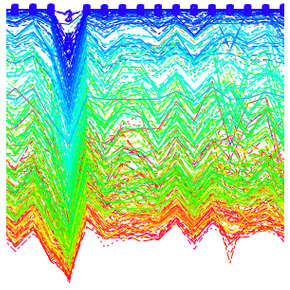
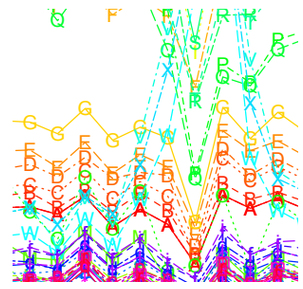
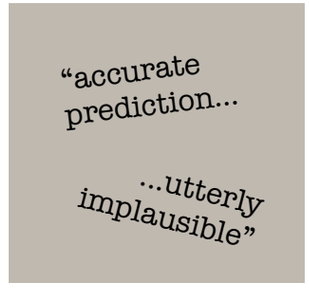
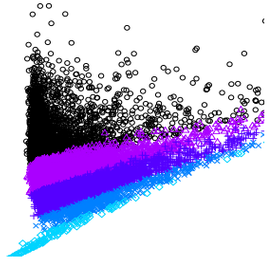
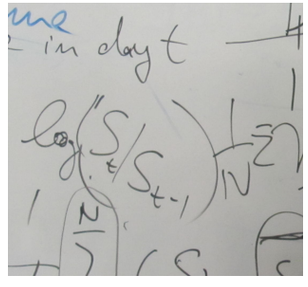


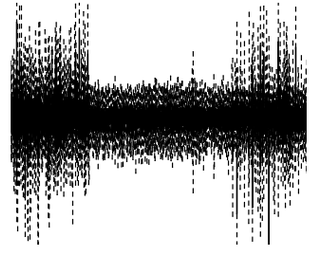


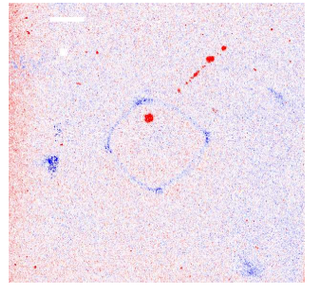


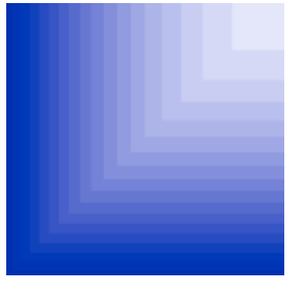


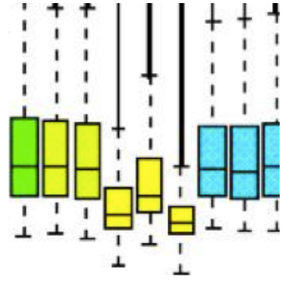
Warwick **Statistics** Research Spotlights











Waterlilies in front of the Maths and Stats building

Contents

Edited by Julia Brettschneider,
Department of Statistics at the University of Warwick

Many thanks to researchers in the Department of Statistics for contributing the material for the following pages, and also to John Rawnsley (Department of Mathematics) for photographs.

Image credits: Waterlilies (p. 2, 3) and Maths & Stats building with rainbow (p. 5) by John Rawnsley; Cover and Atrium 'The Street' (p. 4) by Julia Brettschneider.

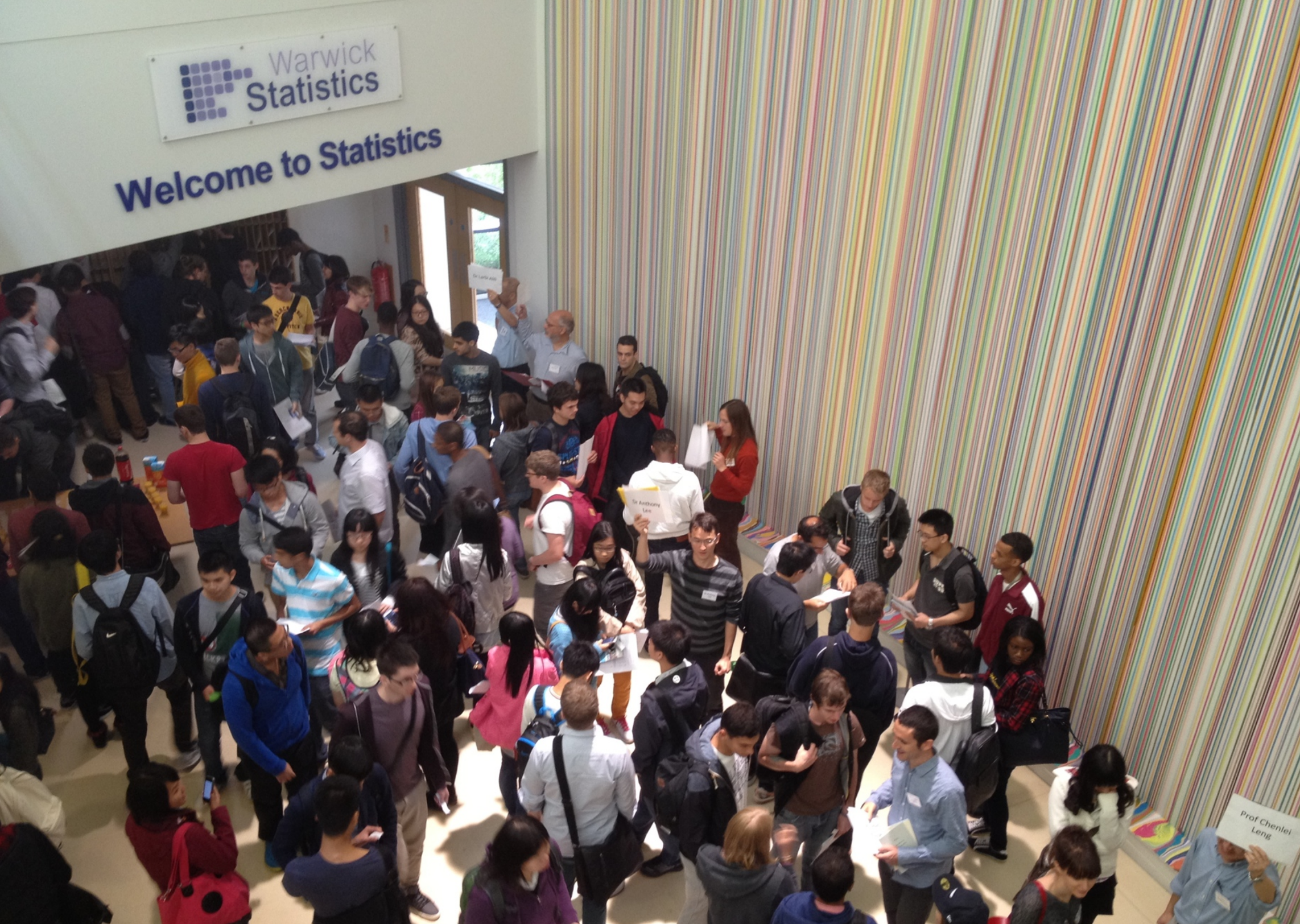
Introduction

This collection of "research spotlights" is a small selection of the innovative work that takes place in Statistics at Warwick. We hope that it gives an indication of the breadth of our research interests, and of the influence that incisive statistical methods can have on the lives and work of just about everyone in the world!

All of the material presented here is from researchers within Warwick Statistics. Often it involves collaborators elsewhere, in Europe or in other continents: our research community is global, as is the reach of our work.

Our students benefit from Warwick's high research profile in statistics and allied disciplines (including biomedical sciences, business analytics, computer science, economics, finance, management science, mathematics and the social sciences) in a variety of ways.

Research informs all of our courses, and keeps them current. The substantial funding that we win for our research helps us to provide a first-class working environment for everyone here, students included. And the wide range of research interests, not only in Statistics but also in the other departments associated with specific degree courses, means that for final-year dissertations our students enjoy an exceptionally rich choice of project topics.


Warwick
Statistics
Welcome to Statistics
Sir Anthony Leo
Prof Chenlei Leng

Getting it Right on Election Night

Professor David Firth, Warwick Statistics

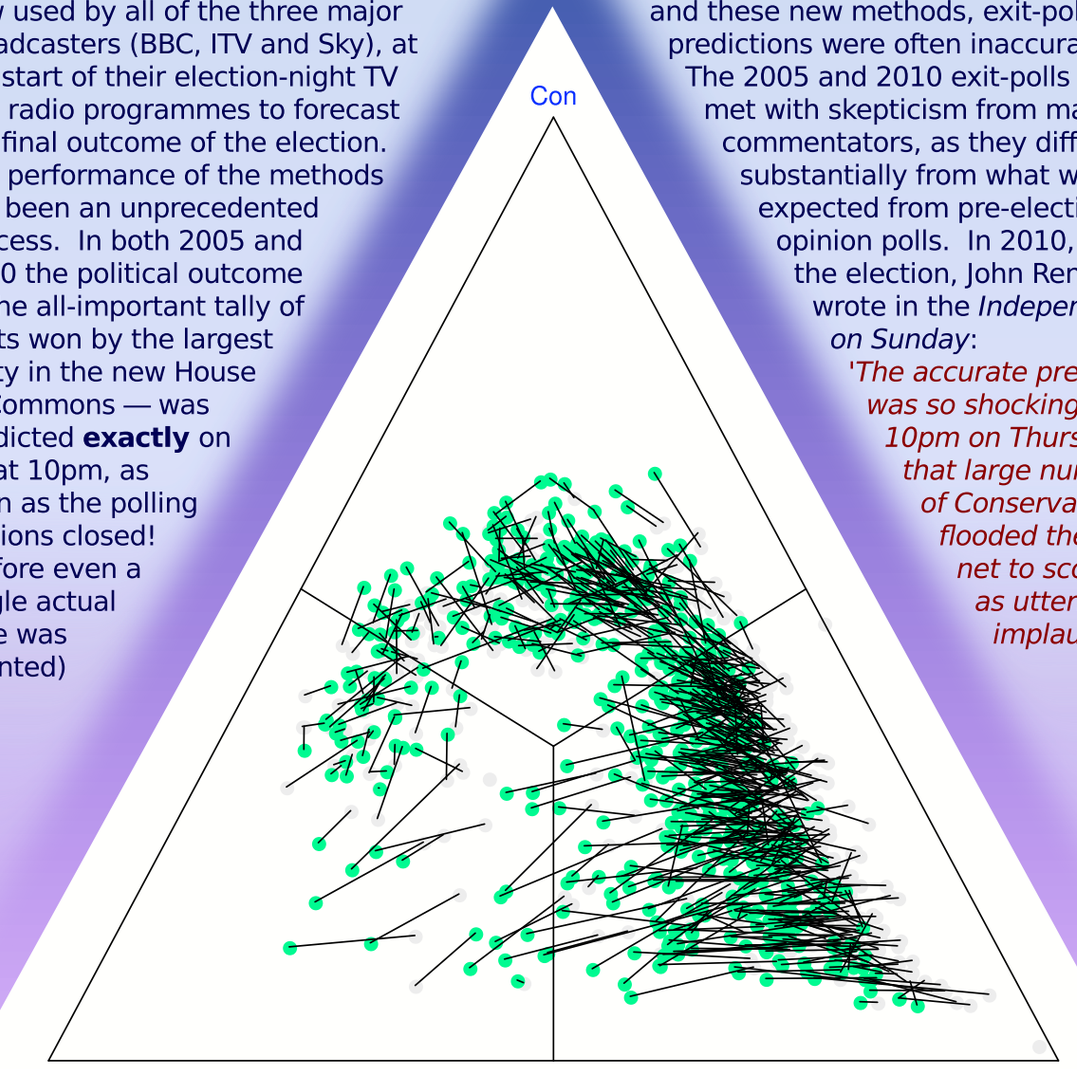
Research by David Firth has developed new statistical methods that have already been highly successful — in full public view — at the last two UK General Elections.

