Predictor selection

Michael K. Tippett

International Research Institute for Climate and Society
The Earth Institute, Columbia University

Statistical Methods in Seasonal Prediction, ICTP
Aug 2-13, 2010

The problem of how best to select which predictors to include in a model is a nontrivial, unsolved one.

> *"All models are wrong but some are useful."*
>
> *–George Box*

The difficulty comes from having to estimate future performance from past behavior.

> *"Past performance is no guarantee of future results."*
>
> *– Any investment document'*

As a forecaster, it is better to know a model has poor skill than to mistakenly think a poor model has good skill.

> *"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so"*
>
> *– Mark Twain*

The problem of how best to select which predictors to include in a model is a nontrivial, unsolved one.

> *"All models are wrong but some are useful."*
> *–George Box*

The difficulty comes from having to estimate future performance from past behavior.

> *"Past performance is no guarantee of future results."*
> *– Any investment document'*

As a forecaster, it is better to know a model has poor skill than to mistakenly think a poor model has good skill.

> *"It ain't what you don't know that gets you into trouble.*
> *It's what you know for sure that just ain't so"*
> *– Mark Twain*

The problem of how best to select which predictors to include in a model is a nontrivial, unsolved one.

> *"All models are wrong but some are useful."*
>
> *–George Box*

The difficulty comes from having to estimate future performance from past behavior.

> *"Past performance is no guarantee of future results."*
>
> *– Any investment document'*

As a forecaster, it is better to know a model has poor skill than to mistakenly think a poor model has good skill.

> *"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so"*
>
> *– Mark Twain*

Problem

Given a pool of candidate predictors, how to do select those to include in a prediction model?

(Why not the model that best fits the data?)

Goal: a model which skillfully predicts *independent* data.

- independent from the the data used to *select* and *train* the model.

Problem

Given a pool of candidate predictors, how to do select those to include in a prediction model?

(Why not the model that best fits the data?)

Goal: a model which skillfully predicts *independent* data.

- independent from the the data used to *select* and *train* the model.

Problem

Given a pool of candidate predictors, how to do select those to include in a prediction model?

(Why not the model that best fits the data?)

Goal: a model which skillfully predicts *independent* data.

- independent from the the data used to *select* and *train* the model.

Problem

Given a pool of candidate predictors, how to do select those to include in a prediction model?

(Why not the model that best fits the data?)

Goal: a model which skillfully predicts *independent* data.

- independent from the the data used to *select* and *train* the model.

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic, all-powerful method.
- All can be tricked by screening
- Avoid methods that that are prone to constructing spurious relations. (How to check?)
- Include “screening” in predictor selection procedure.

Indirect methods (no use of independent data, depend on SSE):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic, all-powerful method.
- All can be tricked by screening
- Avoid methods that that are prone to constructing spurious relations. (How to check?)
- Include "screening" in predictor selection procedure.

Indirect methods (no use of independent data, depend on SSE):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic, all-powerful method.
- All can be tricked by screening
- Avoid methods that that are prone to constructing spurious relations. (How to check?)
- Include "screening" in predictor selection procedure.

Indirect methods (no use of independent data, depend on SSE):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic, all-powerful method.
- All can be tricked by screening
- Avoid methods that that are prone to constructing spurious relations. (How to check?)
- Include "screening" in predictor selection procedure.

Indirect methods (no use of independent data, depend on SSE):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic, all-powerful method.
- All can be tricked by screening
- Avoid methods that that are prone to constructing spurious relations. (How to check?)
- Include "screening" in predictor selection procedure.

Indirect methods (no use of independent data, depend on SSE):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Outline

- List some common methods
- Apply them to a simple example.
- Important: no magic, all-powerful method.
- All can be tricked by screening
- Avoid methods that that are prone to constructing spurious relations. (How to check?)
- Include "screening" in predictor selection procedure.

Indirect methods (no use of independent data, depend on SSE):

- F-test
- Mallow's C_P
- AIC, BIC

Direct methods (apply models to independent data):

- Split the data.
- Cross-validation

Example: DJF temperature

Predictand (y)

- Average Dec-Feb 1962-2003 temperature over land. (42 years)

Predictors (x)

- Climatology
- Sep-Nov NINO 3.4.
- Trend

File Edit View History Bookmarks Tools Help

http://iridl.ldeo.columbia.edu/

Most Visited data Seminars AMS Journals Web of Knowledge OED MathWorld E-Journals CLIO: Basic Search Media M

IRI/LDEO Climate Data Library

IRI

Data Library

expert

Finding Datasets

Browse Datasets
Browse Maproom
By Category
By Source
By Search

Help Resources

Tutorial
Statistical Analysis Tutorial
Ingrid Function Documentation
Questions and Answers
help

IRI/LDEO Climate Data Library

The IRI/LDEO Climate Data Library contains over 300 datasets from a variety of earth science disciplines and climate-related topics. It is a powerful tool that offers the following capabilities at no cost to the user:

- access any number of datasets;
- create analyses of data ranging from simple averaging to more advanced EOF analyses using the Ingrid Data Analysis Language;
- monitor present climate conditions with maps and analyses in the Maproom;
- create visual representations of data, including animations;
- download data in a variety of commonly-used formats, including GIS-compatible formats.

