Missing Values Compensation in Duplicates Detection Using Hot Deck

Abdulrazzak Ali (dowsan1@yahoo.com)
University of Aden https://orcid.org/0000-0002-8004-6791
Nurul A. Emran
Universiti Teknikal Malaysia Melaka Fakulti Teknologi Maklumat dan Komunikasi
Siti A. Asmai
Universiti Teknikal Malaysia Melaka Fakulti Teknologi Maklumat dan Komunikasi

Research

Keywords: Duplicates Detection, Incomplete Data Set, Clustering, Sorting Key, Compensation Method

Posted Date: April 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-390519/v1

Version of Record: A version of this preprint was published at Journal of Big Data on August 21st, 2021.
See the published version at https://doi.org/10.1186/s40537-021-00502-1.

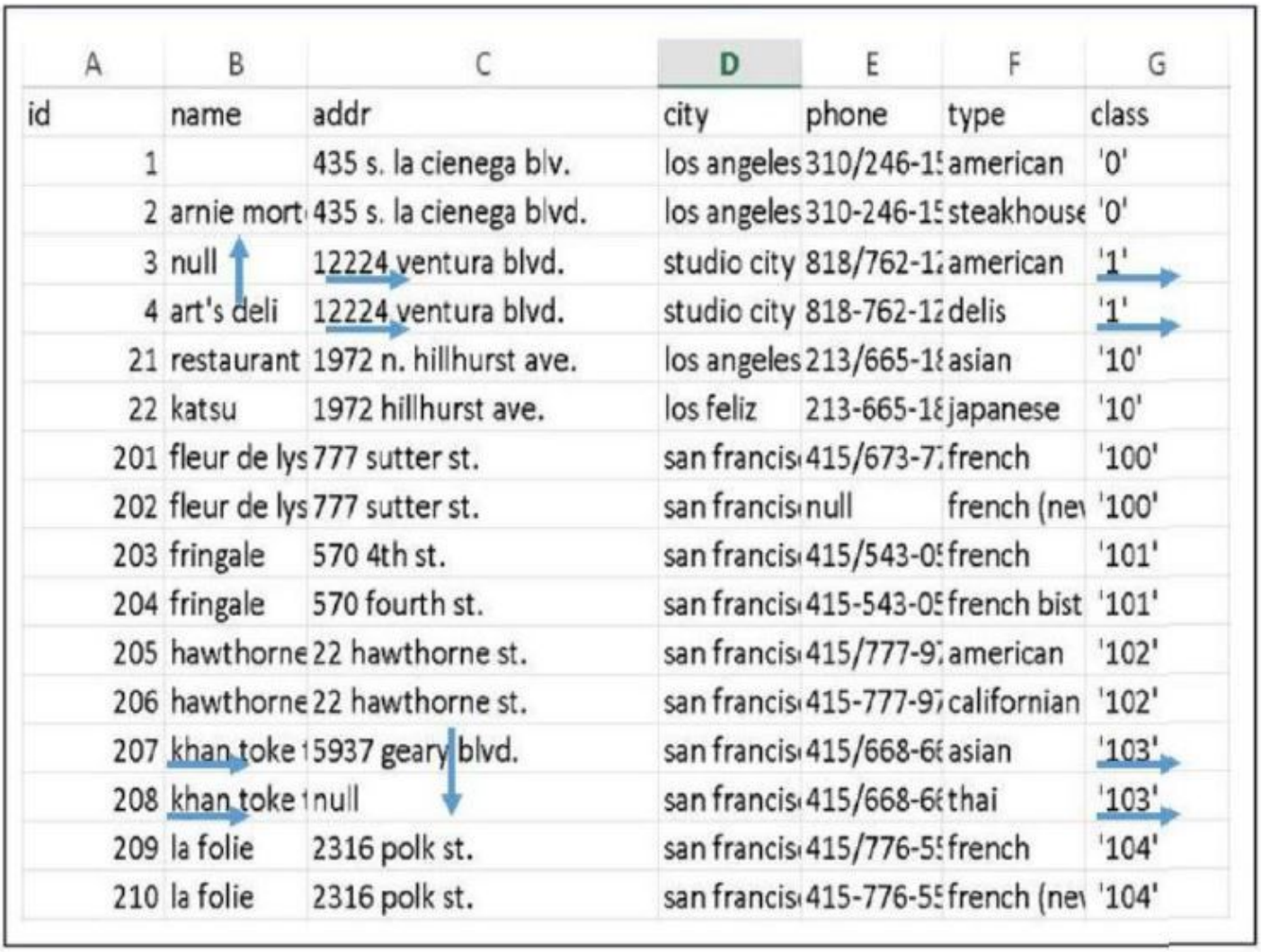
Abstract

Duplicate record is a known problem within the datasets especially within databases of huge volumes. The accuracy of duplicates detection determines the efficiency of the duplicates removal process. Unfortunately, the effort to detect duplicates becomes more challenging due to the presence of missing values within the records. This is because, during the clustering and matching process, missing values can cause records that are similar to be assigned in a wrong group, causing the duplicates left undetected. In this paper, we present how duplicates detection can be improved even though missing values are present within a data set using our Duplicates Detection within the Incomplete Data set (DDID) method. We hypothetically add the missing values to the key attributes of two datasets under study, using an arbitrary pattern to simulate both complete and incomplete data sets. We analyze the results to evaluate the performance of duplicates detection using the Hot Deck method to