



Cost estimation result consistency: Implications for SBSE

Marc Roper
Sukumar Letchmunan
Murray Wood

Dept. Computer and Information Sciences
University of Strathclyde



Cost Estimation

Given a project with various parameters:

$$P(X_1, X_2, \dots, X_n) \rightarrow \text{£}$$

Basic approaches:

- Algorithmic (COCOMO etc.)
- Historical data (Expert Judgement, Statistics and Machine Learning)
 - Large number of approaches – which one to use?
- Interested in the domain of web applications



Systematic Literature Review Results

Study	Size Measures	Prediction Measure	Pred Tech	Best Techniques
1	Web Objects, Function Points	MRE, Pred(25), Boxplot residuals	OLS, Allate Systems	OLS- Web Objects
5	Length Measures, Functional Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	LR, RT, SR, ABE, RT&LR, RT&ABE	LM – RT&ABE, FM - SR
16	Length, complexity, functionality	Boxplot residuals	LR, SR	No single Technique
36a	Web Objects, Tukutuku Measures, Length Measures, Functional Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	SR, CBR	LM- SR , TM- CBR
36b	Tukutuku Measures	MMRE, MdMRE, Pred(25)	SR,CBR, CART	None of them superior
37	Tukutuku Measures	MMRE, Pred(25)	SR,CBR	SR & CBR -Single Co.
38a	Tukutuku Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	SR,BN	BN
38b	Tukutuku Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	SR, CBR, BN	SR
38c	Tukutuku Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	SR, CBR, BN	SR
41	Tukutuku Measures	MMRE, MdMRE, Pred(25)	SR, CBR	Single company datasets
42	Tukutuku Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	SVR, SR, CBR, BN	SVR
42a	Tukutuku Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	SVR, SR, CBR, BN	SVR with LinLog
42b	Tukutuku Measures	MMRE, MdMRE, Pred(25), Boxplot residuals	SVR, SR, CBR	SVR



What about data set characteristics?

- Importance stressed many years ago (Shepperd and Kadoda 2001)
- Suggestion that different techniques perform better on certain types of data. e.g.
 - “Messy” data (non-linear, discontinuous, outliers etc.) -> CBR
 - “Non-messy” – Stepwise regression
- Hard to extract from publications so explored further using a number of datasets generated from subsets of the Desharnais dataset
 - publicly available dataset in Promise repository



Characteristics of data subsets

NORMAL	Normal-15
	Normal-50
NORMAL + HIGH POSTIVE KURTOSIS	Normal-15HPK
	Normal-50HPK
NORMAL + HIGH NEGATIVE KURTOSIS	Normal-15HNK
	Normal-50HNK
NORMAL + OUTLIERS	Normal-15Out2
	Normal-50Out4
SKEWED	Skewed-15
	Skewed-50
SKEWED + OUTLIERS	Skewed-15Out2
	Skewed-50Out
POSTIVE SKEWED	Skewed-15PS
	Skewed-50PS



Techniques and Accuracy Measures

- Prediction Techniques
 - Linear Regression
 - RBF Network
 - SVR
 - SVR-Poly
 - RepTrees
 - CBR
- Prediction Accuracy Measures
 - MAE
 - MMRE
 - Pred(25)



