UNITED STATES
DEPARTMENT OF VETERANS AFFAIRS

Application of Jaro-Winkler String Comparator in Enhancing Veterans Administrative Records

Hyo Park, Eddie Thomas, Pheak Lim
The Office of Data Governance and Analytics
Department of Veterans Affairs
FCSM, 2018

Introduction

- U.S. Veterans Eligibility Trends and Statistics (USVETS) - an integrated Veterans database using administrative records from internal and external data sources
- Social Security Number (SSN) was used as the primary unique identifier to link records across sources.
- Records linked by SSN alone resulted in matching records from different individuals.
- Utilized string comparator in conditional matching of multiple data sources to improve the data quality.

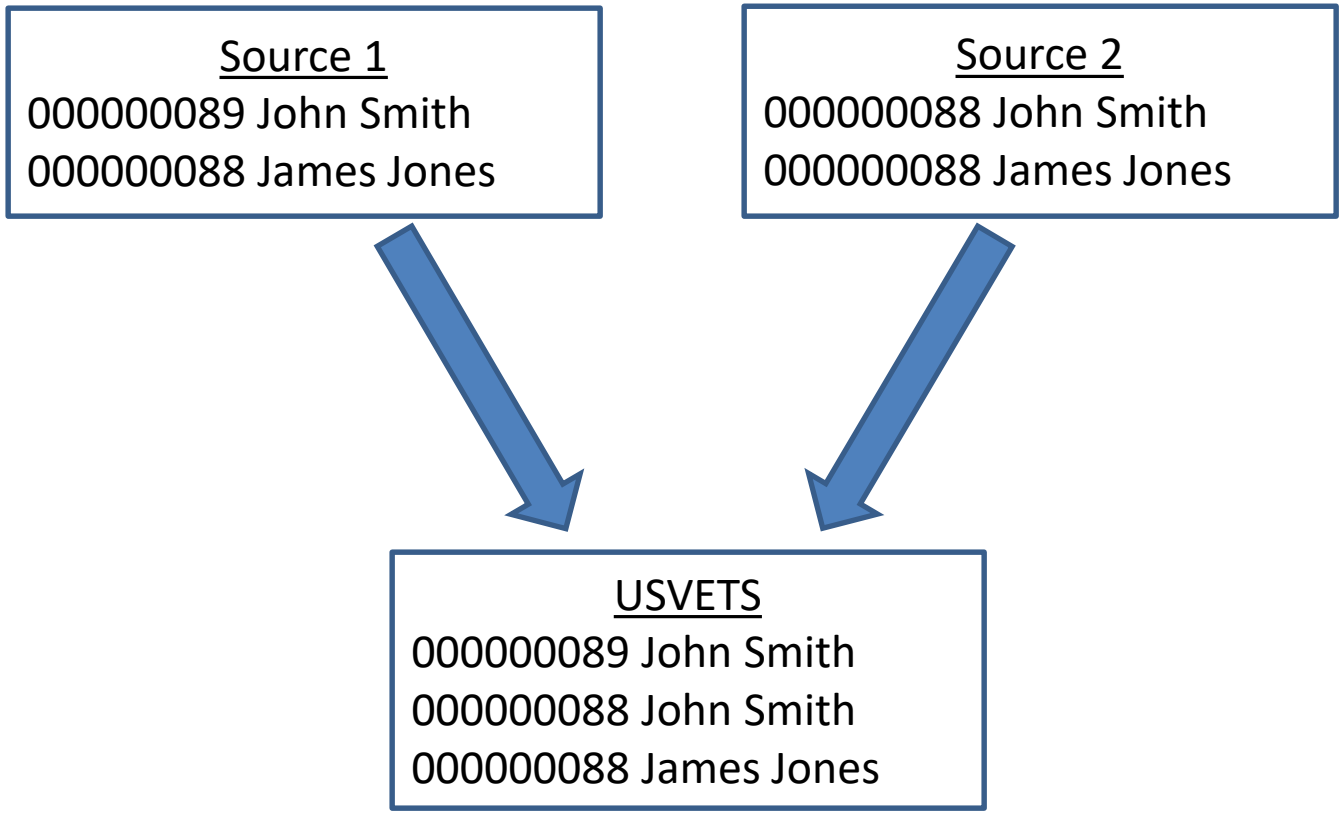
Potential Sources of Name Variation

- Incorrect Data Entry e.g. erroneously typed SSNs
- Misspelled names
- Given Names vs. Nick Names
- Maiden Names vs. Married Names
- Ethnic Names vs. English Names

