# STATalk 2026

## Book of Poster Abstracts

### Poster Presenters

Irene Ariante
Anna Ballario
Cecilia Bergamini
Teresa Bortolotti
Martino Bussa
Michele Cavazzutti
Rossana Ciarfaglia
Simone Colombara
Alessandro Colombi
Niccolò D'Agaro
Alessia D'Ambrosio
Marco Francesco De Sanctis
Riccardo De Santis
Claudio Del Sole
Ilenia Di Battista
Chiara Di Maria
Michela Frigeri
Giulio Grossi
Gaia Gubelli
Maria Grazia Manco
Leonardo Marchesin
Andrea Mascaretti
Emanuele Masillo
Adelajda Matuka
Marta Nesteruk
Riccardo Pajno
Edoardo Pandolfo
Luca Perego
Daniele Petrone
Giuliana Polo
Riccardo Racca
Irene Rotondo
Elena Sabbioni
Muhammad Amir Saeed
Silvia Scarpa
Andrea Teruzzi

Irene Ariante (Università degli Studi di Napoli Federico II)

# Beyond prediction: a configurational perspective on persistent default risk in SMEs

**Abstract:** *This study reconceptualizes corporate default as a cumulative process of organizational deterioration rather than a discrete financial event. By integrating research on organizational decline with advanced machine learning, the paper introduces a diagnostic framework centered on the divergence from strategic financial normality. Strategic financial normality represents the historically viable configuration of financial relationships associated with long term sustainability within a specific institutional context. The analysis develops the concept of configurational financial misalignment to capture the progressive accumulation of structural vulnerabilities that precede formal insolvency. The empirical investigation utilizes a large panel of 32,801 Italian small and medium sized enterprises observed between 2016 and 2024. This context is particularly salient due to the high informational opacity and bank dependency characteristic of SMEs, which amplifies the importance of accounting based signals. The research employs a hybrid analytical approach combining unsupervised anomaly detection via Deep Isolation Forest with supervised XGBoost classification to identify systematic deviations from sustainable configurations. To ensure the practical utility of the findings, the study applies explainable artificial intelligence techniques, specifically SHAP values, to transform complex statistical patterns into structured diagnostic insights. The results demonstrate that anomaly scores derived from configurational deviations increase progressively as firms approach default and remain systematically higher among firms exposed to persistent risk. This temporal pattern suggests that financial vulnerability emerges through a structured and evolving process rather than through isolated fluctuations in individual ratios. Capital structure and profitability consistently emerge as the most stable drivers of vulnerability, while the anomaly score serves as a key synthetic predictor of underlying misalignment. By linking financial architectures to trajectories of decline, the study advances a configurational understanding of default risk and demonstrates how interpretative analytical tools can support early strategic diagnosis in high uncertainty environments.*

Anna Ballario (Politecnico di Torino; Istituto Nazionale di Ricerca Metrologica)

# Bayesian hierarchical modeling for reliable large-scale sensors deployments and applications

**Abstract:** *Reliability of large populations of sensors is a major challenge in modern industrial production and monitoring systems. The widespread deployment of low-cost Micro-Electro-Mechanical Systems (MEMS) sensors across multiple domains requires calibration strategies that ensure metrological reliability while remaining scalable. Traditional laboratory calibration procedures become impractical for large volumes, motivating the development of statistical approaches. This work develops a Bayesian framework for the virtual calibration of large sensor batches. A earlier published model exploits information from a reference, laboratory-calibrated batch to infer the calibration properties of new production lots. By calibrating only a small subset of sensors, the model estimates key parameters characterizing the entire batch, enabling reliable assessments while reducing calibration effort. A hierarchical extension is introduced by incorporating a Beta hyperprior distribution on the probability of detecting out-of-tolerance sensors. This formulation allows the integration of prior industrial knowledge while explicitly controlling parameter variability. It relaxes deterministic assumptions about batch quality and improves the flexibility of reliability estimates. The selection of hyperprior parameters is therefore crucial to ensure consistency with prior knowledge. The study evaluates reliability metrics under different model specifications and proposes alternative measures to address identified limitations. Validation is conducted on a dataset of 100 MEMS accelerometers calibrated at INRiM. Results show that weakly informative hyperpriors increase the influence of observed data: when fewer defective sensors are observed than expected, higher reliability and lower uncertainty are obtained; conversely, when more defects are detected, uncertainty increases significantly. This highlights the importance of consistency between prior assumptions and observed data. Future work will incorporate cost and utility functions to model producer's and consumer's risks, supporting decision-making processes. These developments will help balance the trade-off between increased reliability estimates and higher posterior uncertainty.*

Cecilia Bergamini (Politecnico di Torino)

# A comparative study of statistical methods for high-dimensional multi-omics data integration

**Abstract:** *This work is developed within the context of a Master's thesis and aims to compare and evaluate different statistical methods for the integration of heterogeneous omics data, with the goal of assessing their strengths and limitations when applied to a specific high-dimensional dataset. The analysis focuses on transcriptomic and proteomic data provided by the Telethon Foundation, with application to Ehlers-Danlos Syndrome, where the number of variables largely exceeds the number of samples and poses significant methodological challenges in identifying disease-related patterns. The datasets include 14 samples, with transcriptomic measurements on the order of 10,000 genes and proteomic data comprising approximately 5,000 proteins. Given the availability of prior information on sample group membership (healthy vs. affected individuals), both unsupervised and supervised multivariate techniques are considered. In particular, we first apply Multi-Omics Factor Analysis (MOFA, implemented in the MOFA2 R package), an unsupervised Bayesian latent factor model that decomposes multiple data matrices into shared and view-specific latent factors, capturing sources of variability across omics layers through a probabilistic matrix factorization framework. As a supervised counterpart, we use sparse Partial Least Squares Discriminant Analysis (sPLS-DA, implemented in the mixOmics R package), which is based on PLS regression and incorporates sparsity constraints to perform simultaneous classification and feature selection. Model performance for the supervised approach is assessed within a cross-validation framework. Preliminary results show that the supervised approach (sPLS-DA) achieves better performance in terms of group discrimination and interpretability compared to MOFA. Ongoing work focuses on refining the comparative analysis, developing an improved integrative methodological framework tailored to the characteristics of the datasets, and enhancing biological interpretation through downstream gene ontology analysis.*

Teresa Bortolotti (Politecnico di Milano)