Results

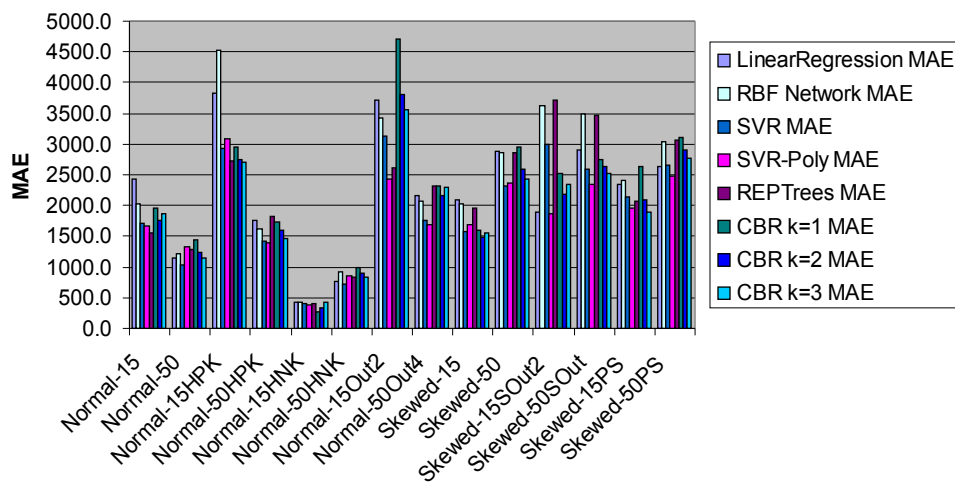
	LinearRegression			RBF Network			SVR			SVR-Poly			REPTrees			k=1			CBR k=2			k=3		
	MAE	MMRE	Pred	MAE	MMRE	Pred	MAE	MMRE	Pred	MAE	MMRE	Pred	MAE	MMRE	Pred	MAE	MMRE	Pred	MAE	MMRE	Pred	MAE	MMRE	Pred
Normal-15	2439.4	139.43	0.27	2030.4	116.17	0.40	1716.3	98.11	0.40	1665.3	95.19	0.20	1543.2	88.21	0.47	1953.9	54.50	0.33	1764.4	59.10	0.40	1872.0	61.40	0.33
Normal-50	1149.6	88.60	0.46	1211.8	93.39	0.52	1027.1	79.17	0.50	1319.1	101.67	0.36	1289.4	99.38	0.46	1445.2	49.90	0.36	1238.9	44.40	0.44	1152.5	43.80	0.48
Normal-15HPK	3826.0	133.24	0.13	4525.0	157.58	0.27	2924.8	101.85	0.33	3075.1	107.09	0.20	2717.9	94.65	0.40	2940.4	56.80	0.40	2741.6	56.90	0.47	2710.3	61.80	0.40
Normal-50HPK	1754.2	104.61	0.40	1623.5	96.81	0.44	1412.4	84.23	0.52	1388.7	82.81	0.50	1814.2	108.19	0.44	1740.3	45.10	0.38	1597.8	45.90	0.46	1457.4	43.30	0.50
Normal-15HNK	437.0	106.40	0.20	428.9	104.42	0.27	400.1	97.42	0.20	390.6	95.09	0.40	410.7	100.00	0.20	279.8	28.90	0.60	336.9	36.00	0.47	438.3	46.50	0.33
Normal-50HNK	762.4	82.29	0.52	930.7	100.47	0.40	715.3	77.22	0.54	851.8	91.95	0.40	840.5	90.72	0.42	1001.3	58.00	0.42	899.8	58.00	0.40	841.6	57.00	0.48
Normal-15Out2	3715.2	128.16	0.33	3414.9	117.79	0.33	3122.2	107.70	0.33	2433.9	83.96	0.60	2610.7	90.05	0.33	4703.1	113.70	0.27	3803.8	88.90	0.33	3559.2	83.60	0.27
Normal-50Out4	2170.8	101.92	0.38	2071.0	97.23	0.44	1759.3	82.60	0.34	1686.6	79.19	0.38	2326.3	109.23	0.26	2329.8	59.70	0.32	2155.6	64.90	0.30	2286.7	70.80	0.30
Skewed-15	2105.3	119.16	0.27	2036.6	115.28	0.20	1569.4	88.83	0.33	1698.4	96.13	0.40	1968.2	111.40	0.33	1605.8	48.90	0.40	1497.5	53.40	0.47	1548.4	52.60	0.33
Skewed-50	2883.8	86.25	0.28	2863.7	85.64	0.32	2315.3	69.25	0.28	2374.4	82.03	0.34	2865.8	85.71	0.32	2939.4	84.90	0.32	2581.4	66.20	0.22	2431.5	61.60	0.32
Skewed-15Out2	1902.7	48.59	0.33	3615.4	92.34	0.27	2999.5	76.61	0.27	1874.1	47.86	0.40	3716.7	94.92	0.13	2527.0	66.80	0.33	2185.9	57.90	0.33	2339.1	53.30	0.27
Skewed-50Out	2905.3	77.44	0.22	3489.3	93.00	0.30	2592.5	69.10	0.32	2348.4	62.59	0.38	3472.8	92.57	0.30	2754.6	66.60	0.28	2636.5	62.80	0.28	2532.1	65.90	0.28
Skewed-15PS	2348.7	98.02	0.33	2413.0	100.71	0.13	2132.7	89.00	0.13	1966.4	82.06	0.20	2077.9	86.72	0.20	2635.1	104.40	0.20	2105.1	89.00	0.20	1887.2	80.30	0.27
Skewed-50PS	2646.4	80.18	0.34	3030.5	92.54	0.22	2649.6	80.90	0.26	2483.3	75.83	0.42	3072.3	93.82	0.28	3115.3	63.40	0.38	2902.8	67.70	0.22	2781.0	67.70	0.26



