Evaluation wit the Minimum Interval of Equivalence

Bayesian Networks in Software Testing

Javier Dolado
U. País Vasco/Euskal Herriko Unibertsitatea

Daniel Rodríguez
Universidad de Alcalá

PRESI TIN2013-46928-C3-1-R, TIN2013-46928-C3-2-R

Two problems we are researching

- **How to compare estimation methods?**

  Article finished in collaboration with Harman's group

- **The use of Bayesian concepts and Bayesian Networks in Software Testing**

  Article in progress with several collaborators

Create k Folds in N datasets

Apply M Estimation Methods with different sets of parameters

Select intervals

DATASET 1

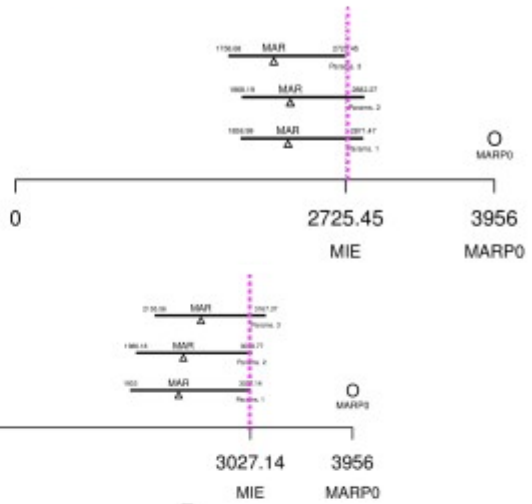
Est	Effort
670	691
912	902
218	274
595	479
267	308
344	301
229	234
190	171
869	333
109	159
289	238
616	373
557	308
416	558
578	861
438	423

seed1

T
t-f1
T
T
T
t-f2
t-f3
T
T

GP
LR
MLP
RTree

Bootstrap CIs



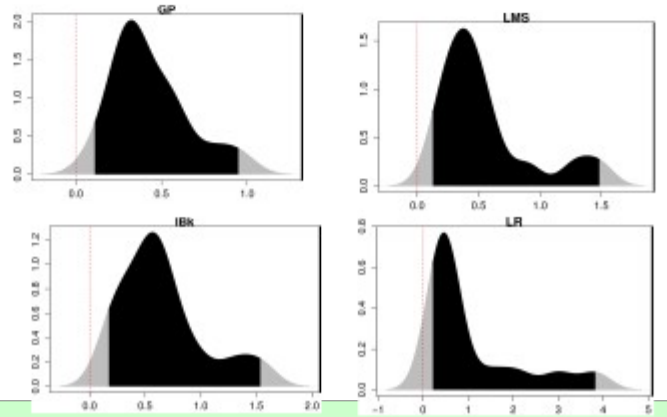
MARP0s

Compute the k·N·M MIEratios

DATASET N

Est	Effort
670	691
912	902
218	274
595	479
267	308
344	301
229	234
190	171
869	333
109	159
289	238
616	373
557	308
416	558
578	861
438	423

Compute the Credible Intervals of the MIEratios for the M methods



N datasets and M methods

One dataset and One method

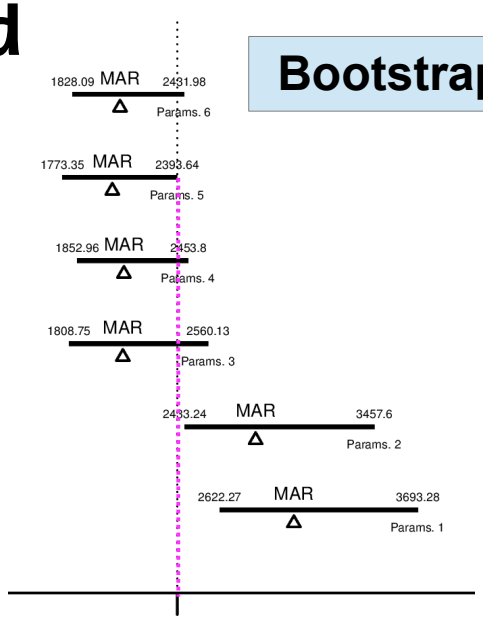
Dataset 1

	Est Effort
670	691
912	902
218	274
595	479
267	308
344	301
229	234
190	171
869	333
109	159
289	238
616	373
557	308
416	558
578	861
438	423

