Prof.ssa Maria Ferrante

Comparison of Bootstrap Confidence Interval Methods Small Area Estimation

The importance of developing reliable sub-population indicator estimates has increased significantly. Surveys usually provide information for broader areas, such as countries or administrative divisions. However, there is a growing need for estimates at a finer level of detail. Due to financial constraints that prevent expanding sample sizes, alternative methods, known as Small Area Estimation (SAE), are used. SAE methods encompass various statistical techniques to obtain reliable estimates for small sub-populations or geographic regions. These techniques are necessary when the variability of the direct estimator, like the Horvitz-Thompson estimator, is too large to produce reliable results (for a comprehensive review, see Rao and Molina, 2015 and Tzavidis et al., 2018).

The SAE literature has produced a large number of papers describing different techniques to

estimate the MSE (see among others Field and Welsh, 2007; Gonzalez-Manteiga et al., 2008; Liu

et al., 2022). The most common method is a parametric bootstrap. Even though the amount of

literature about bootstrap in SAE is considerable, less attention has been given to the estimation of bootstrap confidence intervals. There are various ways to compute bootstrap confidence intervals (Efron and Tibshirani, 1994; Chernick, 2011; and Jung et al., 2019), but only one has been usually applied in SAE (Liu et al., 2022).

In this project, we propose to develop simulation studies in which various methods of bootstrap

confidence intervals are compared to define the best possible choice concerning SAE models.

A first approach will be to test the three methods reported in Jung et al. (2019) for parametric

bootstrap on the baseline models in SAE (Rao and Molina, 2015, Ch. 6 and 7). The research
could then be generalized to more complex models (i.e., non-linear models) and also to the nonparametric bootstrap.

References

- Chernick, M. R. (2011). Bootstrap methods: A guide for practitioners and researchers. JohnWiley & Sons.
- Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. Chapman; Hall/CRC.
- Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3), 369–390.
- Gonzalez-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D., & Santamaria, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78(5), 443–462.
- Jung, K., Lee, J., Gupta, V., & Cho, G. (2019). Comparison of bootstrap confidence interval methods for GSCA using a monte carlo simulation. *Frontiers in psychology*, 10, 2215.
- Liu, Y., Liu, X., Pan, Y., Jiang, J., & Xiao, P. (2022). An empirical comparison of various MSPE estimators and associated prediction intervals for small area means. *Journal of Statistical Computation and Simulation*, 1–27.
- Rao, J., & Molina, I. (2015). Small area estimation. John Wiley & Sons, Inc.
- Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal*
  Statistical Society: Series A (Statistics in Society), 181(4), 927–979.

Prof.ssa Maria Ferrante

Construct a synthetic population for unit-level small area estimation models

Small Area Estimation (SAE) methods encompass a variety of techniques designed to obtain
reliable estimates for small sub-populations when the sample size is too low to yield accurate
results using a classical Horvitz-Thompson estimator (for a comprehensive review, see Rao and
Molina, 2015 and Tzavidis et al., 2018). SAE models leverage strengths from neighboring areas
and auxiliary information. When the auxiliary information is at the individual level, these are
referred to as unit-level SAE models.

One of the main limitations in the development of SAE methods, especially at the unit level,
is the access to individual data necessary for simulation studies, which are often used to test the
efficacy of a method. These data are frequently subject to privacy restrictions and are often replaced with synthetic populations (Ferrante and Pacei, 2017). Synthetic populations are simulated datasets from which it is possible to extract samples of various sizes, and for which all parameters are known. In other words, one possible solution to overcome the problem of the confidentiality constraints is synthetic data, which mimics the original observed data and preserves the relationships between variables without containing any disclosive records. Techniques to produce synthetic populations are well summarized in Taylor et al. (2016), Nowok et al. (2016) and Templ et al. (2017).

In this project, we propose to create a synthetic population to be published and made directly

accessible to the international statistical community. This population will be based on real data
from European surveys and will be generated using appropriate statistical approaches to create a
realistic representation that closely mirrors the original population.