## Conformal classification with tight marginal coverage in noisy settings

**Abstract:** *Conformal prediction is a nonparametric method widely applied in regression, classification, and outlier detection, providing valid predictive inference with finite-sample coverage guarantees. Marginal coverage, in particular, is a fundamental objective in conformal inference, ensuring that prediction sets contain the correct label for a predefined proportion of future test points. However, these guarantees rely on the assumption of data exchangeability, which is often violated in real-world applications due to distribution shifts, outliers, and label noise. In this work, we address the limitations of conformal classification in the presence of label contamination and propose novel adaptive methodologies that automatically adjust for noise to restore marginal coverage. Our approach builds on results from empirical process theory to derive correction factors that account for the inflation of coverage caused by label noise. These adaptive calibration methods not only ensure nominal marginal coverage but also maintain informative prediction sets, even in challenging classification tasks with a large number of classes or severe class imbalance. The proposed method relies on the assumption that the contamination process is known, and we complement it with a strategy to estimate the contamination from the data. The effectiveness of our approach is demonstrated through extensive experiments on synthetic and real-world datasets, including CIFAR-10H and BigEarthNet.*

Martino Bussa (Università degli Studi di Milano Bicocca)

## Prediction of mortality and hospitalisation in heart failure using healthcare administrative data and random forest models

**Abstract:** *Introduction: Heart failure (HF) is a clinical syndrome associated with reduced quality of life, high healthcare resource use, and premature mortality. In the era of big data, machine learning (ML) has emerged as a tool for analysing complex datasets, and its use in predicting outcomes in HF has increased in recent years. Among ML methods, Random Forest (RF) is a flexible approach for classification and risk stratification. In many ML applications, calibration is not evaluated, although good calibration is essential for potential clinical use.*

*Objective: To develop an RF predictive model for all-cause mortality and hospitalisation in adult HF patients, using healthcare administrative data. Methods: Through linkage of healthcare administrative databases from ATS Milan (disease-specific exemption, hospital discharge records, and drug prescription records), resident patients with HF were identified as of 31/12/2018. The cohort was divided into a training set and a test set using a 75:25 ratio. Weighted and unweighted RF models were developed, using weights inversely proportional to class frequency. Model performance was evaluated on the independent test set using the area under the receiver operating characteristic curve (AUROC). The classification threshold was selected using the F2-score in order to favour sensitivity. Model calibration was assessed using the Brier score, calibration intercept and slope, and was graphically explored using a LOESS calibration curve with bootstrap uncertainty bands.*
*Results: A total of 70,249 patients with HF were identified (median age 78 years; 52% men). In 2019, 7,037 all-cause deaths (10.0%) and 22,527 all-cause hospitalisations (32.1%) occurred. The unweighted RF model showed good discriminative ability for mortality, with an AUROC of 0.7929 (95% CI: 0.7826–0.8031), whereas discrimination for hospitalisation was moderate, with an AUROC of 0.6587 (95% CI: 0.6502–0.6672). The corresponding AUROCs for the weighted RF models were 0.7859 (95% CI: 0.7756–0.7961) for mortality and 0.6583 (95% CI: 0.6498–0.6668) for hospitalisation. In the test set, the unweighted RF models produced mean predicted risks close to the observed event rates, whereas the weighted RF models tended to overestimate risk. Conclusion: In patients with HF identified through healthcare administrative data, unweighted RF models showed better calibration than weighted models and maintained good discrimination for mortality. These findings support the potential use of healthcare administrative data for prognostic risk stratification. The integration of clinical data may further improve model performance, particularly for hospitalisation.*

Michele Cavazzutti (Charles University)

## Voronoi depth for partially observable data on multidimensional domains

**Abstract:** *We propose a depth measure for functional data defined over non-convex multidimensional domains, characterized by complex irregular geometry of the support. Thanks to the use of an appropriate Voronoi tessellation of the support of the data, the method enables the analysis of functional data that are observed irregularly over their domain, effectively managing partial observability, even when the functional datum is missing over large portions of the domain. The proposed depth does not require*

*prior reconstruction of the data. It also achieves desirable asymptotic statistical properties while exhibiting high computational efficiency. We illustrate our method through the analysis of Earth surface temperatures from the CESM2 Large Ensemble Community Project. The proposed depth is employed to investigate extreme temperature patterns over the 1850–1975 baseline, enabling visualization via bagplots and functional boxplots. The results are consistent with the established literature on climate change.*

Rossana Ciarfaglia (Politecnico di Torino)

## MCMC-based Bayesian inference with ODE Systems and compositional data

**Abstract:** *This master's thesis addresses the problem of parameter estimation in a kinetic model of carbon dioxide and carbon monoxide methanation over a nickel-based catalyst, within the context of Power-to-Gas technologies for the conversion of electrical energy into stable energy carriers. The dataset consists of 900 experiments, obtained by varying temperature, pressure, and the composition of the reacting mixture. However, the available observations regarding the composition of the mixture concern only the inlet and outlet conditions of the reactor: no measurements are available along the internal spatial domain.*
*The model describes, through a system of ordinary differential equations, the information along the reactor, defined as a spatial axis along which the solution of such equations provides an internal concentration profile related to the variation of the species. This structure makes the issue of model identifiability central, and the results show the consistency of the model through the comparison between simulated outputs and experimental data. The extension of this deterministic structure is carried out through a Bayesian procedure implemented in the Julia programming language, in which the deterministic model defines the mean of the distribution of the observed data. Since the quantities of interest are compositional, the inference is conducted in the log-ratio space of molar fractions, adopting a Gaussian error model consistent with compositional constraints. The likelihood function is therefore defined on log-ratio transformations, while prior distributions play a regularization role in the estimation of the 23 model parameters: normal priors are considered for transformed parameters, constrained priors for parameters subject to physical limits, and inverse-gamma priors for the noise variance. Inference is performed using a Metropolis-Hastings within Gibbs algorithm with adaptive proposal. The developed model is aimed at recovering kinetic and thermodynamic parameters, evaluating the ability of the MCMC procedure to reconstruct the true values in the presence of noise in the observations.*

Simone Colombara (Politecnico di Milano)

