Nieuwsbrief no. 33
oktober 2004

Vereniging voor Ordinatie en Classificatie

Voorzitter: Patrick Groenen, Econometrisch Instituut, Erasmus Universiteit Rotterdam, Postbus 1738, 3000 DR Rotterdam. (groenen@few.eur.nl)
Secretaris: Marieke Timmerman, RU Groningen, Heymans Instituut (DPMG), Grote Kruisstraat 2/1, 9712 TS Groningen (m.e.timmerman@ppsw.rug.nl)
Penningmeester: Paul Arents, Quest International, Sensory Sciences & Consumer Acceptance, Huizerstraatweg 28, 1411 GP Naarden (paul.arents@questintl.com). Postbankrekening 161723 t.n.v. Vereniging voor Ordinatie en Classificatie, Naarden. Bankrekening nummer 777-5952385-56 Bacob Bank t.n.v. VOC, Naarden
Redactie: Mark de Rooij, Faculteit Sociale Wetenschappen, Universiteit Leiden, Postbus 9555, 2300 RB Leiden. rooijm@fsw.leidenuniv.nl
VOC-home page: http://www.voc.ac

Van de voorzitter

Het is zover: de VOC jubileert voor de derde keer. Al vijftien jaar stimuleert de VOC actief classificatie en ordinatie. De volgende bijeenkomst wordt al weer de dertigste. We hebben een keur van vooraanstaande nationale en internationale sprekers gehad op onze bijeenkomsten. Er zijn bijeenkomsten geweest zonder thema en andere gericht op speciale thema's, zoals mengselmodellen, data mining, IRT en robuuste statistiek. De VOC heeft zich ontwikkeld tot een vereniging waarin uiteenlopende disciplines samenkomen om ideeën op het gebied van classificatie en ordinatie uit te dragen en wetenschappelijk werk op dit terrein te stimuleren. Gezien de activiteiten van VOC-leden en hun wetenschappelijke output is de VOC daarin de afgelopen vijftien jaar uitstekend geslaagd.

Het bereiken van de vijftienjarige leeftijd dient goed te worden gevierd. En dat gebeurt ook. Elders in deze nieuwsbrief vind je het uitgebreide programma van de tweedaagse jubileumbijeenkomst te Driebergen met als thema 'alles op zijn plaats'. Als speciale buitenlandse spreker zal daar optreden Andreas Buja (voormalige AT&T Bell-Labs, tegenwoordig Wharton School, Univ of Pennsylvania). Daarnaast zullen naast twee oud-voorzitters van de VOC (Willem Heiser en Henk Kiers) en andere VOC-leden er spreken. Het belooft een prachtige jubileumbijeenkomst te worden. Ik nodig hierbij ieder (potentieel) VOC-lid van harte uit om deel te nemen. Op naar nog eens vijftien jaar VOC en meer!

Patrick Groenen

In dit nummer:

VOC Jubileum-congres: *Alles op zijn plaats*

November 11-12, 2004

De Bergse Bossen, Driebergen

Thursday, November 11

13.50-14.00	Welcome
14.00-14.40	Elffers: *Boeven als buurman: twee typen ruimtelijke invloedsmodellen in de criminologie.*
14.40-15.20	Heisterkamp: *Assessing health impact of sources of airpollution using Bayesian space-time models.*
15.20-16.00	Wehrens: *Clustering image data.*
16.00-16.30	Pauze
16.30-17.30	Kiers: *Visualizing dependency of bootstrap confidence intervals for methods yielding spatial configurations.*
17.30-18.30	Drinks
18.30-20.30	Dinner
20.45-21.45	Heiser: *From Archimedes to Benzécri: How the center of gravity and the moment of inertia entered into statistics.*

Friday, November 12

9.00-10.00	Buja: *Nonlinear dimension reduction.*
10.00-10.30	Pauze
10.30-11.10	Bijmolt: *Country and consumer segmentation: multi-level latent class analysis of financial product ownership.*
11.10-11.50	Van Deun: *Spatial representation of preference and other ranking data.*
11.50-12.30	Van de Velden: *Correspondence analysis of rating data.*
12.30-13.30	Lunch
13.30-14.10	Debba: *Segmentation techniques applied in deriving an optimal sampling scheme.*
14.10-14.50	De Gruyter: *Geostatistical classification of agricultural fields on the basis of nitrate leaching.*
14.50-15.30	Lesaffre: *Correcting for misclassification in caries research.*
15.30-	Afscheid met koffie

Abstracts of Jubilee Meeting. November 11-12, 2004, Driebergen

Henk Elffers & Wim Bernasco (NSCR): ***Boeven als buurman: twee typen ruimtelijke invloedsmodellen in de criminologie***

No abstract (soon available at www.voc.ac)

Prof. dr. Henk Elffers (1948) studeerde wiskundige statistiek en waarschijnlijkheidsrekening aan de Universiteit van Amsterdam en promoveerde op een fiscaal-psychologisch onderwerp aan de Erasmus Universiteit Rotterdam. Hij is thans themacoordinator van het onderzoeksprogramma 'Spreiding en Verplaatsing van Criminaliteit' aan het Nederlands Studiecentrum Criminaliteit en Rechtshandhaving NSCR te Leiden en hoogleraar rechtspsychologie aan de Universiteit Antwerpen. Zijn onderzoeksbelangstelling gaat uit naar de psychologie van de regelnavolging, rationele keuzetheorie, ruimtelijke aspecten van criminaliteit en rechtshandhaving, en naar de rol van statistiek in het strafproces.

Siem Heisterkamp (RIVM): ***Assessing health impact of sources of airpollution using Bayesian space-time models***

We use spatio-temporal models to relate hospital discharge for acute myocardial infarction and bronchitis in the years 1991-1993 to noise and distance from Schiphol airport. The goodness of fit of the different spatial models was assessed using expected predicted deviance. In this paper we will explain why these models are used in epidemiology, how they are used and what we can learn from it.

Simon Heisterkamp is senior-statistician at the National Institute for Public Health and the Environment in Bilthoven (The Netherlands). His interests are in Bayesian statistics and its application in spatial statistics, prediction using time series of infectious disease data and analysis of micro-array data.

Ron Wehrens (KU Nijmegen): ***Clustering image data.***

Automatic segmentation of multivariate images can be achieved by clustering individual pixels. In this presentation, I will focus on model-based clustering, where the data are described by a mixture of multivariate normal distributions. This is a versatile and easily applicable method which gives suggestions on the optimal clustering model, and information on the uncertainty in specific regions of the image. In many applications, these are properties of tremendous importance. Several other characteristics of clustering multivariate images deserve attention. First, because of the sheer size of images, many clustering methods are not

directly applicable. Model-based clustering, for instance, is quite slow, and in some implementations uses a hierarchical clustering for the initialisation. We will show how to deal with this. Second, it may be difficult to detect small clusters; these tend to be overwhelmed by the sheer amount of pixels in larger clusters. For this, an iterative strategy has been developed. And finally, in some cases significant improvements may be obtained by incorporating spatial information, i.e. information on the location of the pixel in the image or information on the classification of neighbouring pixels.