Projection of the predicted 2010 election outcome onto the tower of Big Ben, shortly after the 10pm close of polling stations. Photo courtesy of Gary White.

The new approach to exit polling is now used by all of the three major broadcasters (BBC, ITV and Sky), at the start of their election-night TV and radio programmes to forecast the final outcome of the election. The performance of the methods has been an unprecedented success. In both 2005 and 2010 the political outcome — the all-important tally of seats won by the largest party in the new House of Commons — was predicted **exactly** on air at 10pm, as soon as the polling stations closed! (before even a single actual vote was counted)

In previous elections, before 2005 and these new methods, exit-poll predictions were often inaccurate. The 2005 and 2010 exit-polls *also* met with skepticism from many commentators, as they differed substantially from what was expected from pre-election opinion polls. In 2010, after the election, John Rentoul wrote in the *Independent on Sunday*:

'The accurate prediction was so shocking, at 10pm on Thursday, that large numbers of Conservatives flooded the internet to scorn it as utterly implausible...'



This 3-party 'electoral triangle' shows the chaotic movement of individual parliamentary constituencies in successive general elections. The key to predictive success is a good **probabilistic model** for these changes in vote-share among the main political parties.

The new methods are fully described and analysed in
Exit Polling in a Cold Climate by J Curtice and D Firth
Journal of the Royal Statistical Society, 2008

For more information, please see
http://warwick.ac.uk/exitpolling

GAME THEORETIC ANALYSIS OF RAINBOW OPTIONS

PROFESSOR VASSILI KOLOKOLTSOV

Definition and motivation

The term goes back to Rubinstein, who describes it as a combination of a variety of assets much as a rainbow is a combination of a variety of colours. The payoff is determined by a combination of them. Multi-asset products are attractive because of inherent risk diversification, cost efficiency and opportunities for hedging against correlation.

M. E. Rubinstein, Somewhere Over the Rainbow, RISK, Nov. 1991

A sports-betting analogy

You are in a baseball tournament with three fields. One game is halfway through, a second is just starting and a third starts in an hour. You earn a profit if you pick all three winners, but you get nothing if any one team you pick is a loser.

www.investopedia.com/terms/r/rainbowoption.asp

Examples of rainbow options

- *"Best of assets or cash"*: delivering maximum of two risky assets and cash at expiry
- *"Call on max"*: holder has right to purchase maximum asset at strike price at expiry
- *"Put on min"*: holder has right to sell minimum of the risky assets at the strike price at expiry.
- *"Put 2 and call1"*: an exchange option to put a predefined risky asset and call the other risky asset

W. Margrabe (1978), The value of an option to exchange one asset for another, The Journal of Finance 23

H. Johnson (1987), Options on the maximum or the minimum of several assets, Journal of Financial and Quantitative Analysis 22

Payoffs

Best of assets or cash	$\max(S_1, S_2, \ldots, S_n, K)$
Call on max	$\max(\max(S_1, S_2, \ldots, S_n) - K, 0)$
Call on min	$\max(\min(S_1, S_2, \ldots, S_n) - K, 0)$
Put on max	$\max(K - \max(S_1, S_2, \ldots, S_n), 0)$
Put on min	$\max(K - \min(S_1, S_2, \ldots, S_n), 0)$
Put 2 and Call 1	$\max(S_1 - S_2, 0)$

Hedging and the fair value

The fair value of an option is the price at which both buyer and seller expect to break even. This is based on a mathematical model. The key idea for determining the fair value of an option is *hedging*, i.e. constructing a self-financing portfolio of financial derivatives that perfectly replicates the payoff value at the time of expiry.

J. M. Harrison & S. R. Pliska (1981), Martingales and stochastic integrals in the theory of continuous trading, Stochastic Processes and their Applications 11

Pricing rainbow options

For rainbow options, a range of methods has been used to determine the fair price. Doing this is hard, due to insufficient knowledge of the correlation structure of the multiple assets. Computational methods such as Monte Carlo can yield approximate solutions.

P. Ouwehand, Pricing Rainbow Options, WILLMOTT magazine

A game theoretic approach

The evolution of the capital can alternatively be described by as a dynamic n-step game of the investor. The approach invokes interval models and makes use of risk-neutral probability measures. Solutions are obtained by a risk-neutral evaluation of the options in minimax (robust control) situations.

The method allows transaction costs and incomplete markets to be considered. It leads to an explicit formulation and new numeric schemes. Taking a continuous time limit yields a nonlinear degenerate and/or fractional Black-Scholes type equation.

Z. Hucki, V. Kolokoltsov, Pricing of rainbow options: game theoretic approach. Int. Game Theory Review 9:2 (2007)

V. N. Kolokoltsov, Game theoretic analysis of incomplete markets: emergence of probabilities, nonlinear and fractional Black-Scholes equations. http://arxiv.org/abs/1105.3053, Risk and decision analysis, Volume 4, Number 3, 2013

MATHEMATICAL MODELS OF FINANCIAL BUBBLES

PROFESSOR DAVID HOBSON

Bubbles in history

There are examples of bubbles throughout history, including the Dutch tulip mania, the South Sea Bubble and more recent examples: the DotCom bubble, the Credit bubble, and the UK and the US housing bubbles. When a bubble is followed by a CRASH! it invariably leads to suffering.

Great Financial Scandals, Sam Jaffa, Robson Books 1998

What is a bubble?

It occurs when investors are prepared to pay more for an asset than it is worth, and there is a divergence between price and the fundamental value. Why might this happen? One explanation is the "greater fool" theory: It is rational to knowingly pay for an asset if you believe you will be able to find a greater fool in the future who will pay even more.

A Random Walk Down Wall Street, B. Malkiel, Norton, 2003

Where do bubbles come from?

"We find that whole communities suddenly fix their minds upon one object, and go mad in its pursuit; that millions of people become simultaneously impressed with one delusion, and run after it, till their attention is caught by some new folly more captivating than the first. Men, it has been well said, think in herds; it will be seen that they go mad in herds, while they only recover their senses slowly, and one by one."

Extraordinary Popular Delusions & the Madness of Crowds, C. McKay, 1841

Stylised facts about bubbles

- Asset price bubbles coincide with increases in volatility and trading volume (the Roaring 20s and the Internet Bubble)
- Asset price bubbles coincide with financial or technological innovations (from railroads to biotechnology)
- Asset price bubble implosions (crashes) coincide with increases in asset supply (the South Sea Bubble, CDOs and CDO squareds)

Speculation, Trading and Bubbles, J. Scheinkman, SSRN, 2013

Detecting bubbles

It is often very difficult to recognise a bubble until after it has burst. After all, in order to decide if there is a bubble you need to be able to calculate the fundamental value. But the local martingale model provides a way to identify bubbles whilst they are happening and before they burst.

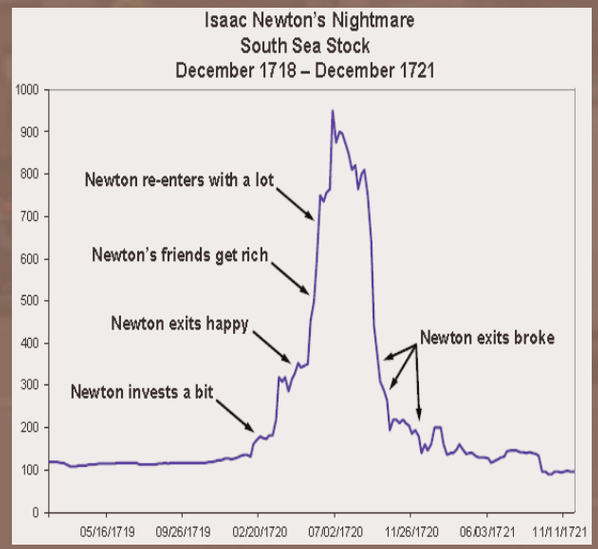
How to Detect an Asset Bubble, R.Jarrow, Y. Kchia and P. Protter, SSRN, 2011

Models of bubbles

The modelling of bubbles in financial mathematics has focussed on local martingale models. A martingale is a random or stochastic process which, on average, is as likely to go up as much as it goes down, and so on average stays the same. (The study of random walks and martingales is one of the highlights of undergraduate probability theory.)

Bachelier studied martingales in his ground-breaking thesis on the theory of speculation. A local martingale has the same fair game property, but only locally (which is enough to rule out arbitrage), and globally prices can fall, even on average.

