SEMINARIO DEL 08/06

Ruggieri Andrea

Stranieri Francesco

MAD Lab

INTRODUZIONE AI TEST ASIA

Approccio metodologico

Metriche di valutazione

Generazione dei dataset

APPROCCIO METODOLOGICO – ASIA

Il dataset originale di ASIA conta 5000 dati e 8 diverse variabili.

I test eseguiti sono stati effettuati prendendo in considerazione dataset di dimensioni diverse

100	200	300	400	500	1000	1500	2000

APPROCCIO METODOLOGICO – ASIA

Il dataset originale di ASIA conta 5000 dati e 8 diverse variabili.

Per ogni dimensione testata sono state calcolate **10 repliche.** Ad ogni replica, il dataset utilizzato risulta essere diverso

100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2
rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3
rep4	rep4	rep4	rep4	rep4	rep4	rep4	rep4
rep5	rep5	rep5	rep5	rep5	rep5	rep5	rep5
rep6	rep6	rep6	rep6	rep6	rep6	rep6	rep6
rep7	rep7	rep7	rep7	rep7	rep7	rep7	rep7
rep8	rep8	rep8	rep8	rep8	rep8	rep8	rep8
rep9	rep9	rep9	rep9	rep9	rep9	rep9	rep9
rep10	rep10	rep10	rep10	rep10	rep10	rep10	rep10

APPROCCIO METODOLOGICO – ASIA

*Una volta ottenuti questi risultati è stata effettuata una **media dei valori** e sono stati calcolati gli **intervalli di confidenza** relativi alla grandezza del dataset fissando **alpha = 0.05***

APPROCCIO METODOLOGICO – ASIA

Sono stati eseguiti 3 diverse tipologie di test:

% correct replacement
Differenza in valore assoluto
Kullback-Leibler divergence

A breve approfondiremo tutte e tre le diverse tipologie di test

APPROCCIO METODOLOGICO – ASIA

Un iper-parametro fondamentale nell'esecuzione dei test è il parametro **prop**

prop indica la percentuale di celle missing presenti nel dataset.

Ad esempio: se il dataset ASIA contiene 5000 dati e ha 8 variabili, il numero totale di celle è 40.000. Fissando prop = 0.05, il numero totale di celle missing risulta essere di circa 2000

Numero missing values sul dataset ASIA al variare della prop (con tolleranza)			
0,05	0,1	0,15	0,2
[1900; 2100]	[3850; 4150]	[5800; 6200]	[7700; 8300]

Un dataset viene ritenuto valido se il numero dei missing values rientrano all'interno di questi intervalli. Vedremo successivamente come i dataset sono stati generati

APPROCCIO METODOLOGICO – ASIA

Comunque, i test sono stati effettuando variando la percentuale delle celle missing per ogni tipologia di test

	Prop			
	0,05	0,1	0,15	0,2
% correct replacement	SI	SI	SI	SI
Differenza in valore assoluto	SI	SI	NO	SI
Kullback-Leibler divergence	SI	SI	NO	SI

APPROCCIO METODOLOGICO – ASIA

% correct replacement

Prop = 0.05							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2							
rep3							
rep4							
rep5							
rep6							
rep7							
rep8							
rep9							
rep10							

Prop = 0.1							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2							
rep3							
rep4							
rep5							
rep6							
rep7							
rep8							
rep9							
rep10							

Prop = 0.15							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2							
rep3							
rep4							
rep5							
rep6							
rep7							
rep8							
rep9							
rep10							

Prop = 0.2							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2
rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3
rep4	rep4	rep4	rep4	rep4	rep4	rep4	rep4
rep5	rep5	rep5	rep5	rep5	rep5	rep5	rep5
rep6	rep6	rep6	rep6	rep6	rep6	rep6	rep6
rep7	rep7	rep7	rep7	rep7	rep7	rep7	rep7
rep8	rep8	rep8	rep8	rep8	rep8	rep8	rep8
rep9	rep9	rep9	rep9	rep9	rep9	rep9	rep9
rep10	rep10	rep10	rep10	rep10	rep10	rep10	rep10

Differenza in valore assoluto

Prop = 0.05							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2							
rep3							
rep4							
rep5							
rep6							
rep7							
rep8							
rep9							
rep10							

Prop = 0.1							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2							
rep3							
rep4							
rep5							
rep6							
rep7							
rep8							
rep9							
rep10							

Prop = 0.2							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2
rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3
rep4	rep4	rep4	rep4	rep4	rep4	rep4	rep4
rep5	rep5	rep5	rep5	rep5	rep5	rep5	rep5
rep6	rep6	rep6	rep6	rep6	rep6	rep6	rep6
rep7	rep7	rep7	rep7	rep7	rep7	rep7	rep7
rep8	rep8	rep8	rep8	rep8	rep8	rep8	rep8
rep9	rep9	rep9	rep9	rep9	rep9	rep9	rep9
rep10	rep10	rep10	rep10	rep10	rep10	rep10	rep10

Kullback-Leibler

Prop = 0.05							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2							
rep3							
rep4							
rep5							
rep6							
rep7							
rep8							
rep9							
rep10							

Prop = 0.1							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2							
rep3							
rep4							
rep5							
rep6							
rep7							
rep8							
rep9							
rep10							

Prop = 0.2							
100	200	300	400	500	1000	1500	2000
rep1	rep1	rep1	rep1	rep1	rep1	rep1	rep1
rep2	rep2	rep2	rep2	rep2	rep2	rep2	rep2
rep3	rep3	rep3	rep3	rep3	rep3	rep3	rep3
rep4	rep4	rep4	rep4	rep4	rep4	rep4	rep4
rep5	rep5	rep5	rep5	rep5	rep5	rep5	rep5
rep6	rep6	rep6	rep6	rep6	rep6	rep6	rep6
rep7	rep7	rep7	rep7	rep7	rep7	rep7	rep7
rep8	rep8	rep8	rep8	rep8	rep8	rep8	rep8
rep9	rep9	rep9	rep9	rep9	rep9	rep9	rep9
rep10	rep10	rep10	rep10	rep10	rep10	rep10	rep10

APPROCCIO METODOLOGICO – ASIA

Ci sono tre varianti dell'algoritmo EM che sono stati valutati:
EM HARD: Assegnamento 'hard' alle celle missing
EM SOFT: Assegnamento 'soft' alle celle missing ma senza limitare il numero di iterazioni possibili
EM SOFT FORCED: Assegnamento 'soft' alle celle missing **limitando il tempo di computazione** e forzando la terminazione in un numero contenuto di iterazioni

E' importante sottolineare che per ogni replica, è stato considerato lo stesso dataset per tutti e tre gli algoritmi EM

APPROCCIO METODOLOGICO – ASIA

Prima di entrare nel dettaglio dei test effettuati, introduciamo gli ouput di ciascuna tipologia di test

	Tipologia di analisi		
	Risultati generali su tutto il dataset	**Generale su tutto il dataset normalizzato [0,1]**	**Analisi node by node**
% correct replacement	**NO**	**SI**	**NO**
Differenza in valore assoluto	**SI**	**SI**	**SI**
Kullback-Leibler divergence	**SI**	**SI**	**SI**

METRICHE DI VALUTAZIONE – ASIA

La prima tipologia di test effettuati prende il nome di % correct replacement

Si confrontano i valori missing rimpiazzati dall'algoritmo EM con la Ground Truth disponibile e si valuta:

$$\%\ correct\ replacement = \frac{Number\ of\ correct\ replacement}{Number\ of\ missing\ values}$$

% correct replacement assume un valore da [0, 1]. Un valore prossimo all'1, indica che l'algoritmo EM rimpiazza correttamente tutti i missing values presenti nel dataset, 0 viceversa.

METRICHE DI VALUTAZIONE – ASIA

% correct replacement ha diversi vantaggi e svantaggi

Molto facile da interpretare	Non sempre il dataset che funge da GT è disponibile
Può essere interpretata come misura di accuratezza dell'algoritmo EM	Non tiene conto delle **distribuzioni di probabilità del modello**: se la probabilità è uniforme, l'errore deve essere considerato meno grave rispetto a una distribuzione di probabilità che tende ai valori estremi

METRICHE DI VALUTAZIONE – ASIA

% correct replacement ha diversi vantaggi e svantaggi

Molto facile da interpretare	Non sempre il dataset che funge da GT è disponibile
Può essere interpretata come misura di accuratezza dell'algoritmo EM	Non tiene conto delle **distribuzioni di probabilità del modello**: se la probabilità è uniforme, l'errore deve essere considerato meno grave rispetto a una distribuzione di probabilità che tende ai valori estremi

Sbagliare a classificare una variabile i cui valori tendono ad assumere una probabilità uniforme è meno grave rispetto a classificare in modo errato una variabile in cui un valore tende ad avere la probabilità vicino ad 1

METRICHE DI VALUTAZIONE – ASIA

Difference in absolute value

A differenza della metrica precedente, il confronto non avviene a livello di dati rimpiazzati ma avviene a livello di rete bayesiana ottenuta dalla computazione dell'algoritmo EM

$$Difference\ in\ absolute\ value = \sum_{x_{prob} \in BN_{EM}, y_{prob} \in BN_{GT}} |x_{prob} - y_{prob}|$$

Difference in absolute value assume un valore da [0, #prob]. Un valore prossimo all'0, indica che non ci sono differenze tra le distribuzioni di probabilità della rete bayesiana con quella considerata essere la ground truth

METRICHE DI VALUTAZIONE – ASIA

Difference in absolute value

Esiste anche una versione normalizzata della difference in absolute value appena introdotta

$$Difference\ in\ absolute\ value = \frac{\sum_{x_{prob} \in BN_{EM}, y_{prob} \in BN_{GT}} |x_{prob} - y_{prob}|}{card(x_{prob})}$$

Questa nuova misura assume valori [0,1]. Un valore prossimo allo 0, indica che le due distribuzioni di probabilità sono uguali, mentre con un valore prossimo all'1, vice versa

METRICHE DI VALUTAZIONE – ASIA

Difference in absolute value

Difference in absolute value (per come è stata definita) ha la limitazione di essere generale su tutta le rete bayesiana e non tiene conto delle singole distribuzioni di probabilità

In questo ambito di ricerca, potremmo essere interessati a valutare la **differenza in valore assoluto tra i singoli nodi**, per cercare di comprendere quali nodi risultano avere una differenza in valore assoluto più alta rispetto agli altri

Nella parte di presentazione dei risultati, quindi si mostreranno anche i risultati ottenuti considerando i nodi singolarmente

METRICHE DI VALUTAZIONE – ASIA

Kullback-Leibler divergence

Siamo ora interessati a valutare la perdita di informazione che si ottiene attraverso l'esecuzione dei diversi algoritmi EM

Anche se, in letteratura **Kullback-Leibler divergence**, prende il nome di **Kullback-Leibler distance**, è un errore considerare questa metrica come **metrica di distanza** in quanto è **asimmetrica**

Ricordiamo infatti, che una distanza deve avere tre proprietà fondamentali:

- **Non negatività**: d(x,y) >= 0 per ogni x e y
- **Simmetria:** d(x,y) = d(y,x) per ogni x e y
- **Diseguaglianza triangolare:** d(x,z) <= d(x,y) + d(y,z)

METRICHE DI VALUTAZIONE – ASIA

Kullback-Leibler divergence

Asimmetria

$$KLD[P(y)||P(x)] = \sum_{i}^{N} P(y_i) log \frac{p(y_i)}{p(x_i)}$$

$$\neq$$

$$KLD[P(x)||P(y)] = \sum_{i}^{N} P(x_i) log \frac{p(x_i)}{p(y_i)}$$

METRICHE DI VALUTAZIONE – ASIA

Kullback-Leibler divergence

Chiamiamo ora **p(x)** un vettore di densità di probabilità ricavato dall'esecuzione dell'algoritmo EM
Chiamiamo anche **p(y)** un vettore di densità di probabilità associato alla **reale distribuzione dei dati**

Allora:

$$KLD[P(y)||P(x)] = \sum_{i}^{N} P(y_i) log \frac{p(y_i)}{p(x_i)}$$

GENERAZIONE DEL DATASET – ASIA

Ad ogni replica, i dataset sono stati generati utilizzando la libreria MICE di R.

Come introdotto precedentemente, i missing values sono stati definiti a livello di cella e non di dato.

Ad ogni replica, EM HARD, EM SOFT e EM SOFT FORCED sono stati eseguiti attraverso lo stesso dataset ottenuto dalla libreria MICE. Ciò ha garantito che i risultati erano maggiormente confrontabili tra di loro

GENERAZIONE DEL DATASET – ASIA

I tre step per la generazione dei dataset validi sono:

- Selezione dei N dati dal dataset
- Definizione di pattern per la generazione dei missing values
- Validazione del dataset appena creato

GENERAZIONE DEL DATASET – ASIA

Selezione dei N dati dal dataset

Problema: il dataset iniziale conta 5000 dati, come faccio a generare campioni di 100, 200, 300, 400, 500, 1000, 1500 e 2000 dati?

Ad ogni replica, i dati vengono scelti in maniera causale variando il seed

GENERAZIONE DEL DATASET – ASIA

Selezione dei N dati dal dataset

Problema: Come si generano valori casuali partendo da un campione ottenuto dal dataset iniziale?

ampute

From mice v3.9.0 by Stef van Buuren

99.99th Percentile

Generate Missing Data For Simulation Purposes

This function generates multivariate missing data in a MCAR, MAR or MNAR manner. Imputation of data sets containing missing values can be performed with `mice` .

Usage

```
ampute(
  data,
  prop = 0.5,
  patterns = NULL,
  freq = NULL,
  mech = "MAR",
  weights = NULL,
  std = TRUE,
  cont = TRUE,
  type = NULL,
  odds = NULL,
  bycases = TRUE,
  run = TRUE
)
```

patterns A matrix or data frame of size #patterns by #variables where `0` indicates a variable should have missing values and `1` indicates a variable should remain complete. The user may specify as many patterns as desired. One pattern (a vector) or double patterns are possible as well. Default is a square matrix of size #variables where each pattern has missingness on one variable only (created with `ampute.default.patterns`). After the amputation procedure, `md.pattern` can be used to investigate the missing data patterns in the data.

bycases Logical. If TRUE, the proportion of missingness is defined in terms of cases. If FALSE, the proportion of missingness is defined in terms of cells. Default is TRUE.

GENERAZIONE DEL DATASET – ASIA

Selezione dei N dati dal dataset

Problema: Come si generano valori casuali partendo da un campione ottenuto dal dataset iniziale?

Ad ogni replica, i patterns vengono generati in maniera causale. Dalla matrice 8*8 vengono scelti randomicamente m celle e queste vengono impostate a 0. La scelta dell'm è casuale ad ogni replica e per prop = 0.05, questo valore viene fissato da 0.1 a 0.2

Per mantenere la riflessività, tutti i valori della diagonale principale della matrice patterns sono stati fissati pari a 0

Successivamente, attraverso il metodo impute viene generato il dataset con i missing values che diventeranno l'input per gli algoritmi EM

GENERAZIONE DEL DATASET – ASIA

Validazione del dataset appena creato

Problema: Questo meccanismo garantisce che su 800 celle, esattamente 40 celle risultano essere missing fissando prop 0.05?

NO, prop=0.05 non significa che vengono definiti esattamente 40 celle missing ma che il numero di missing values tenderà ad avvicinarsi a quella proporzione. Tuttavia, il numero di celle missing può anche essere 10 in determinate situazioni (in base ai dati e al pattern definito=.

Viene quindi fissato un valore di tolleranza che varia al variare della dimensione del dataset considerato. **Un dataset viene ritenuto valido, se il numero di missing values è all'interno dell'intervallo definito dal valore di tolleranza.**

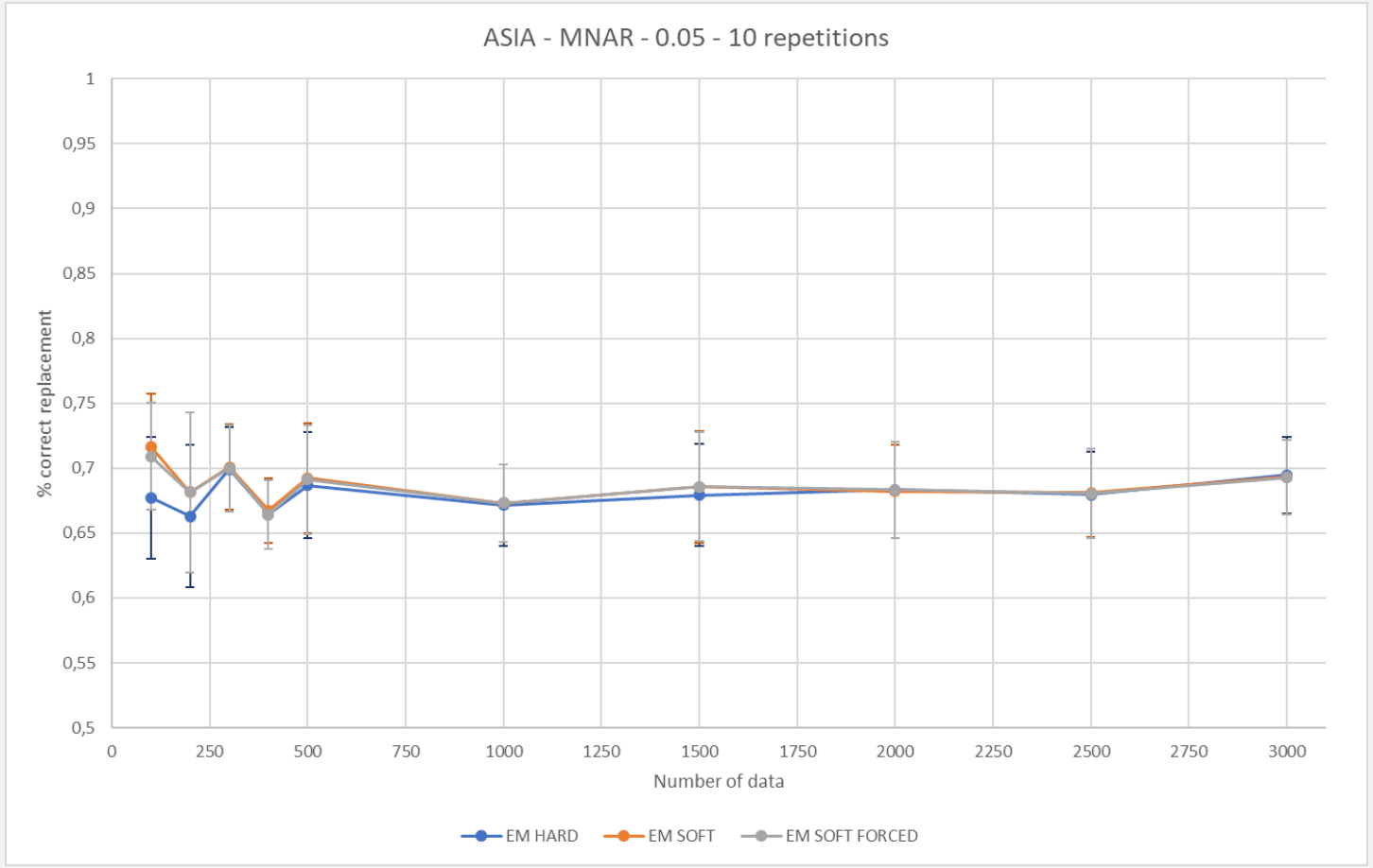
Se il dataset non viene ritenuto valido, si riesegue la procedura di generazione dei missing values con lo stesso pattern, se dopo 10 iterazioni, non è stato trovato un dataset valido, il pattern viene cambiato

