DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Rank-Biased Precision Reloaded: Reproducibility and Generalization

Nicola Ferro and Gianmaria Silvello

Information Management Systems Research Group
Department of Information Engineering
University of Padua

{ferro,silvello}@dei.unipd.it

37th European Conference on Information Retrieval - ECIR 2015
Vienna, Austria, 29 March - 02 April 2015

Outline

- A view on Reproducibility (and Generalization)
- Rank-Biased Precision
- Repeatability
- Reproducibility and Generalization
- Wrapping Up

A view on Reproducibility (and Generalization)

Reproducibility

- No research paper can ever be considered to be the final word, and the *replication* and corroboration of research results is key to the scientific process

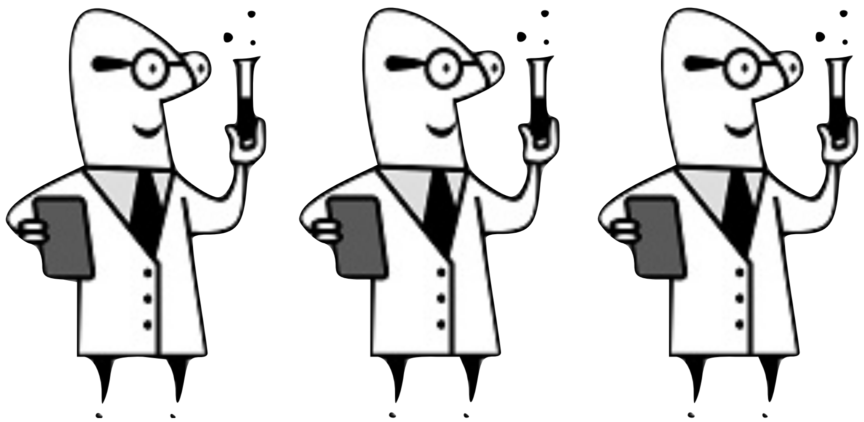
[Nature, `http://www.nature.com/nature/focus/reproducibility/`]

- The basic principle is that, given an experiment, an <u>independent</u> researcher should be able to *replicate* it, under the same conditions, and achieve the same results

[`http://explorable.com/reproducibility`]

~~Repeatability~~ Reproducibility

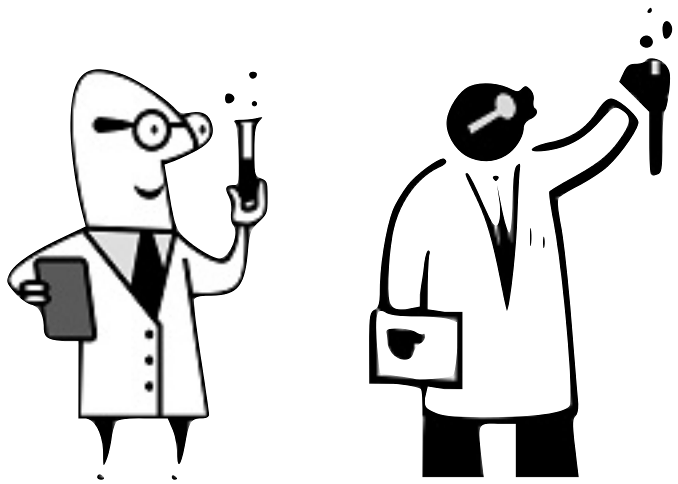
- *Repeatability*: researchers repeat the experiments to test and verify the results



~~Repeatability~~ Reproducibility

- *Repeatability*: researchers repeat the experiments to test and verify the results

- *Reproducibility:*
  - completely independent from the original study
  - generate "identical" findings
  - leads to *Generalization* whose aim is to apply the experimental findings to new situations in order to determine their validity in a different context with different variables



RBP:

Rank-Biased Precision

RBP

- The original paper: A. Moffat and J. Zobel, Rank-Biased Precision for Measurement of Retrieval Effectiveness, *Transactions On Information Systems*, 27(1): 1-27, 2008.

- Impact:

  - > 80 citations in the ACM DL

  - > 190 citations in Google Scholar

  - > 100 citations in Scopus

Rank-Biased Precision for Measurement of Retrieval Effectiveness

2

ALISTAIR MOFFAT
The University of Melbourne
and
JUSTIN ZOBEL
RMIT University and NICTA Victoria Research Laboratory

A range of methods for measuring the effectiveness of information retrieval systems has been proposed. These are typically intended to provide a quantitative single-value summary of a document ranking relative to a query. However, many of these measures have failings. For example, recall is not well founded as a measure of satisfaction, since the user of an actual system cannot judge recall. Average precision is derived from recall, and suffers from the same problem. In addition, average precision lacks key stability properties that are needed for robust experiments. In this article, we introduce a new effectiveness metric, *rank-biased precision*, that avoids these problems. Rank-biased precision is derived from a simple model of user behavior, is robust if answer rankings are extended to greater depths, and allows accurate quantification of experimental uncertainty, even when only partial relevance judgments are available.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models, search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms: Experimentation, Measurement, Human Factors

Additional Key Words and Phrases: Recall, precision, average precision, relevance, pooling

ACM Reference Format:

Moffat, A. and Zobel, J. 2008. Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inform. Syst. 27, 1, Article 2 (December 2008), 27 pages. DOI = 10.1145/1416950.1416952 http://doi.acm.org/10.1145/1416950.1416952

This work was supported by the Australian Research Council.
Authors' addresses: A. Moffat, Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia; email: alistair@csse.unimelb.edu.au; J. Zobel, Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia; email: jz@csse.unimelb.edu.au.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2008 ACM 1046-8188/2008/12-ART2 $5.00 DOI 10.1145/1416950.1416952 http://doi.acm.org/10.1145/1416950.1416952

ACM Transactions on Information Systems, Vol. 27, No. 1, Article 2, Publication date: December 2008.

Why a TOIS paper? / Why RBP?

- Our goal is to start to understand what reproducibility means for IR evaluation.

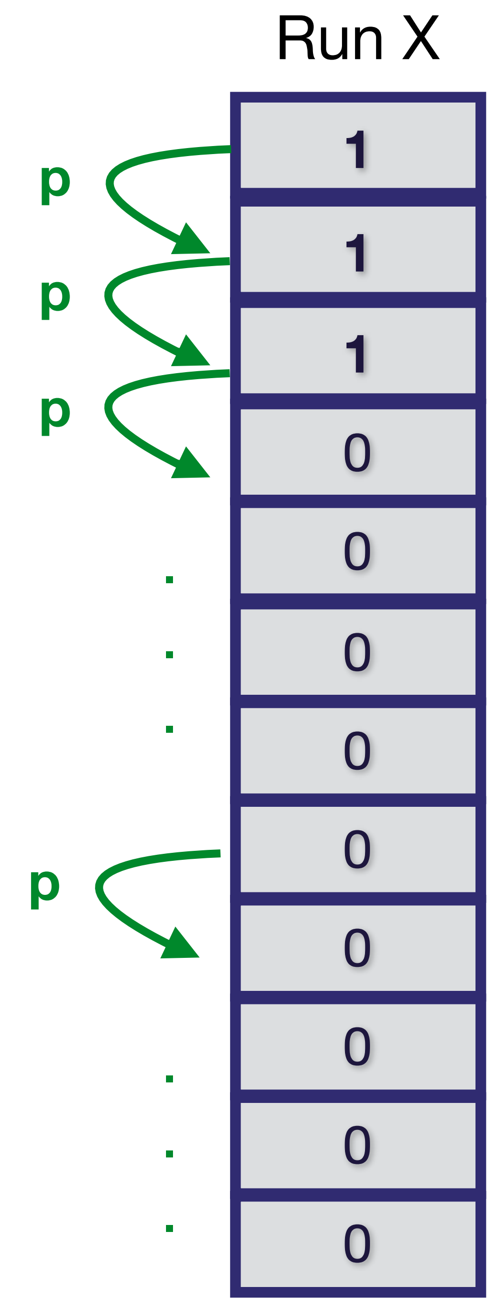
- Therefore, we need to be able to reduce the confounding factors (e.g. poor experimental design) to focus on issues raised only by reproducibility