Local Martingales, Bubbles and Option Prices, A. Cox and D. Hobson, Finance and Stochastics, 2005



PROSPECT THEORY AND THE DISPOSITION EFFECT

DR VICKY HENDERSON

1. Behavioural bias

Lab experiments run by psychologists over the last few decades have uncovered a wealth of biases in the way we make decisions under uncertainty. Our decisions systematically violate the predictions of rational expected utility theory.

The topic has recently caught more attention by the media, e.g. on 24.2.2014, BBC2 aired *Horizon* episode *"How you really make decisions"* focussing on these biases and their implications for society.

- *Statistics* is used to analyse large datasets to identify evidence of biases.
- *Probabilistic models* can be designed to capture and model behaviour under biases.

2. Disposition effect

Investors disproportionately sell winners and are reluctant to sell assets trading at a *loss* relative to purchase price.

There is widespread evidence of the disposition effect in experimental lab studies, in datasets from real estate markets, in traded option markets, in executive stock options, and in corporate investment decisions, to name a few (see references).

3. Prospect theory & Nobel Prize

The pioneering psychologists Kahneman and Tversky proposed 1979 a new model for decision making under risk based on their experimental evidence. Kahneman received the Nobel Memorial Prize in Economics in 2002.

4. Prospect theory: more details

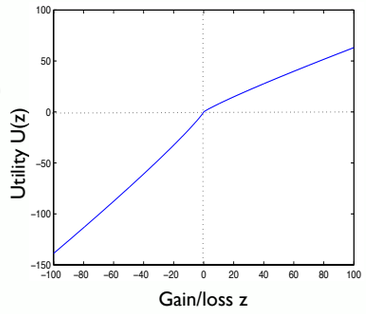
- Utility is defined over *gains* and *losses* relative to a reference point and exhibits *loss aversion*.
- *Concave* over gain, *convex* over losses (S-shaped)
- *Probability weighting*: people overweigh the tails of the distribution.

Kahneman & Tversky (1979) proposed:

$$U(z) = z^{\alpha}; z \geq 0$$
$$U(z) = -\lambda(-z)^{\alpha}; z \leq 0$$

where z is dollar gain or loss (with reference 0).

They obtain estimates:
$\alpha = 0.88, \lambda = 2.25$



5. An explanation?

Shefrin and Stratman (1985) argued, via intuition, that prospect theory explained disposition behaviour. Why? An investor who is facing a loss will tend to *gamble* on the possibility of breaking even (due to convexity of S-shape over a loss).

However, it is not so simple! A number of researchers have attempted to capture this in stochastic models recently. Henderson (2012) proposes and solves a stochastic optimal stopping model to describe the trading decision of an investor according to prospect theory.

6. A model and solution

Key idea is to solve an optimal stopping problem:

$$sup_{\tau}\mathbb{E}U(Y_{\tau} - R)$$

where asset price Y is a time homogeneous diffusion, R denotes the reference level/purchase price of the asset and τ is the stopping time at which the asset is sold.

- Using time homogeneity and a martingale transformation, the structure of the solution is to stop (sell) when Y exits an interval. Consider such stopping times and choose "best" interval.
- Several possible scenarios emerge. The most interesting is when the investor has a two-sided threshold and thus may sell at a gain or at a loss, depending on which threshold the price reaches first.
- We show the rate of selling at a gain is much higher than the rate of selling at a loss - hence a pronounced *disposition effect* emerges.

7. Where next?

- Explore impact of probability weighting on investor trading decisions (with Alex Tse, PhD student)
- Dynamic model involving "Realization" utility
- Incorporate other behavioural biases into stochastic models of investor trading or portfolio optimization

References

Henderson V., 2012, Prospect Theory, Liquidation and the Disposition Effect, Management Science, 58, 2.

Kahneman D. and A. Tversky, 1979, Prospect Theory: An Analysis of Decision under Risk, Econometrica, 46, 171-185.

Odean T., 1998, Are Investors Reluctant to Realize their Losses?, Journal of Finance, 53, 5.

Shefrin H. and M. Statman, 1985, The Disposition Effect to Sell Winners Too Early and Ride Losers Too Long, J. of Finance, 40

USING BAYESIAN NETWORKS FOR FORENSIC DNA INFERENCE

DR ANJALI MAZUMDER AND PROFESSOR STEFFEN LAURITZEN

Introduction

What is a BN?

A Bayesian Network (BN) is a graphical model for expressing probabilistic relationships among a set of measured variables. Each node in the graph represents a variable (or event), and has a table of probabilities associated with that variable. Arrows between nodes describe associations between the variables in the graph.

Why Use a BN?

BNs provide graphical representations of very complex problems. They provide a computational alternative to complex algebraic manipulations required to solve these problems. BNs can be used as a tool by forensic scientists and lawyers to analyze evidence, construct and communicate evidence, and assess the value of possible investigating ways.

Motivation

Forensic scientists are often called upon in courts to give expert testimony, for example, source of a DNA sample. Prior to making a (often numerical) judgment about the source of the sample, they are routinely challenged to make decisions under uncertainty, for instance: which genetic marker to type or how many to type.

BN formulation

Target node Q, represents a query variable, generally impossible or difficult to observe, with a finite number of states, and information node X. X represents evidence variable, generally observable, with a fine number of states. The direct edge shows that the query of interest Q is relevant for evidence X.

Often, the evaluation of evidence involves estimating unknown quantities, say Q, from some given observations X. Often there is a quest for data to reduce the uncertainty which is seldom cost free.

Thus, the Inference or reasoning problem assesses the weight of evidence, and the planning or decision-making problem determines the value of observing the evidence.

Inference problem

The task of reasoning involves calculating the weight of evidence in the form of a likelihood ratio:

$$LR = \frac{\Pr(X|q_1)}{\Pr(X|q_2)}.$$

In determining which evidence contributes to the query, we define the informativeness I_q for this scenario as:

$$I_q = H(Q) - H(Q|X),$$

where $H(X)$ denotes the entropy of the distribution of X

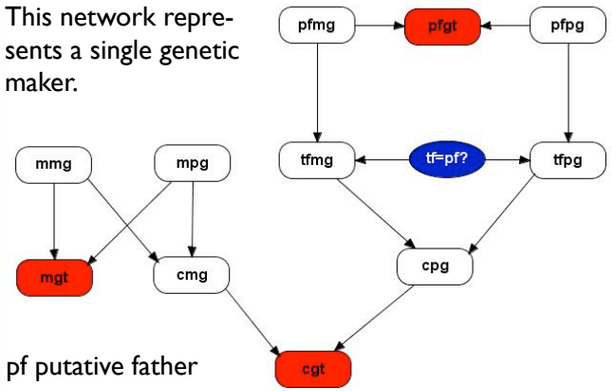
$$H(x) = p_x \log p_x,$$

a measure of the total uncertainty of the distribution. The

BN Representation for Paternity

Consider a simple paternity identification problem, shown using a BN representation where query Q is represented by node tf=pf? The query of interest is whether the true father is the putative father or a man drawn randomly from the population.

This network represents a single genetic maker.



pf putative father

gt (observed) genotype

pg (unobserved) paternal gene

Mg (unobserved) maternal gene

To obtain probabilities of interest, we enter evidence into the observed nodes and propagate the evidence.

Results

Inference Problem

Suppose we have the following case evidence at a single genetic maker, FES: cgt = {12,12}, mgt = {10,12}, and pfgt = {10,12}, and population allele frequencies for 10 and 12 are, respectively, 0.28425 and 0.25942. The *LR* in favour of paternity based on this data at this marker:

$$LR = \frac{\Pr(\text{cgt,mgt,pfgt}(|q_1)}{\Pr(\text{cgt,mgt,pfgt}|q_2)} = \frac{0.6584}{0.3416} = 1.9274.$$

LRs for each genetic maker is calculated and multiplied to give the joint LR, providing the weight of evidence.

Suppose we need to make a decision to observe either one of two genetic makers. The quantity I_q measures the reduction in uncertainty in Q due to observation of the genotypes of the associated indivuals. For paternity

$$I_q = H(Q) - H(Q|\text{CGT}, \text{MGT}, \text{PFGT}),$$

measures contribution of a genetic marker. For Caucasian population, paternity informativeness I_q of genetic markers FGA and THO1 are 0.3672 and 0.2576, which can aid the forensic scientist to decide which genetic marker(s) to measure based on expected reduction of uncertainty.

Conclusions

BNs are a useful tool for DNA evidence evaluation, allowing scientists to calculate laborious marginal and conditional probabilities easily. With software programs such as HUGIN, complex networks are relatively simple to create and calculations readily accessible accounting for complex dependence relationships between variables. BNs provide a vehicle to communicate and investigate the value of evidence for any forensic query of interest.

References

Dawid, A.P., Mortera, J., Piscali, V., and Boxel, D.V. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics* 29, 577-595.

Lauritzen, S.L.and Mazumder, A. (2008) Informativeness of genetic markers for forensic inference – An information theoretic approach. *Forensic Science International: Genetics Supplement Series* 1, 652-653.

Bayesian Networks for Food Security

Professor James Q. Smith, Dr. Martine J. Barons and Xiaoyan Zhong
Contact: Martine.Barons@warwick.ac.uk go.warwick.ac.uk/MJBarons

Warwick Statistics

THE UNIVERSITY OF WARWICK

"Food security exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life" (FAO, 1996)

1. Motivation

Food security is influenced by a wide variety of factors, e.g. climate, farming and subsidies, business practice, energy costs, politics, etc. When designing policies to promote food security, it is necessary to take account of all the relevant factors. A decision support system needs to be capable of combining these in a coherent way.

2. Why Bayesian Networks?

•Decisions may need to be made under uncertainty, e.g. crop yield, price.
•Many elements of the system depend on and affect each other.
•It is impossible for a single decision-maker to be expert in all the topics

3. The sugar industry example

Sugar can be used for food or biofuel production, and is grown as beet in the UK and as cane largely in Brazil, making it an interesting example.

