VOC Newsletter 60
October 2019

Vereniging voor Ordinatie en Classificatie / Dutch-Flemish Classification Society

Chairman: Mark de Rooij, Universiteit Leiden, Faculteit der Sociale Wetenschappen, Instituut Psychologie, Sectie Methodologie & Statistiek, Postbus 9555, 2300 RB Leiden, Nederland (rooijm@fsw.leidenuniv.nl)

Secretary: Katrijn Van Deun, Universiteit van Tilburg, Faculteit Sociale Wetenschappen, Departement Methoden en Technieken van Onderzoek, Postbus 90153, 5000 LE Tilburg, Nederland (K.VanDeun@uvt.nl)

Treasurer: Tom Wilderjans, Universiteit Leiden, Faculteit der Sociale Wetenschappen, Instituut Psychologie, Methodologie & Statistiek, Postbus 9555, 2300 RB Leiden, Nederland (t.f.wilderjans@fsw.leidenuniv.nl)

Editor: Pieter Schoonees, Erasmus Universiteit Rotterdam, Rotterdam School of Management, Department of Marketing Management, Postbus 1738, 3000 DR Rotterdam, Nederland (schoonees@rsm.nl)

VOC website: https://www.voc.ac
Postbankrekening (IBAN) NL86 INGB 0000 161723 t.n.v. Vereniging voor Ordinatie en Classificatie.

VOC Jubilee Meeting

21 – 22 November 2019 Wageningen

Fletcher Hotel-Restaurant De Wageningsche Berg
Generaal Foulkesweg 96, 6703 DS Wageningen

Thursday, 21 November 2019

11:00 - 11:30	Arrival and welcome
11:30 - 12:15	Session 1
12:15 - 13:30	Lunch
13:30 - 15:00	Session 2
15:45 - 18:00	Session 3
18:00 - late	Drinks and dinner

Friday, 22 November 2019

07:00 - 09:00	Breakfast
09:00 - 10:30	Session 4
11:15 - 12:45	Session 5
12:45 - 14:00	Lunch
14:00 - 15:30	Session 6
15:30 - 15:40	Closing

In this issue:

Registration details for the VOC Jubilee Meeting

Those who would still like to join the VOC Jubilee Meeting are kindly requested to register through our website **https://voc.ac/meeting/**. Details are provided through the website. A limited number of spots are still available.

From the President

In 1989 the VOC was founded and therefore we celebrate its 30st anniversary this year. Now, 30 years after the foundation we can wonder whether the VOC is alive and kicking or not. The question can be answered by either yes or no. Yes, because the field of supervised and unsupervised classification and of data visualization is booming business. We see data science bachelor and master programs appear at many universities, and data science – yes, that's what the VOC is all about! Simultaneously we must admit the answer is no, because 1) the number of members of the VOC decreases; and 2) the number of participants in our spring and fall meetings declines. The board is thinking about how to turn the tide, but we don't have the definite answer (yet).

For the anniversary meeting the board, together with Paul Eilers (thanks Paul for all your work!), invited speakers from different disciplines and from different countries but with one characteristic in common: they are all distinguished speakers and researchers. We hope that this program brings together many VOC members and possibly some new members. More information about the program can be found in this newsletter. Please have a look at it, become inspired, and enroll for the meeting at our website. If possible, bring a colleague.

Hope to see you in Wageningen,

Mark de Rooij

Jubilee Meeting Sponsors

Thank you to Biometris at Wageningen University & Research and Smit Consult for sponsoring the VOC Jubilee Meeting!

BIOMETRIS
Quantitative Methods brought to Life

Smit Consult
Consultancy in statistiek

Publications

Biondi, B., Van der Lans, I.A., Mazzocchi, M., Fischer, A.R.H., Van Trijp, H.C.M., & Camanzi, L. (2019). Modelling consumer choice through the random regret minimization model: An application in the food domain. *Food Quality and Preference, 73*, 97-109. https://doi.org/10.1016/j.foodqual.2018.12.008.

ter Braak, C.J.F. (2019). New robust weighted averaging- and model-based methods for assessing trait-environment relationships. *Methods in Ecology and Evolution, 00*, 1-10. https://doi.org/10.1111/2041-210X.13278.

