

ONE-SOURCE MATCHING PROCESS

Test Environment
Match designer
Match passes
Blocking columns
Match commands
Comparison types
Agreement and disagreement weights
Composite weight
Cutoff values
Histogram
Test results table
Weight override
Variable special handling
Matching record types
Pass statistics
Total statistics

Test Environment

Before you can run test match passes, you must define the test environment.

Test Environment

To maximize update efficiency uncheck items whose name, format, and content have not changed...

Sample Information

☒ Data Sample Data Set: /opt/IBM/QUALITYSTAGE/files/block_match_cols_sample.ds ...

☐ Reference Sample Data Set: ...

Frequency Information

☒ Data Frequency Data Set: /opt/IBM/QUALITYSTAGE/files/block_match_cols_freq_sample.ds ...

☐ Reference Frequency Data Set: ...

☒ Maximum Frequency Value: 100

Test Results Database

Using Data Connection Object: (none)

ODBC Data Source Name: MDDb

Username: db2inst1

Password:

Load Test Save Clear

Execution Environment

Configuration File Location: /opt/IBM/InformationServer/Server/Configurations/OneLogicalNode.apf ...

Environment Variables: (use default)

Update Cancel Help

Match Designer

The Match Designer is a tool for performing the iterative process of defining a match, testing it against sample data, viewing results and statistics, and then fine-tuning the match.

Match Designer - Specification: msTest_20220610

Compose Total Statistics

Save Add Pass Remove Pass Test All Passes Configure Specification Show Pass Holding Area

Match Type: One-source Dependent

Match Pass Holding Area

Match Pass: psVATPerson

Pass Definition Pass Statistics

Blocking Columns: CIVAT_IIS

Match Commands: EnQLast_Name, EnPrLast_Name, EnPrFather_Name, CIVAT_IIS, Fixphone_IIS, Cellphone_IIS, PrCustomer_Type_IIS_changed, QIPassport_IIS, QIEmail_Uppercase, Billing_TerraKey

Cutoff Values: Match: 20, Clerical: 7

Test Results:

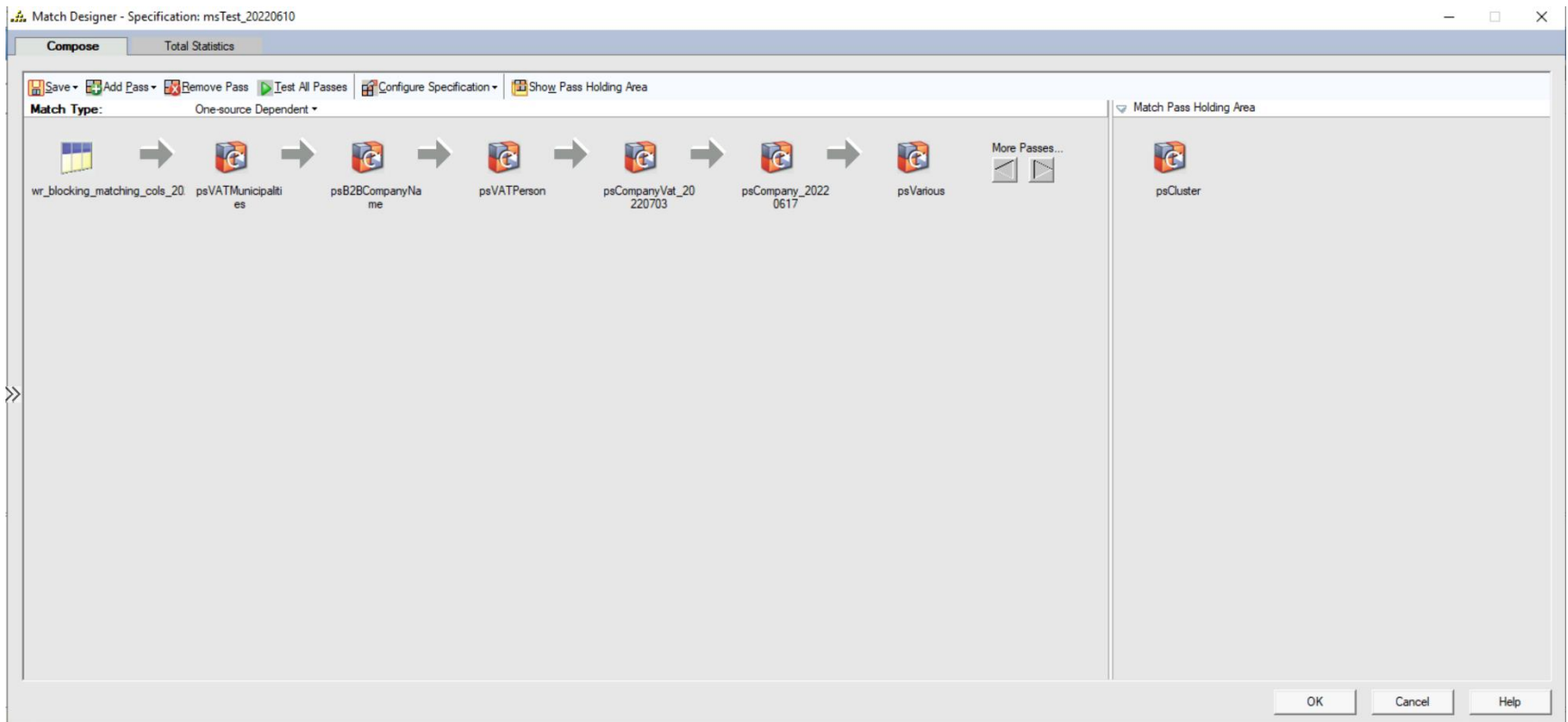
Business_Partner_ID	SetID	Record Type	Weight	DataID	EnPrFirst_Name	EnQLast_Name	EnPrLast_Name	EnPrFather_Name	CIVAT_IIS	Fixphone_IIS	Cellpho
1100007559	4	MP	76.41	4	ANASTASIOS	KARAPOSTOLIS	KARAPOSTOLIS	EMANUIL	010833910	2310456290	
	4	DA	38.66	5299	ANASTASIOS	KARAPOSTOLIS	KARAPOSTOLIS		010833910	2310456290	6
1100013578	25	MP	78.86	25	TAMARA	EMINIDU TAMARA	EMINIDU	ALEKSIOS	065468676	2310440859	
1100013593	25	DA	47.07	7474	TAMARA	EMINIDU TAMARA	EMINIDU		065468676	2310828013	
1100013585	25	DA	78.86	5014	TAMARA	EMINIDU TAMARA	EMINIDU	ALEKSIOS	065468676	2310440859	
1100013671	93	MP	68.91	93	STILIANOS	MARONIDIS	MARONIDIS		036466172	2310449501	6
	93	DA	46.72	6315	STILIANOS	MARONIDIS	MARONIDIS		036466172	2310449501	6
1100009427	109	MP	65.95	109	ATHANASIOS	PAPATHANASIU	PAPATHANASIU	HRISOSTOMOS	044739086		
1100009439	109	DA	65.95	9990	ATHANASIOS	PAPATHANASIU	PAPATHANASIU	HRISOSTOMOS	044739086		
1103967543	119	MP	80.26	119	STERGIOS	DIAMADIS	DIAMADIS	ATHANASIOS	116517809	2310566457	6
1103967543	119	DA	80.26	10603	STERGIOS	DIAMADIS	DIAMADIS	ATHANASIOS	116517809	2310566457	6
1100013764	137	MP	57.55	137	IOANIS	HARISIU IOANIS	HARISIU	NIKOLAOS	032320801	2310431113	

OK Cancel Help

Match passes

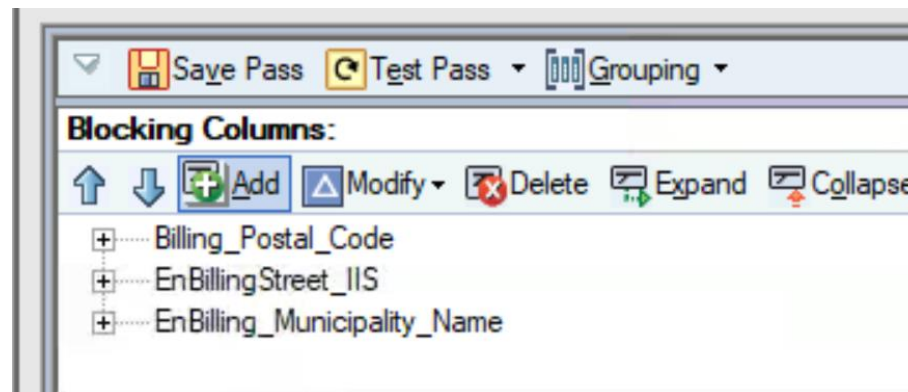
You define the columns to compare and how to compare them.

You also define the criteria for creating blocks.