Ron Wehrens (KU Nijmegen). Ron Wehrens is verbonden aan de vakgroep Analytische Chemie van de Radboud Universiteit Nijmegen. Zijn onderzoek beweegt zich in het veld van de chemometrie, dat wil zeggen de toepassing van multi-variate statistiek en globale optimalisatie op chemische systemen. Voorbeelden van toepassingen zijn clustering van moleculen, het voorspellen van chemische of biologische activiteiten, of identificatie en quantificatie van stoffen, meestal op basis van verschillende soorten spectrale informatie

Henk Kiers (RU Groningen) & Patrick Groenen (Erasmus Universiteit): ***Visualizing dependency of bootstrap confidence intervals form methods yielding spatial configurations***

Several techniques exist for summarizing data by means of a graphical configuration of points in a low-dimensional space. The most common examples are multidimensional scaling (MDS) and principal component analysis (PCA). Usually, such analyses are applied to data for a sample drawn from a population, while the researcher often hopes that the configuration (at least roughly) holds for the full population. For instance, a PCA may be carried out on the scores of a sample of subjects on a particular set of variables, while, when PCA is used to display the relations between the variables in a plot, it is hoped that this plot also holds (roughly) for the whole population. To assess how accurate the sample based plot is as a representation for the population, confidence intervals or ellipsoids can be constructed around each plotted point (representing a variable). For this purpose, it has been proposed to use a bootstrap procedure. This procedure gives a full configuration for each bootstrap sample, so we end up with a great many configurations that jointly display variation of the configuration of the variables upon resampling. Usually, the variation is displayed by considering different variables separately. That is, for each individual variable, its location in the low-dimensional space in all bootstrap samples is assessed, and the variation of these locations is represented, for instance, by confidence ellipsoids. However, such a procedure ignores the dependency of variation of *different* variables across bootstrap samples. To display how variation of different variables depends on each other, we propose to visualize bootstrap configurations in a temporally smooth way (movie). Problems encountered then are: How to smooth the transitions from configuration to configuration, and, related to this, how to order the configurations. These problems and some first solutions will be described and demonstrated in the presentation.

Het onderzoek van **Henk Kiers** richt zich op technieken voor multivariate data analyse, zoals twee- en drieweg componenten-analyse. Hij is, op dit terrein, hoogleraar bij de vakgroep Psychologie aan de RuG. Als vroegere voorzitter van de VOC en huidige president van de IFCS is hij altijd nauw betrokken geweest bij "Data Analyse en Classificatie".

Willem J. Heiser (Leiden University): ***From Archimedes to Benzécri: How the center of gravity and the moment of inertia entered into statistics.***

Spatial or geometrical models, such as a Euclidean distance model, a hierarchical tree model, or a factor analysis model, usually serve to account for derived measures of (dis-) similarity, association, or correlation between objects, categories, or variables, respectively. They approximate or highlight aspects of the data. However, there are also basic geometrical models that accommodate all possible data distributions of specific kinds. For any given data set, they express the data geometrically, rather than numerically. Examples of interesting data spaces are the simplex for relative frequencies over a set of categories, and the permutation polytope for rankings of a set of options, but of course, we may simply consider points on a line, too. In this context, the paper looks back into history to trace the origins of two major descriptors of data distributions, the expected value and the variance.

Willem J. Heiser studied psychology in Leiden and completed his dissertation “Unfolding Analysis of Proximity Data” there in 1981. After a post-doc year at Bell Telephone Labs in Murray Hill, New Jersey, he was appointed professor of data theory at Leiden University in 1989. His research focuses on the analysis of multivariate categorical data using multidimensional scaling and classification techniques. He was invited as a visiting professor by the Universidad de Granada, the Universidad de Santiago de Compostela, the University of Exeter, and the Université de Haute Bretagne. He was elected president of the Psychometric Society (2003-2004), is a former editor of Psychometrika (1995-1999), and is the current editor of the Journal of Classification (2002-present).

Andreas Buja & Lisha Chen (The Wharton School, University of Pennsylvania, Philadelphia): ***Nonlinear dimension reduction***

Nonlinear dimension reduction has been a topic of interest on and off for at least half a century. Among the better known approaches are: the continuous-variable versions of the GIFI system, the PRINCALS program, and multiple correspondence analysis; Kruskal-Shepard

multi-dimensional scaling (MDS) when used for dimension reduction; princial curves and surfaces as introduced by Hastie and Stuetzle.
Recently computer scientists in machine learning have proposed novel nonlinear dimension reduction schemes. One proposal, called InfoMap, is just classical Torgerson-Young MDS applied to a novel distance matrix. The other proposal, called 'Locally Linear Embedding' or LLE, is conceptually more novel in that it attempts to recreate local affine relations among neighboring points.
In this talk we will show that Kruskal-Shepard distance scaling makes a strong competitor of InfoMap and LLE when applied to a localized distance matrix augmented by repulsive force between non-local object pairs. We call the resulting method 'Local MDS' or 'LMDS. Localizing MDS by using only distances between near neighbors has been attempted many times and always been shown to be unstable to the point of uselessness. However, the repulsive force proposed here often does a good job at spreading out and stabilizing point configurations. On some of the illustrative datasets used in the InfoMap and LLE articles, LMDS shows superior performance in that it reveals more detail than the two other methods.

Andreas Buja is chaired Professor in the Statistics Department at the Wharton School, Univ of Pennsylvania, Philadelphia. He is interested in machine learning, in particular boosting, as well as multi-dimensional scaling, multivariate analysis, and data visualization. Previous employment, in reverse order: AT&T Labs, AT&T Bell Labs, Bellcore, Salomon Brothers, Univ. of Washington, Stanford University and Stanford Linear Accelerator (visiting faculty), Children's Hospital of Zurich (research associate).

Tammo Bijmolt (RU Groningen), Leo Paas & Jeroen Vermunt (Tilburg University): ***Country and consumer segmentation: multi-level latent class analysis of financial product ownership***

The financial services sector has internationalized over the last few decades. Important differences and similarities in financial behavior can be anticipated between both consumers within a particular country and those living in different countries. For companies in this market, the appropriate choice between strategic options and the resulting international performance may critically depend on the cross-national market structure of the various financial products. Insight into country segments and international consumer segments based on domain-specific behavioral variables will therefore be of key strategic importance. We present a multi-level latent class framework for obtaining simultaneously such country and consumer segments. In an empirical study we apply this methodology and several alternative modeling approaches to data on ownership of eight financial products. Information is available for fifteen European countries, with a sample size of about 1000 consumers per country. We find that both country segments and consumer segments are highly interpretable. Also, consumer segmentation is related to demographic variables such as age and income. Our conclusions feature implications, both academic and managerial, and directions for future research.