Durieux, J., & Wilderjans, T.F. (2019). Partitioning subjects based on high-dimensional fMRI data: comparison of several clustering methods and studying the influence of ICA data reduction in big data. *Behaviormetrika, 46*, 271-311. https://doi.org/10.1007/s41237-019-00086-4.

van der Hoef, H., & Warrens, M.J. (2019). Understanding information theoretic measures for comparing clusterings. *Behaviormetrika, 46*, 353–370.

Hoekstra, R., Vugteveen, J., Warrens, M.J., & Kruyen, P.M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology*, *22*, 351-364.

Mellgren, K., Nierop, A.F.M., & Abrahamsson, J. (2019). Use of multivariate immune reconstitution patterns to describe immune reconstitution after allogeneic stem cell transplantation in children, Biology of Blood and Marrow Transplantation, *25(10)*, 2045-2053. doi.org/10.1016/j.bbmt.2019.06.018.

de Raadt, A., Warrens, M.J., Bosker, R.J., & Kiers, H.A.L. (2019). Kappa coefficients for missing data. *Educational and Psychological Measurement*, *79*, 558-576.

Sok, J., Van der Lans, I.A., Hogeveen, H., Elbers, A.R.W., & Oude Lansink, A. (2018). Farmers' preferences for Bluetongue vaccination scheme attributes: An integrated choice and latent variable approach. *Journal of Agricultural Economics,* *69*(2), 537-560. https://doi.org/10.1111/1477-9552.12249.

Vendrig, N.J., Hemerik, L., Pinter, I.J. & Ter Braak, C.J.F. (2019). Relating ultrasonic vocalizations from a pair of rats to individual behavior: A composite link model approach. *Statistica Neerlandica*, *73*, 139-156. http://dx.doi.org/10.1111/stan.12144.

Warrens, M.J. (2019). Similarity measures for 2x2 tables. *Journal of Intelligent and Fuzzy Systems*, *36*, 3005-3018.

Warrens, M.J., & De Raadt, A. (2019). Properties of Bangdiwala's B. *Advances in Data Analysis and Classification*, *13*, 481–493.

Yuan, B., Heiser, W., & de Rooij, M. (2019). The δ-Machine: Classification-Based on Distances Towards Prototypes. *Journal of Classification*, 1-29. https://doi.org/10.1007/s00357-019-09338-0.

Programme: VOC Jubilee Meeting
Wageningen, 21 – 22 November 2019

Fletcher Hotel-Restaurant De Wageningsche Berg
Generaal Foulkesweg 96, 6703 DS Wageningen

THURSDAY, 21 NOVEMBER 2019

11:00	Arrival	
11:25	Welcome	
11:30	**Ron Wehrens**	Adapting the self-organising map to the big data era
12:15	Lunch	
13:30	**Eva Ceulemans**	Kernel change point detection on the running statistics: a flexible, comprehensive and user-friendly tool
14:15	**Peter Bühlmann**	Stabilizing high-dimensional regression
15:00	Coffee break	Room check-in from 15:00
15:45	**Laura Bringmann**	The (un)necessity of complex statistics in an N=1 world
16:30	**Niël le Roux**	Implementing a biplot based multivariate data-driven industrial performance index
17:15	**Tim Offermans**	Improving the statistical assessment of industrial process quality: towards industry 4.0 and beyond
18:00	Drinks	
19:00	Dinner	

FRIDAY, 22 NOVEMBER 2019

07:00	Breakfast	Runs until 10:00
09:00	**Aurélie Lemmens**	Managing churn to maximize profits
09:45	**Anders Skrondal**	The role of conditional likelihoods in mixed-effects modeling
10:30	Coffee break	Check out from room until 11:00
11:15	**Sophie Swinkels**	Statistical challenges in nutrition studies: dealing with multi-target, non-adherence and ethical constraints
12:00	**Michael Greenacre**	Logratio analysis versus correspondence analysis: and the winner is ...?
12:45	Lunch	
14:00	**Tom Snijders**	Networks in social contexts: the settings model
14:45	**Anne-Laure Boulesteix**	Computational statistics and open science
15:30	Closing	End at 15:40

* Each talk lasts 45 minutes, including 10 minutes for discussion.

