XVIII Conferenza Italiana degli Utenti di *Stata*

Milano, 25-26 Settembre 2025

PROGRAMMA | 25 SETTEMBRE

8:30-9:00 Registrazione dei partecipanti

9:00-10.25 SESSIONE I - EXPLOITING THE POTENTIAL OF *STATA* 19, I

Linking frames in *Stata* • Jeff Pitblado, Executive Director, Statistical Software, *StataCorp*

This presentation gives an overview of data **frames** in *Stata*. I demonstrate the basics of working with multiple datasets in *Stata*. I cover most of the -frames- suite of commands, touching on frame creation and management, linking frames, copying variables from linked frames, alias variables, and working with a set of frames.

The new **cate** command: an overview • Giovanni Cerulli, Institute for Research on Sustainable Economic Growth, Italian National Research Council (IRCrES - CNR)

This presentation offers a concise overview of the **cate** command, a new tool introduced in *Stata 19* for estimating Conditional Average Treatment Effects (CATEs). CATEs quantify how the impact of a treatment varies across individuals or subgroups defined by observed characteristics, thus enabling a more nuanced understanding of treatment-effect heterogeneity and supporting the design of targeted policy interventions.

The **cate** command delivers three types of estimates:

- **IATEs** (Individualized Average Treatment Effects), which are specific to each observation,
- **GATEs** (Group Average Treatment Effects), computed for user-defined groups,
- **GATESs** (Sorted Group Average Treatment Effects), which aggregate treatment effects over quantiles of the IATE distribution.

To compute these estimates, **cate** jointly models the outcome and treatment assignment processes using flexible machine learning or parametric methods. Users can choose among lasso, random forests, or traditional regression, with cross-fitting to improve robustness. Estimation of the CATEs is then performed using either a partialing-out (PO) estimator or an augmented inverse probability weighting (AIPW) approach, both available with linear and nonparametric options.

10:25-10:45 Pausa caffè

10:45-12.15 SESSIONE II - COMMUNITY CONTRIBUTED, I

xtbreak: Testing and Estimating Structural Breaks in Time Series and Panel Data in *Stata* • Jan Ditzen, Libera Università di Bolzano

Identifying structural change is a crucial step in analysis of time series and panel data. The longer the time span, the higher the likelihood that the model parameters have changed as a result of major disruptive events, such as the 2007–2008 financial crisis and the 2020 COVID–19 outbreak. Detecting the existence of breaks, and dating them is therefore necessary, not only for estimation purposes but also for understanding drivers of change and their effect on relationships. This talk introduces a new community contributed command called **xtbreak**, which provides researchers with a complete toolbox for analysing multiple structural breaks in time series and panel data. **xtbreak** can detect the existence of breaks, determine their number and location, and provide break date confidence intervals. A special emphasis of the talk will be put on the Python integration to gain speed advantages.

Variance's Components in Panel Data • Maria Elena Bontempi, *Alma Mater Studiorum* - Università degli Studi di Bologna

A preliminary and crucial step in any empirical research on panel data, whether longitudinal, time-series-cross-section or multilevel, is to study the nature and relevance of the components that influence the variability of the variables, in particular the dependent variable. Each panel dataset can be considered as a set of grouped data, whether these are temporal observations nested within individuals or individuals nested within groups and supergroups. The fundamental steps for guiding the modelling strategies to be adopted are: break down the total variability into variances between and within clusters, also in terms of percentage shares; assessing whether there are relevant common factors within clusters and, in the case of temporal observations, whether these are stationary or not; comparing the relevance and significance of group and individual effects depending on whether they are considered fixed or random.

fffuroot: Implementing in *Stata* unit-root and stationarity tests with smooth breaks approximated by flexible Fourier forms • Giovanni Bruno, Università Commerciale L. Bocconi, Milano

This work describes the *Stata* implementation of unit-root and stationarity tests with flexible Fourier forms as in Enders and Lee (2012a), (2012b) and Becker, Enders and Lee (2006).