TEST ASIA

Presentazione dei risultati

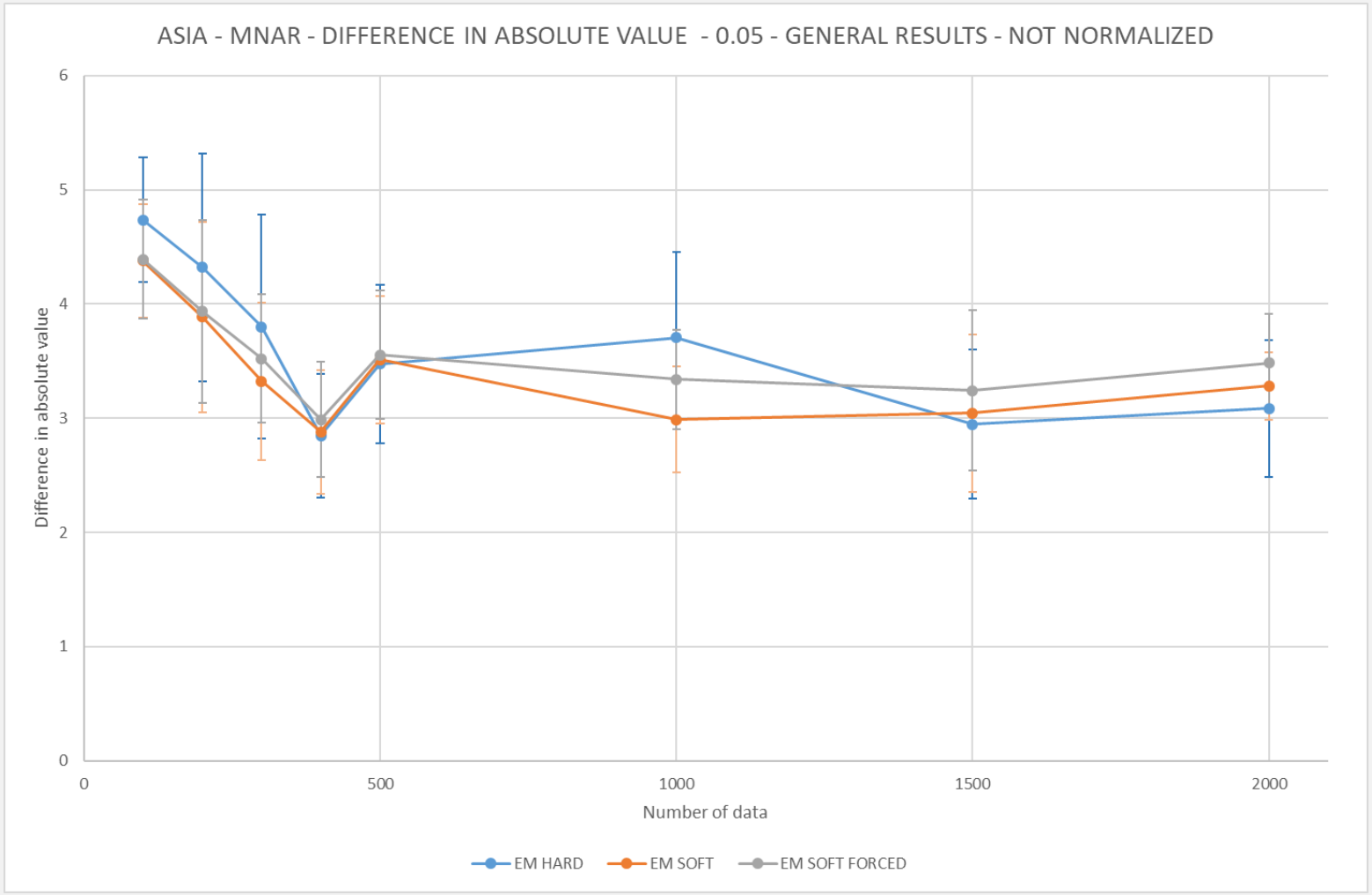
% CORRECT REPLACEMENT – ASIA

Prop = 0.05



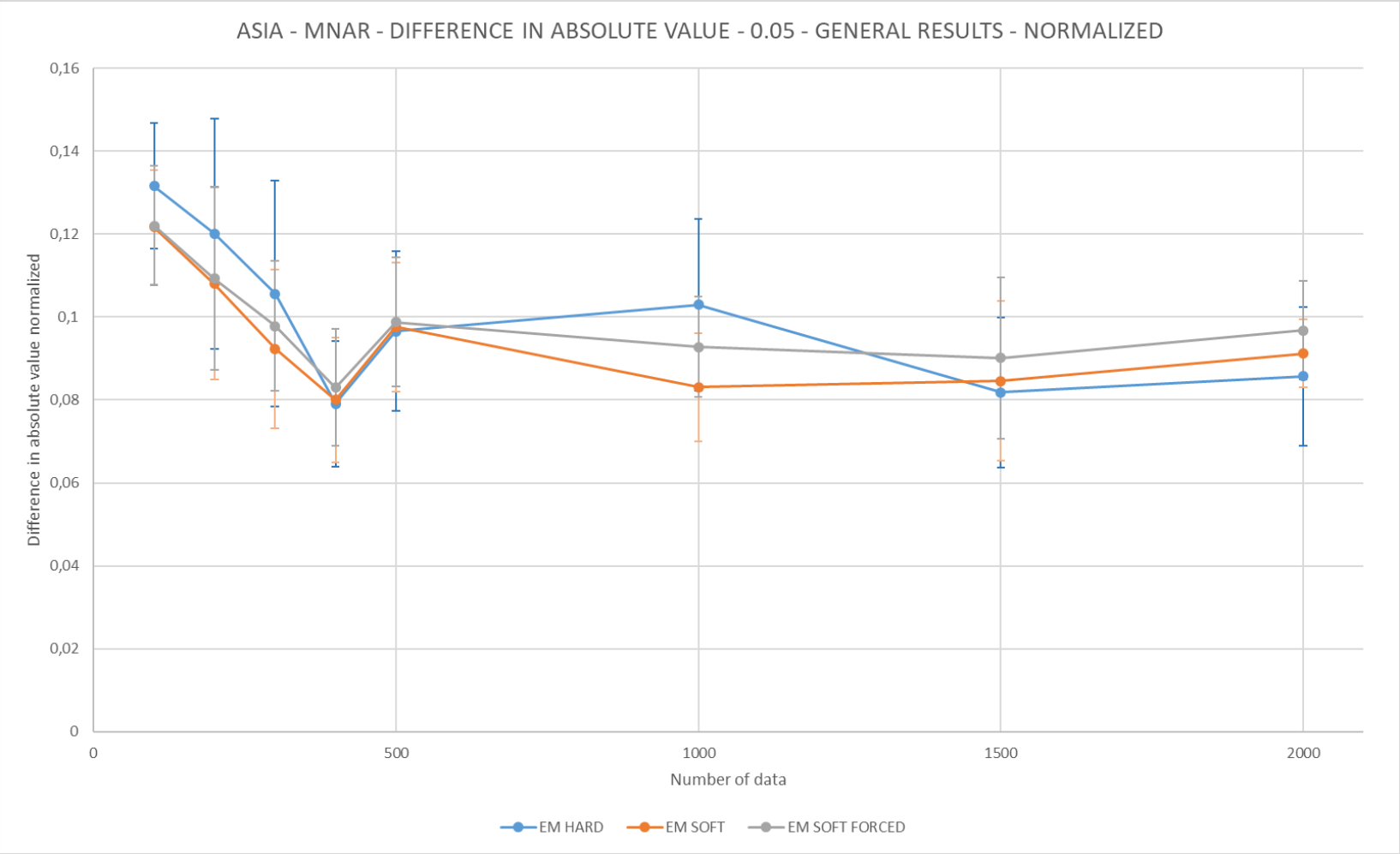
DIFFERENZA IN VALORE ASSOLUTO – ASIA

Prop = 0.05



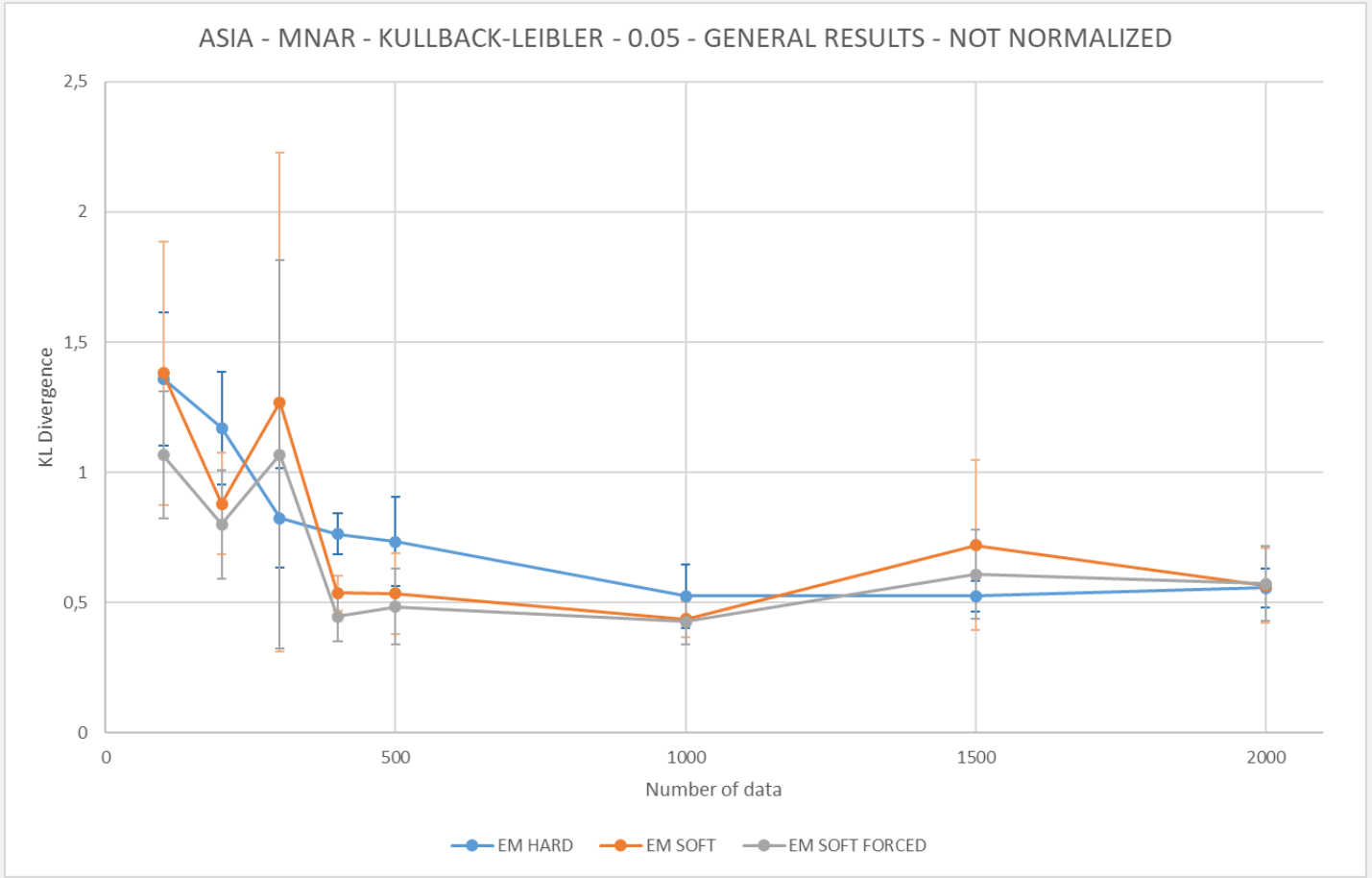
DIFFERENZA IN VALORE ASSOLUTO – ASIA

Prop = 0.05 - NORMALIZED



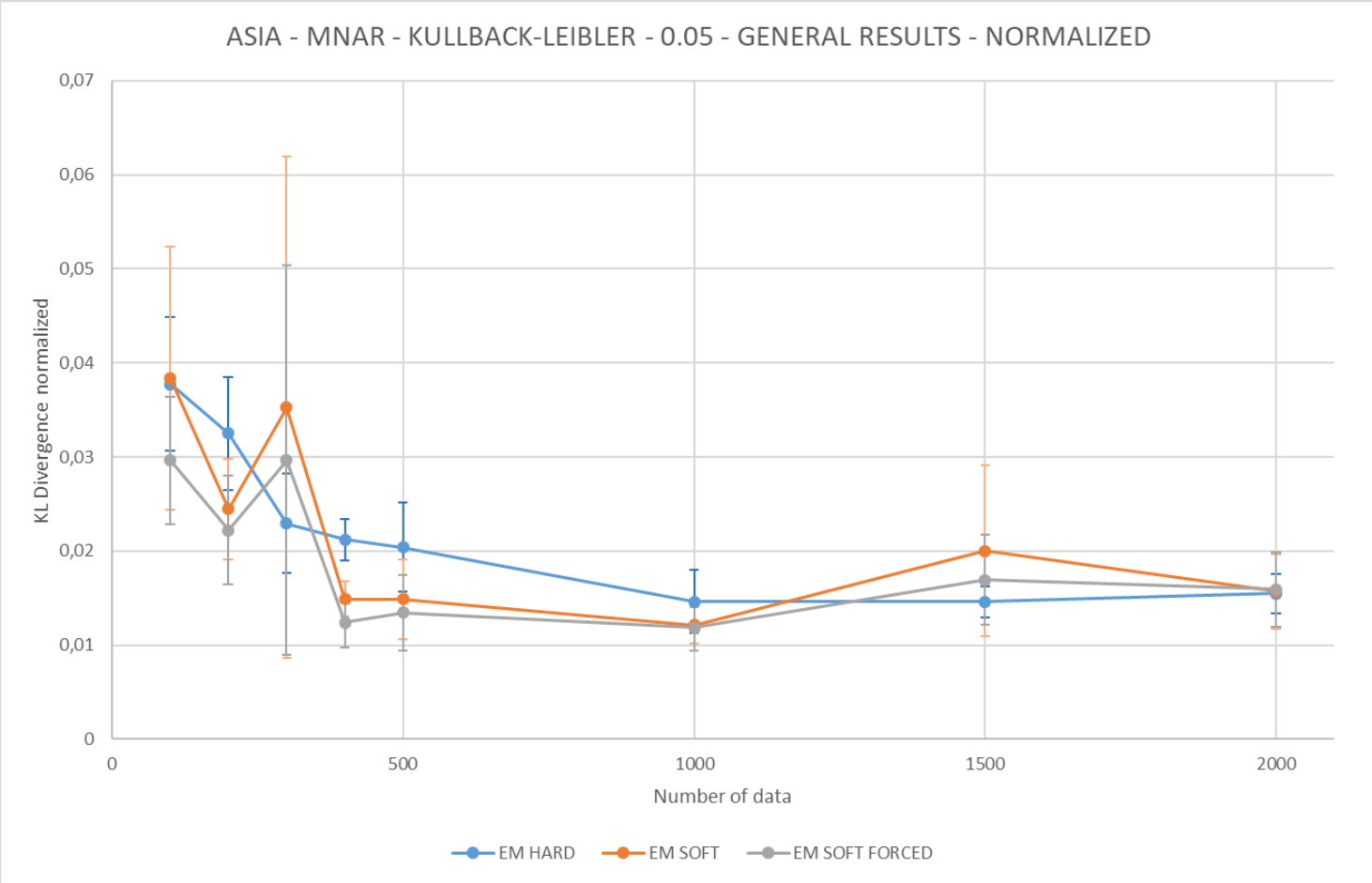
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.05



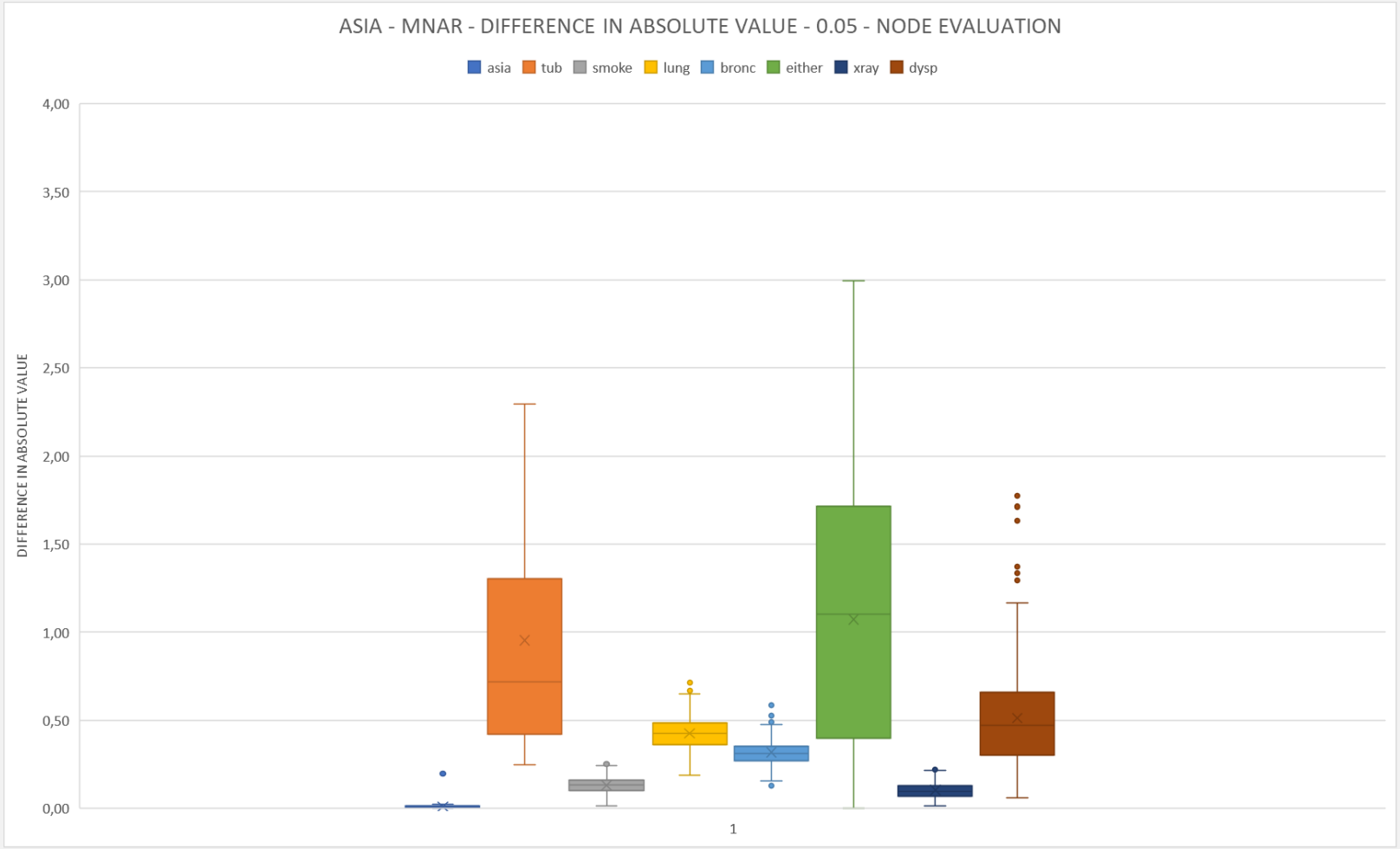
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.05 - NORMALIZED



