

# 2010

**DAG voor STATISTIEK en BESLISKUNDE**

donderdag 1 april 2010

Vrije Universiteit, Amsterdam

---

hoofdsprekers

**Ernst Wit**

**&**

**Bradley Efron**



# DAG VOOR STATISTIEK EN BESLISKUNDE 2010

## Toegang

De lezingen zijn vrij toegankelijk voor alle belangstellenden, ook van buiten de Vereniging.

## Locatie

De Dag voor Statistiek en Besliskunde is dit jaar te gast bij de Vrije Universiteit te Amsterdam. De VU is gevestigd aan de De Boelelaan 1105, 1081 HV Amsterdam Buitenveldert. De Dag voor Statistiek en Besliskunde zal plaatsvinden in het Hoofdgebouw, auditorium en zalen.

Het ochtendprogramma vindt plaats in het Auditorium, de middaglezingen in de collegezalen 8A-00, 8A-04 en 11A-05, de slotlezing in collegezaal 8A-00. De zalen zijn bewegwijzerd. Voor routebeschrijving en plattegrond zie <[www.vu.nl/organisatie/index.cfm](http://www.vu.nl/organisatie/index.cfm)>.

## Lunch en borrel

De lunch kan in de VU locatie worden genuttigd (voor eigen rekening). De slotborrel wordt aangeboden door de Vereniging.

## Taal

De voertaal voor de algemene gedeelten is Nederlands. De meeste lezingen zijn echter in het Engels.

## Organisatiecomité

Richard Gill, Jacqueline Meulman, Cees Diks, Gerrit Stemerding en de Sectiecoördinatoren

## Informatie

Cees Diks, e-mail <[c.g.h.diks@uva.nl](mailto:c.g.h.diks@uva.nl)>, <<http://home.uva.nl/c.g.h.diks>>

## VVS

Vereniging voor Statistiek en Operationele Research, Postbus 244, 6700 AE Wageningen, telefoon 0317 - 419572, fax 0317 - 421364, <admin@vvs-or.nl>. Raadpleeg onze site <www.vvs-or.nl> over hoe u lid kunt worden van de VVS of een abonnement kunt nemen op een van de VVS-periodieken.

## Algemene Leden Vergadering

De Algemene Leden Vergadering vindt plaats voorafgaand aan het lezingenprogramma. De vergaderstukken kunnen vanaf 2 weken voor de vergadering worden gedownload van de website van de VVS-OR, of worden opgevraagd bij de secretaris van de vereniging.

## Wandelgangen

De Dag voor Statistiek en Besliskunde van de VVS-OR is dé ontmoetingsplaats voor mensen die de statistiek en de operations research een warm hart toedragen danwel in de genoemde vakgebieden werkzaam zijn; leden en niet-leden ontmoeten elkaar in de wandelgangen. Ook kunt u de stands van enkele exposanten bezoeken.

## JMP

Cosinus Computing BV zal tijdens de lunchpauze een demonstratie geven van JMP, een softwareprogramma voor exploratieve gegevensanalyse, met speciale aandacht voor de toepassingsmogelijkheden. De presentatie vindt plaats van 12.45 tot 13.30 uur in collegezaal 8A-04.



*De Dag voor Statistiek en Besliskunde 2010 wordt mede mogelijk gemaakt door de Vrije Universiteit en Cosinus Computing BV, Drunen.*



# PROGRAMMA

## DAG VOOR STATISTIEK EN BESLISKUNDE 2010

donderdag 1 april 2010 in Amsterdam

- 09.15 – 10.10 Algemene Leden Vergadering
- 10.15 – 10.30 Pauze met koffie en thee
- 10.30 – 10.35 Opening Dag voor Statistiek en Besliskunde 2010
- 10.35 – 11.30 Lezing door Ernst Wit (zie pagina 5)
- 11.30 – 11.45 Uitreiking VVS-Scriptieprijs
- 11.45 – 12.30 Uitreiking Van Dantzigprijs en voordracht door de winnaar
- 12.30 – 13.30 Lunch (en JMP-demonstratie in collegezaal 8A-04)
- 13.30 – 15.30 Parallelsessies (zie pagina 8 e.v.)  
*De parallelsessies worden georganiseerd door de Secties van de VVS-OR. Ook is er een speciale sessie, georganiseerd door het Centraal Bureau voor de Statistiek*
- 13.30 – 14.30 Centraal Bureau voor de Statistiek, Economische Sectie, Sectie Mathematische Statistiek
- 14.30 – 15.30 Nederlands Genootschap voor Besliskunde, Sociaal-Wetenschappelijke Sectie en Biometrische Sectie
- 15.30 – 15.45 Pauze met koffie en thee
- 15.45 – 16.45 Lezing door Bradley Efron (zie pagina 6)
- 16.45 – 17.30 Borrel