ESTIMATION METHOD 1

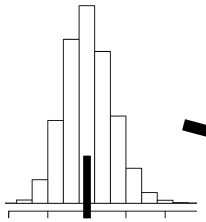
Parameters 1
Parameters 2
…
Parameters n

Bootstrap CIs



Select interval

Compute MARP0



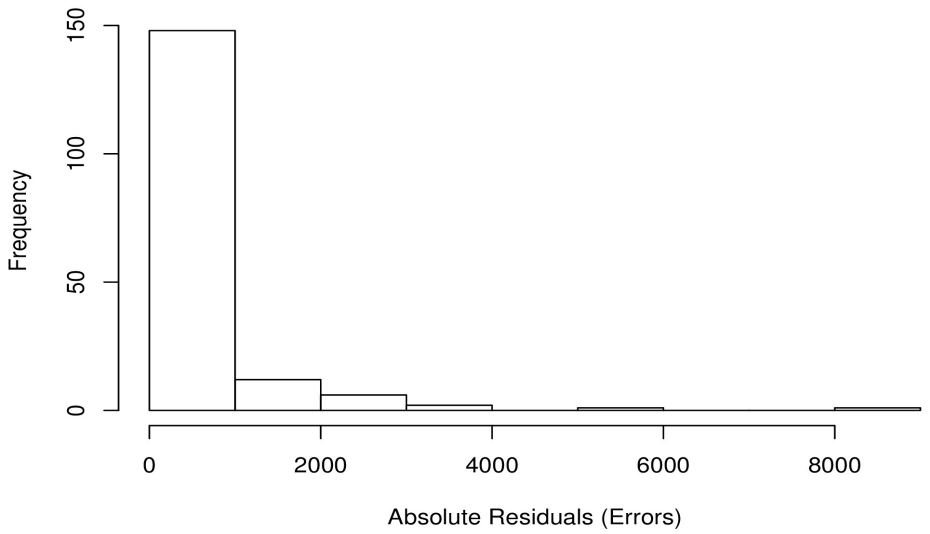
Compute MIEratio

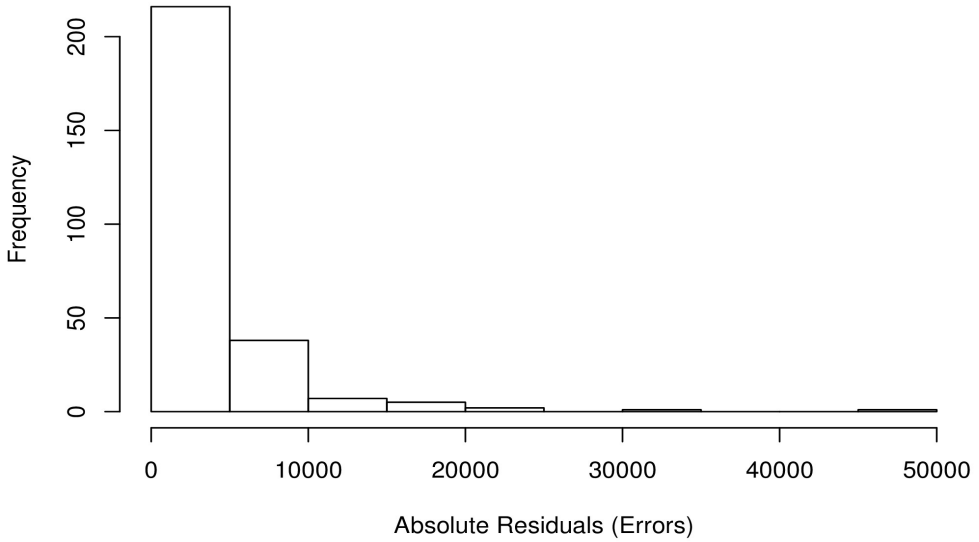
Process followed for **one dataset** and **one estimation method**:

1. Select a set of parameters for the estimation method (“Parameters 1”)
2. Bootstrap a Confidence Interval for the Mean Absolute Error (“MAR”)
3. Repeat the process for “n” sets of parameters
4. Select the best interval, i.e. , that one with the upper limit closest to 0
5. Compute the MARP0
6. Compute the MIEratio