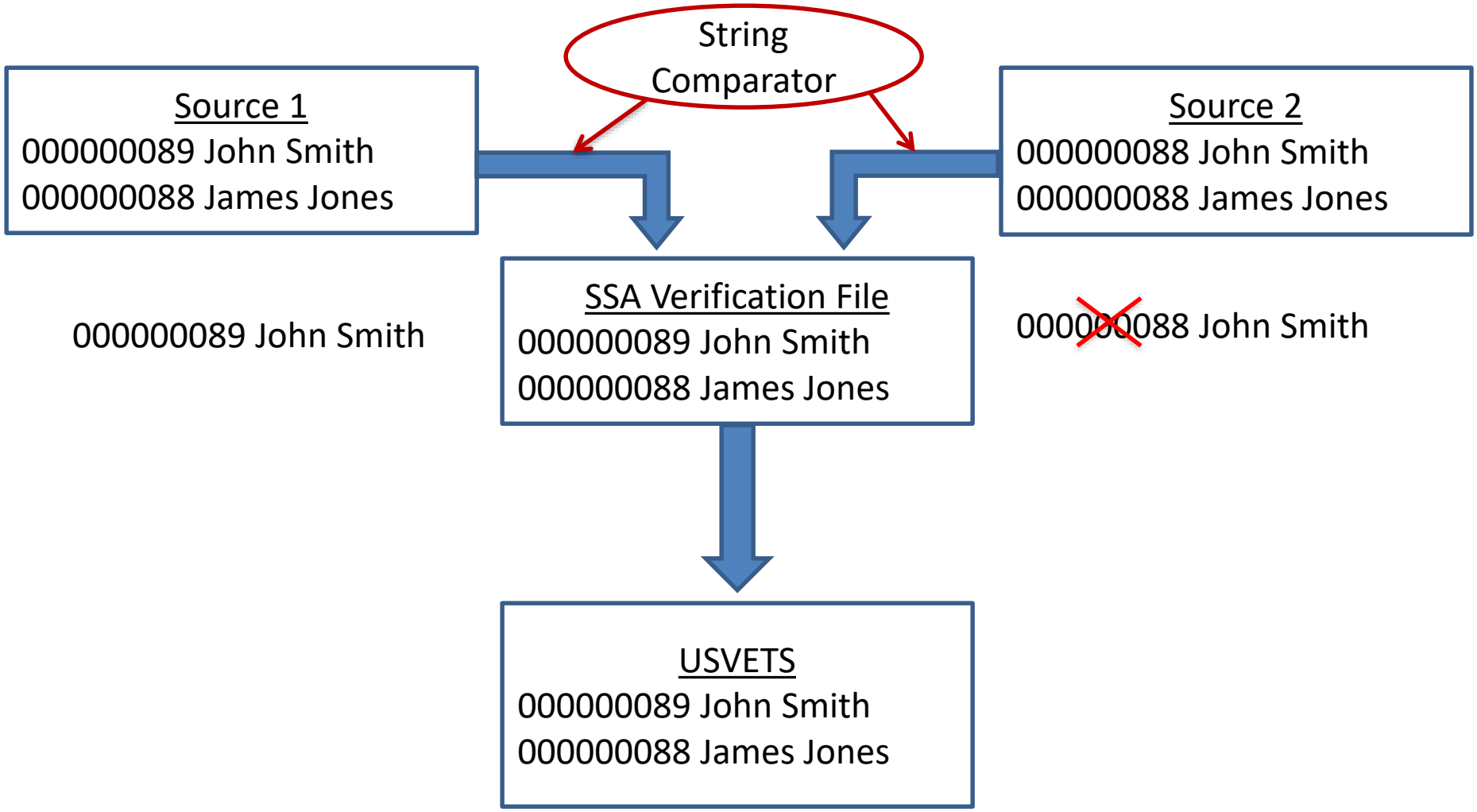
SSA Verification File

- SSN, names, genders, and dates of birth sent to SSA for verification.
- SSA returns the input fields along with verification codes
- Produces the SSAV final dataset that contains SSA validated information, SSN, and best names
- Final dataset includes all necessary variables to validate the SSN for each source.