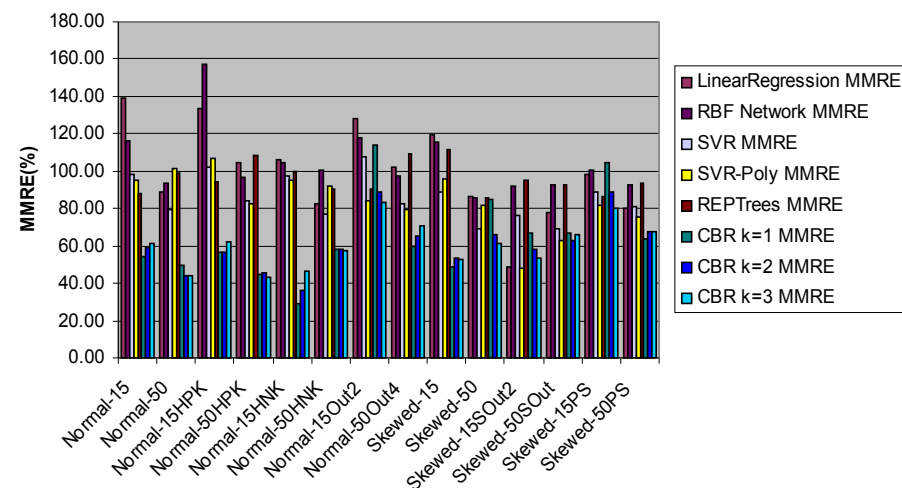
Issues for SBSE #1

- Apparent interaction between dataset, traditional accuracy measures, and prediction technique.

MAE vs Group of Dataset



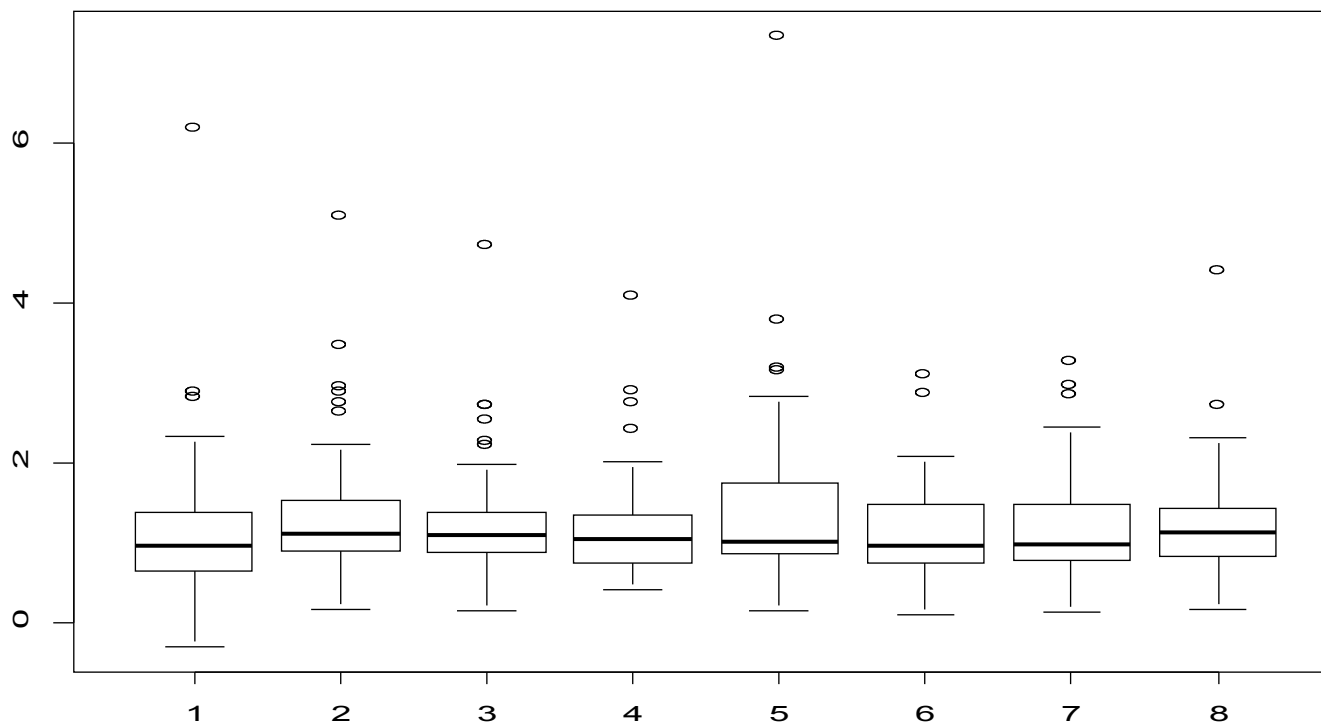
MMRE vs Group of Dataset





Preferable Accuracy Measures

- Boxplots of Z and of residuals





Issues for SBSE #2

- Boxplots can be compared and ranked
 - Consider median, box length, tail length, outlier values etc...
- Hard to aggregate into a single value
 - => for the design of objective functions