## Multivariate Bayesian prediction of metabolic syndrome

**Abstract:** *Metabolic Syndrome (MetS) is a complex clinical condition characterized by the simultaneous presence of multiple interconnected risk factors. The early identification of these alterations is crucial for effective prevention. This study presents a Bayesian statistical framework to estimate and predict the risk of MetS by analyzing a large longitudinal dataset of voluntary blood donors (AVIS Milano), encompassing repeated clinical measurements and lifestyle data. We propose a multivariate mixed-effects model that jointly analyzes the five diagnostic components of MetS. This architecture explicitly captures the covariance structure among biomarkers, enabling the model to "borrow strength" across outcomes for the robust imputation of missing responses (waist circumference). The model employs a grouped horseshoe prior for structured covariate selection and patient-specific random intercepts to account for baseline heterogeneity. To translate the complex probabilistic output into a practical clinical tool, we developed a "traffic light" classification algorithm. By evaluating posterior predictive mean against an optimal threshold (derived by maximizing Youden's J statistic), visits are stratified into low (green), potential (yellow), or high (red) risk zones. Compared to traditional discriminative machine learning algorithms, our generative Bayesian approach ensures rigorous uncertainty quantification, excellent sensitivity (successfully flagging 100% of at-risk donors in the alert zones), and enhanced medical interpretability. Ultimately, this provides clinicians with a transparent, theoretically sound tool natively capable of handling missing data to support targeted preventive interventions during routine screenings.*

Alessandro Colombi (Università Bocconi)

## Bayesian model-based inference for modal missing species and features

**Abstract:** *Many discovery problems require assessing whether additional sampling is likely to reveal new categories with non-negligible prevalence. We study this question through the lens of the maximum unseen probability, defined as the largest probability among categories not yet observed in the sample, and construct one-sided interval estimates for this quantity. Unlike existing distribution-free approaches, which provide universal but often*

*overly conservative bounds, we adopt a Bayesian nonparametric perspective. By specifying flexible priors, we obtain closed-form expressions for posterior upper bounds that explicitly model structural features of the underlying population, such as power-law behavior or a finite alphabet of unknown size. Our framework encompasses the two most common sampling schemes: Bernoulli product models, widely used for presence–absence (incidence) data, and multinomial sampling models, which arise naturally in abundance settings. In both cases, the resulting bounds are substantially sharper than their frequentist worst-case counterparts. We illustrate the proposed methodology through an application to criminal network data concerning the activities of the 'Ndrangheta in Northern Italy.*

Niccolò D'Agaro (Università degli Studi di Milano Bicocca)

## Critical evaluation of the performance of different spatial segmentation methods in distinct simulated scenarios

**Abstract:** *Mass Spectrometry Imaging (MSI) is a powerful technology that enables spatially resolved molecular profiling of biological tissues. However, the high dimensionality and spatial complexity of MSI data require robust computational frameworks to extract meaningful clinical information. Within this context, clustering-based segmentation techniques have emerged as a critical statistical tool for accurately partitioning tissues into distinct, biologically relevant regions. This study compares several unsupervised spatial clustering methods characterized by distinct grouping criteria, including traditional point-based algorithms (k-means, Gaussian Mixture Model) and advanced methods that incorporate spatial information (Spatially Aware and Spatially Aware Structurally Adapted k-means, Spatial Shrunken Centroids, Bayesian Analytics for Spatial Segmentation). Is also introduced a new approach using a Gaussian Graphical Mixture Model (GGMM) and relying on Hidden Markov Random Fields to account for the spatial nature of the dataset. The performance of these methods is tested across multiple simulated scenarios characterized by varying levels of noise and structural complexity. The further analysis of a real MALDI-MSI epileptic brain sample confirms that accounting for spatial dependencies and probabilistic structures significantly enhances segmentation performance. This work provides a critical benchmark for selecting appropriate segmentation strategies, ultimately supporting reliable tissue characterization and laying the groundwork for advanced histopathological biomarker analysis.*

Alessia D'Ambrosio (Università degli Studi di Napoli Federico II)

## Data science for behaviour understanding in maintenance of industrial machinery

**Abstract:** *Predictive maintenance is a critical challenge for industry today, where continuous monitoring of machinery can prevent costly failures and optimize operational efficiency. This work proposes a data-driven framework for condition monitoring of industrial machinery, integrating machine learning methodologies and statistical inference to improve fault prognosis and anticipate malfunctions. The case study concerns an industrial pumping system consisting of several electromechanical units operating in a parallel configuration. The data comes from a network of wireless sensors that detect temperature, pressure, vibration, and ambient weather conditions. The proposed analytical framework is divided into three phases. The first phase aims at the identification and characterization of the operating conditions of the machinery: DBSCAN clustering is applied to identify the operating regimes, Principal Component Analysis to interpret the clusters obtained and classification trees to extract operating rules. The second phase evaluates the impact of external environmental conditions on operation: through statistical tests and Variable Importance analysis, it is verified whether pressure and atmospheric temperature determine significant differences in performance. The third phase analyzes the interaction between jointly operating machinery, employing Granger Causality Analysis to identify causal relationships, regression to estimate direction and magnitude, and Cross-Correlation Analysis to identify time delays. The results show that the pumping units have differentiated start-up dynamics, with distinct operating regimes identifiable through clustering. External weather conditions are statistically significant in determining variations in the performance of machinery. Finally, bilateral coordination between the units emerges, with compensatory effects on the workload and synchronous responses for thermal and pressure variables.*

Marco Francesco De Sanctis (Politecnico di Milano)

## Physics-informed mixed-effects regression for spatio-temporal environmental data

**Abstract:** *The analysis of spatio-temporal environmental data, such as air pollutants in the atmosphere, is characterized by complex physical dynamics and heterogenous external factors which influence the*

*underlying phenomenon. For instance, sensors monitoring nitrogen dioxide are based on distinct measurement technologies, which introduces a natural group structure in the data. This work introduces a semiparametric mixed-effects regression model regularized by partial differential equations. The proposed framework embeds prior physical knowledge into the statistical estimation process, capturing diffusion and advection phenomena driven by wind streams. Through the inclusion of random effects, the model disentangles group-specific systematic biases, such as sensor-related noise originating from the different monitoring technologies, from the underlying true signal. Finally, fixed effects account for external natural and anthropic factors, including altitude and population density across the spatial domain. The proposed semiparametric estimation procedure is based on the minimization a penalized loss functional that balances data fidelity with physics-informed spatial and temporal penalties. Asymptotic Gaussianity of both parametric and nonparametric estimators is investigated. Finally, the model is applied to analyze hourly nitrogen dioxide data in Lombardy region (Italy) to identify critical pollution areas and support environmental monitoring strategies.*

Riccardo De Santis (Università degli Studi di Padova)