Record Matching without SSA Verification File



Using SSA Verification to select correct name



Cleaning Source Data

- Over 20 different source data ingested
- All source data are first cleaned and formatted
- Cleaning and Formatting Macro Programs:
  - ✓Remove invalid SSNs and invalid names (e.g., DEMO, DONOT)
  - ✓converts character dates to MM/DD/YYYY
  - ✓converts dollar fields to SAS DOLLAR formats
  - ✓Verifies SSN and name formats
  - ✓Create a suffix name field
  - ✓Clean date fields and convert to SAS date formats

Validation Process

- Verify each SSN and name combination
- If possible, verify gender and date of birth
- Create strings based on name and date-of-birth for source file and SSA Verification File
- Apply sting comparators to compare the two strings for each record

Datasets and Samples

- U.S. Veterans Eligibility Trends and Statistics (USVETS)
- 590,233 Unique Records from Chapter 33 Education Benefits (FY2016) Linked to SSA Verification File
- 13,145,484 Unique Records from Veterans Affairs/Department of Defense Identity Repository (VADIR) Active Component File Linked to SSA Verification File
- A random sample of 1,000 records for each source for classification analysis and ROC Curve Analysis, where string match scores are less than 1.

String Comparators

- Jaro Distance
- Jaro-Winkler
- Levenshtein Edit Distance

Jaro Distance

- $\boldsymbol{d_j} = \frac{\mathbf{1}}{\mathbf{3}}\left(\frac{\boldsymbol{m}}{|\boldsymbol{S_1}|} + \frac{\boldsymbol{m}}{|\boldsymbol{S_2}|} + \frac{\boldsymbol{m-t}}{\boldsymbol{m}}\right)$ [1]
- $|\boldsymbol{S_1}|$ and $|\boldsymbol{S_2}|$ are lengths of $\mathbf{S_1}$ and $\mathbf{S_2}$ respectively;
- *m* is a number of matching characters; and
- t is a number of transpositions.
- Two characters are called matching if the one from the string $\mathbf{S_1}$ agrees with another one from the string S_2 which is located not farther than $\left[\frac{\max\left(|\boldsymbol{S_1}|,|\boldsymbol{S_2}|\right)}{2}\right] - 1$

Jaro-Winkler

- $\boldsymbol{d_w = d_j + (1 - d_j)\left(\frac{m-(p+1)}{|S_1|+|S_2|-2(p-1)}\right)}$ [1]
- $\boldsymbol{d_j}$ is a Jaro Distance
- $\boldsymbol{p}$ is the length of common prefix, up to 4 characters
- Adjust for similar characters, common prefix, and longer string

Levenshtein Edit Distance

- The minimum number of edit steps required to convert one string to the other
- edit steps include insertion, deletion, and substitution.
- $\boldsymbol{x_e = 1 - \frac{e}{n}}$ [1]
- *e* is the edit of the two strings
- *n* is the maximum edit length between two strings

Examples

S1=PHEAKDEYLIM, S2=PHEAKLIM

Jaro Distance:

|S1|=11, |S2|=8, m=8, t=0

$$d_j = \frac{1}{3}\left(\frac{8}{11} + \frac{8}{8} + \frac{8-0}{8}\right) = 0.9091$$

