Enhancing Record Linkage Production Data Quality

Douglass Huang, Director of Engineering

Todd Johnsson, CEO

ADI, LLC

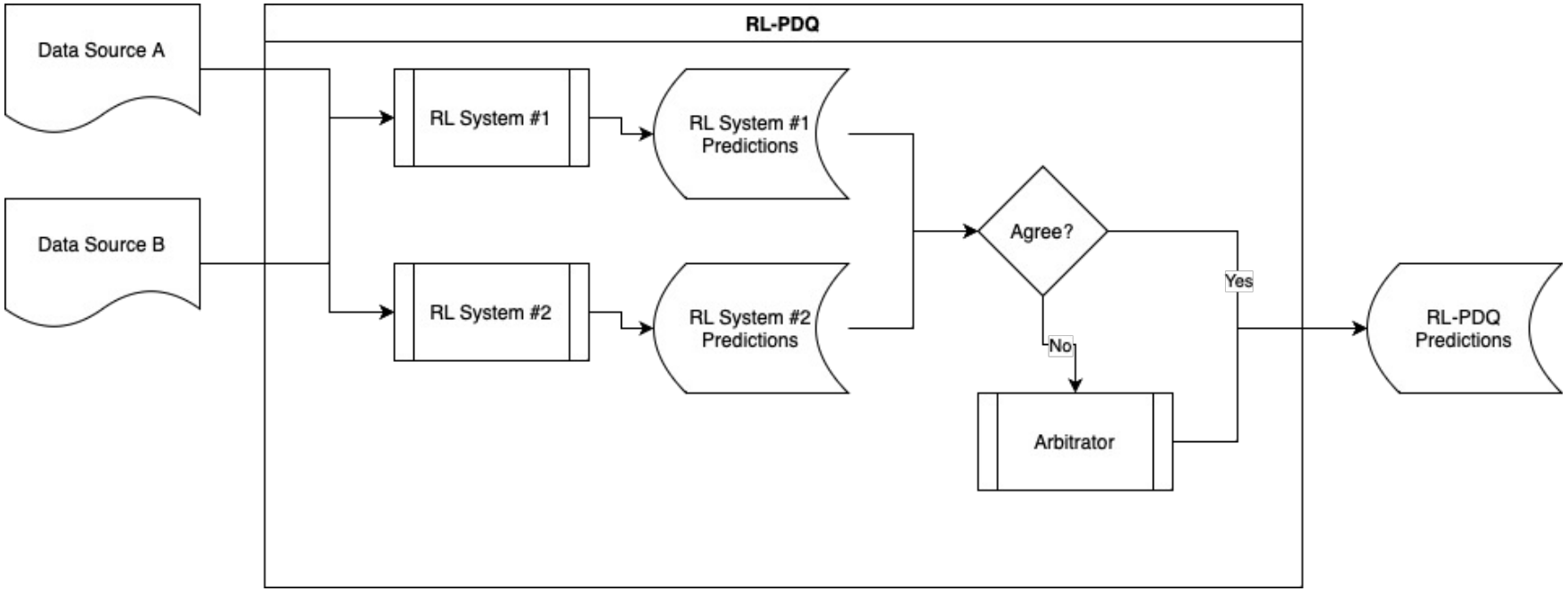
Introduction

- Record Linkage Production Data Quality (RL-PDQ) is a methodology for enhancing any existing record linkage system
- Architecture-independent
- Works by integrating the original system into a larger process
- Doesn’t disrupt core design or operation of original system
- Drives substantial accuracy and data completeness improvements

Background

- Production Data Quality (PDQ) independent verification and validation (IV&V) Tool
  - Developed and operated by ADI for 2010 Census
  - Assessed Decennial Response Integration System (DRIS) paper data capture quality
  - Achieved truth error rate ≥80% lower than equivalent double-key and verify (DK&V) operation
  - 96% reduction in manual effort compared to DK&V
- RL-PDQ applies PDQ's core principle to record linkage

RL-PDQ Block Diagram



See Patent No: US 10,067,976 B2 [Method for Enhancing Record Linkage Production Data Quality]