## Inference on multiple quantiles in regression models by a rank-score approach

**Abstract:** *This work tackles the challenge of performing multiple quantile regressions across different quantile levels and the associated problem of controlling the familywise error rate, an issue that is generally overlooked in practice. We propose a multivariate extension of the rank-score test and embed it within a closed-testing procedure to efficiently account for multiple testing. Then we further generalize the multivariate test to enhance statistical power against alternatives in selected directions. Theoretical foundations and simulation studies demonstrate that our method effectively controls the familywise error rate while achieving higher power than traditional corrections, such as Bonferroni.*

Claudio Del Sole (Università degli Studi di Milano Bicocca)

## A Bayesian nonparametric approach to the multi-armed bandit problem in trait allocation models

**Abstract:** *In trait allocation models, each observation may display multiple features and may have different levels of belonging for each feature. For example, multiple individuals from distinct species may be observed within a time or spatial window. When data are collected from multiple populations, the same feature may appear in two different observations both within and across populations. We consider the problem of maximizing the cumulative number of distinct observed features by sequentially selecting the population from which the next observation is sampled. This task can be naturally framed as a multi-armed bandit problem, where payoffs correspond to the number of newly discovered features. We propose a Bayesian nonparametric approach relying on hierarchical gamma processes, which promotes borrowing of information among populations by establishing a common set of features, while accounting for their heterogeneity through population-specific parameters. The tractable posterior characterization facilitates the implementation of a Thompson sampling strategy to effectively balance exploration and exploitation. This standard approach is compared with a simpler strategy that selects the population with the highest posterior estimate for the number of new features, conditionally on populations parameters. For further comparison, we also consider a novel frequentist estimator, derived along the lines of the popular Good-Turing estimator for species sampling model. The performances of the proposed algorithms are assessed through simulation studies, and compared on two real datasets: the first contains tree species counts from ecological survey plots at various locations in Japan, the second comprises sentence-level words counts from multiple books.*

Ilenia Di Battista (Politecnico di Milano)

## Modeling non-stationarity via PDE penalization: an application to mobility data

**Abstract:** *This work presents a novel methodology for modeling anisotropy and non-stationarity in spatio-temporal phenomena. Many real-world processes across diverse fields, such as telecommunications and urban development, are inherently non-stationary and anisotropic, exhibiting*

*varying statistical properties across both spatial and temporal domains. Traditional spatio-temporal models often assume stationarity, limiting their ability to capture such variability. The proposed methodology introduces a nonparametric spatio-temporal regression model that incorporates a partial differential equation regularization to effectively model anisotropy and non-stationarity. This regularization term is expressed through a second-order linear differential operator, providing the flexibility to integrate problem-specific knowledge. The utility of this model is demonstrated through its application to the Telecom Italia dataset, which consists of mobile phone data collected over a fine spatio-temporal grid. The analysis focuses on the metropolitan area of Milan, where spatial dynamics are strongly influenced by proximity to highways and roads, which serve as preferential signal directions, and temporal dynamics reflect daily and weekly human activity cycles. The resulting signal patterns are highly complex, displaying localized spatio-temporal behaviour driven by geography and urban mobility. This highlights the importance of accounting for non-stationarity when analyzing, for instance, mobility patterns. Information on non-stationary anisotropy is mathematically encoded using diffusion tensors, positive definite matrices that provide a powerful representation of directional dependencies. However, working with tensors poses several mathematical challenges, such as preserving their positive definite property during operations like averaging or interpolating, as exemplified by modeling road intersections in the Telecom dataset. To address these challenges, we employ the Log-Euclidean metric to develop a mathematical framework that enables tensor operations while allowing the derivation of eigenvalue and eigenvector properties of tensors obtained through averaging.*

Chiara Di Maria (Università degli Studi di Palermo)

## Investigating spillover mechanisms in soil via causal mediation under spatial interference

**Abstract:** *Mediation analysis is widely used to investigate the mechanisms linking exposures to outcomes, but standard approaches assume independent units, an assumption often violated in environmental and soil studies where spatial dependence induces interference. In such settings, exposures and mediators at one location may affect outcomes at neighbouring sites, making it difficult to disentangle direct and indirect pathways. We propose a causal mediation framework under spatial interference that decomposes the total effect into four components: local direct, spillover direct, local indirect, and spillover indirect effects. Identification extends sequential ignorability to incorporate kernel-weighted*

*neighbourhood summaries of both exposure and mediator, and estimation is performed via Monte Carlo g-computation combined with spatial block bootstrap. A simulation study based on Gaussian process confounding shows that local effects can be estimated with low bias and variance across a wide range of scenarios, whereas spillover effects are more sensitive to the strength and spatial scale of confounding. Explicit modelling of residual spatial correlation improves estimation of spillover effects under strong, short-range dependence. We illustrate the framework on a dataset of 703 soil profiles from Sicily, analysing the causal pathway from clay content to gravimetric field capacity mediated by cation exchange capacity, considering both vertical spillover across soil depths and spatial spillover across sites. Results indicate that local effects are estimated more precisely than spillover effects, consistent with the predominantly local nature of the underlying pedological processes.*

Michela Frigeri (Libera Università di Bolzano)

## A spatially informed latent position model

**Abstract:** *Mediation analysis is widely used to investigate the mechanisms linking exposures to outcomes, but standard approaches assume independent units, an assumption often violated in environmental and soil studies where spatial dependence induces interference. In such settings, exposures and mediators at one location may affect outcomes at neighbouring sites, making it difficult to disentangle direct and indirect pathways. We propose a causal mediation framework under spatial interference that decomposes the total effect into four components: local direct, spillover direct, local indirect, and spillover indirect effects. Identification extends sequential ignorability to incorporate kernel-weighted neighbourhood summaries of both exposure and mediator, and estimation is performed via Monte Carlo g-computation combined with spatial block bootstrap. A simulation study based on Gaussian process confounding shows that local effects can be estimated with low bias and variance across a wide range of scenarios, whereas spillover effects are more sensitive to the strength and spatial scale of confounding. Explicit modelling of residual spatial correlation improves estimation of spillover effects under strong, short-range dependence. We illustrate the framework on a dataset of 703 soil profiles from Sicily, analysing the causal pathway from clay content to gravimetric field capacity mediated by cation exchange capacity, considering both vertical spillover across soil depths and spatial spillover across sites. Results indicate that local effects are estimated more precisely than spillover effects, consistent with the predominantly local nature of the underlying pedological processes.*

Giulio Grossi (Università di Firenze)