Tammo H.A. Bijmolt is Professor of Marketing Research at the Department of Marketing, Tilburg University, The Netherlands; but as of September 2004 at the University of Groningen. His research interest are a fruitful combination of methodology and conceptual marketing issues. Among his major research themes are meta-analysis, multidimensional scaling, and modelling of consumer choice behavior. He published papers in several international journals, such as: Journal of Marketing Research, Journal of Consumer Research, International Journal of Research in Marketing, Journal of Classification, and Multivariate Behavioral Research.

Katrijn Van Deun (KU Leuven): ***Spatial representation of preference and other ranking data***

Ranking data can be represented in Euclidean space by a high- dimensional structure called the permutation polytope. In this presentation, it will be discussed how low-dimensional representations can be derived from the permutation polytope that hold information on the relation between a judge and the different objects he ranked, and also on the relation among the different objects and among the different judges. First, two representations that are based on projecting the polytope on a low-dimensional subspace are presented: these are known as the principal components biplot and the vector model of unfolding. Here, another type of low-dimensional representation will be introduced that is based on a two-step approach. First, distances are measured in the permutation polytope extended with the objects and second, these distances are subjected to an ordinal multidimensional scaling analysis. The different low-dimensional representations are compared using an empirical example.

Katrijn Van Deun behaalde haar diploma in de Psychologische Wetenschappen aan de Katholieke Universiteit Leuven. Momenteel werkt zij er aan het departement Psychologie waar zij een proefschrift voorbereidt over het degeneratie-probleem bij multidimensionele ontvouwing. Ze is er ook betrokken bij het onderwijs van multivariate methoden en variantie-analyse.

Michel van de Velden (Erasmus University):
Correspondence analysis of rating data.

Correspondence analysis is a popular method for data visualization. The typical formulation and derivation of correspondence analysis is based on the analysis of a two-way contingency table. However, mathematically there are no objections against the application of correspondence analysis to any type of nonnegative data matrix. In this presentation, we consider the analysis of rating data using correspondence analysis. Rating data can be used to identify preferences or perceptions of a

group of subjects concerning a set of objects. For example, in marketing research, consumers can indicate preferences concerning products (or product attributes) by assigning ratings. In correspondence analysis the relationships between the products (as indicated in the ratings) are then depicted graphically, allowing a quick and easy interpretation. However, as noted by van de Velden (2000: In "*Innovations in multivariate statistical analysis"*, eds. R.D. Heijmans, D.S.G. Pollock and A. Satorra) and Torres and Greenacre (2002: *International Journal of Research in Marketing*), there exist different approaches to correspondence analysis of rating data. Some theoretical aspects of these differences were studied in van de Velden (2004: *Journal of Classification*). Here we will focus on the practical consequences of the differences.

Michel van de Velden is werkzaam als post-doc bij het Econometrisch Instituut van de Erasmus Universiteit Rotterdam. In 2000, promoveerde hij aan de Universiteit van Amsterdam op het proef-schrift "Topics in Correspondence Analysis". Na zijn promotie werkte hij enige tijd aan de Rijksuniversiteit Groningen. Daarna was hij twee jaar werkzaam als Marie Curie Fellow aan de Universitat Pompeu Fabra, te Barcelona. Michel's onderzoeksinteresses liggen op het gebied van de multivariate statistiek, in het bijzonder, theoretische en praktische aspecten van correspondentie analyse en aanverwante methoden. Zijn werk verscheen onder andere in Linear Algebra and its Applications en Journal of Classification.

Pravesh Debba, Alfred Stein, Freek van der Meer and Arko Lucieer (ITC International Institute for Geo-Information Science and Earth Observation, Enschede): ***Segmentation techniques applied in deriving an optimal sampling scheme.***

An optimized sampling scheme is presented which is useful in selecting samples that represent different categories of interest without any presampling field data. This method uses the iterated conditional modes algorithm (ICM) as an unsupervised segmentation technique to create several homogeneous categories. Within each category, simulated annealing is applied as an optimization technique by minimizing the mean shortest distance between sampling points. The number of sampling points in each category was proportional to the size and variability of the category.
To test the methodology, a generated image with several categories was used. Most categories resulted in an almost equilateral triangular design of the sampling points, thereby enforcing an even spread of the samples within each category. The derived sample points in effect will have image characteristics, for example, gray tone, texture, reflectivity or pattern, depending on the type of segmentation performed.
The combination of previously well formulated techniques such as the ICM for image segmentation and simulated annealing for optimized sampling, results in an elegant and powerful tool in designing optimal sampling schemes using remote sensing images.

Pravesh Debba was born in Durban, KwaZulu-Natal, South Africa in 1969. He received a B.Sc. degree (Mathematics and Statistics) and B.Sc. (Hons)(Statistics) from the University of Durban-Westville, Durban, KwaZulu-Natal, South Africa in 1991 and 1992 respectively. He received an M.Sc. degree in Biostatistics from Limburgs Universitair Centrum, Diepenbeek, Limburg, Belgium in 1998. He started his career as a junior lecturer in the department of statistics at the University of Durban-Westville in 1993 and continued at the University of South Africa, Pretoria, Gauteng, South Africa, from 1994 until 1999. He then joined the School of Mathematical and Statistical Sciences at the University of KwaZulu-Natal, Durban, KwaZulu-Natal, South Africa in 2000, where he is presently employed as a lecturer. He is currently pursuing the Ph.D. degree at the ITC International Institute for Geo-Information Science and Earth Observation, Enschede, The Netherlands. His research interests are on designing optimal sampling schemes using remote sensing images.

Jaap de Gruijter (Alterra, Wageningen University and Research Centre): ***Geostatistical classification of agricultural fields on the basis of nitrate leaching***

Leaching of nitrates from soils used for agriculture to the groundwater is one of the major public health issues in The Netherlands. The government policy is to limit nitrate emission to the groundwater by regulating the nitrogen application by individual farmers. As sandy soils with low groundwater tables are more susceptible to nitrate leaching than other soils, the government decided to set a special, more severe limitation for nitrogen application on dry sandy soils. As this extra limitation has substancial negative financial consequences for the farmers in question, the mapping of these susceptible soils is a soil survey task linked to a hot political dossier. As the extra nitrogen limitation applies to agricultural parcels (management units), the problem is to indicate which parcels are on dry sandy soil. Two questions have to be resolved.
Firstly, how are "parcels on dry sandy soil" to be defined? Research on leaching led to the following definition. A susceptible *soil* has texture-class 'sand', Mean Highest Groundwater table (MHG) deeper than 60 cm below surface, and Mean Lowest Groundwater table (MLG) deeper than 120 cm. This definition applies to a point in the field, but properties vary within parcels. Therefore a susceptible *parcel* was defined as having at least 2/3 of its area on susceptible soil.
Secondly, how can we classify parcels as 'susceptible' or 'not susceptible'? The available data are: (a) observations at sample points on texture, MHG and MLG, (b) spatially exhaustive maps of ancillary variables such as soil type, altitude and drainage class. To this end we developped a geostatistical method, referred to as