http://ak.picdn.net/shutterstock/

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

RBP

- User model: a user always starts from the first document in a ranking and then s/he progresses with probability *p (persistence parameter)*



$$\text{RBP} = (1-p)\sum_{i=0}^{d} r_i \cdot p^{i-1}$$

$$0 \leq p \leq 1 \qquad p = \begin{cases} 0.5 & \text{fairly persistent user} \\ 0.8 & \text{persistent user} \\ 0.95 & \text{very persistent user} \end{cases}$$

p = 0.5 $\quad RBP(X) = 0.5(0.5^0 + 0.5^1 + 0.5^2) = 0.88$

p = 0.8 $\quad RBP(X) = 0.2(0.8^0 + 0.8^1 + 0.8^2) = 0.49$

p = 0.95 $\quad RBP(X) = 0.05(0.95^0 + 0.95^1 + 0.95^2) = 0.14$

Repeatability

What do we need to reproduce?

- Experiments are based on the TREC-05, 1996, Ad-Hoc collection
    - 61 runs, 50 topics, binary relevance, ~530K docs
    - released by the *National Institute of Standard and Technology* (NIST): http://trec.nist.gov/

What do we need to reproduce?

- Three main experiments have been conducted to explore how RBP behaves:
  - Kendall's tau correlation with shallow pools (depth 100 and 10)
  - Upper and lower bounds for RBP varying the p parameter (0.5, 0.8 and 0.95)
  - Discriminative power: t test and Wilcoxon test

1^{st} set of experiments to be reproduced

1. Kendall's correlation coefficients from the systems ordering generated by pair of metrics and by considering two pool depths (10 and 100)

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

1st set of experiments to be reproduced

1. Kendall's correlation coefficients from the systems ordering generated by pair of metrics and by considering two pool depths (10 and 100)

Pool at depth 10 was calculated by exploiting original assessments but applying them to a reduced set of documents (the union of the first 10 documents of each run).

This downsampling technique is deterministic

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

1st set of experiments to be reproduced

1. Kendall's correlation coefficients from the systems ordering generated by pair of metrics and by considering two pool depths (10 and 100)

Rank-Biased Precision for Measurement of Retrieval Effectiveness • 2:9

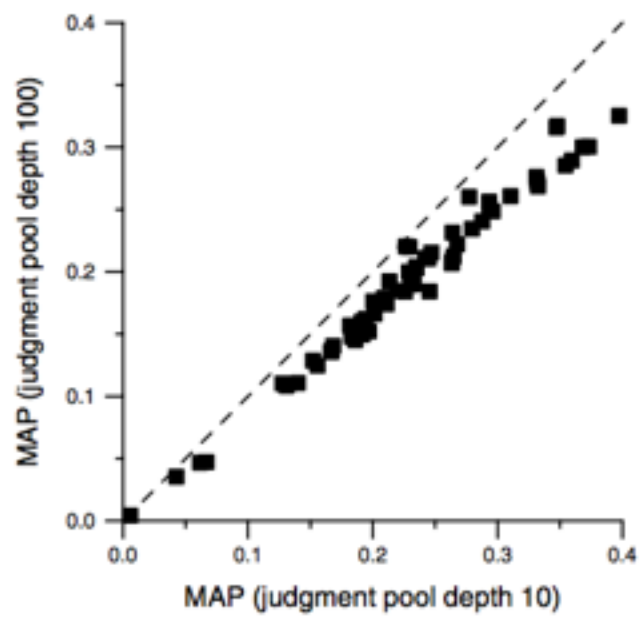


Fig. 2. Mean average precision of 61 TREC-5 systems, using relevance judgments compiled using two different pool depths. The dotted line is the identity relationship, with points below the line showing systems for which average precision decreased when additional documents were judged. The nonlinearity of the decrease shows that the ordering of systems is also affected.

2:16 • A. Moffat and J. Zobel

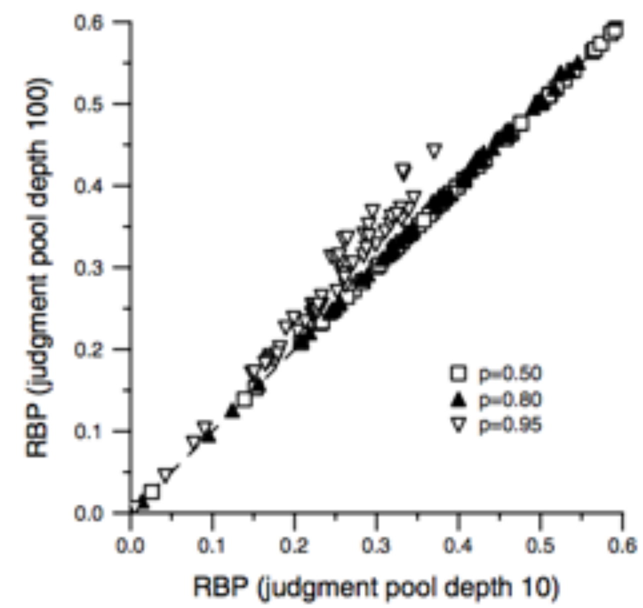


Fig. 4. Rank-biased precision of 61 TREC-5 systems, for three different values of p, using relevance judgments compiled using two different pool depths. Rank-biased precision at $p = 0.5$ and $p = 0.8$ is stable when the pool depth is increased from 10 documents per system to 100 documents. At $p = 0.95$ the RBP scores increase (and never decrease) when the pool depth is increased.