12:15-13:00 SESSIONE III - *STATA* TIPS AND TRICKS

xtplot2 • Jan Ditzen, Libera Università di Bolzano

The command **xtplot2** investigates the structure of panel dataset with respect to unbalancedness and values using heatplots. It allows the researcher a quick and efficient way to gain insights into the structure

splitting • Automating Episode Splitting: Introducing the **splitting** command for *Stata* • Davide Bussi, Università di Milano-Bicocca

Event History Analysis (also known as survival analysis) is a well-established analytical tool in the social sciences and research more broadly, and it is particularly useful when researchers aim to estimate the effect of time-varying variables. Survival analysis is well-supported in *Stata* via numerous built-in commands. In particular, **stsplit** facilitates breaking the time axis into episodes in order to include time-varying covariates in the analysis. While **stsplit** is straightforward to use when the time axis must be split at the point a change occurs in a dichotomous variable, the procedure becomes less intuitive when dealing with polytomous variables.

xtgetpca • Jan Ditzen, Libera Università di Bolzano

The command **xtgetpca** extracting principal components in panel data is common, however no *Stata* solution exists. **xtgetpca** fills this gap. It allows for different types of standardization, removal of fixed effects and unbalanced panels.

13:00-14:00 Pranzo

14:00-15:20 SESSIONE IV - EXPLOITING THE POTENTIAL OF *STATA* 19, II

Meta-analysis in *Stata* • Gabriela Ortiz, Senior Applied Econometrician, *StataCorp*

Many studies attempt to answer similar research questions. For instance, you may have results from studies asking, What is the association between unemployment and mental health? Or you may have results from studies asking, How does motherhood affect women's wages? The results from different studies may be inconclusive or conflicting. Meta-analysis is a statistical technique for combining the results from several similar studies. It allows us to explore the variation across studies and, when appropriate, provide a single estimate for the effect size of interest. In this presentation, I show how to use the **meta** suite of commands to perform meta-analysis in *Stata*.

Consensus Clustering in *Stata* • Carlo Drago, Università Niccolò Cusano

This work considers consensus clustering in *Stata*, combining bootstrapped k-means with hierarchical clustering based on a co-association matrix. The method addresses the possible inherent instability of partitioning-based clustering by aggregating results from multiple bootstrap samples, improving robustness and reproducibility. In this respect, at each iteration, k-means clustering is applied, and the results are collected in a large-scale cluster assignment matrix. A consensus matrix is then created to measure the co-occurrence of observations within the same cluster across all iterations. This matrix is transformed into a dissimilarity structure and in this way subjected to hierarchical clustering in order to obtain a final, stable partition.

This framework shows how consensus clustering can be performed robustly and efficiently in *Stata*. It uses a combination of *Stata* routines, bootstrap sampling, and optimized Mata routines to compute the co-association matrix, ensuring computational efficiency. The approach is broadly applicable to clustering tasks in the social sciences, economics, epidemiology, and other fields where cluster stability is critical.

15:20-15:35 Pausa caffè

15:35-16:35 SESSIONE V - COMMUNITY CONTRIBUTED, II

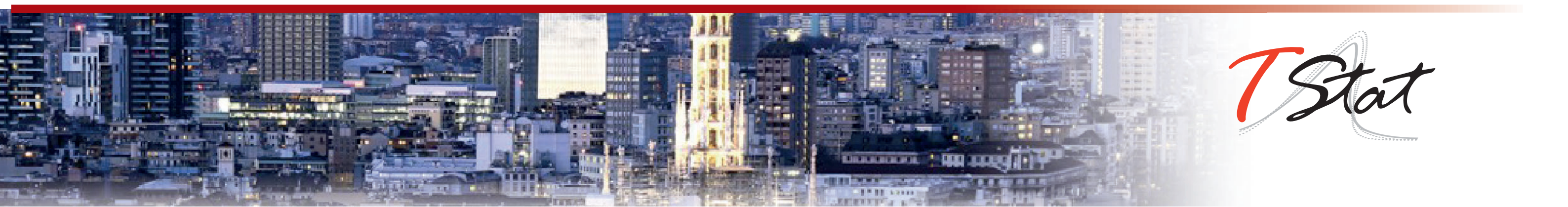
outdetect: Outlier detection for inequality and poverty analysis • Giulia Mancini, Università degli Studi di Sassari