Estimation Errors

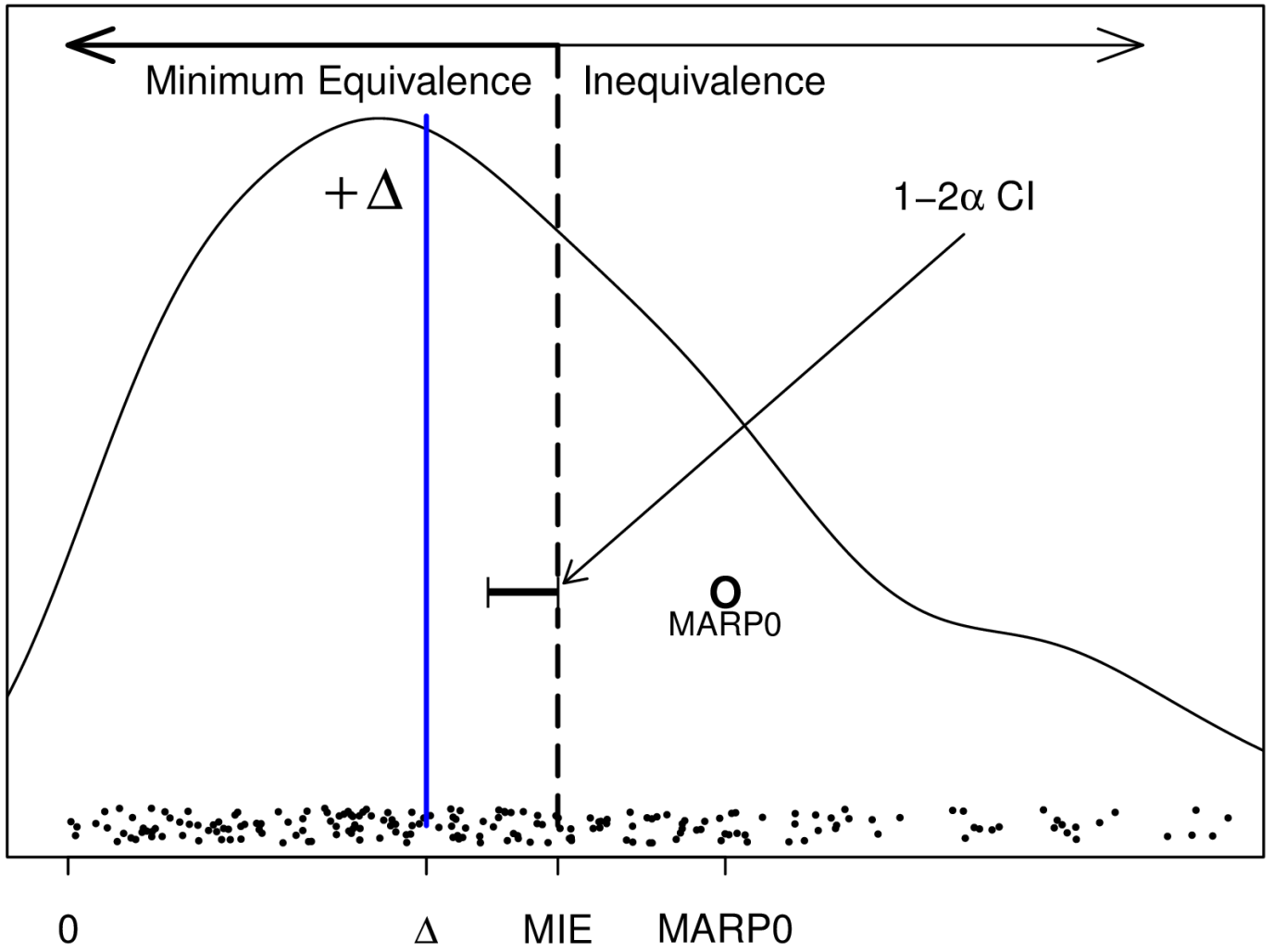
- The distribution of the absolute errors of the estimations is non-normal
- Therefore, many of the usual measures based on the arithmetic mean are not valid
- We use geometric means and bootstrap (non-parametric method)





