

Random reflections on data analysis and modelling

Kelvyn Jones
School of Geographical Sciences
University of Bristol

- Informal
- Arises out of questions
- Three sections
 - some organizing principles
 - some design issues (EPI)
 - developments in modelling
(GLM → GAM → GLLAMM)
- Resources and software
 - especially open-source R
 - [The Comprehensive R Archive Network](http://www.stats.bris.ac.uk/R/)
<http://www.stats.bris.ac.uk/R/>
 - Not.....neural nets, kriging.....

Some Organizing Principles I: The GEOGRAPHICAL DATAFRAME

= DATA + STRUCTURE

pH value	Geology	Distance from scarp	Agricultural Practice	Sample	Instrument	Field
5.4	Non-chalk	101.1	Pasture	1	I	A
6.9	Chalk	150.8	Pasture	2	II	A
5.7	Chalk	160.8	Pasture	3	II	A
4.3	Non-chalk	230.3	Tree	1	I	B
4.4	Non-chalk	245.8	Tree	2	III	B
7.1	Chalk	62.1	Arable	15	I	Z

NOTE

- DATA is **Mixture** of types of measurement
 - discrete & continuous:
- Different **Types** of variables (EPI)
 - outcome/ response
 - exposure: primary variable of causal interest
 - covariates/cofactors: need to condition on
 - structure ~ time/space /context, measurement instrument
- **Structure** needs to be taken into account in the analysis

Some Organizing Principles II: The PURPOSE of Modelling

What is the **quantitative** relationship between the outcome and exposure **conditional** on the co-factors and co-variates?

What is the effect on precipitation of 1 metre increase in altitude taken account of distance to the coast; rain-shadow or not?

GENERAL STRUCTURE OF A MODEL

- DATA = SIGNAL + NOISE
- DATA = SMOOTH + ROUGH
- RESPONSE = SYSTEMATIC TREND + VARIATION
- RESPONSE = FIXED + RANDOM

Response *Fixed* *Random*
RAIN ~ Alt + Shad + Dist + (Residual)

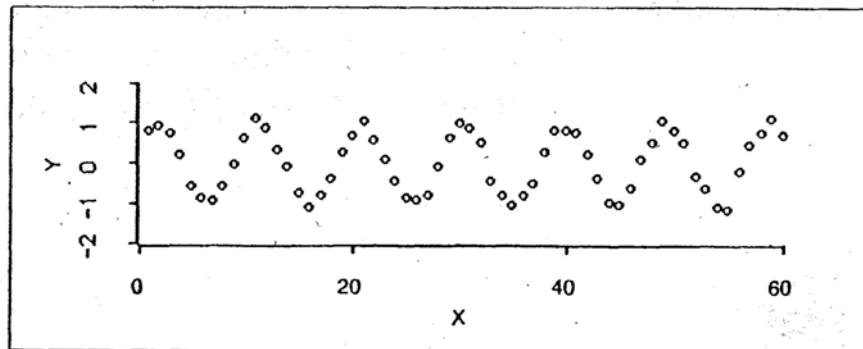
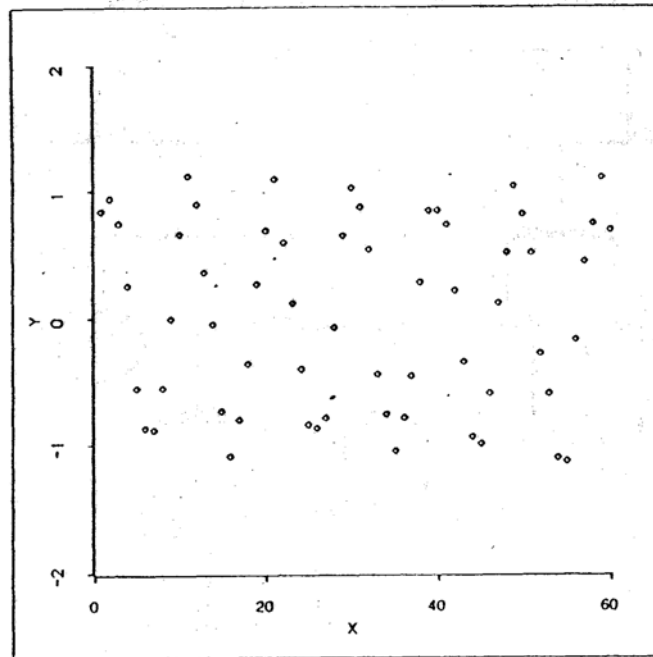
Fixed = **Averages**

Residual = **Distribution** summarised by a variance

NB

- Both aspects are important
- Possibility of “cross-contamination”

SIGNAL & Noise?