# Plenaire lezingen

**ERNST WIT**

*University of Groningen*

## War on Error

The callous attack on the Western mathematical World by Islamic mathematical extremists started on 9/11 in the year 808 with the translation of Euclid's elements in Arabic by Al-Ḥajjāj ibn Yūsuf ibn Maṭar (786-833). These Islamic mathematicians then went on to terrorize the population with the introduction of algebra, arithmetic, non-euclidean geometry and trigonometry, integral and differential calculus, mathematical physics and astronomy, and finally cryptography. Europe and its allies were powerless to do anything against these fanatic people willing to give up their lives for mathematics.

Our story starts in 1560 with when a brave Italian lawyer, Gerolamo Cardano, begins to break up the power of the international network of terrorist organizations, known under the umbrella as al-Gebra, from the inside out. Solving cubic and quartic equations, he then went on to publish the first systematic treatment on probability, which opened the eyes of the Western World. Although the subsequent development of statistics has been linked to human right abuses, we will show in this talk that the statistical preemptive war in the 18th, 19th and 20th centuries has *not* violated a more narrow interpretation of international law.

**Ernst Wit** is the Chair of Statistics and Probability at the University of Groningen since 2008. He has degrees in Statistics and Philosophy from the University of Chicago and the Pennsylvania State University, respectively. He has worked at the University of New South Wales, University of Glasgow and Lancaster University, before coming to Groningen. His statistical interests are in the areas of history and philosophy of statistics and probability theory, as well as networks and high-dimensional inference with applications to genomics.



**Bradley Efron**

*Stanford University*

## The future of Indirect evidence

Familiar statistical tests and estimates are obtained by the direct observation of cases of interest: a clinical trial of a new drug, for instance, will compare the drug's effects on a relevant set of patients and controls. Sometimes, though, indirect evidence may be temptingly available, perhaps the results of previous trials on closely related drugs. Very roughly speaking, the difference between direct and indirect statistical evidence marks the boundary between frequentist and Bayesian thinking. Twentieth-century statistical practice focused heavily on direct evidence, on the grounds of superior objectivity. Now, however, new scientific devices such as micro arrays routinely produce enormous data sets involving thousands of related situations, where indirect evidence seems too important to ignore. Empirical Bayes methodology offers an attractive direct/indirect compromise. There is already some evidence of a shift toward a less rigid standard of statistical objectivity that allows better use of indirect evidence. The talk features some examples from current practice, with a little bit of futuristic speculation.

**Bradley Efron** (b. St Paul, Minnesota, 1938) is the Max H. Stein Professor of Statistics and Biostatistics at Stanford University's School of Humanities and Sciences and the Department of Health Research and Policy with the School of Medicine. He completed his undergraduate work in mathematics at the California Institute of Technology, and earned his doctorate in statistics from Stanford in 1964, joining the Stanford faculty that same year. He was Associate Dean for the School of Humanities and Sciences from 1987 to 1990, served a term as Chair of the Faculty Senate as well as three terms as Chair of the Department of Statistics, and continues as Chairman of the Mathematical and Computational Sciences Program. He has served as president of the American Statistical Association and of the Institute of Mathematical Statistics. He is a past editor of the Journal of the American

Statistical Association and is presently the founding editor of the *Annals of Applied Statistics*.

Among the numerous honours that Efron has received are Fellowships of the American Academy of Arts and Sciences, the American Statistical Association, the Institute of Mathematical Statistics, the Royal Statistical Society, the International Statistical Institute and the MacArthur Fellows Program of the John D. and Catherine T. MacArthur Foundation. He is a member of the U.S. National Academy of Sciences, a recipient of the Ford Prize of the Mathematical Association of America and of both the Wilks Medal and the Noether Prize of the American Statistical Association. Efron was awarded the 1998 Parzen Prize for Statistical Innovation by Texas A&M University, and the first-ever Rao Prize for outstanding research in statistics by Pennsylvania State University in 2003. He received the 2005 National Medal of Science “for his contributions to theoretical and applied statistics, especially the bootstrap sampling technique; for his extraordinary geometric insight into nonlinear statistical problems; and for applications in medicine, physics and astronomy.”