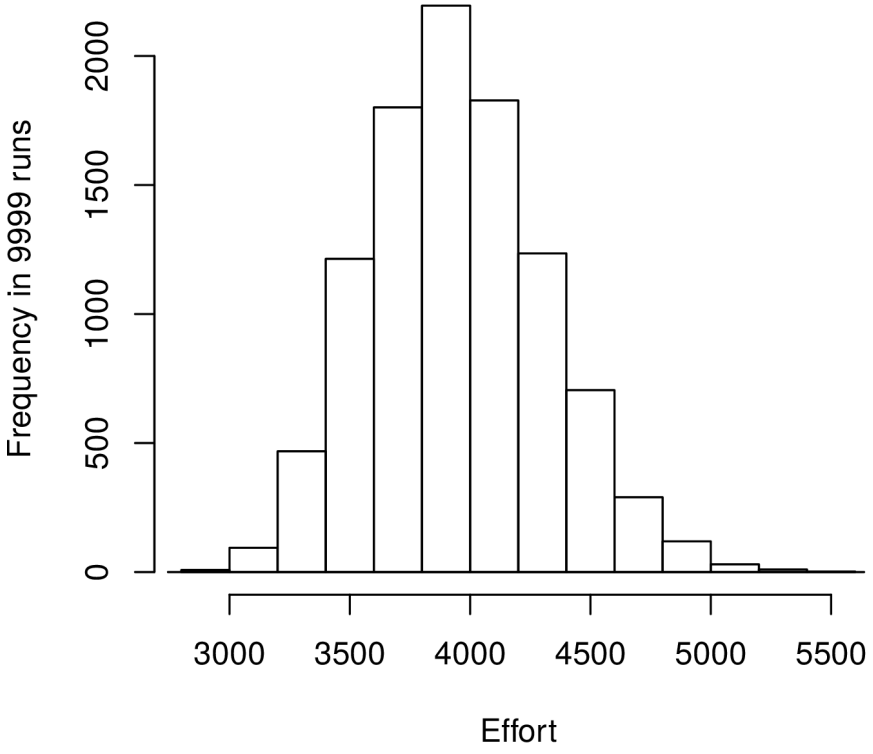
Selection of the intervals

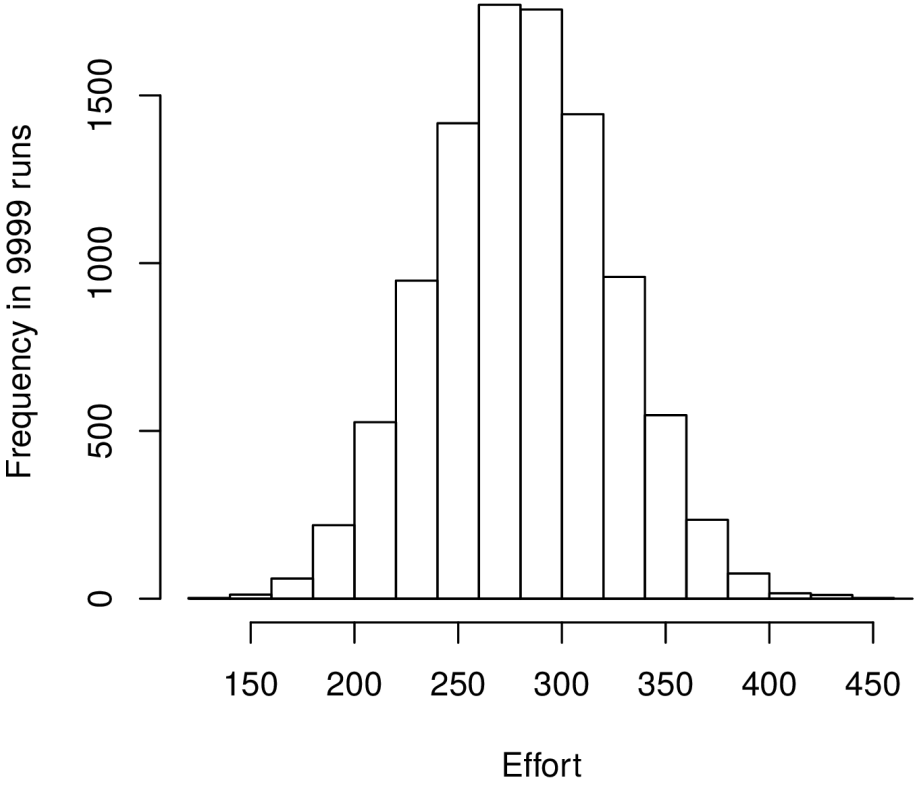
- We compute the 1-2α confidence intervals by resampling methods (with bootstrap)
- We select the closest intervals to the origin 0



