Akaike, Htrotugu (1973).

Maximum likelihood identification of Gaussian autoregressive moving average models
Biometrika: 60(2),255–265.

Schwarz, Gideon (1978).

Estimating the dimension of a model
The annals of statistics: 6(2), 461–464.

Wahba, Grace (1985).

A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem
The annals of statistics: 13(4), 1378–1402.

Shi, Peide and Tsai, Chih-Ling (2002).

Regression model selection – a residual likelihood approach
Journal of the Royal Statistical Society: Series B (Statistical Methodology): 64(2), 237–252.

Yu, Lean and Lai, Kin Keung and Wang, Shouyang and Huang, Wei (2006).

A bias-variance-complexity trade-off framework for complex system modeling
International Conference on Computational Science and Its Applications: 2006, 518-527

Hurvich, Clifford M. and Tsai, Chih-Ling (1989).

Regression and time series model selection in small samples
Biometrika: 76(2), 297–307.

References

Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome (2009).
The elements of statistical learning: prediction, inference and data mining. Second Edition.
Springer.

Wood, Simon N. (2017).
Generalized additive models: an introduction with R (2nd Edition)
Chapman and Hall/CRC.

Rissanen, Jorma (1983).
A universal prior for integers and estimation by minimum description length
The Annals of statistics: 11(2),416–431.

Lever, Jake Krzywinski and Martin Altman, Naomi (2016).
Points of significance: model selection and overfitting
Nature Methods: 13, 703–704.

Brenning, Alexander (2012).
Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package "sperrorest".
2012 IEEE International Geoscience and Remote Sensing Symposium: 5372-5375.

Oweis, Theib and Hachum, Ahmed (2006).
Water harvesting and supplemental irrigation for improved water productivity of dry farming systems in West Asia and North Africa.
Agricultural Water Management: 80(1-3), 57–73.

Landau, S., Mitchell, R. A. C. , Barnett, V., Colls, J. J., Craigon, J., Payne, R. W. (2000).
A parsimonious, multiple-regression model of wheat yield response to environment
Agricultural and forest meteorology; 101(2-3), 151–166.

Wolpert, D.H., Macready, W.G. (1997).
No Free Lunch Theorems for Optimization
IEEE Transactions on Evolutionary Computation 1, 67.