

metasurvey: pipelines reproducibles para el análisis de encuestas complejas en R

Metaprogramación, recetas y estimación con pesos replicados

Palabras clave— Reproducibilidad, Bootstrap, BRR, Jackknife, Metaprogramación, Encuestas por muestreo

Resumen

metasurvey estructura el análisis de encuestas complejas en pipelines reproducibles con metaprogramación y recetas. Integra **survey** para estimación bajo diseño y admite pesos replicados (bootstrap, BRR/Fay, jackknife), priorizando trazabilidad de extremo a extremo. Ofrece: (i) API para declarar diseños (pesos base y replicados); (ii) *recipes* que armonizan variables entre olas/años; y (iii) *workflows* que generan indicadores y errores estándar por período y subpoblación. Incluye una API opcional de recetas con backend MongoDB para versionar, validar y compartir definiciones entre equipos. El paquete está en desarrollo activo, desarrollado por el autor, con documentación y ejemplos en <https://metasurveyr.github.io/metasurvey> y código en <https://github.com/metasurveyR/metasurvey>. Como caso de estudio, se estiman indicadores laborales mensuales y trimestrales en la ECH (Uruguay) con replicación bootstrap.

Introducción y motivación

La comparación temporal en encuestas oficiales suele verse afectada por cambios de cuestionario y definiciones, induciendo deriva entre olas y reimplementaciones ad hoc. Proponemos pipelines reproducibles con metadatos que capturan, desde el diseño muestral hasta la estimación e informe, los pasos y parámetros aplicados, favoreciendo replicabilidad, mantenimiento y colaboración.

Metodología y API

El paquete se apoya en **survey** para diseños con estratificación y conglomeración, y en **svrepdesign** para varianza por replicación (Lumley, 2004,1; Rust & Rao, 1996). La API permite:

- Declarar pesos base, FPC y pesos replicados (bootstrap, BRR/Fay, jackknife) (Binder, 1983; Fay, 1989; Rao & Wu, 1988).
- Definir *recipes* de recodificación/derivación (p.ej., PEA, PO, PD) versionadas y reutilizables.
- Ejecutar *workflows* que producen estimadores (tasas, ratios, medias) y sus errores, registrando insumos y software.
- (Opcional) Gestionar un repositorio de recetas y metadatos vía API con backend MongoDB (Bradshaw, Brazil y Chodorow, 2020) para versionado semántico, validación y reutilización entre proyectos.

Compatible con sintaxis *tidy* (**dplyr/srvyr**) (Wickham et al., 2019; Ellis, 2017) y orquestación con **targets** (Landau, 2021), en entornos aislados con **renv**.

Caso de estudio

Aplicamos *metasurvey* a la Encuesta Continua de Hogares (Uruguay), declarando un esquema panel con replicación bootstrap para estimar indicadores laborales mensuales y trimestrales. Las transformaciones se encapsulan en *recipes* por edición/año, facilitando armonización y documentación.

Interoperabilidad, comunidad y documentación

Para la comunidad de R y actores públicos/privados, un catálogo de recetas en MongoDB permite:

- Compartir y citar definiciones estandarizadas, reduciendo duplicación y ambigüedades.
- Reutilizar transformaciones entre proyectos y equipos, con control de versiones y trazabilidad.
- Validar consistencia de series temporales con pruebas automáticas.

Documentación, guías y API: <https://metasurveyr.github.io/metasurvey>

Resultados y conclusiones

El pipeline replicó indicadores con varianzas consistentes entre olas, reduciendo reescritura y errores de integración. La separación entre (i) diseño, (ii) transformaciones y (iii) estimación permitió cambios controlados, extensión a nuevos dominios y revisión por pares más efectiva. Trabajo futuro: calibración pos-estratificada, diagnósticos de diseño y plantillas de informes automatizados.

Referencias

- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3), 279–292.
- Bradshaw, S., Brazil, E., & Chodorow, K. (2020). *MongoDB: The Definitive Guide* (3rd ed.). O'Reilly Media.
- Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Survey Research Methods Section*, ASA, 212–217.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(8).
- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley.
- Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401), 231–241.
- Rust, K., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3), 283–310.
- Wickham, H., Averick, M., Bryan, J., Chang, W., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.
- Ellis, G. (2017). *srvyr: 'dplyr'-Like Syntax for Summary Statistics of Survey Data*. R package. <https://CRAN.R-project.org/package=srvyr>
- Landau, W. M. (2021). The *targets* R package: a dynamic Make-like function network for reproducibility and high-performance computing. *Journal of Open Source Software*, 6(57), 2959.