3a.Influence diagram

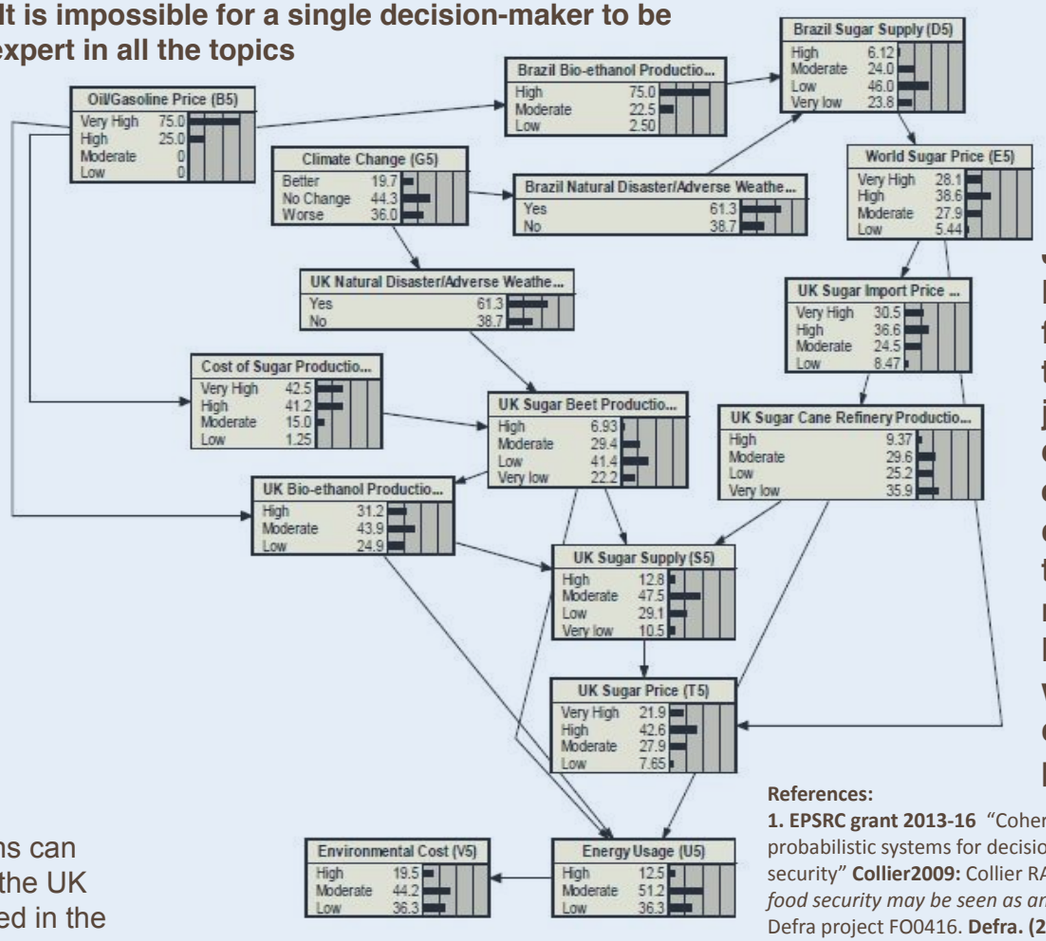
An influence diagram captures the most important elements' identified with the help of sugar industry experts.

3b. Elicitation

This Bayesian network was built using the experts' opinions about how much changes in one variable would affect another. Professor Ben Richardson is a Warwick expert in the sugar industry and was a valuable source of information along with experts from China, giving an international perspective.

3c. Bayesian network

After the expert opinion is added, 'What if..?' questions can be asked. This Bayesian network (right) shows how the UK sugar supply and UK sugar price is likely to be affected in the medium term (5 years) if there was a sharp rise in oil prices.



Jim Smith and Martine Barons are finding out when the probabilistic judgments of different panels of experts can be coherently drawn together and measures for the lack of coherence when full coherence is not possible[1].

References:
1. EPSRC grant 2013-16 "Coherent inference over a network of probabilistic systems for decision support with applications to food security" **Collier2009:** Collier RA et al. (2009). *Identify reasons why food security may be seen as an issue requiring specific attentions.* Defra project FO0416. **Defra. (2008).** *Ensuring the UK's Food Security in a Changing* World. A Defra Discussion Paper. London: Defra. **FAO1996:** Food and Agriculture Organization, World Food Summit, 1996

Emergency Planning, Response and Recovery for Nuclear Accidents

Professor Simon French and Dr. Nikolaos Argyris

25 years – Chernobyl to Fukushima

In the years since Chernobyl much work has been done to improve processes for emergency planning, management and recovery in the event of a nuclear accident. Since the International Chernobyl Project of 1990-91. Simon French has been involved with much of this:

- The introduction of socio-economic criteria into emergency planning and recovery to supplement criteria relating to radiation exposure and financial cost.
- The design of a decision support system, RODOS, which is now implemented in several European countries, especially in relation to the uncertainty handling, data assimilation and evaluation of options.
- Public communication in relation to local, regional, national and international handling of an emergency.
- The use of stakeholder engagement and public participation in emergency planning and recovery.

But fundamental Issues remain …

- The *linear* ***hypothesis*** which relates the risk of health impact to low levels of (chronic) exposure is misunderstood by many stakeholders and the media.
- Throughout much of an emergency there are very *significant* ***uncertainties***, yet the concept of intervention levels treats these in a very naïve fashion.
- The handling of Chernobyl and Fukushima have **set *public expectations*** about appropriate levels of response that may not be feasible in other contexts.
- ***Equity*** issues are poorly *explored* – or perhaps it would be better to say *articulated* – in emergency planning.

Management of Nuclear Risk Issues: Environmental, Financial and Safety (NREFS)

The NREFS project forms part of the UK-India Civil Nuclear Power Collaboration. The consortium consists of 4 UK academic partners (City, Cambridge, Manchester and Warwick Universities) with direct and collaborative links to the Atomic Energy Commission of India.

The consortium is using a wide variety of methods to explore the important issues in the evaluation of strategies to mitigate the effects of nuclear emergencies. The Warwick team is focussing on the use of scenario-based multi-criteria decision analysis (MCDA).

Objectives

- Develop and apply the J-value framework to post-accident mitigation, particularly for a large nuclear accident.
- Use real options analysis as a tool for judging the cost of instituting an exclusion zone following a severe nuclear accident.
- Use objective methods to assess nuclear power plant siting and liability insurance.
- Use scenario-based multi-criteria decision analysis to investigate differences between recommendations from the objective methods and decisions being taken on the ground.
- Integrate the results from the various methods into recommendations.

References

- C. Niculae, S. French and E. Carter (2004). "Emergency Management: Does it have a sufficiently comprehensive understanding of decision-making, process and context?" *Radiation Protection Dosimetry* **109(1-2): 97-100.**
- S. French and C. Niculae (2005). "Believe in the Model: Mishandle the Emergency." *Journal of Homeland Security and Emergency Management* ***2(1): 1–16.***
- E. Carter and S. French (2006). "Are current processes for nuclear emergency management in Europe adequate?" *Journal of Radiation Protection* ***26: 405-414.***
- K. N. Papamichail and S. French (2013). "25 years of MCDA in nuclear emergency management." *IMA Journal of Mathematics in Management* ***24(4): 481-503***.

RISK PERCEPTION AND DECISION-MAKING IN CANCER

DR JULIA BRETTSCHNEIDER ET AL

1. Cancer treatment

In many common cancers, initial local treatment (e.g. surgery) is followed by adjuvant treatment to prevent recurrence. This includes options such as hormonal treatment and chemotherapy. The latter is costly and has massive side effects, including serious risks. To strike the right balance, a careful assessment of the recurrence risk is desirable.

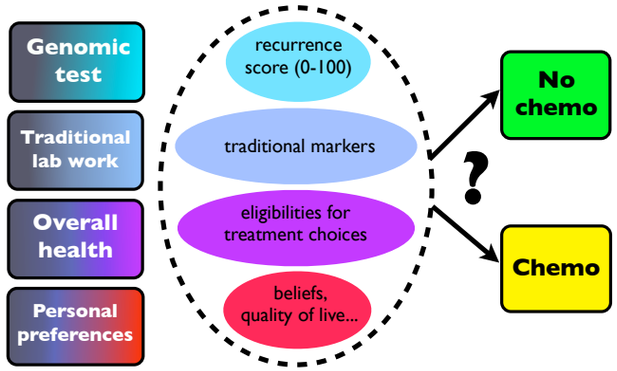
Cancers are heterogeneous diseases and recurrence risks vary between subtypes. In addition to traditional markers such as tumour size and type, novel genomic technologies provide molecular information for further improvement of recurrence risk estimates and hence further individualisation of the treatment decision.

2. Genomic recurrence tests

In the last decade, prognostic tests based on multivariate gene expression measurements have been developed for common cancers such as breast, colon and prostate. Initial validation studies have demonstrated their prognostic value and extensive clinical trials are under way.

An example is Oncotype DX for breast cancer (e.g. [1]). Based on a panel of 22 genes, the test returns a risk score between 0 and 100 which is associated with probabilities of recurrence free survival rates. It is usually communicated in a simplified way: low, intermediate or high risk.

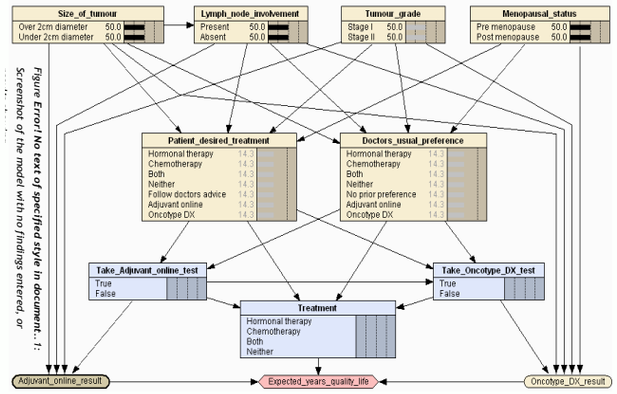
3. Decision task



4. Models

- Complex decision under uncertainty and ambiguity
- Multiple information sources for recurrence
- Shared decision making

Methods include decision trees and Bayesian networks.