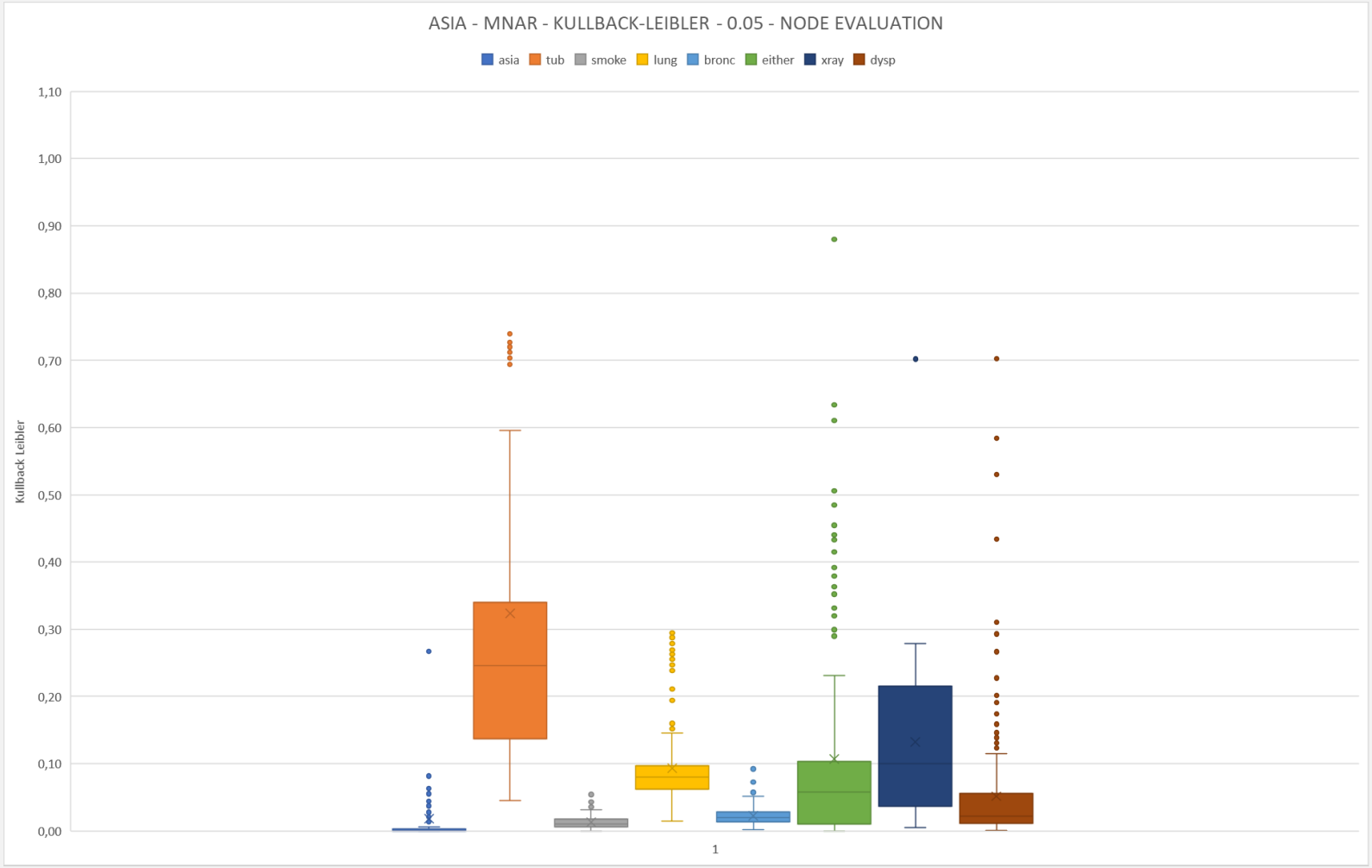
DIFFERENCE IN ABSOLUTE VALUE – ASIA

Prop = 0.05 – NODE EVALUATION



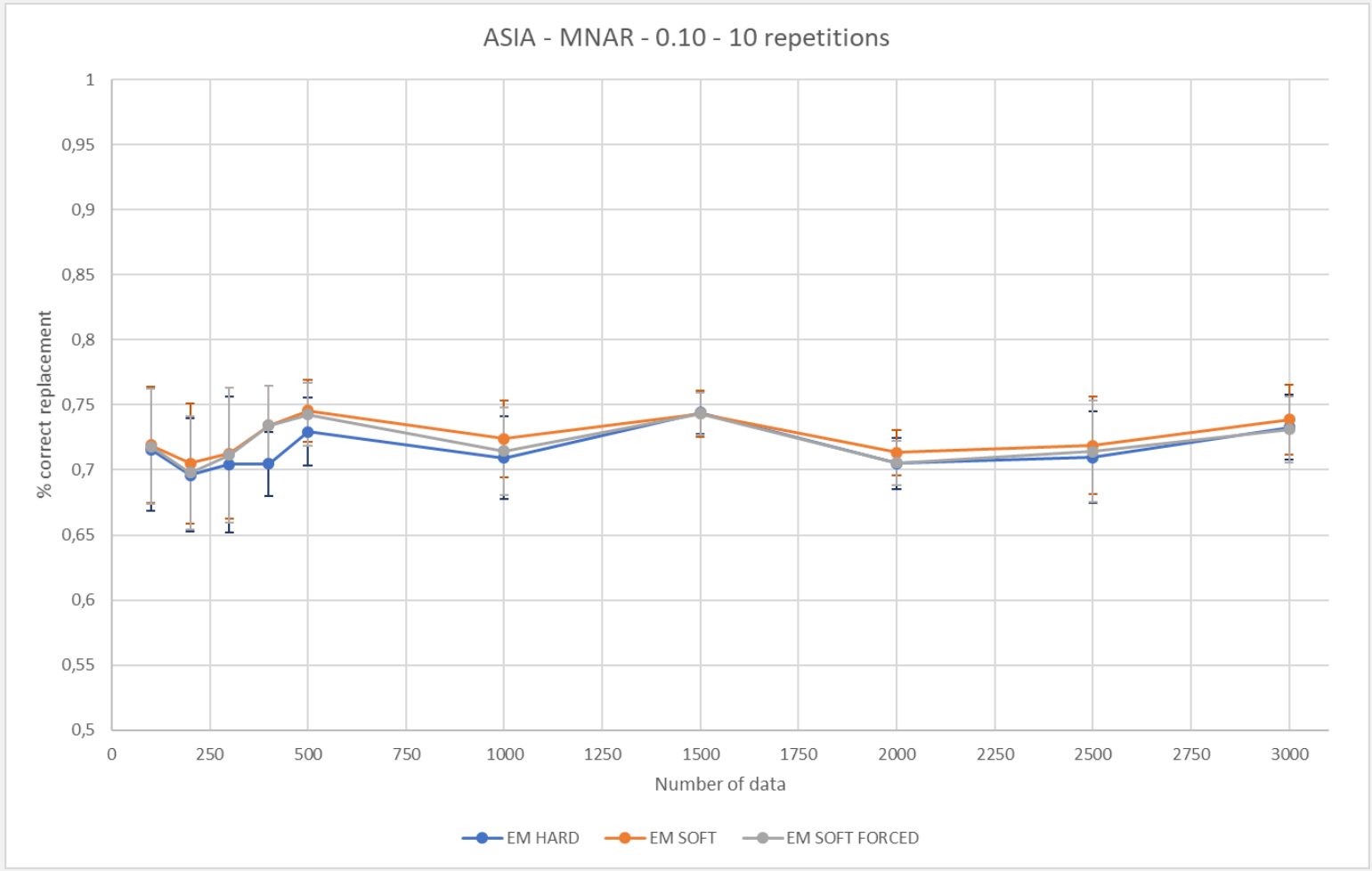
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.05 – NODE EVALUATION



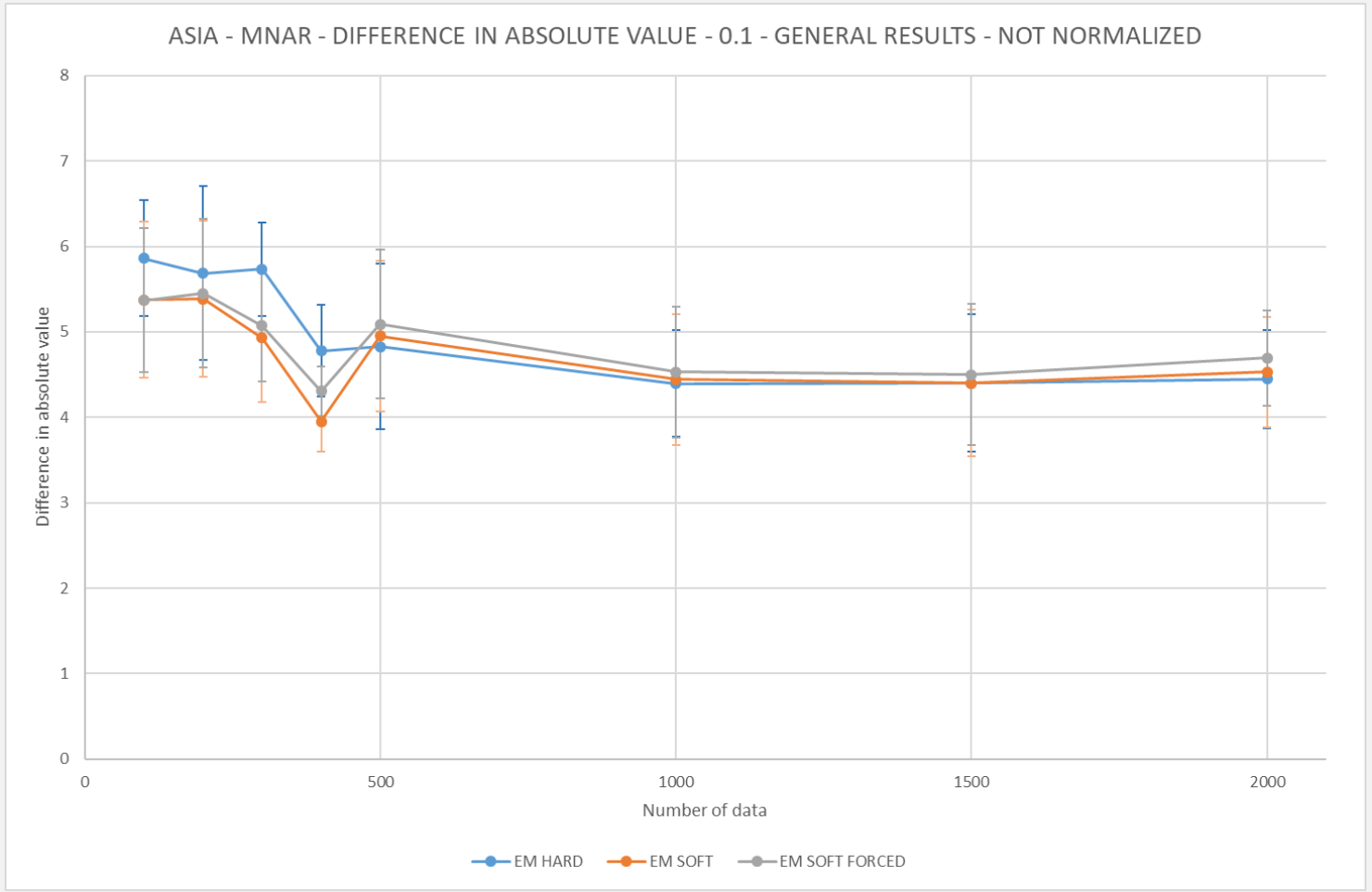
% CORRECT REPLACEMENT – ASIA

Prop = 0.1



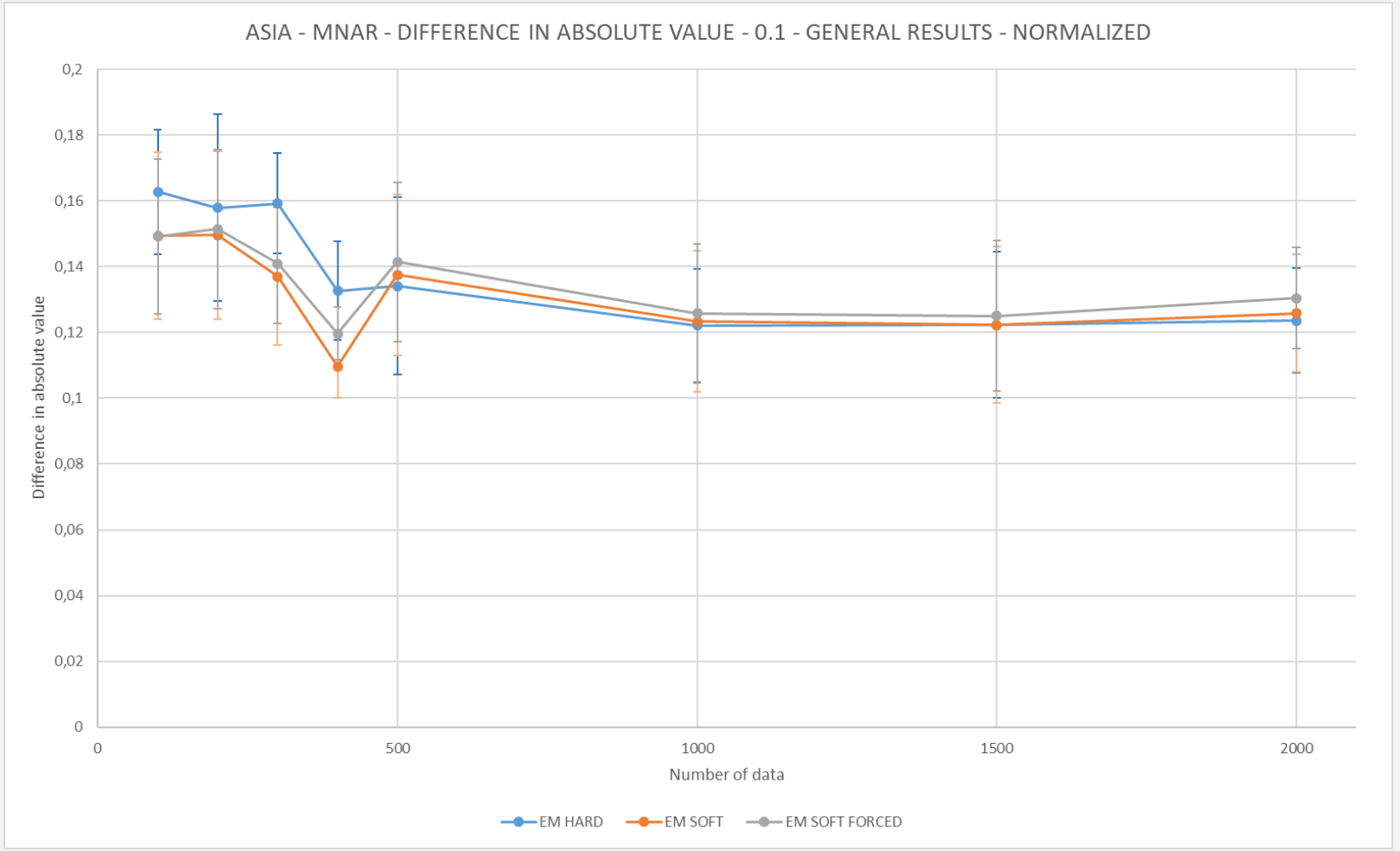
DIFFERENZA IN VALORE ASSOLUTO – ASIA

Prop = 0.1



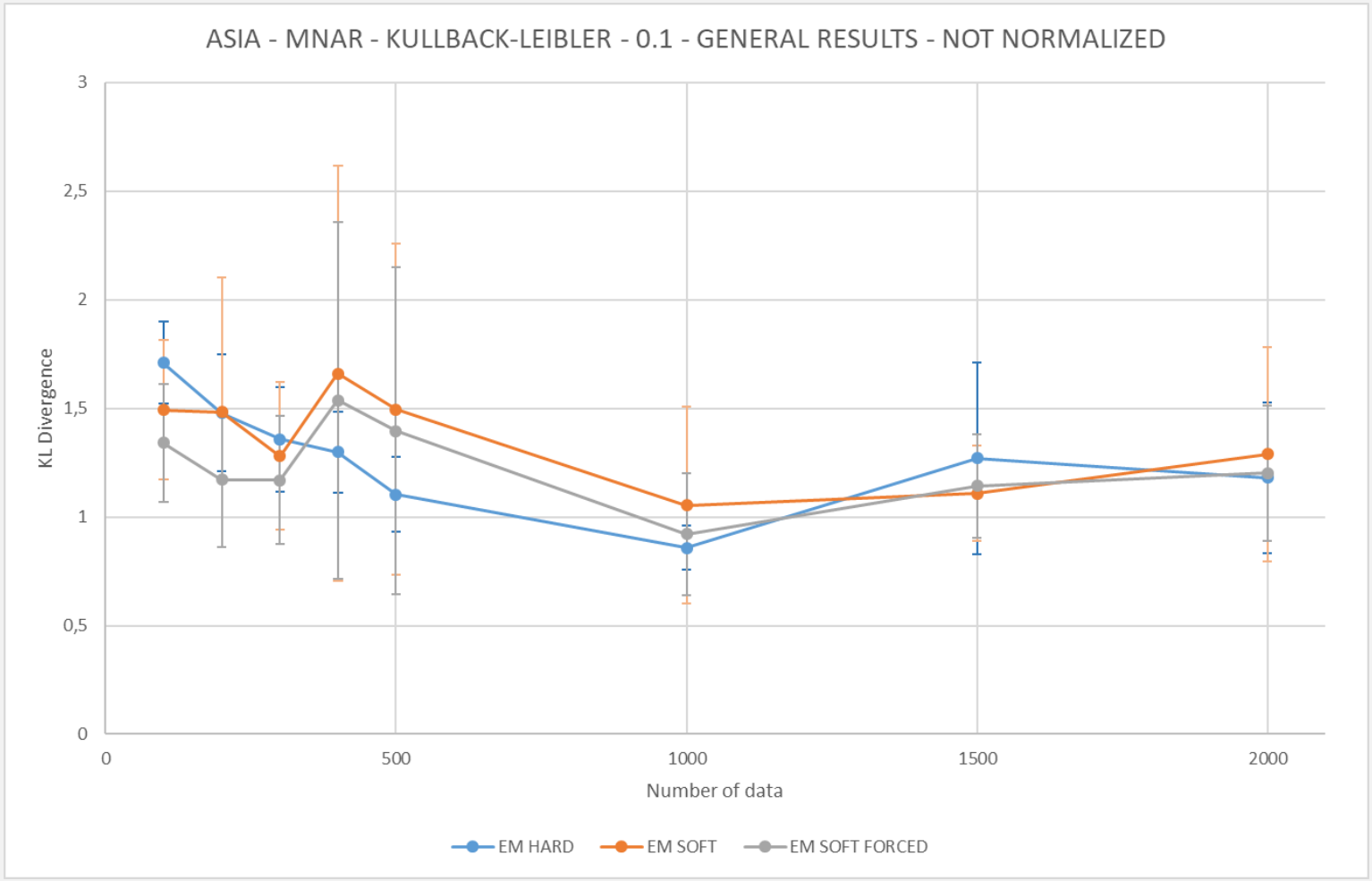
DIFFERENZA IN VALORE ASSOLUTO – ASIA

Prop = 0.1 - NORMALIZED



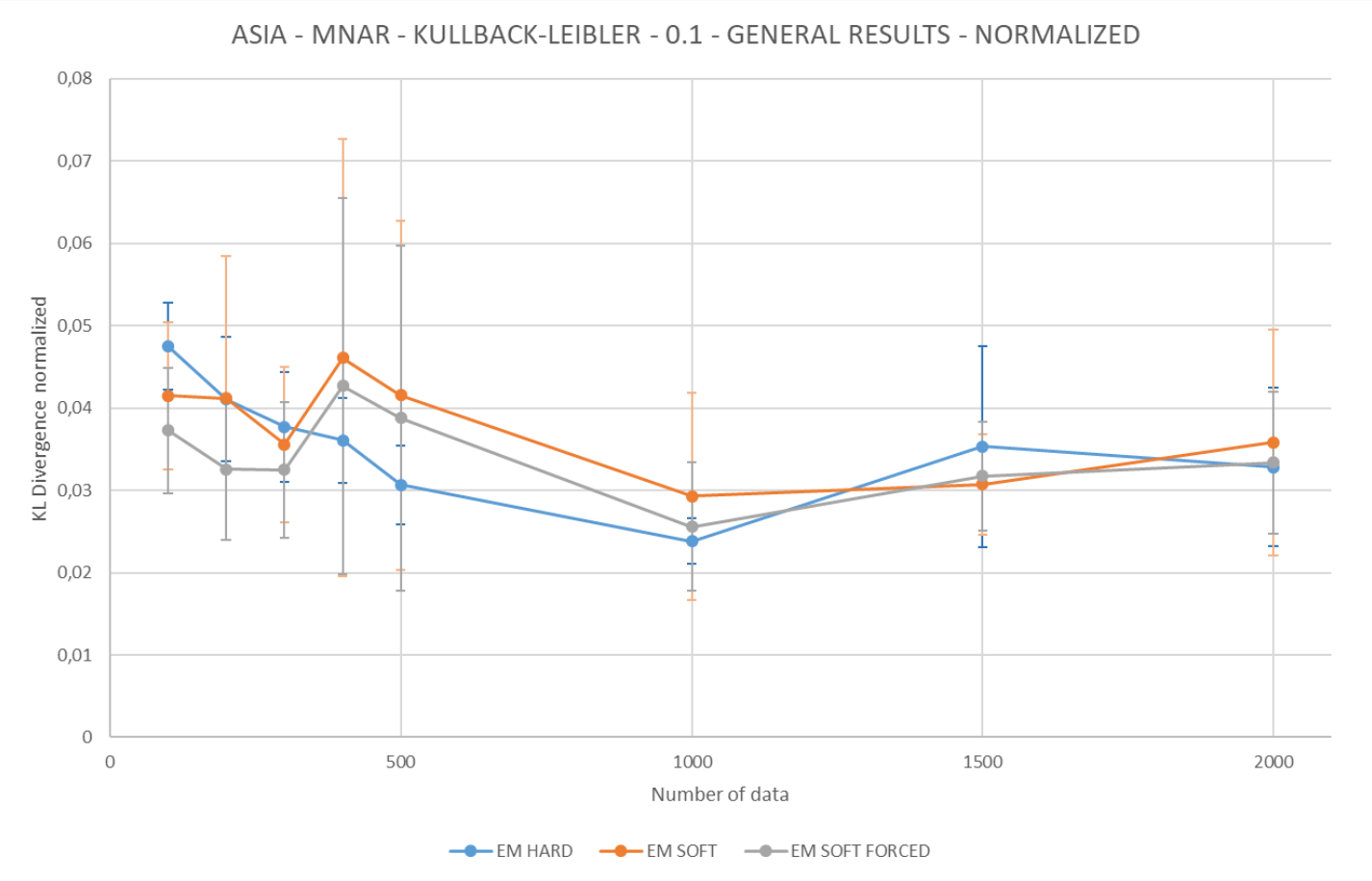
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.1