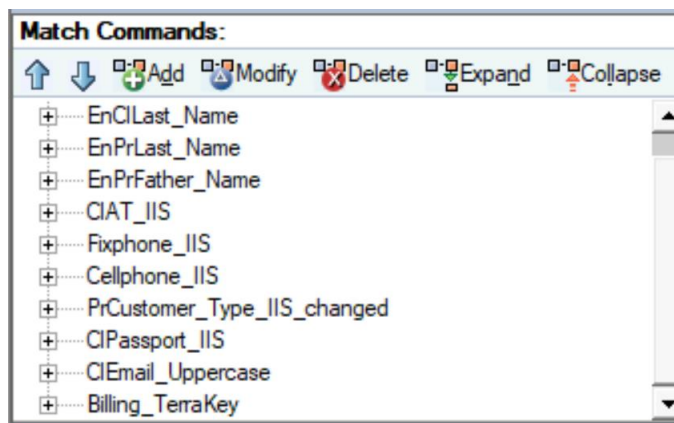
Blocking columns

Blocking columns are used to **create subsets or blocks** of input data records that are likely to be associated. The records that have the **same values** in the blocking columns are compared to only one another. Blocks make the **matching process faster and more efficient**.



Match Commands

Matching columns establish how records within the block are compared.



Match Command Window

Match Command

Specify comparison type and match column(s)...

Name: CIVAT_IIS

Available Comparison Types:
CNT_DIFF - Counting errors in fields

Available Data Columns:

Name	Type	Description
Row	Integer	
SOURCE	VarChar	
CIVAT_IIS	NVarChar	
CIAT_IIS	NVarChar	
CIDOY_IIS	NVarChar	
CIPassport_IIS	NVarChar	
CIEmail_Uppercase	NVarChar	
Cluster_CAddressQ3Sim09	Integer	
Cluster_CNameQ3Sim09	Integer	
Contract_Account_ID	NVarChar	
Billing_Postal_Code	NVarChar	
Billing_TerraKey	NVarChar	
EnBilling_Address_Region	NVarChar	
EnBilling_Street	NVarChar	
EnBilling_Municipality_Nam	NVarChar	
EnTrBilling_Region_Name	NVarChar	
EnTrBilling_Street	NVarChar	
TrBilling_Postal_Code	NVarChar	
EnCICompany_Name	NVarChar	

☒ Columns
☐ Vectors

☐ Reverse

Selected Column(s):

Source	Name
Data	CIVAT_IIS

Command Options

m-prob: .9

u-prob: .01

Param 1: 1.

[Weight Overrides](#)

OK Cancel Help

Comparison Types

CHAR: Compares two strings on a character-by-character basis. If one string is shorter it pads the shorter column with trailing blanks.

CNT_DIFF: Compares string of numbers. Param 1 indicates the number of differences that will be tolerated.

UNCERT: Evaluates the similarity of two character strings by using an algorithm that is based on information theory principles.

MULT_ALIGN: Scores the similarity of two sequences of terms.

- Similarity of the terms
- Order of similar terms in their original sequence
- Proximity of similar terms in their original sequence

MULT_UNCERT: Compares all words using a string comparison algorithm based on information theory principles.

NAME_UNCERT: Truncate the longer string.

David, Dav

Comparison Types

Required Parameter

The following parameter is required:

- **Param 1.** The cutoff threshold, which is a number 0 - 900.
 - 900. The two strings are identical.
 - 850. The two strings can be safely considered to be the same.
 - 800. The two strings are probably the same.
 - 750. The two strings are probably different.
 - 700. The two strings are almost certainly different.

M prob

Optional parameter

M probability reflects the importance and reliability of the data in a column.

The valid range is greater than or equal to 0.0001-0.9999.

Default value: 0.9

Examples of m probability values:

- For most columns, use the default value
- For highly important columns, use value 0.999
- For moderately important columns, use value 0.95
- For columns with poor reliability (such as street direction), use value 0.8

U prob

Optional parameter

U probability reflects the probability that the data in a column accidentally agrees.

The valid range is greater than or equal to 0.0001-0.9999.