'conditional Gaussian co-simulation with uncertain trends'. The steps are as follows.
(1) Develop a multiple linear regression model to predict MHG from the available ancillary variables, and similarly a model for MLG. These models represent the (bivariate) trend.
(2) Fit variogram models to the calculated MHG and MLW regression residuals, representing the spatial autocorrelation structure of the residuals.
(3) Generate a large number (300) of pairs of correlated MHW and MLW fields (values arranged on a fine grid) by Monte Carlo simulation, conditional on the data at the sample points, using the variograms and the trend models. The variation within and between the generated fields represents the uncertainty about regression residuals between sample points, as well as the uncertainty about the regression parameters.
(4) Post-process the simulation results by determining, for each parcel separately, the frequency by which the pairs of fields meet the classification criteria for 'susceptible'. Classify the parcel as 'susceptible' if this frequency is at least 95%, a confidence level chosen by the government in order to balance the farmer's financial risk of false 'susceptible' classification against the ecological risks of false 'not susceptible' classification.
As this method is computationally very demanding, special attention was given to the number of simulations needed in step 3. The method was succesfully tested in a number of test areas and is now being applied on a routine basis.

Jaap de Gruyter werkt bij het Centrum Bodem van Alterra, onderzoeksinstituut voor het landelijk gebied, en onderdeel van Wageningen Universiteit & Research Centre (WUR). Hij houdt zich bezig met statistische methoden voor ruimtelijke inventarisatie en monitoring van natuurlijke hulpbronnen zoals bodem, grondwater, vegetatie, bos en landschap.

Emmanuel Lesaffre (Biostatistical Centre, KU Leuven): ***Correcting for misclassification in caries research***

In large epidemiological (dental) studies often several observers are involved. It is customary to highlight the between-observer agreement with a kappa-statistic. However, the kappa-statistic cannot distinguish between variability and bias in scoring. When a benchmark scorer is available it is preferred to report sensitivity and specificity of the examiners with respect to the benchmark scorer. However, best is to correct for misclassification. We will discuss the issue of correcting for misclassification in models for count data. Frequentist and Bayesian approaches can be used.
The approaches are applied to the first year's data of a dental longitudinal study (Signal Tandmobiel® Study) performed in Flanders from 1996 to 2001. More specifically, caries experience in 7-year old Flemish children showed an East-West geographical trend with more caries in the East. However, the different scoring behavior of the sixteen dental examiners complicated the interpretation of this trend. Controlling for examiner in a classical way largely removed the geographic trend. On the other hand, correcting for misclassification (re-) established the East-West gradient.

Emmanuel Lesaffre works at the Biostatistical Centre from the Katholic University of Leuven. His research focusses on clinical trials, repeated measurements, survival analysis, statistics in dentistry, en meta-analyses.

Personalia

Gedurende het Akademisch jaar 2004-2005 is Willem J. Heiser 'Fellow-in-residence' op het NIAS (Netherlands Institute for Advanced Study in the Humanities and Social Sciences) te Wassenaar. Hij blijft bereikbaar via de gewone adressen.

Cosinus computing
Consultancy in Statistics and Numerics
Adviesbureau voor statistische en numerieke analyses

"Cosinus Computing BV, Adviesbureau voor statistische en numerieke analyses en gevestigd te Waalwijk, ondersteunt software voor statistische analyses zoals Stata, GenStat, EViews en Gauss. Al deze producten kunnen goed worden gebruikt bij analyse van ruimtelijke gegevens en analyse met gebruik van ruimtelijke modellen. Het doet Cosinus Computing deugd het jubileumcongres te sponsoren. Wij wensen de Vereniging nog vele succesvolle jaren toe."

Agenda

17 - 19 November 2004. Aarau, Switzerland. Swiss Statistics Meeting 2004. http://www.statoo.ch/sst04/

5 - 6 December 2004. Ein-Gedi, Israel. The 3rd winter workshop on statistics and computer science - Scientific applictaions of Bayesian Analysis. http://www.cri.haifa.ac.il-/events/2004/csstat/csstat04.htm

6 - 10 December 2005. Atlantic City, USA. 60th Annual Deming Conference on Applied Statistics. httpo://www.demingconference.com

13 - 17 December 2004. San Francisco, USA. American Geophysical Union: Data Mining in the Geosciences. http://www.agu.org/meetings/fm04/

4 - 6 January 2005. International conference on recent advances in statistics. http://home.iitk.ac.in/~kundu/conference.html

6 - 8 January 2005. Gainesville, Florida, USA. University of Florida Seventh Annual Winter Workshop: Longitudinal Data Analysis. http://www.stat.ufl.edu/symposium/2005/longitudinal/index.html

12 -14 January 2005. Bormio, Italy. Second International IMS/ISBA Joint Meeting. http://alien.eco.uninsubria.it/IMS-ISBA-05/

20 - 23 March 2005. Austin, Texas, USA. 2005 ENAR Spring Meeting. http://www.enar.org

29 March - 1 April 2005. Auckland, New Zealand. 14th International Workshop on Matrices and Statistics. http://iwms2005.massey.ac.nz/

05 - 12 April 2005. Sydney, Australia. 55th Session of the International Statistical Institute (ISI)

15 April. Leiden. Voorjaarsbijeenkomst VOC. Thema: Hoog-dimensionale data

21 - 23 April 2005. Newport Beach, California, USA. Fifth SIAM International Conference on Data Mining. http//www.siam.org/meetings/sdm05 .

06 - 08 June 2005. Parma, Italy. 5th Biennial Meeting of the Classification and Data Analysis Group of the Italian Statistical Society. http://www.cladag2005.unipr.it/

12 - 16 June 2005. San Antonio, Texas, USA. 2005 international symposium on forecasting. http://www.isf2005.org

12 - 15 June 2005. Saskatoon, Canada. 2005 Annual Meeting of Statistical Society of Canada. http://www.ssc.ca

15 - 18 June 2005. Izmir, Turkey. Ordered Statistical Data: Approximations, Bounds and Characterizations

21 - 24 June 2005. Fairbanks, USA. Joint annual meeting of the Western North America Region (WNAR) of the International Biometric Society and the Institute of Mathematical Statistics (IMS) http://www.uaf.edu/wnar/

26 June - 1 July, 2005. Santa Barbara, USA. 30th Conference on Stochastic Processes and their Applications. http://www.pstat.ucsb.edu/projects/spa05/

25 - 29 July, 2006. Ljubljana, Slovenie. IFCS conference. http://www.classification-society.org/

2 - 4 August 2006, Ås, Norway. 8th Sensometrics Meeting: Imagine the senses. http://www.sensometric.org/-pages/meet.htm

Boekbespreking

Applied functional data analysis: Methods and case studies. J.O. Ramsay & B.W. Silverman (2002). Springer-Verlag, New-York.