1st set of experiments to be reproduced

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

1. Kendall's correlation coefficients from the systems ordering generated by pair of metrics and by considering two pool depths (10 and 100)

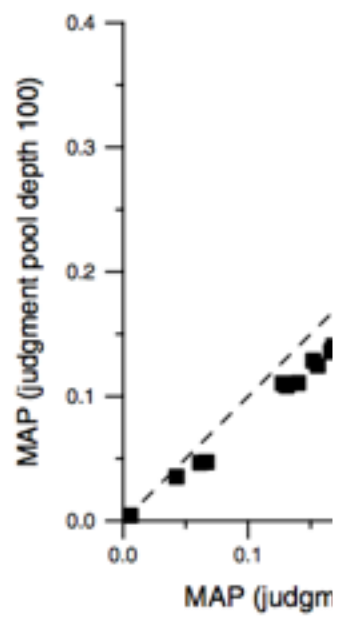
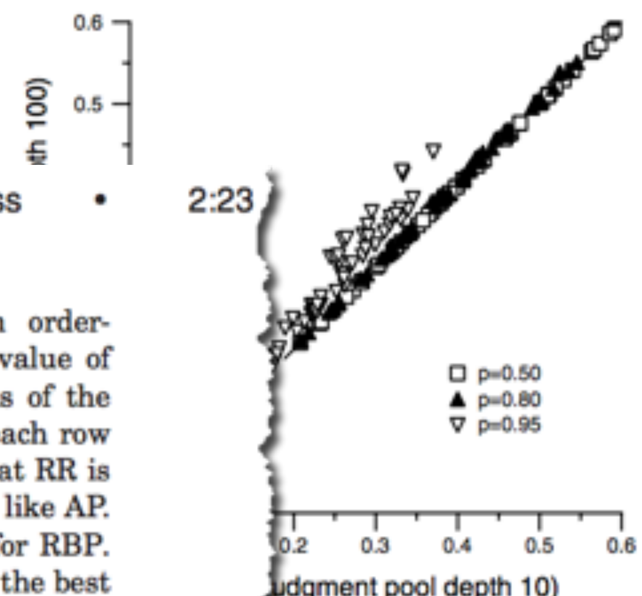


Fig. 2. Mean average precision of 61 TREC-5 sys
two different pool depths. The dotted line is the i
showing systems for which average precision decr
The nonlinearity of the decrease shows that the o

Table III.
Kendall's τ correlation coefficients calculated from the system orderings generated by pairs of metrics using the 61 TREC-5 runs. A value of 1.0 indicates perfect agreement between the two metrics, in terms of the system ordering that they produce. The largest (nonself) value in each row is highlighted in boldface, with the top part of the table showing that RR is most like P@10; that P@10 is most like P@R; and that P@R is most like AP. The bottom group of rows shows the same correlation coefficients for RBP. When $p = 0.5$, RBP behaves most like RR. When RBP uses $p = 0.8$, the best agreement is with P@10. When RBP uses $p = 0.95$, there is good agreement with all of P@10, P@R, and AP.

Metric	Pool depth	Kendall's τ, pool depth 100			
		RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**



ystems, for three different values of p, using relevance
depths. Rank-biased precision at $p = 0.5$ and $p = 0.8$
rom 10 documents per system to 100 documents. At
decrease) when the pool depth is increased.

Before everything: How to import the data?

- TREC-05 is a public experimental collection composed by 61 runs shared using the following well-known format:

  <topic id> <q0> <document id> <rank> <score> <run id>

- The standard library `trec_eval` employed by TREC imports runs as follows (`trec_eval` ordering):
  - items are sorted in descending order by <u>score</u> and descending lexicographical order of <u>document-id</u> when scores are tied

How to import the data?

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

- It is possible to specify different import orders
    - original ordering: the runs are imported as they were submitted to the campaign without performing any additional ordering
- `trec_eval` does not implement RBP and in the paper the importing order of the run is not specified

The import ordering effect

Metric	Pool depth	Kendall's τ, pool depth 100			
		RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**

Kendall's Tau correlations in the RBP original paper

The import ordering effect

Metric	Pool depth	Kendall's τ, pool depth 100 RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**

Kendall's Tau correlations in the RBP original paper

Reproduced results

Numbers in bold are those which are at least 1% different from the correlations in the original RBP paper

treceval ordering

Metric	depth	depth 100 RR	P@10	P@R	AP
RR	10	0.997	0.842	0.748	0.732
P@10	10	0.840	1.000	0.861	0.845
P@R	100	0.746	0.861	1.000	0.908
RBP.5	10	0.926	0.858	0.764	0.755
RBP.8	10	0.888	0.930	**0.819**	0.809
RBP.95	10	0.778	0.882	0.877	0.896
RBP.95	100	0.793	0.916	0.895	**0.859**
nDCG	100	0.765	0.831	0.877	0.915

original ordering

Metric	depth	depth 100 RR	P@10	P@R	AP
RR	10	0.997	0.841	0.747	0.730
P@10	10	0.840	1.000	0.860	0.844
P@R	100	**0.769**	0.861	1.000	0.907
RBP.5	10	0.924	0.858	**0.776**	0.755
RBP.8	10	0.889	0.929	**0.828**	0.809
RBP.95	10	0.779	0.880	**0.905**	0.894
RBP.95	100	0.792	0.913	**0.850**	**0.859**
nDCG	100	0.763	0.829	**0.886**	0.913

The import ordering effect

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Metric	Pool depth	Kendall's τ, pool depth 100 RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**

Kendall's Tau correlations in the RBP original paper

Reproduced results

Numbers in bold are those which are at least 1% different from the correlations in the original RBP paper

Metric	depth	treceval ordering, depth 100 RR	P@10	P@R	AP
RR	10	0.997	0.842	0.748	0.732
P@10	10	0.840	1.000	0.861	0.845
P@R	100	0.746	0.861	1.000	0.908
RBP.5	10	0.926	0.858	0.764	0.755
RBP.8	10	0.888	0.930	**0.819**	0.809
RBP.95	10	0.778	0.882	0.877	0.896
RBP.95	100	0.793	0.916	0.895	**0.859**
nDCG	100	0.765	0.831	0.877	0.915

Metric	depth	original ordering, depth 100 RR	P@10	P@R	AP
RR	10	0.997	0.841	0.747	0.730
P@10	10	0.840	1.000	0.860	0.844
P@R	100	**0.769**	0.861	1.000	0.907
RBP.5	10	0.924	0.858	**0.776**	0.755
RBP.8	10	0.889	0.929	**0.828**	0.809
RBP.95	10	0.779	0.880	**0.905**	0.894
RBP.95	100	0.792	0.913	**0.850**	**0.859**
nDCG	100	0.763	0.829	**0.886**	0.913

The import ordering effect

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Metric	Pool depth	Kendall's τ, pool depth 100 RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**

Kendall's Tau correlations in the RBP original paper

Reproduced results

Numbers in bold are those which are at least 1% different from the correlations in the original RBP paper

Metric	depth	treceval ordering depth 100 RR	P@10	P@R	AP

Metric	depth	original ordering depth 100 RR	P@10	P@R	AP

Lesson learned #1

The import ordering of the runs may influence the experimental results and it <u>should</u> be explicitly specified as in other contexts data manipulation/cleaning are reported

Measure Parameters

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Metric	Pool depth	Kendall's τ, pool depth 100			
		RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**

- The calculation of nDCG is influenced by:
  - the weighting schema
  - the log base of the discounting function
- These parameters are not specified in the paper
  - weighting schema: R = 1, NR = 0
  - log base = 2

Measure Parameters

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Metric	Pool depth	Kendall's τ, pool depth 100			
		RR	P@10	P@R	AP
RR	10	0.997	**0.841**	0.749	0.733
P@10	10	0.839	1.000	**0.861**	0.846
P@R	100	0.748	0.861	1.000	**0.905**
RBP, $p = 0.5$	10	**0.925**	0.858	0.768	0.758
RBP, $p = 0.8$	10	0.887	**0.930**	0.822	0.812
RBP, $p = 0.95$	10	0.778	0.880	0.874	**0.897**
RBP, $p = 0.95$	100	0.791	**0.913**	0.896	0.863
NDCG	100	0.763	0.831	0.878	**0.916**