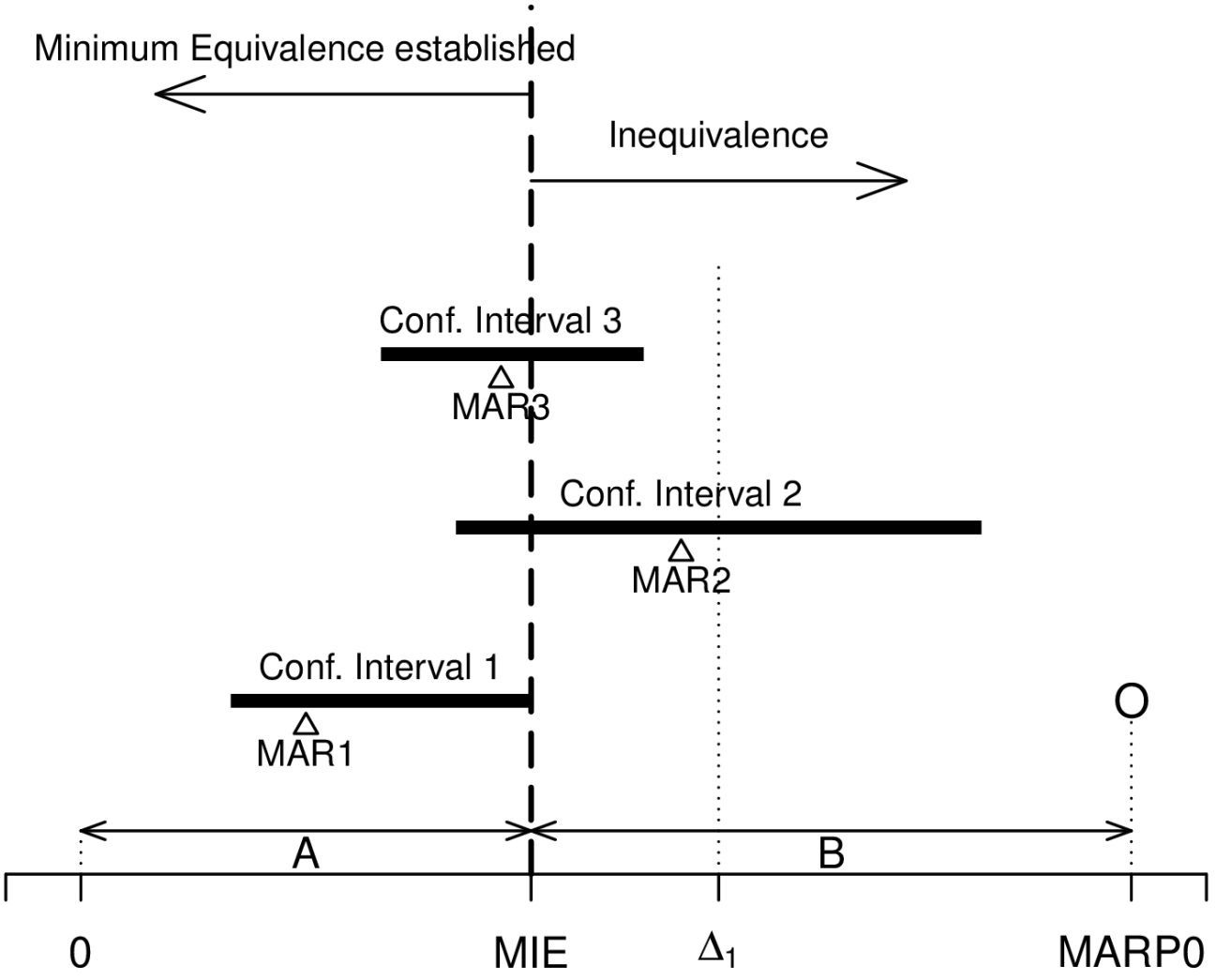
MARP0s: random estimation

- In order to compare the estimations with “something” we resort to the concept of random estimation (Hyndman, Shepperd, MacDonell)
- Random estimation is the “worst estimation” that could be performed in a dataset