compensate for the missing values in the key attributes. We hypothesize that by using Hot Deck, there is a performance improvement in duplicates detection. The performance of the DDID is compared with an early duplicates detection method (called DuDe) in terms of its accuracy and speed. The findings of the experiment show that, even though the data sets are incomplete, DDID is capable to offer better accuracy and faster duplicates detection as compared to a benchmark method (called DuDe). The results of this study contribute to duplicates detection under incomplete data sets constraint.

Full-text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures



A	B	C	D	E	F	G
id	name	addr	city	phone	type	class
1		435 s. la cienega blv.	los angeles	310/246-1!	american	'0'
2	arnie mort	435 s. la cienega blvd.	los angeles	310-246-1!	steakhouse	'0'
3	null	12224 ventura blvd.	studio city	818/762-1:	american	'1'
4	art's deli	12224 ventura blvd.	studio city	818-762-1:	delis	'1'
21	restaurant	1972 n. hillhurst ave.	los angeles	213/665-1!	asian	'10'
22	katsu	1972 hillhurst ave.	los feliz	213-665-1!	japanese	'10'
201	fleur de lys	777 sutter st.	san francis	415/673-7:	french	'100'
202	fleur de lys	777 sutter st.	san francis	null	french (ne	'100'
203	fringale	570 4th st.	san francis	415/543-0!	french	'101'
204	fringale	570 fourth st.	san francis	415-543-0!	french bist	'101'
205	hawthorne	22 hawthorne st.	san francis	415/777-9:	american	'102'
206	hawthorne	22 hawthorne st.	san francis	415-777-9:	californian	'102'
207	khan toke	5937 geary blvd.	san francis	415/668-6!	asian	'103'
208	khan toke	null	san francis	415/668-6!	thai	'103'
209	la folie	2316 polk st.	san francis	415/776-5!	french	'104'
210	la folie	2316 polk st.	san francis	415-776-5!	french (ne	'104'