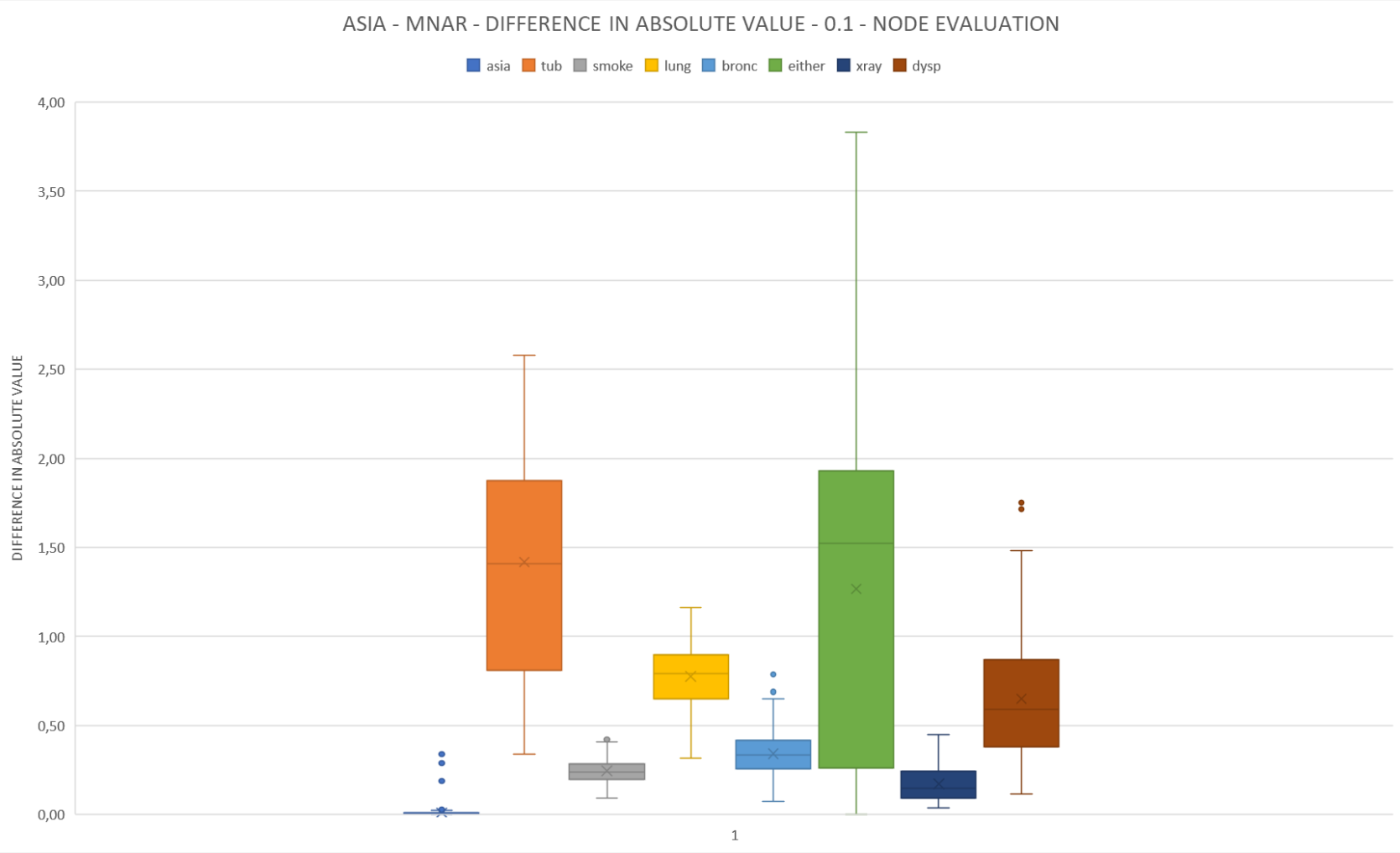
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.1 - NORMALIZED



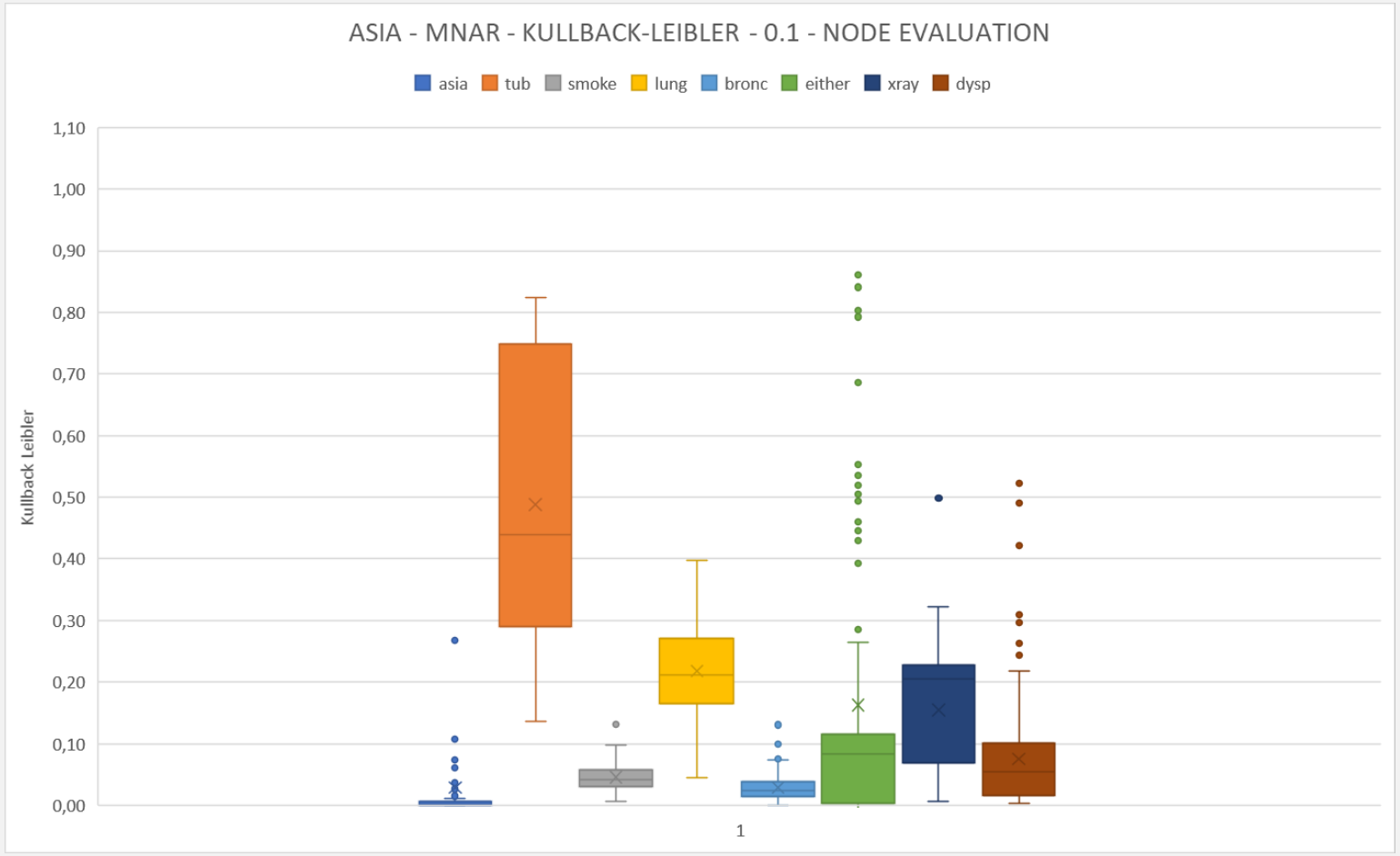
DIFFERENCE IN ABSOLUTE VALUE – ASIA

Prop = 0.1 – NODE EVALUATION



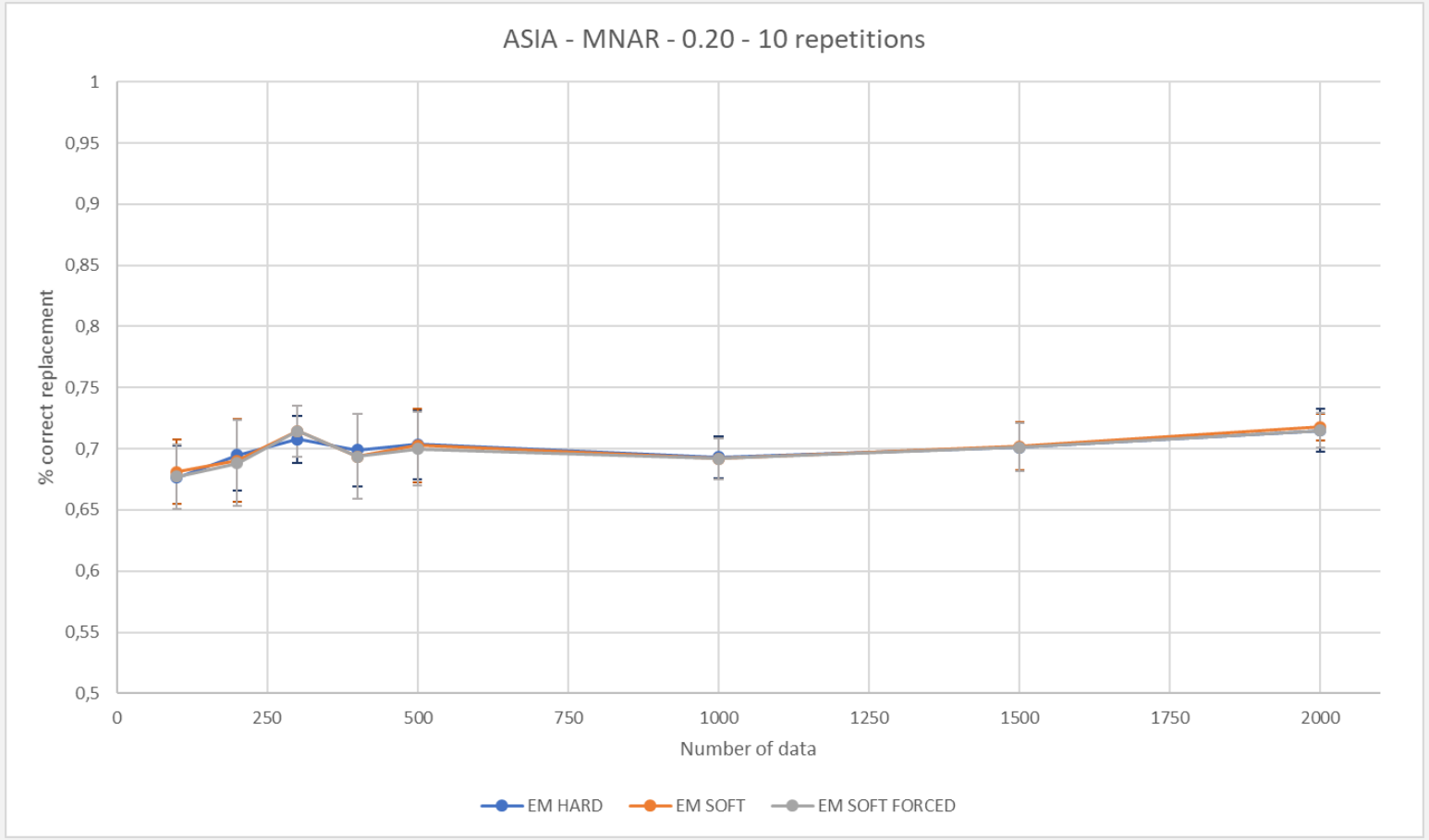
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.1 – NODE EVALUATION