Are you new to the world of climate data? Check out our Introduction to Climate Data page.

What's New

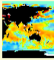
NOAA ESRL 20th Century Reanalysis Version 2 (extended) NOAA ESRL 20th Century Reanalysis Version 2 six-hourly data for 1871-2008. The analysis is performed with the Ensemble Filter as described in Compo et al. (2006) based on the method of Whitaker and Hamill (2002). Observations of surface pressure and sea level pressure from the International Surface Pressure Databank station component version 2 (Gleason et al. 2008), ICOADS (Woodruff et al. 2009), and the International Best Track Archive for Climatic Stewardship (IBTrACS, Kruk et al. 2010) were assimilated every six hours.

The IRI Server now has a local copy of all the mean fields; the IRI copy of the spread will be extended as needed. This should address some of the performance problems we were having.

Published: Mon, 24 May 2010 14:39:02 GMT

TRMM_3B42 Precipitation Estimates The combined instrument rain calibration algorithm (3B-42) uses an optimal combination of 2B-31, 2A-12, SSMI, AMSR and AMSU precipitation estimates (referred to as HQ), to adjust IR estimates from geostationary IR observations. Near-global estimates are made by calibrating the IR brightness temperatures to the HQ estimates

Mon
Global



Map
A collecti
and analy
monito
conditions.
of the map
the figures
sourc

ENS
Informati
Niño-S
Osci

Done

File Edit View History Bookmarks Tools Help

http://iridl.ldeo.columbia.edu/expert/SOURCES/.UEA/.CRU/

Most Visited data Seminars AMS Journals Web of Knowledge OED MathWorld E-Journals CLIO: Basic Search Media M

dataset: UEA CRU

Data Library

Finding Data
Tutorial
Questions and Answers
Function Documentation
help

UEA CRU[**Hulme** **Global** **New** **TS3p0** **TS2p1** **Jones**]

```
expert
SOURCES .UEA .CRU
```

ok reset

served from IRI/LDEO Clima

SOURCES UEA CRU

UEA CRU

UEA CRU: Climatic Research Unit.

Documents

overview an outline showing sub-datasets of this dataset
CRU Home Page

Datasets and variables

Global Historical monthly precipitation dataset for global land areas.
Hulme UEA CRU Hulme[**Global**]
Jones Land air temperature and sea surface temperature anomalies.
New Mean surface climate data over global land areas, including tercile and percentile data.
TS2p1 Mean surface climate data over global land areas, including tercile and percentile data.
TS3p0 TS3.0 Pre-Release (Interim) Data: Mean surface climate data over global land areas.

Done

File Edit View History Bookmarks Tools Help

http://iridl.ldeo.columbia.edu/expert/SOURCES/.UEA/.CRU/.TS2p1/.monthly/.mean/.temp/

Most Visited data Seminars AMS Journals Web of Knowledge OED MathWorld E-Journals CLIO: Basic Search Media M

data: UEA CRU TS2p1 monthly ...

IRI

Data Library

Finding Data

Tutorial

Questions and Answers

Function Documentation

UEA CRU TS2p1 documentation

help

T X Y

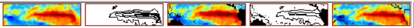
UEA CRU TS2p1 monthly mean temp[X Y | T] M M M

```
expert
SOURCES .UEA .CRU .TS2p1 .monthly .mean .temp
```

ok

reset

NEW Views



old Viewer

Data Selection Filters Data Files Tables

served from IRI/LDEO Climat

SOURCES UEA CRU TS2p1* monthly mean temperature

UEA CRU TS2p1 monthly mean temp: temperature data

monthly mean temperature from UEA CRU TS2p1: Mean surface climate data over global land areas, including tercile and percentile data

Independent Variables (Grids)