Met dit boek willen de auteurs illustreren dat en hoe functionele data-analyse toepasbaar is binnen inhoudelijk erg uiteenlopende domeinen. Na een zeer summiere inleiding over functionele data -1 bladzijde- wordt in Hoofdstuk 1 een overzicht gegeven van de gevalsstudies die aan bod komen in de overige 11 hoofdstukken. Een greep uit de onderwerpen: longitudinale criminaliteits-gegevens, de vorm van beenderen afkomstig van een paleontologische studie, psychologische testgegevens, reactietijden bij ADHD patiënten, dynamische gegevens over de positie van de wijsvinger tijdens het jongleren, enz. De lezer zal snel door hebben dat in functionele data-analyse de basiseenheden geen individuele observaties zijn (bijvoorbeeld: de criminele activiteit van persoon x op tijdstip 1, op tijdstip 2, enz.), maar functies (de evolutie van de criminele activiteit van persoon x doorheen de tijd). De kernvraag is: hoe kunnen deze functies goed geschat worden op basis van een beperkt aantal observaties? Waarbij ‘goed’ staat voor: vloeiend maar toch nauw aansluitend bij de gegevens. ‘Maar toch’, want tussen beide criteria is er een trade-off (cfr. de bias-variance trade-off).
In Hoofdstuk 2 tot en met 12 wordt telkens een voorbeeld uitgewerkt. De volgorde is niet zomaar lukraak gekozen. De auteurs volgen een learning-by-example benadering. In elk hoofdstuk wordt een techniek uit de functionele data-analyse voorgesteld, met de meer geavanceerde technieken in de latere hoofdstukken. Elk hoofdstuk bevat een korte samenvatting, meestal gevolgd door een bibliografie en een technische sectie. In deze technische secties gaan Ramsay en Silverman soms (meestal) kort door de bocht. Wie de gebruikte technieken wil begrijpen in al hun wiskundige finesses, zal op andere bronnen beroep moeten doen. ‘Functional Data Analysis’ van dezelfde auteurs, bijvoorbeeld.
Een techniek die in verschillende voorbeelden terugkomt is functionele principale componenten analyse. Net als klassieke PCA heeft functionele PCA tot doel om de belangrijkste bronnen van variabiliteit in de gegevens op te sporen. Het analogon van de discrete vector van

componentgewichten uit klassieke PCA is bij functionele PCA een continue principale component gewichtsfunctie. Andere technieken die aan bod komen zijn o.a. het gebruik van cross-validatie om de smoothing parameter te bepalen; het construeren van phase-plane plots (een plot van de tweede t.o.v. eerste afgeleiden doorheen de tijd) om de dynamiek van een proces in kaart te brengen; en time warping: het mappen van individuele curven op een standaardcurve door ze horizontaal uit te rekken en te verschuiven.
De meeste datasets en code (Matlab/S-plus) zijn te vinden op het web. Zelf had ik meer succes op http://www.stats.ox.ac.uk/~silverma/fdacasebook/ en http://www.psych.mcgill.ca/faculty/ramsay/fda.html dan op de springer site die vermeld staat in het boek. Het boek is aan te bevelen als een eerste kennismaking met het domein van functionele data-analyse. Wie al meer beslagen is, kan zich tegoed doen aan het (her)analyseren van de interessante datasets.

Frank Rijmen

Procrustes Problems. J.C. Gower & G.B. Dijksterhuis (2004). Oxford University Press.

Met het verschijnen van *Procrustes Problems* is er voor het eerst een omvangrijk overzichtswerk beschikbaar gekomen voor het breed toepasbare vakgebied van de Procrustes-methoden. Er zijn wel eerder pogingen ondernomen dit gebied in kaart te brengen maar er is niets van een vergelijkbaar formaat. In 14 hoofdstukken en 6 appendices wordt de lezer de weg gewezen in het oerwoud van methoden die met elkaar gemeen hebben dat matrices via bepaalde rotaties, transformaties, en/of schaling maximaal met elkaar in overeenstemming worden gebracht, bij voorkeur via minimalisatie van residu-kwadratensommen. Dit boek is doordrenkt van least-squares methoden.

Het boek begint met een terugblik op de voorgeschiedenis, en eindigt met een overzicht van toepassingen en toekomstperspectieven. In de tussenliggende hoofdstukken wordt uitgebreid ingegaan op de techniek van orthogonale Procrustesrotatie, Procrustes-projectieproblemen (transformatie via niet-vierkante orthonormale matrices), en Oblieke Procrustesproblemen. Er is een apart hoofdstuk over gegeneraliseerde Procrustesanalyse, waarbij het erom gaat meer dan twee matrices met elkaar in overeenstemming te brengen. Voorts zijn er hoofdstukken over weging, schaling en ontbrekende waarden, over kanskapitalisatie bij Procrustesmethoden, over variantie-componenten bij Procrustes-oplossingen, en over biplots. De appendices geven deels inleidende definities, die even goed in de tekst zelf hadden kunnen worden opgenomen, deels technische uitwerkingen die de auteurs op een plaats wilden parkeren waar ze niemand in de weg staan. Zo bevat Appendix E resultaten van grensverleggend onderzoek die belangwekkend genoeg lijken om aan een psychometrisch tijdschrift aan te bieden. Het boek beschikt over nuttige auteurs- en onderwerpindex, en is rijk gelardeerd met ingekaderde boxen waarin algorithmen voor de behandelde methoden worden geschetst.

Over het algemeen is dit een helder geschreven boek, maar het heeft wel een sterke hang naar technische complicaties. Daarmee lijkt het primair bedoeld voor specialisten die zelf methoden bestuderen en ontwikkelen. De puur op toepassingen gerichte lezers zullen grote delen van dit boek te technisch vinden. Zij zullen echter toch baat hebben bij de vele algoritmen die ze gemakkelijk zelf kunnen programmeren. Ook als naslagwerk heeft dit boek waarde voor beide typen gebruikers. Als studieboek is het aan de pittige kant. Ik verwacht dat de hulp van een toegewijde docent geen overbodige luxe zal zijn.