# Programma's van de secties *voor abstracts zie pagina 12 en verder*

## Speciale Sessie georganiseerd door het CBS

*Collegezaal 8A-00*

- 13.30-14.00 **Sander Scholtus, Centraal Bureau voor de Statistiek**  
Een bootstrapmethode voor een combinatie van registers en steekproeven
- 14.00-14.30 **Bart Buelens, Virginie Blaess, Centraal Bureau voor de Statistiek**  
Kleinedomeinschatters toegepast op de Veiligheidsmonitor Rijk



## **Economische Sectie**

*Collegezaal 8A-04*

### **theme ECONOMIC PREDICTIONS IN TIMES OF FINANCIAL CRISIS**

13.30-14.00 **Jasper de Jong, Centraal Planbureau**

Ramen in tijden van crises

14.00-14.30 **Rene Segers, Richard Paap & Dick van Dijk, Erasmus  
University Rotterdam**

Evaluating the Consistency and Timeliness of Business  
Cycle Indicators



## Sectie Mathematische Statistiek

*Collegezaal 11A-05*

### theme FORENSIC STATISTICS

13.30-14.00 **Julia Mortera, Università Roma Tre, Rome, Italy**  
Bayesian Networks for Complex DNA mixture analysis

14.00-14.30 **Robert Cowell, City University, London**  
Auto generation of large Bayesian networks for problems in forensic genetics



## Nederlands Genootschap voor Besliskunde

*Collegezaal 8A-00*

14.30-15.30 **Rene Haijema, Wageningen University & Research centre**  
**Nikky Kortbeek, Universiteit van Amsterdam**  
A decision support tool for efficient blood platelet production



## Sociaal-Wetenschappelijke Sectie

*Collegezaal 8A-04*

14.30-15.00 **Jeroen K. Vermunt, Joost R. van Ginkel, L. Andries van der Ark & Klaas Sijtsma, Tilburg University**

Multiple imputation of incomplete categorical data using latent class analysis

15.00-15.30 **Marieke E. Timmerman, University of Groningen**

Bootstrap Confidence Intervals in Component Models



## Biometrische Sectie

*Collegezaal 11A-05*

### theme STATISTICAL ASPECTS IN MODELING THE SPREAD OF INFECTIOUS DISEASES

14.30-15.00 **Jacco Wallinga, RIVM**

Tracking the novel influenza A/H1N1v pandemic

15.00-15.30 **Niel Hens, Universiteit Antwerpen | Universiteit Hasselt**

A study of the early stage of the A/H1N1v pandemic in Europe and the mitigation effect of school closure



# ABSTRACTS

**Bart Buelens, Virginie Blaess**, Centraal Bureau voor de Statistiek  
**Kleinedomeinschatters toegepast op de Veiligheidsmonitor Rijk**

De Veiligheidsmonitor Rijk (VMR) wordt vanaf 2005 door het CBS uitgevoerd en verschaft informatie over thema's van veiligheid, slachtofferschap van misdrijven en politiefunctiearen. De steekproefomvang per politiedistrict is te klein om nauwkeurige directe schattingen voor de politiedistricten te maken. Daarom zijn bij de Veiligheidsmonitor Rijk modelmatige schattingsmethoden voor kleine domeinen toegepast. Hulpvariabelen komen uit politieregisters en andere administratieve bronnen en zijn nu gebruikt om de regionale schattingen te verbeteren. Door middel van modelselectiecriteria zijn optimale modellen geselecteerd uit een grote verzameling van mogelijke modellen. Met deze optimale modellen kunnen voor enkele doelvariabelen nauwkeurigere regionale schattingen gemaakt worden dan met directe schattingen.



**Robert Cowell**, City University, London  
**Auto generation of large Bayesian networks for problems in forensic genetics**

In recent years, Bayesian networks have proved useful in analysing problems in forensic genetics. Typically these networks can be quite large, involving hundreds of nodes, or probability tables with perhaps millions of entries. I shall talk about some software I have written to automate this process to reduce the chances of errors.



**Rene Haijema**, Wageningen University & Research centre

**Nikky Kortbeek**, Universiteit van Amsterdam

### **A decision support tool for efficient blood platelet production**

The production and issuing of blood platelets in blood banks is complicated by short-term perishability, highly uncertain demands and weekend production stops. An optimal production strategy balances shortages and outdating and respect that younger pools are of a higher quality. International studies show that at many other blood banks this figure is more than 15%. Our approach shows in two case studies that outdating can be reduced to less than 1-2%. Optimal production volumes are obtained by solving a (downsized) Markov decision problem (MDP). Nearly-optimal rules can be read by simulating the MDP strategy. Initiated by the Dutch South-East Blood Bank, the MDP-Simulation approach is applied in practice. The software initially developed for research and demonstration is extended to a user-friendly software tool. This tool, called TIMO (Thrombocytes Inventory Management Optimizer), has meanwhile been implemented and is in use for more than one year.