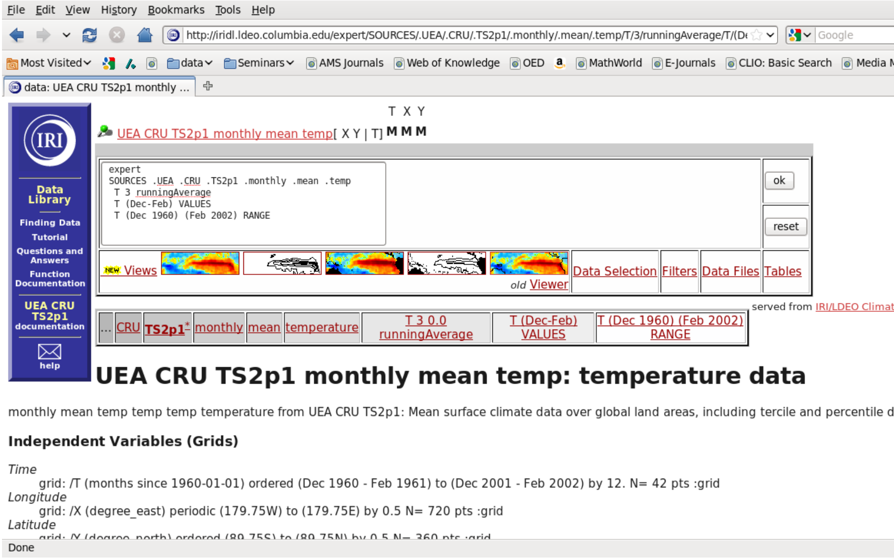
Time
grid: /T (months since 1960-01-01) ordered (Jan 1901) to (Dec 2002) by 1. N= 1224 pts :grid

Longitude
grid: /X (degree_east) periodic (179.75W) to (179.75E) by 0.5 N= 720 pts :grid

Latitude
grid: /Y (degree_north) ordered (89.75S) to (89.75N) by 0.5 N= 360 pts :grid

Other Info

Done



File Edit View History Bookmarks Tools Help
http://iridl.ldeo.columbia.edu/expert/SOURCES/.UEA/.CRU/.TS2p1/.monthly/.mean/.temp/T/3/runningAverage/T/(De
Most Visited data Seminars AMS Journals Web of Knowledge OED MathWorld E-Journals CLIO: Basic Search Media
data: UEA CRU TS2p1 monthly ...
IRI
Data Library
Finding Data
Tutorial
Questions and Answers
Function Documentation
UEA CRU TS2p1 documentation
help
T X Y
UEA CRU TS2p1 monthly mean temp[X Y | T] M M M
expert
SOURCES .UEA .CRU .TS2p1 .monthly .mean .temp
T 3 runningAverage
T (Dec-Feb) VALUES
T (Dec 1960) (Feb 2002) RANGE
ok
reset
NEW Views
old Viewer
Data Selection
Filters
Data Files
Tables
served from IRI/LDEO Climat
... CRU TS2p1* monthly mean temperature T 3 0.0 runningAverage T (Dec-Feb) VALUES T (Dec 1960) (Feb 2002) RANGE
UEA CRU TS2p1 monthly mean temp: temperature data
monthly mean temp temp temp temperature from UEA CRU TS2p1: Mean surface climate data over global land areas, including tercile and percentile d
Independent Variables (Grids)
Time
grid: /T (months since 1960-01-01) ordered (Dec 1960 - Feb 1961) to (Dec 2001 - Feb 2002) by 12. N= 42 pts :grid
Longitude
grid: /X (degree_east) periodic (179.75W) to (179.75E) by 0.5 N= 720 pts :grid
Latitude
Done

File Edit View History Bookmarks Tools Help

http://iridl.ldeo.columbia.edu/expert/SOURCES/.UEA/.CRU/.TS2p1/.monthly/.mean/.temp/T/3/runningAverage/T/(De

Most Visited data Seminars AMS Journals Web of Knowledge OED MathWorld E-Journals CLIO: Basic Search Media M

UEA CRU TS2p1 monthly mean ...

Data Library
Finding Data
Tutorial
Questions and Answers
Function Documentation
UEA CRU TS2p1 monthly mean temp dataset
help

UEA CRU TS2p1 monthly mean temp Data Files

This dataset has bytes (4.3545600E07 41.52832MB) of data in it, which should give you a rough idea of the size of any file that you ask

Download Data To Specific Software

ingrid	The Postscript-based software on which the Data Library is built.
CPT	Climate Predictability Tool More information
ferret	Interactive computer visualization and analysis software. More information
GrADS	Grid Analysis and Display System More information
matlab	Data analysis and visualization software. More information
NCL	NCAR Command Language More information
WinDisp	A public domain software package for the display and analysis of satellite images, maps and associated databases, with an e on early warning for food security. More information

Other Available File Formats

Full Information Formats These files contain all of the available metadata.	
OPeNDAP	A system which downloads data directly to software, such as matlab, Ferret, GrADS, etc. Specific instructions available in the table above. Note: OPeNDAP was formerly known as DODS (Distributed Oceanographic Data More Information
netCDF (network Common Data Form)	A commonly supported self-describing data format. More Information

Partial Information Formats

Done

Example: DJF temperature

Predictand (y)

- Average Dec-Feb 1962-2003 temperature over land. (42 years)

Predictors (x)

- Climatology
- Sep-Nov NINO 3.4.
- Trend

Consider 4 possible sets of predictors.

