

Towards a Quality Framework for Gender-Related Citizen Data: Challenges, Principles and Perspectives

Monica Pratesi, University of Pisa

Citizen-generated data (CD) is an increasingly important source of evidence for monitoring gender equality and supporting the Sustainable Development Goals (SDGs). Produced by civil society organizations, community networks, and citizens, these data can capture lived experiences and emerging issues often missed by traditional statistical systems, including gender-based violence, unpaid care work, discrimination, and the realities of marginalized populations.

However, the growing use of citizen data raises important quality challenges. Traditional quality frameworks in official statistics, based largely on probability sampling and the Total Survey Error paradigm, do not fully reflect the diversity of citizen data sources, methods, and objectives.

Gender-related citizen data is not simply sex-disaggregated data; it also reflects how power relations and inequalities shape both social realities and data production. Quality must therefore extend beyond statistical accuracy.

This paper proposes a flexible quality framework based on three dimensions: Trustworthiness, Inclusivity, and Utility. Drawing on collaborations between National Statistical Offices and Civil Society Organizations, it explores how these dimensions can support rigorous and inclusive quality principles that recognize lived experience as a valuable form of evidence for policy and advocacy.

Combining probability and nonprobability samples to estimate linear model coefficients

Char Hilgers, German Institute for Economic Research (DIW Berlin), Socio-Economic Panel

Nonprobability sampling is increasingly preferred for surveys due to much lower costs and ease of implementation, compared to large probability samples. However, probability samples remain important because they generally provide more accurate estimates than nonprobability samples and provide benchmark information for nonprobability surveys.

There is a growing body of research on the integration of probability and nonprobability samples to enjoy the strengths of both.

Several approaches have been developed to integrate probability and nonprobability samples, including pseudo-weighting, blended calibration, use of informative priors, and doubly robust estimators.

We present a comprehensive comparison of the pseudo-weighting, blended calibration, informative prior, and composite estimator approaches, to evaluate their efficacy in estimation of parameter values and reduction of variance in linear model coefficients, in contrast to the probability sample alone.

We apply the methods to a real-data probability and nonprobability sample measuring beliefs in conspiracy myths in Germany. We compare maximum likelihood estimates of regression coefficients and 95 percent confidence intervals for each method, with outcome variables on conspiracy beliefs and predictors covering demographic, education, and employment

information collected in both samples.

Digital Breadcrumbs of Migration: Data-Driven Insights into Immigrants' Online Activities

Jisu Kim, Utrecht University

In an era of pervasive digital connectivity, online platforms generate vast amounts of data that can provide valuable insights into the experiences and behaviors of immigrant communities. This talk presents a data-driven approach for studying immigrants and their online activities through the analysis of digital breadcrumbs derived from social media. Rather than relying on a single indicator, the approach integrates multiple signals embedded in digital traces, including linguistic patterns, cultural references, migration-related content, social network structures, geolocation information, and temporal activity patterns. By combining these diverse data dimensions, the approach enables a richer understanding of how immigrants engage, interact, and form communities in digital environments. The findings contribute to broader discussions on migration, integration, and social dynamics in the digital age, while highlighting the opportunities and challenges of using online data to study population groups.

Geocoding Uncertainty and Spatial Bias: Implications for Multi-Source Research

Manuela Schmidt, RPTU Kaiserslautern-Landau

Geocoding is a key step for linking survey respondents' addresses with administrative records, regional statistics, and geospatial context data. Systematic spatial errors, such as centroid placement for incomplete addresses or address sparsity in rural areas, propagate into downstream analyses that rely on geocoded locations, potentially introducing measurement error and spatial bias that vary systematically across regions.

This study presents a nationwide quantitative assessment of commonly used geocoding services in Germany, covering open-data and commercial providers. The analysis is based on two datasets: an authoritative set of hospital addresses with government-provided coordinates serving as the ground truth, and a large sample of randomly selected residential-like addresses extracted from OpenStreetMap. This design allows both accuracy assessment and relative comparison of geocoding consistency under realistic conditions, without relying on personal address data. The analysis also explores how geocoding error and provider disagreement vary with urban-rural context, population density, proximity to national borders, and address characteristics.

Building on this, the talk discusses implications for multi-source data integration: How spatial uncertainty in geocoding affects the reliability of linked contextual variables such as regional typologies, administrative unit assignments, and environmental indicators, and how it may interfere with survey-based research.