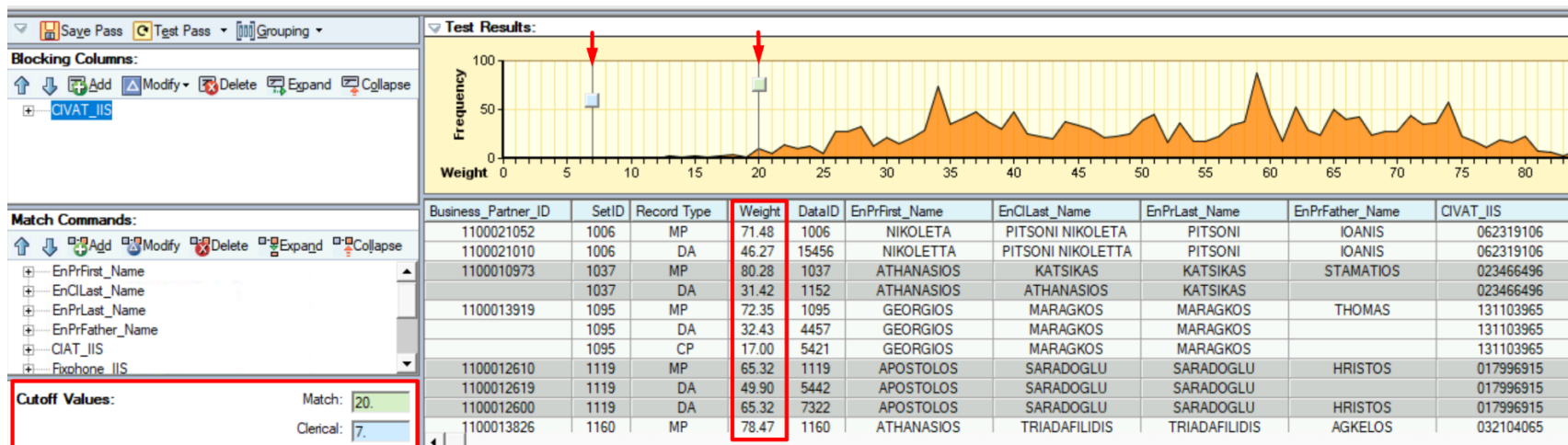
Default value: 0.01

Examples of u probability values:

- For age, use 0.02
- For gender, use 0.5

Matching record types

Match (MP)	Master record
Clerical (CP)	The duplicates that fall in the <u>clerical range</u> .
Duplicate (DA)	The duplicate records that are <u>above the match cutoff</u> .
Residuals or nonmatched (RA)	The records that are not master, duplicate, or clerical records.



Composite weights

For each record pair that you want to compare, a composite weight is computed.

The *composite weight* is the sum of the individual weights for all the match comparisons. It is the measure of confidence that the two records are a match.

Weight Comparison

Match command names:	EnPrFirst_Name	EnCILast_Name	EnPrLast_Name	EnPrFather_Name	CIAT_IIS	Fixphone_IIS
Data column names:	EnPrFirst_Name	EnCILast_Name	EnPrLast_Name	EnPrFather_Name	CIAT_IIS	Fixphone_IIS
Variable special handling for this data column:	CRITICAL MISSINGOK	CRITICAL MISSINGOK	CRITICAL MISSINGOK	CRITICAL MISSINGOK	NOFREQ	NOFREQ
Match comparison type:	NAME_UNCERT	MULT_ALIGN	MULT_UNCERT	NAME_UNCERT	UNCERT	MULT_UNCERT
Data importance and reliability [m-prob]:	0.9	0.9	0.9	0.9	0.9	0.9
Probability of accidental agreement [u-prob]:	0.01	0.01	0.01	0.01	0.01	0.01
Parameter / mode settings for this comparison:	850	850	850	850	850	850
Weight Overrides...	[none]	[none]	[none]	[none]	[none]	[none]
Replace weights with these values:	-	-	-	-	-	-
Add / subtract these values to / from weights:	-	-	-	-	-	-
Scale weights based on these values:	-	-	-	-	-	-
Weight comparison master record with a weight of 76.41:	ANASTASIOS	KARAPOSTOLIS	KARAPOSTOLIS	EMANUIL	Φ150216	2310456290
Duplicate record with a composite weight of 38.66:	ANASTASIOS	KARAPOSTOLIS	KARAPOSTOLIS			2310456290
Contribution made by this column to the composite weight:	7.21	4.95	12.81	0.00	0.00	6.49
Default agreement / disagreement weights:	7.21 / -3.31	13.89 / -3.32	12.81 / -3.32	0.16 / -0.98	[not available]	[not available]

☒ Show Match Definition

Agreement and disagreement weights

For each match comparison, the matching process calculates an agreement weight and a disagreement weight.

- The agreement weight is a positive value.
- The disagreement weight is a negative value.
- Agreement weights add to the composite weight, and disagreement weights subtract from the composite weight.
- The higher the score is; the greater the agreement is.
- *Partial weight* is assigned for non-exact or fuzzy matches.
- Missing values have a default weight of zero.