Hoewel het verschijnen van dit boek van harte mag worden toegejuicht en grote delen ervan zeer leerzaam genoemd kunnen worden, zijn er ook wel tekortkomingen. Mijn belangrijkste bezwaar is dat er nogal wat bronnen in de literatuur te noemen zijn waarvan de auteurs onvoldoende profijt hebben getrokken. Dat blijkt o.a. als er informele of nodeloos omslachtige bewijzen voor optimaliteit van bepaalde oplossingen worden gegeven, terwijl betere alternatieven beschikbaar zijn. Zo beschouw ik het bewijs van pag. 41 (PCA) als onvolledig (ook na correctie van typfouten). Een zeer eenvoudige manier om het bewijs te compleet te maken is het afleiden van een bovengrens voor de te maximaliseren functie. Dat kan o.a. via een ongelijkheid van Poincaré (zie bijv. Magnus & Neudecker, 1999) of via een generalisatie van de stelling van Kristof (zie bijv. Ten Berge, 1993).
Die laatste stelling zou ook op diverse andere plaatsen van nut hebben kunnen zijn. Zo wordt op pagina's 87 t/m 90 het "dubbele Procrustesprobleem" opgelost, met behulp van de speciaal voor dit probleem geschreven Appendix F. Het probleem is het maximum te vinden van $\mathrm{tr}(\mathbf{X}_2'\mathbf{Q}_2'\mathbf{X}_1\mathbf{Q}_1)$, waarbij $\mathbf{X}_1$ en $\mathbf{X}_2$ van orde $p\times q$ zijn met $p\geq q$, over orthonormale matrices $\mathbf{Q}_1$ $(q\times q)$ en $\mathbf{Q}_2$ $(p\times p)$. De oplossing kan in drie regels worden afgeleid. Regel 1 definieert de singuliere waarden-ontbindingen $\mathbf{X}_i=\mathbf{P}_i\mathbf{\Gamma}_i\mathbf{R}_i'$, $i=1,2$, met $\mathbf{P}_i$ een $p\times q$ matrix. Regel 2 luidt dat $\mathrm{tr}(\mathbf{\Gamma}_1\mathbf{\Gamma}_2)$ een bovengrens is voor de functie, dankzij de gegeneraliseerde stelling van Kristof. Als tenslotte op regel 3 geconstateerd wordt dat de bovengrens wordt bereikt als $\mathbf{Q}_1$ en $\mathbf{Q}_2$ de op pag. 90 vermelde vorm hebben, is het bewijs compleet.
Deze en andere bezwaren betreffende de bewijsvoering zullen alleen voor de technisch geïnteresseerde lezer een rol spelen. Maar er zit ook hier en daar een omissie in het weergeven van bronnen waarvan de op toepassing gerichte lezer profijt had kunnen trekken. Soms worden artikelen op zo'n manier besproken dat het lijkt alsof de auteurs de latere ontwikkelingen niet meer hebben bijgehouden. Geen mens zal het ze kwalijk nemen. De vakliteratuur op het gebied van de Procrustesproblemen is zeer omvangrijk. Gower en Dijksterhuis hebben ons met dit boek in belangrijke mate geholpen er een weg in te vinden.

Magnus, J.R., & Neudecker, H. (1999). *Matrix differential calculus with applications in Statistics and Econometrics*. Wiley: New York.

Ten Berge, J.M.F. (1993). *Least squares optimization in multivariate analysis*. Leiden: DSWO.

Jos ten Berge

Publicaties en rapporten

Bancsi LFJMM, Broekmans FJM, Looman CWN, Habbema JDF & te Velde ER (2004). Predicting poor ovarian response in IVF: use of repeat basal FSH measurement. *Journal of Reproductive Medicine*, 49, 187-194.

Bechger TM, & Maris G (2004). Structural Equation Modeling of Multiple Facet Data: Extending Models for Multitrait-Multimethod Data. *Psicológica*, 25, 235-252. (http://www.uv.es/psicologica/)

Bechger TM, Verstralen HHFM, Verhelst ND & Maris G (2004). Equivalent linear logistic test models: A rejoinder. *Psychometrika*, 69, 219-220.

Currie I, Durban M & Eilers PHC (2004) Efficient smoothing of d-dimensional arrays. *Proceedings of the 19th International Workshop on Statistical Modelling*, Florence.

De Boer EJ, Den Tonkelaar I, Burger CW, Looman CWN, van Leeuwen FE & te Velde ER (2004). The number of retrieved oocytes does not decrease during consecutive gonadotrophin-stimulated in vitro fertilisation cycles. *Human Reproduction*, 19, 899-904.

De Bruin JP, Dorland M, Spek ER, Posthuma G, van Haaften M, Looman CWN & te Velde ER (2004). Age-related changes in the ultrastructure of the resting follicle pool in human ovaries. *Biology of Reproduction*, 70, xxx-xxx.

De Gucht V, Fischler B & Heiser WJ (2004). Personality and affect as determinants of medically unexplained symptoms in primary care – A follow-up study. *Journal of Psychosomatic Research*, 56, 279-285.

De Gucht V, Fischler B & Heiser WJ (2004). Neuroticism, alexithymia, negative affect, and positive affect as determinants of medically unexplained symptoms. *Personality and Individual Differences*, 36, 1655-1667.

De Jong AE, Morreau H, Van Puijenbroek M, Eilers PHC, Wijnen J, Nagengast FM, Griffioen G, Cats A, Menko FH, Kleibeuker JH & Vasen HFA (2004). The role of mismatch repair gene defects in the development of adenomas in patients with HNPCC. *Gastroenterology*, 126, 42-48.

Eilers PHC (2004). Parametric time warping. *Analytic chemistry*, 76, 404-411.

Eilers PHC (2004). The shifted warped normal model for mortality. *Proceedings of the 19th International Workshop on Statistical Modelling*, Florence.

Eilers PHC & Dijksterhuis GB (2004). A parametric model for time-intensity curves. *Food Quality and Preference*, 15, 239-245.

Eilers PHC & Goeman JJ (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics*, 20, 623-U82.

Eilers PHC, van Soolingen D, Lan NTN, Warren RM & Borgdorff MW (2004). Transposition rates of Mycobacterium tuberculosis IS6110 restriction fragment length polymorphism patterns. *Journal of Clinical microbiology*, 42, 2461-2464.

Engels GI, Duijsens IJ, Haringsma R & Van Putten CM (2003). Personality disorders in the elderly compared to four younger age groups: A cross-sectional study of community residents and mental health patients. *Journal of Personality Disorders,* 17, 447-459.

Fredriks AM, van Buuren S, Jeurissen SE, Dekker FW, Verloove-Vanhorick SP & Wit JM (2004). Height, weight, body mass index and pubertal development reference values for children of Moroccan origin in the Netherlands. *Acta Paediatrica*, 93, 817-24.

Giordani P & Kiers HAL (2004). Principal Component Analysis of symmetric fuzzy data. *Computational Statistics and Data Analysis*, 45, 519-548.

Gower JC & Dijksterhuis GB (2004). *Procrustes problems*. Oxford Statistical Science Series. Oxford: Oxford University Press.

Groenen PJF & Koning A (2004). Generalized Bi-additive Modelling for Categorical Data. 8 pp. *Econometric Institute Report* EI 2004-05.

Groenen PJF & Koning A (2004). A new model for visualizing interactions in analysis of variance. 17 pp. *Econometric Institute Report* EI 2004-06.

Groenen PJF & van de Velden M (2004). Multidimensional Scaling. 14 pp. *Econometric Institute Report* EI 2004-15.