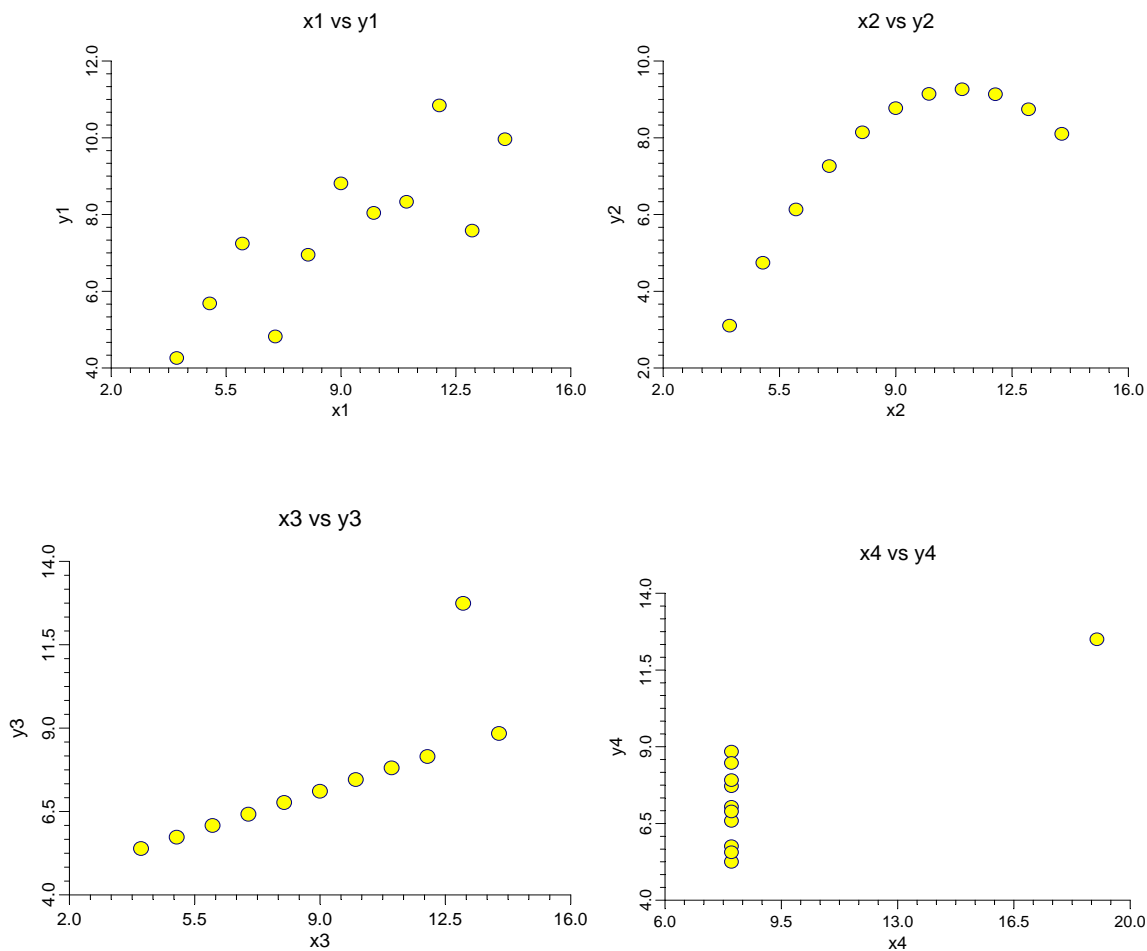
W. S. Cleveland. (1994) *The Elements of Graphing Data*. Hobart Press,

- Aspect ratio: height divided by width:
 - Top = 1
 - Automatic by Trellis graphs in R *banking to 45 degrees*

Some Organizing Principles III: The **IMPORTANT** of **GRAPHS**

- Allow to you see general picture & detail
Simultaneously

What is relation between Y and X in the 4 graphs?



In regression terms?

$$y = 3 + 0.5x ; R\text{-sq} = 0.67; \text{ same } t \text{ and } F$$

The IMPORTANCE of GRAPHS

- Plot

- before you model : outliers
- after you model: comparative size of fixed effects
- **diagnostics** for inappropriate model, over-influential data, case-deletion statistics (one fit!)