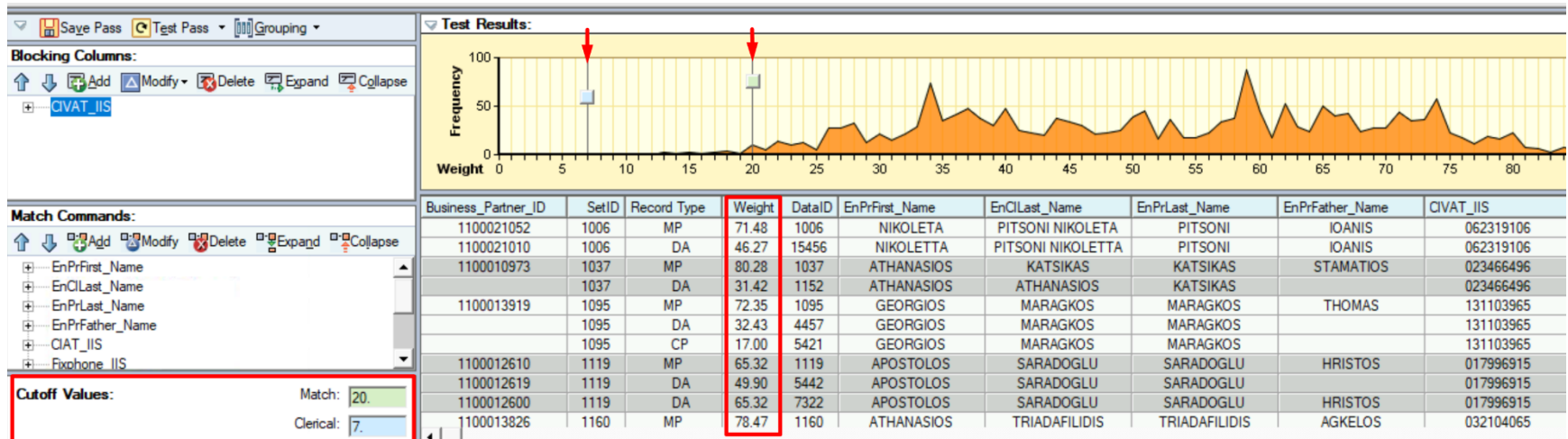
Weight Comparison

Match command names:	EnPrFirst_Name	EnCLast_Name	EnPrLast_Name	EnPrFather_Name	CIAT_IIS
Data column names:	EnPrFirst_Name	EnCLast_Name	EnPrLast_Name	EnPrFather_Name	CIAT_IIS
Variable special handling for this data column:	CRITICAL MISSINGOK	CRITICAL MISSINGOK	CRITICAL MISSINGOK	CRITICAL MISSINGOK	NOFREQ
Match comparison type:	NAME_UNCERT	MULT_ALIGN	UNCERT	NAME_UNCERT	UNCERT
Data importance and reliability [m-prob]:	0.9	0.9	0.9	0.9	0.9
Probability of accidental agreement [u-prob]:	0.01	0.01	0.01	0.01	0.01
Parameter / mode settings for this comparison:	700	850	800	700	850
Weight Overrides...	[none]	[none]	[none]	[none]	[none]
Replace weights with these values:	-	-	-	-	-
Add / subtract these values to / from weights:	-	-	-	-	-
Scale weights based on these values:	-	-	-	-	-
Weight comparison master record with a weight of 86.49:	ATHANASIA	DURTOGLU	DURTOGLU	ATHANASIOS	AK876028
Duplicate record with a composite weight of 41.05:	ATH	DURTOGLU ATH	DURTOGLU		
Contribution made by this column to the composite weight:	8.58	7.05	18.92	0.00	0.00
Default agreement / disagreement weights:	18.88 / -3.32	19.77 / -3.32	18.92 / -3.32	1.00 / -2.46	[not available]

☒ Show Match Definition

Cutoff Values

Match and clerical cutoffs are thresholds that determine how to categorize scored record pairs.



Cutoff Values

Record pairs with composite weights equal to or greater than the match cutoff are considered matches.

Record pairs with composite weights equal to or greater than the clerical cutoff but less than the match cutoff are called *clerical pairs*.

