Modelling European Agriculture with Climate Change for Food Security
— a FACCE JPI knowledge hub —

Statistical learning approach for modelling the effects of climate change on oilseed rape yield

Sassari, Italy *April 2014*

Behzad Sharif
Jørgen E. Olesen
Kirsten Schelde

Aarhus University
Department of Agroecology

Outline

- Introduction & Background
- Materials and Methods
- Results
- Discussion

Introduction

- An introduction to statistical learning with applications in R (James et al) Springer

What is statistical learning?

Assume X is a set of features and Y is a set responses:

- Set of approaches to estimate f in a way that

$Y = f(X) + \varepsilon$

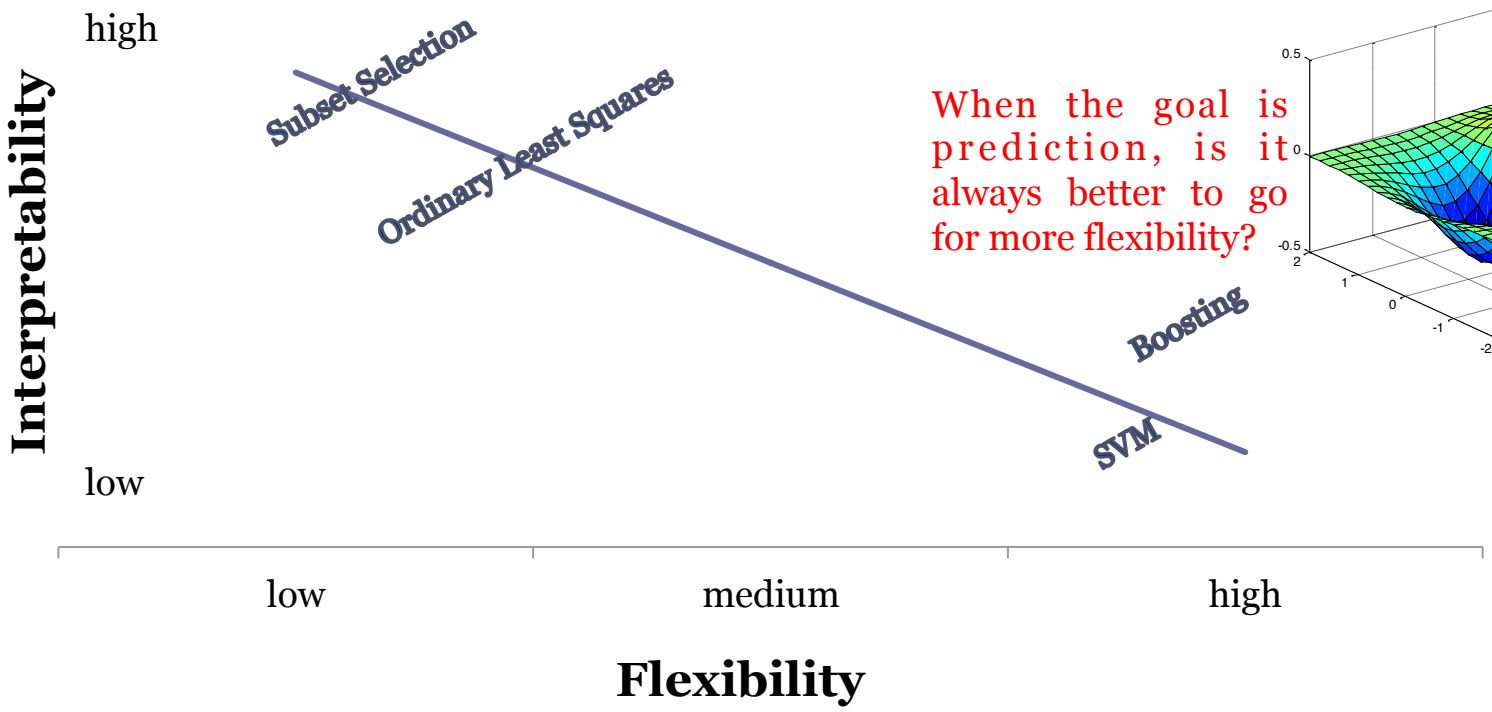
- Systematic information that X provides about Y