# Between a rock and a hard place: mitigating latent spatial confounding and spillover in synthetic control method

**Abstract:** *Standard Synthetic Control Methods (SCM) assume a convex combination of control units can reconstruct a treated unit's counterfactual trajectory by matching pre-treatment characteristics. However, this approach is inherently space-blind, completely ignoring geography. In spatial data governed by latent factor models, this creates a "Spatial Trilemma" with two opposing failure modes.First, the absence of spatial constraints leads to Spatial Mismatch. By minimizing only the pre-treatment prediction error, the model may select geographically distant donors that fit the pre-intervention curve purely through statistical noise, causing overfitting. In the post-treatment period, this generates severe bias due to latent spatial confounding. Second, restricting the donor pool exclusively to geographic neighbors—necessary to capture similar latent structures—inevitably introduces Spillover Contamination. Nearby units often experience the treatment's spillover effects, violating the Stable Unit Treatment Value Assumption (SUTVA).To solve this trade-off, the project proposes a Spatially Penalized Synthetic Control Method (SP-SCM). This estimator introduces an exogenous distance-based penalty directly into the weight optimization. By tuning this parameter via a predefined tolerance threshold, researchers accept a controlled, slight increase in pre-treatment error—an "insurance premium"—to enforce spatial coherence. This regularizes the synthetic control, successfully navigating the trade-off between spatial overfitting and spillover risks.The methodology is supported by a sensitivity analysis framework based on partial identification. Monte Carlo simulations demonstrate that SP-SCM outperforms standard estimators (Naive, Donut, and ASCM) across multiple scenarios. Finally, the method is empirically validated by assessing the 2009 tourism revitalization along Tuscany's Via Francigena route.*

Gaia Gubelli (Università degli Studi di Milano Bicocca)

# Modeling dependence structure in MALDI-MSI data: a copula-based statistical framework for thyroid lesion analysis

**Abstract:** *The application of spatial omics techniques, such as MALDI-MSI, to biopsy-derived tissue samples allows to capture complex, high-dimensional molecular information together with the native spatial arrangement of the tissue. These data typically exhibit strong spatial dependence and non-Gaussian distributions, including skewness and zero inflation. Standard statistical approaches are often inadequate in this setting, as they ignore spatial dependence and rely on restrictive assumptions that may lead to biased inference. This study aims to develop and compare different methodologies that explicitly incorporate spatial information and account for complex distributions, with a particular focus on the usage of copula models. Each tissue sample consists of a spatial grid which records multiple molecular signals for each coordinate. For each sample, the dependence structure between molecules is translated into a graph model, with nodes representing molecules and edges depicting pairwise association. To account for spatial information, three approaches are considered. Firstly, B-spline regression is employed to remove large-scale spatial trends before estimating associations. Second, Spatial autoregressive (SAR) models are used to capture local spatial autocorrelation. A third strategy operates on original data: a novel model combining copula functions with Zero-Adjusted Gamma (ZAGA) marginal distributions is introduced to handle zero-inflated and skewed intensity distributions typically of MALDI-MSI data. Resulting networks are compared using graph-based distances. Analyses indicate that accounting for spatial information and non-typical distributions affects the association structure between molecules. Strategies based on spatial preprocessing (B-splines and SAR) and copula models allow to capture different aspects of the data, leading to graph models with distinct topological properties. By applying the methodology to real clinical biopsy samples of different thyroid lesion types, substantial inter-sample heterogeneity is observed. While some differentiation between major pathological groups can be detected, a consistent clustering across all classes is not observed.*

Maria Grazia Manco (Università degli Studi di Bari Aldo Moro)

## Seasonal cetacean calf hotspots in the Gulf of Taranto using INLA

**Abstract:** *The Gulf of Taranto (northern Ionian Sea) is recognized as a critical habitat for several cetacean species, with the striped dolphin (Stenella coeruleoalba), common bottlenose dolphin (Tursiops truncatus), and Risso's dolphin (Grampus griseus) representing some of the most frequently observed species in the study area. However, this basin is also characterized by intense and diverse anthropogenic pressures, including military activities, aquaculture, fishing, and maritime tourism. Understanding how these pressures intersect with critical life-history stages, such as calving, is essential for effective conservation. This study aims to identify, for each of the three target species, the hotspot areas of calf presence in different seasons. Data collected between 2013 and 2023 were analyzed. We implemented a presence-only marked point process model within a spatio-temporal Bayesian framework using the R package inlabru, accounting for barrier effects and spatial thinning. The model integrated presence data with a mark representing the calves' proportion within each dolphin pod. A set of physico-chemical covariates characterizing the Gulf of Taranto was used to model habitat suitability and calf distribution. Our results provide an initial quantitative perspective on season-specific patterns of calf occurrence, potentially informing future conservation and management discussions in a basin subject to intense human pressures.*

Leonardo Marchesin (Politecnico di Milano)

## Spatial statistical models on linear networks for sea surface temperature hot-spot identification

**Abstract:** *This work introduces a streamlined geostatistical framework for modeling sea-surface temperature (SST) variability by directly incorporating major ocean currents into the spatial covariance structure of the field. We construct a physics-informed network across the marine environment, where nodes represent locations and edges map dominant flow pathways. Over this network, we establish a Gaussian random field utilizing a convolution-based covariance process to build in a constructive way the covariance operator of the field. Parameter estimation is stabilized through a regularized inference scheme, ensuring robust covariance estimates. By evaluating residuals and inferring their covariance through this new model,*

*we drive scenario-based Monte Carlo simulations to generate full-distribution SST forecasts across multiple emission trajectories. This parsimonious, physically grounded approach enhances uncertainty quantification—facilitating the early detection of thermal hot-spots and joint exceedance probabilities—making it highly adaptable for diverse climate impact assessments and marine resource planning.*

Andrea Mascaretti (Scuola Internazionale Superiore di Studi Avanzati)

## Probabilistic local-geometric alignment of data representations

**Abstract:** *We introduce a probabilistic method to optimise the similarity between two representations of the same data, such as an image and its description. Our approach learns a diagonal metric that aligns local geometric structure across the two data spaces using triplet distance comparison constraints. The method relies on two key assumptions: (i) meaningful representations preserve neighbourhood relationships, and (ii) in multimodal representations only a small subset of features contributes to this structural alignment. To capture uncertainty, identify the active feature set, and regularize estimation, we adopt a shrinkage prior within a Bayesian metric-learning framework. This yields a flexible posterior over metrics that reveals both the strength and the sources of agreement between the two representations. We validate the method by conducting experiments on both simulated and real data.*

Emanuele Masillo (Sapienza Università di Roma)