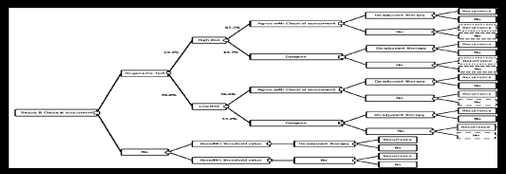
5. Risk perception and biases

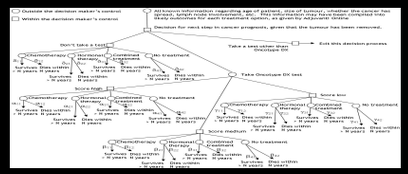
Many empirical investigations have demonstrated that people do not always follow the normative rules of probability. Their perception and processing of risk is subject to biases [2] affecting their decision making. Examples for scenarios:

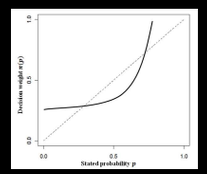
- ***Distortion:*** *Refined differentiation of probabilities close to 0 or 1 compared to those around 0.5, e.g. 3%-0% is perceived as much bigger than 43%-40%.* Potentially, this could lead to overuse of adjuvant treatment while neglecting risks and alternative opportunities.
- ***Ambiguity avoidance:*** *Preference for known risks versus unknown risks even if that leads to a disadvantage.* Potentially, this leads to opting for chemo at all costs.
- ***Confirmation bias:*** *Seeking or altering information to confirm existing beliefs.* Potentially this interferes with the decision to take optional tests and impacts the interpretation and weighing of risk information.

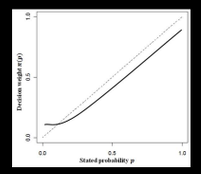
6. Conclusions

- Fallacies may lead to avoiding or overdoing prognostic tests and to misinterpreting results
- Empirical studies on decision-making behaviour (both physicians and patients) needed.
- Room for comprehensive modelling including "irrational" aspects
- Need for improved risk communication and systematic decision support









[1] Paik et al (2004), NEJM 351(27)

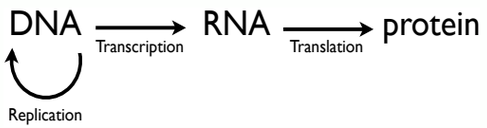
[2] Tversky and Kahneman (1973), Judgement under Uncertainty: Heuristics and Biases, Science New Series, Vol. 185, No. 4157

GENOMIC TECHNOLOGIES: CAN WE TRUST THE DATA?

DR JULIA BRETTSCHNEIDER, DR FRANCOIS COLLIN, DR BEN BOLSTAD AND PROFESSOR TERRY SPEED

1. What is in your genes?

Why are brain cells different from liver cells even though they have the same DNA? Do genes determine the states of cells and organisms? How?



- Gene expression = amount of RNA produced
- Genes interact with the environment
- Genes act in concert

2. Gene expression profiling

Exploratory studies to shed light on complex genetic processes compares expression levels of all genes across stages and conditions, e.g.:

- to find genes involved in cellular processes (e.g. cell cycle, circadian clock)
- to refine diagnosis and prognosis to individualise treatment decisions (e.g. tumour classification, adjuvant treatment)

High-dimensional measurements

Microarray technology quantifies expression levels of all genes simultaneously in one biochemical experiment on a glass slides using 14-20 oligonucleotide probes per gene.

3. Are such finding reproducible?

Microarray data quality assessment (QA) issues:

- Distinguish biological from technical variation
- Massive parallel measurements
- Multi-stage measurement process
- Systematic errors worse than random errors
- No agreement on parametric model
- Repositories swamped with low quality data

4. Quality assessment (QA)

Data preprocessing includes background correction, normalisation, and probe intensity summarisation by iterative reweighted least squares fit of a log linear model. Obtained weights correspond to the reliability of probe j in microarray i. Converted to colours plotted in an array they form **quality landscapes**.

Normalized Unscaled Standard Error:

$$NUSE(\hat{\mu}_i) = \frac{\sqrt{\sum_j w_{ij}^2}}{W_i} \Big/ \text{Median}_i \left\{ \frac{\sqrt{\sum_j w_{ij}^2}}{W_i} \right\}$$

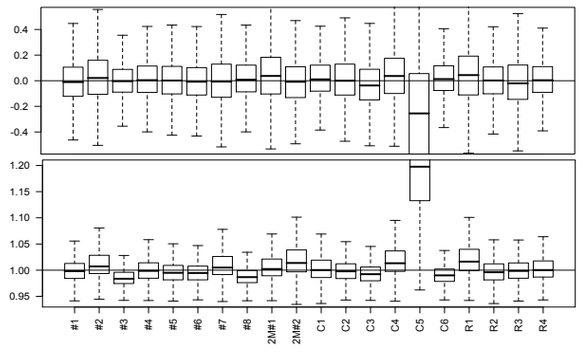
Relative Log Expression (RLE): For each gene, log ratio of its expression to median expression (across all microarrays in experiment).

Typically, expression levels of most genes do not change, same numbers are up and down regulated. **Good quality microarrays**:

- Median(RLE) close to 0 and small IQR(RLE)
- Median(NUSE) close to 1 and small IQR(NUSE)

5. Example for microarray QA

Experiment with 20 microarrays. Quality landscapes show minor local defects in C1, C2, damaged area/overall low quality in C5. Numerical quality scores visible in boxplots of RLE (top) and NUSE (below) also indicate that C5 is an outlier in terms of data quality.

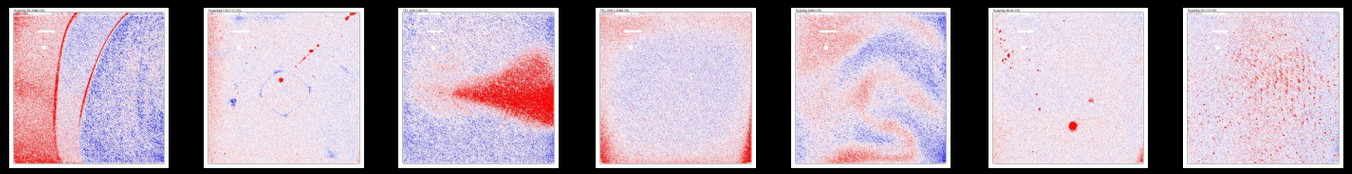


6. Conclusions

The high-dimensional QA toolbox is useful for

- detecting **outliers** and **patterns** related to experimental conditions
- spotting **spatial defects**
- ensuring **reproducibility** of studies
- decreasing errors in molecular medical **diagnosis** and **prognosis**

Some of the methods be applicable to **other kinds of high-dimensional data.**



[1] J Brettschneider et al, Quality assessment for short oligonucleotide microarray data, Technometrics, August 2008, Vol. 50, No. 3 (with Discussion)

[2] www.plmimagegallery.bmbolstad.com

Handwriting recognition using neural networks and rough paths

Dr. Ben Graham
Department of Statistics and Centre for Complexity Science
University of Warwick

Machine learning

The challenge of machine learning sounds rather simple: to devise algorithms for computers that can solve problems that humans find quite easy. For example, reading letters of the alphabet drawn by hand. In practice, it can be rather difficult. For instance, Chinese handwriting has over 7,000 different symbols in widespread use.

Neural networks

Convolutional neural networks are a special kind of computational structure designed for processing two-dimensional images [1]. First developed for reading digits on envelopes and bank cheques, as computers have become more powerful, they have been refined to be able to recognize a vast range of objects, from road signs to pictures of everyday objects.

Online character recognition with rough paths

Online character recognition refers to reading handwriting captured on a tablet computer—the pen stroke is stored as a function embedded in $\mathbb{R}^2$. Convolutional neural networks can accepts a variety of different type of information. The rough path signature, developed by Kuo-Tsai Chen and Terry Lyons, provides a very powerful way of capturing the information contained in a path: it is tensor quantity defined as a collection of iterated integrals:

$$X^k = \int_{0<u_1<\cdots<u_k<1} dX_{u_1} \otimes \cdots \otimes dX_{u_k}.$$

For different k, different kinds of information about the path are expressed. It is natural to ask if these features are useful in the context of machine learning.

Above: Three character written by ten different writers.

Results for Chinese characters

The CASIA OLHWDB1.1 dataset is a benchmark for character recognition algorithms. Previously, the best result for a convolutional network was 5.61%. Using features from the signature reduced the test error to 3.59%. In addition, as the signature features can be stored in a sparse grid, this allows the speed of character recognition to be improved substantially [2].
For the ICDAR 2013 Chinese Handwriting Recognition Competition, researchers were invited to submit a computer program that would then have to read 225,000 handwritten Chinese characters from a secret database. To measure the difficulty, humans were challenged to read part of the test set, with the lowest error rate being 4.81%. Using signature features, a computer program beat this with an error rate of 2.61% [3].

[1] LeCun, Bottou, Bengio, Haffner, 1998

[2] Graham 2013 http://arxiv.org/abs/1308.0371

[3] http://www.nlpr.ia.ac.cn/events/CHRcompetition2013/competition/Home.html

FUNCTIONAL CLUSTERING FOR MORPHOLOGICAL ANALYSIS OF ELECTROCARDIOGRAPH CURVES

DR F. IEVA, PROFESSOR A. M. PAGANONI, DR D PIGOLI AND DR V. VITELLI

1. Introduction

Cardiovascular diseases are one of the main causes of death all over the world. An early diagnosis is essential for good prognosis. Automatic classification for teletransmitted electrocardiogram (ECG) traces is desirable. This is a pilot analysis of ECG traces of patients whose 12-leads pre-hospital ECG has been sent by ambulances to 118 (Italian Emergency Number) Dispatch Centre of Milan.

The statistical analysis consists of preliminary steps followed by the clustering of denoised and aligned ECGs. A diagnostic procedure based on ECG morphology is proposed to classify patients and predict pathologies. We focus here on the identification of Left or Right Bundle Branch Block.

2. The PROMETEO project

Anticipating diagnostic time, reducing infarction complications and optimizing the number of hospital admissions are three main goals of the PROMETEO project (a project for the effective use of electrocardiogram transmitted from emergency unit in the area of Milan, Italy).

Thanks to the partnerships with Azienda Regionale Emergenza Urgenza and Abbott Vascular, ECG machinery with GSM transmission have been installed on all Basic Rescue Units of Milan urban area to obtain a preliminary diagnosis with an automatic procedure.