References

- Ferrante, M. R., & Pacei, S. (2017). Small domain estimation of business statistics by using multivariate skew normal models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4), 1057–1088.
- Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*, 74, 1–26.
- Rao, J., & Molina, I. (2015). Small area estimation. John Wiley & Sons, Inc.
- Taylor, J., Moon, G., & Twigg, L. (2016). Using geocoded survey data to improve the accuracy of multilevel small area synthetic estimates. *Social Science Research*, 56, 108–116.
- Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of synthetic complex data: The R package simpop. *Journal of Statistical Software*, 79(10), 1–38.
- Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4), 927–979.

Prof Matteo Farnè

Multimodal fake news detection models: an empirical comparison

Fake news is not just about text; it also uses images, videos, and social signals to deceive people (Yan et al., 2024). Text-based features are not sufficient, as they do not explain a key aspect of deception, such as when an image and text do not match (Rout et al. 2025). This has changed the way we think about and solve the problem of fake news. The task is no longer simple text matching but an anomaly detection problem, where the goal is to identify when a piece of information deviates from the expected distribution of truthful content (Rout et al., 2025; Xue et al., 2021). Some models that can understand different types of information, such as words and images, use complex and diverse parts to do their job (Yan et al., 2024). For example, the MCNN model has different parts for text and image features and a special part that compares them directly to see if they match (Xue et al., 2021). This is important for detecting fake news. Another example is the BTCM model, which uses two parts that transform text and image information, respectively (Yan et al., 2024). It also has a new use of a completely unrelated or tampered image to manipulate perceptions. Other works propose a model that not only fuses textual and visual features but also incorporates user information, such as historical posts and preferences, to improve the recognition of fake news (Hao et al., 2025) or a model that aligns and integrates different types of data (text and images) by using contrastive learning and optimal transport to effectively merge features from various modalities (Shen et al., 2024).

The present RAship project has the goal to make an extensive literature review of multimodal fake news detection models, to compare their performances statistically on proper datasets, and to identify gaps in this literature to propose enhanced solutions.

Bibliography

Yan, Y., Fu, H., & Wu, F. (2024). Multimodal social media fake news detection based on 1dccnet attention mechanism. Electronics, 13(18), 3700.

Rout, J., Mishra, M., & Saikia, M. J. (2025). Towards Reliable Fake News Detection: Enhanced Attention-Based Transformer Model. Journal of Cybersecurity and Privacy, 5(3), 43.

Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management, 58(5), 102610.

Hao, X., Xu, W., Huang, X., Sheng, Z., & Yan, H. (2025). MFUIE: A Fake News Detection Model Based on Multimodal Features and User Information Enhancement. EAI Endorsed Transactions on Scalable Information Systems, 12(1).

Shen, X., Huang, M., Hu, Z., Cai, S., & Zhou, T. (2024). Multimodal Fake News Detection with Contrastive Learning and Optimal Transport. Frontiers in Computer Science, 6, 1473457

Prof. Daniele Ritelli

Methods of solving differential equations with Lie symmetries using computer algebra

A very practical approach is followed on the subject, given that the existence of Lie symmetries, which allow the simplification of a given differential equation, consists of solving accessory differential equations, which are treated using computer algebra.

Computer algebra in the symbolic treatment of nonlinear differential equations using special functions

After introducing the student to the use of software for symbolic calculus, the internship aims at using it for the treatment of differential equations describing the behaviour of nonlinear oscillators. Tutorials illustrating the theoretical approach underlying the treatment of the models will be an integral part of the training. The candidate must be willing to acquire the mathematical tools necessary for the study of the models.

Prof. Michele Scagliarini

Methods for Monitoring Time Between Events and Amplitude Data

While many control charts have been developed for monitoring the time interval (T) between the occurrences of an event, many other charts are employed to examine the magnitude (X) of the event (E). These two types of control charts have usually been investigated and applied separately.

Time Between Events and Amplitude (TBEA) control charts are a combined scheme for monitoring the time interval T of an event E as well as its amplitude X.

