

International Journal of Population Data Science

Journal Website: www.ijpds.org



Rethinking data linkage through machine learning techniques

Tiziana Tuoto¹ and Laura Tosco¹

¹Istat, Rome, Italy

In this presentation we propose a combination of machine learning techniques to perform record linkage processes in the presence of high-volume data. Indeed, record linkage is a complex process composed of several steps, each of them exploiting different techniques. Machine learning algorithms are usually exploited to reduce the pair space search, in the well-know Fellegi and Sunter probabilistic framework. However, this framework hardly manages high volume data and extends to the case of linking simultaneously multiple files. Our proposal relies on efficient algorithms based on approximate nearest neighbour (ANN) techniques to compute the distance between records and reduce the links search space,

combined with linear optimization algorithms and graph techniques for the final selection of the links. The proposed solution is implemented in open-source language. It dramatically overcomes the computational performances of other fast and scalable linkage packages. Diagnostics to evaluate the linkage results are provided, and it shows to be very effective in avoiding false links while reducing the risk of missing true matches. The proposal has been tested in several scenarios, including different sizes of the linking files, different overlaps between them, and different levels of errors in the matching variables, allowing us to guide practitioners in setting the overall linkage strategy in real life cases.