Extreme values are common in survey data and represent a recurring threat to the reliability of both poverty and inequality estimates. The adoption of a consistent criterion for outlier detection is useful in many practical applications, particularly when international and intertemporal comparisons are involved. In this article, we discuss a simple univariate detection procedure to flag outliers. We present **outdetect**, a command that implements the procedure and provides useful diagnostic tools. The output of **outdetect** compares statistics obtained before and after the exclusion of outliers, with a focus on inequality and poverty measures. Finally, we carry out an extensive sensitivity exercise where the same outlier detection method is applied consistently to per capita expenditure across more than 30 household budget surveys. The results are clear and provide a sense of the influence of extreme values on poverty and inequality estimates.

rdlasso: A *Stata* Command for High-Dimensional Regression Discontinuity Designs • Marianna Nitti, Università La Sapienza Roma

The command **rdlasso** implements Regression Discontinuity Designs (RDD) with high-dimensional covariates in *Stata*. The procedure is based on the methodology developed by Kreiss and Rothe (2023), and extends it to both sharp and fuzzy designs. Covariate selection is performed through a Lasso-based local estimation, ensuring valid inference under approximate sparsity.

The command is built using *Stata*'s Python integration via



the sfi module and automates all steps of the estimation process—from covariate selection to bandwidth choice and bias-corrected treatment effect estimation. The syntax allows for flexible user control while remaining fully embedded in the *Stata* environment.

rdlasso enables *Stata* users to apply machine learning techniques for causal inference without requiring programming in external platforms such as R or Python. The command generates output variables that can be used for further post-estimation analysis within the same session. An option automatically distinguishes between sharp and fuzzy designs, making the tool both user-friendly and methodologically complete. The implementation is illustrated through a step-by-step example and an empirical application. The command contributes to the growing set of tools for modern causal analysis in *Stata*, particularly in high-dimensional settings.

16:35-17:40 SESSIONE VI - EXPLOITING THE POTENTIAL OF *STATA* 19, II

Automated Data Extraction from Unstructured Text Using LLMs: A Scalable Workflow for *Stata* Users • Loreta Isaraj, Institute for Research on Sustainable Economic Growth, Italian National Research Council (IRCrES - CNR)

In several data-rich domains such as finance, medicine, law, and scientific publishing, most of the valuable information is embedded in unstructured textual formats, from clinical notes and legal briefs to financial statements and research papers. These sources are rarely available in structured formats suitable for immediate quantitative analysis. This presentation introduces a scalable and fully integrated workflow that employs Large Language Models (LLMs), specifically ChatGPT 4.0 via API, in conjunction with Python and *Stata* to extract structured variables from unstructured documents and make them ready for further statistical processing in *Stata*.

As a representative use case, I demonstrate the extraction of information from a SOAP clinical note, treated as a typical example of unstructured medical documentation. The process begins with a single PDF and extends to an automated pipeline capable of batch-processing multiple documents, highlighting the scalability of this approach. The workflow involves PDF parsing and text pre-processing using Python, followed by prompt engineering designed to optimize the performance of the LLM. In particular, the temperature parameter is tuned to a low value (e.g., 0.0–0.3) to promote deterministic and concise extraction, minimizing variation across similar documents and ensuring consistency in output structure.

Once the LLM returns structured data, typically in JSON or CSV format, it is seamlessly imported into *Stata* using custom .do scripts that handle parsing (insheet), transformation (split, reshape), and data cleaning. The final dataset is used for exploratory or inferential analysis, with visualization and summary statistics executed entirely within *Stata*. The presentation also addresses critical considerations including the computational cost of using commercial LLM APIs (token-based billing), privacy and compliance risks when processing sensitive data (such as patient records), and the potential for bias or hallucination inherent to generative models. To assess the reliability of the extraction process, I report evaluation metrics such as cosine similarity (for text alignment and summarization accuracy) and F1-score (for evaluating named entity and numerical field extraction).

By bridging the capabilities of LLMs with *Stata*'s powerful analysis tools, this workflow equips researchers and analysts with an accessible method to unlock structured insights from complex unstructured sources, extending the reach of empirical research into previously inaccessible text-heavy datasets.