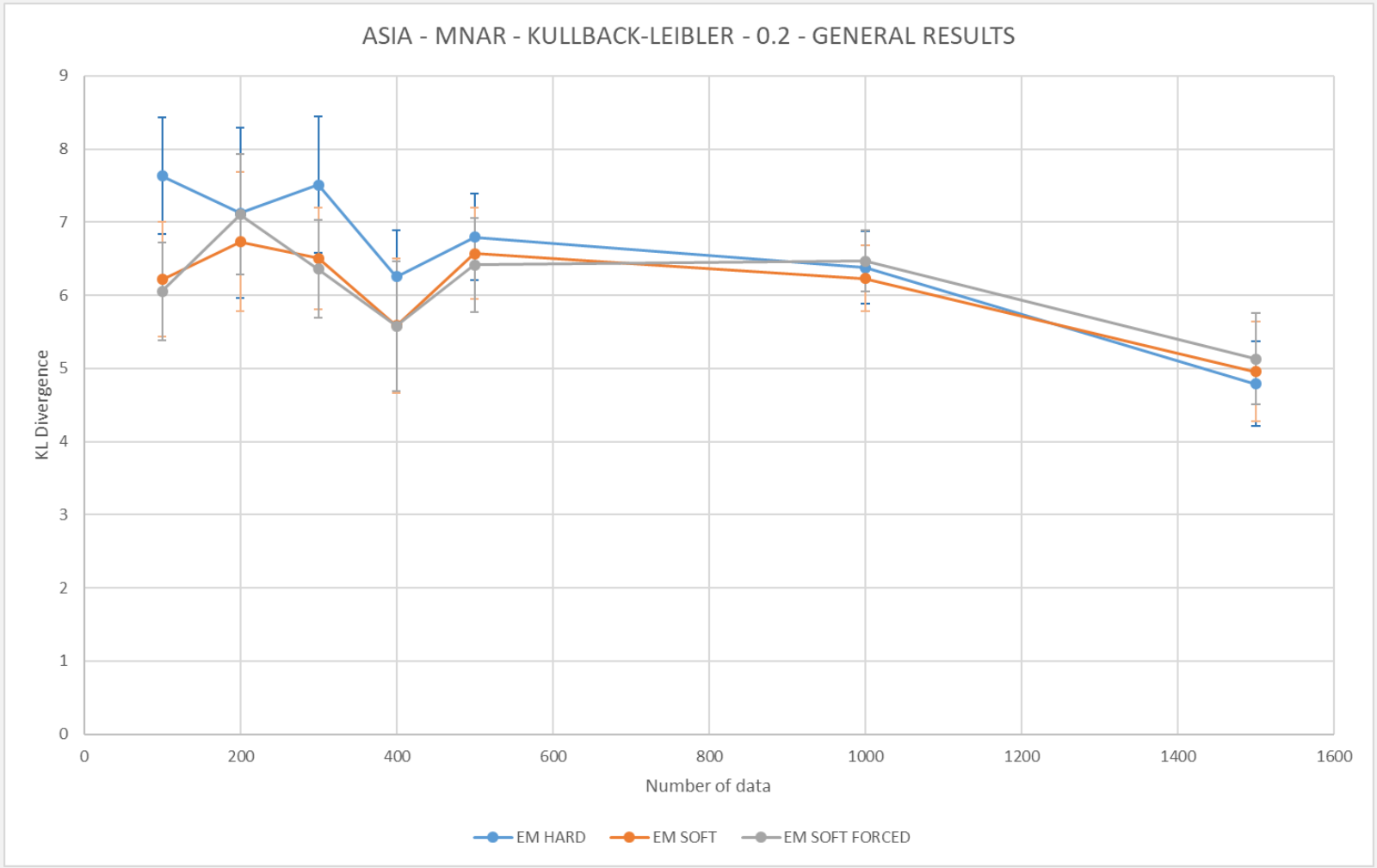
% CORRECT REPLACEMENT – ASIA

Prop = 0.2



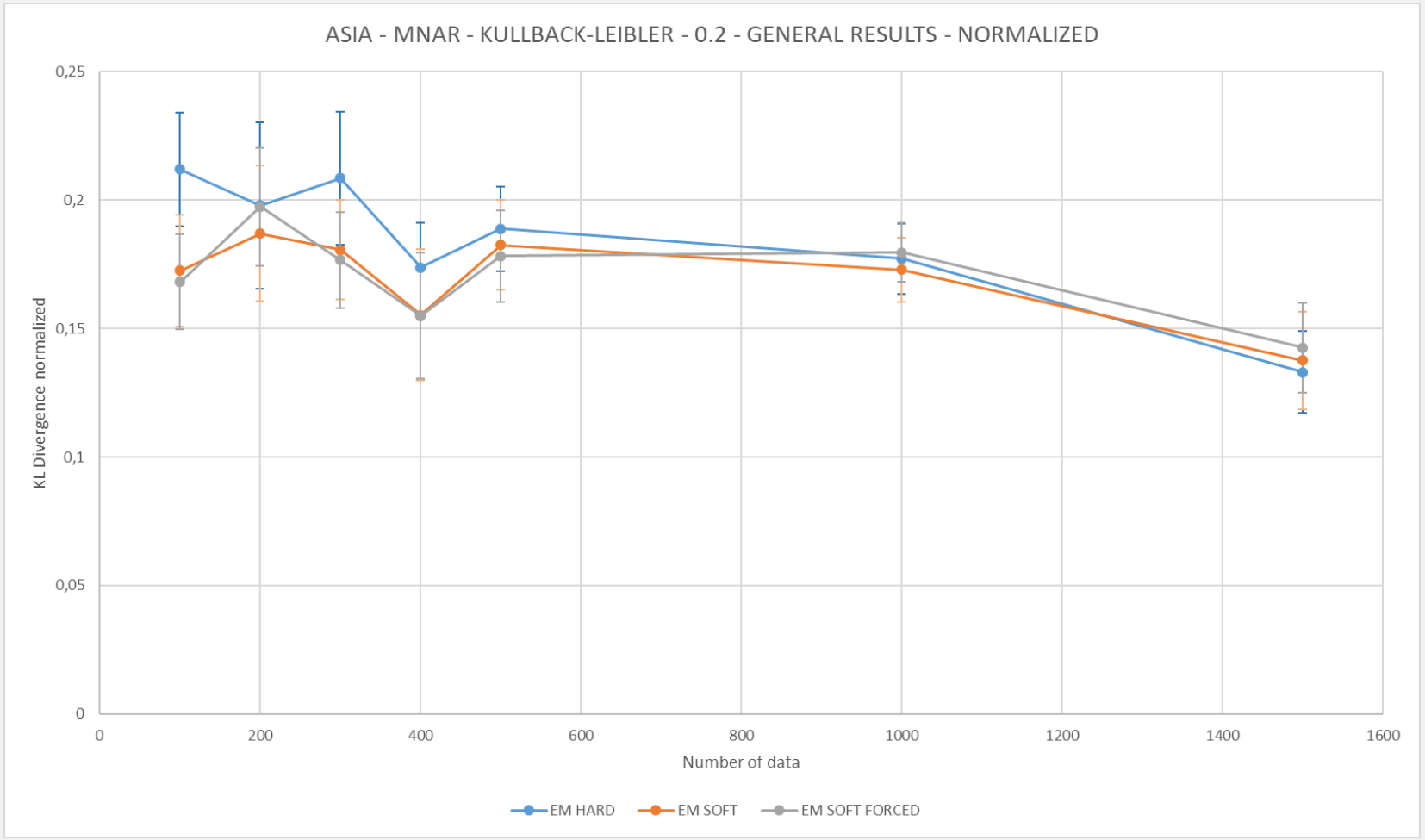
DIFFERENZA IN VALORE ASSOLUTO – ASIA

Prop = 0.2



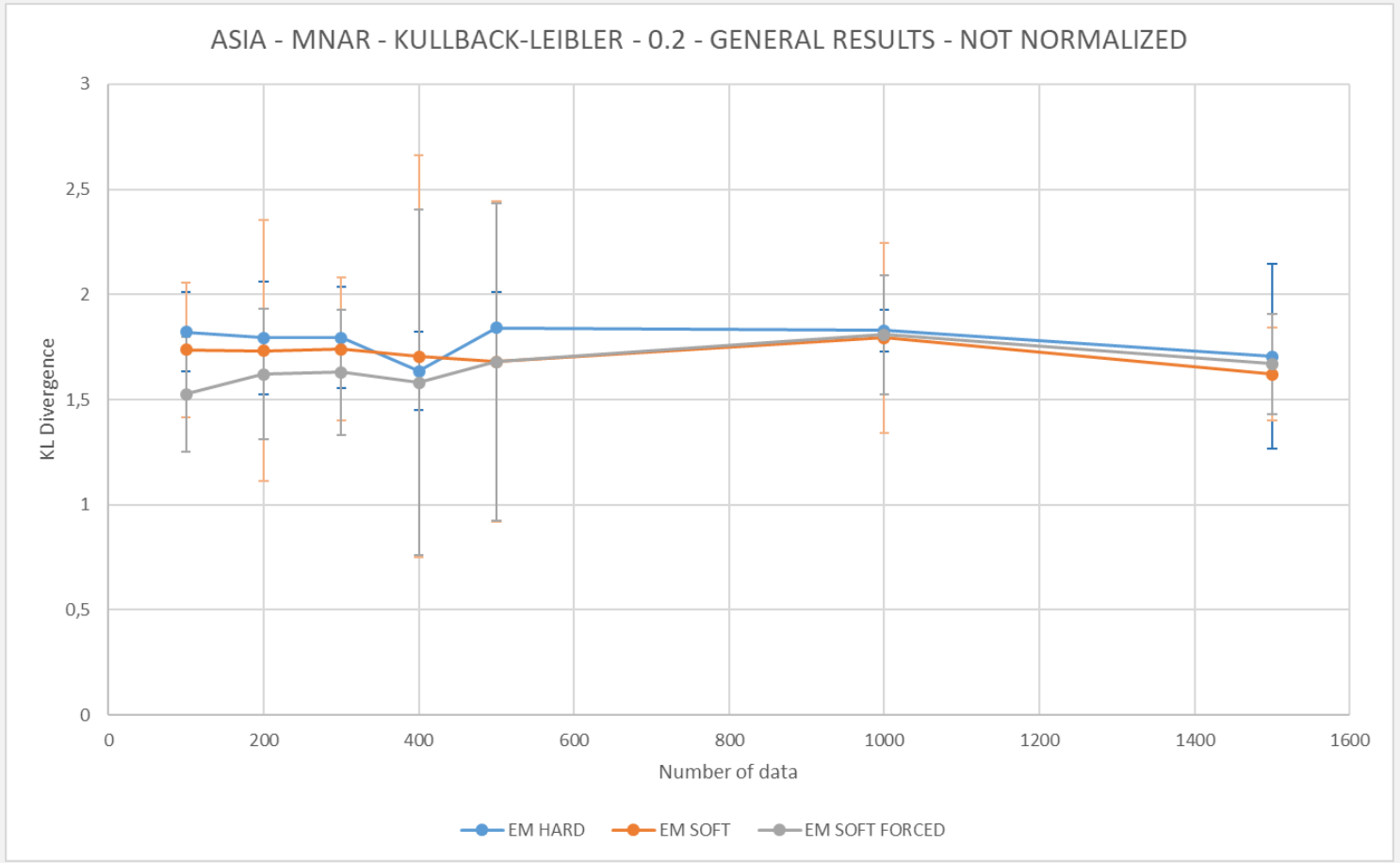
DIFFERENZA IN VALORE ASSOLUTO – ASIA

Prop = 0.2 - NORMALIZED



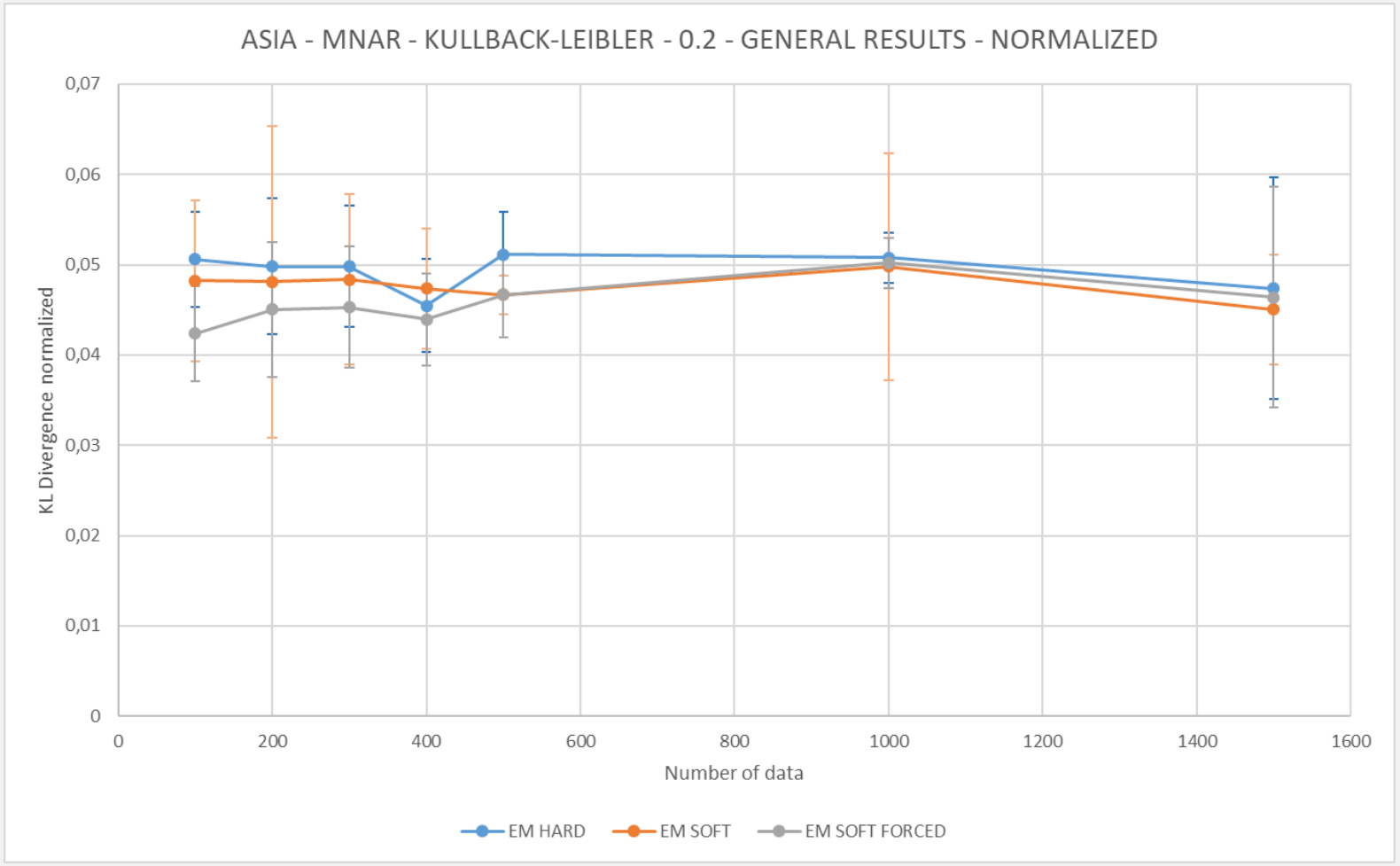
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.2



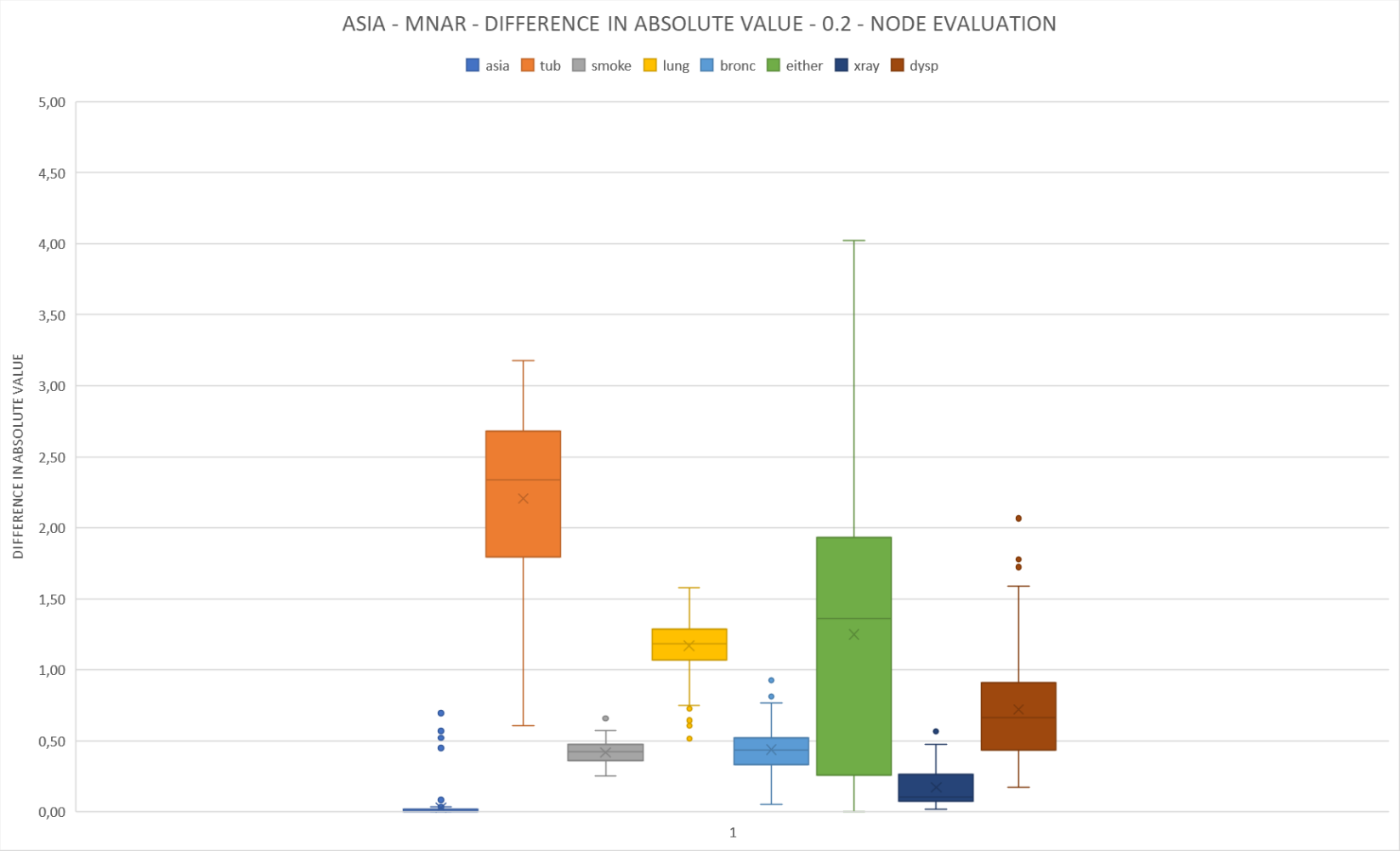
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.1 - NORMALIZED