## A copula-based framework for doping detection: an application to NADO Italia data

**Abstract:** *Doping control in sport represents a global challenge and is coordinated internationally by the World Anti-Doping Agency (WADA). Athletes are monitored longitudinally through both direct detection of prohibited substances in blood and the Athlete Biological Passport (ABP) profile over time. The current analytical framework for implementing the ABP, known as ADAPTIVE, relies on univariate Bayesian probabilistic models that analyze each biomarker separately. The model is based on a hierarchical structure, assuming Gaussian and log-Gaussian distributions to account for both intra- and inter-individual athletes variability. Although effective in operational settings, this approach is limited by restrictive parametric*

*assumptions and by the implicit assumption of independence among biomarkers, which may not hold in physiological contexts. In particular, the joint behavior of the two biomarkers remains unexplored, and the assumption of independence between them is unlikely to hold in physiological settings. This study aims to extend the ADAPTIVE framework to a multivariate setting that jointly models multiple biomarkers, explicitly accounting for their dependence structure and individual athlete characteristics through copula models, and enhancing the classification phase via multivariate posterior predictive regions. Using a real dataset of the Italian National Anti-Doping Organization (NADO Italia) athletes, the study investigates the marginal distributional properties and characterizes the dependence structure of primary biomarkers in the hematological module. The proposed framework provides a more flexible and robust alternative to the current implementation, addressing key limitations related to parametric assumptions and independence. By incorporating multivariate dependencies, the proposed methodology is shown to enhance the ability to detect abnormal biomarker patterns while preserving the Bayesian hierarchical structure of the original ADAPTIVE model.*

Adelajda Matuka (Università degli Studi di Bologna)

## From disaster to development: the economic lesson of the 2019 earthquake in Albania through a VECM lens

**Abstract:** *This study examines the impact of the 2019 earthquake on Albania's economic growth from 1990 to 2024, using a set of macroeconomic variables and the Vector Error Correction Model (VECM). Time-series analysis indicates that all key macroeconomic variables negatively affect economic growth. The adverse effect of gross capital formation suggests that the Albanian economy is not highly productive, likely due to resource misallocation, limited technological capacity, and poor institutional quality. Government spending on final consumption indicates the economy is currently facing recurrent expenses. The adverse impact of trade reflects the country's ongoing trade deficit, especially after COVID-19 restrictions and the earthquake. It shows weak economic integration and limited export diversification, while inflation undermines investor confidence and reduces purchasing power. Most significantly, the negative influence of central government debt is apparent in high debt-servicing costs, indicating that public borrowing is becoming a constraint on growth. This suggests that debt payments crowd out funds that could be used for efficient investments and other vital sectors for sustainable development. The findings emphasise*

*the urgent need for coordinated policies to enhance public investment performance, strengthen fiscal and macroeconomic resilience, and diversify exports. Implementing budget reforms, improving institutional effectiveness, and better resource allocation are essential to enhancing Albania's economy.*

Marta Nesteruk (Università degli Studi di Milano Bicocca)

## Complementary discriminative signals in high-dimensional proteomics: a comparison of abundance- and network-based models

**Abstract:** *Preoperative diagnosis of thyroid nodules is a challenging clinical problem, as frequent indeterminate biopsy results often lead to unnecessary surgical interventions. In this setting, Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) provides high-dimensional proteomic profiles that may improve diagnostic stratification. However, the analysis of such data critically depends on the statistical framework adopted, as different models may capture distinct aspects of the underlying signal structure. This study investigates the extent to which abundance-based and network-based methods identify complementary discriminative information in MALDI-MSI thyroid data. Three approaches are compared for feature selection and classification: Elastic Net classification, sparse Partial Least Squares Discriminant Analysis (sPLS-DA) and a Copula Graphical Model for heterogeneous mixed data. While the first two methods target differential feature abundance, the graphical model characterizes dependency structures among variables. Applied to a dataset comprising five thyroid diagnostic categories, the methods selected substantially different feature sets, with limited overlap across approaches, suggesting that they capture complementary biological information. To further evaluate these differences, an in silico simulation framework was developed to generate synthetic high-dimensional datasets under controlled scenarios varying in mean shifts, partial correlation structures, marginal distributions, and observation-to-variable ratios. Quadratic Discriminant Analysis was incorporated to enable classification based on graphical model outputs. Preliminary results suggest that abundance-based classifiers perform best when class separation is primarily driven by marginal mean differences, but their performance deteriorates when this condition is not met, whereas network-based approaches retain discriminative power when differences arise from dependency structure alone. Overall, the study highlights the complementary strengths of abundance- and network-based modeling*

*strategies and provides a statistical framework for understanding when each approach is most informative in high-dimensional biomarker discovery.*

Riccardo Pajno (Università degli Studi di Milano Bicocca)

## A double bootstrap procedure for uncertainty quantification in small area estimation using satellite data

**Abstract:** *Satellite data are increasingly employed in statistical research, yet their use in small area estimation remains limited. Indeed, satellite information is usually collected in large frequently updated open source databases. A key challenge is that satellite information is typically provided as regular spatial rasters, whereas official statistics require estimates aggregated according to administrative units. This mismatch gives rise to the Change of Support Problem, requiring the harmonization of data across differing spatial domains. In this framework, our research aims to obtain small area estimates of the agricultural emissions in northern Italy, using a Fay-Herriot model with spatially correlated random effects, relying exclusively on a satellite-derived covariate. Specifically, our target quantity is the agricultural carbon footprint, which is collected by institutions in the RICA database at the sample level. As a covariate, we incorporate ammonia emissions collected from satellite by Copernicus, and its spatial misalignment is addressed using block kriging. However, the alignment procedure introduces additional variability that may propagate to the final estimates. To this end, we propose a double parametric bootstrap algorithm, where the target variable is simulated from the fitted Fay-Herriot model, while the covariate is simulated from a gaussian random field conditionally to the block kriging. Specifically, we present two versions of the algorithm, one to estimate the standard errors of the predictors and the other to construct p-values for hypothesis testing. Finally, the performance of the methods is evaluated through diagnostic analyses, demonstrating their consistency and effectiveness, and some empirical results on agricultural carbon footprint estimation are discussed.*

Edoardo Pandolfo (Sapienza Università di Roma)

## Modeling university outcomes: a hidden Markov model approach

**Abstract:** *In this paper, we propose a hidden Markov model to investigate late graduation and dropout outcomes for university students by accounting for latent activity and inactivity periods in their careers. Data are based on the whole cohort of University La Sapienza students at the faculty of Economics enrolled in the academic year 2017/18. Model estimation is performed via a Bayesian approach. A Gibbs sampler algorithm is used to produce posterior estimates.*