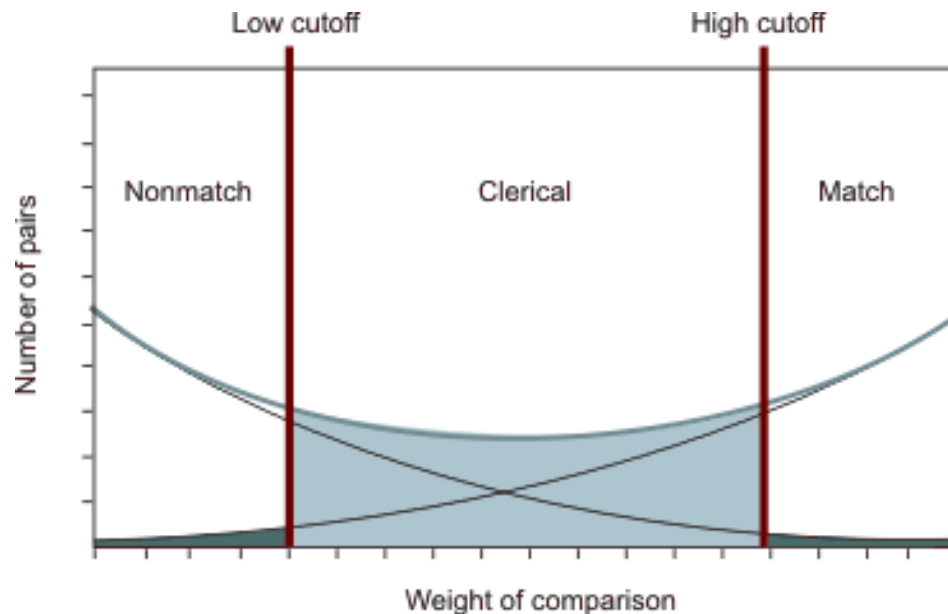
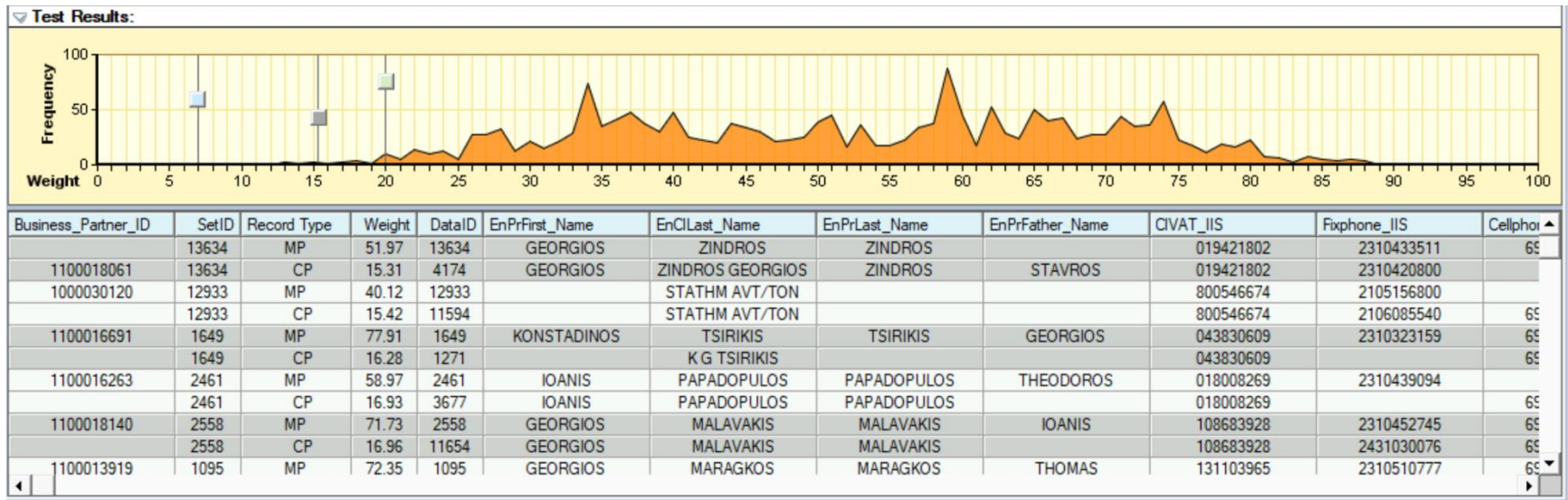


Figure 1. Histogram of weights

Histogram

The Frequency/Weight histogram at the top of the Test Results pane is a graphical representation of the distribution of weights assigned by the run of a match pass.