The aim of this project is to study the implementation of such monitoring algorithms in the R environment. The developed methodology will be applied on both simulated and real data.

The internship will be divided into three phases:

- first phase dedicated to study Shewhart Time-Between-Events-and-Amplitude Control Charts and their implementation in R;
- second phase dedicated to study the effect of the correlation between T and X;
- third phase aimed to implement a non-parametric EWMA control chart for Monitoring TBEA.

References

D. Rahali, P. Castagliola, H. Taleb, and M.B.C. Khoo. Evaluation of Shewhart Time-Between-Events-and-Amplitude Control Charts for Several Distributions. Quality Engineering, 31(2):240-254, 2019. doi: 10.1080/08982112.2018.1479036.

D. Rahali, P. Castagliola, H. Taleb, and M.B.C. Khoo. Evaluation of Shewhart Time-Between-Events-and-Amplitude Control Charts for Correlated Data. Quality and Reliability Engineering International, 37(1):219-241, 2021. doi: 10.1002/qre.2731.

S. Wu, P. Castagliola, and G. Celano. A Distribution-Free EWMA Control Chart for Monitoring Time-Between-Events-and-Amplitude Data. Journal of Applied Statistics, 48(3):434-454, 2021. doi:10.1080/02664763.2020.1729347.

Z. Wu, J. Jiao, and H. Zhen. A Control Scheme for Monitoring the Frequency and Magnitude of an Event. International Journal of Production Research, 47(11):2887-2902, 2009.

Prof.ssa Silvia Pacei

Misura della resilienza

Le crisi globali del 2008 e 2020 hanno determinato una situazione di incertezza attorno ai sistemi economici e finanziari. La ripresa poi, com'è noto, è stata particolarmente rallentata in Europa, rispetto agli USA, benché con una certa variabilità nella velocità della ripresa tra i paesi europei. I dati delle indagini sulle famiglie evidenziano la loro preoccupazione per il periodo di incertezza economica che stanno vivendo e per la loro capacità di recuperare eventuali perdite. Tali preoccupazioni influiscono su diversi comportamenti umani.

Ad oggi pochi autori si sono proposti di misurare la resilienza (Asheim et al. 2020; Cissé e Barrett, 2018). L'obiettivo di questo progetto è proporre una definizione di resilienza e nuove misure di resilienza calcolabili a livello di individuo e in grado di soddisfare alcune proprietà auspicabili. Queste misure possono essere calcolate utilizzando i dati delle indagini sulle famiglie EU-SILC o Banca d'Italia. Le misure di resilienza possono poi essere impiegate in modelli per le determinanti di diversi comportamenti sociali.

Riferimenti bibliografici

- Asheim, G. B., Bossert, W., D'Ambrosio, C., & Vögele, C. (2020). The measurement of resilience. *Journal of Economic Theory*, *189*, 105104.
- Cissé, J. D., & Barrett, C. B. (2018). Estimating development resilience: A conditional moments-based approach. *Journal of Development Economics*, *135*, 272-284.

Prof.ssa Silvia Pacei

THE MEASUREMENT OF FINANCIAL RESILIENCE

Household financial shocks can result from different reasons, for example: loss of employment, ill health of a family member, relationship breakdown, the loss of a partner, or an unexpected large expense. To cope with these shocks, households need access to sufficient liquid assets or emergency savings, or be able to borrow from financial institutions, wider family or friends. For those without sufficient liquid assets or emergency savings to cover a financial shock, the effects on financial well-being will be longer lasting. Some households are less resilient to financial shocks than others. This may be because they have low levels of savings, have limited access to affordable credit, already hold high levels of debt or lack the skills required to manage household budgets.

Financial resilience is difficult to estimate because it is a dynamic concept – the ability to recover quickly from an income or expenditure shock. This means that we have to turn to dynamic measures of resilience. So far, few authors have attempted to measure financial resilience (Brandolini et al. 2009; McKnight and Rucci, 2020). The aim of this project is to propose a definition of financial resilience and a measure of financial resilience that can be calculated at the individual level and that can satisfy some desirable properties.