Luca Perego (Politecnico di Milano)

## Early childhood education and women's employment in Italian provinces

**Abstract:** *This project investigates how the relationship between Early Childhood Education and Care (ECEC) services and women's employment varies across Italian provinces. To model this relationship, we employ a dataset with data at the provincial level (NUTS3) from Istat for the year 2022, including census data, labour survey data, urban, socioeconomic indicators and childcare-related variables. We begin by identifying the most relevant predictors of women's employment using a Bayesian regression model with a horseshoe prior. The selected variables include the net migration of graduates, employment growth, coverage of ECEC services, the share of employees in women-dominated sectors (education, care and health), and the share of workers in the manufacturing sector. We add the women's employment rates and additional childcare-related variables to these covariates, and subsequently perform distance-based clustering using Flexclust, a clustering algorithm which considers both geographical distances and similarity in the covariates. Finally, we estimate regression models within each cluster to assess how the association between ECEC-related factors and women's employment differs across distinct contexts. The analysis highlights whether ECEC availability, coverage and participation in ECEC services, and related contextual conditions are linked to women's employment in different ways depending on the socio-economic and spatial profile of provinces. Our preliminary results suggest that the association between these covariates and women's employment differ across the Italian peninsula. The findings contribute to the literature on*

*gender, labour markets, and childcare policy by showing that the relationship between ECEC and women's employment is not uniform across space. More broadly, the project offers a data-driven framework for identifying territorial policy needs and for informing social investment strategies in Italy.*

Daniele Petrone (Sapienza Università di Roma)

## Spatiotemporal analysis for the identification of measles risk areas in Italy: 2013–2024

**Abstract:** *After the reduction in measles circulation during the COVID-19 pandemic, Italy has experienced a resurgence of infections, highlighting that, despite high national and regional first-dose vaccination coverage, areas with accumulated susceptibility may persist. Timely identification of high-risk areas is essential to guide interventions and elimination strategies. This study describes the spatiotemporal distribution of measles in Italy at the municipal level and identifies clusters (groups of contiguous provinces) with higher-than-expected incidence, assessing their evolution over time. The analysis includes possible, probable, and confirmed cases reported to the National Integrated Measles–Rubella Surveillance System, with symptom onset between 01/01/2013–31/12/2019 and 01/01/2023–31/12/2024, with municipality of domicile or, if unavailable, residence. Clusters were identified using a spatiotemporal scan statistic based on Neill et al. (2005). This method, using a Poisson model, detects areas where observed infections significantly exceed expected counts. Only statistically significant clusters were retained (99,999 Monte Carlo simulations, p-value <0.05), and their intensity was assessed using relative risk (RR). Of 16,261 notifications, 14,895 cases (91.6%) were included. The analysis identified 74 significant clusters, involving 440 municipalities and 7,759 cases (52.1%). Clusters were generally small but persistent, with high RR: median 5.23 (IQR 3.46–9.80), indicating incidence several times higher than expected. High-risk areas were concentrated in the North-West and Sicily, while clusters were sporadic in Central and North-Eastern Italy and rare in Southern mainland Italy and Sardinia. In Italy, measles risk is strongly localized in space and time, with many cases concentrated in recurrent municipal hotspots. Integrating spatiotemporal analyses with vaccination coverage data may improve surveillance and support targeted interventions in high-risk municipalities.*

Giuliana Polo (Sapienza Università di Roma)

# The impact of school proximity on educational dropout and early labor market entry among refugee youth in Uganda

**Abstract:** *Understanding the determinants of educational attainment and early labour market entry is a key concern in demographic research, particularly in low-income and displacement settings. Furthermore, education is often associated with health outcomes and use of healthcare services, underscoring the broader importance of access to social services. Within this framework, this study investigates the association between school proximity and education level and work outcomes among young people living in refugee camps in Uganda. Specifically, it examines whether shorter distance to schools is linked to lower education dropout rates and delayed entry into the labour market. The analysis draws on data from the Uganda Resilience Index Measurement and Analysis (RIMA) dataset from Food and Agricultural Organization of the United Nations, covering the period 2017-2021 across refugee settlements and host communities in 11 districts. The dataset provides a comprehensive assessment of refugees' food security, well-being, and resilience, capturing household characteristics, access to resources such as land and services, and livelihood strategies within a policy context aimed at promoting self-reliance and socio-economic integration. Regression models are employed to estimate the relationship between school proximity and outcomes, controlling for key demographic and socioeconomic factors. School proximity is measured as distance to the nearest educational facility, while outcomes include years of schooling, school attendance, and early labour market participation. The study hypothesizes that shorter distances to schools are associated with prolonged educational engagement and reduced early entry into work. By providing empirical evidence on the role of spatial access in shaping youth life courses, this research aim to inform policies that enhance education access and support refugee integration and human capital development.*

Riccardo Racca (Politecnico di Torino)

## Bayesian inference for nonhomogeneous hidden semi-Markov models

**Abstract:** *We propose a novel Bayesian framework for the estimation of nonhomogeneous hidden semi-Markov models, where the underlying state dynamic is governed by time-varying covariates and sojourn times. Our methodology employs a Markov Chain Monte Carlo inference procedure that naturally accommodates the hierarchical structure of latent-state models, avoiding both approximations and computational burden induced by augmented-state formulations, typically required in sequential data methods. The proposal is first validated on simulated data and subsequently applied to a benchmark dataset. Particularly, the case study focuses on a bivariate time series of wind and wave directions recorded by the Ancona buoy in the Adriatic Sea, with wind speed included as a time-varying exogenous covariate.*

Irene Rotondo (Università degli Studi di Torino)

## Spatial factor models for multivariate prediction on river networks

**Abstract:** *Anthropogenic activities result in increased nutrient loadings in rivers, exacerbating eutrophication and causing a deterioration of water quality. Trophic status is commonly assessed through nutrient and dissolved oxygen concentrations, whose strong dependence structure reflects the underlying eutrophication process. Specifically, enhanced nutrient concentrations stimulate algal blooms, leading to higher oxygen consumption and consequent depletion of dissolved oxygen. This study aims to provide a spatial factor modelling framework for multivariate prediction of eutrophication indicators over Piedmont's river network. This approach conjugates latent factor modelling with kriging on stream networks by leveraging factor analysis and spatial covariance structures derived from Spatial Stream Network (SSN) models, accounting for stream distance and flow connection. A spatial latent common factor representing eutrophication is assumed to be underlying four observed variables, namely nitrate nitrogen, ammoniacal nitrogen, total phosphorus, and dissolved oxygen. The model is fitted via the Expectation-Maximization (EM) algorithm, resulting in estimates of factor loadings and latent factor scores at sampled sites. Subsequently, SSN spatial covariance structures are used to model*