- Climatology
- Climatology & Sep-Nov NINO 3.4.
- Climatology & Trend
- Climatology & Sep-Nov NINO 3.4.& Trend

F-test

Compare the SSE of a P-predictor model with that of the 1-parameter reference model. (What is the 1-parameter model?)

Reference forecast = "climatology" (1-parameter model).

$$f = \frac{\frac{SSE_1 - SSE_P}{P-1}}{\frac{SSE_P}{N-P}} = \frac{SSE_1 - SSE_P}{SSE_P} \frac{N-P}{P-1}$$

where

- $SSE_1 = \sum_{i=1}^{N} (Y_i - \overline{Y})^2$ is the sum of squared error for the climatology forecast.
- $SSE_P = \sum_{i=1}^{N} (Y_i - Y_{Pi})^2$ is the sum of squared error for the model with P predictors,
- N is the sample size.

F-test

Compare the SSE of a P-predictor model with that of the 1-parameter reference model. (What is the 1-parameter model?) Reference forecast = "climatology" (1-parameter model).

$$f = \frac{\frac{SSE_1 - SSE_P}{P-1}}{\frac{SSE_P}{N-P}} = \frac{SSE_1 - SSE_P}{SSE_P}\frac{N-P}{P-1}$$

where

- $SSE_1 = \sum_{i=1}^{N}(Y_i - \overline{Y})^2$ is the sum of squared error for the climatology forecast.
- $SSE_P = \sum_{i=1}^{N}(Y_i - Y_{Pi})^2$ is the sum of squared error for the model with P predictors,
- N is the sample size.

F-test

$$f = \frac{\frac{SSE_1 - SSE_P}{P-1}}{\frac{SSE_P}{N-P}}$$

- Under the null hypothesis that the P-parameter model is not better than the 1-parameter model, f has an F distribution with parameters $(P-1, N-P)$.
- Compute the associated $\alpha = Prob(F > f)$ probability value.
- Find the model with the lowest α.
- Check that α is smaller than some limit (5%). If α exceeds the limit, use climatology forecast.

F-test

A correction is needed for multiple comparisons.

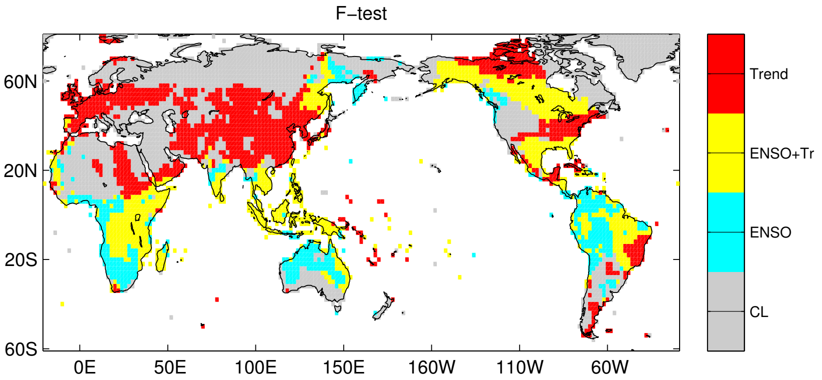
$$\alpha \to \alpha/(m-1)$$

Not quite right (not independent).

Modest values of m lead to very strict requirements on the significance level.

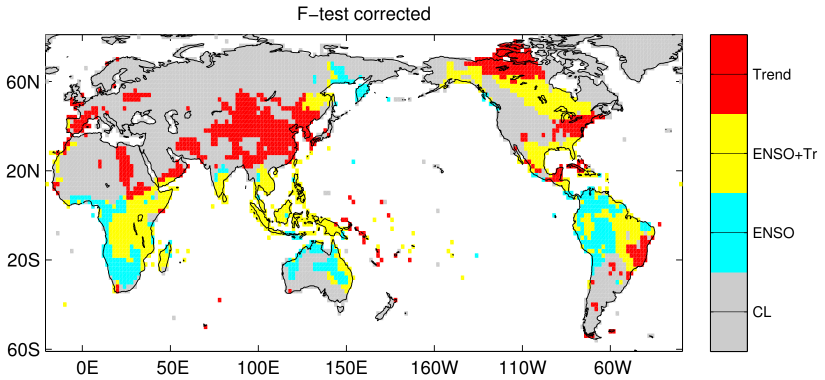
Example: DJF temperature

Models selected at each gridpoint using the F-test ($\alpha \leq 0.05$)



Example: DJF temperature

Models selected at each gridpoint using the F-test ($\alpha \leq 0.05/3$)



Mallow's C_P

$$C_P = \frac{SSE_P}{MSE_{full}} - N + 2P$$

where

- $SSE_P = \sum_{i=1}^{N}(Y_i - Y_{Pi})^2$ is the sum of squared error for the model with P predictors,
- Y_{pi} is the predicted value of the i-th observation of Y from the model with P predictors.
- $MSE_{full} = \frac{1}{N-K}\sum_{i=1}^{N}(Y_i - Y_{Ki})^2$ is the residual mean square of the model using the complete set of K predictors
- N is the sample size.

