

Submission for the 2020 IAOS Prize for Young Statisticians

Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis

Ms. Kenza Sallier

Methodologist

Statistics Canada

www.statcan.gc.ca

Ottawa, Canada

Abstract

Under the Modernization programme Statistics Canada has recently undertaken, the Agency is to put forward data access solutions that present greater analytical value to Canadians while maintaining its core values of protecting confidentiality of respondents' information. One avenue currently explored is Data Synthesis as a means of delivering synthetic data *with high analytical value* to users. At the time of writing, Statistics Canada has publicly released synthetic versions of two different datasets related to census, mortality and cancer information. In both cases, synthetic data were generated using the R package *synthpop*. This paper describes the use of Data Synthesis as a proof of concept for modernizing Statistics Canada's data access solutions. More specifically, the methodology underlying Data Synthesis is presented as well as the aforementioned synthetic data projects and some of the undergoing research.

Keywords: Synthetic data, data access, confidentiality, digital government, machine learning.