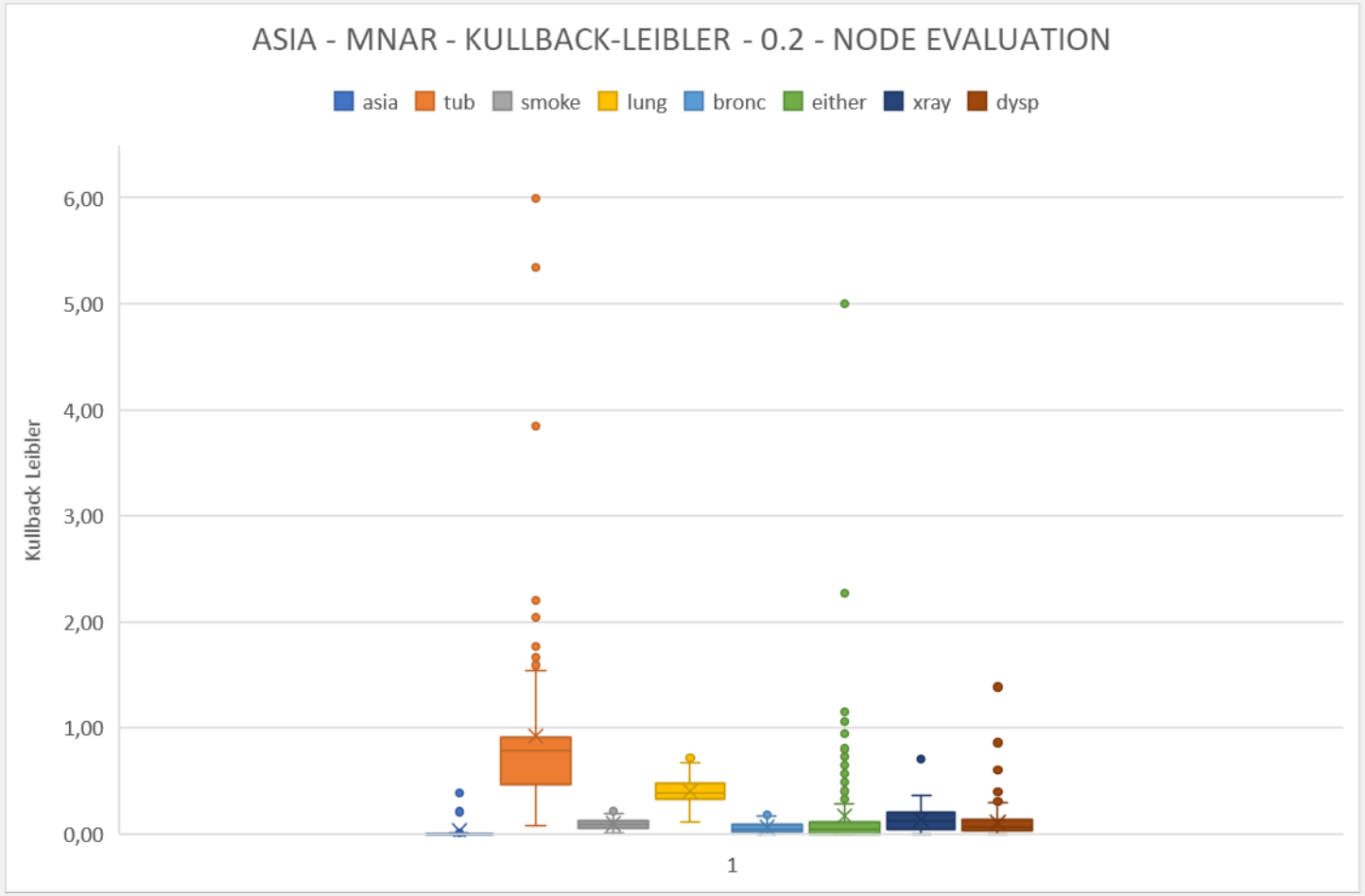
DIFFERENCE IN ABSOLUTE VALUE – ASIA

Prop = 0.2 – NODE EVALUATION



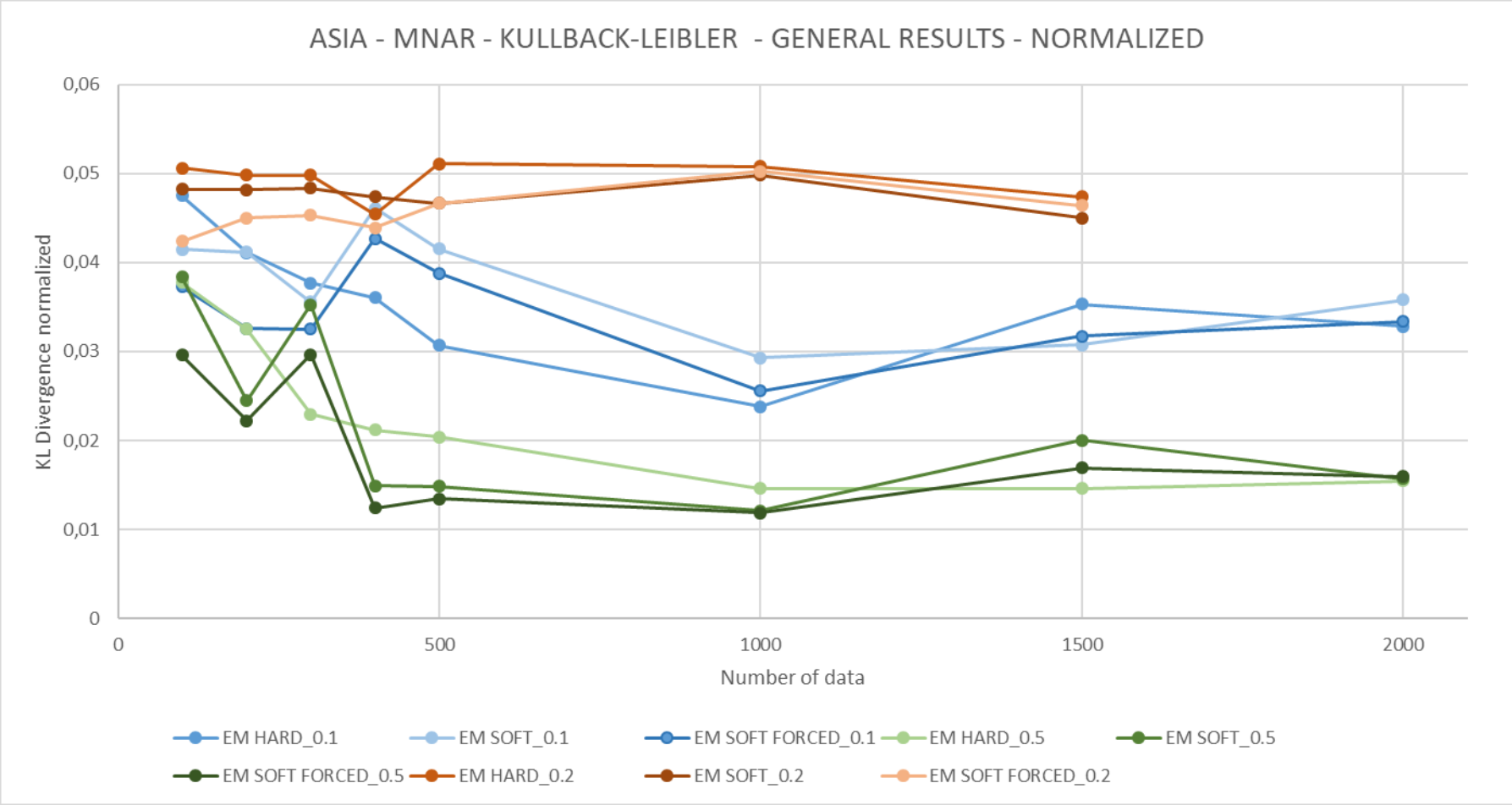
KULLBACK LEIBLER DIVERGENCE – ASIA

Prop = 0.1 – NODE EVALUATION



DIFFERENCE IN ABSOLUTE VALUE – ASIA

RISULTATI GENERALI

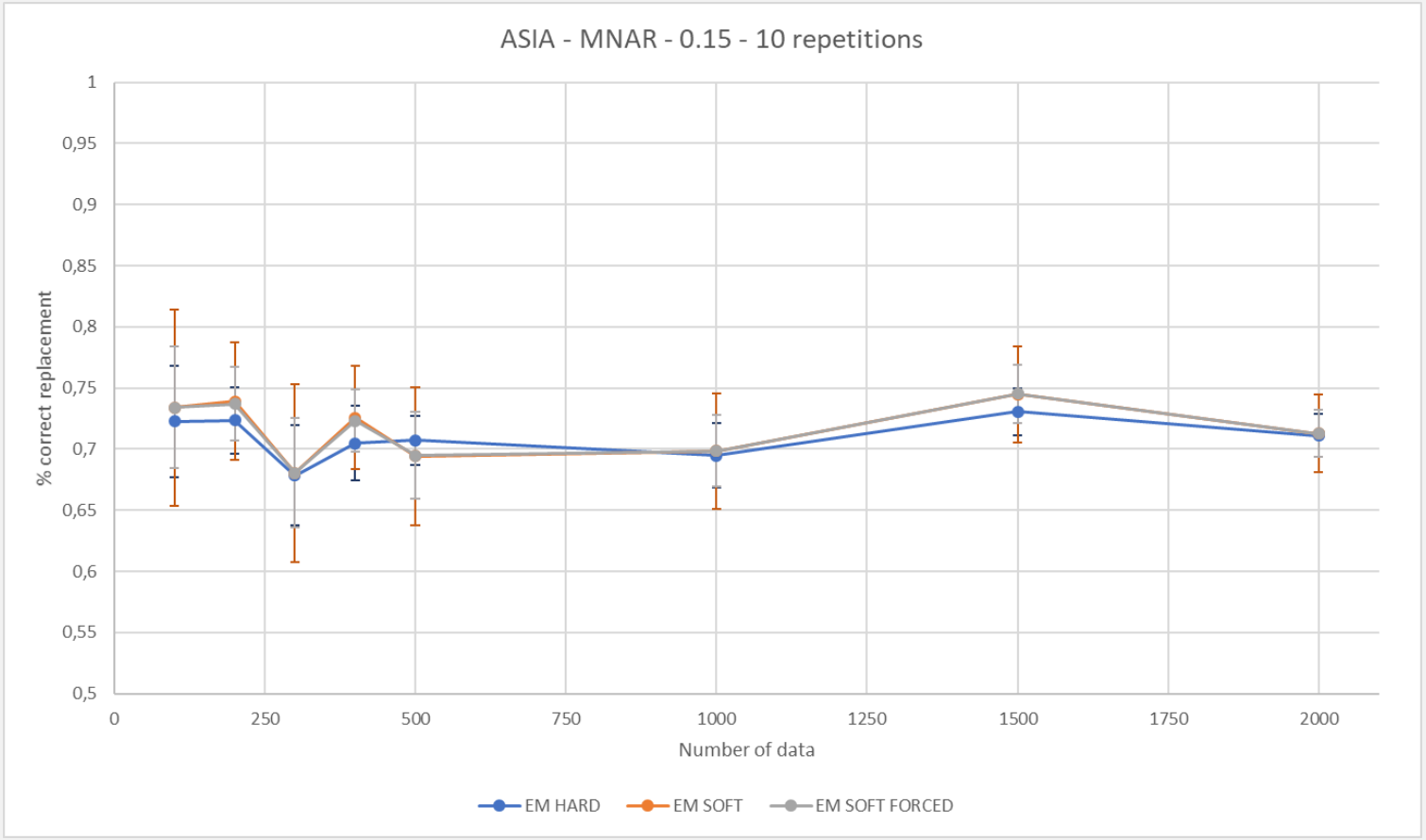


ALTRI RISULTATI

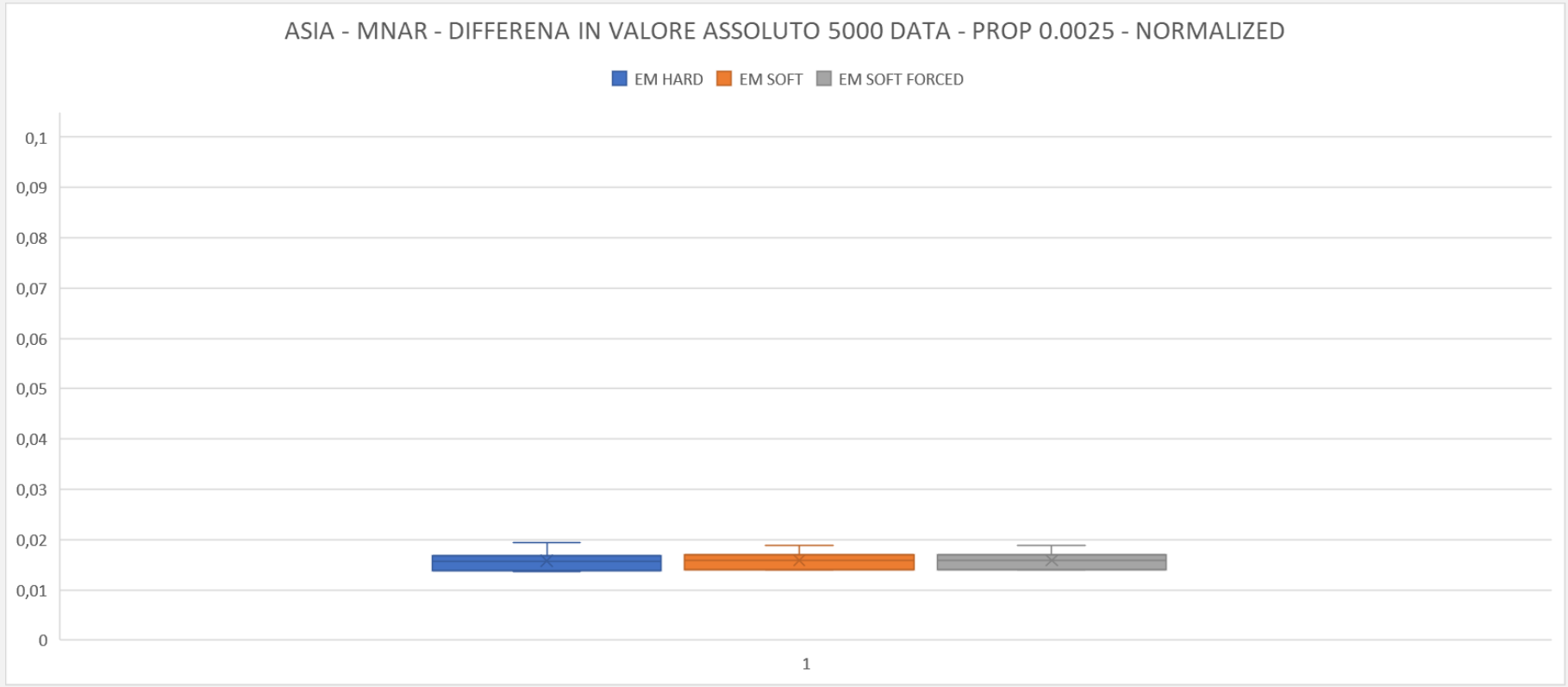
Presentiamo ora i risultati delle diverse analisi effettuate

% CORRECT REPLACEMENT – ASIA

Prop = 0.15



DIFFERENZA IN VALORE ASSOLUTO – ASIA



INTRODUZIONE AI TEST ALARM

Approccio metodologico

Metriche di valutazione

Generazione dei dataset

APPROCCIO METODOLOGICO – ALARM

La maggior parte delle considerazioni generali fatte precedentemente su ASIA valgono anche per la rete ALARM.

Tuttavia, in ALARM cambiano alcuni parametri di simulazione

APPROCCIO METODOLOGICO – ALARM

Il primo iperparametro che cambia in ALARM è **prop**

Ad esempio: se il dataset ALARM contiene 20000 dati e ha 37 variabili, il numero totale di celle è 740.000. Fissando prop = 0.05, il numero totale di celle missing risulta essere di circa 37000

Numero missing values sul dataset ALARM al variare della prop (con tolleranza)	
0,01	0,05
[7.000; 7.800]	[35.000; 39.000]

Un dataset viene ritenuto valido se il numero dei missing values rientrano all'interno di questi intervalli. Vedremo successivamente come i dataset sono stati generati

APPROCCIO METODOLOGICO – ALARM

Comunque, i test sono stati effettuando variando la percentuale delle celle missing per ogni tipologia di test

	Prop	
	0,01	0,05
% correct replacement	SI	SI
Differenza in valore assoluto	SI	SI
Kullback-Leibler divergence	SI	SI

APPROCCIO METODOLOGICO – ALARM

Anche in questo caso, le metriche vengono valutare su 3 varianti dell'algortimo EM

- **EM HARD** (numero massimo di iterazioni pari a 5 per motivi computazionali)
- **EM SOFT** (numero massimo di iterazioni pari a 10 per motivi computazionali)
- **EM SOFT FORCED** (termina alla stessa iterazione di EM HARD, si potrebbe tuttavia stimare l'iterazione di arresto dell'algoritmo)

APPROCCIO METODOLOGICO – ALARM

Prima di entrare nel dettaglio dei test effettuati, introduciamo gli output di ciascuna tipologia di test

	Tipologia di analisi		
	Risultati generali su tutto il dataset	**Generale su tutto il dataset normalizzato [0,1]**	**Analisi node by node**
% correct replacement	**NO**	**SI**	**NO**
Differenza in valore assoluto	**SI**	**SI**	**SI**
Kullback-Leibler divergence	**SI**	**SI**	**SI**

APPROCCIO METODOLOGICO – ALARM

A differenza di ASIA, le grandezze dei campioni dei dataset in ALARM sono: 200, 400, 600, 1000 e 1500

Grandezza del dataset	Numero missing values sul dataset ALARM al variare della prop (con tolleranza)	
	0,01	0,05
200	[59; 89]	[350; 390]
400	[118; 178]	[592; 888]
600	[178; 266]	[888; 1.332]
1000	[296; 444]	[1.580; 2.100]
1500	[444; 666]	[2.300; 3.025]

APPROCCIO METODOLOGICO – ALARM

Per quanto riguarda le metriche di valutazione e la generazioni dei dataset, valgono le stesse indicazioni esposte precedentemente

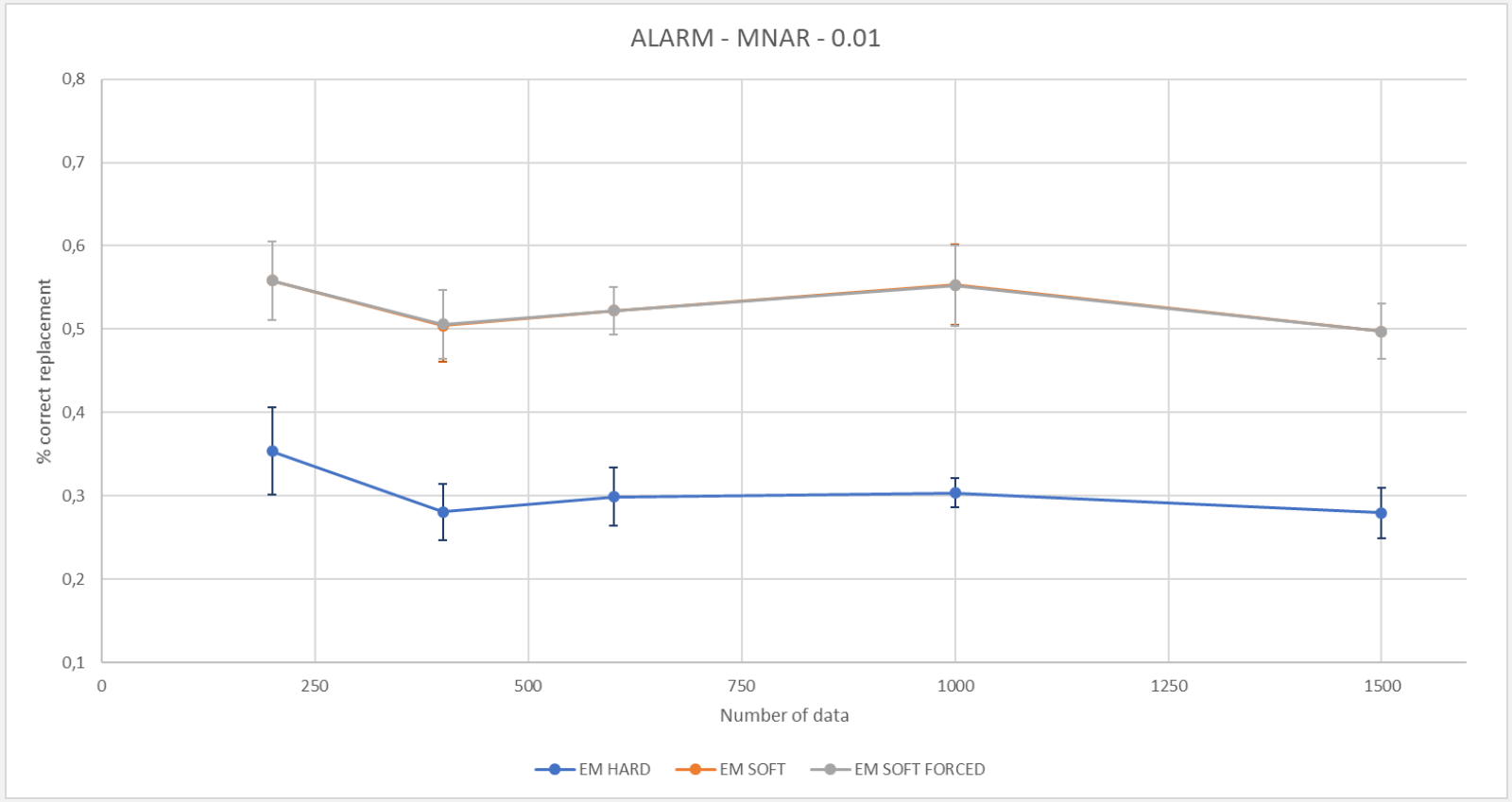
APPROCCIO METODOLOGICO – ALARM

ASIA E ALARM a confronto		
	ASIA	**ALARM**
Numero di variabili	8	37
Numero dati	5000	20000
Prop (proporzione missing values)	0,05; 0,1; (0,15); 0,2	0,01; 0,05
Dimensioni campioni	100; 200; 300; 400; 500; 1000; 1500; 2000; (2500)	200; 400; 600; 1000; 1500
Numero di repliche ad iterazione	10	8 (per motivi di computazione)
Numero di iterazioni massime eseguibili	Max. 20	5 per EM HARD, 10 per EM SOFT
Analisi effettuate	% correct replacement Differenza in valore assoluto Kullback Leibler Divergence	% correct replacement Differenza in valore assoluto Kullback Leibler Divergence

