Pierfrancesco Alaimo di Loro (LUMSA)
a joint work with Dankmar Böhning, Sujit K. Sahu

The standard Poisson Auto-Regression framework considers static coefficients and does not incorporate any spatio-temporal dependence on the parameters governing the process dynamic. However, unobserved space-time variability is a very relevant component when dealing with observations organized on neighboring areal units. We consider a more flexible specification that can adjust for local deviations from the general pattern while borrowing information from adjacent areas. The model, specified in a Bayesian framework, might suffer from computational bottlenecks and convergence issues that can make its Bayesian estimation unfeasible even for modest data sizes. Therefore, we implement it in STAN to jointly update all the parameters and improve mixing, while adopting a novel sparse-matrix representation to attain improved computational performances. The computational advantage and the model performances have been validated through a simulation study.

Martina Amongero (Università di Torino)

Prostatectomized patients are at risk of developing cancer recurrences. During a follow-up period of years, they are monitored for Prostate-Specific Antigen (PSA) resurgence, which is an indicator of tumor progression. The so-called PET-PSMA (Positron Emission Tomography with Prostate-Specific Membrane Antigen) scan is an expensive and complex procedure that can be used to evaluate the effective presence of a tumor.
To optimize the benefit/risk ratio, patients should be referred to a PET-PSMA only when the evidence collected by the PSA is strong enough. A single high value of PSA is not sufficient evidence, and instead, the pattern of PSA evolution is usually monitored over time.
This work addresses the problem of estimating more precisely the probability of a positive PET (POSPET) and the optimal time to recommend a PET exam based on patients' history. To this aim, we build a Hierarchical Bayesian model that describes, jointly, the growth curve of the PSA and the POSPET. With our proposal, we process all past and present information about the patients' PSA measurements and PET-PSMA results and give an informed estimate of the optimal time, improving current practice.

Giulio Biscardi (Università di Firenze)
a joint work with Chiara Marzi, Alessandra Mattei, Aitana Lertuxundi, Michela Baccini

INTRODUCTION
The dose-response curve's shape, which describes air pollution's health effects, has important regulatory implications, and its knowledge is indispensable for health impact assessments. From a casual perspective, when exposure is continuous, one can account for confounding and estimate an average dose-response curve using the generalized propensity score (GPS).

OBJECTIVE
In this work, we propose a method that combines machine learning techniques with a GPS-based matching procedure to study the short-term effects of PM10 on mortality. Analyses are conducted on data already analyzed in the literature, on the city of Milan (2003-2006) and on data from two areas in the Spanish Basque Country (2010-2019): the city of Donostia and an area including three neighboring valleys characterized by industrial-type pollution.
METHOD
Under the Stable Unit Treatment Value Assumption, for each day in the study period, it is possible to define a set of Potential Outcomes (PO) for mortality, i.e., the number of deaths that would be observed on that specific day at different PM10 levels. Since only the PO corresponding to the level actually measured PM10 is observed for each day, the POs have multiple missing data. The missing POs were imputed under a local ignorability assumption by GPS matching. First, we estimated the GPS on each day and for a predefined set of pollutant levels through a boosting model for daily PM10 concentrations given confounders. Then, we imputed the missing POs through a matching procedure between days based on a bivariate distance defined by GPS and PM10 values. Finally, we estimated the mean dose-response curve by constructing a spline on the imputed POs. We calculated the confidence intervals (CI) with a non-parametric bootstrap.
RESULT
The estimated curves for natural, cardiovascular, and respiratory mortality in both datasets show an increasing trend, with a steeper slope for lower PM10 concentrations. Consistent with the literature, the estimates indicate that daily exposures above 20 μg/m3 in Milan caused 3983 deaths [90% CI: 1475-7084], 2382 from cardiovascular causes, and 604 from respiratory causes. Impact estimates for Donostia and the valleys are affected by considerable uncertainty.
CONCLUSIONS
This study deepens our understanding of the causal relationship between PM10 and mortality and explores an alternative method for estimating the dose-response curve that has the advantage of not requiring the specification of a model for the outcome and allows for robust GPS estimates. However, due to its non-parametric nature, this method may be inefficient when the average number of daily events is low.