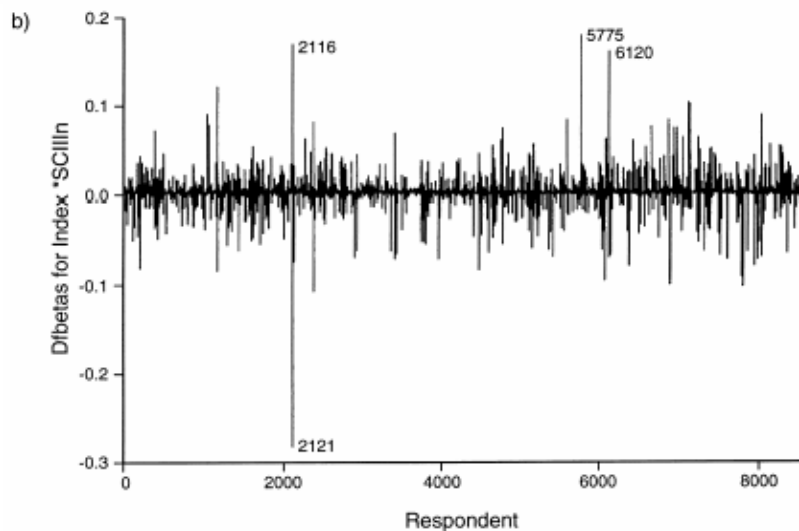
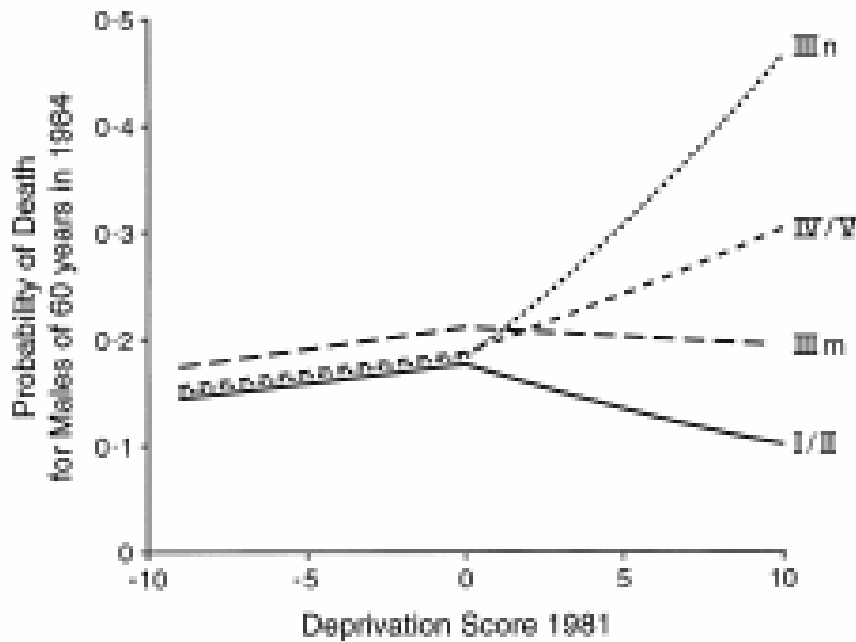


Fig. 5. Influential observations for deprivation social-class interactions.

Some Design Issues I

STATISTICAL POWER

- What does finding a non-significant result mean?
 - Either no-effect
 - Or study insufficiently **powerful** to detect effect

- The **Opposite** is also a problem
 - trivial effects are found to be significant
 - waste of resources in collecting ‘too much info’

Four outcomes of a hypothesis test

	Null Hypothesis	
Decision	True	False
Fail to reject H_0	Correct decision Prob = $1 - \alpha$	<i>Type II error</i> Prob = β
Reject H_0	<i>Type I error</i> Prob = α	correct decision Prob = $1 - \beta$

Type 1 error: When the H_0 is true & you reject it, finding significant results where there aren't any; controlled by α : the **significance** level

Type II error : not rejecting the H_0 when you should have done, controlled by β , largely ignored

Power probability of identifying significant effect when one really exists, determined by $1 - \beta$

WHAT DETERMINES POWER?

Power is **increased** in the following circumstances.

- little **noise** in the system; clear signal
- the **effect** is substantial
- α is set leniently (0.05 and not 0.01)
- large sample **size**

Can usually only do something about size

Power analysis

Prospectively: how large should my study be?

Retrospectively: when non-significant' effect found; is there sufficient power ?

Problem!

- Need to know size of effect
- Cohen's solution for very large range of procedures
 - defines small, medium and large effects
 - as ratio of effect to variation

Cohen, J (1988) *Statistical power analysis for the behavioural sciences* Lawrence Erlbaum, New Jersey

Software

PS, freely available from:

<http://www.mc.vanderbilt.edu/prevmed/ps/>

Power Calculator online

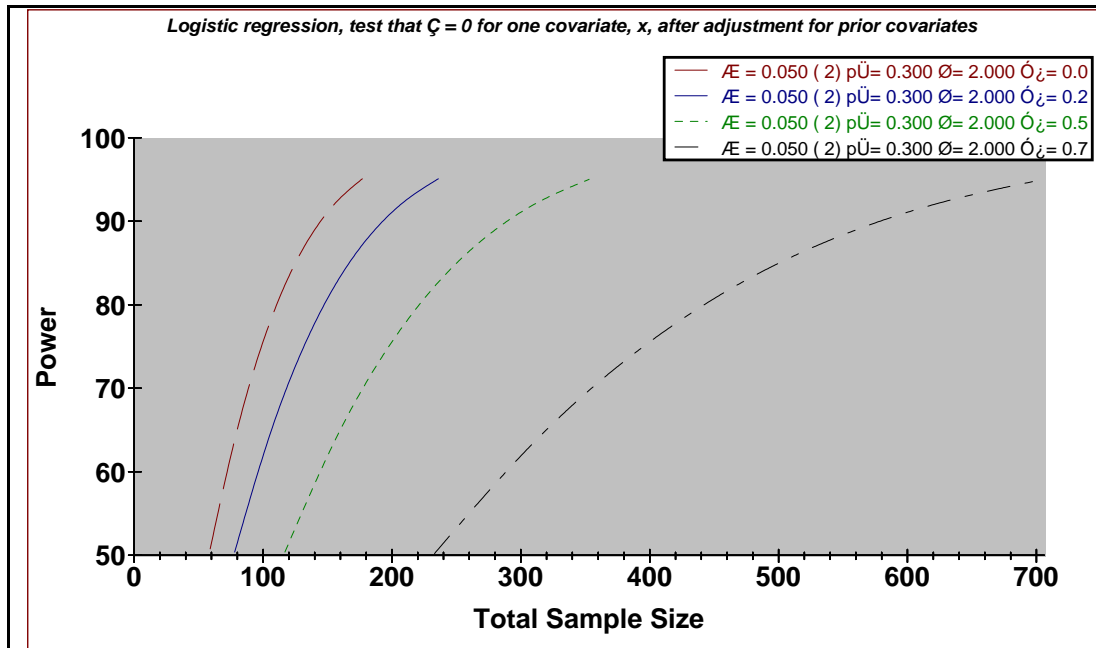
<http://ebook.stat.ucla.edu/calculators/powercalc/>

*G*Power* freely available from

<http://www.psych.uni-duesseldorf.de/aap/projects/gpower/index.html>;

'compromise power' between Type I and II errors.

EXAMPLE



- *nQuery Advisor*® industry standard

<http://www.statsol.ie/nquery/nquery.htm>

- **Binary outcome:** Dead or alive
- **Binary exposure :** Polluted or not
- **Effect size:** odds of 2; exposure doubles risk of outcome
- α : two-sided 0.05
- **Confounding** correlation between exposure and covariates linked to both exposure and outcome
 R^2 between exposure & covariates (0, 0.2, 0.5, 0.7)
- **Results** No confounding: 111 cases for power of 80%
 R^2 is 0.75, need 443 cases
- **NB** **diminishing returns of increasing size**

CASE-COMPARISONS & EFFICIENCY

Question: is there any protective effect for BCG vaccination scar on leprosy (Leprosy & TB: similar bacillus)

Study design 1: prevalence study in Central Africa (Huge!)

Data	BCG Scar	Cases	Population
	Present	101	46,028
	Absent	159	34,594
	Total	260	80,622

Results	Size	OR	Low 95% OR	High 95% OR
	80,622	0.48	0.37	0.62

odds of leprosy halved (0.48) in presence of BCG scar

Study design 2: What if case comparison design?

Drawing comparisons at random from non-cases

Comparisons

BCG Scar	Cases	a) 1:1	b) 2:1	c) 5:1
Present	101	148	296	740
Absent	159	112	224	560
Total	260	260	520	1300

Results	Size	OR	Low 95% OR	High 95% OR
	80,622	0.48	0.37	0.62
	260	0.48	0.33	0.69
	520	0.48	0.35	0.66
	1300	0.48	0.36	0.64

- same result is found as when all 80,000 used;
- little gain in precision (either 1,300 or 80,000)
- C-C highly efficient for rare outcomes

Developments in modelling I

GENERAL LINEAR MODEL

- Developments on the **right-hand side** of the equation
- **Linear** model (**continuous** predictors, **additive** and **linear**)

Eg $\text{Rain} \sim \text{Alt} + \text{Dist} + (\text{Residual})$

Results <- lm(Rain ~ Alt + Dist) *R-code*

- **General** linear model

1. **Factors** as predictors (qualitative states)

$\text{Rain} \sim \text{Alt} + \text{Dist} + \mathbf{SHAD} + (\text{Residual})$

$\text{Rain} \sim \text{Alt} + \text{Dist} + \mathbf{FullShad} + \mathbf{PartShad} + (\text{Residual})$

2. **Interactions** between variables

$\text{Rain} \sim \text{Alt} + \text{Dist} * \mathbf{SHAD} + (\text{Residual})$

3. **Non-linearity** as polynomial function

$\text{Rain} \sim \text{Alt} + \mathbf{Poly}(\text{Dist}, 3) + (\text{Residual})$

$\text{Rain} \sim \text{Alt} + \text{Dist} + \text{Dist}^2 + \text{Dist}^3 + (\text{Residual})$

4. **In combination**

$\text{Rain} \sim \text{Alt} + \mathbf{Poly}(\text{Dist}, 3) * \mathbf{SHAD} + (\text{Residual})$