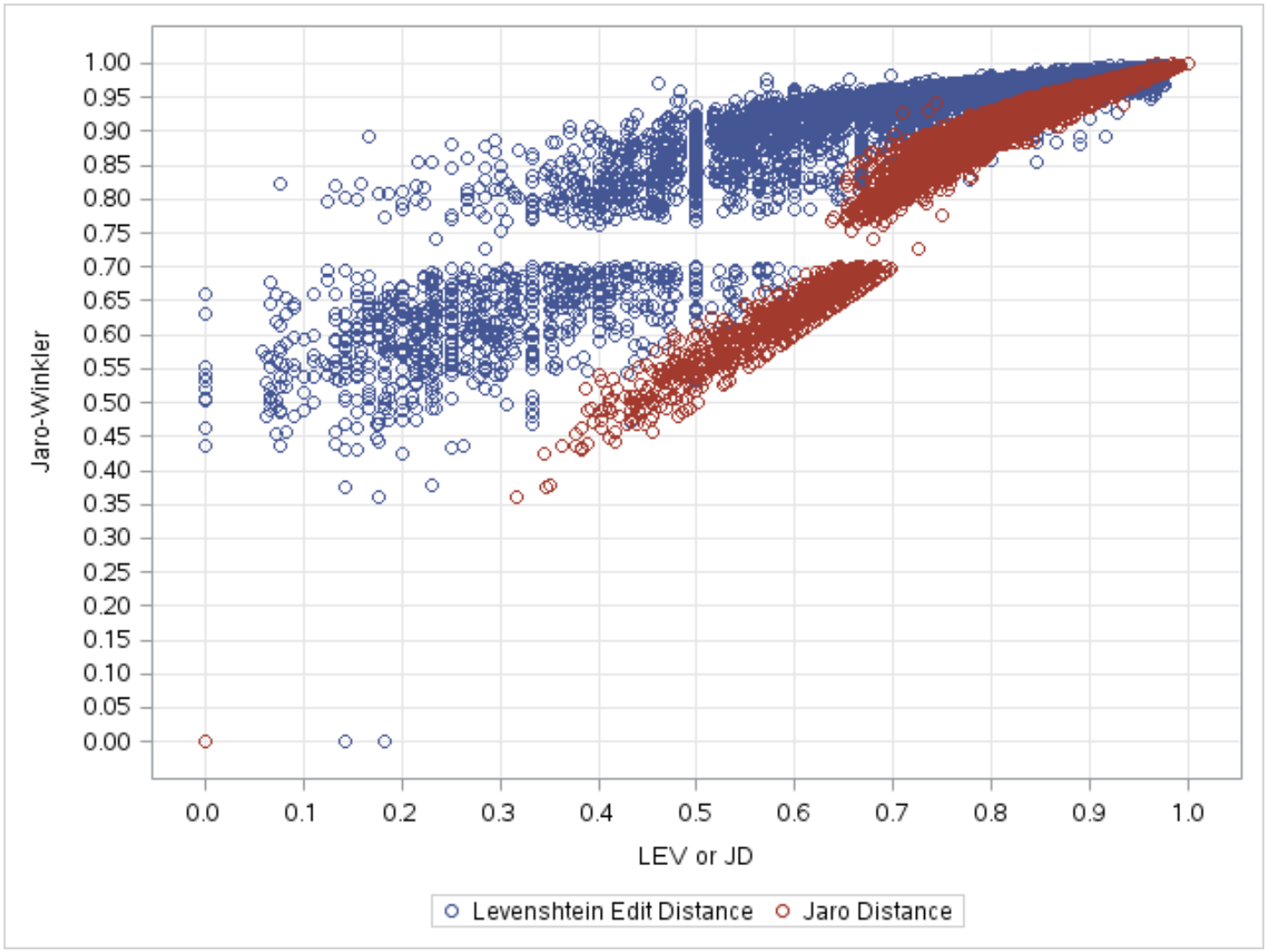
Jaro-Winkler: *p* = 4

$$d_w = 0.9091 + (1 - 0.9091)\left(\frac{8-(4+1)}{11+8-2(4-1)}\right) = 0.9301$$

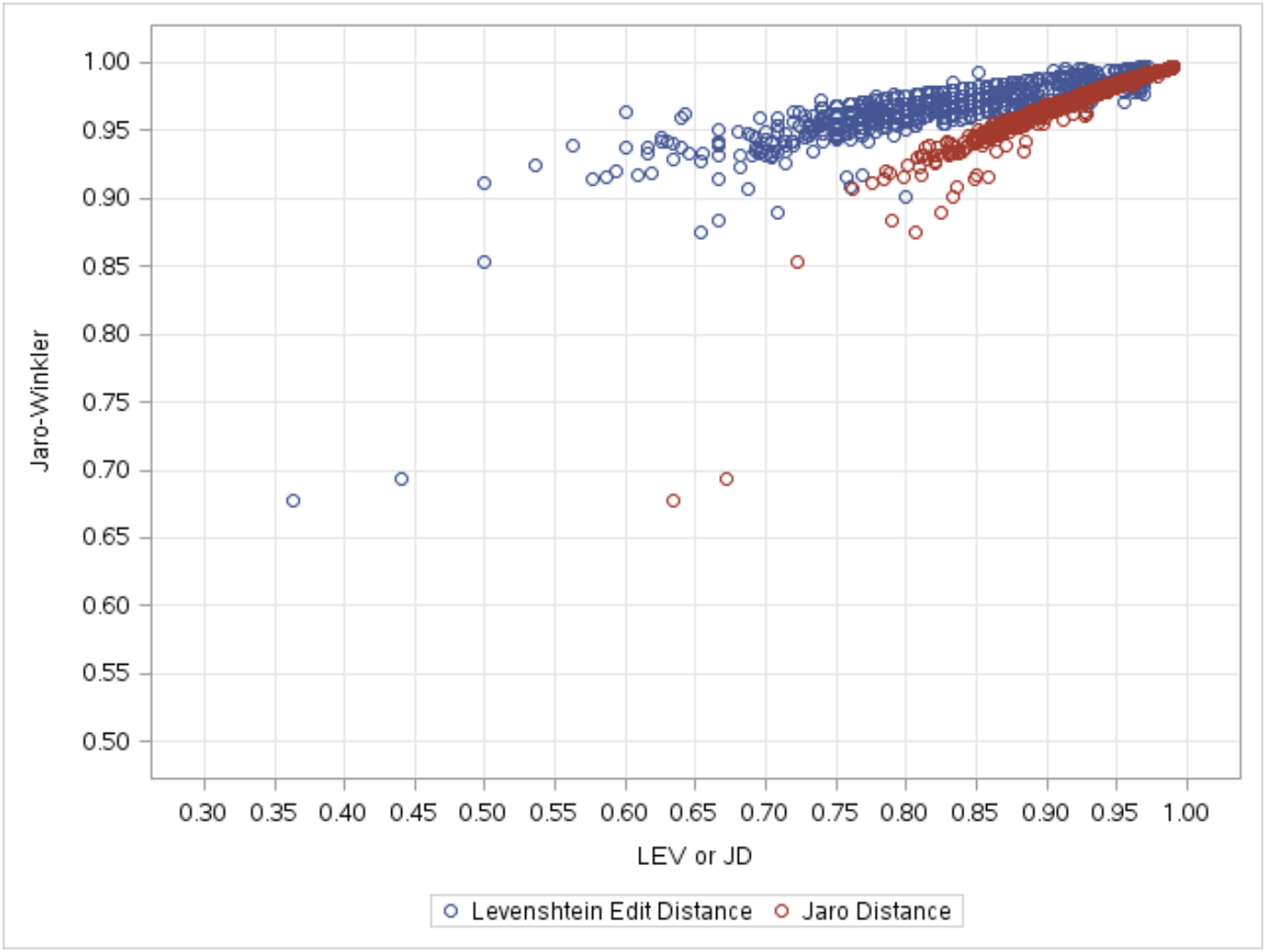
Levenshtein Edit Distance:

$$x_e = 1 - \frac{3}{11} = 0.7273$$

Comparing Jaro-Winkler with Jaro Distance and Levenshtein Distance for Chapter 33



Comparing Jaro-Winkler with Jaro Distance and Levenshtein Distance for VADIR



Classification Table [2]

	Observed Outcome		
Expected Outcome	P (Positive)	N (Negative)	
Matched	TP	FP	PP
Unmatched	FN	TN	PN
	OP	ON	TOT

True Positives (TP)	the number of cases which were correctly classified to be positive
False Positives (FP)	the number of cases which were incorrectly classified as positive (Type I Error)
True Negatives (TN)	the number of cases which were correctly classified to be negative
False Negatives (FN)	the number of cases which were incorrectly classified as negative (Type II Error)