Arianna Burzacchi (Politecnico di Milano)
a joint work with Bellanger, L., and Le Gall, K., and Stamm, A., and Vantini, S.

In the present world, multiple stakeholders have developed interest in the collection and analysis of mobility data. While new data sources generate large databases of high-resolution mobility data, their application is limited by privacy constraints imposed by governments and data providers to protect the sensitive information of the users. The research presented in this poster introduces a new approach for generating synthetic mobility data from GPS trajectory records, addressing privacy concerns while maintaining data utility. The sequences of GPS data are modeled as multidimensional curves, describing variations in latitude, longitude, and time-to-end within an appropriate mathematical space. Then, framing the analysis in the

Functional Data Analysis (FDA) domain, a new synthetic curve is generated for each of the original by applying the proposed new method for functional data synthesis. Specifically, candidate trajectories are evaluated based on their proximity to a reference function in terms of both amplitude and phase features. Instances with the shortest amplitude and phase distances from the reference function are selected and averaged within the space of Square-Root Velocity Functions (SRVFs). The synthesized curves are then re-mapped back to the original functional space, retaining their interpretability in the context of GPS trajectories. The resulting synthetic dataset is not associated to specific individuals, and hence is fully privacy compliant. Moreover, it shows a high fidelity to the original data: it captures the spatiotemporal structure and offer insights into the underlying mobility patterns.

Ylenia Francesca Buttigliero (Università di Torino)
a joint work with Matteo Ruggiero, Filippo Ascolani

We consider a hidden Markov model for discretely observed binary data, with underlying unobserved dynamic probabilities driven by a one-dimensional Wright-Fisher diffusion. We leverage on recent results on the posterior distribution of the diffusion state given data collected at two time points, to investigate a non-informative prior specification for inference at an intermediate time. Our findings describe explicitly the probability distribution of the data points retention for inference at this intermediate time.

Nicoletta D'Angelo (Università di Palermo)
a joint work with Giada Adelfio

`stopp` is a novel R package specifically designed for the analysis of spatio-temporal point patterns which might have occurred in a subset of the Euclidean space or on some specific linear network, such as roads of a city. It represents the first package providing a comprehensive modelling framework for spatio-temporal Poisson point processes. While many specialized models exist in the scientific literature for analyzing complex spatio-temporal point patterns, we address the lack of general software for comparing simpler alternative models and their goodness of fit. The package's main functionalities include modelling and diagnostics, together with exploratory analysis tools and the simulation of point processes. A particular focus is given to local first-order and second-order characteristics. We aim to welcome many further proposals and extensions from the R community.

Claudio Del Sole (Università Bocconi)
a joint work with Antonio Lijoi, Igor Prünster

Bayesian nonparametric models often assume some kind of homogeneity among the observations, motivated by the exchangeability assumption in de Finetti's representation theorem. In presence of multiple groups of observations, homogeneity within each group is usually modeled through partial exchangeability. Instead, including continuous covariates within a fully nonparametric regression model represents a more challenging task, and existing models in the literature face a trade-off between flexibility in modelling the latent partition structure, its analytical tractability, and its consistency for new observations. This work introduces a novel class of covariate-dependent random probability measures, arising from the normalization of suitable random measures, which depend on covariates through a kernel structure: specifically, the jumps of a common discrete random measure are rescaled via multiplication by a similarity kernel. A noteworthy example arises when the distribution of such random measure is a specific transformation of the distribution of a stable completely random measure. This construction induces a random partition model with dependence on covariates, which is characterized by great flexibility while retaining some analytical tractability, thanks to the introduction of suitable latent variables; moreover, it is inherently consistent for new observations. Both the partition probability function and the posterior distribution of the common random measure are derived in closed form, conditionally on such latent variables. A marginal Gibbs sampler, based on a generalized Pólya urn scheme, is also developed for posterior computation, together with a conditional slice sampling algorithm. Our proposal can be effectively exploited as a clustering or species sampling model which incorporates information available through both discrete and continuous covariates; in addition, it may represent the building block for the construction of nonparametric regression models.