Groenen PJF & Van der Velde M (2004). Inverse correspondence analysis. *Linear Algebra and its Applications*, 388, 221-238. (Also appeared as Econometric Institute Report EI 2002-31.)

Groenen PJF (2004). Visualisatie met dynamische meerdimensionele schaling. In: A.E. Bronner (Ed.), *Ontwikkelingen in het marktonderzoek, Jaarboek 2004*, MarktOnderzoekAssociatie, pp. 183-196.

Heiser WJ & Busing FMTA (2004) Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage (2004), pp. 25-48.

Hirasing RA, Fredriks AM, van Buuren S, Verloove-Vanhorick SP & Wit JM (2003). Toegenomen prevalentie van overgewicht en obesitas bij Nederlandse kinderen en signalering daarvan aan de hand van internationale normen

en nieuwe referentiediagrammen. *Tijdschrift voor Jeugdgezondheidszorg*, 34, 82-87.

Hukshorn CJ, Lindeman JHN, Toet KH, Saris WHM, Eilers PHC, Westerterp-Plantenga MS & Kooistra T (2004). Leptin and the proinflammatory state associated with human obesity. *Journal of clinical endocrinology and metabolism*, 89, 1773-1778.

Hulst J, Joosten K, Zimmerman L, Hop W, van Buuren S, Büller H, Tibboel D & van Goudoever J (2004). Malnutrition in critically ill children: from admission to 6 months after discharge. *Clinical Nutrition*, 23,223-32.

Kaczmarek K, Walczak B, de Jong S & Vandeginste BGM (2004). Preprocessing of 2-D gel electrophoresis images. *Proteomics,* 4, 2377-2389.

Kaczmarek K, Walczak B, de Jong S & Vandeginste BGM (2003). Comparison of image-transformation methods used in matching 2D gel electrophoresis images. *Acta Chromatographica*, 13, 7-21.

Kaczmarek K, Walczak B, de Jong S & Vandeginste BGM (2003). Matching 2D Gel Electrophoresis Images. *Journal of Chemical Information and Computer Science*, 43, 978-986.

Kiers HAL (2004). Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics*, 18, 22- 36.

Klinkert ER, Broekmans FJM, Looman CWN & te Velde ER (2004). A poor response in the first in vitro fertilization cycle is not necessarily related to a poor prognosis in subsequent cycles. *Fertility and Sterility*, 81, 1247-53.

Meerding WJ, Looman CWN, Essink-Bot ML, Toet H, Mulder S & van Beeck EF (2004). Distribution and determinants of health and work status in a comprehensive population of injury patients. *Journal of Trauma*, 56, 150-161.

Meulman JJ, Van der Kooij AJ & Heiser WJ (2004). Principal Components Analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage, pp. 49-70.

Nusselder WJ & Looman CWN (2004). Decomposition of differences or changes in health expectancies. *Demography*, 41, 315-334.

Pietersz R & Groenen PJF (2004). Rank reduction of correlation matrices by majorization. 21 pp. *Econometric Institute Report* EI 2004-11.

Seegers G, Van Putten CM & Vermeer HJ (2004). Effects of causal attributions following mathematics tasks on student cognitions about a subsequent task. *Journal of Experimental Education*, 72, 307-328.

Smilde A, Westerhuis JA & de Jong S (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17, 323-337.

Ten Berge JMF (2004). Partial uniqueness in Candecomp/Parafac. *Journal of Chemometrics*, 18, 12-16.

Ten Berge JMF (2004). Simplicity and typical rank of three-way arrays, with applications to TUCKER-3 analysis with simple cores. *Journal of Chemometrics*, 18, 17-21

Ten Berge JMF, Sidiropoulos ND & Rocci R (2004). Typical rank and INDSCAL dimensionality for symmetric three-way arrays of order I×2×2 or I×3×3. *Linear Algebra and Applications*, 388, 363-377.

Van Beilen M, Kiers HAL, Bouma A, van Zomeren EH, Withaar FK, Arends J & van den Bosch RJ (2003). Cognitive deficits and social functioning in schizophrenia: A clinical perspective. *The Clinical Neuropsychologist*, 17, 507-514.

van Buuren S & Tennant A (2004). Response Conversion for the Health Monitoring Program. *TNO-rapport* 2004.145. TNO Preventie en Gezondheid, Leiden. ISBN 90-5986-082-9 (82 pp.)

van Buuren S, Bonnemaijer-Kerckhoffs DJA, Grote FK, Wit JM & Verkerk PH (2004). Many referrals under Dutch short stature guidelines. *Archives of Diseases in Childhood*, 89, 351-253.

van Buuren S, van Dommelen P, Zandwijken GRJ, Grote FK, Wit JM & Verkerk PH (2004). Towards Evidence Based Referral Criteria for Growth Monitoring. *Archives of Diseases in Childhood*, 336-341.

van Dommelen P, de Gunst MCM, van der Vaart AW, van Buuren S & Boomsma DI (2004). Groeidiagrammen voor lengte, gewicht en 'body mass index' voor tweelingen in de peutertijd. *Nederlands Tijdschrift voor de Geneeskunde*, 148, 1345-1350.

Van Herk H, Poortinga YH, & Verhallen ThMM (2004). Response Styles In Rating Scales: Evidence of Method Bias in Data from 6 EU Countries. *Journal of Cross-Cultural Psychology*, 35, 346-360.

van Wezel B, Bruil J, van Buuren S & Fredriks AM (2004). Signaleren van overgewicht en (secundaire) preventie bij jeugdigen. *Nederlands Tijdschrift van Diëtisten*, 59, 42-46.