Based on traditional methods, inspired by data mining and machine learning

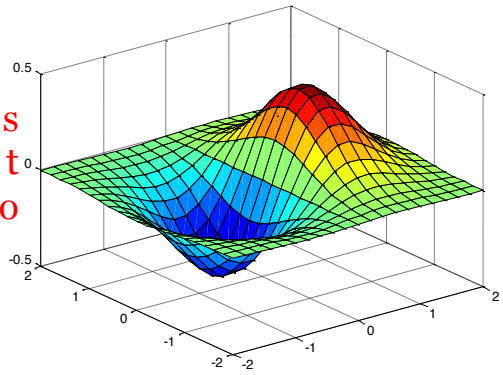
How would we like to use our learning?

- Prediction
  - When we are more interested in the results and its accuracy
- Inference
  - When we are more interested in understanding:
    - Which predictors are important?
    - What is the relationship?

Flexibility vs. Interpretability



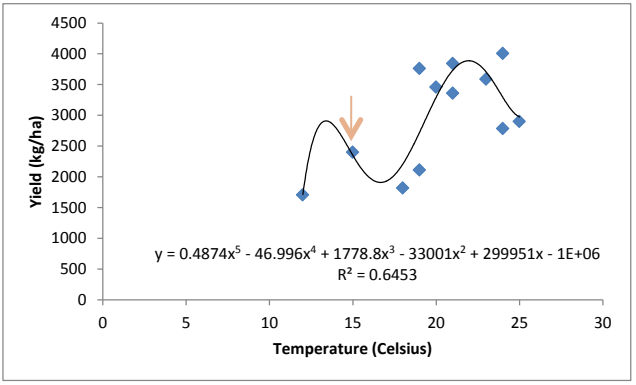
When the goal is prediction, is it always better to go for more flexibility?