Kristina Dorofeeva (Università di Firenze)
a joint work with Daniele Vignoli, Raffaele Guetto, Elisa Brini

The global fertility rate decline in Western countries is compounding current problems with an aging population, a decrease in the number of indigenous people, and a rise in the flow of migrants from developing countries. Further research is required to ascertain the significance of various fertility-related factors. In this project, our objective is to examine the likelihood of having a first child in Italy, considering the employment status and economic circumstances of both partners in a couple. We use data from the Italian section of the EU-SILC (Statistics on Income and Living Condition), 2004-2023, accounting for its longitudinal nature. The results of a regression analysis will reveal how income levels and the type of employment contract affect the first childbirths among couples. In general, it is anticipated that in the Italian context, a permanent occupation for both partners is associated with a higher fertility rate. However, the existence of alternative job typologies for women may also have an impact.

Paolomaria Fabrizio (Università di Firenze)
a joint work with Patrizio Lodetti, Raffaele Guetto, Daniele Vignoli

Media, intended as aggregators of news and opinions, represent the main source of information regarding economic issues for most individuals. This article builds quantitative and qualitative indices of media economic narratives based on data on the coverage of the economy in Italian newspapers from 2009 to 2020. The data stem from a the LexisNexis database, which includes newspaper articles from a 54 European countries over the past 20 years. The study analyses both the amount of economic coverage and the sentiment conveyed by the news, with the purpose of creating comprehensive indices reflecting the media narrative on the overall state of the economy. These indices will be constructed exploiting a machine learning approach, through the implementation of two Natural Language Processing models with the language model BERT: the first model will discriminate between ”in-topic” (talking about economy) and ”off-topic” news, whereas the second model will assign a sentiment (positive, neutral, or negative) to the ”in-topic” news. In addition, the article examines the reliability and validity of the constructed indices through a comparison with objective indicators of the Italian economic situation. The results show that the economic trends described in the media are not always aligned with actual economic trends. The article also discusses possible applications of these indices in population studies, such as fertility research, and demonstrates the potential of media-based indices as a complementary tools for economic analysis.

Dalila Failli (Università di Firenze)
a joint work with Maria Francesca Marino, Francesca Martella

Biclustering concerns the simultaneous partitioning of units and variables into homogeneous blocks of rows and columns in a data matrix. In detail, this approach is often used to analyze large data matrices in which the relationships between rows and columns can be considered symmetrical. A common area of application concerns the field of genetics, where the biclustering approach can be used to identify groups of genes that are co-expressed under subsets of experimental conditions. A novel model-based biclustering approach for multivariate data is introduced exploiting a finite mixture of generalized latent trait models. The proposed model allows us to cluster units into subsets, called components, via a finite mixture specification. Within each component, subsets of variables, called segments, are identified by a flexible and parsimonious specification of the linear predictor in terms of a row-stochastic vector. The model is designed to handle both qualitative and quantitative variables with (conditional) distribution in the Exponential Family. The integration of a multidimensional, continuous latent trait in the linear predictor allows us to account for the residual dependence between multivariate outcomes from the same unit. In addition, the proposal allows for the inclusion of covariates in the latent layer of the model to determine their impact on component formation. We employ an EM-type algorithm for maximum likelihood estimation of model parameters,

together with Gauss Hermite quadrature in order to approximate multidimensional integrals whose closed-form solutions are not available.

Luisa Ferrari (Università di Bologna)
a joint work with Massimo Ventrucci

The class of Latent Gaussian models (LGM) is widely employed in many fields of application, such as epidemiology, environmental sciences, and ecology. While the Bayesian implementation offers both computational and analytical advantages, the traditional priors used on the variance parameters of these models (i.e., vague i.i.d. Inverse-Gammas) have been found to cause overfitting. Recently, new proposals have emerged in the literature. Most notably, the Penalized Complexity framework has proven quite successful for variance parameters, among others. Despite achieving an increase in performance, the basic assumption of i.i.d. priors still represents the default approach and may hinder a better use of available prior information about the relative contributions of the model terms. To address this, the Hierarchical Decomposition (HD) framework by Fuglstad et al. (2020) has developed a method to design intuitive joint priors on the variance parameters, in a way such that prior information and model structure awareness are exploited. Thanks to this novel approach, users can easily and intuitively include knowledge about the supposed relationships between different effects. However, the original HD method has so far had a limited scope of application. The joint prior can be correctly derived only for a subset of all the effects that can be included in an LGM. We propose a generalization specifically aimed at extending the use of HD priors to models that include generic effects with Intrinsic Gaussian Markov Random Field priors (IGMRFs) and non-stationary ones. One of the main advantages of this extension is the inclusion of P-Spline effects for continuous covariates, as well as linear and higher polynomial effects. In practice, this work opens the use of HD priors to new fields, such as community ecology models, which often require non-linear effects of covariates and potentially non-stationary spatio-temporal correlation.