Route description to the VOC Jubilee Meeting location

Hotel name and address:

Fletcher Hotel-Restaurant De Wageningsche Berg
Generaal Foulkesweg 96, 6703 DS Wageningen

The website of the conference hotel is available at:
https://www.hoteldewageningscheberg.nl/en/

Travel information is provided at:
https://www.hoteldewageningscheberg.nl/en/contact-route

Route planner for public transportation
Plan your public transport route using https://9292.nl/en.

Public transportation
The conference hotel is located 10km from the nearest train station (Station Ede-Wageningen). Bus 352 (operated by Breng) do stop a 7-minute walk away from the hotel.

You can reach the hotel either by bus from Station Ede-Wageningen, or by bus from Station Arnhem Centraal. Travelling through Station Ede-Wageningen is likely faster (depending on your origin) but requires changing buses again to Bus 352 at Wageningen Bus Station. From Station Arnhem Centraal it is possible to take Bus 352 directly to the hotel.

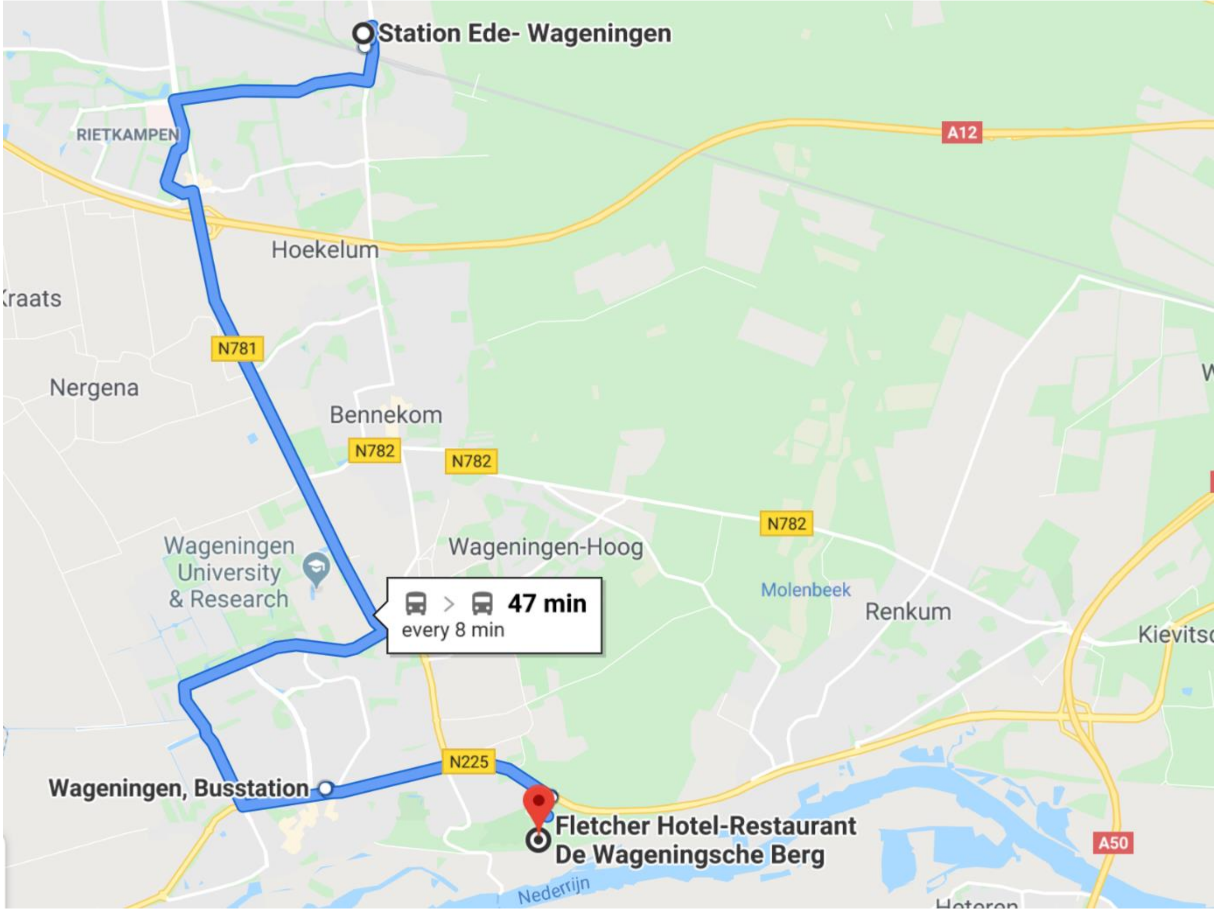
Public transportation in the Netherlands is best utilized using an OV chipkaart (chip card). These can be used on all public transport modes, including trains and buses. Anonymous OV chip cards can be obtained for €7.50 from vending machines or service desks at train or major bus stations. The anonymous OV chip cards must be preloaded with additional credit before use. Disposable chip cards, which are valid for a limited period of time and are already preloaded with a travel product, can be bought (usually with cash only) on buses.