Figure 1

An example of missing values compensation

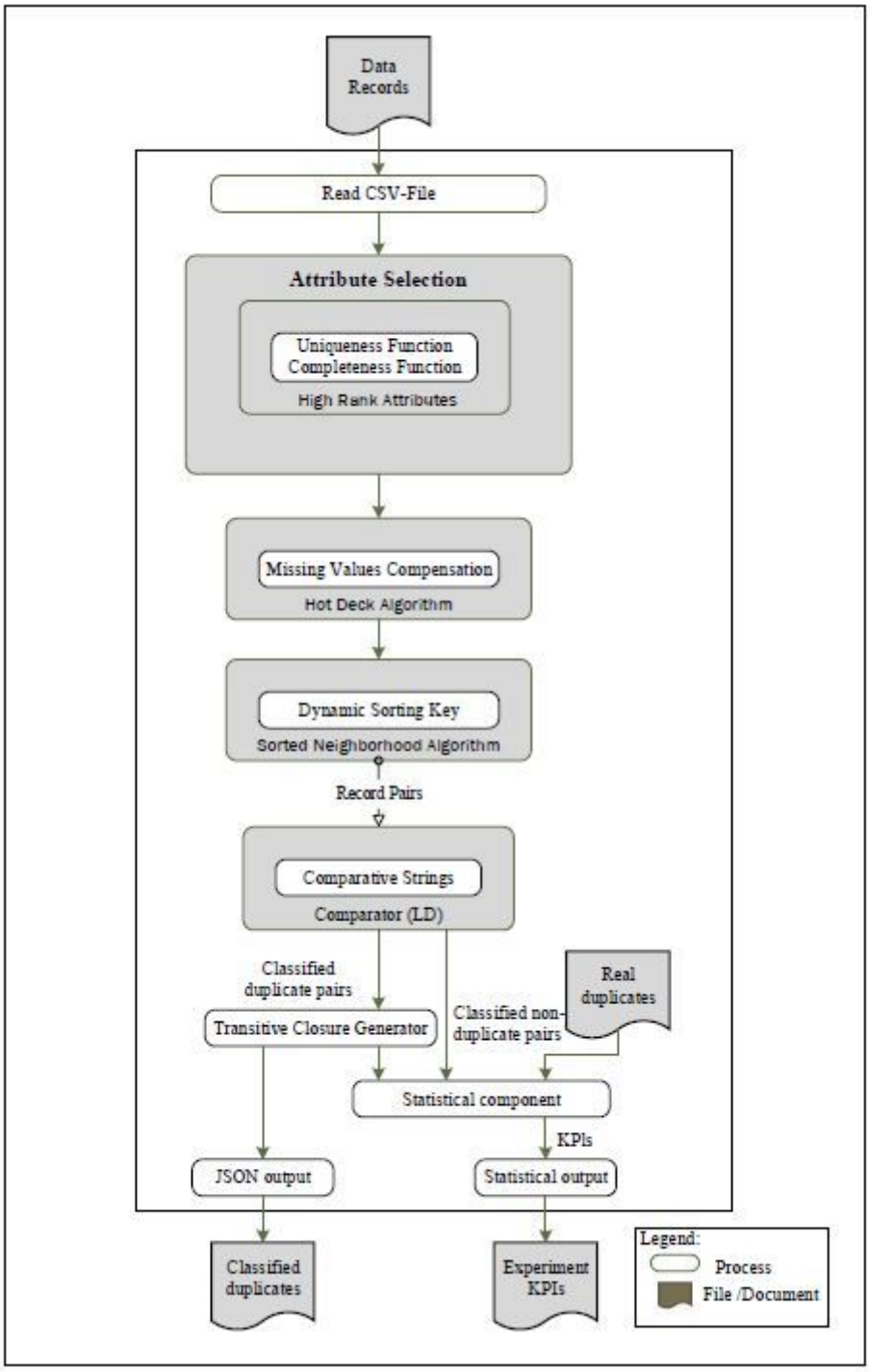


Figure 2

Experiment Work-Flow

DuDe implementation on a Restaurant dataset

Elapsed Time in ms:	5328	Number of Data Records:	865	True Positives:	78.0
Max. Memory in KB:	13,890	Number of Comparison Candidates:	16245	False Positives:	19.0
Min. Memory in KB:	13,890	Number of actual comparisons:	16245	False Negatives:	34.0
Avg. Memory in KB:	13,890	Number of real Duplicates:	112	True Negatives:	734.00
Declared Duplicates:	97				
Recall: 0.70	Precision: 0.80	F-Measure: 0.75	Accuracy 0.94		

BUILD SUCCESSFUL (total time: 9 seconds)

DDID implenentation on a Restaurant dataset

Elapsed Time in ms:	2092	Number of Data Records:	865	True Positives:	97.0
Max. Memory in KB:	20,098	Number of Comparison Candidates:	16245	False Positives:	0.0
Min. Memory in KB:	20,098	Number of actual comparisons:	16245	False Negatives:	15.0
Avg. Memory in KB:	20,098	Number of real Duplicates:	112	True Negatives:	753.00
Declared Duplicates:	97				
Recall: 0.87	Precision: 1.00	F-Measure: 0.93	Accuracy 0.98		

BUILD SUCCESSFUL (total time: 2 seconds)

A. Restaurant

DuDe implenentation on a CD dataset

Elapsed Time in ms:	12298	Number of Data Records:	9763	True Positives:	205.0
Max. Memory in KB:	224,700	Number of Comparison Candidates:	477162	False Positives:	101.0
Min. Memory in KB:	50,271	Number of actual comparisons:	477162	False Negatives:	94.0
Avg. Memory in KB:	125,487	Number of real Duplicates:	299	True Negatives:	9363.00
Declared Duplicates:	306				
Recall: 0.69	Precision: 0.67	F-Measure: 0.68	Accuracy 0.98		