Vincenzo Gioia (Università di Trieste)
a joint work with Gioia Di Credico, Francesco Pauli

Heat waves (HWs) are of utmost importance in climate studies due to the potential effect of prolonged high temperatures on ecological systems and human health. HWs are generally defined as extended periods during which the temperature stays above a threshold, although the precise definition to adopt in a study also depends on the type of impact, which is primarily of interest (see, e.g., Perkins-Kirkpatrick and Lewis, 2020). In the literature, data-driven quantile-based fixed or varying thresholds approaches are prevalent in classifying HWs periods. Upon adopting a specific definition, examining the frequency, intensity, and duration of HWs is pertinent. However, this is a complex task due to the limited availability of data resulting from the

extreme characteristics of HWs and their varied occurrence across different regions. Motivated by this relevant challenge, we expand the related work by Shaby et al. (2016) by proposing a Bayesian distributional regression Markov-switching model for enhancing the probabilistic classification of the HW regime. In particular, we focus on tail behaviour, and we assume one of the regimes to be Gaussian and the second one may be an extreme value distribution. The proposal is illustrated by analyzing the maximum daily temperatures recorded in four meteorological stations of the Italian region Friuli Venezia Giulia. We obtain similar posterior probability estimates of being in the HW regime, positive trends, and akin seasonality effects across sites. Further, the well-known European summer 2023 heat waves correspond to the periods associated with a high probability of being in the HW regime.

References:
1. Perkins-Kirkpatrick, S.E., and Lewis, S.C. (2020). Increasing trends in regional heatwaves. Nature Communications, 11, 3357
2. Shaby, B.A., Reich, B.J. Cooley, D., and Kaufman, C.G. (2016) A Markov-switching model for heat waves. The Annals of Applied Statistics, 10 (1), 74 - 93.

Cosimo Grazzini (Università di Firenze)
a joint work with Giulia Cosenza, Daniele Castellana, Michela Baccini, Elena Pilli, Giulia Cereda

The identification of human remains involved in large-scale events, such as mass disasters, mass graves from past armed conflicts, or missing persons, poses a significant challenge for forensic experts. Furthermore, in cases where no investigative leads, reference profiles, or database hits are available, the ability to identify remains drastically decreases. Providing additional genetic information, such as BioGeographical Ancestry (BGA), could support investigative activities by guiding searches for victims' relatives and/or enabling identification. An individual's BGA, the biological component of ethnicity, can be inferred from his/her DNA, particularly using single-nucleotide polymorphism (SNP) markers. While most major works aim to target BGA at continental level, only a few attempts have been made to go below this macro level, with these attempts showing lesser accuracy in discriminating BGA. Our aim is to classify individuals at a finer level through an innovative panel of SNPs, coupled with supervised Machine learning (ML) methods. Starting with a panel of 3233 SNPs from 3591 individuals with known BGA at a high level of detail, we apply several supervised classification methods, chosen according to their advantages in high dimensional qualitative data, such as Naive Bayes, Random Forest, XGboost and Support Vector Classifier. The best values of the hyperparameters for each method were chosen through nested cross-validation. The results show that ML methods can be useful in classifying individuals by BGA, having good discriminating capacity and performance that are all the better the more sophisticated the method is, allowing to reach an intermediate classification of individuals in macro areas at a level that is intermediate between nations and continents. In an effort to improve the accuracy of inter-continental classification, additional informative SNPs for BGA classification will be

evaluated and selected to provide practitioners with a kit of genetic markers that can be used in real forensic cases.