3. The ECG signal

The most common clinical ECG-system consists of the following 12 leads measuring **voltage differences** between pairs of electrodes:

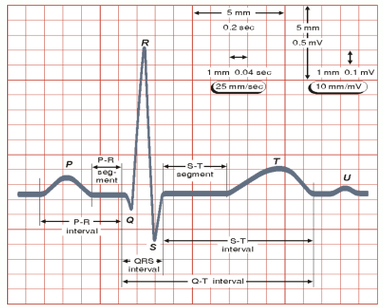
I, II, III

aVR, aVL, aVF

V1, V2, V3,

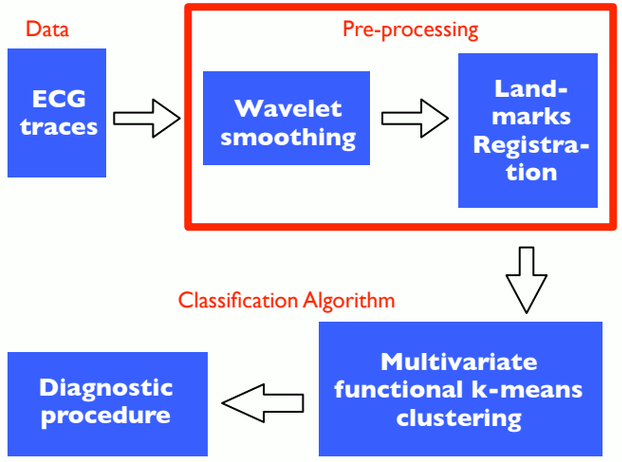
V4, V5, V6

Typical healthy ECG trace of the heartbeat on lead I:

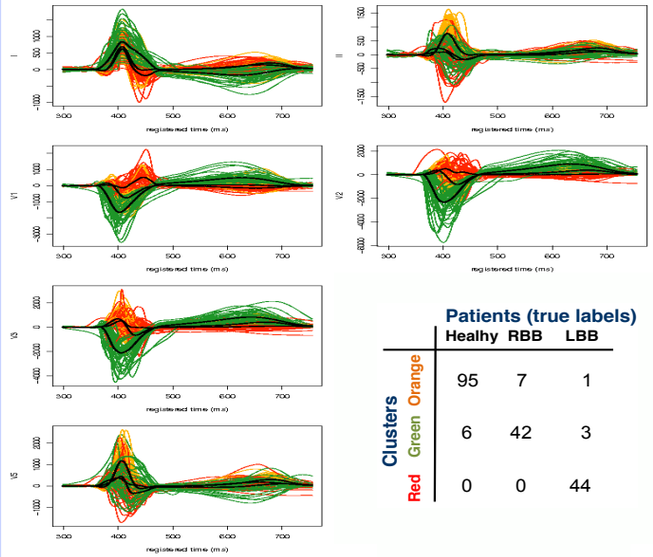


4. Data & Methods

DATABASE for pilot analysis (Median ECG traces records): **101** Healthy traces, **48** Left Bundle Branch Block (LBB) trace and **49** Right Bundle Branch Block (RBB) traces.



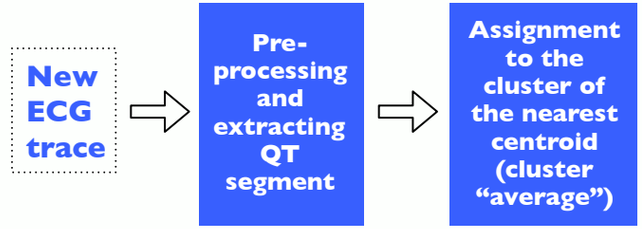
5. Clustering of QT segments



Patients (true labels)

Clusters	Healhy	RBB	LBB
Orange	95	7	1
Green	6	42	3
Red	0	0	44

6. Conclusion: diagnostic procedure



References

Ieva, F., Paganoni, A.M., Pigoli, D. and Vitelli, V. (2013), Multivariate functional clustering for the morphological analysis of electrocardiograph curves, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62.

Ramsay, J.O. and Silverman, B.W. (2005), *Functional Data Analysis* (2nd ed.), Springer, New York.

Professor Wilfrid Kendall and Dr Elke Thönnes

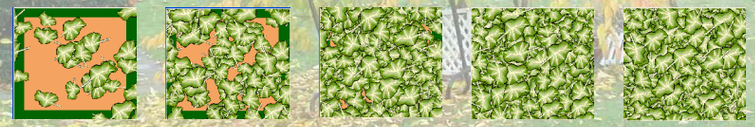
The falling leaves of Fontainebleau

1: *French geologists, working at the École des Mines in the forest of Fontainebleau, formulated the "dead leaves" model as a versatile random pattern to fit to their geological data.*

The model is defined as follows: look at the pattern formed by dead leaves falling randomly on the ground. The "dead leaves" model is the random pattern which is the statistical equilibrium, obtained in the limit for very large time.

2: *The "dead leaves" model is beautiful and simple, and easy to analyze. (Of course in geological applications one replaces the dead leaves by appropriate patterns — crystalline shapes to match the geological sample of interest!)*

Here is a sequence of images of dead leaves, showing how the pattern builds up. A close look shows that it keeps changing as the leaves keep falling ...



3: *Can we ever sample exactly from the statistical equilibrium? Or are we condemned only ever to obtain approximations?*

We can do better with a simple change of perspective. Alter your point of view, from that of a French geologist, to that of a hibernating mammal in its burrow: look up from the ground not down from the sky!

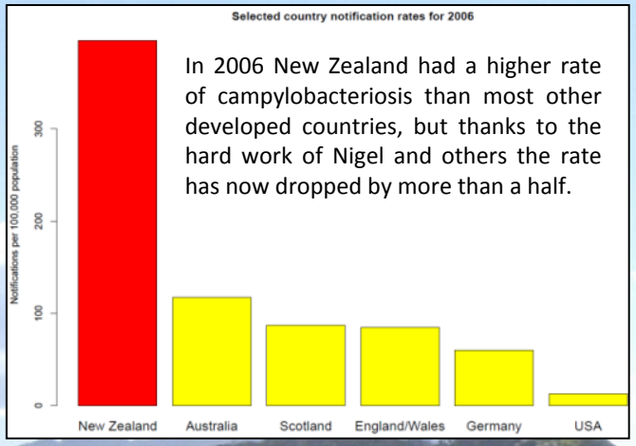
4: *An easy statistical argument (needs first year Warwick probability) shows the limiting patterns have the same statistics, whether observed by small mammal or French geologists. But from the small mammal's point of view the pattern eventually stops changing ... supplying an exact draw from equilibrium!*

You can find out more at
http://www.warwick.ac.uk/go/wsk/perfect_programs

Mapping cases of campylobacteriosis in New Zealand

Dr Simon Spencer – *Assistant Professor in Statistics for Analytical Sciences, University of Warwick*
Joint work with Dr Petra Müllner *and* Professor Nigel French *at Massey University in New Zealand*

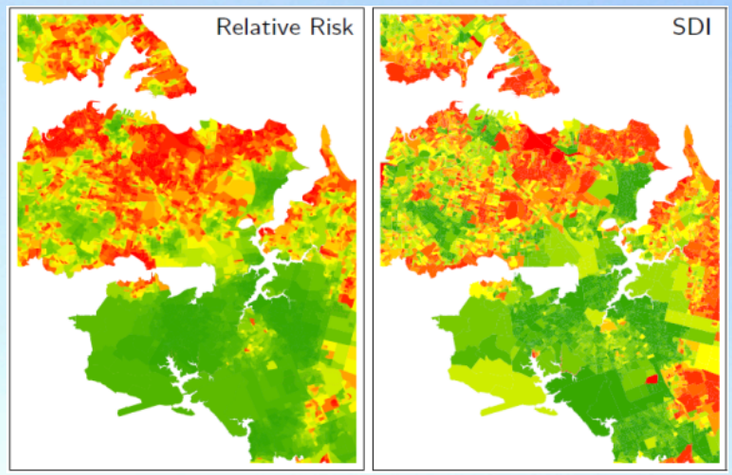
Introduction



In 2006 New Zealand had a higher rate of campylobacteriosis than most other developed countries, but thanks to the hard work of Nigel and others the rate has now dropped by more than a half.

Social deprivation

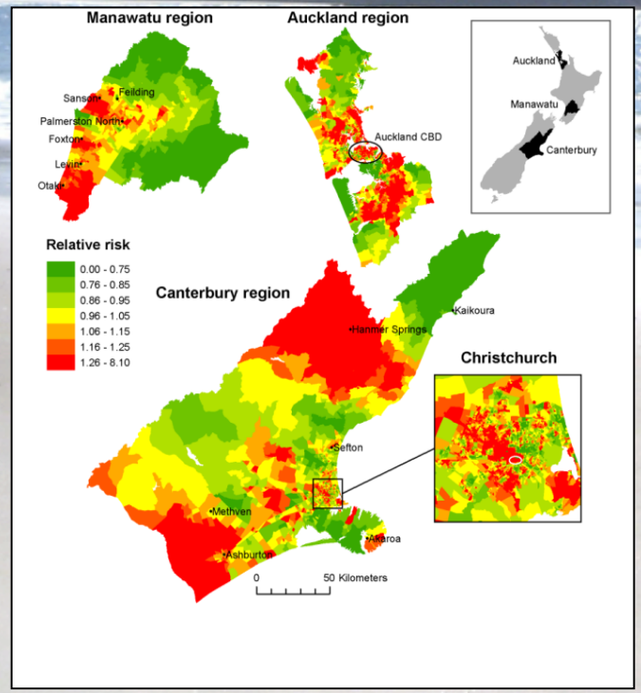
By comparing the risk of infection with the Social Deprivation Index (SDI) we showed that in deprived areas people are less likely to report *Campylobacter* infection. This may be because it costs money for adults to see their doctor in New Zealand.



