IV Workshop on
Methodologies for official statistics
a tentative discussion

Marco Alfò

Dipartimento di Scienze Statistiche
Sapienza University, Rome - IT

ISTAT, Rome, 1 december, 2025

Session 1: Statistical models for OS

1. Haziza & Dagdoug, "Statistical inference for a finite population mean with machine learning-based imputation for missing survey data"
2. Di Cecco, Filipponi & Guarnera, "A contamination model for multivariate zero-inflated data"
3. Salvati, Schirripa Spagnolo, Bugallo & Morales, "Small area estimation via spatio-temporal M-quantile modeling"

Haziza & Dagdoug

- Challenge ML improves imputation for missing survey data but typically converges slower than parametric models, making valid inference difficult (bias, variance underestimation, no CLT).
- Solution Adapt double/debiased machine learning (DML) framework to survey sampling, combining outcome regression and propensity score models in a doubly robust (DR) estimator.
- Key Mechanism Neyman orthogonality ensures the DR estimator achieves root-n consistency even when ML models converge at a slower rate, enabling asymptotically valid inference.
- Variance Estimation A cross-fitted variance estimator is proposed that avoids bias due to overfitting bias and yields near-nominal coverage.
- Results The cross-fitted DR estimator is nearly unbiased with accurate variance estimates and correct coverage across ML methods; standard DR estimator seems to underestimate, due to *overfitting*

Di Cecco et al.

- Problem Zero-inflated multivariate data often contain *erroneous* values - both zeros and positive values; standard Tobit model is unable to capture excess zeros or heterogeneous error variability.
- Proposal The Intermittently Contaminated Tobit (ICT) model introduces a mixture of two Tobit components, with constant mean and different variances to identify units affected by intermittent errors.
- Mechanism A latent binary indicator is used to denote a component with inflated variance (error). This framework jointly models zeros and positives, enabling detection of *anomalous* units anywhere in the distribution.
- Estimation An EM algorithm is used for parameter estimation and provides unit-level error probabilities and expected true values; this may be used for enabling selective editing.
- Application Italian enterprise data, ICT outperforms standard Tobit and zero-inflated Tobit (lower AIC/BIC) and adjusts expected values, improving data quality.

Salvati et al.

- Objective The paper develops a spatio-temporal M-quantile model for small area estimation in the presence of spatial, temporal dependence and outliers.
- Proposal Extending standard M-quantile regression by considering area-level spatial and time-specific effects, improving robustness and prediction stability over time.
- Robustness M-quantile methods are robust to outliers and heavy-tailed distributions, making the approach suitable for noisy administrative or survey data.
- Estimation Parameters are estimated using an IRLS algorithm combining MQ regression with spatio–temporal weights, allowing flexible prediction across areas and periods.
- Application Simulation and empirical studies demonstrate improved accuracy over competing small-area methods, especially in settings with contamination or strong dependence patterns.

Some issues – Haziza and Dagdoug - 1

- Contribution vs. accessibility required assumptions such as the rate condition under cross-fitting and the boundedness of estimated propensity scores may be challenging.
- Tuning The proposed framework heavily relies on appropriate tuning for both the outcome and the propensity score models, essential for the convergence condition. Concerns about reproducibility and implementation at scale.
- Computational complexity Cross-fitting, multiple ML model fitting, residual variances estimation introduce substantial computational load, especially in large-scale surveys. Is it feasible with thousands of variables and millions of records?

Some issues – Haziza and Dagdoug - 2

- Flexibility vs. instability The cross-fitted variance estimator is a major advancement over traditional methods, but it also introduces new sources of instability. Any empirical exploration of remedies based on trimming and/or stratification on PS?
- Simulation study, design The authors explore a variety of ML algorithms and missingness mechanisms; the study focuses exclusively on simple random sampling without replacement and relatively simple models.
- Simulation study, scope Nonlinear functions without massive interactions, which should be the focus for ML techniques. A formal comparison with additive models may be of interest?
- Simulation study, generalizability Last, the study uses sample sizes ($n = 2.000$ and $n = 16.000$) that may not fully reflect modern large-scale surveys.

Some issues – Di Cecco et al. - 1

- Zero inflation It applies only to discrete or (rigid) semi-continuous data; the Tobit model is for data censored due to a *limit of detection*. Accidentally, it adapts to semi-continuous data with a mass at zero which, however, is not due to *inflation*.
- Restrictive distributional assumptions The Intermittently Contaminated Tobit (ICT) model is an elegant mixture of two Tobit components with shared means but different variances. While conceptually appealing, the approach relies heavily on strong and unverifiable assumptions on the latent process
- Computational complexity and scalability The EM algorithm is mathematically rigorous but computationally heavy, and unstable for high-dimensional q and/or large datasets, especially without constraints on the covariance matrices.

Some issues – Di Cecco et al. - 2

- Two-step (hurdle-type) models While ICT improves upon the standard Tobit and ZIT by allowing two sources of zero values, both are still motivated by latent Gaussian censoring. The model does not consider *structural* zeros, and lead to mis-classifying units as *contaminated*.
- Interpretability The model labels units as *contaminated* when they are predicted to belong to the component with inflated variance, but this may not be equivalent to detecting *true* data errors.
- Empirical application The application on Italian enterprise data demonstrates practical feasibility, but it presents several limitations (dimensionality, zero fraction, square root transform, differences between observed and predicted values, etc.)

Some issues – Salvati et al. - 1

- Innovation vs. Feasibility The model structure is computationally demanding: large spatio-temporal weight matrices, repeated cross-validation for bandwidth selection may be an issue when working with thousands of areas or long time series.
- Weight specification The models rely heavily on spatio-temporal weights combining spatial decay and AR-based weights. How sensitive predictions and MSE estimates are to the choice of spatial kernel, the bandwidth selection procedure, or the AR-order chosen?
- MSE estimation The proposed RMSE estimators - particularly $mse_{stmq,1}$ - tend to underestimate RMSE as the proportion of out-of-sample subregions increases. Any strategy to quantify how this bias impacts the accuracy of uncertainty quantification in practice?

Some issues – Salvati et al. - 2

- Simulation design The simulation setup - $D = 25$ regions, $\mathcal{K}_d = 4$ subregions each, $T = 8$ time periods - represents a highly stylized spatial layout. It does not reflect the complex patterns in the real AQI data (over 1,000 counties, varying sampling patterns).
- Interpretation The region-specific M-quantile coefficients θ_d are central to the plug-in predictors, yet their interpretation is not sufficiently discussed. Since θ_d is influenced by both spatial and temporal weights, its meaning as *representative* probability level becomes increasingly complex.

...Thanks for Your attention & patience...