General linear model (cont)

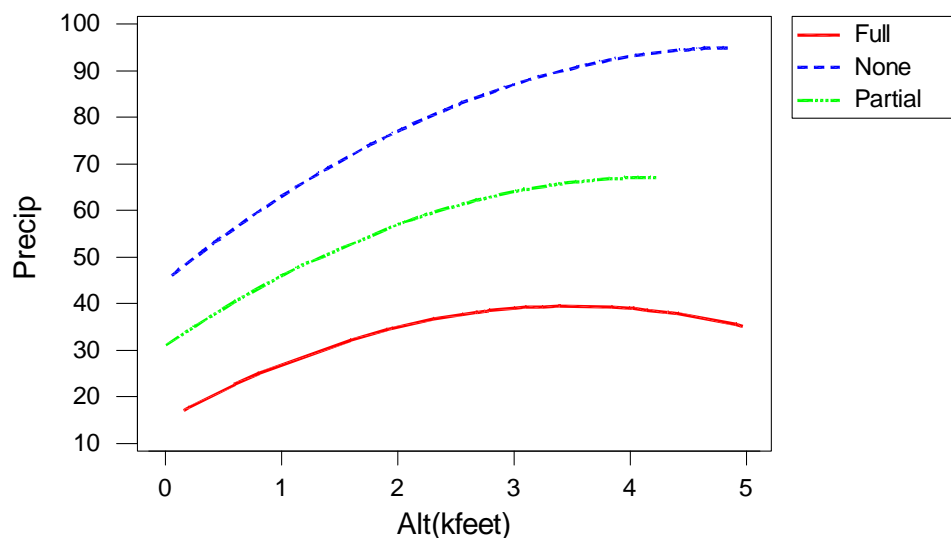
- **Dataframe**

Case	Precip	Alt(kfeet)	Shad	Full	Partial	F*Alt	P*Alt	Alt-sq
1	29.9	0.598	Full	1	0	0.598	0.000	0.3576
2	61.8	3.535	Partial	0	1	0.000	3.535	12.4962
3	71.2	2.695	Partial	0	1	0.000	2.695	7.2630
4	30.5	1.906	Full	1	0	1.906	0.000	3.6328
5	84.4	3.877	None	0	0	0.000	0.000	15.0311
6	64.8	1.402	None	0	0	0.000	0.000	1.9656
7	89.0	4.119	None	0	0	0.000	0.000	16.9662
100	83.6	3.136	None	0	0	0.000	0.000	9.8345

- **Dummy variable trick:** drop one of the qualitative states (here: None) and 1/0 code the others

Results

$$\text{Precip} = 42.8 + 20.3\text{Alt} - 25.1\text{Full} - 15.8\text{Partial} - 7.20\text{F*Alt} - 2.97\text{P*Alt} - 1.91\text{Alt-sq}; \quad \text{R-Sq} = 87.9\%$$



Developments in modelling II

GENERALIZED LINEAR MODEL

- Developments on the **left-hand side** of the equation (incorporates all that on the right)
- So far Y is **continuous** but it need not be....some examples

Type	Example	Property	Model
Count	No of rain days in a year	Positive values; discrete count	Log link and Poisson or NBD distribution
Binary	Eroding/ Not eroding	Discrete qualitative states	Logit/Probit/clog Link & binomial distribution
Ordered & unordered categorical	Eroding, stable, deposition	Multiple discrete states	Logit and multinomial distribution
Time to event	Time to death	Positive & censored	Cox model
Multiple states & episodes	Single , married, divorced	Recurrent discrete events	Event–history models

Nelder, J. A. & Wedderburn, R. W. M. (1972) Generalized linear models, *Journal of the Royal Statistical Society, A* 135(3): 370-384.

Cox, DR 1972 . Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* 34 187 220

Allison, P D (1984) *Event History Analysis*, Sage

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models* Chapman and Hall, London

EXAMPLE OF LOGIT MODEL: BINARY OUTCOME

- **Dataframe** (Atkinson and Massari, 1998 Susceptibility to landsliding in Central Apennines *Computers and Geosciences* , 373-385) KJ's re-analysis

Pixel(20*20m)	Lslide	Geology (7cat)	Dip (5 cat)
1	Yes	LAML(base)	5-20
2	No	CSSL	45-80
3	No	CSSL	20-45 (base)
1900	No	SCSD	20-45

- **Model**

```
Res <- glm(Lslide ~ Geology + Dip, binomial)
```

- **Results**

Variable	Odds	Prob	No of sites	Chi²	Prob
Geology				46.42	0.00
<i>LAML</i>	<i>1.0</i>		<i>249</i>		
CSSL	3.52	0.00	967		
LIMB	4.18	0.04	18		
LMTL	1.00	0.99	183		
MASC	0.42	0.11	242		
MLTL	2.11	0.12	78		
SCSD	1.21	0.67	163		
Dip				4.12	0.39
<i>20-45</i>	<i>1.0</i>		<i>1127</i>		
45-80	1.30	0.23	262		
5-20	1.43	0.26	177		
O'turned	0.90	0.65	285		
Vertical	1.52	0.26	49		

Geology is important Clay and sandstone 3.5, Limestone with marls below 4.2 in Cf Layered & Massive Limestone 1.

