QUAIL: QUality Analyser for Interpreting Linkage

Leah Maizey, Sarah Collyer, Tim Gammon & Gavin Thomson.

Office for National Statistics

Streamlining the quality review process in data linkage: automation of sample creation and metric computation

Background & Aims

- Metrics are needed to understand the quality of a linked product.
- No standardised method of computing sample size for error estimation is used in ONS.
- Quality metric estimation typically involves manual entry of results from multiple samples/files that have been clerically reviewed. This can be burdensome and error-prone.

QUAIL aims to:

1. Automate the creation of samples for clerical review using researched sampling methodologies.
2. Automate the computation of quality metrics (precision, recall and associated variance estimates).

Stratification → Sampling → Clerical Review → Precision & Recall Tool

1. Stratification

Candidates	Stratum
High	1
	2
	3
Prob. score	4
	5
	6
	7
Low	8

- Probabilistic scoring of candidate links.
- Strata are based on the division of probabilistic scores.
- Optimal stratification to account for variance across strata.
- Characteristics within/across strata are important.
- Candidate links created for False Positive (FP) and False Negative (FN) review separately.

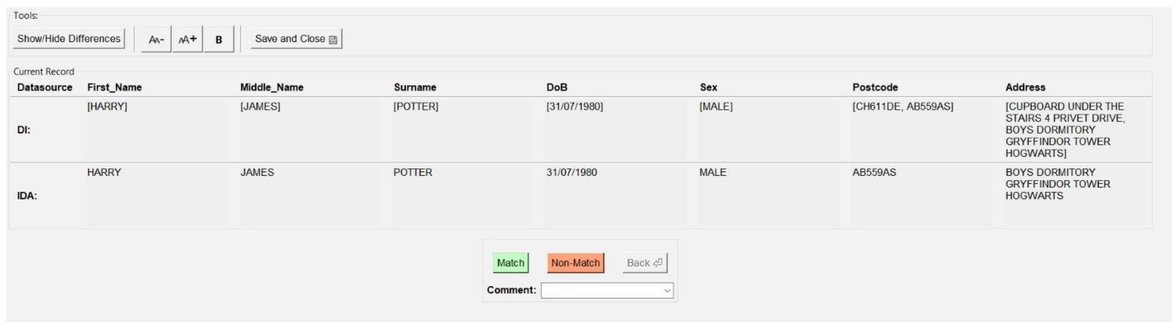
2. Sampling

- Bayesian approach is under development.
- Offers flexibility and use of data to inform expected error.
- Probabilistic analysis of the relationship between sample size and accuracy estimates.
- Incorporates prior knowledge of expected error to predict error rates.
- Intuitive interpretation of uncertainty in estimates.
- Allows for continuous updating of estimates given new data.
- Investigating methods for arbitrary priors:

Aspect	Numerical Method (Markov Chain Monte Carlo)	Conjugate Priors (Beta Distribution)
Robustness	High – copes well with biased or irregular data	Low – assumes unbiased, independent data
Computational Cost	High – due to iterative simulations	Low – due to analytical solutions
Adaptability of Priors	High – works with a range of data and parameters	Medium -- constrained to a choice of beta distributions
Ease of Implementation	Tuning and diagnostics needed in implementation	Straightforward to run with choice of initial priors

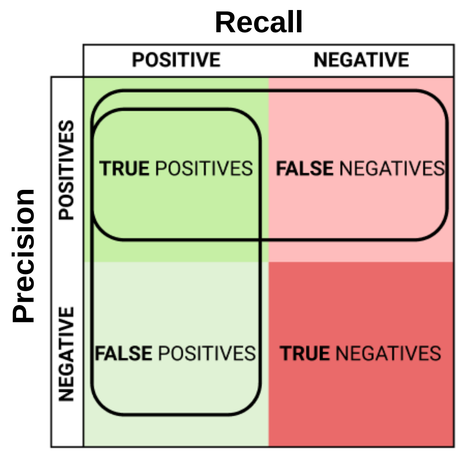
3. Clerical Review

- Use of ONS' Clerical Resolution Online Widget (CROW).
- Reviewers determine if link is a match or not.
- Decisions recorded in .csvs.



4. Precision & Recall Tool

- Automates the computation of quality metrics post clerical review (precision, recall and variance estimates).



- Coded in PySpark.
- Input .csvs from CROW.
- Create estimates of FPs and FNs per stratum.
- Computes precision, recall and variance across strata.

- Development needed:
  - Incorporate Bayesian variance estimates.
  - Create user interface.

Future Directions

- Finalise generalised stratification and sampling approaches.
- Further development of the Precision and Recall tool.
- Development with respect to use cases and longitudinal linkage.
- Recode in Python to enable wider use.
- Expand QUAIL to include descriptive statistics and bias estimates (e.g., as per ONS' Bias Analysis Tool).

References:

Ellis, J., & Wyatt, R. (2009). A comparison of Bayesian and Frequentist approaches for estimating WFD classification errors. *Environment Agency*.

Shipsey, R. & Spakulova, I. (2021). Bayesian Approach to Sampling Applied to False Negative Assessment of Census to CCS Matching. *Internal ONS paper*.

Contact: datalinkage@ons.gov.uk