**Niel Hens**, Universiteit Antwerpen, Universiteit Hasselt

### **A study of the early stage of the A/H1N1v pandemic in Europe and the mitigation effect of school closure**

When the A/H1N1v pandemic found its way to Europe in the late spring/early summer of 2009, it was typified by sporadic cases and isolated self-limiting outbreaks linked to importations with one notable exception in the UK where a generalized epidemic took place. No such major epidemic was reported elsewhere in Europe until the autumn, during which time

the UK had its second wave of infection. To understand the country-specific differences in the progression of the pandemic, we postulated and tested a series of epidemiological hypotheses by relating the daily number of secondary infections by means of the reproductive number to weather patterns, importations, susceptibility patterns and school closure. It was apparent that the school closure during the summer had a mitigating effect on the spread of the infection in the UK. To quantify the impact of school closure on the pandemic, we conducted several analyses showing that school closure could slow down the pandemic but should be combined with other mitigation strategies to contain the epidemic and involves a huge economical cost.



**Jasper de Jong**, Centraal Planbureau  
**Ramen in tijden van crises**

Hoe komen wij tot een raming en hoe kan het dat onze ramingen in 2008 voor 2009 niet heel nauwkeurig (eufemisme) waren.



**Julia Mortera**, Università Roma Tre, Rome, Italy  
**Bayesian Networks for Complex DNA mixture analysis**

We show how probabilistic expert systems can be used to analyse forensic identification problems involving DNA mixture traces using peak area information. This information can be exploited to make inferences regarding the genetic profiles of unknown contributors to the mixture, or for evaluating the evidential strength for a hypothesis that DNA from a particular person is present in the mixture. We will also present an extension of the Bayesian network for taking account artifacts such as allelic dropout, stutter bands and silent alleles when interpreting DNA profiles from a sin-

gle and from a pair of mixture traces. We illustrate the use of the network on a published criminal casework example. This is joint work with Robert Cowell and Steffen Lauritzen



**Sander Scholtus**, Centraal Bureau voor de Statistiek

### **Een bootstrapmethode voor een combinatie van registers en steekproeven**

Bij het maken van statistieken wordt steeds vaker uitgegaan van administratieve data, aangevuld met steekproefwaarnemingen. De gebruikelijke nauwkeurigheidsmaten, zoals de vertekening en de variantie van een schatting, zijn in dergelijke situaties niet eenvoudig te bepalen. Resampling-methoden, zoals de bootstrap, bieden een mogelijke oplossing voor dit probleem. Op het CBS is een bootstrapmethode ontwikkeld voor het bepalen van de nauwkeurigheid van schattingen uit het zogenaamde Opleidingsniveaubestand. De informatie in dit bestand is afkomstig uit een aantal partiële opleidingsregisters én meerdere jaargangen van de Enquête Beroepsbevolking. In de presentatie wordt de bootstrapmethode toegelicht en geïllustreerd met voorbeelden uit het Opleidingsniveaubestand.



**Rene Segers, Richard Paap & Dick van Dijk**, Erasmus University  
Rotterdam

### **Evaluating the Consistency and Timeliness of Business Cycle Indicators**

We develop a Markov switching panel data model to simultaneously estimate the individual lead times of a large panel of leading indicator variables. The model relates the turning points of the indicators to the turning points of a reference series, where it is assumed that the cycle of

the reference series coincides with the business cycle. An important feature of the model is that the lead times of the indicators are allowed to be different at business cycle peaks and at troughs.

The modelling framework is applied to The Conference Board's Composite Coincident Index (CCI) and the ten components included in its Composite Leading Index (CLI). Our results suggest that the indicators building permits, stock prices, money supply and consumer expectations are the most timely and consistent indicators among the ten, having average lead times of nine to ten months at peaks and four to five months at troughs with standard deviations of about one month. We apply the model to construct a new, synchronized composite leading index by shifting the ten indicators according to their lead times before aggregation. We show that the synchronized index is more consistent than the CLI and yields better in-sample predictions of the business cycle chronology as determined by the NBER.