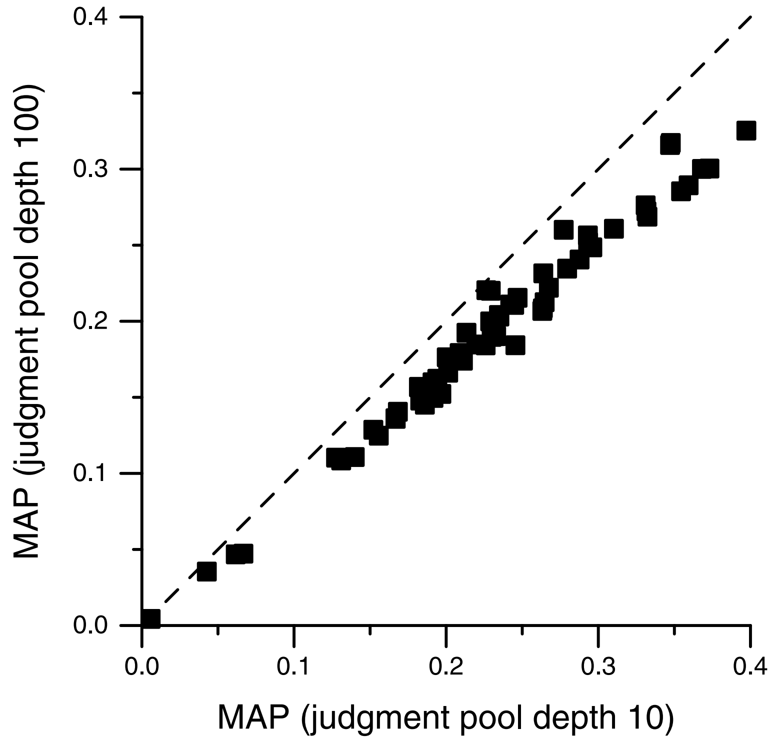
- The calculation of nDCG is influenced by:
  - th
  - th
- The

Lesson learned #2

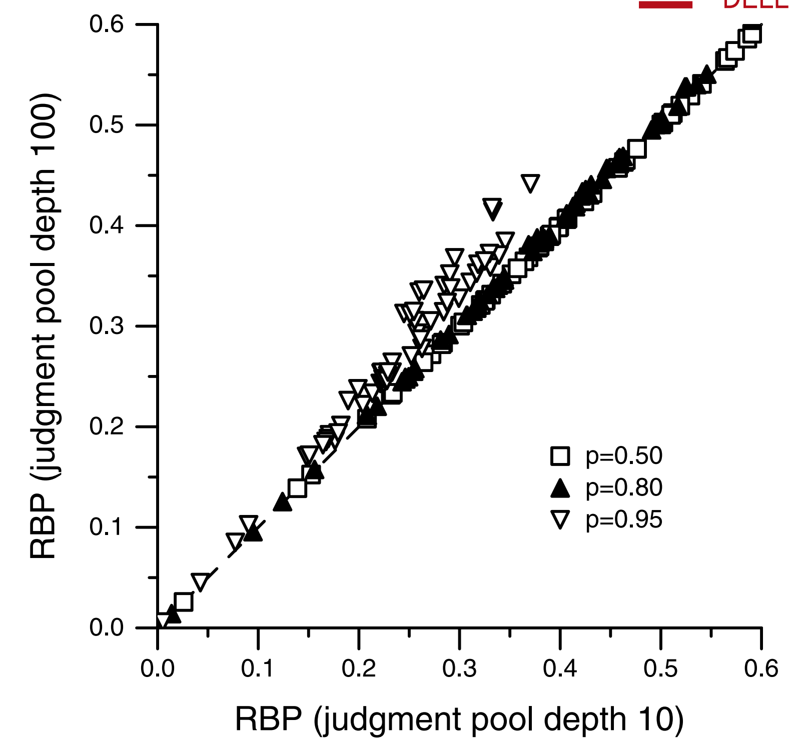
The measure parameters <u>should</u> be specified; e.g. weighting schema for the DCG family and the persistence value for RBP

  - weighting schema: R = 1, NR = 0
  - log base = 2

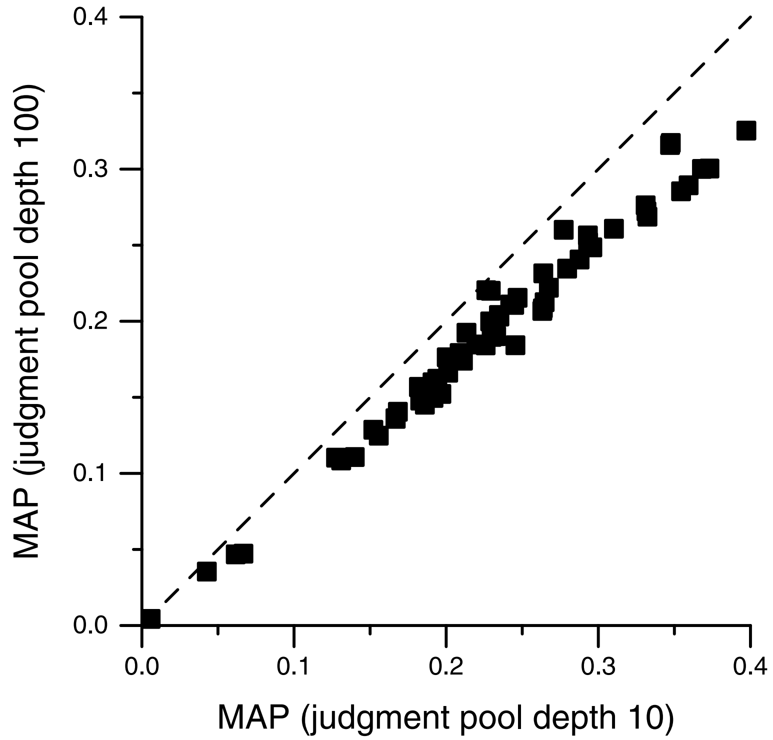
Pool Downsampling



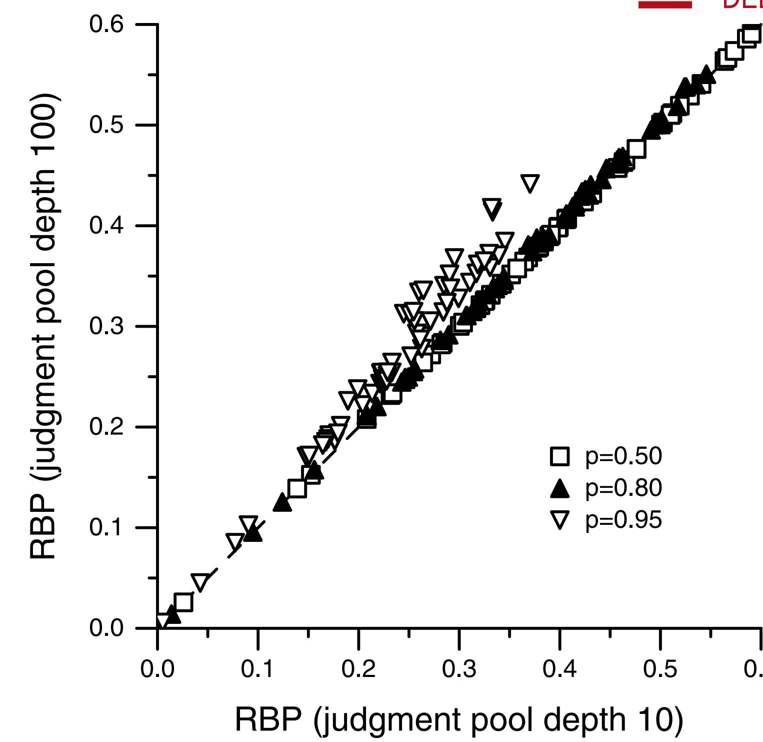
Pool downsampling effect in the RBP original paper