TEST ALARM

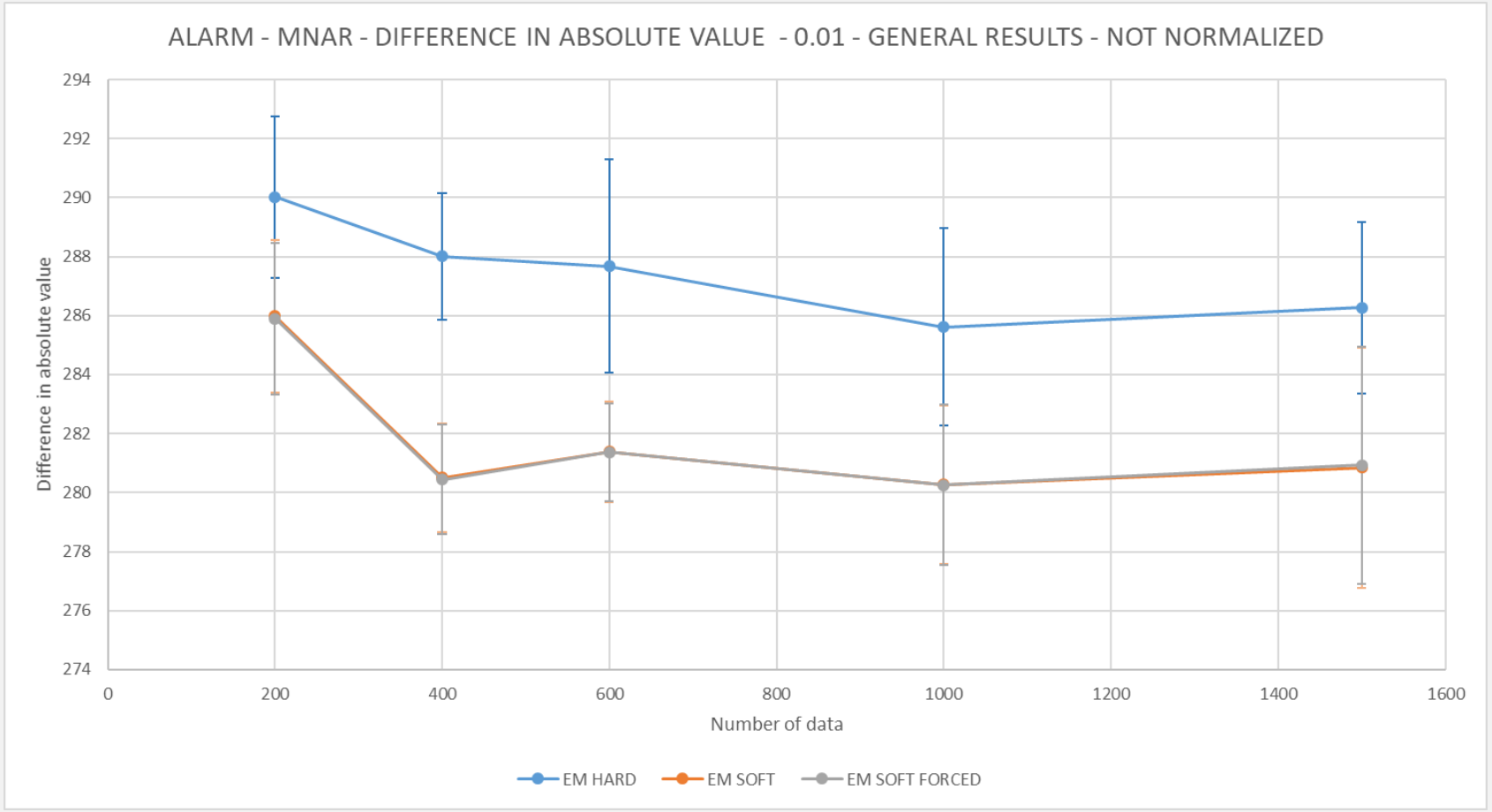
% CORRECT REPLACEMENT – ALARM

Prop = 0.01



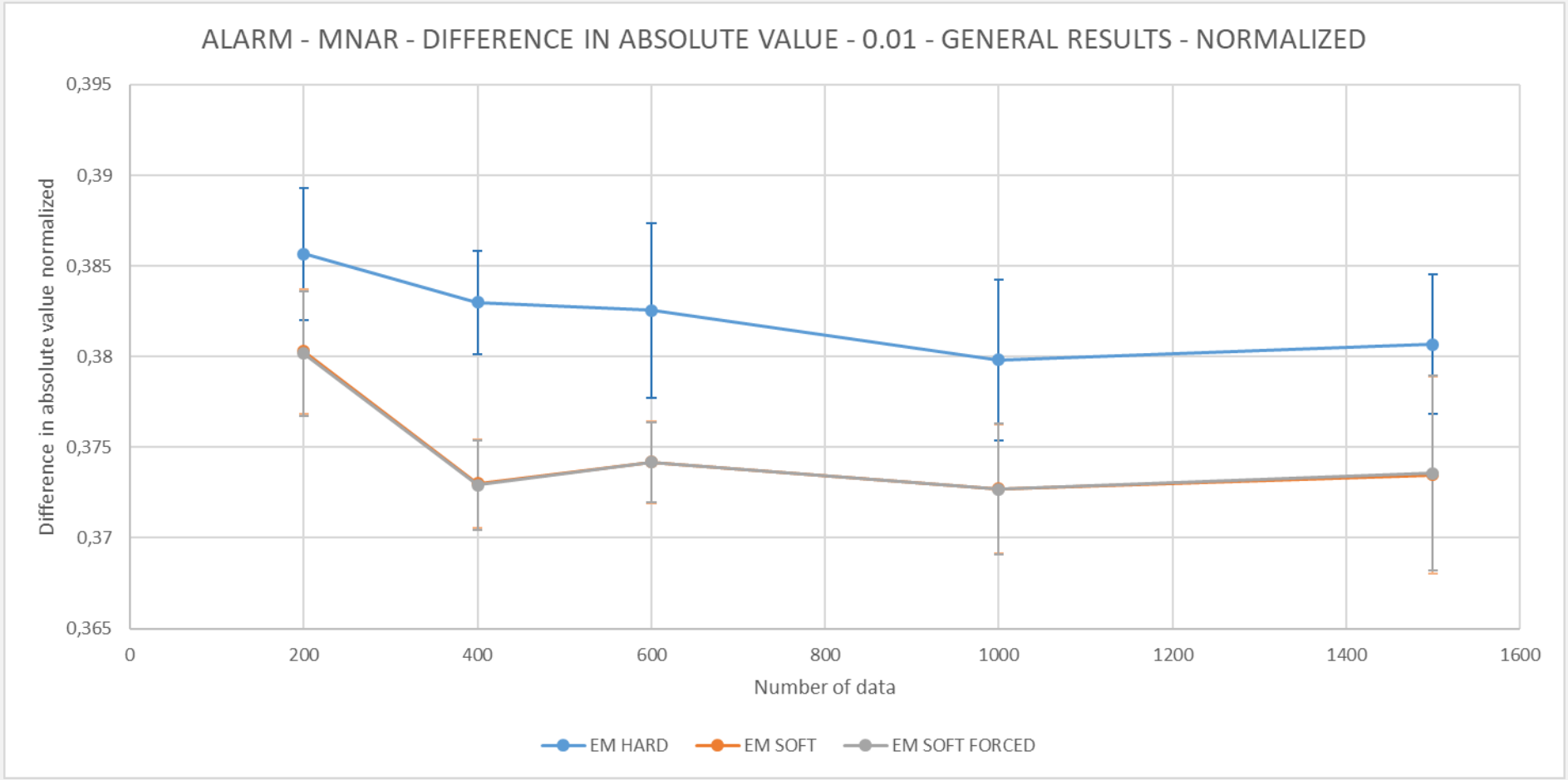
DIFFERENZA IN VALORE ASSOLUTO – ALARM

Prop = 0.01



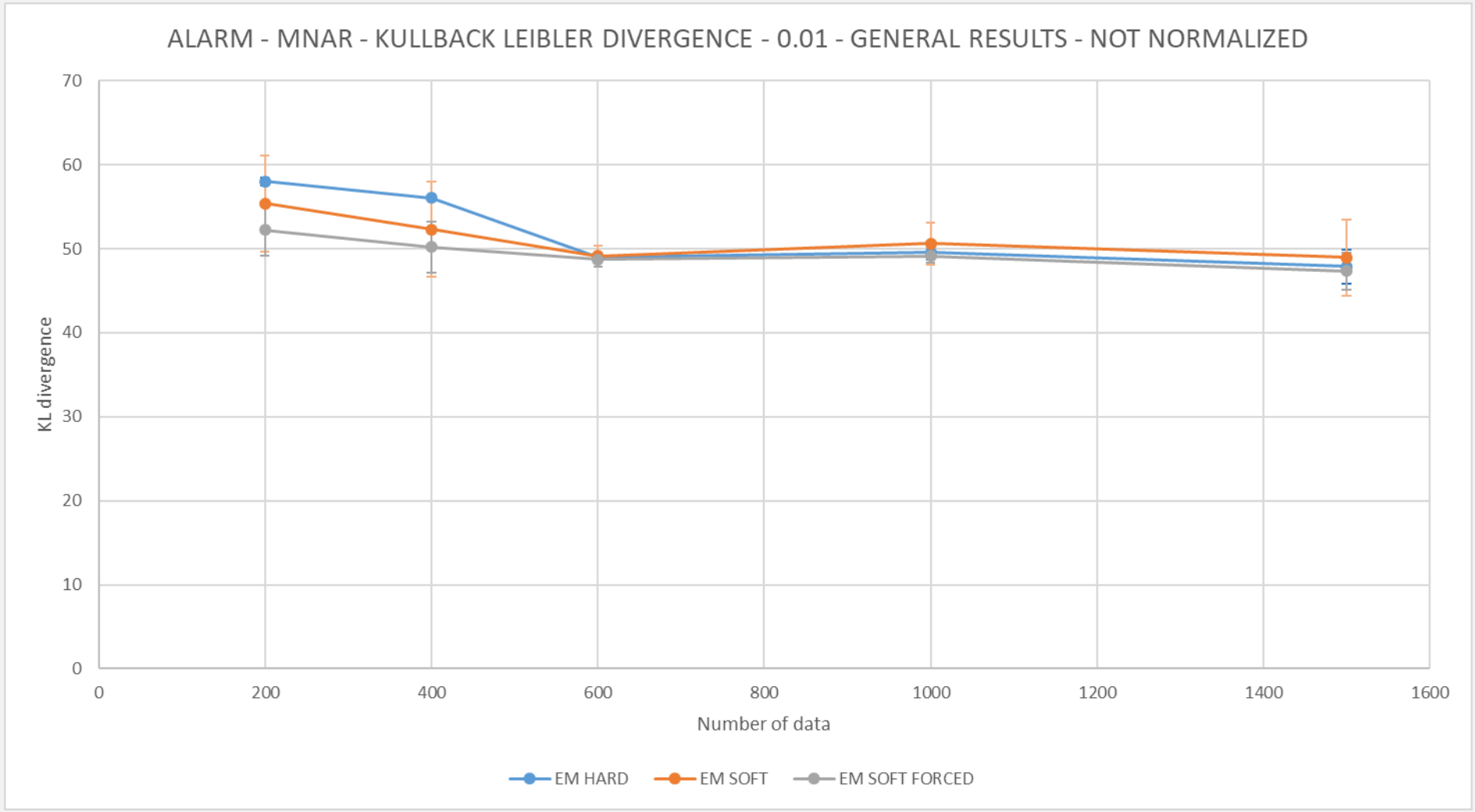
DIFFERENZA IN VALORE ASSOLUTO – ALARM

Prop = 0.01 - NORMALIZED



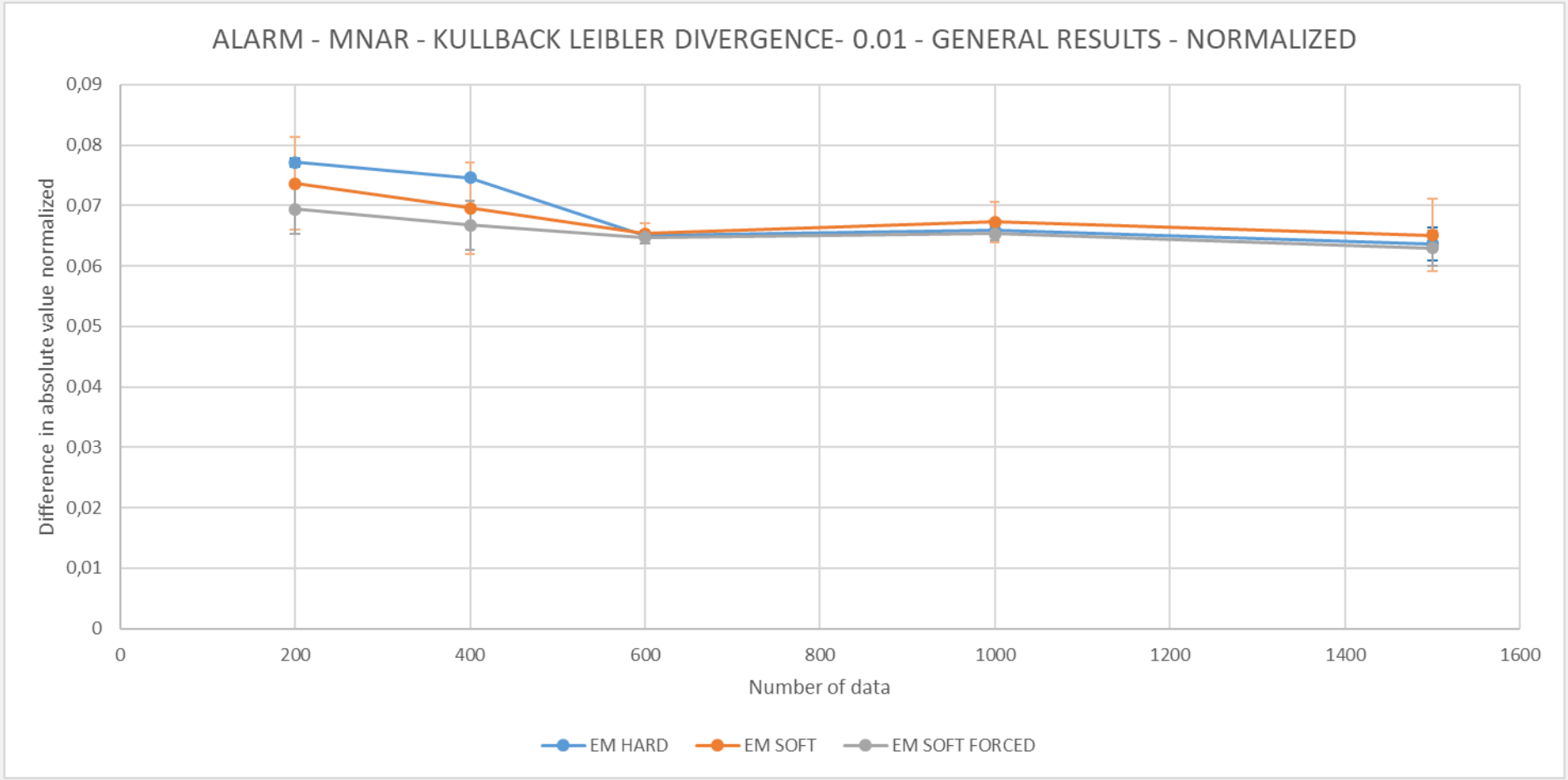
KULLBACK LEIBLER DIVERGENCE – ALARM

Prop = 0.01



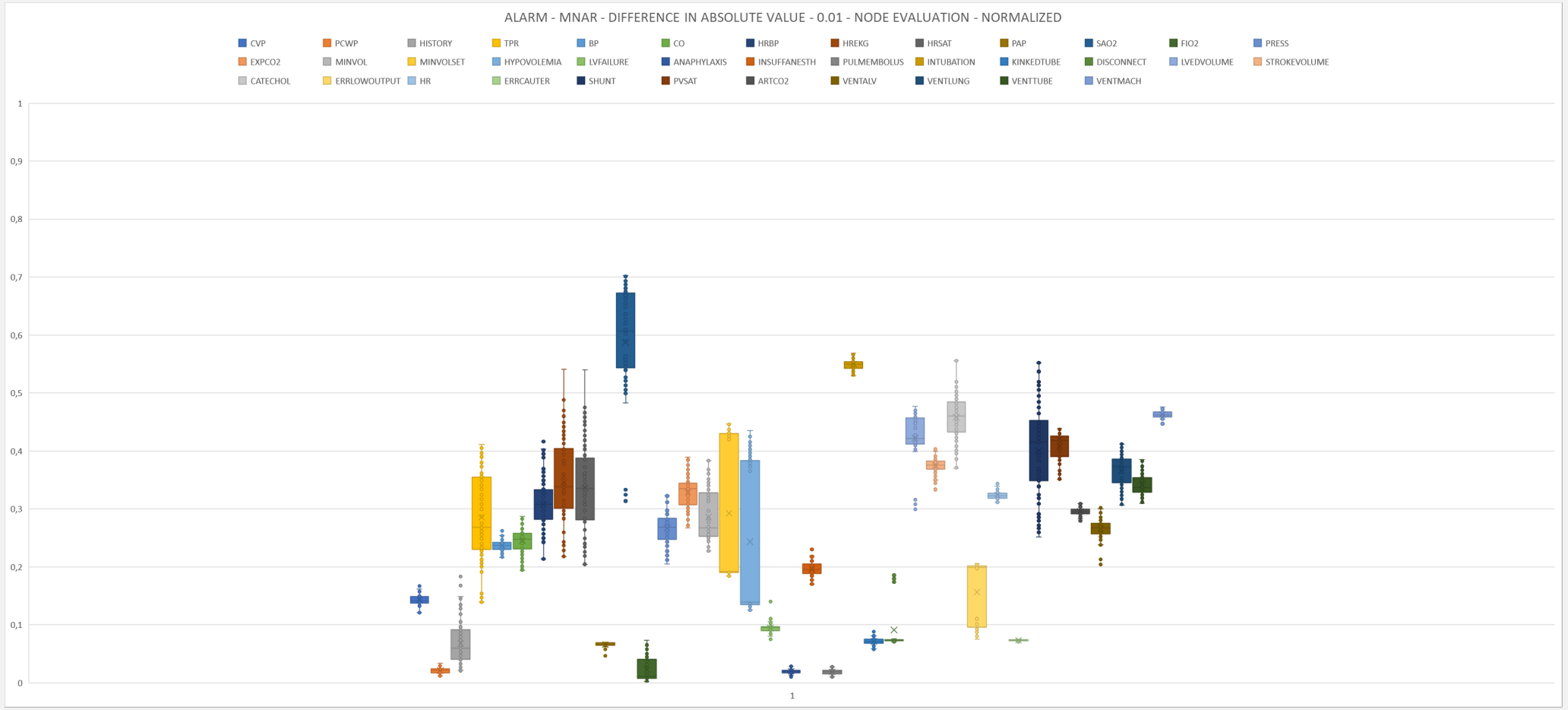
KULLBACK LEIBLER DIVERGENCE – ALARM

Prop = 0.01 - NORMALIZED



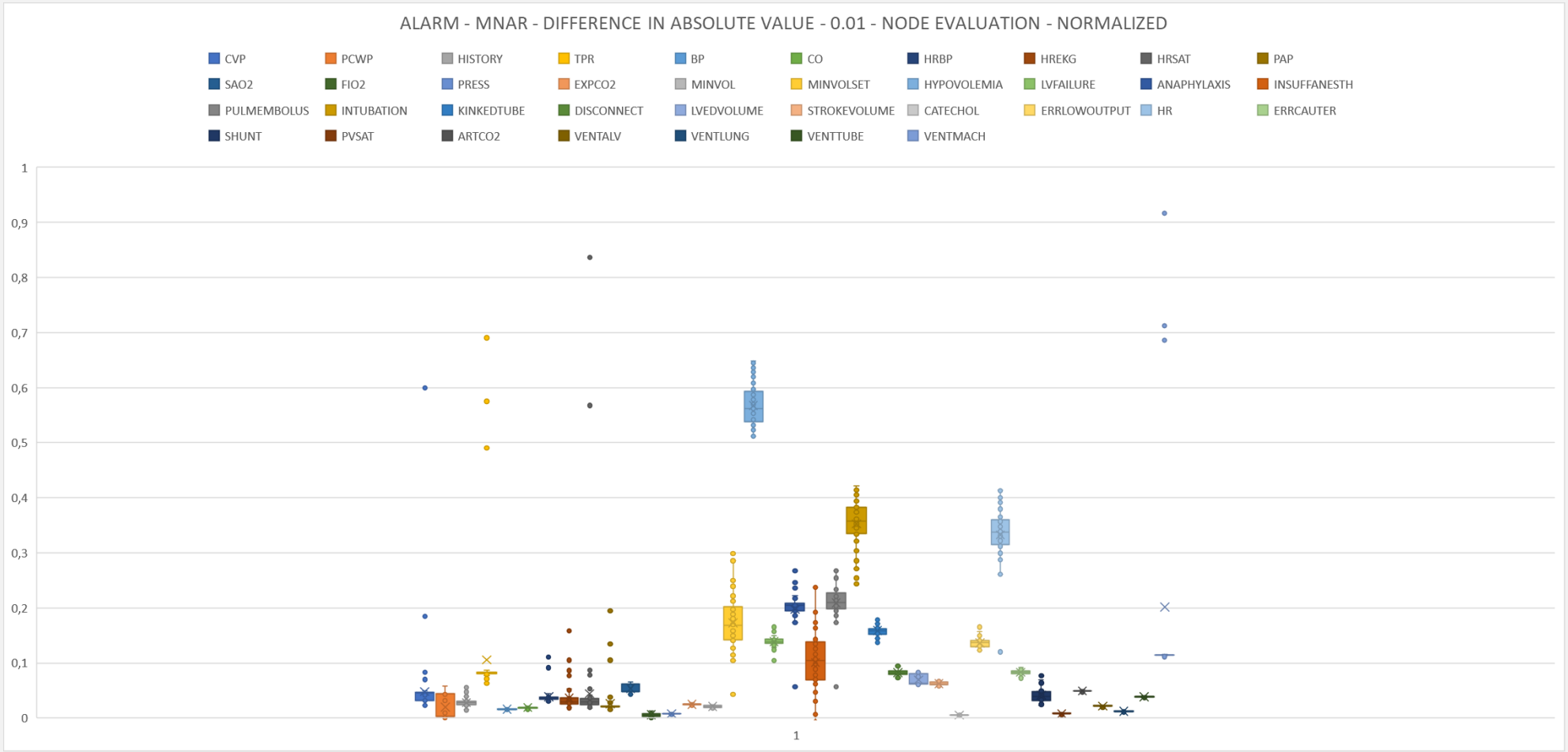
DIFFERENCE IN ABSOLUTE VALUE – ALARM

Prop = 0.01 – NODE EVALUATION



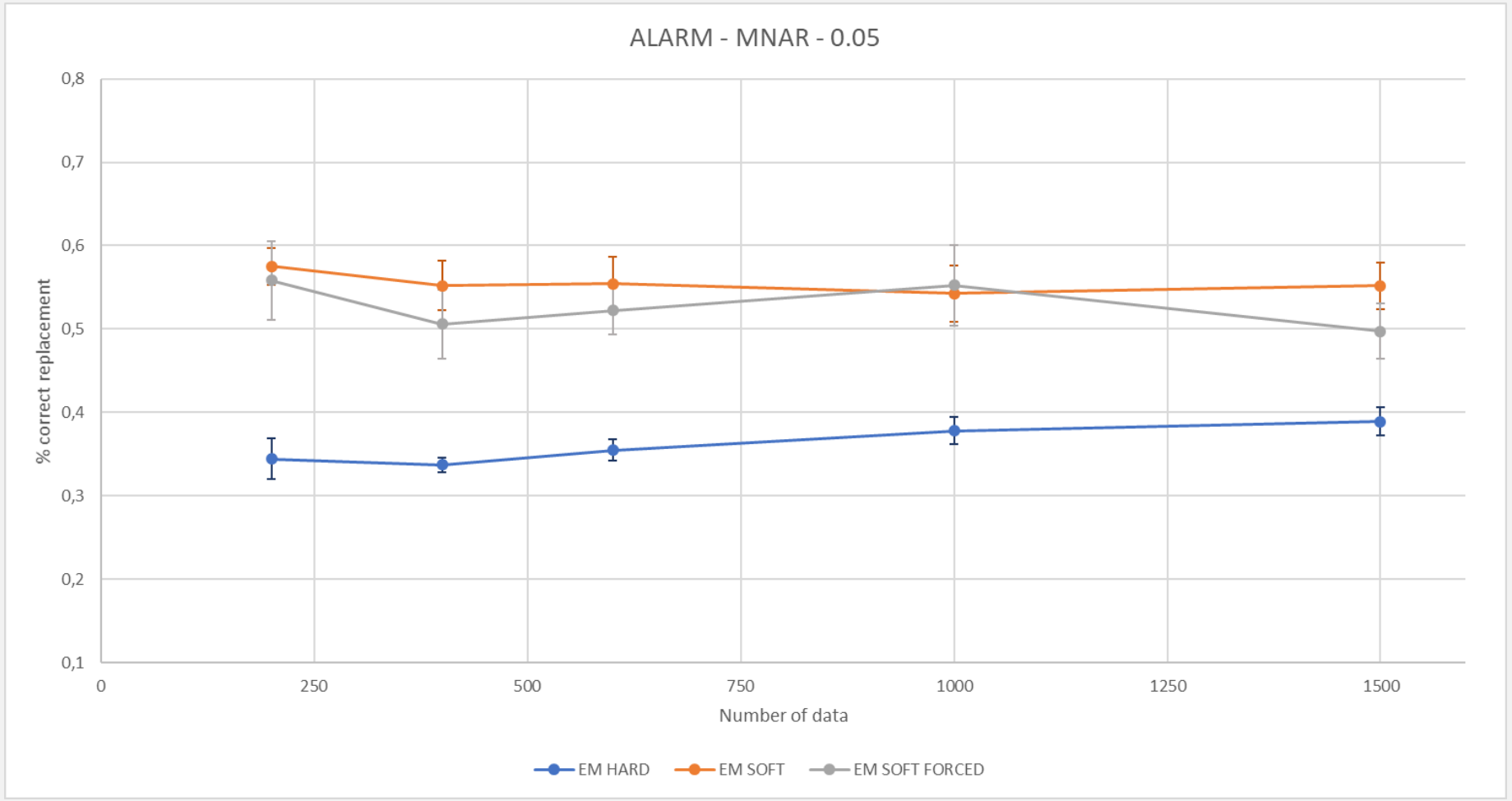
KULLBACK LEIBLER DIVERGENCE – ALARM

Prop = 0.01 – NODE EVALUATION



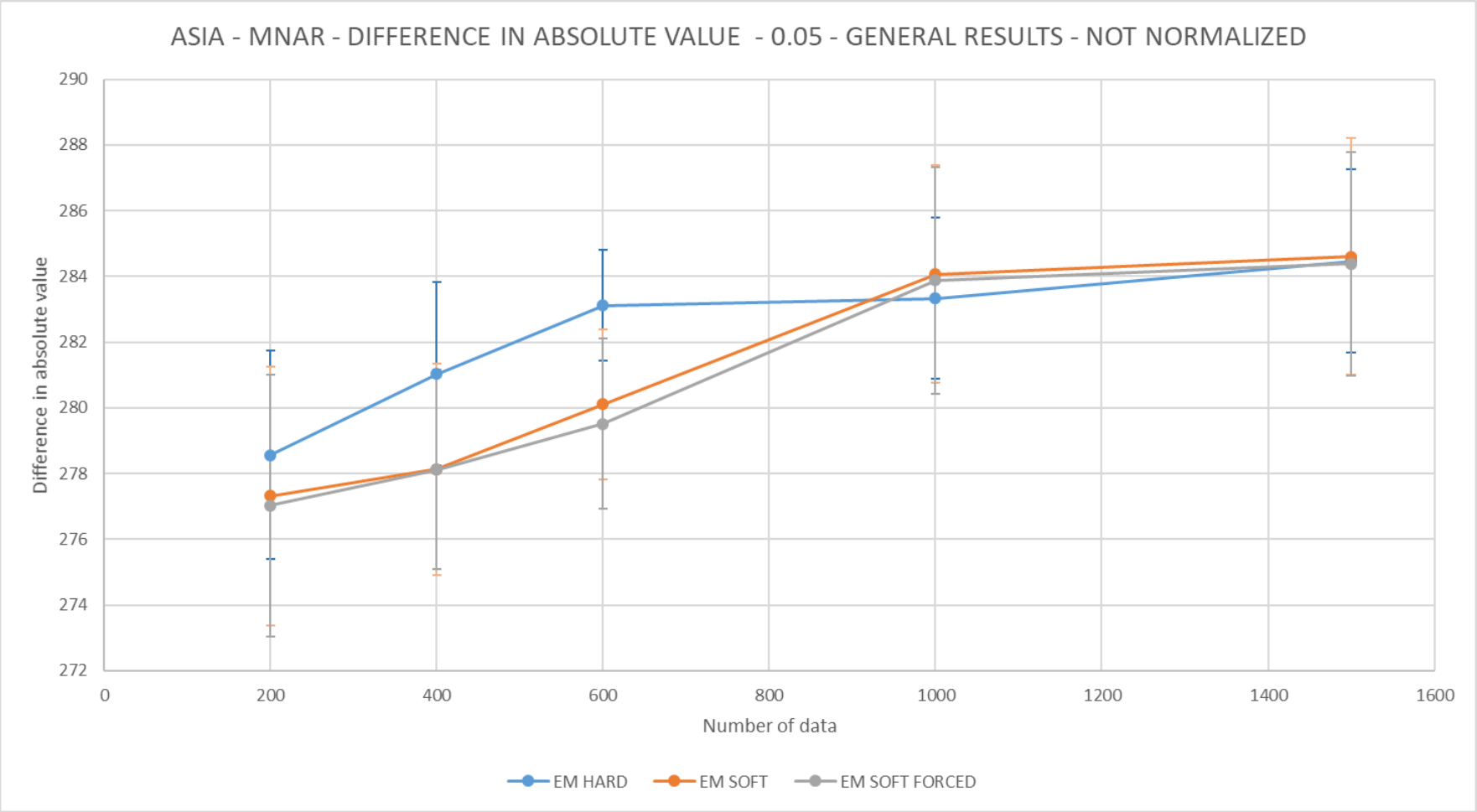
% CORRECT REPLACEMENT – ALARM

Prop = 0.05



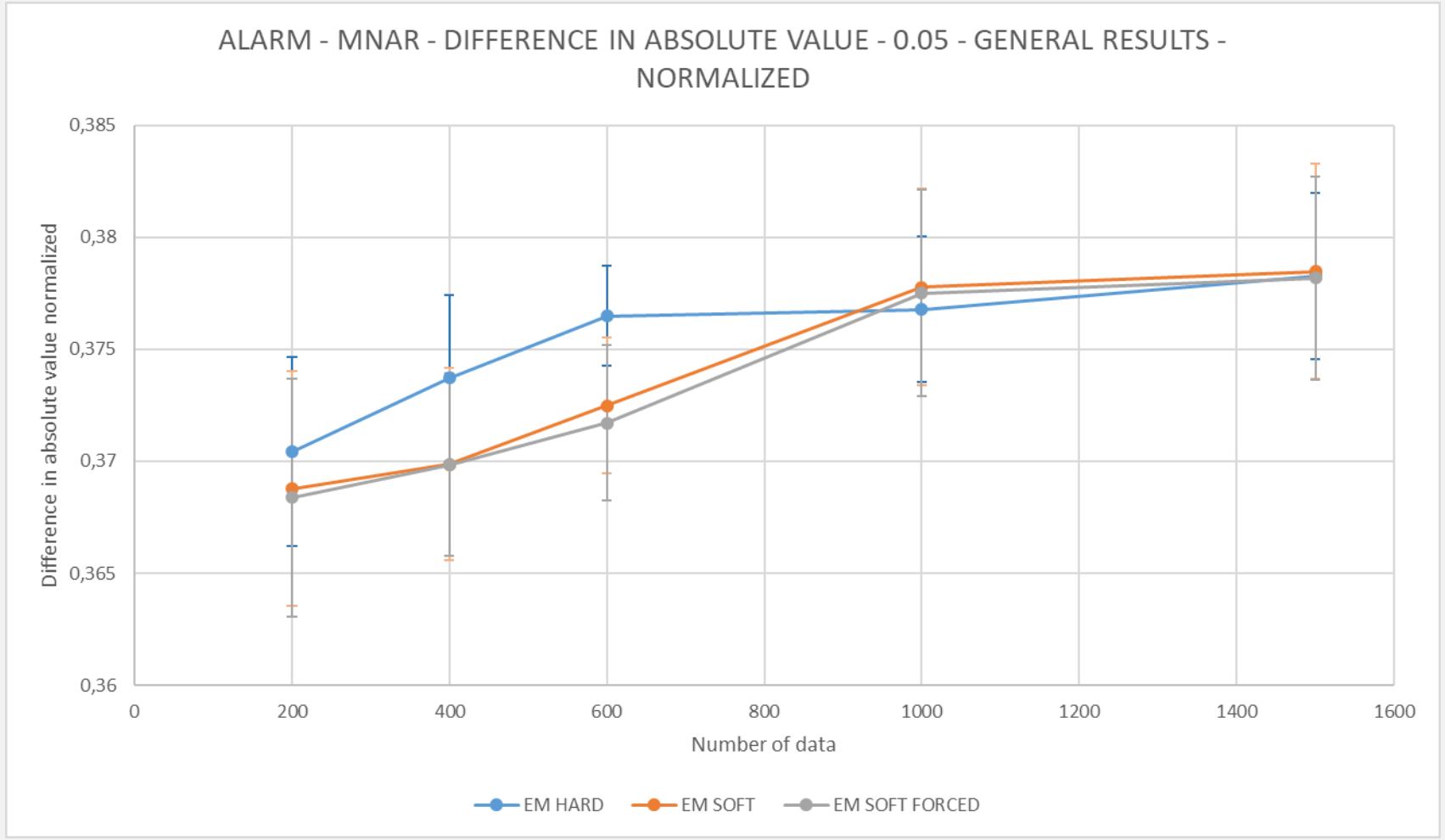
DIFFERENZA IN VALORE ASSOLUTO – ALARM

Prop = 0.05



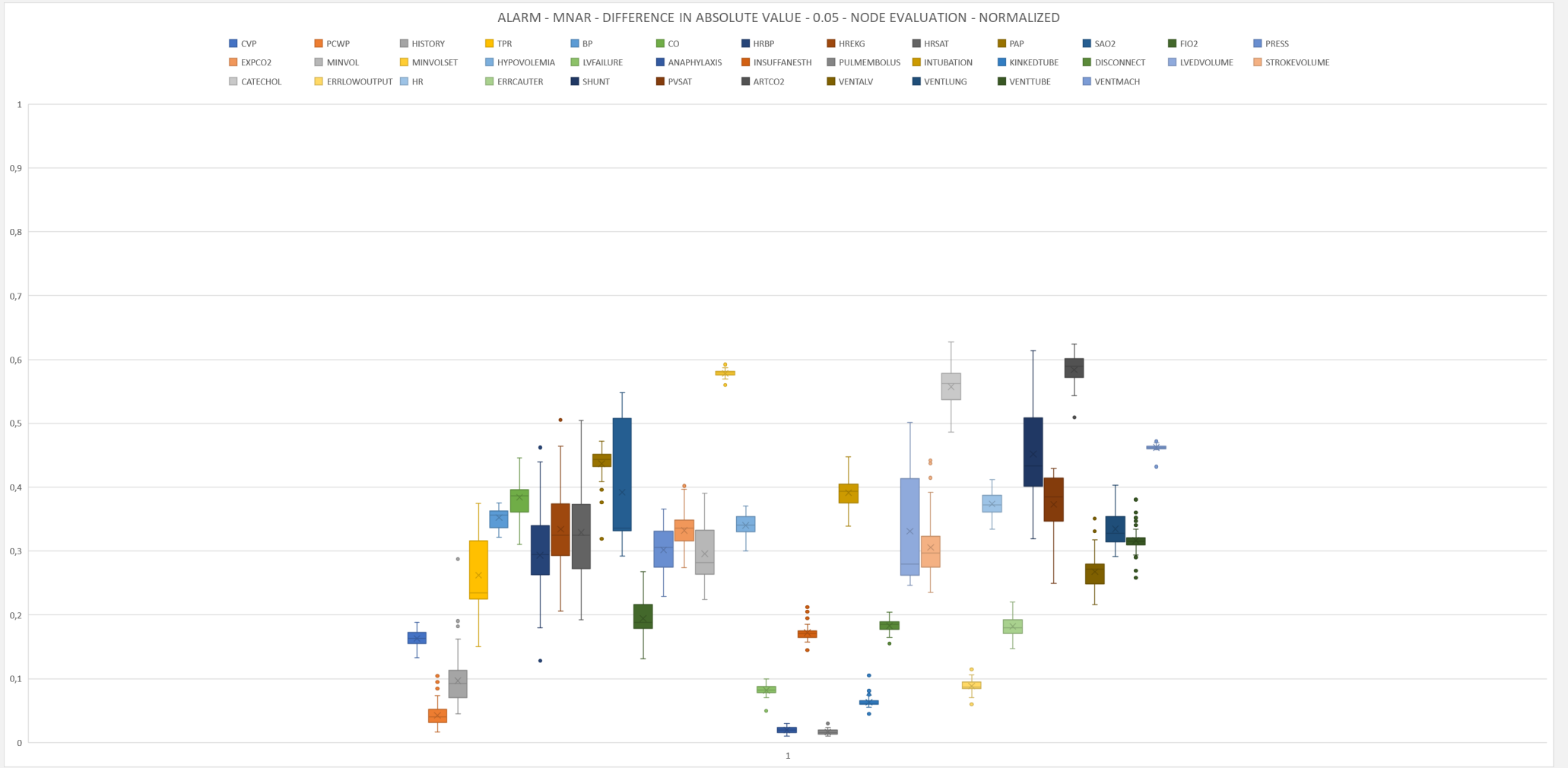
DIFFERENZA IN VALORE ASSOLUTO – ALARM

Prop = 0.05 - NORMALIZED



DIFFERENCE IN ABSOLUTE VALUE – ALARM

Prop = 0.05 – NODE EVALUATION



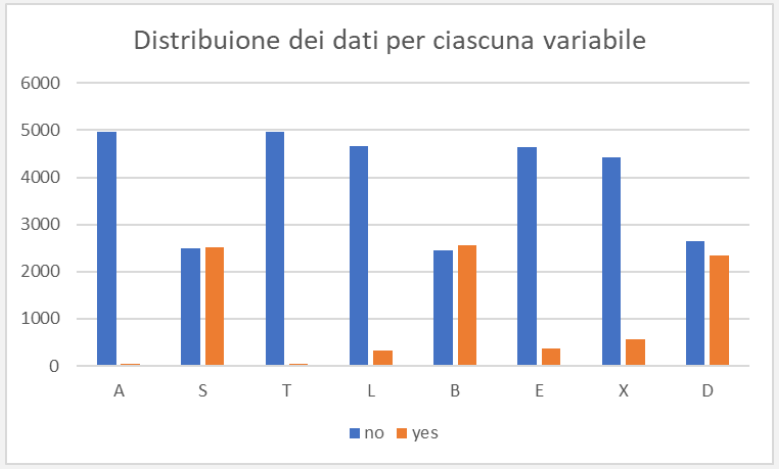
KULLBACK LEIBLER DIVERGENCE – ALARM

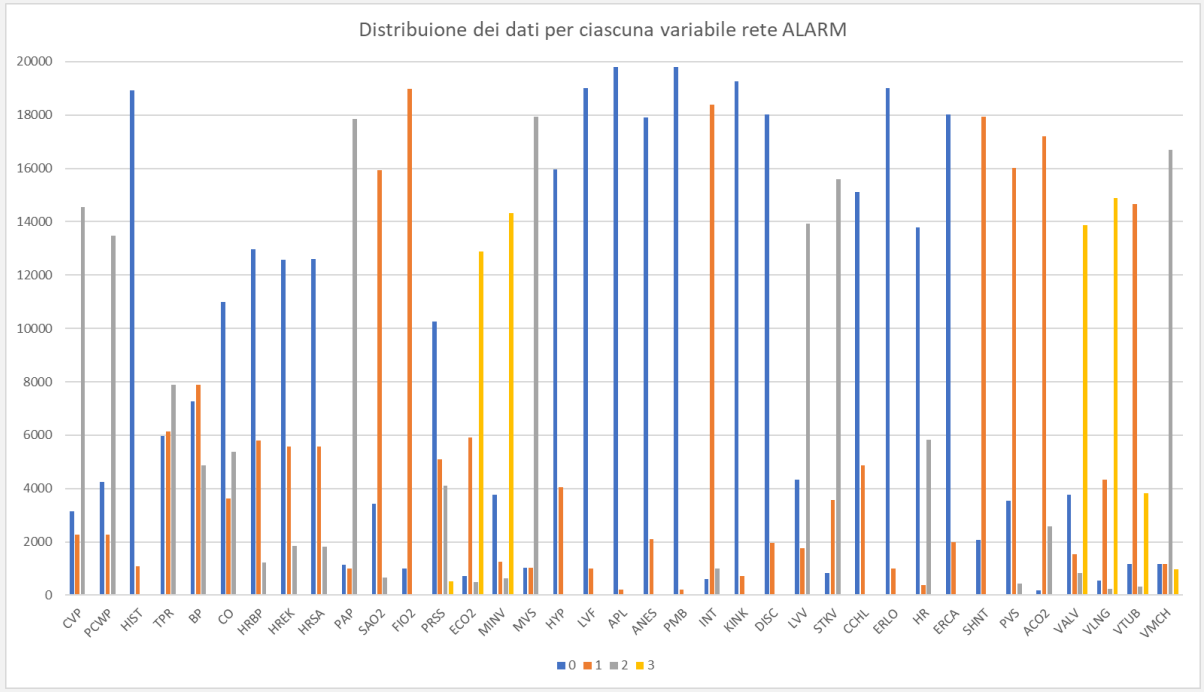
Con l’obiettivo di verificare la correttezza dei valori ottenuti sulla rete ALARM con prop = 0.05, abbiamo momentaneamente ommesso i risultati della KL Divergence

RIASSUNTO E CONCLUSIONI

CONCLUSIONI E RIASSUNTO