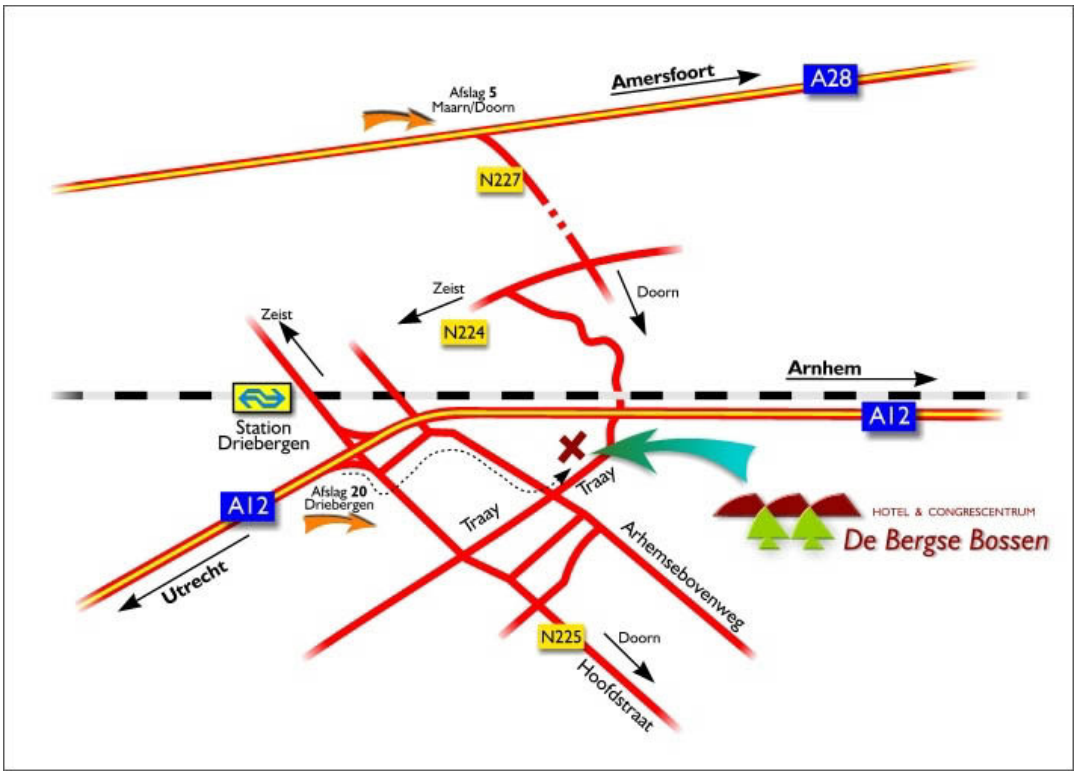
Vestering N & van Buuren S (2004). Two new Bayesian item selection methods and their application in Computerized Classification Testing. *TNO-rapport* 2004.176. TNO Preventie en Gezondheid, Leiden. ISBN 90-5986-093-4 (41 pp.)

Vos AM, Meima A, Verver S, Looman CWN, Bos V, Borgdorff MW & Habbema JDF (2004).High incidence of pulmonary tuberculosis persists a decade after immigration, the Netherlands. *Emerging Infectious Diseases*, 10, 736-739.

Wu W, Guo Q, Massart DL, Boucon C & de Jong S (2003). Structure preserving feature selection in PARAFAC using a genetic algorithm and Procrustes analysis. *Chemometrics and Intelligent Laboratory Systems*, 65, 83-95.

Xu QS, de Jong S, Lewi PJ & Massart DL (2004). Partial Least Squares Regression With Curds and Whey. *Chemometrics and Intelligent Laboratory Systems*, 71, 21-31.

Routebeschrijving (http://www.debergsebossen.nl):



Uit de richting Amsterdam, Den Haag, Utrecht:

Bij knooppunt Oudenrijn richting Arnhem (A12). U neemt afrit 20 Driebergen/Zeist en gaat onderaan de afrit rechts richting Driebergen. Bij de eerste verkeerslichten (na 50 meter) links af (Loolaan) richting Austerlitz. Op de rotonde de tweede weg rechts. Na twee kilometer op de kruising met verkeerslichten links richting Austerlitz waarna u op de Traaij na ongeveer een kilometer Hotel en Congrescentrum "De Bergse Bossen"aan uw linkerhand aantreft.

Uit de richting Arnhem:

Richting Utrecht (A12). U neemt afrit 20 Driebergen/Zeist en gaat onderaan de afrit links richting Driebergen. Bij de tweede verkeerslichten (na 50 meter) links af (Loolaan) richting Austerlitz. Op de rotonde de tweede weg rechts. Na twee kilometer op de kruising met verkeerslichten links richting Austerlitz waarna u op de Traaij na ongeveer een kilometer Hotel en Congrescentrum "De Bergse Bossen"aan uw linkerhand aantreft.

Uit de richting Amersfoort/Zwolle:

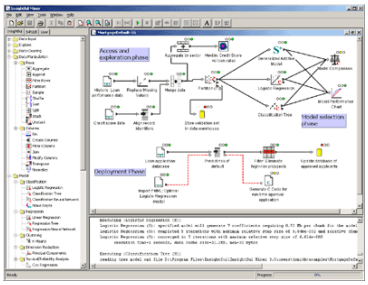
U neemt op de (A28) afrit 5 Maarn/Amersfoort-Zuid richting Maarn/Doorn. U volgt deze weg (N227) tot aan de kruising met de (N224) en gaat daar rechts richting Austerlitz/Zeist. U volgt de borden Driebergen (eerste afslag links). Na ongeveer drie kilometer bij het binnenrijden van de bebouwde kom treft u “De Bergse Bossen” aan uw rechterhand aan.

Per openbaar vervoer:

Uitstappen op station Driebergen/Zeist. Taxi’s zullen door de VOC georganiseerd worden.

INSIGHTFUL MINER

A HIGHLY SCALABLE DATA ANALYSIS WORKBENCH



Insightful Miner is a highly scalable data mining and analysis workbench that gives new analysts and skilled modelers the ability to deploy predictive intelligence throughout the enterprise. Insightful Miner increases support for large data environments with new versions for Windows and Solaris servers and adds many new features that allow data analysts and data miners to easily build and deploy analytic applications that boost product performance and improve the efficiency of critical business processes.

Insightful Miner Deploys Predictive Applications to Non-technical Users and Scales Many Legacy S-PLUS Models to Large Data Sets

"We use Insightful Miner to extend sophisticated portfolio analysis capabilities to our less technical decision makers… Insightful Miner lets us leverage our custom methods developed in S-PLUS and its visual workmap interface makes it easy to share these techniques with those who must perform the analysis, but don't necessarily need to learn how to program."
William Alexander, Sr. VP of Consumer Deposit Risk Management, Bank of America

Insightful Miner Turns the Complex and Rigorous Into Easy and Sharable

"Insightful Miner fits our knowledge discovery needs very well. It easily handles very large data sets and its intuitive, self-documenting interface makes the entire analysis process more efficient. Approaching analysis from multiple perspectives using Insightful's suite of statistical tools makes our inferences more reliable."
Dr. Drew Griffin Levy, Director, Exploratory Data Analysis, Pfizer

Key Benefits

- **Build powerful predictive models without programming**
  Wire components together to form self-documenting workflows, then quickly build and evaluate multiple models to choose the best.
- **Full support for every step of the data mining and analysis process**
  Explore and model large data sets with tightly integrated components that handle all your data preparation, modeling, assessment and deployment needs.
- **Data mining software that adapts to your changing needs**
  The comprehensive feature set handles most data mining problems right out of the box. But unlike rigid, pre-made analytic applications, Insightful Miner can be customized to meet specific business needs by tuning built-in algorithms or adding custom components tailored to the task at hand.
- **Deploy results easily into your organization**
  A wide variety of deployment options fit the needs of every organization. From publishing Web-ready graphics and reports to decision makers, to scoring large data sets via batch processes or integrating C code generated from predictive models into other analytic applications.

CANdiensten
020-5608400
info@can.nl
www.can.nl