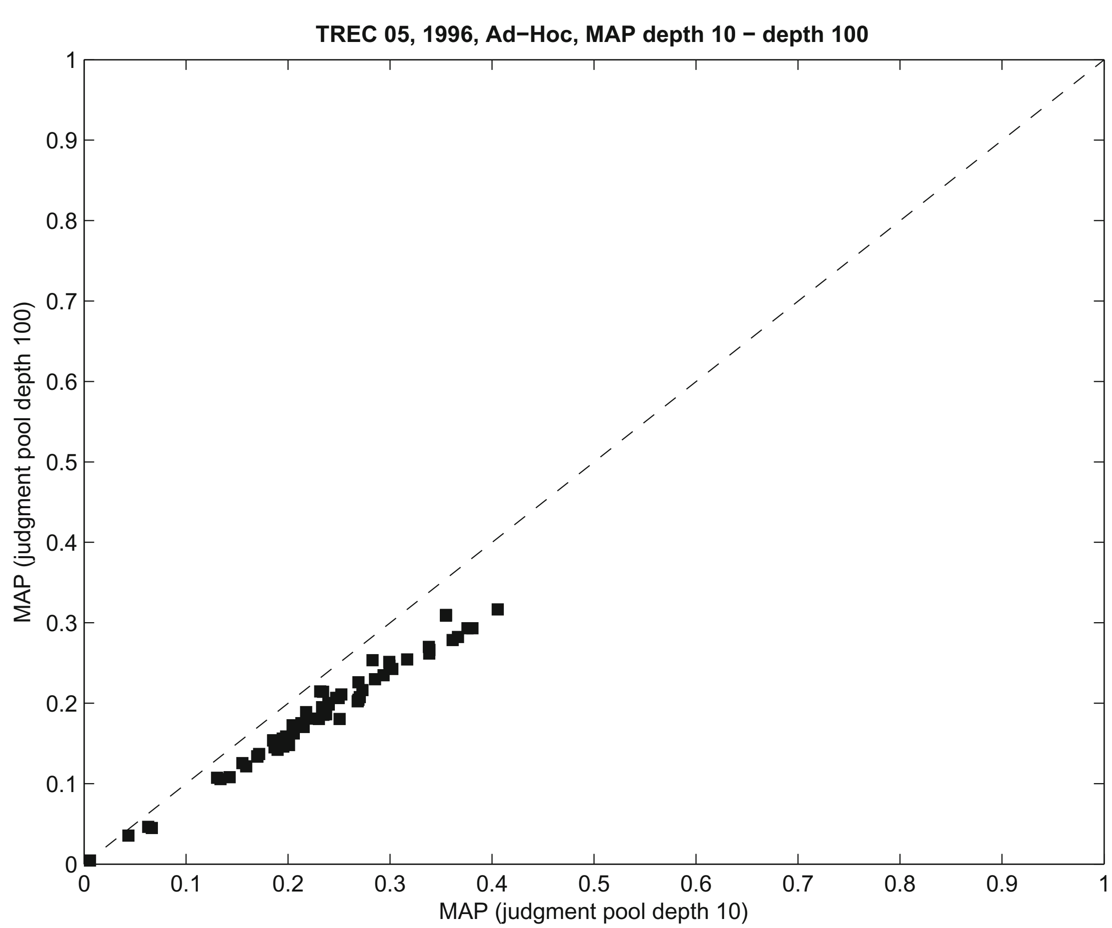
Pool Downsampling

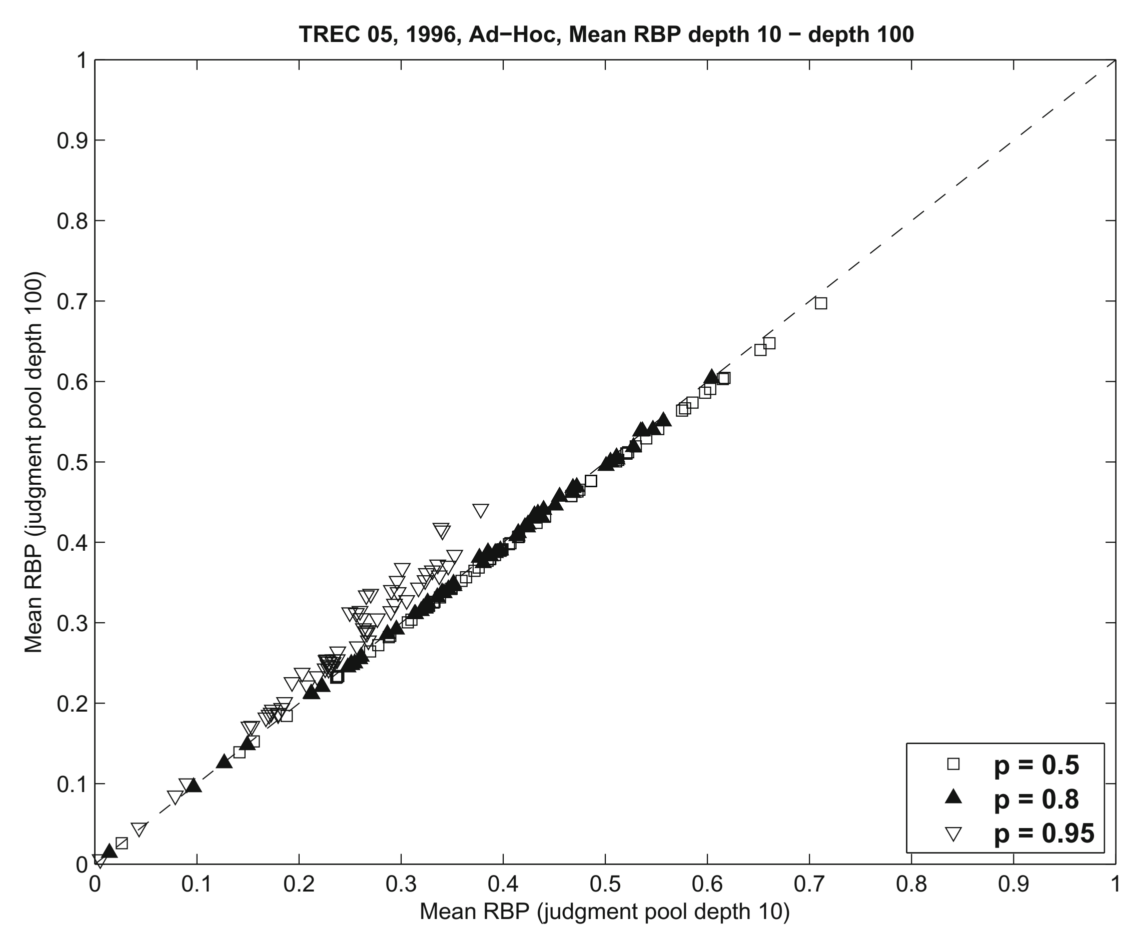


Pool downsampling effect in the RBP original paper



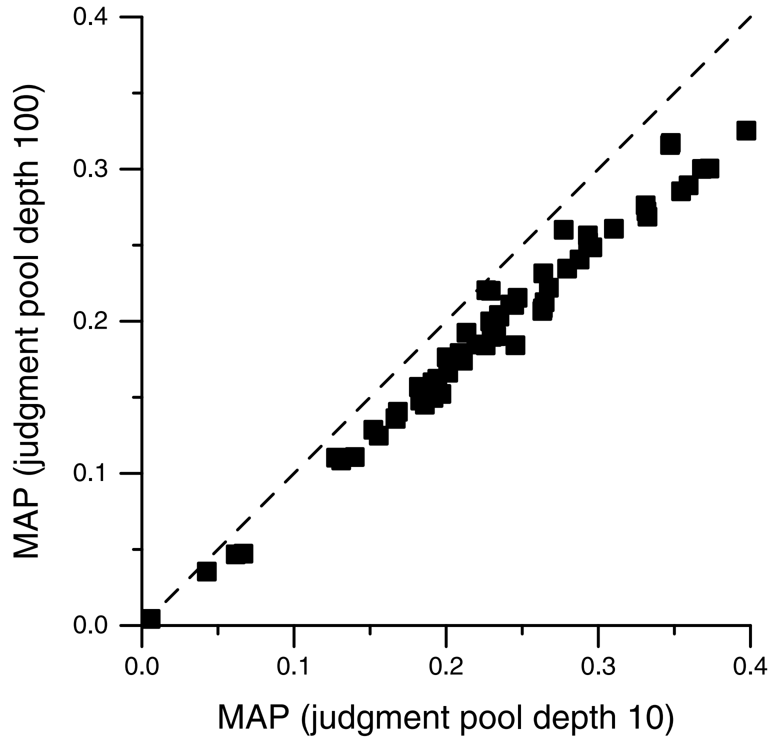