More information is available at:
https://www.ov-chipkaart.nl/everything-about-travelling/different-types-of-passenger/tourists.htm

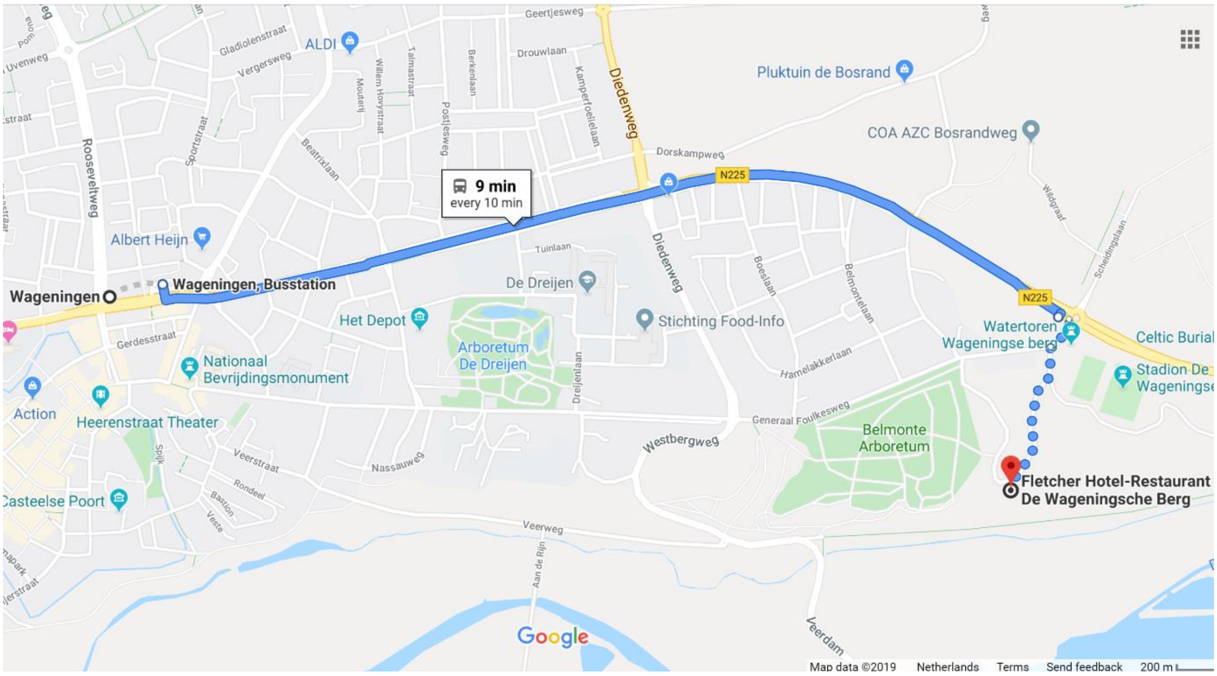
Car parking
Parking is available free of charge on the hotel premises.

By bicycle
The conference hotel is located about 10km from the nearest train station (Ede-Wageningen), and will take about 35 minutes to reach by bicycle from this station.