Mallow's C_P

$$C_P = \frac{SSE_P}{MSE_{full}} - N + 2P$$

If the extra variables are noise (no more variables needed)

$$E\left[\frac{SSE_P}{MSE_{full}}\right] = (n-p)\frac{\sigma_P}{\sigma_{full}} = n - p$$

and

$$E\left[C_P\right] = p$$

If the extra variables are useful (not enough variables in model), $\sigma_P > \sigma_{full}$ and

$$E\left[C_P\right] > p$$

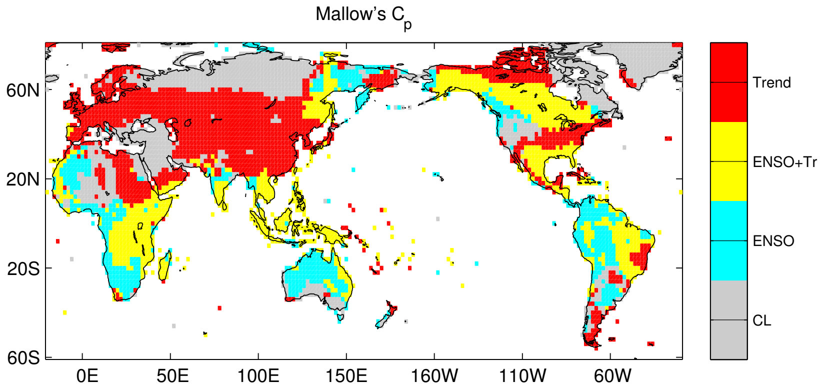
The model with the lowest C_P value approximately equal to P is the most "adequate" model.

Strategies:

- Minimize C_p.
- Graphical

Example: DJF temperature

Models selected at each gridpoint using Mallow's C_P.



AIC

Information theory measure of the difference between model and truth.

- To estimate parameters, find the most likely model (best fit) given the observations.
- This maximized likelihood (fit) is biased.
- Likelihood (fit) increases as the number of predictors increases. AIC corrects for this bias.

General case

$$AIC = -2 \log L + 2P$$

where L is the maximized likelihood of a model with P parameters.

Can be applied to any model where L is known. (Not just regression).

AIC

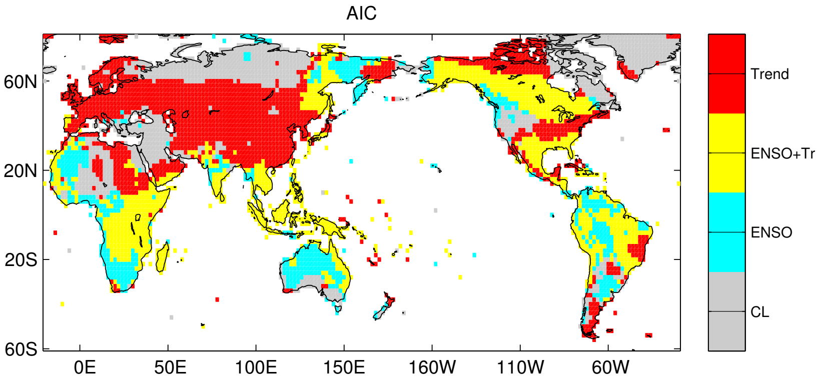
For linear regression (neglecting some constants),

$$AIC = N \log SSE_P + 2P$$

- $SSE_P = \sum_{i=1}^{N} (Y_i - Y_{Pi})^2$ is the sum of squared error for the model with P predictors,
- AIC rewards fit, penalizes complexity.
- Choose model that minimizes AIC.
- Differences in AIC are relevant.
  $\Delta < 2$ small.
  $4 < \Delta < 7$ large.
  $\Delta > 10$ very large.

Example: DJF temperature

Models selected at each gridpoint using AIC.



Corrected AIC

Correction for small sample size.
AIC is an approximation.
AICc is more accurate for small sample size.