Reproduced Results



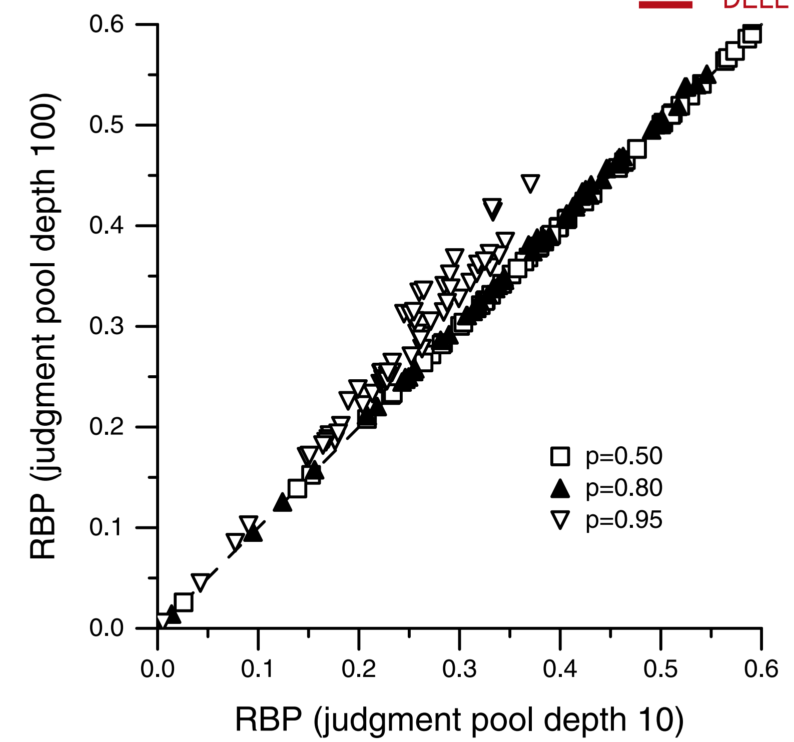


Pool Downsampling

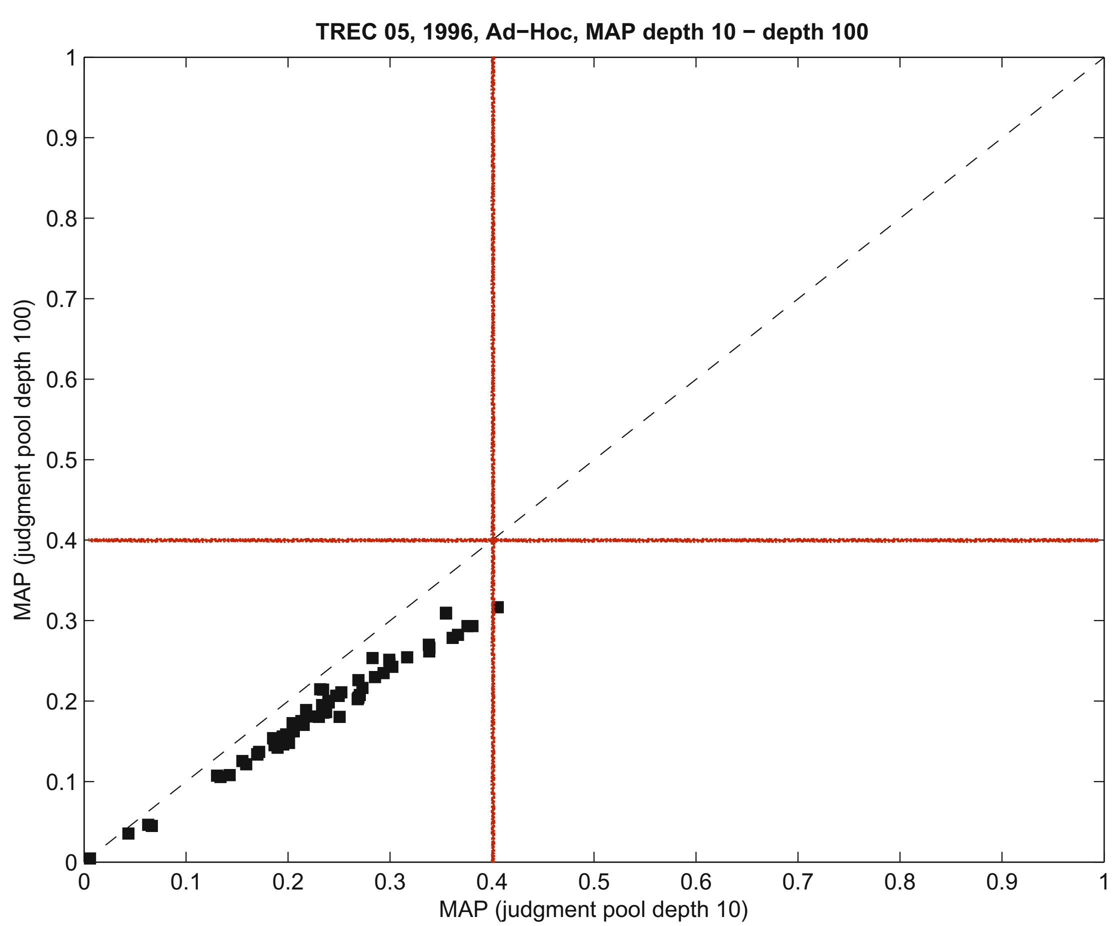
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