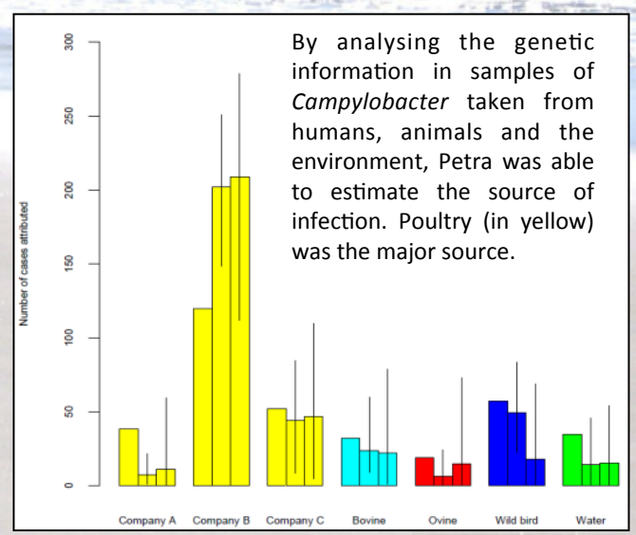
Summary

Campylobacteriosis is a form of food poisoning caused by eating improperly cooked contaminated meat, drinking contaminated water or though direct contact with animals. In this study we used the location of *Campylobacter* cases to produce a risk map which highlights areas at risk of infection

Risk map of *Campylobacter* infection

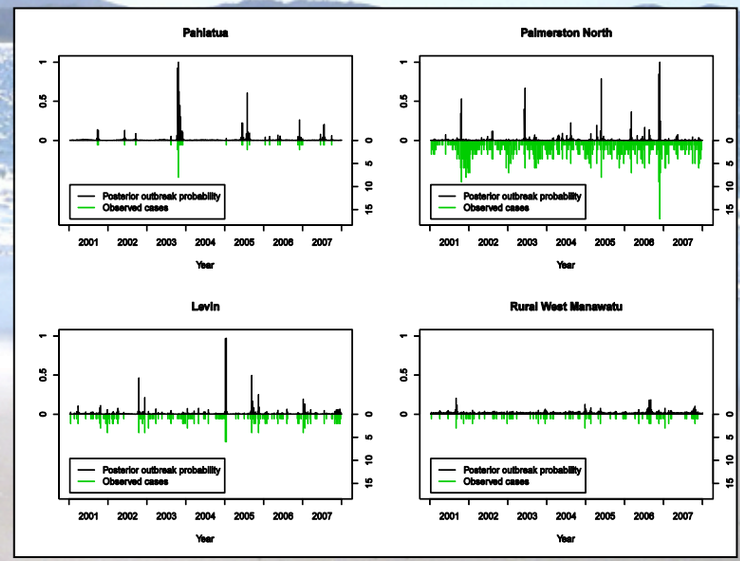


Source of Infection



By analysing the genetic information in samples of *Campylobacter* taken from humans, animals and the environment, Petra was able to estimate the source of infection. Poultry (in yellow) was the major source.

Outbreak detection



We developed sophisticated statistical tools to identify localised outbreaks of campylobacteriosis for further investigation. (Outbreak probability in black, number of cases in green).

Life expectancy for people with cerebral palsy

Professor Jane Hutton

Cerebral palsy

Cerebral palsy is damage to the immature brain which leads to impairment of walking and movement, and can also lead to other impairments. It is the most common cause of physical disability in children.

Factors affecting life expectancy, world-wide

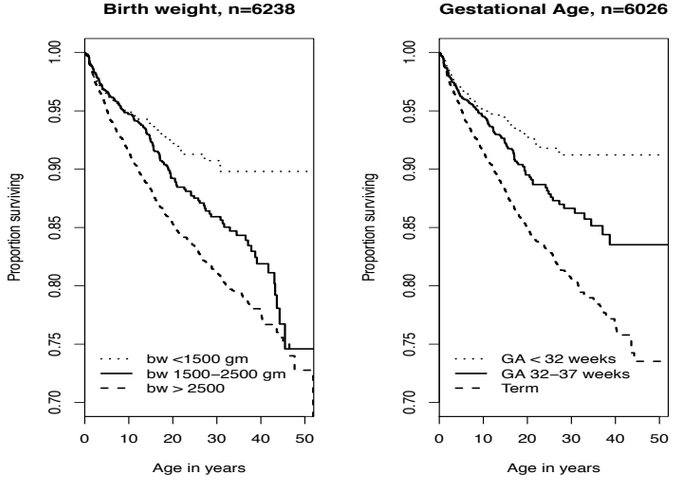
Ambulation, general mobility, lower limb function, manual dexterity, upper limb function, mental ability, visual ability, birth weight, gestational age, growth restriction

UKCP registers used for detailed results

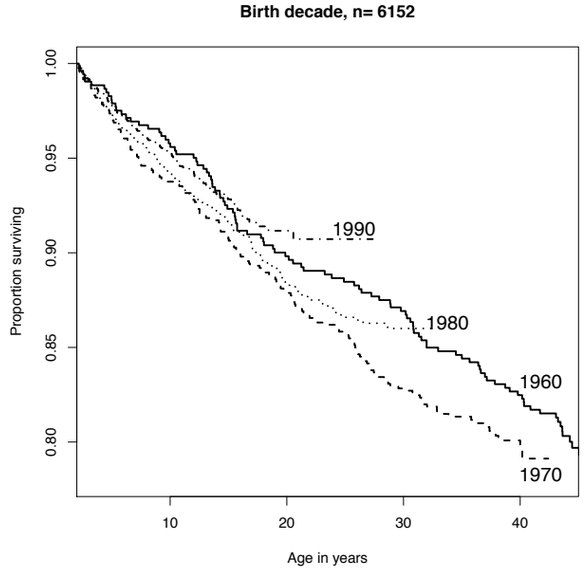
Geographically based cohorts.

1. Merseyside and Cheshire Cerebral Palsy Register, 1966-1991
2. North of England Collaborative Cerebral Palsy Survey, 1960-1999
3. Northern Ireland Cerebral Palsy Register, 1981-2008
4. 4Child (Oxford Register of Early Childhood Impairments) 1984-1997
5. Cerebral Palsy Register for Scotland 1984-1990

Life time: Birth weight, gestation



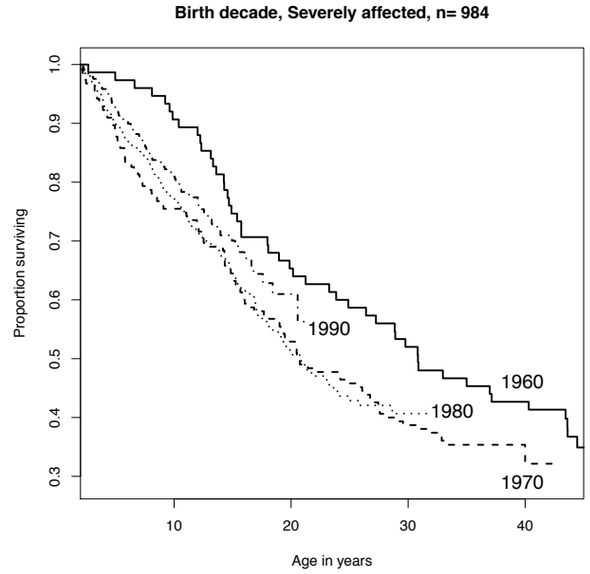
Life expectancy: Secular trend



Life expectancy: individual impairments

	Percentage living from age 2 to			
	10 yrs	20 yrs	30 yrs	40 yrs
Ambulation: Not severe	99	99	97	95
Wheelchair required	86	69	58	52
Upper limb: Not severe	99	98	97	94
Cannot feed or dress	83	62	49	42
Mental ability: IQ $\geq$ 50	99	98	97	95
IQ $<$ 50	85	97	61	55
Vision: Not severe	98	95	93	90
$<$ 6/60 in better eye	80	58	50	44

Secular trend, severely affected



Conclusions

1. People with cerebral palsy who have no severe impairments have similar life expectancy to the general public.
2. About half of people with cerebral palsy who have four severe impairments will reach adulthood, age 18 years.
3. There is no obvious improvement in survival for people with cerebral palsy born in the 1990s over previous decades.

References

1. JL Hutton, Outcome in cerebral palsy: life expectancy, Paed. & Child Health, 2008 ,vol. 18
2. K Hemming, JL Hutton, A Colver and MJ Platt, Regional variation in survival of people with cerebral palsy in the United Kingdom, Pediatrics, vol. 116, 2005

Searching for the Saxon Perch

Professor Wilfrid Kendall

Canterbury - Hauptschiff der Kathedrale von Canterbury
Photo taken by ABrocke on May 25th 2005.
Licenced under CCA-SA 3.0.

1: Anglo-Saxon England undertook some **very big building projects**.

From *Current Archaeology*, 24 May 2007:
"A major Anglo-Saxon cathedral has been revealed – directly under the flagstones of the nave of Canterbury Cathedral. To everyone's surprise, the Anglo-Saxon Cathedral was almost as big as its Norman successor."

Bartholomew, J. G.
1860-1920

2: Question (John Blair of Queen's College, Oxford):
In Anglo-Saxon building projects across England, is there evidence that the builders used the same system of measurement, whether building in Northumberland or Canterbury?

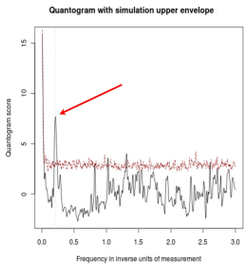
Prime suspect for Mercian projects:
The short Anglo-Saxon Perch (about 4.6m).

3: Finding an answer

Blair has supplied measurements from 66 points on transects of floor plans of 5 Anglo-Saxon buildings.
Is there evidence that these measurements were based on multiples (plus random noise) of a standard "Saxon Perch"?
Use an (approximate) statistical model based on directional statistics (von Mises distribution): large values of

$$\Psi(q) = \sqrt{(2/N)}\, \Sigma_i \cos(2\pi X_i / q)$$

indicate evidence for q being a "quantum" or "module" (cf: D.G.Kendall, 1974; contrast Freeman, 1976).
Which q to select?
Use *DGK quantogram* to assess by simulation (modify to allow for dependence between measurements from same baseline).



4: Plot quantogram $\Psi(q)$ against $1 / q$ (use all possible differences of measurements). Details in Kendall (2013).

Evidently there is an isolated peak, but is it tall enough to take seriously?

Yes: compared to 99% upper envelope using 499 simulations (red curve), the peak is clearly much higher.

There is reasonable evidence for a "Saxon Perch".