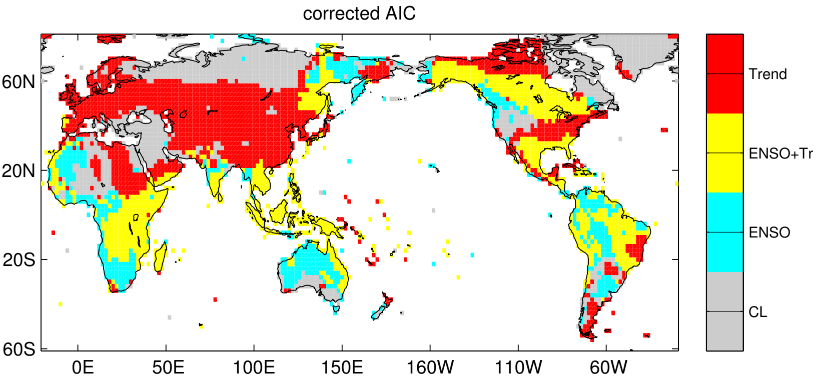
Should be used always (especially. for $N/P < 40$)

$$\mathrm{AICc} = N \log SSE_P + 2P + \frac{2P(P+1)}{N-P-1}$$

Rewards fit, penalizes complexity a little more.

Example: DJF temperature

Models selected at each gridpoint using AICc.



BIC

Approximation to Bayes factor with equally likely priors.
(AIC = Bayes factor with "savvy" prior).

General case

$$BIC = -2 \log L + P \log N$$

where L is the maximized likelihood of a model with P parameters.

$$[AIC = -2 \log L + 2P]$$

(Which picks simpler models? Why?)

BIC

For linear regression (neglecting some constants),

$$BIC = N \log SSE_P + P \log N$$

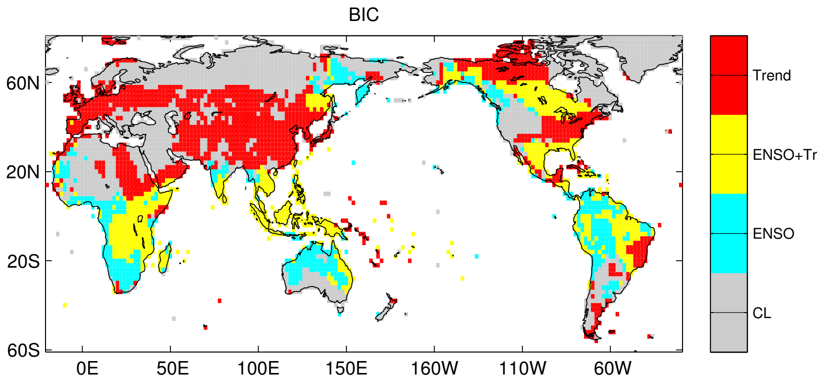
Rewards fit, penalizes complexity more than AIC.

May under-fit in small-moderate sample sizes.

AIC vs. BIC? Unsettled.

Example: DJF temperature

Models selected at each gridpoint using BIC.



Data splitting method

- Train the model on half of the data.
- Make forecasts the other half of the data.
- Choose the model with the best skill.

Is this picking the model with the best fit?

A third data set is needed to evaluate the skill of the selected model. Why?

The third data set ...

A screening example.

- Your model = 20 random numbers. `rnorm(20)`
- Generate many such models.
- Check how well each one fits the last 20 years of AIR.
- Pick the one that does best.

Skill in the selection data set is high.

Skill in an independent data set (and real skill) would be low.

Moral:

1. Avoid looking at many models.
2. Model selection and skill estimation are separate.

Avoid procedures that lead to the skill in the "third data set" being very different from that in the selection data set. (How to check?)

The third data set . . .

A screening example.

- Your model = 20 random numbers. `rnorm(20)`
- Generate many such models.
- Check how well each one fits the last 20 years of AIR.
- Pick the one that does best.

Skill in the selection data set is high.

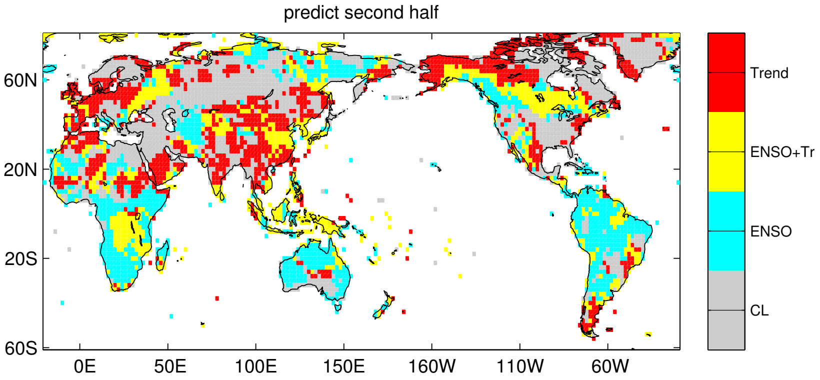
Skill in an independent data set (and real skill) would be low.
Moral:

1. Avoid looking at many models.
2. Model selection and skill estimation are separate.

Avoid procedures that lead to the skill in the “third data set” being very different from that in the selection data set.
(How to check?)