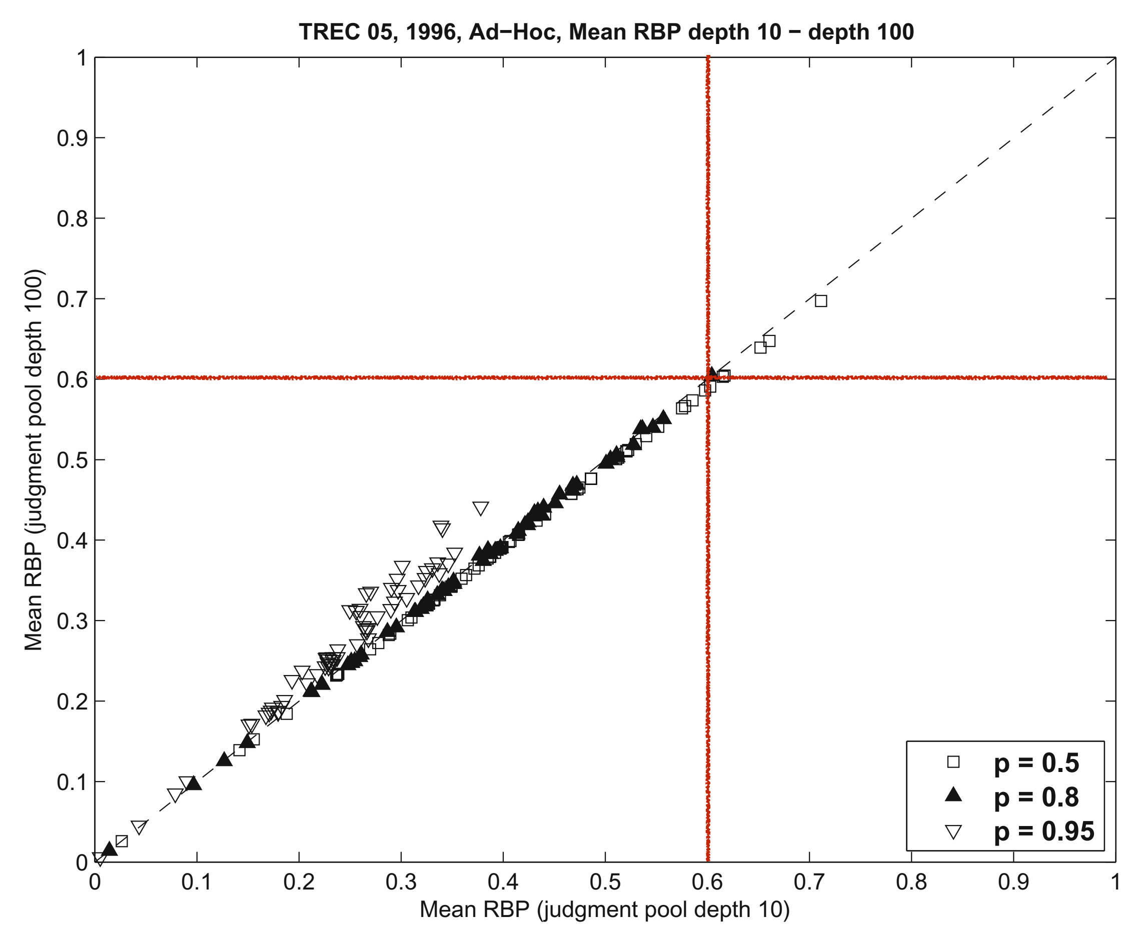


Pool downsampling
effect in the
RBP
original paper

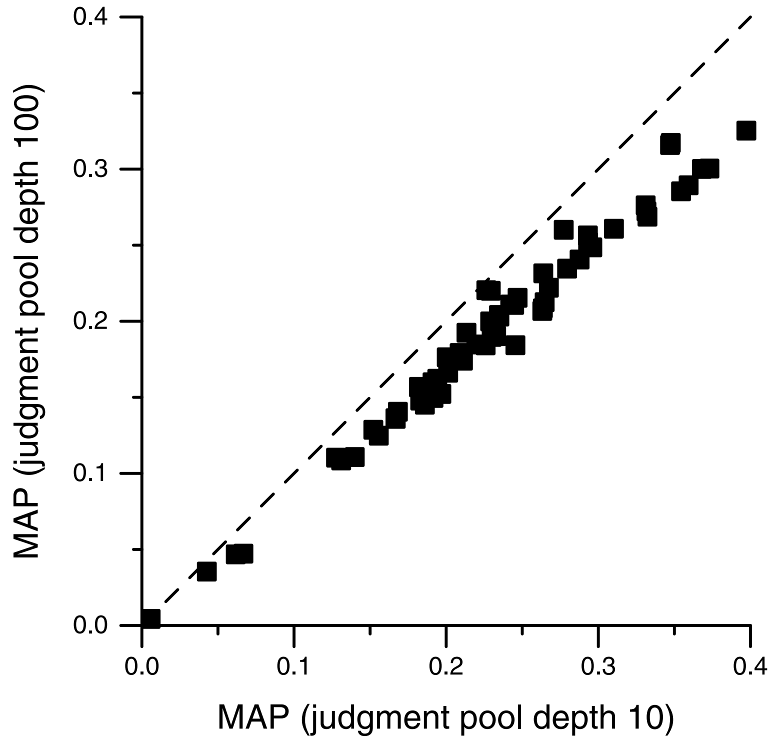