Selection of the intervals of equivalence



Ordered Table of the MIEratios

- After all computations we get something like the table below

Usual measures used MMRE, Pred(0.25) are not correlated with our current values

Method	Dataset	$\overline{MAR}_{P_0}$	MAR	gMAR	MMRE	MdMRE	Prd(0.25)	MIE	SA	**MIEratio**
GP	CSC(fld1)	2354.836	540.467	104.638	0.164	0.115	0.667	231.078	0.770	**0.109**
LMS	CSC(fld1)	2354.836	558.943	129.680	0.167	0.157	0.733	258.696	0.763	**0.123**
IBk	CSC(fld1)	2354.836	644.667	171.551	0.187	0.129	0.667	344.652	0.726	**0.171**
MLP	CSC(fld2)	2228.334	515.881	163.979	0.242	0.178	0.533	345.462	*0.768*	**0.183**
LMS	CSC(fld2)	2228.334	594.199	186.718	0.229	0.250	0.533	371.337	*0.733*	**0.200**
IBk	CSC(fld2)	2228.334	532.507	191.963	0.242	0.179	0.600	371.378	*0.761*	**0.200**
LR	Maxwell(fld3)	16636.643	2477.946	1609.117	0.552	0.094	0.667	2947.914	*0.851*	**0.215**
GP	Maxwell(fld1)	7346.358	2132.182	332.289	0.240	0.241	0.571	1344.906	0.710	**0.224**
MLP	CSC(fld1)	2354.836	596.663	239.629	0.272	0.226	0.533	433.527	*0.747*	**0.226**
GP	CSC(fld2)	2228.334	527.827	225.176	0.325	0.215	0.600	417.198	*0.763*	**0.230**
LMS	ISBSG(fld2)	4367.901	2125.949	666.624	0.814	0.486	0.156	868.887	0.513	**0.248**
MLP	China(fld3)	5331.634	2136.744	781.631	1.291	0.582	0.260	1083.683	*0.599*	**0.255**
M5P	Maxwell(fld3)	16636.643	2904.573	1667.502	0.461	0.132	0.833	3395.997	*0.825*	**0.256**
Telecom1	Telecom1	270.725	86.389	29.999	0.343	0.191	0.722	57.192	*0.681*	**0.268**
LMS	China(fld3)	5331.634	2206.317	834.287	1.630	0.550	0.220	1155.405	*0.586*	**0.277**

Table 2: This table shows the summary results for the best 15 values in ascending order of MIEratio ($\alpha = 0.05$).