Example: DJF temperature

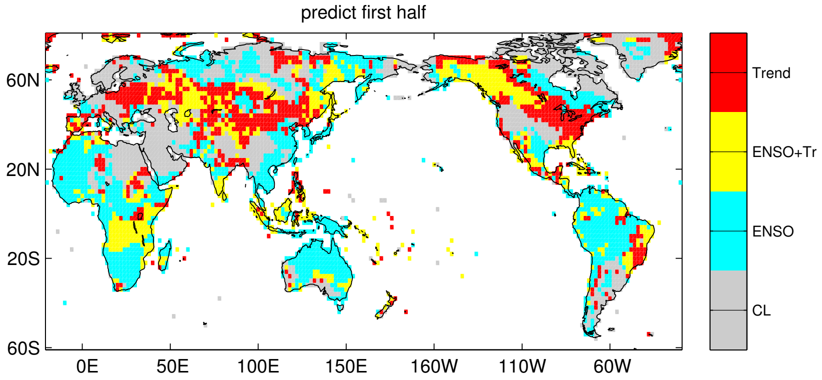
Models trained using 1962-1982 and selected at each gridpoint using skill 1983-2003.



Why noisier?

Example: DJF temperature

Models trained using 1983-2003 and selected at each gridpoint using skill 1962-1982.



Why noisier?

Cross validation

A method for mimicking actual forecasting.

An alternative to splitting the data.

- Remove some number K of samples from the data set.
- Estimate the model on the remaining $N - K$ samples.
- Use that model to predict the K left-out samples.
  - Sometimes a set of K contiguous in time samples are left out and only the middle one is predicted to deal with temporal correlation.[More later]
- Repeat.

Often $K = 1$. Leave-one-out cross-validation.

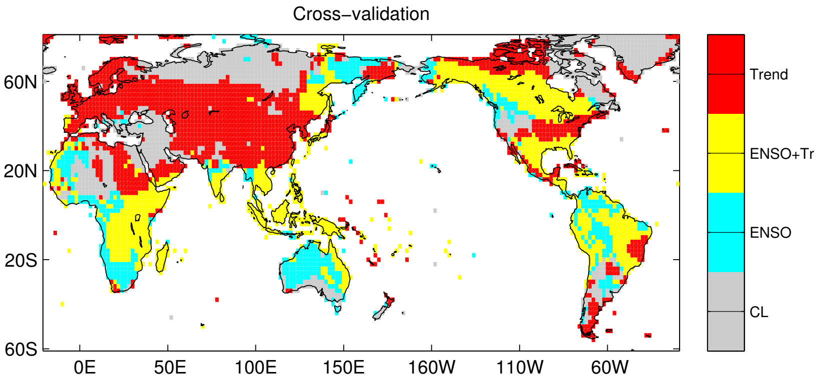
Illustration: Cross-validation in R

```
ypred = y+NA
for(ii in 1:N) {
   out = (ii-1):(ii+1)
   training = setdiff(1:N,out)
   xcv = x[training]
   ycv = y[training]
   model.cv = lm(ycv ~ xcv)
   ypred[ii] = predict(model.cv,list(xcv=x[ii]))
}
```

- R has built-in cross-validation routines
- More efficient method for leave-one-out.

Example: DJF temperature

Models selected at each gridpoint using leave-one-out cross-validation.



Summary of methods

Two types of methods

Balance between fit and number of predictors.

- F-test
- Mallow's C_P
- AIC (corrected), BIC

Apply model to independent data:

- Split data
- Cross-validation

Summary of methods

Two types of methods

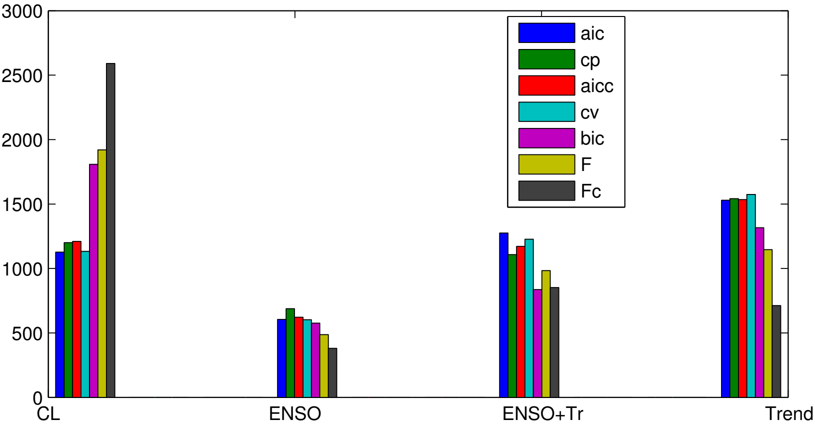
Balance between fit and number of predictors.