Francesca Labanca (Università di Firenze)
a joint work with Anna Gottard

Toroidal data, arising from observations on a p-vector of circular variables, are characterized by their support on the p-torus. In various applications, the multivariate wrapped normal distribution has emerged as a popular choice for modeling data on a hypertorus. However, the probability density function of this distribution involves an infinite series, as it is derived by wrapping around the p-torus a p-dimensional normal distribution. To address this problem, we propose an Indirect Inference approach that leverages a Normal auxiliary model defined on the real space, sharing the same dimension as the parameter space of the target model. The Indirect Inference estimators exhibit properties of consistency and asymptotically Gaussian distribution. Notably, this approach circumvents the analytical and computational complexities typically associated with standard procedures, as it is constructed from the multivariate normal maximum likelihood estimators, selected for their analytical and numerical simplicity. To assess the finite sample behavior, we conduct a Monte Carlo numerical study exploring various scenarios, including known mean but unknown variance-covariance matrix, and unknown mean and variance-covariance matrix with and without a plug-in circular mean. In the simulation results the proposed method yields efficient and accurate estimates for the parameters of the multivariate wrapped normal distribution, particularly for small to medium variances.

Alessio Lachi (Università di Firenze)
a joint work with Giulia Carreras, Cecilia Viscardi, Andrea Saltelli, Michela Baccini

Cigarette smoking has still a significant impact on population morbidity and mortality. This study introduces an innovative approach to enhance the reliability of inferences and forecasts produced by the Smoking Habits Compartmental (SHC) model, a compartmental model for simulating smoking dynamics. While compartmental models like SHC offer a valuable framework for understanding complex systems and projecting public health dynamics, they suffer from limitations related to stringent model assumptions and parameter identifiability. The proposed methodology aims to robustify inference, forecasting, and the assessment of tobacco control policies by systematically incorporating uncertainties present in model definition and data. This involves a strategy of error propagation analysis and a variance-based Global Sensitivity Analysis (GSA). The GSA provides insights into how the variance of model outputs can be attributed to uncertainties in model inputs, utilizing sensitivity indexes. The study underscores the importance of considering all sources of uncertainty in the modeling process, especially when crafting forecasts under hypothetical scenarios for guiding public health policies. The proposed robustification procedure, incorporating GSA, contributes to a more

comprehensive assessment of variability, aiding in the identification of influential data subsets and variables. The paper confirms previous findings regarding model parameters but also highlights the dependence of inference on the calibration window used, signaling potential issues with assumptions about the constancy of transitions between compartments over time. The research concludes with an evaluation of alternative tobacco control policies for Tuscany, emphasizing the substantial impact of uncertainty on policy effectiveness. Overall, the study advocates for the integration of uncertainty analysis and GSA in modeling processes, providing a more nuanced understanding of the robustness of public health projections and guiding the development of intricate models.

Maria Francesca Morabito (Università di Firenze)
a joint work with Serena Verbena, Valentina Tocchioni, Benedetta Emanuela Palladino

This study aims to explore the dynamics of ethnic victimization at school. Previous research focused on the harmful consequences of ethnic bullying while devoting little attention to the role of social capital resources. Rare suggestions indicated the protective function of teachers' support and school climate for bullying victims. We seek to shed light on the role of individual- and school-level social capital factors in shaping the behavior of ethnically bullied students. We employ a large dataset of Italian high school students collected in 2021, 2022, and 2023. We select participants who reported being bullied based on their ethnicity (N=55,289). Through multilevel latent class analysis, we evaluate the impact of social capital factors on the behavior of ethnically bullied students by obtaining specific profiles of bullying victims based on their behaviors. We employ the five subscales of the Strengths and Difficulties Questionnaire (SDQ) referring to emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behavior as the response variables. The key explanatory variables on social capital resources at both individual and school levels refer to the perception of students about school climate, teachers' responses to bullying, and school context.
Preliminary findings suggest that most social capital factors mitigate the severity of emotional symptoms, conduct and relationship problems of ethnically bullied students, as well as their hyperactivity. Instead, those factors that promote the expression of victims' emotions and enhance their consciousness tend to amplify these behaviors. Finally, social capital factors seem to enhance prosocial behaviors among ethnically bullied students. Having obtained several classes of ethnically bullied students based on their behaviors, we proceed to explore the effects of social capital resources within each latent class. Our findings can lay the groundwork for educational policies and interventions aimed at promoting students' adaptation and well-being through the maximization of social capital.