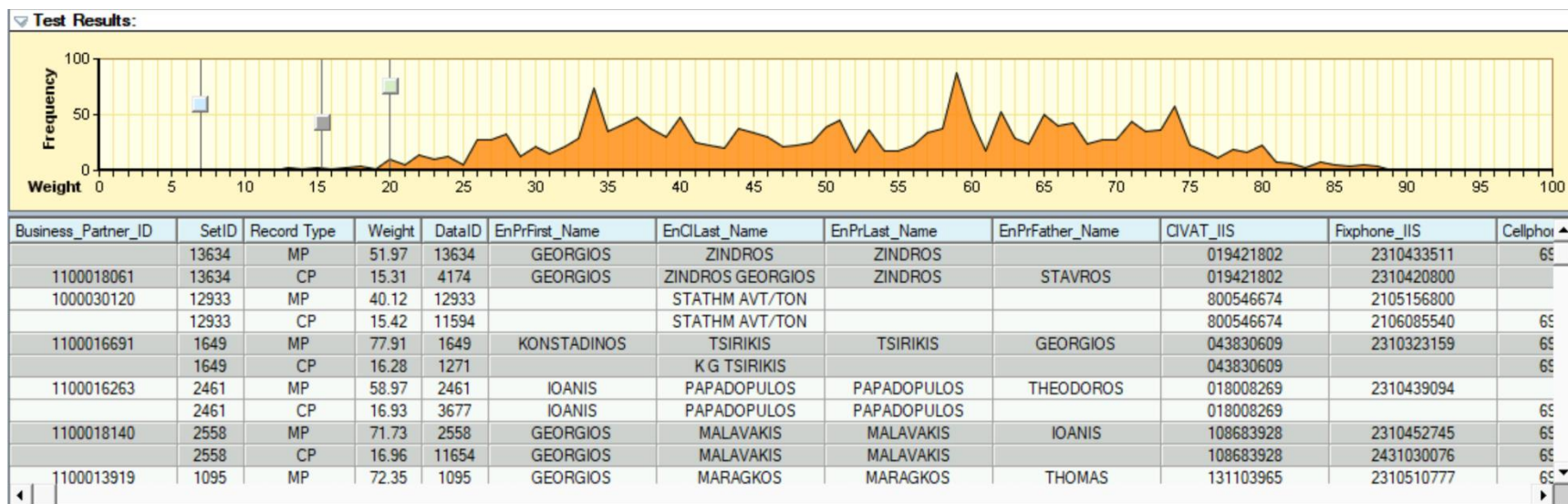


Test results table

Displays rows of data initially ordered in match sets by **SETID**.

Tasks:

- Search for information in the table.
- Sort and group the results in a variety of ways.
- View column details.
- Compare weights.
- Change the column display.



Master record selection and group construction

- Match processes one block of records at a time.
- Each record in a block of records is compared to every other record in the block.
- Each master record is used to create a group.

Process:

1. All pairs of records within a block are scored.
2. From the records that have not been added to a group, the record pair with the highest composite weight over the cutoff thresholds is selected.
3. The record from the pair with the highest score when compared to itself is designated as the master record.
4. Other records are then added to create the group:
 $\text{weight} < \text{match cutoff} = \text{match duplicate}$
 $\text{clerical cutoff} < \text{weight} > \text{match cutoff} = \text{clerical duplicate}$
5. The process is repeated within the block until there are no remaining pairs of ungrouped records whose weight is above either of the cutoffs.

Weight Override

Weight Overrides



Enter one or more values for each weight override...

Compose Weight Override☒ Replace☐ Add☐ ScaleAgreement Weight [AW]: Disagreement Weight [DW]: Data Source Missing Weight [AM]: Reference Source Missing Weight [BM]: Both Missing Weight [XM]: Conditional Data Source Value [AV]: Conditional Reference Source Value [BV]:

Add Override

Summary of Weight Overrides

A/R/S	AV	BV	AW	DW	AM	BM	XM
Add			20.				

Delete Override

OK

Cancel

Help

Weight Override (cont.)

The Weight Overrides window lets you change the calculated weights for missing value situations or for specific combinations of values for a particular match command.

- **Replace.** Select to replace the weight calculated for the column or columns with a weight that you specify.
- **Add.** Select to add the weight that you are specifying to the weight calculated for this column or columns.
- **Scale.** Like **Replace** override, but **Scale** also preserves probabilistic scoring.
- **Agreement Weight (AW).** The agreement weight if the values for the column agree and are not missing.
- **Disagreement Weight (DW).** The disagreement weight if the values for the column disagree and are not missing.
- **Data Source Missing Weight (AM).** The weight when the value on the data record is missing.
- **Reference Source Missing Weight (BM).** The weight when the value on the reference record is missing.
- **Both Missing Weight (XM).** The weight when values are missing on both records.
- **Conditional Data Source Value (AV).** Enter the value, enclosed in single quotation marks ('), that is expected in a column on the data source or the word ALL.

Variable Special Handling (global configuration)

Variable Special Handling

Select the desired Actions and corresponding Columns that require special handling...

Define Special Handling

Actions: (click arrow to view) Table Definition Name: **wr_blocking_matching_cols_20220930_2**

Available Data Columns:

Name	Type	Description
Row	Integer	
SOURCE	VarChar	
CIVAT_IIS	NVarChar	
CIAT_IIS	NVarChar	
CIDOY_IIS	NVarChar	
CIPassport_IIS	NVarChar	
CIEmail_Uppercase	NVarChar	
Cluster_CAddressQ3Sim09	Integer	
Cluster_CNameQ3Sim09	Integer	
Contract_Account_ID	NVarChar	
Billing_Postal_Code	NVarChar	

Column1:

Summary of Special Handling

Action	Source	Column Name(s)
CRITICAL MISSINGOK	Data	B2B_TerraKey
CRITICAL MISSINGOK	Data	CIVAT_IIS
NOFREQ	Data	CIEmail_Uppercase
NOFREQ	Data	CIVAT_IIS_copy
CRITICAL	Data	PrCustomer_Type_IIS

Variable Special Handling (cont.)