From Station Ede-Wageningen, take a bus to Bus Station Wageningen, and change there to Bus 352:



From Wageningen Bus Station, take Bus 352 direction Arnhem CS, and get off at Wageningse Berg:



Vereniging voor Ordinatie en Classificatie
Dutch/Flemish Classification Society
VOC

VOC Jubilee Meeting
21 – 22 November 2019

Wageningen, the Netherlands

Fletcher Hotel-Restaurant De Wageningsche Berg
Generaal Foulkesweg 96, 6703 DS Wageningen

Book of Abstracts

BIOMETRIS
Quantitative Methods brought to Life

Smit Consult
Consultancy in statistiek

Scope

The Dutch/Flemish Classification Society, VOC, aims at communicating scientific principles, methods, and applications of ordination and classification. The VOC is a member of the International Federation of Classification Societies (IFCS).

Adapting the self-organising map to the big data era

Ron Wehrens
Wageningen University & Research

Self-organising maps have become popular tools for clustering and visualizing multivariate data. The kohonen package in R, a popular implementation, has recently been completely redesigned [1], leading to significant performance improvements, a.o. through parallellization. In addition, it is now possible to apply user-defined distance functions, which makes the approach useful in new domains and often removes the need for data transformations.

In this presentation I will highlight some of the key improvements in the new version of the package, and show examples of dedicated distances leading to more interpretable results. In addition, examples will be presented on objects represented by several individual data layers. In this era of Big Data data from several different sources and with different characteristics are often combined - the kohonen package uniquely presents simple and powerful methods of applying SOM methodology to such data sets.

References:
[1] R. Wehrens and J. Kruisselbrink. Flexible self-organising maps in kohonen v3.0. *Journal of Statistical Software*, 87, 2018.

Kernel change point detection on the running statistics: A flexible, comprehensive and user-friendly tool

Eva Ceulemans
KU Leuven

In many scientific disciplines, studies have demonstrated that on top of the mean, signaling other types of changes is crucial to better capture and understand an event. For example, in emotion psychology, it has been uncovered that it is not only response patterning (i.e., simultaneous change in means) but also response synchronization (i.e., change in the correlations) that characterize response concordance during an emotional episode. In psychopathology research, recent evidence revealed that changes in three statistics, namely, the variance, autocorrelation and correlation, can serve as early warning signs before relapse to depression. In this presentation, we will present KCP-RS, a change point detection tool that can be tailored to capture changes not only in the means but in any statistic that is relevant to the researcher. KCP-RS implements KCP (Kernel Change Point) detection on the running statistics, a derived time series reflecting the statistics of interest. These running statistics are extracted by sliding a window across the time series, and in each window, computing the statistics value. Next, we will put forward a KCP-RS workflow to guide researchers in how to carry out the analysis when multiple running statistics need to be tracked. Finally, using stocks return data and physiological time series, we will introduce the R package we recently built to make KCP-RS freely and easily accessible to applied researchers.

Authors:
Eva Ceulemans, Jedelyn Cabrieto, Kristof Meers, Janne Adolf, Peter Kuppens, Francis Tuerlinckx (*KU Leuven)*

Stabilizing high-dimensional regression

Peter Bühlmann
ETH Zurich

The common notion of replicability of statistical discoveries deals with generalization from a data set to a new unobserved population from the same data-generating distribution (and is typically quantified by some statistical uncertainties). We discuss the problem in the context of regression when the new population comes from a different distribution than the one generating the observed data. We present a principled way of stabilizing regression with a simple yet effective regularization scheme to achieve replicability in such settings: it builds on distributional robustness and borrows ideas from causality. We highlight the potential and limitations of the approach and provide some illustrations on bio-medical data.

The (un)necessity of complex statistics in an N=1 world

Laura Bringmann
University of Groningen

More intensive longitudinal data is becoming available in which people such as patients with a clinical disorder are measured over a long time period, for example, 3 times a day for several months. This requires more complex modeling techniques. Or does it? In this talk, I will discuss different visions on what to do with N=1 data.

Implementing a biplot based multivariate data driven industrial performance index

Niël le Roux
Stellenbosch University

We introduce a novel approach which utilises Generalized Orthogonal Procrustes Analysis to find the optimal units and time period to employ as a reference set for the multivariate monitoring of multiple production processes simultaneously. Principal Component Analysis (PCA) and Canonical Variate Analysis (CVA) theory and biplots are evaluated and extended for the real-time monitoring of a complex industrial plant. Therefore, given a period where all the industrial processes (gasifiers) are in control, the following aspects are addressed:

- The selection of the optimal principal component combinations to utilize for possible biplot scaffoldings or visualizations.
- The optimal set of axes to include for each principal component combination for purposes of visualization.
- The identification of critical variables responsible for process performance deviations for each principal component combination presented.

