International Journal of Population Data Science

Journal Website: www.ijpds.org

Establishing an International Data Linkage Repository Workgroup Toward a Benchmarking Repository

Kum, H-C[1], Leonard, S[2], Akgün, Ö[3], Alexander, T[2], Antonie, L[4], Levenstein, M[2], and O'Hara, A[5]

[1]Texas A&M University
[2]ICPSR, University of Michigan
[3]University of St Andrews
[4]University of Guelph
[5]Stanford University

Introduction

Access to real data with diverse attributes is critical for effective development of any data analytic algorithm. Benchmarking data repositories have all been vital to the development of research communities focused on algorithm development. This work reports on the development of such a data repository for record linkage.

Objectives and Approach

Establishing a common benchmarking repository of real data can propel a field to the next level of rigor by facilitating comparison of different algorithms, understanding what type of algorithms work best under certain real data conditions and problem domains, promoting transparency and replicability of research, and creating incentives for proper citations for contributions. In addition, benchmarking repositories can bring together the diverse stakeholders (e.g., computer scientists, statisticians, data custodians, data users including social, behaviour, economic, and health (SBEH) scientists) that can advance the field more effectively than could researchers from any single discipline.

Results

In Fall 2016, international leaders in record linkage formed a Data Linkage Repository workgroup (DLRep) to establish a benchmarking data repository for record linkage. The workgroup is working in collaboration with The Inter-university Consortium for Political and Social Research (ICPSR) to host the site data repository planned for release in Summer 2018. The repository for record linkage research will house various types of real data that require linking with metadata, unique handles for citations, proposed algorithms for evaluation criteria, and a platform for posting, sharing, and comparing results as well as citations of relevant papers. Some datasets will have the gold standard published that researchers can evaluate their results against. Other datasets will gather results to build the gold standard as a community.

Conclusion/Implications

Record linkage methodology is important to domains where data needs to be integrated from multiple sources, including diverse disciplines. Establishing an international interdisciplinary research community around a benchmark data linkage repository to validate and compare linkage algorithms is crucial to fully realizing the social benefits of data about people.

http://dx.doi.org/10.23889/ijpds.v3i4.835