International Journal of Population Data Science

Journal Website: www.ijpds.org

# Assessing data linkage solutions in high-volume settings

Laura Tosco[1] and Tiziana Tuoto[1]

[1]Istat, Rome, Italy

To design a record linkage strategy, different factors need to be taken into account, e.g. the size of the linking files, the level of errors in the matching variables, the unknown match rate. All these aspects impact on several choices: either a deterministic or probabilistic approach, (e.g. the traditional Fellegi and Sunter, a Bayesian framework, some machine learning algorithms), solutions for reducing and filtering the search space of the links, the choice of the matching variables and the metrics for comparison, the setting of resources for manual revision and checks of the results and ambiguous cases. When dealing with large volume data, as is often the case in national agencies, everything becomes even more complicated since some solutions are not scalable. In this presentation, we explore some solutions based on machine learning techniques that are very convenient for large amount of data, due to their scalability and computational efficiency. We appraise the machine learning solutions, comparing them to the traditional probabilistic solution and underlining benefits and limits. The comparison includes elements like the file size, the level of errors in the matching variables, the matching rate, along with the traditional diagnostics for the linkage results, like the false match and missing match rates, time of execution and computation resource required.

https://doi.org/10.23889/ijpds.v11i5.3725