Developments in modelling III

GENERALIZED ADDITIVE MODEL

- **Relaxes** the non-linearity assumption
- Much more **flexible** form than a polynomial
- **Data-driven** via tuning parameter
 - do different degrees of roughness improve the fit?
 - Smoothest possible = straight line

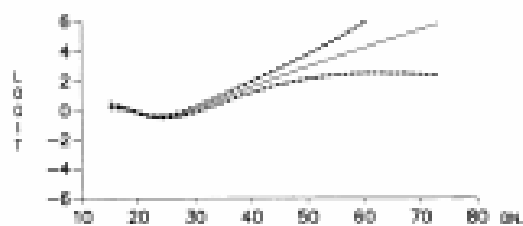
Example Jones and Almond (1992) ‘The possibilities of Generalized Additive Models’, *Transactions of the Institute of British Geographers*, 17, 434-447 uses Spot

Model:

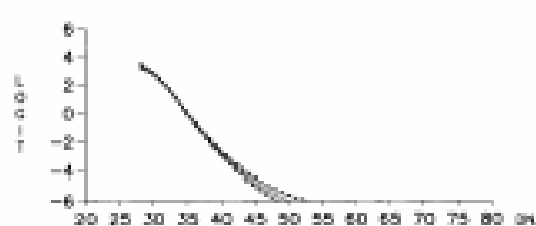
```
Res <-gam(Wood ~s(band1,5) +s(band2,5)+.., binomial)
Plot.gam(Res)
```

Results

Band 1 – Green



Band 2 – Red



Band 3 – Infra-red

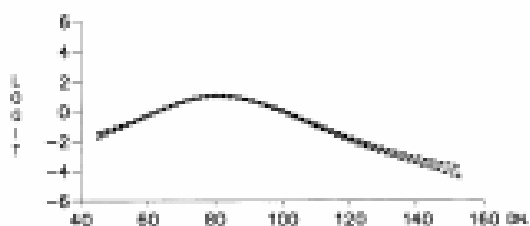


TABLE III. Analysis of seasonal presence/absence

Term	Smooth DF	Estimate	Pseudo T-ratio	Non-Linear P-val
Intercept	1	16.3		
Band 1	4-03	0.021	2.5	0.0000
Band 2	4-82	-0.49	-48.9	0.0000
Band 3	4-94	-0.015	-12.3	0.0000

Developments in modelling IV

TREE REGRESSION

Origins

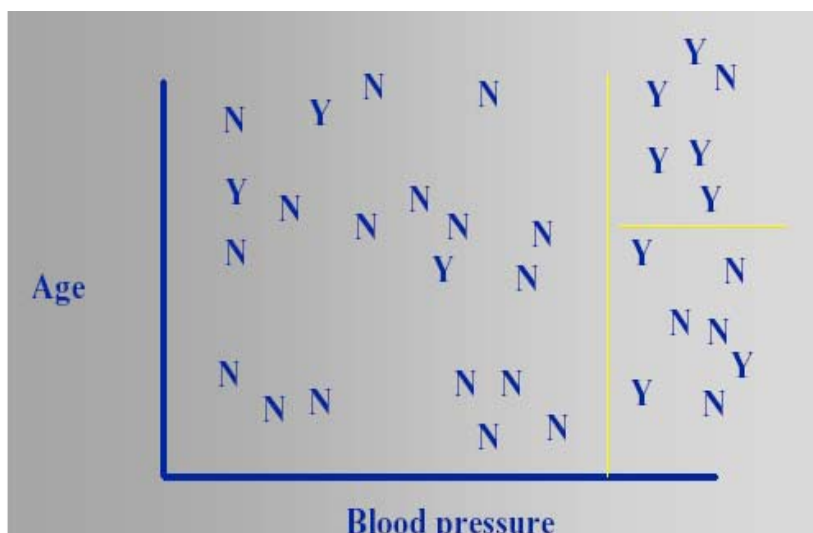
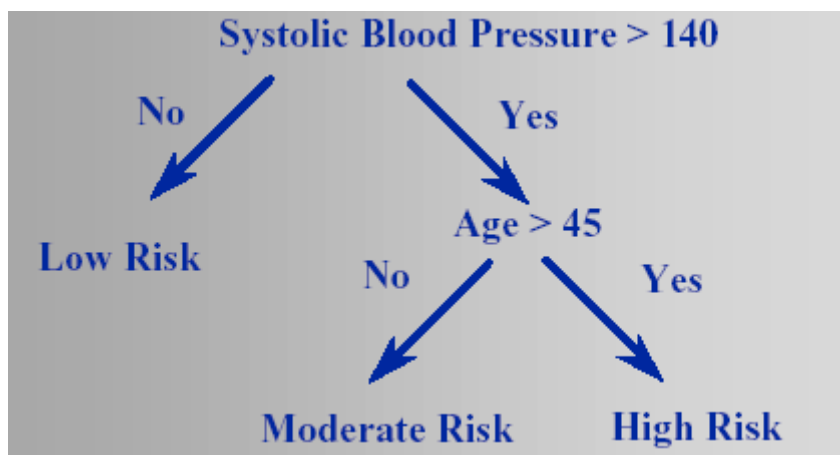
- AID
- Data mining especially marketing
- Non-additive and non-linear behaviour captured