- F-test
- Mallow's C_P
- AIC (corrected), BIC

Apply model to independent data:

- Split data
- Cross-validation

Frequencies of the models selected



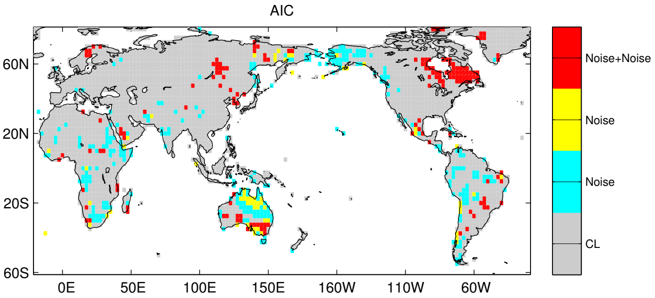
- AIC, AICc, C_p and cross-validation agree at 90% of the gridpoints.
- BIC and F-test agree in 93% of the gridpoints.
- F-test “corrected” for multiple comparisons is very strict.

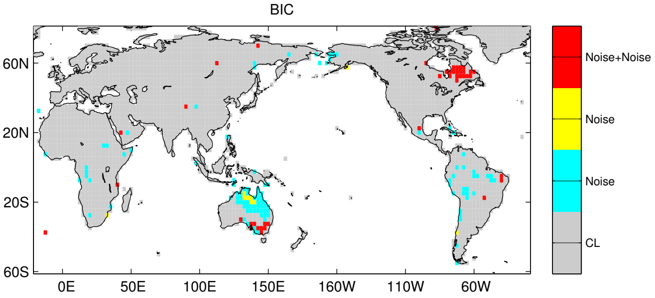
How effective are the methods?

Apply them to models with random predictors.

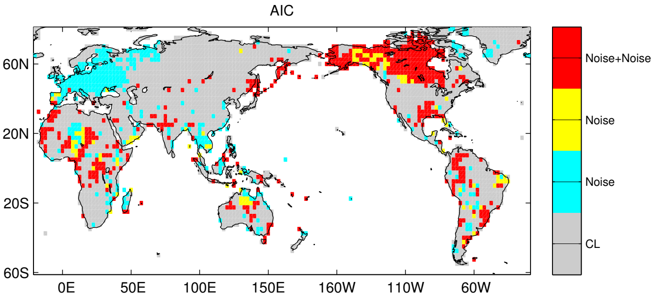
Performance across methods is more similar than different.

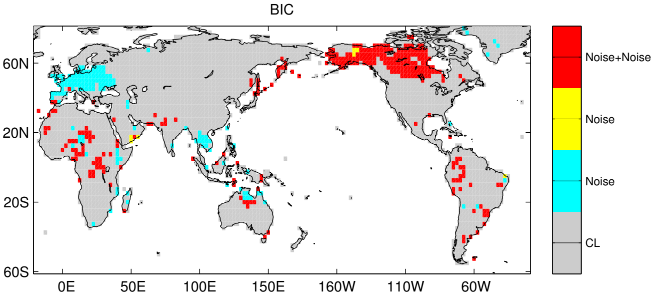
Example: DJF temperature



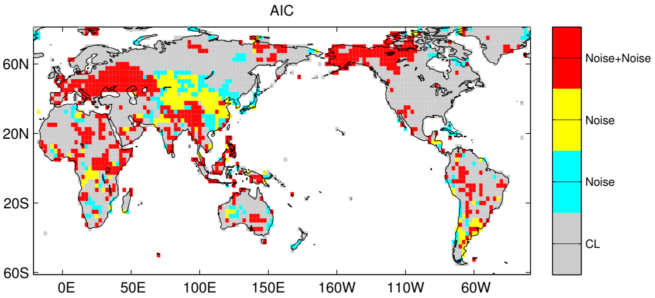


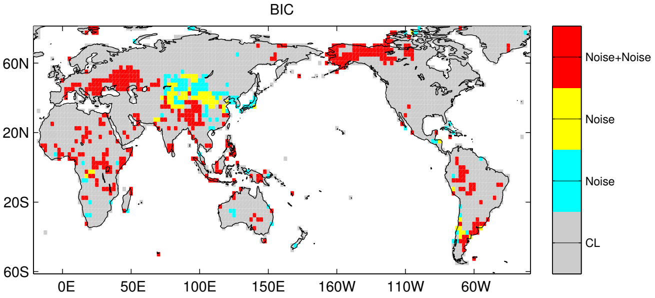
Example: DJF temperature





Example: DJF temperature





Moral

- Many predictor selection methods.
- All can be fooled given enough chances.

What can be done to avoid mishaps?