This measure can be calculated using data from the Bank of Italy survey. The financial resilience measure can then be used in models for the determinants of different social behaviours.

References

Brandolini, A., Magri, S., and Smeeding, T. (2009). ‘Asset-related measures of poverty and economic distress’, working paper, https://www.researchgate.net/publication/253948770_Asset-related_measures_of_poverty_and_economic_stress.

McKnight A., and Rucci M. (2020), The financial resilience of households: 22 country study with new estimates, breakdowns by household characteristics and a review of policy options, Centre for Analysis of Social Exclusion and London School of Economics, CASE/219.

Prof. Luca Trapin

Analysis of illiquidity risk premium

Liquidity is a fundamental property of a well-functioning market, and lack of liquidity is generally at the heart of many financial crises and disasters. The financial economics literature hypothesizes the existence of an illiquidity premium in the market, i.e. investors require higher returns to hold illiquid stocks (Amihud and Mendelson, 1986). Liquidity is an elusive concept. It is not observed directly and cannot be captured in a single measure (Amihud and Mendelson, 1991). Using several proxies of liquidity, numerous studies have documented the existence of a positive relationship between stock returns and stock illiquidity, thus confirming empirically the existence of an illiquidity premium. Amihud (2002) shows that the existence of a premium is not only in cross-section but also in time series, i.e. future expected stock returns are increasing in expected illiquidity .

This project aims at investigating the conclusions on the illiquidity premium using ‘’liquidity factors’’ instead of ‘’liquidity proxies’’ (Hallin et al., 2011). The research student will have to: (i) build a large dataset of low-frequency liquidity proxies (Goyenko, 2009) for a large set of U.S. stocks using Eikon Refinitiv; (ii) extract liquidity factors from the liquidity proxies using factor models (Stock and Watson, 2002); (iii) run regression analysis for the identification of the illiquidity premium (Amihud, 2002).

References

1. Amihud, Y., & Mendelson, H. (1986). Asset pricing and the bid-ask spread. Journal of financial Economics, 17(2), 223-249.
2. Amihud, Y., & Mendelson, H. (1991). Liquidity, asset prices and financial policy. Financial Analysts Journal, 47(6), 56-66.

3. Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. Journal of financial markets, 5(1), 31-56.
4. Goyenko, R. Y., Holden, C. W., & Trzcinka, C. A. (2009). Do liquidity measures measure liquidity?. Journal of financial Economics, 92(2), 153-181.
5. Hallin, M., Mathias, C., Pirotte, H., & Veredas, D. (2011). Market liquidity as dynamic factors. Journal of econometrics, 163(1), 42-50.

Prof. Luca Trapin

Project Title
Interpretable Generative AI for Operational and Cyber Risk

Project Description:
This project aims to develop an innovative, interpretable framework for simulating operational and cyber risk losses using generative models. Unlike conventional Generative Adversarial Networks (GANs) that rely on opaque neural network architectures, this research explores interpretable generative mechanisms rooted in well-established statistical models commonly used in the operational risk literature. The goal is to construct a GAN-style architecture where the generator is replaced by a parametric statistical model. This model will be flexible enough to reproduce heavy tails and skewness typical of operational and cyber risk data, while remaining interpretable and grounded in the domain knowledge of risk management and extreme value theory. The discriminator component will be used to assess the goodness-of-fit of the generated samples relative to empirical data, enabling adaptive learning and calibration of the statistical generator.

Research Objectives:
1. Develop and implement interpretable generator models within a GAN-like architecture.

2. Explore loss functions and adversarial training schemes suitable for heavy-tailed data.

3. Validate the approach using publicly available datasets on operational and cyber risk losses.

4. Compare the performance of the interpretable GAN with traditional deep learning GANs and standard statistical approaches in terms of statistical fit and interpretability.

Ideal Candidate Profile:

We are looking for a motivated research assistant with a strong interest in applied risk modeling, extreme value theory or heavy-tailed distributions, programming in R or Python, generative AI models (GANs or similar).