Using further statistical theory, we estimate this "Saxon Perch" to be 4.75m ± 0.26m. This is satisfyingly close to suggestions made by Anglo-Saxon historians.

Statistics gets everywhere!

5: What's next?

- Examining another set of building measurements from Wessex and Northern France (expectation: we will find evidence for a *different* module);
- Develop image-analysis methods to get measurements directly from maps;
- Study other phenomena, e.g. patterns in the spatial distribution of Anglo-Saxon placenames (Giacomo Zanella, PhD work in progress).

References:

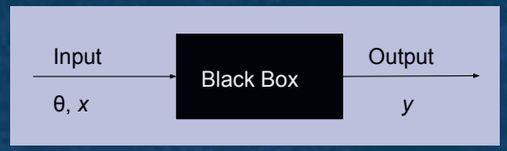
- Blair, J. (2013). Grid-planning in Anglo-Saxon Settlements: the Short Perch and the Four-Perch Module. *Anglo-Saxon Studies in Archaeology and History*, **18**, 18–61.
- Freeman, P. R. (1976). A Bayesian analysis of the megalithic yard. *Journal of the Royal Statistical Society. Series A,* **139**(1), 20–55.
- Kendall, D. G. (1974). Hunting Quanta. *Philosophical Transactions of the Royal Society, London, Series A,* **276**(1257), 231–266.
- Kendall, W. S. (2013). Modules for Anglo-Saxon constructions: Appendix to "Grid-Planning in Anglo-Saxon Settlements: the Short Perch and the four-Perch Module" by John Blair. *Anglo-Saxon Studies in Archaeology and History,* **18**, 55–57.

Simulation and Inference

Dr Anthony Lee

Simulation

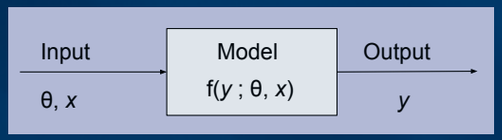
To better understand a complex, stochastic process, we might run a computer program that mimics it. The phenomenon could be mechanical, biogeochemical, financial, economic, astronomical...



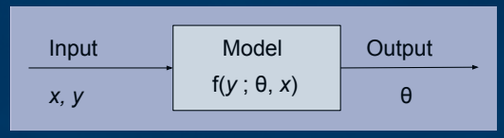
The input consists of a source of randomness, as well as parameters θ and covariates x. We might believe that the simulator is realistic because it can, for some θ, accurately describe the phenomenon of interest. However, we may not know which θ most closely corresponds to reality.

Inference

The problem of determining which values of θ make the simulator accurately reflect reality is a problem of statistical inference. The simulator specifies a statistical model for observed outputs. There are statistical methods for inferring appropriate values of θ given a statistical model.



Statistical inference can be viewed as taking as input x and y and outputting one or many values of θ that are consistent with the data. Usually we have a function f that tells us how likely the data is for given θ and x.



The problem

Most standard methods require at least the computation of f(y; θ, x). This can be a problem when the simulator is very complicated. In many cases, f(y; θ,x) may not even be expressible analytically.

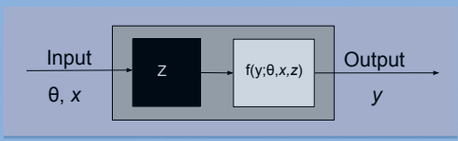
Examples

- Evolutionary branching processes, like those in population genetics.
- Hidden Markov models, where our observations are noisy measurements of a stochastic process at specific times.

Simulation-based inference

Recent research has uncovered a variety of general purpose methods that can be used in situations where parts of or even all of f(y; θ,x) cannot be evaluated. These methods often involve the principled use of simulations from the stochastic process defined by the simulator.

It is often the case that more accurate statements about θ can be made when more of f(y; θ,x) is known.



Some methodology

- Approximate Bayesian computation
- Pseudo-marginal methods
- Particle Markov chain Monte Carlo

Implications

In recent years we have seen a surge in the amount and complexity of data available, in tandem with generative models that are sufficiently rich to explain this data.

By understanding such models, we hope to understand the complex phenomena that shape our world.

These methods are at the forefront of our attempt to meet the challenge of bringing together such models with data to answer modern scientific questions.

References

Marin, J. M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012), Approximate Bayesian computational methods, *Statistics and Computing*, *22*(6)

Andrieu, C., & Roberts, G. O. (2009), The pseudo-marginal approach for efficient Monte Carlo computations, *The Annals of Statistics*, Vol.27, No.2

Phytoplankton bloom in the Black Sea. Image Science and Analysis Laboratory, NASA-Johnson Space Center. The Gateway to Astronaut Photography of Earth. Image ISS035-E-40035.

A Global Positioning System for the Epigenome

Dr Joan Font-Burgada, Dr Oscar Reina, Dr David Rossell and Professor Fernando Azorín

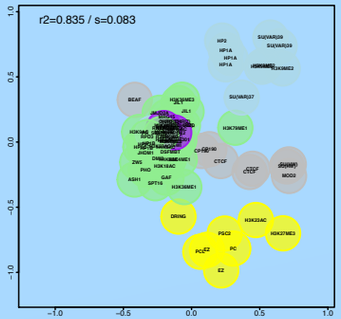
Massive epigenetic databases

Understanding how genomic information is translated into cellular functions constitutes a main challenge in biology. After sequencing genomes of several model organisms, large amounts of data have been gathered regarding different aspects of genome functioning and interaction.

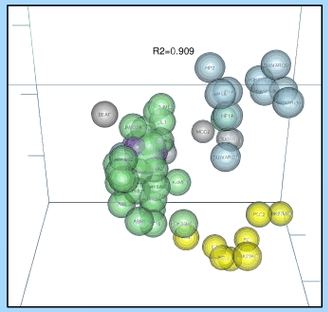
Integrating experimental results and databases on these *epigenetic* factors and genetic elements in a user-friendly manner, amenable to the non-specialist, is a challenge.

Associations between epigenetic factors

2D Map in Drosophila

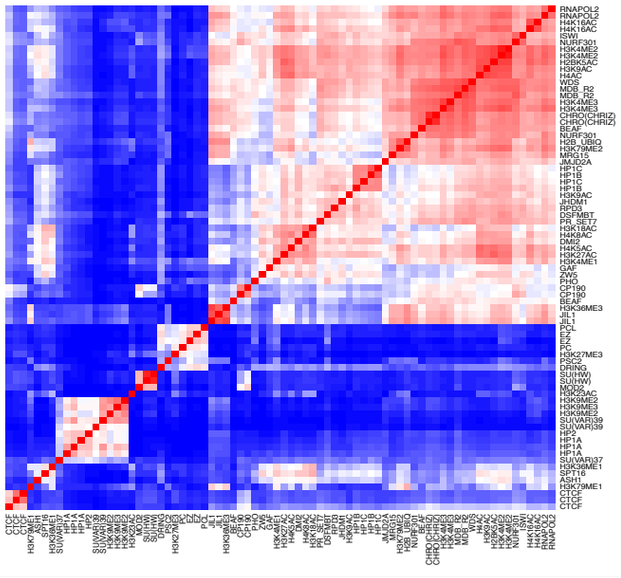


3D Map in Drosophila



Typical analysis (Clustering visualisation)

76 Factors



A statistical challenge

Too much data, hard to get a clear picture. Summarise the info in 2D/3D maps?

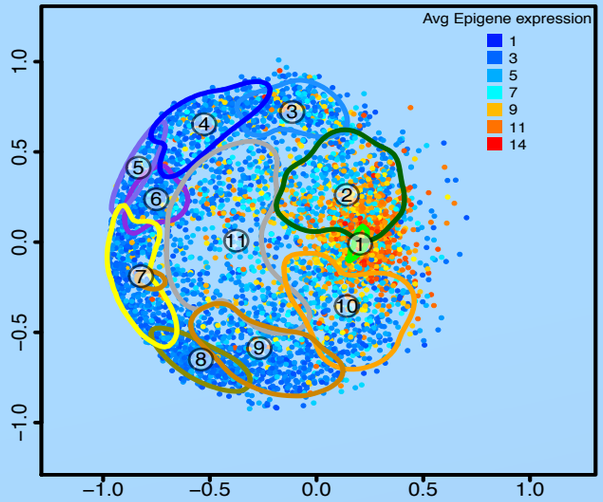
- Combine data from many sources
- Remove systematic biases
- Build maps maximizing information at good computational time
- Give useful biological interpretation

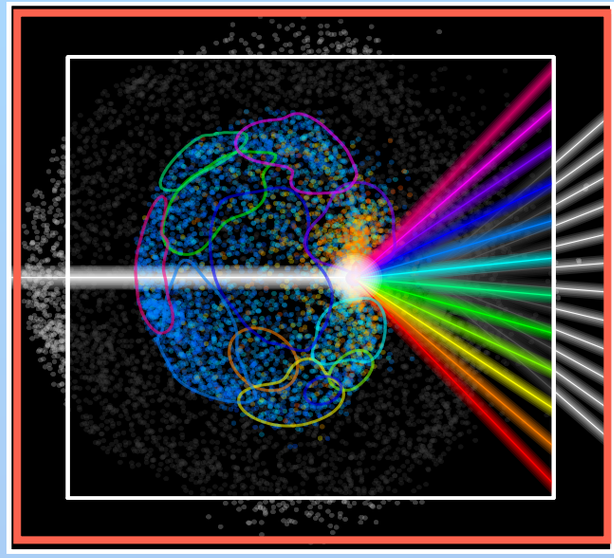
A global positioning system

chroGPS (global chromatin positioning system) integrates and visualizes the associations between epigenetic factors and their relation to functional genetic elements in low-dimensional maps.

Associations between genes

Gene clusters and gene expression levels





J. Font-Burgada, O. Reina, D. Rossell and F. Azorín: chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome, Nucleic Acids Research, Nov. 2013

Department of Statistics, The University of Warwick, Coventry, CV4 7AL, UK Tel: +44 (0)24 7657 4812
Web: go.warwick.ac.uk/statistics | go.warwick.ac.uk/crism | go.warwick.ac.uk/riscu
Email: statistics@warwick.ac.uk | crism@warwick.ac.uk | riscu@warwick.ac.uk