**Marieke E. Timmerman**, University of Groningen

### **Bootstrap Confidence Intervals in Component Models**

The bootstrap methodology can be used to estimate confidence intervals (CI's) for the statistics in a component analysis. As different bootstrap strategies may result in clearly different CI estimates, it is important to define a proper strategy for the problem at hand. In this presentation, an overview will be provided of the key aspects to consider when setting up a bootstrap scheme in component analysis. Attention will be paid to the selection of a resampling scheme, the method to estimate the CI's, and how to deal with possible non-uniqueness of the estimated parameters. The key choices will be illustrated by considering bootstrap strategies for three variants of component analysis into detail.

First, for the simple case of Principal Component Analysis (PCA) for independent two-way data, it appears that CI estimates using different meth-

ods in PCA may diverge highly (Timmerman, Kiers & Smilde, 2007). We explain that this results from differences in both quality, and perspective on the degree of rotational freedom of the population parameters. The results of a comparative simulation study indicate that the bias-corrected and accelerated ( $BC_a$ ) method for bootstrap CI's is preferred over other methods considered, including an approach based on asymptotic standard errors.

Second, the Principal Response Curve model analyzes multivariate data resulting from experiments involving repeated sampling in time. The time-dependent treatment effects are represented by Principal Response Curves (PRCs). Confidence bands for PRCs with good coverage can be obtained with  $BC_a$  intervals using a nonparametric bootstrap, except for the case of exactly zero population PRCs for all conditions (Timmerman & Ter Braak, 2008). As will be explained, this is caused by the sign indeterminacy of the PRCs.

Third, Multilevel Component Analysis analyses multilevel multivariate data. The key question to come up with a proper resampling scheme is which level(s) are considered random. The results of a comparative simulation study show that very large sample sizes are necessary when more than a single level is considered to be random (Timmerman, Kiers, Smilde, Ceulemans, & Stouten, 2009).

References:

- Timmerman, M.E., Kiers, H.A.L. & Smilde, A.K. (2007). Estimating Confidence Intervals in Principal Component Analysis: A comparison between the Bootstrap and Asymptotic Results. *British Journal of Mathematical and Statistical Psychology*, 60, 295-314.
- Timmerman, M.E., Kiers, H.A.L., Smilde, A.K., Ceulemans, E. & Stouten, J. (2009). Bootstrap confidence intervals in Multilevel Simultaneous Component Analysis. *British Journal of Mathematical and Statistical Psychology*, 62, 299-318.
- Timmerman, M.E. & Ter Braak, C.J.F. (2008). Bootstrap Confidence Intervals for Principal Response Curves. *Computational Statistics and Data Analysis*, 52, 1837-1849.



**Jeroen K. Vermunt, Joost R. van Ginkel, L. Andries van der Ark & Klaas Sijtsma**, Tilburg University

## **Multiple imputation of incomplete categorical data using latent class analysis**

We propose using latent class analysis as an alternative to loglinear analysis for the multiple imputation of incomplete categorical data. Similar to log-linear models, latent class models can be used to describe complex association structures between the variables used in the imputation model. However, unlike loglinear models, latent class models can be used to build large imputation models containing more than a few categorical variables. To obtain imputations reflecting uncertainty about the unknown model parameters, we use a nonparametric bootstrap procedure as an alternative to the more common full Bayesian approach. The proposed multiple imputation method, which is implemented in Latent GOLD software for latent class analysis, is illustrated with two examples. In a simulated data example, we compare the new method to well-established methods such as maximum likelihood estimation with incomplete data and multiple imputation using a saturated log-linear model. This example shows that the proposed method yields unbiased parameter estimates and standard errors. The second example concerns an application using a typical social sciences data set. It contains 79 variables that are all included in the imputation model. The proposed method is especially useful for such large data sets because standard methods for dealing with missing data in categorical variables break down when the number of variables is so large.



**Jacco Wallinga**, RIVM

## **Tracking the novel influenza A/H1N1v pandemic**

The novel influenza A/H1N1v virus emerged in Mexico in March 2009 and spread quickly around the world. Infection with the virus caused a wide

range of symptoms that is comparable to infection with “seasonal” influenza A viruses. The pandemic of this new virus presented formidable challenges to those epidemiologist and biostatisticians who had to estimate the current state of the pandemic and convert incoming observations into useful information for decision-making. To project the future trajectory of the epidemic, key epidemiological parameters (generation interval, reproduction number, proportion susceptible) had to be inferred from incoming data. To allocate scarce intervention measures, estimates were needed for the age-specific number of potential infectious contacts. Here, we discuss various approaches that have been developed during the pandemic.