*spatial dependence of the common factor, allowing spatial interpolation at unmonitored locations. Finally, predictions for the four eutrophication variables are derived thanks to the fitted model. This approach is a more computationally viable alternative to cokriging on a stream network, that preserves the correlation among the variables through common factor modelling. Moreover, it enables to assess eutrophication across the stream network in its entirety. Key results depict a heterogenous spatial pattern in eutrophication, with higher nutrients concentrations and lower dissolved oxygen percentages in lowland tributaries of the main rivers of the region, in proximity of agricultural and urban centres, reflecting the effects of increasing anthropogenic pressure on water resources.*

Elena Sabbioni (University of Oxford)

## Factor-analysis estimation of phenome dimensionality under pleiotropy

**Abstract:** *Recent genome-wide studies have identified thousands of DNA variants associated with one or more complex human traits. While it is well known that pleiotropy is pervasive in our genome, meaning that the same gene affects different traits, the global structure of this dependence remains poorly understood. While the majority of previous studies have focused on analyzing a small number of traits simultaneously (typically fewer than 10), our goal is to take a more comprehensive approach and estimate the effective dimensionality of the entire phenome with respect to pleiotropy, defined as the minimal number of traits (or latent components) required to predict all others without loss of accuracy. Our approach utilizes factor analysis to decompose the phenotypic variance–covariance matrix and infer the number of latent factors required to capture shared variation across traits. To select this dimensionality, we adopt a predictive criterion: for each trait, we construct out-of-sample predictions using the remaining traits under varying numbers of factors, and evaluate performance via correlation between predicted and observed values. We then aggregate across traits and estimate the effective dimensionality. We validate the method through simulations across a range of trait numbers and true latent dimensions, and assess robustness to missing data via different imputation strategies. We further apply the framework to high-dimensional proteomic data from the UK Biobank (~3000 proteins), providing an initial characterization of phenome-wide pleiotropic structure.*

Muhammad Amir Saeed (Università degli Studi di Milano Bicocca)

# Bayesian optimization for categorical and mixed variables using a multinomial logit surrogate

**Abstract:** *Bayesian optimization (BO) is a widely used framework for optimizing expensive black-box functions. Most BO methods use Gaussian process (GP) models, which work well for continuous data but struggle when the decision variables include categories or a mix of discrete and continuous elements. In particular, GP-based approaches typically require ad hoc numerical encodings of categorical variables that may fail to capture the structure of discrete decision spaces, leading to suboptimal performance in optimization tasks that involve such variables. In this work, we suggest MNL-BO (Multinomial Logit Bayesian Optimization), a framework for Bayesian optimization that uses a multinomial logit (MNL) model instead of the GP surrogate, which is trained using comparisons of preferences between pairs. The resulting model gives a clear and understandable way to represent different categories while managing continuous, discrete, and categorical variables all in one optimization process. The MNL model generates estimates of usefulness and uncertainty, which are used to create acquisition functions that balance exploring new options with making the best use of known ones. The proposed method is tested on three increasingly difficult optimization problems: a simple categorical test, a traveling salesman problem that involves combinations, and a mixed-variable engineering design problem related to choosing materials for pressure vessel optimization. Multi-run tests provide consistent advantages over random search and exhibit stable convergence behavior across diverse random initializations. Besides comparing with basic methods like local search and traditional metaheuristics, we also look at tree-based Bayesian optimization methods that are based on the Sequential Model-based Algorithm Configuration (SMAC) framework. The results show that the MNL-BO method performs well compared to others when given the same amount of resources, and it also offers an easy-to-understand probabilistic model for making decisions in situations with categorical options. These results show that preference-based surrogate modeling is a useful and flexible way to do Bayesian optimization on problems with categorical and mixed variables.*

Silvia Scarpa (Università degli Studi di Modena e Reggio Emilia)

## Who gets tested and why not? Estimating diagnostic coverage gaps from administrative and laboratory records

**Abstract:** *Laboratory data offer rich information on population health, diagnostic service utilization, and biomarker monitoring. This project leverages 15 years of outpatient laboratory records from the Modena Local Health Authority, covering a province of roughly 700,000 residents served by around 390 general practitioners (GPs), to analyse utilization patterns of two common and clinically relevant biomarkers: complete blood count and creatinine (with derived GFR estimates). Testing behaviour is shaped by individual characteristics — age, sex, comorbidities — but also by supply-side factors such as geographic distance to laboratories and GP prescribing propensity. Inference on disease prevalence conditional on biomarker measurements for the target population is therefore susceptible to selection bias and non-ignorable missingness: individuals who never appear in laboratory records are generally not missing at random, but differ systematically from those who do. The distinctive methodological asset of this project is access to population administrative records, including non-tested individuals, which allows direct characterisation of the non-tester group. Building on this, the project develops a rigorous statistical framework to address nonignorable (MNAR) missingness, exploring existing and emerging techniques from the missing data literature. The primary objectives are: to estimate the diagnostic coverage gap within Modena province; to produce bias-corrected prevalence estimates for the conditions under study; and to identify population subgroups with systematically lower propensity to undergo diagnostic testing. The resulting framework is designed to be generalizable and to serve as a replicable template for integrated health data infrastructures grounded in Real World Data and administrative records.*

Andrea Teruzzi (Università degli Studi di Milano Bicocca)

## Robust Bayesian nonparametric clustering across groups

**Abstract:** *In this work we address the challenge of clustering across partially exchangeable groups of data. Traditional models rely on exact atom sharing within group-specific mixing measures, an overly rigid approach that*

*fragments clusters and forces a trade-off between clustering and density estimation. To overcome this, we propose a flexible mixture model utilizing a novel (normalised) hierarchical shot-noise Cox process (HSNCP) prior. Instead of forcing identical mixture components across groups, the HSNCP allows group-specific components to concentrate around shared centers via a kernel. By redefining a cluster as an atom of the underlying "mother process" rather than a single mixture component, our framework achieves robust across-group clustering while maintaining highly accurate within-group density estimation. Furthermore, we establish strong theoretical guarantees for this model, proving posterior consistency for the data distribution and demonstrating that the posterior of the number of clusters is tight as the sample size increases. Alongside these theoretical results, we introduce an efficient conditional MCMC algorithm for posterior inference and validate the model's superior performance on simulated and real-world datasets.*