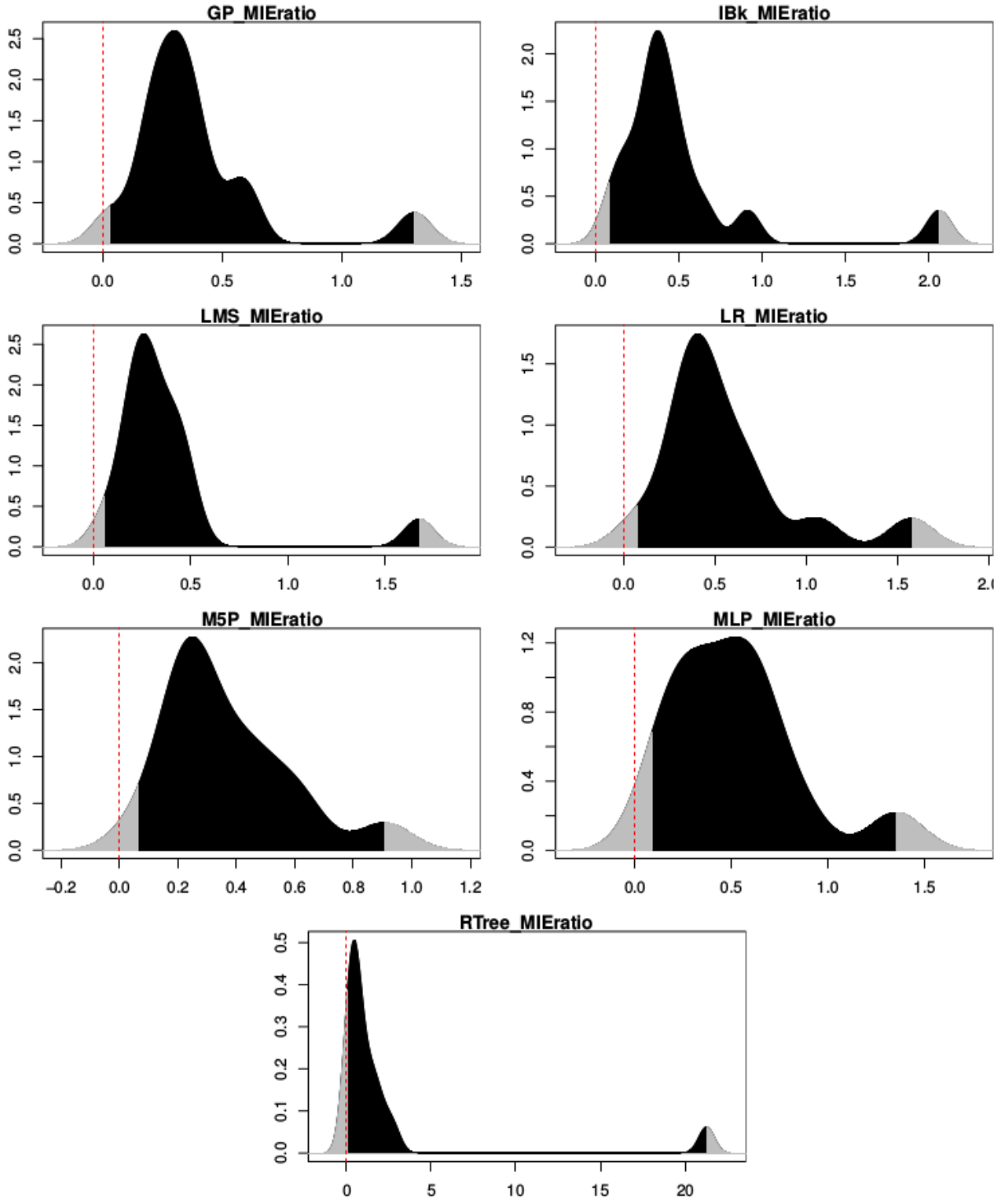


Figure 7: High Posterior Density intervals of the MIEratios of the Methods

	Qtle. 2.5%-97.5%	HPD low-upper	M-Hast. 2.5%-97.5%
GP	0.082-1.057	0.029-1.304	0.279-0.787
IBk	0.114-1.658	0.084-2.06	0.354-0.891
LMS	0.099-1.261	0.057-1.673	0.267-0.645
LR	0.147-1.409	0.075-1.577	0.409-1.005
M5P	0.112-0.806	0.066-0.908	0.275-0.585
MLP	0.12-1.207	0.09-1.356	0.367-0.971
RTree	0.16-15.242	0.099-21.263	0.895-7.867

Table 3: This table shows different probabilistic intervals for each one of the 7 methods ($\alpha = 0.05$) for the data of the MIEratios. Scale is 0-∞. Lower values are better.

	Qtle. 2.5%-97.5%	HPD low-upper	M-Hast. 2.5%-97.5%
GP	0.108-0.749	0.1-0.764	0.316-0.691
IBk	0.118-0.81	0.114-0.85	0.362-0.781
LMS	0.113-0.691	0.106-0.724	0.27-0.559
LR	0.23-3.123	0.204-3.808	0.582-1.586
M5P	0.128-0.778	0.125-0.82	0.3-0.595
MLP	0.182-1.647	0.145-1.99	0.461-0.978
RTree	0.334-1.962	0.329-2.022	0.775-1.837

Table 4: This table shows different probabilistic intervals for each one of the 7 methods ($\alpha = 0.05$) for the means of the MIEratios in 10 runs. Scale is 0-∞. Lower values are better.

The use of Bayesian concepts and Bayesian Networks in Software Testing

Article in progress with several collaborators

- Conference in Bergamo 2015
  - Bayesian Concepts in Software Testing: An Initial Review

  - Now, we focus on Bayesian Networks.
  - Still, doing the review