In addition to the multidimensional visualization of the multiple production processes the proposed monitoring methodology also addresses the challenge to establish for each individual gasifier a real time gasifier performance index (GPI). This GPI should provide a single value which indicates the current health of the gasifier. Three different approaches are considered to address this challenge:

- A fundamental approach followed by process engineers to develop an index consisting of a weighted deviation from a recommended operating point for each process variable.
- A purely data driven (empirical) approach to develop a performance index making use of historical multivariate data collected during periods of good performance.
- The integration of the fundamental and empirical indices to develop a GPI index which uses both the multivariate statistical methodology, as well as the knowledge from subject matter experts.

Authors:
Niël le Roux (Stellenbosch University), Ruan Rossouw & Roelof Coetzer (Sasol Group R&T)

Improving the statistical assessment of industrial process quality: towards industry 4.0 and beyond

Tim Offermans
Radboud University Nijmegen

(Bio)chemical industrial production facilities suffer from many sources of variation, such as raw material variation, weather variation and operator variation. To deal these sources of variation and to optimize the production process, dedicated monitoring and control solutions are necessary. Developing and improving such solutions from the perspective of multivariate data analysis is a major interest in the field of chemometrics, especially since the introduction of industry 4.0.

In the presented research, two key issues in this field are addressed. The first issue is the dynamic synchronization of data measured at a production plant in real-time from different sources with different sampling times. The most commonly used synchronization methods and the quality of statistical models of the data obtained with each of those methods are discussed. The second issue is the monitoring of chemical changes in an industrial batch-reaction. A strategy to monitor these changes on an abstract level and batch-invariant is introduced. This strategy can be used to follow the productivity of the reaction, and allows for a more accurate detection of the end of production. This in turn leads to a more pure product and less waste of energy, time and raw material.

Authors:
Tim Offermans & Jeroen Jansen (Radboud University Nijmegen)

Managing churn to maximize profits

Aurélie Lemmens
Erasmus University Rotterdam

Customer defection threatens many industries, prompting companies to deploy targeted, proactive customer retention programs and offers. A conventional approach has been to target customers either based on their predicted churn probability, or their responsiveness to a retention offer. However, both approaches ignore that some customers contribute more to the profitability of retention campaigns than others. This study addresses this problem by defining a profit-based loss function to predict, for each customer, the financial impact of a retention intervention. This profit-based loss function aligns the objective of the estimation algorithm with the managerial goal of maximizing the campaign profit. It ensures (1) that customers are ranked based on the incremental impact of the intervention on churn and post-campaign cash flows, after accounting for the cost of the intervention and (2) that the model minimizes the cost of prediction errors by penalizing customers based on their expected profit lift. Two field experiments affirm that our approach leads to significantly more profitable campaigns than competing models.

Authors:
Aurélie Lemmens (Erasmus University Rotterdam), Sunil Gupta (Harvard Business School)

The role of conditional likelihoods in mixed-effects modeling

Anders Skrondal
University of Oslo & UC Berkeley

When applicable, constructing a conditional likelihood is one way of handling incidental parameters (whose numbers increases in tandem with the observations) in statistical models. In this talk I will argue that conditional likelihoods have an important role to play in mixed-effects and latent-variable modeling. In particular, such «fixed-effects» approaches can allow protective estimation under common challenges such as (1) unobserved confounding, (2) heteroskedasticity, (3) endogenous sampling, (4) cluster-size dependence, and (5) missing data. I will also discuss some limitations of conditional likelihood estimation.

Statistical challenges in nutrition studies: dealing with multi-target, non-adherence and ethical constraints

Sophie Swinkels
Danone Nutricia Research

Clinical trials investigating the effectiveness of nutritional interventions pose specific statistical challenges. For example, nutritional products often target more than one aspect of health. One challenging goal is to unravel the many-to-many ingredient–outcome relationships, for example, through applying machine learning (e.g. Bayesian networks/graphical modelling) to identify probabilistic pathways from chemical composition of a new compound to real world clinical efficacy. Extending statistically modelling into the manufacturing of nutritional products also presents many challenges, with highly complex recipes and production processes generating rich stochastic behaviour, with the goal to predict ahead of time the classification (in-spec/out of spec) of finished products.