Method

- recursively splitting on the predictors to maximize being able to predict the outcome

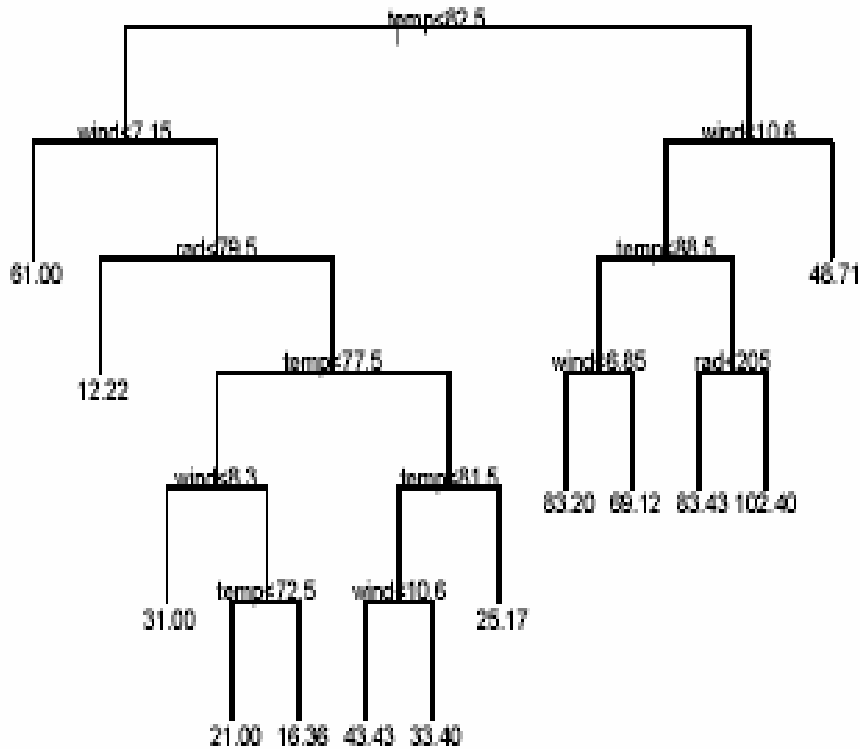
Example 1: classification tree

```
Res <- tree(risk ~ age + pressure)
```



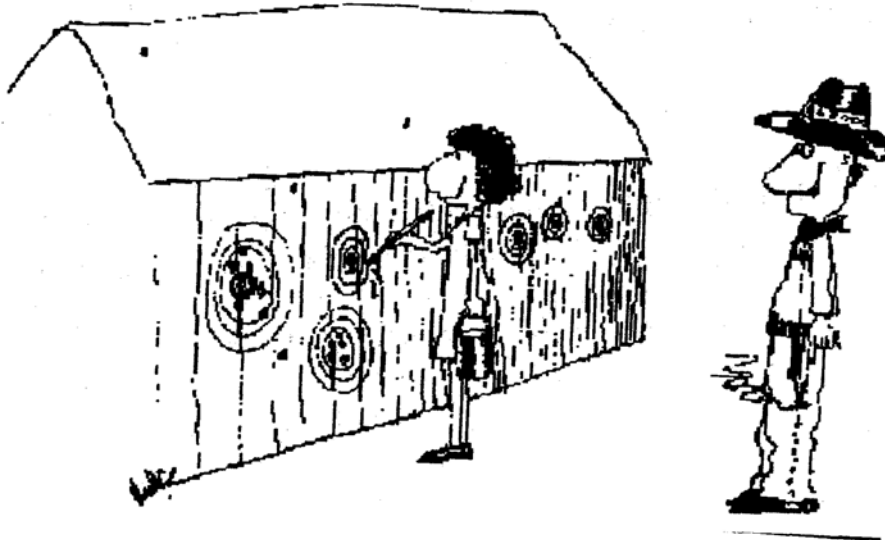
Example 2: regression tree

```
model<-tree(ozone~ temp + wind + rad)
plot(model,type="u")      text(model)
```



EG very low ozone mean = 12.22; relatively low temperatures (<82.5); high wind (> 7.15), low levels of radiation (< 79.5) high ozone mean of 102.4: high temp (>82.5), relatively still days (<10.6), high radiation (> 205).