PP / PN	Predictive Positive / Predictive Negative
OP / ON	Observed Positive / Observed Negative
TOT	Total Sample Size

Analysis of Classification Table - Chapter 33

JW

Cutoff= 0.7	P	N
M (T)	986	0
U (F)	9	5
Total	995	5
Accuracy =		0.991

Cutoff= 0.85	P	N
M	985	1
U	5	9
Total	990	10
Accuracy =		0.994

Cutoff= 0.95	P	N
M	923	63
U	0	14
Total	923	77
Accuracy =		0.937

Jaro

Cutoff= 0.7	P	N
M	986	0
U	9	5
Total	995	5
Accuracy =		0.991

Cutoff= 0.85	P	N
M	970	16
U	0	14
Total	970	30
Accuracy =		0.984

Cutoff= 0.95	P	N
M	486	500
U	0	14
Total	486	514
Accuracy =		0.500

LEV

Cutoff= 0.7	P	N
M	960	26
U	2	12
Total	962	38
Accuracy =		0.972

Cutoff= 0.85	P	N
M	604	382
U	0	14
Total	604	396
Accuracy =		0.618

Cutoff= 0.95	P	N
M	365	621
U	0	14
Total	365	635
Accuracy =		0.379

Definition - Sensitivity, Specificity, and 1- Specificity

- Sensitivity = true positive rate = TP/(TP+FN)
- Specificity = true negative rate = TN/(TN+FP)
- 1- Specificity = false positive rate = FP/(FP+TN)

