Title and abstracts:

Plenary Lecturer 1: Anna Kiriliouk – University of Namur (Belgium)

Title: Estimating failure probabilities for high-dimensional extremes

Abstract:
An important problem in risk management is the estimation of the probability that a random d-dimensional vector X falls into a given extreme "failure set". For example, the components of X may represent a certain environmental index (e.g., maximal hourly wind speeds), observed on d spatial locations. A catastrophic event is commonly due to a combination of variables being extreme simultaneously; hence, an important challenge is to characterize such tail dependence between the components of X. Parametric tail dependence models for high-dimensional environmental data are often based on a spatial correlation function. However, as the dimension d increases, such an approach becomes increasingly complex, whereas models with a small number of parameters tend to oversimplify the tail dependence structure. On the other hand, nonparametric methods suffer from the curse of dimensionality. We focus on a generalisation of the so-called tail pairwise dependence matrix (TPDM), which gives a partial summary of tail dependence for all pairs of components of X. Cooley & Thibaud (2019) showed that a completely positive decomposition of the TPDM of X gives the parameter matrix of a max-linear model whose TPDM is equal to that of X. The max-linear model is a simple but rather popular extreme-value model because of its computational convenience. Moreover, it is dense in the class of d-dimensional multivariate extreme-value distributions. Failure probabilities could therefore be modeled using the max-linear model if we can estimate its parameter matrix through a completely positive decomposition. Unfortunately, exact algorithms for obtaining such decompositions tend to be computationally heavy. We propose an algorithm to obtain approximate completely positive decompositions of the TPDM. It is applicable to dimensions in the order of hundreds; moreover, obtaining multiple decompositions allows us to quantify the model uncertainty. We apply the proposed algorithm to a dataset of maximal wind speeds in the Netherlands.

Plenary Lecturer 2: Mirjam Moerbeek - Utrecht University (Netherlands)

Title: Optimal design of cluster randomized trials

With cluster randomized trials complete groups such as schools, households or family practices are randomized to treatment conditions. Cluster randomized trials are less efficient than individual randomized trials, hence it is important they are designed in an optimal way.
This presentation gives a short introduction to the cluster randomized trial and the statistical model to analyze data from such a trial. It then shows how to calculate the optimal number of clusters and cluster size such that highest power for the test on treatment effect is achieved. Software for calculating the optimal design is demonstrated. After that, an extension is made to multiperiod designs where clusters and the subjects therein are measured multiple times in a longitudinal trial. Examples are the cross-over trial and the stepped-wedge design. The optimal design of such trials under attrition of subjects and/or clusters over time is discussed and it shown how to repair for the loss of efficiency due to attrition.

Plenary Lecturer 3: Piet Daas - Statistics Netherlands (Netherlands)

Title: Big Data and Official Statistics: Challenges and Applications at Statistics Netherlands

Abstract:
The use and application of Big Data in official statistics has made considerable progress at Statistics Netherlands. The major contributors are the increased attention for Big Data in the methodological research program, in the creation of experimental statistics and in its use for regular statistics production. To stimulate this the Center for Big Data Statistics was setup in 2016. The most important research challenges identified are: 1. Concept: What (derived) concept is measured in Big Data? 2. Population: What part of the target population is included in Big Data? 3. Methods: What new methods (or new ways of thinking) are needed? 4. Infra: What infrastructural requirements are needed? The infrastructural (IT) challenge is ignored here. The fact that there is a steady increase in the application of Big Data at the office indicates the need (and progress made) in the study of the research challenges identified. The statistics that make use of Big Data and are either in production or for which an implementation process has started at Statistics

Netherlands are: 1. Using scanner data and scraped prices for the Consumer Price Index 2. Using road sensor data for Traffic Intensity statistics 3. Using website texts for Online Platform Economy statistics 4. Using social media for the Social Tension indicator 5. Using texts of online job advertisement for Vacancy statistics 6. Using solar panel output and weather data for Solar Energy production The presentation will discuss the research challenges and how this has affected the use of Big Data for official statistics at the office.

Plenary Lecturer 4: Holger Dette - Ruhr-Universitaet Bochum (Germany)

Title: Equivalence of regression curves

This paper investigates the problem whether the difference between two parametric models m_1, m_2 describing the relation between a response variable and several covariates in two different groups is practically irrelevant, such that inference can be performed on the basis of the pooled sample. Statistical methodology is developed to test the hypotheses $H_0: d(m_1, m_2) \geq \epsilon$ versus $H_1: d(m_1, m_2) < \epsilon$ to demonstrate equivalence between the two regression curves m_1, m_2 for a pre-specified threshold ϵ, where $d(.)$ denotes a distance measuring the distance. Our approach is based on the asymptotic properties of a suitable estimator $d(\widehat{m_1}, \widehat{m_2})$ of this distance. In order to improve the approximation of the nominal level for small sample sizes a bootstrap test is developed, which addresses the specific form of the interval hypotheses. In particular, data has to be generated under the null hypothesis, which implicitly defines a manifold for the parameter vector. The results are illustrated by means of a simulation study and a data example. It is demonstrated that the new methods substantially improve currently available approaches with respect to power and approximation of the nominal level. The results have been applied in cooperation with the EMA and FDA for comparing dissolution profiles.