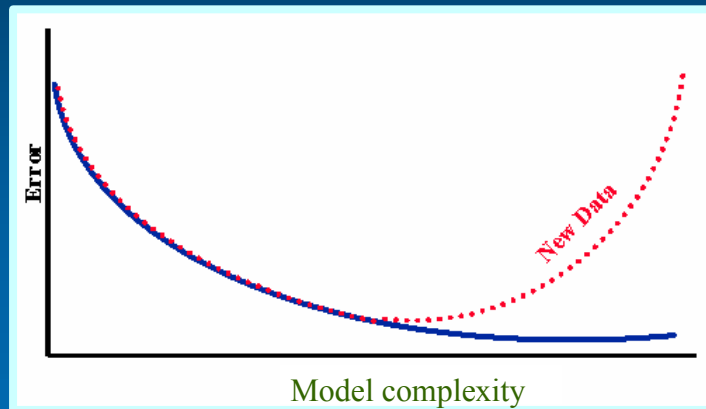
Problem: the Texas sharp shooter



TREE REGRESSION (continued)

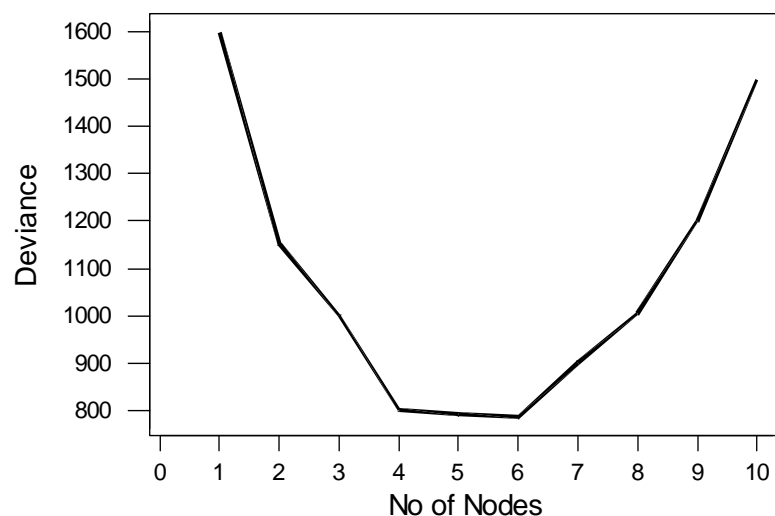
Over-fitting: modelling noise not signal

Overfitting results in the model fitting the training set too well, and in undermining the model's performance on unseen data set (with its own, yet different peculiarities)



Cross-validation (Beirman et al)

- gauges size of tree warranted by data
- random selections are omitted and predicted from tree of the remainder



4-6splits

Developments in modelling V

MULTILEVEL MODELS

- **AKA** as random co-efficient modelling; hierarchical modelling, mixed modelling; highly structured stochastic systems; generalised linear latent and mixed models
- ‘statistical models as a formal framework of analysis with a **complexity of structure** that matches the system being studied;
- GLM models **averages** (fixed) ; multilevel models averages **AND variances** (random)
- **Complexity** = ‘dependencies’ + ‘missingness’
- **Dependencies** due to context (space & time), design (eg clustered sample); measurement
- **Missingness** : not completely balanced data
- Some problems cast as **multilevel models**
 - **unit** diagrams
 - **classification** diagrams

CONCLUSIONS

Realistically complex modelling:

In COMBINATION

Complex **dependencies** due to **structures** + different **types** of responses + different **types** of predictors + **relaxing** parametric assumptions + **diagnostics**

Increasingly: Bayesian-inspired engine for calibration
MCMC
General implementation in R
Specialist software
(MLwiN , BUGS, GeoBugs, Bayes-X)

BUT KISS & caveat emptor;

NB experience with techniques in its **infancy** (Gam onwards!)

NB All based on **assumptions**

“in my experience it is scientists themselves who are the keenest purveyors of statistical snake oil
Matthews, R (2002) **RSSS news**, 30, 1-3.

Some Design Issues II

SENSITIVITY & SPECIFICITY

Synopsis

The aim of this talk is to cover in an informal way some recent and not so recent developments in data analysis. It arises out of questions from physical geography staff and postgraduates; usually after they have collected their data. Consequently I will first talk about research designs especially statistical power, sensitivity, specificity and efficiency. The second half of the talk is about modelling and covers the journey from general linear model to generalised linear model to generalised linear latent and mixed models, taking in generalized additive models and tree regression on the way! I point to appropriate software tools; and given our financial exigencies these will be open-source wherever possible.