Greta Panunzi (Sapienza Università di Roma)
a joint work with Silvia Polettini e Serena Arima

Violence against women remains a prevalent and enduring issue. However, a significant gap exists in obtaining comprehensive and up-to-date data on this widespread yet often underreported problem. Specialized surveys are the most reliable methods of estimating the prevalence and characteristics of gender-based violence, but the lack of recent survey data presents a challenge. In cases like this, police registers can be a source of information, but they suffer from significant underreporting. To address the lack of data, we suggest a Bayesian model that incorporates administrative data and socio- demographic indicators. Our model aims to improve the accuracy of prevalence estimates for gender- based violence in Italy. To model this data, we propose a Poisson regression that accounts for under- reporting using the Pogit model. In this way, we consider the reporting process and address potential biases in administrative data sources. Our approach represents a methodological advancement in the study of gender-based violence, providing insights into the complex dynamics of this phenomenon in Italy. Combining data from multiple sources and using advanced statistical techniques, we aim to provide policymakers and stakeholders with more accurate and actionable information. This will facilitate the development of targeted and effective interventions and policies to combat violence against women.

Kevin Taglialatela Scafati (Università di Firenze)
a joint work with Fabrizia Mealli, Paolo Brunori

During the sixties, a season of historical reforms drastically changed the Italian education system. The largest and most debated of them was the introduction of the single middle school (SMU), a three-year post-elementary program that was intended to be attended by all the Italian students and substituted a fragmented reality composed by a variety of vocational programs and a highly selective academic track. The goal was that, regardless of their origins, students would have the same opportunity to obtain professional success and acquire the human capital necessary to economically thrive later in life. This idea aligns with the definition of equality of opportunity given in the field of normative economics: a system of equal opportunities is realized when everyone has the same expectation regarding the well-being they can achieve in their lives, regardless of factors such as geographical or social origins. SMU approval gave rise to an articulated parliamentary debate, from which it emerged the concern that this intervention, without additional forms of subsidies to families, would not suffice to reach the proposed egalitarian aim. This work provides data-based evidence on the validity of these concerns, evaluating if and to what extent the SMU decree caused a reduction in the inequality of opportunity among Italian pupils. Such empirical exercise involves identifying a convincing well-being dimension of interest, obtaining data on the joint distribution of this individual-level well-being measure and on the key circumstances possible sources of inequality. It also involves detecting two groups of students that overlap in terms of potential confounders, but differ since

one of them was not exposed to the SMU. An unconfounded comparison of these groups based on robust measures of well-being predictability given observed circumstances allow us to investigate the existence of the effect that the reform intended to bring.

Emma Torrini (Università di Firenze)
a joint work with Fabrizia Mealli, Christian Stock

Principal stratification has been proposed as one approach to deal with intercurrent events in the ICH E9 (R1) addendum. However, so far, limited experience exists with principal stratification in the context of pharmaceutical clinical trials, especially concerning situations where interest is in a time-to-event endpoint. Our work is motivated by questions arising in the development of biological drugs in rare immunological diseases, where patients under the investigational treatment may have an immune response that leads to the development of antidrug antibodies (ADAs). If ADAs occur, they can have an impact on the efficacy of the drug. It is thus commonly an important secondary objective in phase II/III trials to explore the efficacy of the drug in patients who develop ADAs. Specifically, it is of interest to estimate the effect of the drug in the principal stratum of patients who would develop ADAs if given treatment. We investigate a basic Bayesian principal stratification approach for exponentially distributed time-to-event endpoints that applies to the described setting. It is based on a latent mixture likelihood and allows consideration of predictors of ADA development. We focus on the estimation of hazard ratios and restricted mean survival times using MCMC methods. Our simulation results show that the methodology can validly estimate the parameters of interest. A publicly available R package is presented that facilitates model fitting and simulation to determine operating characteristics in a given setting. Overall, this work lays theoretical and practical foundations for desirable extensions, including the use of more flexible time-to-event distributions.