DDID implenentation on a CD dataset

Elapsed Time in ms:	11340	Number of Data Records:	9763	True Positives:	236.0
Max. Memory in KB:	123,835	Number of Comparison Candidates:	477162	False Positives:	26.0
Min. Memory in KB:	74,961	Number of actual comparisons:	477162	False Negatives:	63.0
Avg. Memory in KB:	99,398	Number of real Duplicates:	299	True Negatives:	9438.00
Declared Duplicates:	262				
Recall: 0.79	Precision: 0.90	F-Measure: 0.84	Accuracy 0.99		

B. CD

Figure 3

Experiment Results

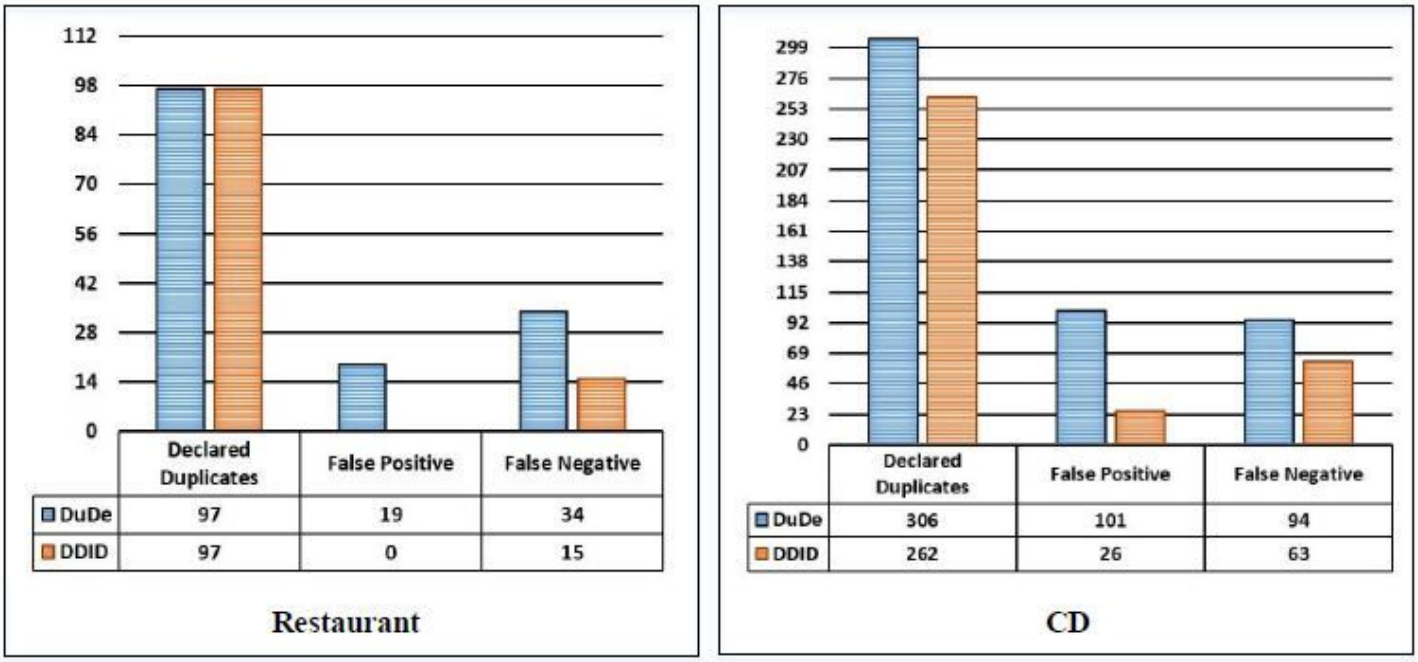


Figure 4

Comparison of Error Categories between DuDe vs DDID

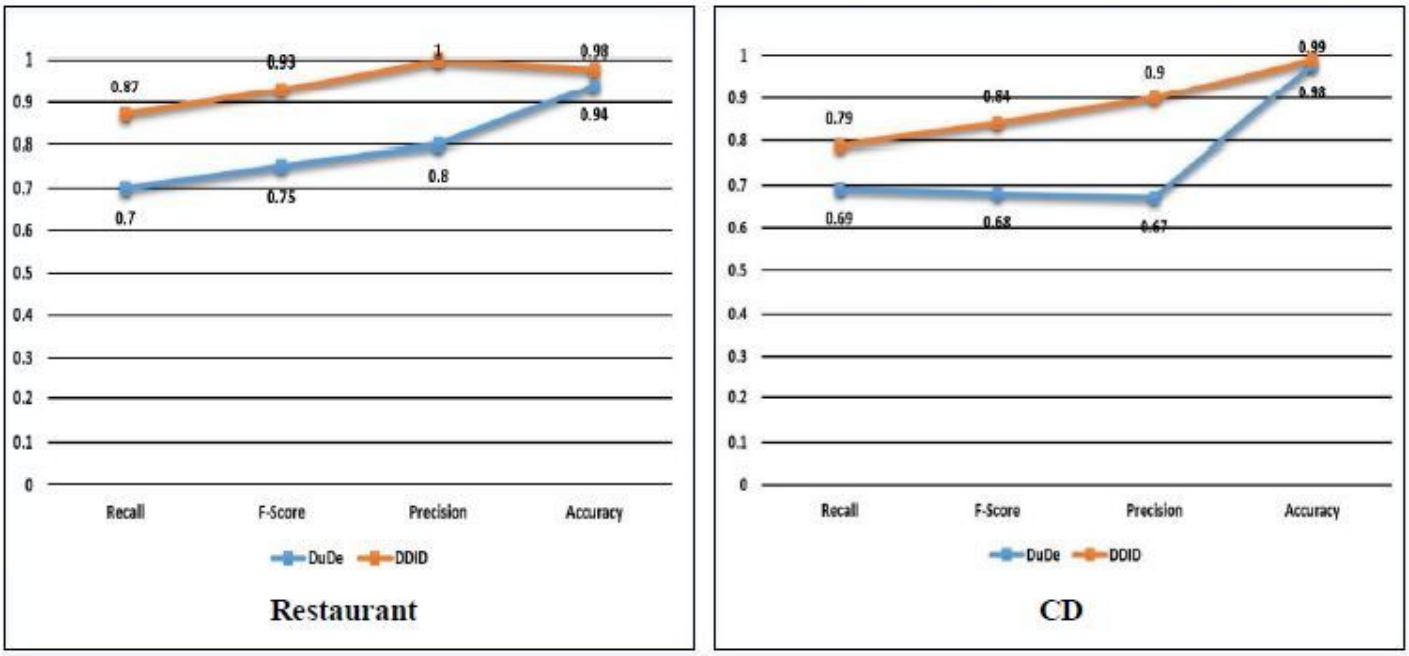
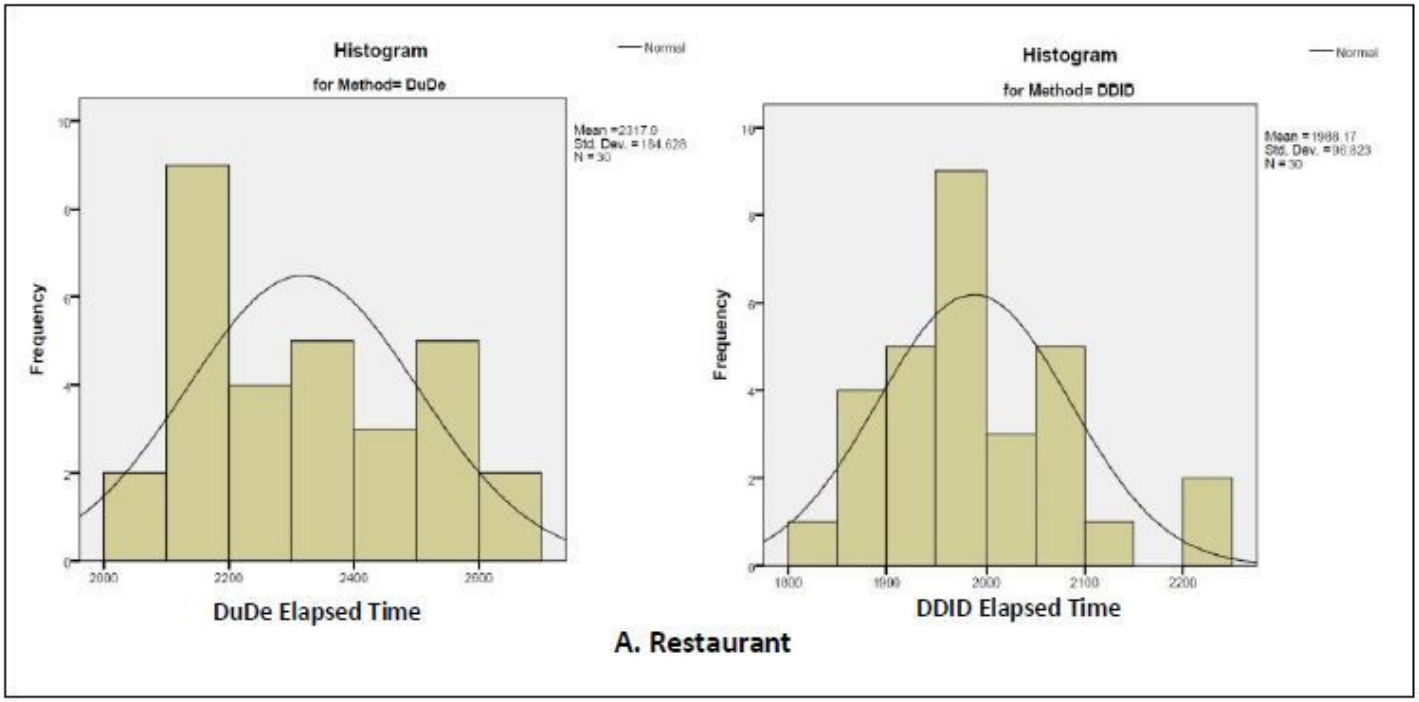