Sensitivity / Specificity Analysis Chapter 33

	JW		Jaro		LEV	
Cutoff	Sensitivity (TPR)	1-Specificity (FPR)	Sensitivity	1-Specificity	Sensitivity	1-Specificity
0.7	0.991	0	0.991	0	0.998	0.684
0.85	0.995	0.1	1	0.533	1	0.965
0.95	1	0.82	1	0.973	1	0.978

Analysis of Classification Table - VADIR

JW

Cutoff= 0.7	P	N
M	990	0
U	8	2
Total	998	2
Accuracy=		0.992

Cutoff= 0.85	P	N
M	990	0
U	8	2
Total	998	2
Accuracy=		0.992

Cutoff= 0.95	P	N
M	905	85
U	2	8
Total	907	93
Accuracy=		0.913

Jaro

Cutoff= 0.7	P	N
M	990	0
U	8	2
Total	998	2
Accuracy=		0.992

Cutoff= 0.85	P	N
M	940	50
U	3	7
Total	943	57
Accuracy=		0.947

Cutoff= 0.95	P	N
M	213	777
U	0	10
Total	213	787
Accuracy=		0.223

LEV

Cutoff= 0.7	P	N
M	944	46
U	5	5
Total	949	51
Accuracy=		0.949

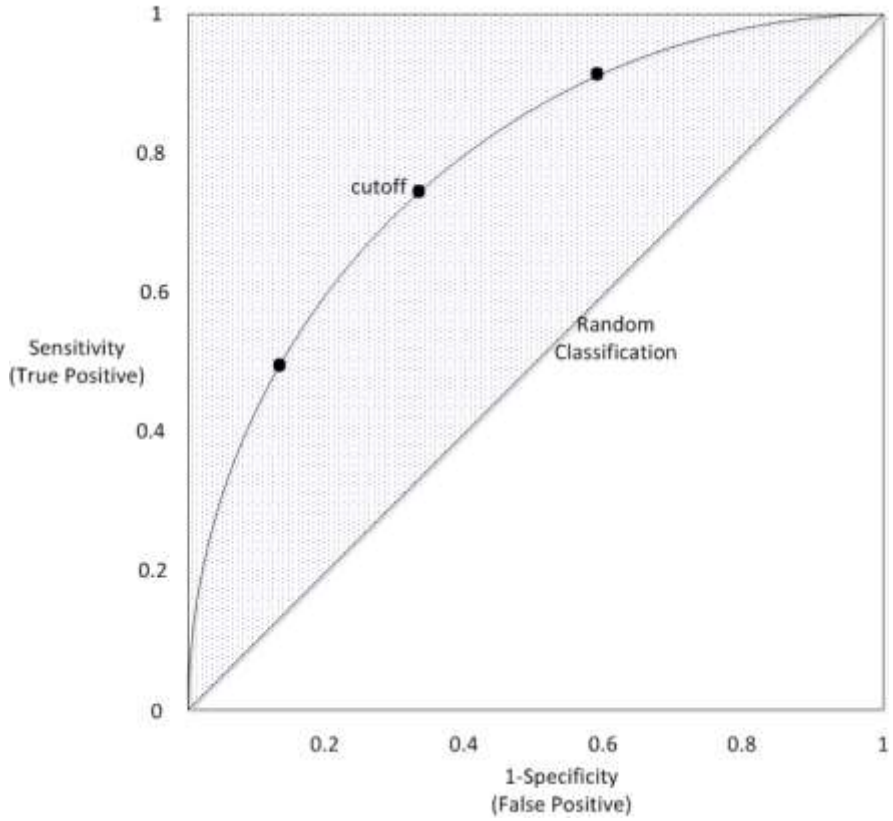
Cutoff= 0.85	P	N
M	393	597
U	2	8
Total	395	605
Accuracy=		0.402