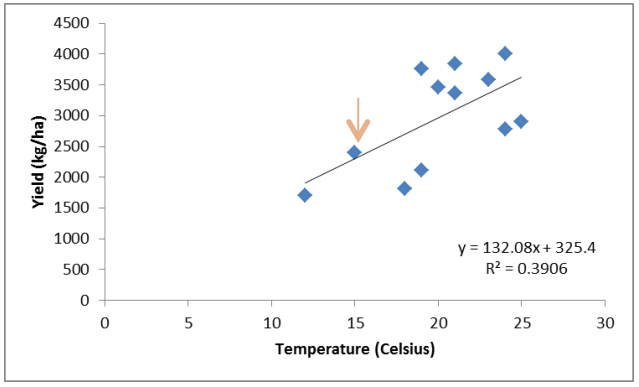


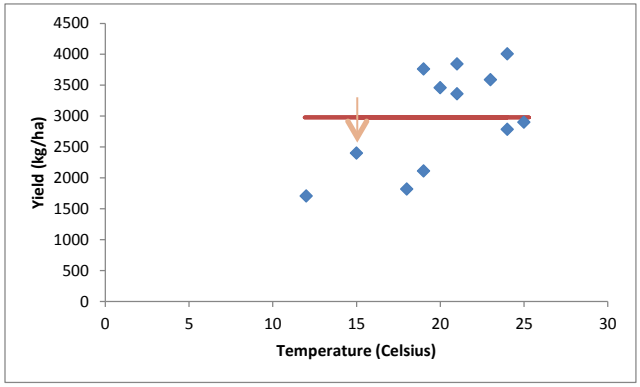
The Bias-Variance Trade-off

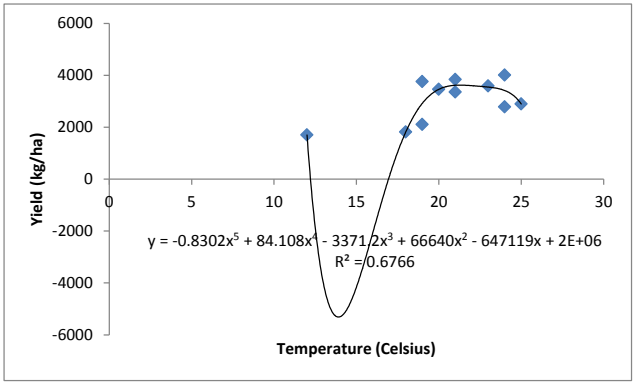
- Is Ordinary Least Squares the most powerful regression technique for prediction?
- Expected test $MSE =$
  $Var(f(x_0)) + Bias(f(x_0))^2 + Var(\varepsilon)$
- More flexible models has less bias but more variance

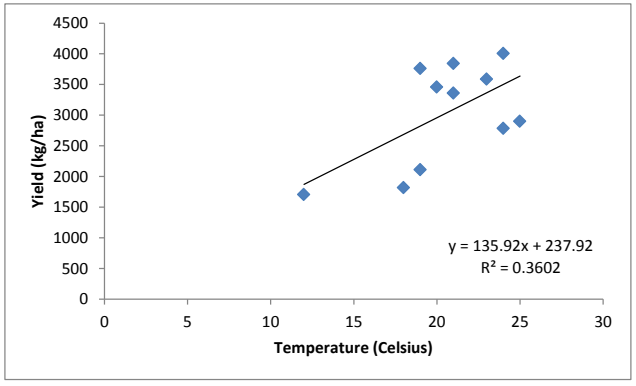
Flexibility vs. Variance

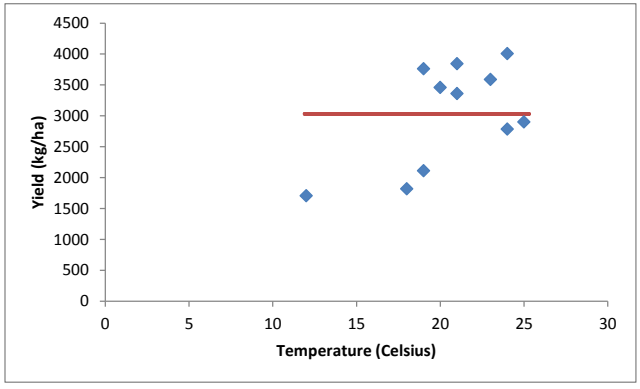












Advanced Regression Techniques

Technique	How does it work?	Significant Advantage
STEPWISE regression	Selects significant predictors	Highly interpretable
PLS Regression	Supervised dimension reduction	Overcoming the problem of a lot of predictors
PCR Regression	Unsupervised dimension reduction	Overcoming the collinearity problem
Ridge Regression	Shrinks coefficients	Powerful in prediction
Lasso Regression	Shrinks coefficients and selects significant predictors	Highly interpretable and more powerful in prediction
Elastic Nets	Combination of Ridge and Lasso	Combination of Ridge and Lasso

Materials and Methods

Oilseed rape data

- A protocol for data collection
  - Yield of oilseed rape
  - Sowing and Harvest Data, soil type, daily climatic data
- Yield data as the response variable, others as explanatory variables.

Dataset

- Denmark (data from other European countries are being received)
- 1992 to 2013
- 17021 observations (656 Standard trials)
- Daily climatic data
- Sowing Date, Harvest Date, Soil Type
- Yield data as the response variable

Future projection

- LARS weather generator
- 50 years of climate data for each time period
- Biweekly averages over the climatic data
- For this case: HADCM3 - SRA1B
  - 2011-2030
  - 2046-2065
  - 2080-2099