- **CLERICAL**: Use when you want a disagreement on the column to cause the record pair to be considered a clerical pair, even if the composite weight is above the match cutoff.
- **CLERICAL [MISSINGOK]**: Use MISSINGOK if a missing value probably is not the cause of the record pair being considered to be forced into clerical review.
- **CRITICAL**: Used when a disagreement on the column causes the record pair to automatically be considered a nonmatch.
- **CRITICAL [MISSINGOK]**: Use MISSINGOK if it is acceptable that one or both values are missing.
- **NOFREQ**: Typically use when a column has high cardinality, such as a national identification number. NOFREQ indicates that no frequency analysis must be performed.
- **CONCAT**: Use when you want to concatenate columns to form one frequency count.

Default Handling for Missing Weights (global configuration)



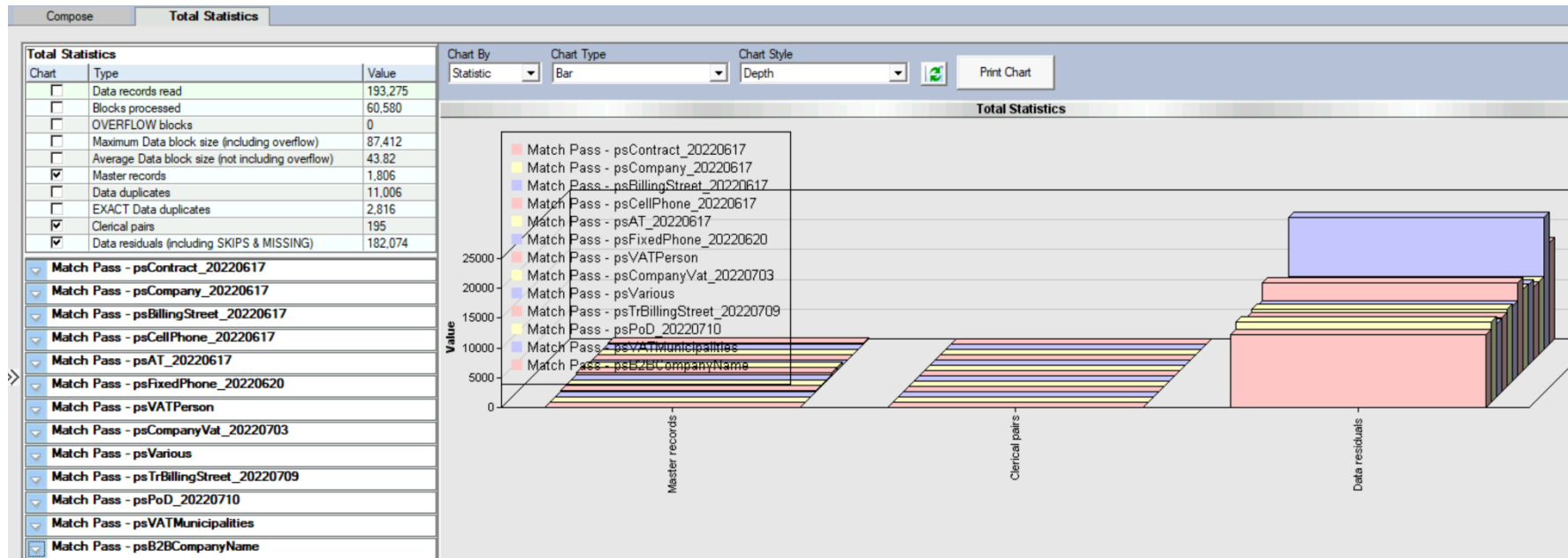
Pass Statistics

This tab displays the statistics about the current test, the statistics about a baseline run (if you created one), and a graphical illustration of the pass statistics.

- Data records read
- Blocks processed
- Average block size (overflow blocks not included)
- OVERFLOW blocks
- Maximum Datablock size(including overflow)
- Master records
- Data duplicates
- Exact duplicate records
- Clerical pairs
- Data residuals(including SKIPS & MISSING)
- Total number of comparisons performed

Total Statistics tab

This tab displays the results of each match-pass test run and the combined results of all the match passes.



END ONE-SOURCE MATCHING PROCESS