SPES Sustainability performances, evidence & scenarios

D4.1 Report on mapping, harmonising and integrating novel data sources for research purposes

January 2024

Funded by European Union's Horizon Europe Programme under Grant Agreement No. 101094551

Authors

Veronica Ballerini – Researcher of the SPES Project, University of Florence
Davide Beraldo – Researcher of the SPES Project, University of Amsterdam
Chiara Bocci – Researcher of the SPES Project, University of Florence
Lisa Braito – Researcher of the SPES Project, University of Florence
Roberta Milana – Interning researcher of the SPES Project, University of Amsterdam
Emilia Rocco – Researcher of the SPES Project, University of Florence
Martin Trans – Researcher of the SPES Project, University of Amsterdam

Contributors and peer reviewers:
Mario Biggeri, University of Florence; Paolo Brunori, London School of Economics & Political Science; Jeroen de Vos, University of Amsterdam; Andrea Ferrannini, University of Florence; András Gábos, TARKI; Luca Lodi, University of Florence; Fabrizia Mealli, European University Institute; Stefania Milan, University of Amsterdam; Gregor Zens, IIASA.

Acknowledgements:
The authors would like to thank all SPES partners for their inputs provided in several SPES meetings and communications. We would like to thank also all interviewees involved in our data collection activities for having shared with us their opinion and insights on the topic.

Cite as:
Ballerini, V., Beraldo, D., Bocci, C., Braito, L., Milana, R., Trans, M. (2024). Report on mapping, harmonising and integrating novel data sources for research purposes. SPES Report no. 4.1, SPES project – Sustainability Performances, Evidence and Scenarios. Florence: University of Florence. Available at: https://www.sustainabilityperformances.eu/publications-deliverables/

Disclaimer

This Report D4.1 for the project SPES has been prepared by the University of Amsterdam and the University of Florence as part of Task 4.1 “Map complex and novel data sources and methods” of Work Package 4.

This task has allowed SPES research partners to identify and provide an evaluation of complex and novel data sources and associated methods that can be repurposed for the development of innovative measurement framework.

Table of contents

Abstract

Sustainability, a complex and multi-dimensional phenomenon at the crossroads of environmental, social, geographical, and economic considerations, becomes increasingly crucial to explore and measure as we work towards a more sustainable future. Over the past decades, the rise of innovative data sources has significantly expanded our ability to assess the various dimensions of sustainability. Thus, this report provides on the one hand a mapping of innovative data sources used to measure the dimensions of sustainability transition and on the other hand an overview of data integration methods which enable the use of this novel data sources for research purposes.

The first part of this research consists in a Large Language Models-enabled distant reading of a large collection of academic texts related to innovative data sources used in the study of sustainability. We developed an expansive operationalization of the notion of sustainability, grounded in the SPES framework, to account for its multi-dimensional and inter-disciplinary character. GPT models have been used at various stages to extract, classify and organise information. The analysis shows which type of innovative data categories have been associated to which dimensions of sustainability. To illustrate this, we also present several selected examples.

The second part of this report explores the evolving domain of data integration methods in official statistics and survey methodology. It investigates the possible methods to reach integration of alternative data sources, e.g., administrative records, satellite data, and web data, highlighting their potential to complement and enrich traditional data sources. It offers insights into the challenges and opportunities associated with data integration, providing a comprehensive overview for a broad audience interested in the evolving landscape of statistical methods.

General introduction

Sustainability is a complex, multi-dimensional phenomenon that lies at the intersection of environmental, social, geographical and economic considerations. As we navigate the path toward a more sustainable future, the need to study and measure the performance of sustainability transitions becomes crucial. In the past decades, the advent of innovative data sources has significantly expanded our ability to gauge the many dimensions of sustainability addressed by the SPES framework (Biggeri et. 2023). The inexorable process of *datafication* (Kitchin, 2022) –i.e., the transformation of more and more domains of society into quantified information ready for collection, processing and circulation, has impacted research practices across fields. Unlike traditional data sources (e.g., surveys and census data), novel data sources are often generated as a by-product of activities of all kinds and are only subsequently (and sometimes creatively) repurposed in innovative ways for research goals. Such a complex and dynamic landscape requires continuous exploration and mapping across disciplinary boundaries, to ensure that both potentials and pitfalls are adequately seized and avoided. Albeit providing unprecedented opportunities, indeed, such new sources of data come with several methodological and epistemological caveats.

Traditional data sources have played a pivotal role in furnishing researchers and practitioners with valuable statistical insights for over 80 years. Their reliability in generating statistics across various domains is fundamental and their role will remain unchanged in the future, at least in those countries where well established statistical offices exist. However, there exists ample opportunity to enrich traditional data by incorporating information from alternative sources. Moreover, various current phenomena such as globalisation, recession, economic growth, climate change, urbanisation, digitalisation, ageing society, conflicts, extreme weather events, and humanitarian crisis and their fast pace and development highlight the need for more timely but at the same time reliable ways to capture their multifaceted nature (Bosco et al., 2022).

This report is structured in two complementary parts. Part 1 focuses on the identification, mapping and illustration of innovative data sources for the study of sustainability transition. It answers the following question: What innovative data sources are currently leveraged to study the many dimensions of sustainability? It consists of exploratory research that leverages Large Language Models (LLMs) to extract, classify and organise information about the use of innovative data sources in the study of sustainability, based on a large corpus of recent academic studies. One of the issues that this analysis highlights is the heterogeneity of data types that can be used for the purpose, and the widespread practice of integrating data of different types. Therefore, Part 2 consists in a systematic review of data integration strategies for the creation of a synthetic dataset and for combining probability and nonprobability samples also in the context of non-conventional data sources. It answers the following questions: What is the state-of-the art in data integration methods? Which existing methods can be leveraged to develop an appropriate harmonisation and integration strategy of both traditional and new data sources? This section aims at providing an overview for a broader audience of the statistical methods developed in the last decades in Official statistics and Survey statistics to deal with data integration issues, at proposing toy/practical examples to illustrate both the issues at stake and the methods suggested to tackle them and supplying available statistical software and packages for their implementation.