Figure 5

Accuracy comparison of Performance between DDID and DuDe



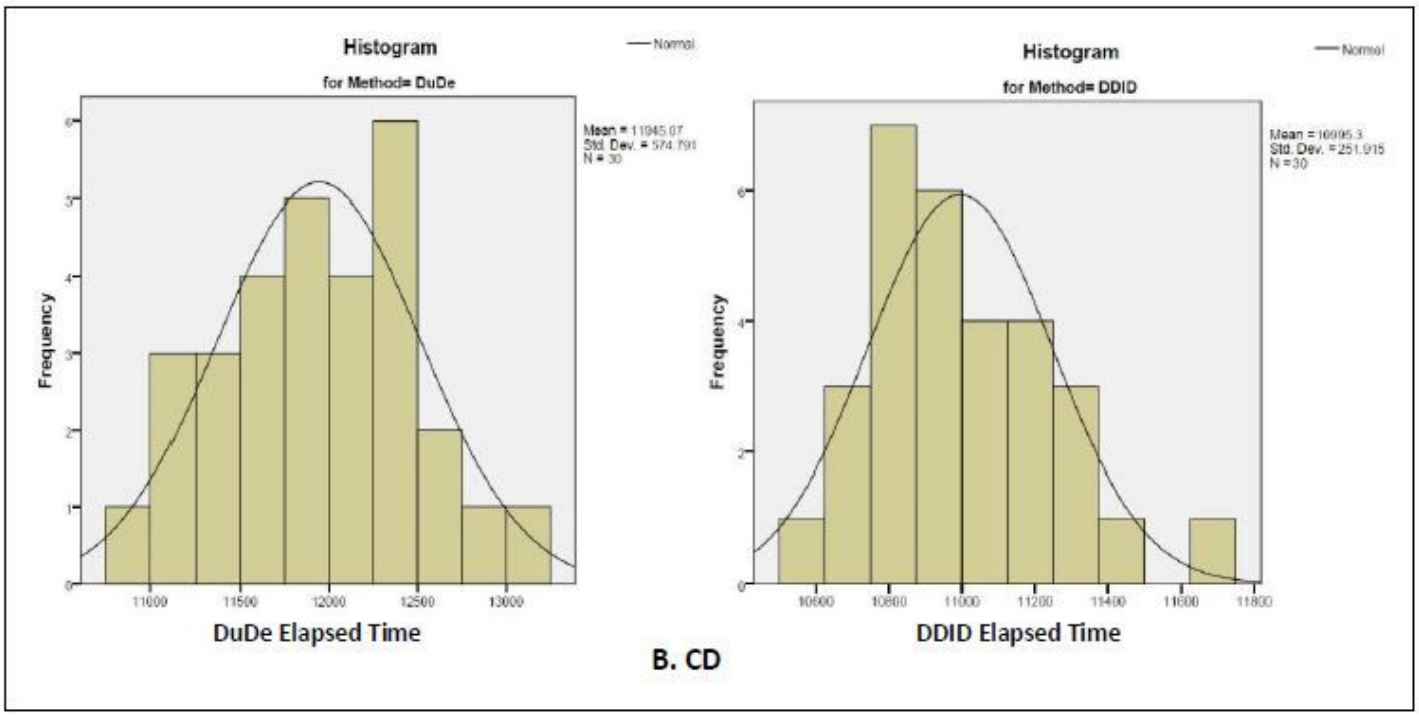


Figure 6

Data Distribution