Operational Variables

- False Negative Rate
  - $FNR = \frac{FN}{P} = \frac{FN}{FN+TP}$
- False Positive Rate
  - $FPR = \frac{FP}{N} = \frac{FP}{FP+TN}$
- True Positive Rate
  - $TPR = \frac{TP}{P} = \frac{TP}{FN+TP} = 1 - FNR$
- True Negative Rate
  - $TNR = \frac{TN}{N} = \frac{TN}{FP+TN} = 1 - FPR$

Basic Concept

- Example:
  - Data Source A: 1,000 records; Data Source B: 1,000 records
  - $P = 1{,}000;\ N = 999{,}000$
  - $FPR_1 = 10/999{,}000 \approx 0.0010\%$
  - $FNR_1 = 10/1{,}000 = 1.0\%$
  - $FPR_2 = 20/999{,}000 \approx 0.0020\%$
  - $FNR_2 = 20/1{,}000 = 2.0\%$
- The likelihood that the two independent RL systems agree on a false prediction is extremely small:
  - $(FNR_1 \times FNR_2 \times P + FPR_1 \times FPR_2 \times N)/(P + N) \approx 0.000020\%$
- RL-PDQ leverages automation to gain efficiency: Where there is agreement between the RL systems, we assume that both systems are correct
- The likelihood that the two independent RL systems disagree is on the order of their respective false positive rates:
  - $1 - (FNR_1 \times FNR_2 \times P + FPR_1 \times FPR_2 \times N)/(P + N) - (TPR_1 \times TPR_2 \times P + TNR_1 \times TNR_2 \times N)/(P + N) \approx 0.0060\%$
- RL-PDQ leverages well-trained human analysts to gain accuracy: Where there is disagreement between the RL systems, we invoke manual review to choose the correct match status

A Possible Scenario for Census 2030

"JASON recommends that the Census Bureau consider starting the 2030 Census with an 'in-office' enumeration of the population using existing government administrative records. That would be followed by a second step using additional data and more traditional methods to find people not present in government records and to 'fill in' variables that might be missing in these records."

- The MITRE Corporation, *Alternative Futures for the Conduct of the 2030 Census*, Nov 2016

Record Linkage Simulation

- Two simulated data sources to link
  - 2020 Census Housing Unit responses
  - IRS U.S. Individual Income Tax Returns (Form 1040) for 2028, filed in 2029
- Record linkage problem: Do a given Census response and a given tax return represent the same household?

Simulated Data Characteristics

- Simplifying assumptions:
  - Households stay intact
  - No births, deaths, marriages, or divorces
  - Only Person 1 from Census response can be primary tax filer
- Data features
  - ~1K Census responses with real addresses and varying household sizes and compositions, using same data model we developed for the 2020 Census
  - 2.5% of Persons 1 have a full name identical to that of another Person 1
  - 10% of persons who have nicknames use their nickname as their first name in the Census response
  - 85% of Persons 1 file a tax return
  - Whenever Person 2 is Person 1's spouse, they file a joint tax return
  - 10% of households move between the 2020 Census and 2028 tax year
- Proof of concept, not limitation of simulated data technology

RL-PDQ Experiment

- Configure and run RL System #1 to link Census responses with IRS returns
- Configure and run RL System #2 (different algorithms, different configuration) to link Census responses with IRS returns
- Compare outputs of two RL systems
  - Agreement → Final Prediction
  - Disagreement
    - Simulate RL-PDQ Arbitrator subsystem with varying error rates

Performance Metrics

- False Discovery Rate
  - $FDR = \frac{FP}{FP+TP}$
- False Negative Rate
  - $FNR = \frac{FN}{P} = \frac{FN}{FN+TP}$