Le reti ALARM e ASIA presentano tantissime differenze: ALARM risulta essere una rete molto più complessa rispetto ad ASIA





CONCLUSIONI

ALGORITMI EM A CONFRONTO

Per concludere, la scelta di quale algoritmo utilizzare tra EM HARD, EM SOFT o EM SOFT FORCED dipende sempre dal contesto e non è possibile stabilire a priori qual è l'algoritmo migliore. Tuttavia, di seguito, vengono riassunte alcune osservazioni:

- Se il dataset è costituito da pochi dati, **EM SOFT** è la soluzione preferibile in quanto l'algoritmo impara su tutti i possibili assegnamenti dei valori alle variabili, non limitando il numero di iterazioni. In dataset molto grossi potrebbe impiegare un alto tempo computazionale;

- Se il dataset è costituito da tanti dati ma la proporzione dei dati missing è bassa, **EM SOFT FORCED** è la soluzione preferibile. Si è osservato (soprattutto nella rete ALARM) che EM HARD potrebbe avere prestazioni più basse quando il numero di dati è limitato;

- Se il dataset è costituito da tanti dati, **EM HARD** è la soluzione preferibile, considerando anche i tempi computazionali. In ALARM si è notato che all'aumentare di PROP, le prestazioni di EM HARD tendono ad aumentare mentre quelle di EM SOFT tendono a diminuire;

- In generale, se la rete presenta **distribuzioni di probabilità uniformi** (massima incertezza), le prestazioni di EM HARD ne risentono molto e EM SOFT risulta essere consigliabile.