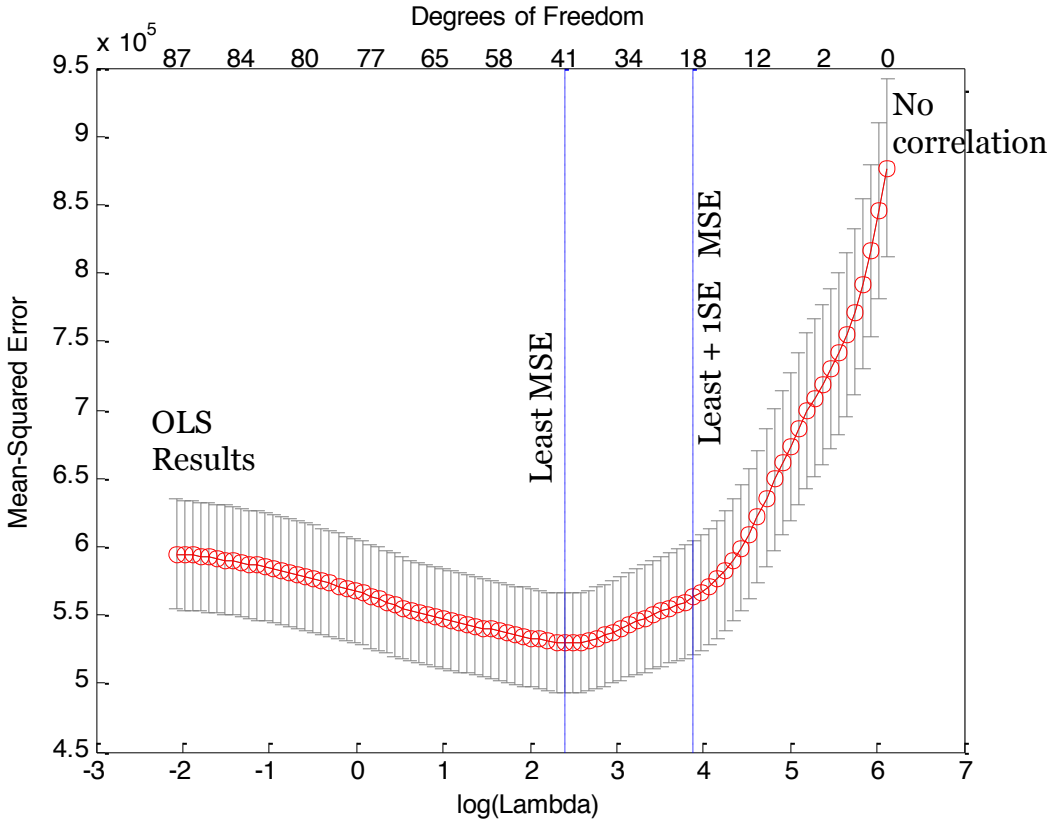
LASSO

- Assume: $Y = XB$ $(Y = b\downarrow 1\ x\downarrow 1 + ... + b\downarrow n\ x\downarrow n)$
- Lasso solves: $\min RSS + \lambda |B|$

		1	2	3	4		58	59	60	61	62	63	64	65
Temp	August	87.9429298	87.9420156	87.9410425	87.9400575	...	78.3955461	77.2110109	75.9110855	74.4842712	72.9184797	71.1999414	69.3382979	67.5399233
	September	-15.169195	-15.162977	-15.156178	-15.14877		-6.3157453	-5.5745207	-4.7607189	-3.8680091	-2.8878353	-1.8123396	-0.5876092	0
	Oktober	-49.177092	-49.166249	-49.154304	-49.141102		-44.39656	-44.452283	-44.513328	-44.580484	-44.654038	-44.734852	-44.837954	-44.945914
	November	97.8902808	97.8528036	97.8115349	97.7659412		79.1937802	79.2047999	79.2166491	79.2300039	79.2443247	79.2602378	79.26239	79.5085851
	December	-15.829326	-15.794928	-15.757063	-15.715275		0	0	0	0	0	0	0	0
	Januar	11.1103131	11.080508	11.0477025	11.0115264		0	0	0	0	0	0	0	0
	Februar	42.6088532	42.5899244	42.5691003	42.5461762		15.9411544	14.8553107	13.6638107	12.3558417	10.920641	9.34534504	7.57380464	5.04364588
	Marts	-20.621868	-20.608043	-20.592861	-20.576206		0	0	0	0	0	0	0	0
	April	313.115035	313.11448	313.113896	313.113335		291.985333	289.894515	287.599924	285.081502	282.317647	279.28426	275.938802	272.291491
	Maj	19.3431159	19.3611172	19.3809496	19.4028619		22.6609591	22.2706706	21.8424213	21.3722836	20.8564316	20.2902074	19.6150221	18.7572055
	Juni	-285.13803	-285.11493	-285.08955	-285.06164		-253.37899	-249.98263	-246.25554	-242.16448	-237.67511	-232.7477	-227.29652	-220.59935
	Juli	154.213441	154.198103	154.181254	154.162756		129.416158	126.747902	123.819665	120.605686	117.078582	113.207456	108.987037	104.313579