Results

System	TP	FP	FDR	ΔFDR	FN	FNR	ΔFNR
RL System #1	926	9	0.01		22	0.023	
RL System #2	930	15	0.016		18	0.019	
RL-PDQ: FDR.a = 0, FNR.a = 0	948	2	0.002	-80%	0	0	-100%
RL-PDQ: FDR.a = 0.002, FNR.a = 0.0058	947.72	2.08	0.00219	-80%	0.28	0.00030	-99%
RL-PDQ: FDR.a = 0.005, FNR.a = 0.012	947.53	2.26	0.00238	-80%	0.47	0.00050	-98%
RL-PDQ: FDR.a = 0.01, FNR.a = 0.023	947.20	2.53	0.00266	-70%	0.80	0.00084	-96%

Discussion of Outcomes

- RL-PDQ Arbitrator workload was 60 record pairs
- RL-PDQ reduced FDR by 70-80%
  - Impacts survey data quality
- RL-PDQ reduced FNR by 96-100%
  - Impacts cost of traditional enumeration efforts

Additional Thoughts

- RL System #1 and RL System #2 should have different operating characteristics (specifically, produce different false positives and false negatives) for best data quality
- RL System #1 and RL System #2 should have lower FDR and lower FNR for best efficiency

Recap

- Record Linkage Production Data Quality (RL-PDQ) is a methodology for enhancing any existing record linkage system
- Architecture-independent
- Works by integrating the original system into a larger process
- Doesn’t disrupt core design or operation of original system
- Drives substantial accuracy and data completeness improvements

Next Steps

- Engage with appropriate personnel who are using record linkage in order to refine use case
- Tailor experimentation to more appropriate potential use

Applicable Patents

- K. Bradley Paxton, William L. DiBacco, Steven P. Spiwak, Craig A. Towne, and Manuel Trevisan. Handprint Recognition Test Deck. Patent No.: US 8,498,485 B2, Filed Apr 13, 2012, Issued Jul 30, 2013.
- Joshua David Glasser and Gary A. Passero. System and Method for Rule-Driven Constraint-Based Generation of Domain-Specific Data Sets. Patent No: US 8,862,557 B2, Filed Mar 12, 2010, Issued Oct 14, 2014.
- Douglass Huang, Steven Paul Spiwak, and K. Bradley Paxton. Method and System for Assessing Data Classification Quality. Patent No: US 8,498,948 B2, Filed Jul 30, 2010, Issued Jul 30, 2013.
- K. Bradley Paxton. Method for Enhancing Record Linkage Production Data Quality. Patent No: US 10,067,976 B2, Filed Mar 17, 2015, Issued Sep 4, 2018.

Bibliography

- Douglass Huang. Balancing truth error and manual processing in the PDQ system. Thesis, Aug 2011.
- Gunnison Consulting Group, Inc., and ADI, LLC. Production Data Quality (PDQ) Executive Report on DRIS 2010 Data Capture Quality. USCB Contract No: YA1323-09CQ-0024, Task Order TO003, Sep 2011.
- Gunnison Consulting Group, Inc., and ADI, LLC. Production Data Quality (PDQ) Final Report on DRIS 2010 Data Capture Quality. USCB Contract No: YA1323-09CQ-0024, Task Order TO003, Sep 2011.
- K. Bradley Paxton and Thomas Hager. Use of Synthetic Data in Testing Administrative Records Systems. 2012 FCSM Research Conference, Jan 2012.
- K. Bradley Paxton. Testing Record Linkage Production Data Quality. 2013 Federal CASIC Workshops, Mar 2013.
- K. Bradley Paxton. Testing Record Linkage Production Data Quality. JSM 2013, Aug 2013.
- K. Bradley Paxton. How Good Is Your Record Linkage System, Really?. 2014 Federal CASIC Workshops, Mar 2014.
- K. Bradley Paxton. Using Record Linkage to Create Big Data? How Good Is It?. JSM 2014, Aug 2014.
- K. Bradley Paxton. Two New Quality Metrics for Measuring Big Data Record Linkage Systems. 2016 Federal CASIC Workshops, May 2016.
- The MITRE Corporation. Alternative Futures for the Conduct of the 2030 Census. census.gov, Nov 2016.
- Todd Johnsson and Beverly Harris. Correlated Simulated Data for Decennial System of Systems Development and Test - Privacy by Design. 2021 Federal CASIC Workshops, Apr 2021.