Other challenges come from the fact that investigators and subjects tend to be rather sloppy in their adherence to protocol instructions when in a nutritional study. This results for instance in a high amount of out of window visits making the arbitrary means covariance pattern models that are used in pharmacological trials less suitable for our data. At Danone Nutricia Research (DNR), we often chose as the primary approach to model trajectories with some parametric function of time. Several models have been implemented like a polynomial random effect model or – as advised by L-BioStat - the Berkey-Reed model. However, how to handle non-monotonic trajectories is still an area of development.

A third example concerns the amount of drop outs. Nutritional interventions are often especially relevant for patients in the start phase of a degenerative disease. In such a population, a certain proportion of subjects will show a worsening of symptoms during the trial and will be put on rescue medication. Typically, the nutritional intervention is hypothesized to influence this process resulting in a drop out process that should be regarded as informative. To tackle this complexity, DNR has implemented competing risk joint models. These models include a non-linear element and random effects which have the consequence that the aggregated treatment effect has the subject-specific or conditional interpretation. However, the primary interest of an intervention study is typically on the population-averaged treatment effect. To derive the marginal estimate, we have proposed a method based on an extension of the approach of Hedeker et al. (2017, Biometrics, doi: 10.1111/biom.12707) using Monte Carlo simulations.

Studies on infant formula have to adhere to the World Health Organisation code of marketing of breast-milk substitutes. At DNR this has led to the implementation of “Best after Breast designs” which generates data with heterogeneity in feeding patterns. To investigate this source of variation we apply clustering of multivariate longitudinal data.

Logratio analysis versus correspondence analysis: and the winner is ...?

Michael Greenacre
Universitat Pompeu Fabra

Compositional data can be fun,
Their values always add up to one.
Drop a category and re-express,
The data change: it’s a mess!
Rather use ratios, then you’re done!

Correspondence analysis is also much fun
Even though its values don't add up to one
But counts you should relativize
To remove effect of sample size
I wonder: which method is better in the long run?

Networks in social contexts: the settings model

Tom Snijders
University of Groningen

When network dynamics is considered in groups that are not quite small, the heterogeneity of the network has to be taken into account. 'Heterogeneity' is a wide term, which can be elaborated in many ways. The 'Settings Model' is a newly developed stochastic actor-oriented model for network dynamics that specifies heterogeneity by distinguishing between the nearby and the far-off. The data structure includes, next to the network itself, a representation of the social context in which the network develops. Settings are regarded as meeting opportunities, structuring the creation of new ties. A setting is defined as a graph, usually non-directed, and rules for tie change depend on the setting. The primary setting is defined endogenously as an extended local network neighbourhood. In addition, the model assumes the existence of so-called meeting settings, given exogenously (e.g., classrooms, departments, the complete graph). The model defines tie changes in the primary setting as multinomial choices, similar to the standard stochastic actor-oriented model. Choices in the meeting settings are binary choices. The Settings Model is conceptually and computationally more attractive for representing network dynamics in larger networks. An overview of the model will be presented with an empirical application.

Background literature:
Tom A.B. Snijders (2017). Stochastic Actor-Oriented Models for Network Dynamics. *Annual Review of Statistics and Its Application*, 4, 343-363.

Computational statistics and open science

Anne-Laure Boulesteix
LMU München

In the first part of my talk, I will give a brief overview of the concept of open science with a special focus on aspects which are particularly relevant to prediction modelling. This includes issues such as the publication of data and code for the purpose of reproducibility, the prevention of "fishing expeditions" and related questionable research practices, and the appropriate representation of uncertainties related to the data analysis strategy. In the second part of my talk, I will argue that open science principles are not only relevant to data analysts working in applied projects, but also, in a perhaps more subtle and not very well-understood way, to methodological researchers (i.e. researchers working on the development of new data analysis methods) in their own research projects. This second part will be illustrated through several empirical meta-research projects with focus on prediction modelling.