		MAE	MMRE	Boxplot Of z	Boxplot Of Residuals
NORMAL	Normal-15	REPTrees	CBR1 CBR2	RepTrees	RepTrees
	Normal-50	SVR	CBR3 CBR2	SVR	SVR
+ HIGH POSTIVE KURTOSIS	Normal-15HPK	CBR3, REPTrees, CBR2	CBR1 CBR2, CBR3	CBR3	RepTrees
	Normal-50HPK	SVRP, SVR	CBR3 CBR1, CBR2	SVRP	SVRP
NORMAL + HIGH NEGATIVE KURTOSIS	Normal-15HNK	CBR1	CBR1	CBR1	CBR1
	Normal-50HNK	SVR, LR	CBR3 CBR2, CBR1	CBR3	SVR
+ OUTLIERS	Normal-15Out2	SVRP	CBR3 CBR2	SVRP	SVRP
	Normal-50Out4	SVRP SVR	CBR1 CBR2	SVRP	SVRP
SKEWED	Skewed-15	CBR2 CBR3, SVR	CBR1 CBR3 CBR2	CBR3	CBR2
	Skewed-50	SVR SVRP	CBR3 CBR2	CBR3	CBR3
SKEWED + OUTLIERS	Skewed-15Out2	SVRP LR	SVRP LR	SVR	SVRP
	Skewed-50Out	SVRP	SVRP CBR2	SVRP	SVRP
POSTIVE SKEWED	Skewed-15PS	CBR3 SVRP	CBR3 SVRP	CBR3	SVRP
	Skewed-50PS	SVRP	CBR1 CBR1, CBR3	SVR	SVRP

Results
using
boxplot
rankings
–
still lacking
conclusion
stability



A Refined Set of Rules

Big/Small Group	B	Big Group
	S	Small Group
Skewness	HS	High Skew >3
	LS	Low Skew >2 but <3
	AS	Acceptable Skew value <2
Kurtosis	HK	High Kurtosis >3
	LK	Low kurtosis >2 but <3
	AK	Acceptable Kurtosis <2
Outlier proportion	HO	High outlier proportion > 0.10
	LO	Low outlier proportion <0.10
Outlier average < or > Median	OAM	Outlier average greater than Median
	MOA	Outlier average lower than Median



Results on New ISBSG Subset

Group	Characteristics Code	Suggestion Prediction	MAE	Boxplot Of Z	Boxplot Of Residuals
G1-15	SLSHKHOAM	CBR	SVRP, CBR2	CBR2	CBR2
G1-30	BASAKHOMOA	SVRP	SVRP	SVRP	SVRP
G2-15	SASAKHOAM	SVRP	SVRP	SVRP	SVRP
G2-30	BASLKHOOAM	SVRP	RBFN	CBR2	RBFN
G3-15	SASAKLOMOA	CBR	SVRP, CBR2	CBR2	SVRP
G3-30	BASAKLOMOA	RBFN	RepTrees, SVRP	RepTrees	RepTrees
G4-15	SLSHKLOOAM	SVRP	SVRP	SVRP	SVRP
G4-30	BASAKHOAM	SVRP	SVRP	SVRP	SVRP
G5-15	SLSHKLOOAM	SVRP	SVRP	SVRP	SVRP
G5-30	BLSHKLOOAM	SVRP	SVRP	SVRP	SVRP



(Obvious) Issues for SBSE #3

- Rules do not necessarily translate between datasets
- Or even within datasets
 - (Not single company)



A Simpler Set of Rules?

Classifier Model	ISBSG	Desharnais	NewSubsetISBSG
BFTree	Outliers < 4.5: SVRP Outliers >= 4.5: CBR3	SVRP	SVRP
DecisionStump	Outliers <= 4.5 : SVRP Outliers > 4.5 : CBR3	Kurtosis <= 4.14 : SVRP Kurtosis > 4.14 : SVR	Skew <= 1.175 : CBR2 Skew > 1.175 : SVRP
J48	Outliers <= 4: SVRP Outliers > 4: CBR3	Kurtosis <= 3.946 Outliers <= 0: SVRP Outliers > 0 Skew <= 1.13: CBR Skew > 1.13: SVRP Kurtosis > 3.946: SVR	Outliers <= 0: CBR2 Outliers > 0: SVRP
REPTree	Outliers < 4.5 : SVRP Outliers >= 4.5 : CBR3	SVRP	SVRP
SimpleCart	Outliers < 4.5: SVRP Outliers >= 4.5: CBR3	SVRP	SVRP



Final Challenges for SBSE

Wider exploration of dataset characteristics

Wider exploration of algorithm parameters