Cutoff= 0.95	P	N
M	111	879
U	0	10
Total	111	889
Accuracy=		0.121

Sensitivity / Specificity Analysis VADIR

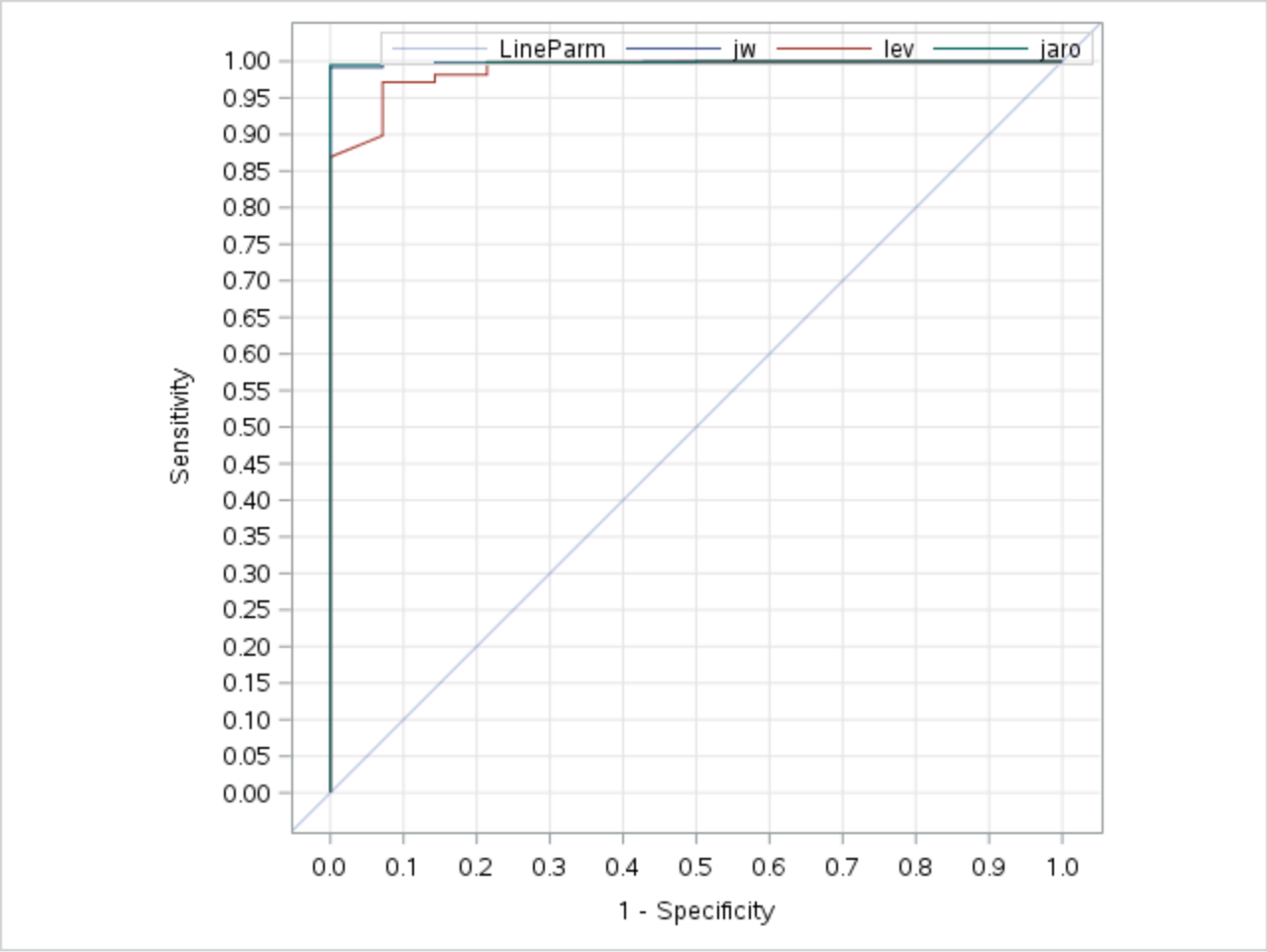
	JW		Jaro		LEV	
Cutoff	Sensitivity	1-Specificity	Sensitivity	1-Specificity	Sensitivity	1-Specificity
0.7	0.992	0	0.992	0	0.995	0.902
0.85	0.992	0	0.997	0.877	0.995	0.9827
0.95	0.998	0.914	1	0.987	1	0.989