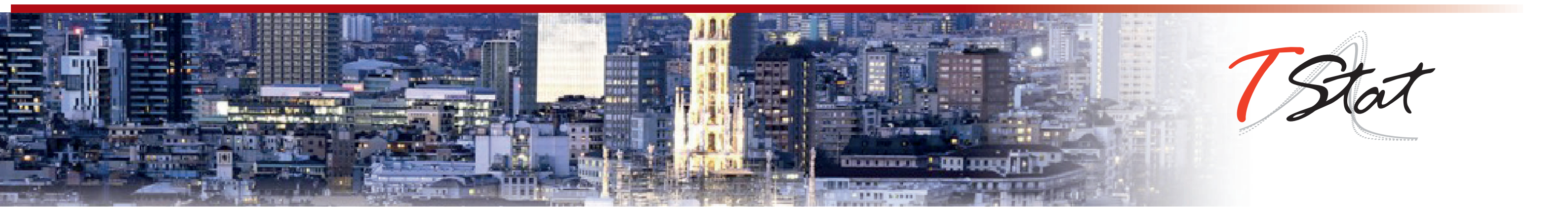
Text Mining and Hierarchical Clustering in *Stata*: An Applied Approach for Real-Time Policy Monitoring. Forecasting and Literature Mapping • Carlo Drago, Università Niccolò Cusano

This work show an applied framework for text mining and clustering in the *Stata* environment and provides practical tools for policy-relevant research in economics and health economics. With the growing amount of unstructured textual data — from financial news and analyst reports to scientific publications — there is an increasing demand for scalable methods to classify and interpret such information for evidence-based policy and forecasting.

A first relevant concept is the *Stata* capacity to be integrated with Python with aim to implement hierarchical clustering from scratch using TF-IDF vectorization and cosine distance. This technique is specifically applied to economic text sources — such as headlines or institutional communications — with the aim to segment documents into a fixed or silhouette-optimized number of clusters. This approach allows researchers to identify patterns on data, uncover latent themes, and organize information for macroeconomic forecasting, sentiment analysis, or real-time policy monitoring.

In the second part, we focus on literature mapping in health economics. Using a curated corpus of article titles related to telemedicine and diabetes, we apply a native *Stata* pipeline based on text normalization and clustering to identify thematic areas within the literature. The approach promotes organized reviews in health technology assessment and policy evaluation and makes evidence synthesis more accessible.

By combining native *Stata* capabilities with Python-enhanced workflows, we provide applied researchers with an accessible and policy-relevant toolkit for unsupervised text classification in multiple domains.



17:40 -18:00 OPEN PANEL DISCUSSION WITH *STATA* DEVELOPERS • JEFF PITBLADO AND GABRIELA ORTIZ, *STATACORP*

La sessione “*Open panel discussion with Stata Developers*” offre ai partecipanti la possibilità di interagire direttamente con la *StataCorp*: sarà possibile evidenziare problemi o limitazioni del software nonché suggerire eventuali miglioramenti o comandi che potrebbero essere inclusi in *Stata.*

20.00 Cena Sociale *(opzionale)*

COMITATO SCIENTIFICO

Una-Louise BELL, TStat - TStat Training • Rino BELLOCCO, Università degli Studi di Milano-Bicocca • Giovanni CAPELLI, Istituto Superiore di Sanità • Giovanni CERULLI, IRCRES-CNR • Jan DITZEN | Libera Università di Bolzano • Maurizio PISATI, Università degli Studi di Milano-Bicocca

SEGRETERIA ORGANIZZATIVA

Monica GIANNI • Via Rettangolo, 12-14 • 67039 Sulmona (AQ)
T. +39 0864 210101 • www.tstat.it | www.tstattraining.eu
formazione@tstat.it | training@tstat.eu

TStat S.r.l. | Distributore esclusivo di *Stata* per: Albania, Bosnia e Erzegovina, Croazia, Grecia, Italia, Kosovo, Macedonia del Nord, Malta, Montenegro, Serbia, Slovenia, Slovacchia.