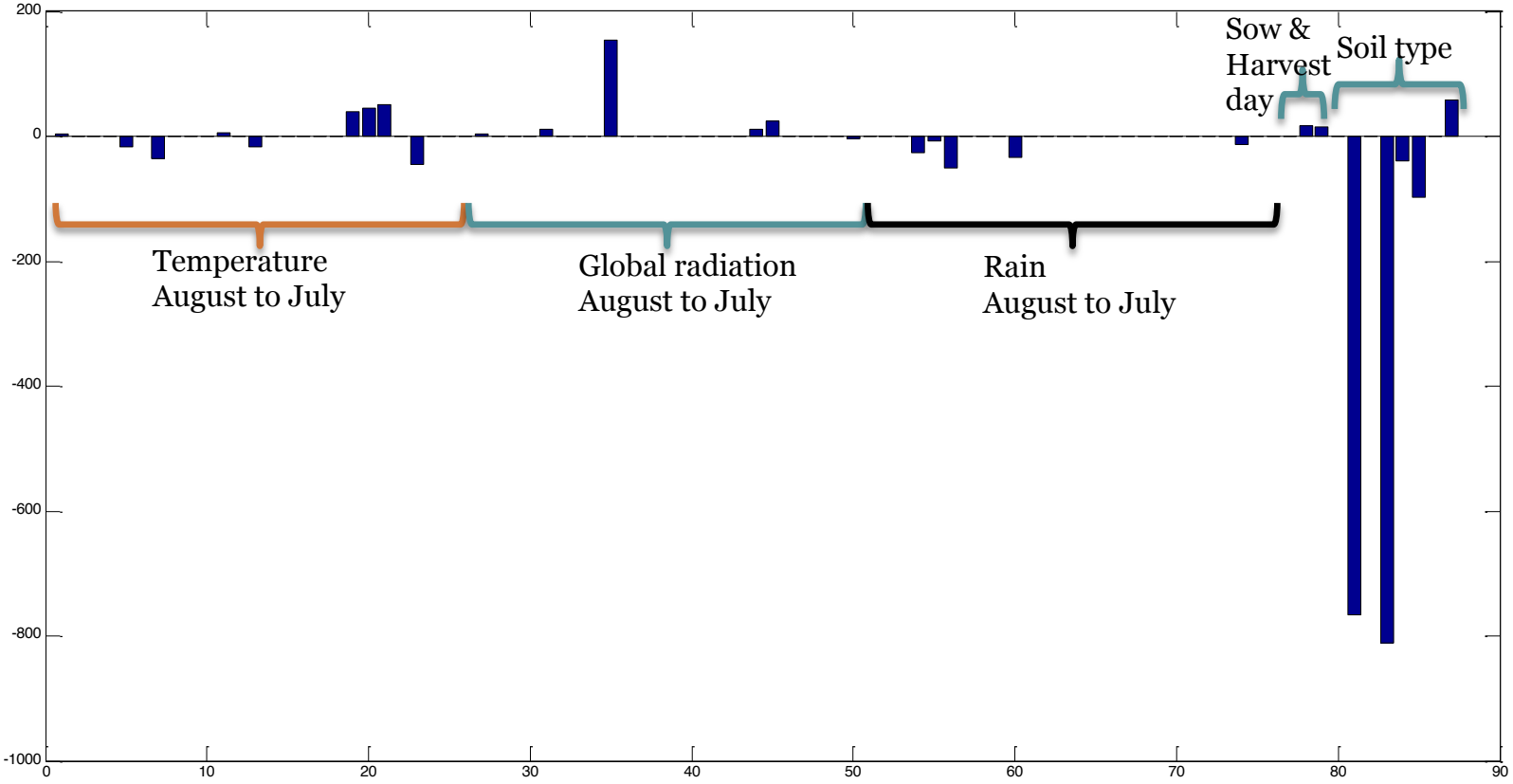
Results

Expected Test MSE

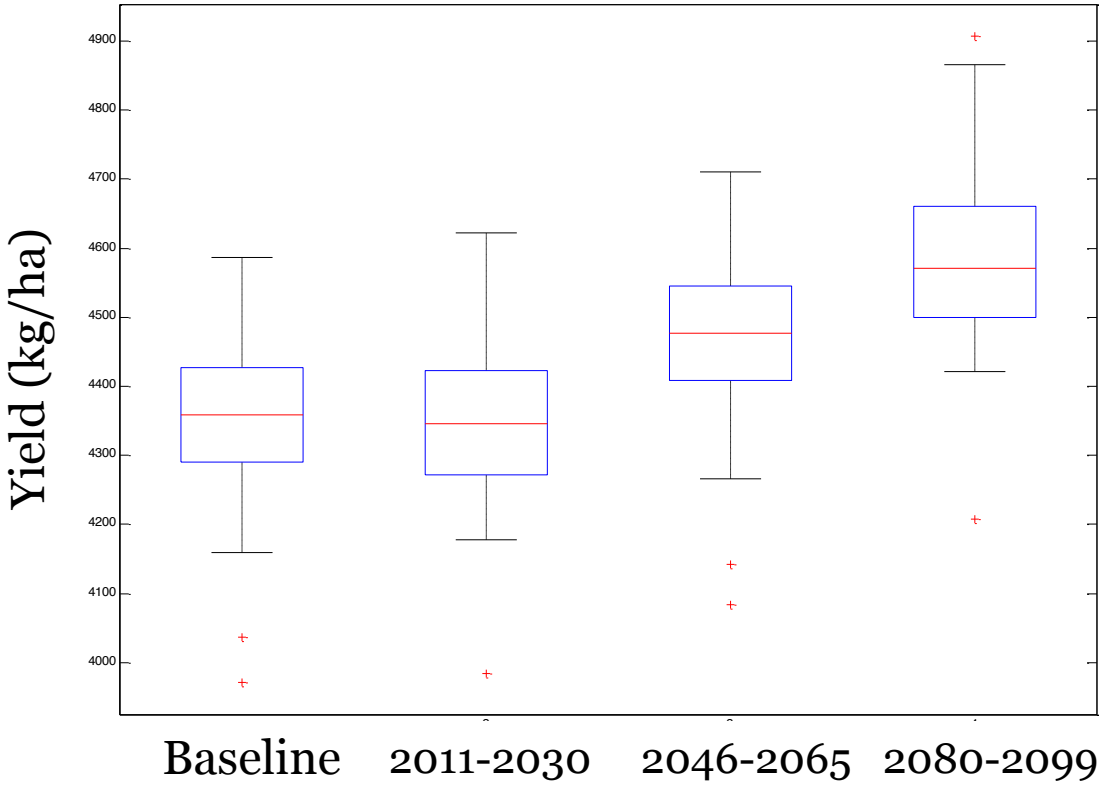


Selection of significant features

- There are 87 explanatory variables in the case of biweekly average over the climatic data.



Yield response to futue climate



- Example for
- Northern Denmark,
- Clayey soil
- no advance in breeding,
- average sow and harvest dates

Conclusion

- Statistical learning methods are useful techniques, especially for larger datasets.
- There are several easy-to-use advanced regression (and other) techniques, available in statistical packages.
- The usefulness of such methods depends on the data, spatial scales, objectives and complexities.

Thank you!