Receiver Operating Characteristic (ROC) Curve



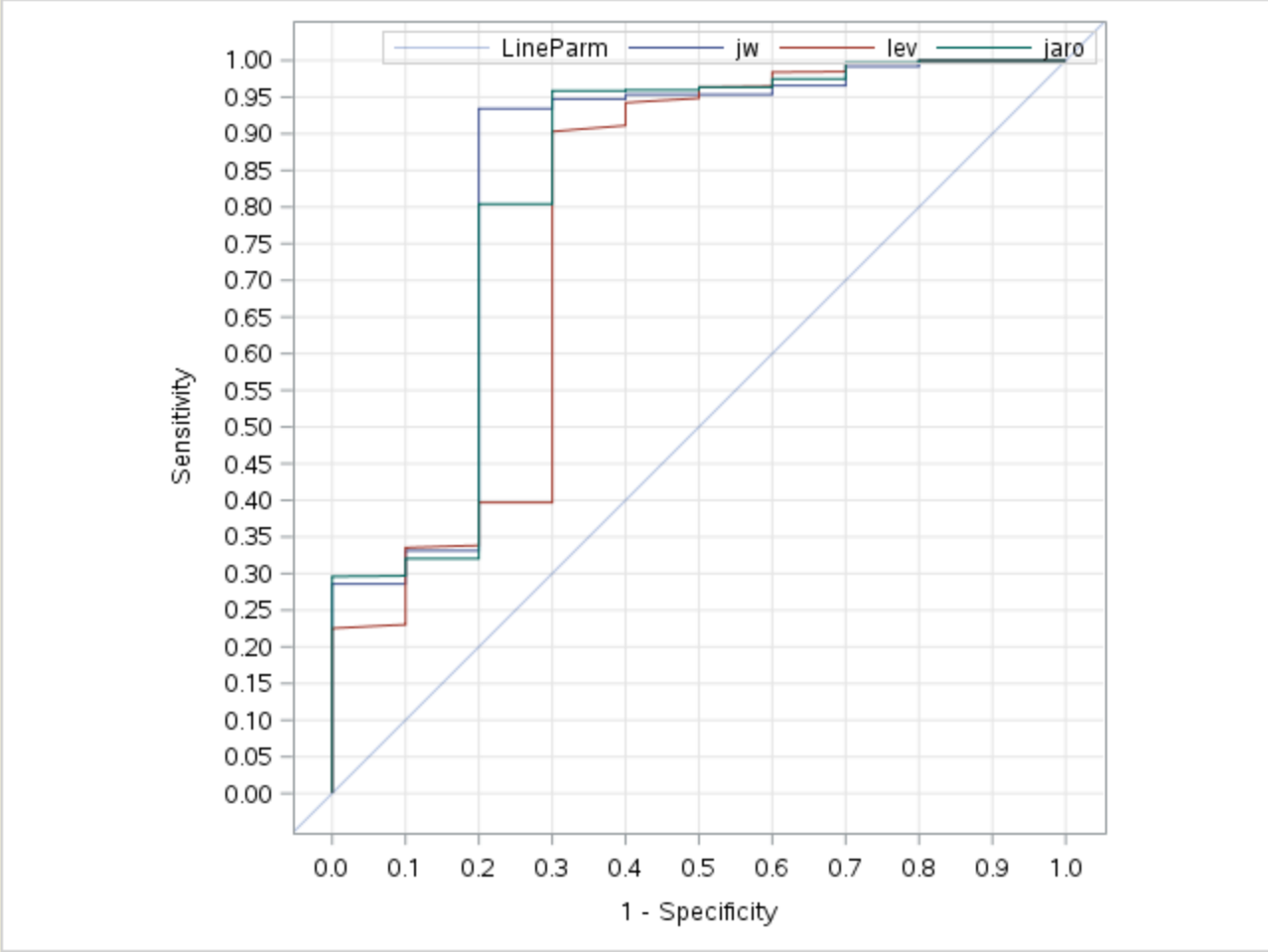
ROC Curve and Area Under Curve - Chapter 33

TYPE	AUC
JW	0.9986
Jaro	0.9988
LEV	0.9881



ROC Curve and Area Under Curve - VADIR

TYPE	AUC
JW	0.836
Jaro	0.827
LEV	0.776



Conclusion and Future Work

- Jaro-Winkler and Jaro Distance performed considerably better than Levenshtein Distance, especially at high cutoff points.
- String comparator can enhance the quality of identity matching over that of solely based on SSN.
- Explore name variations due to ethnic names
- Explore the selection of a threshold that will results in optimal identity matching quality

References

[1] Yancey, W.E. (2005), "Evaluating String Comparator Performance for Record Linkage," research report RRS 2005/05 at http://www.census.gov/srd/www/byyear.html).

[2] http://www.real-statistics.com/descriptive-statistics/roc-curve-classification-table/classification-table/