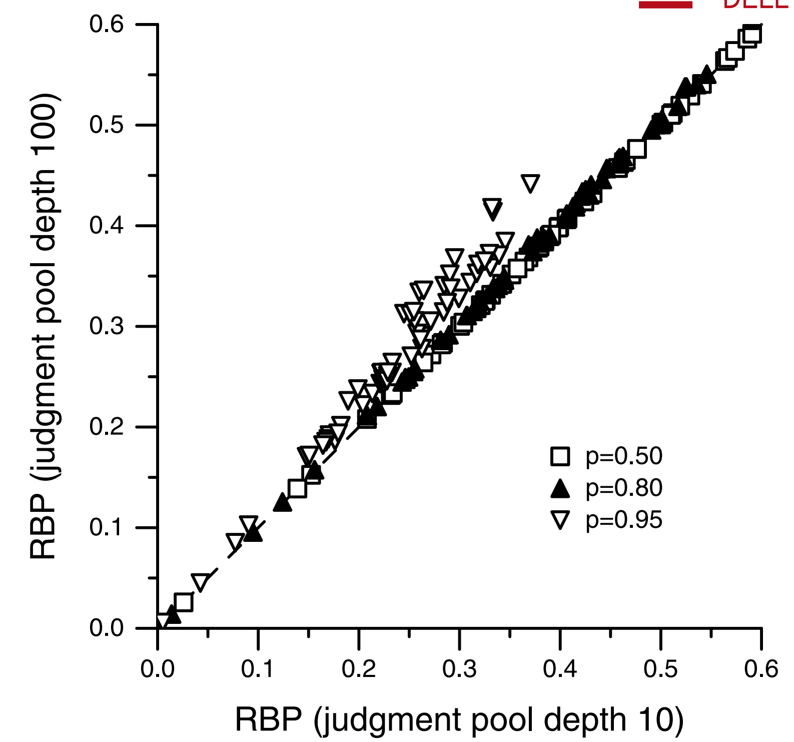
Reproduced Results



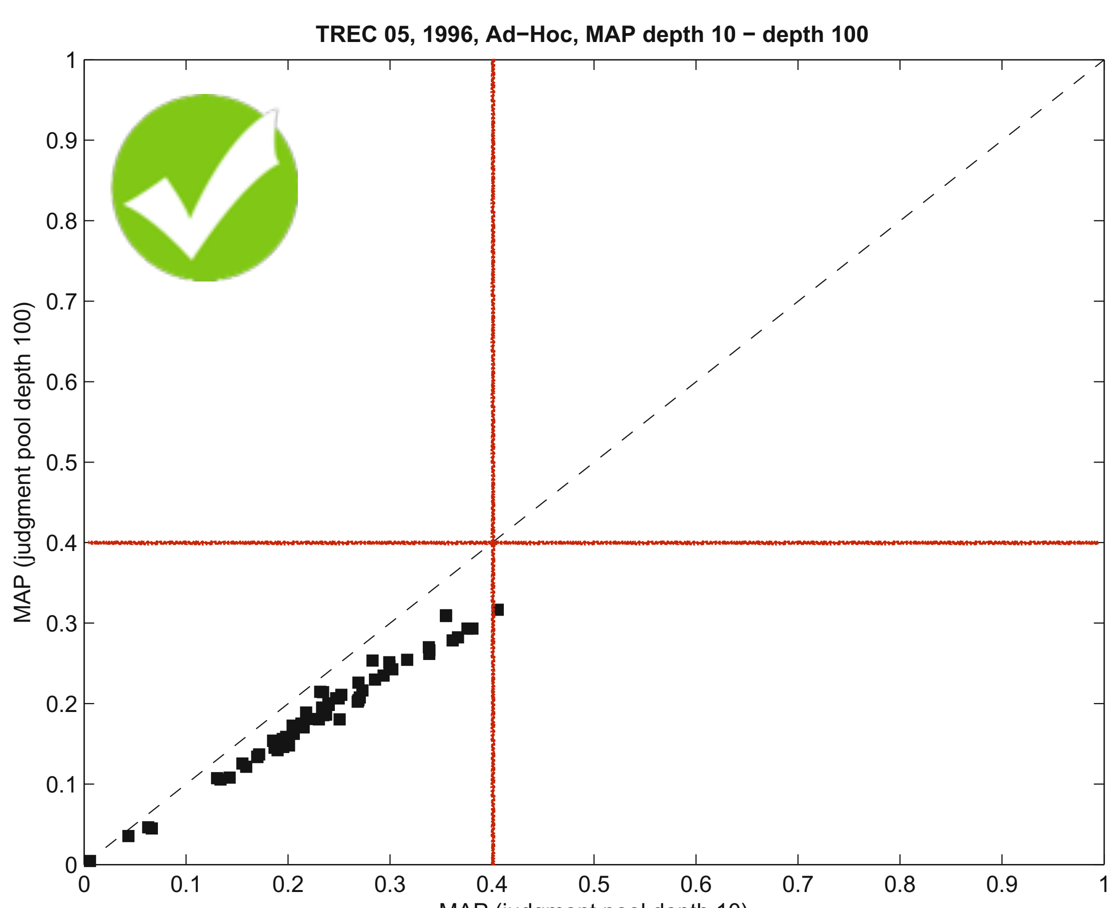


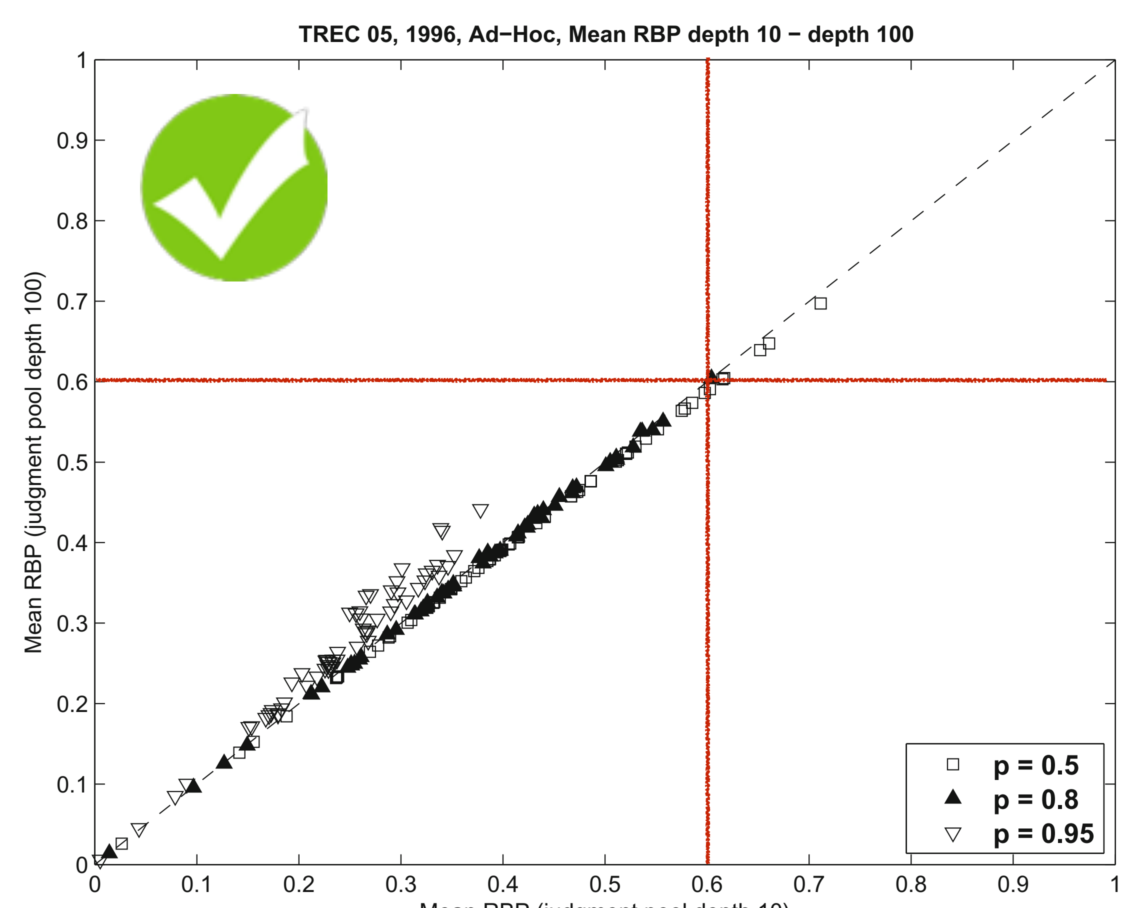
Pool Downsampling



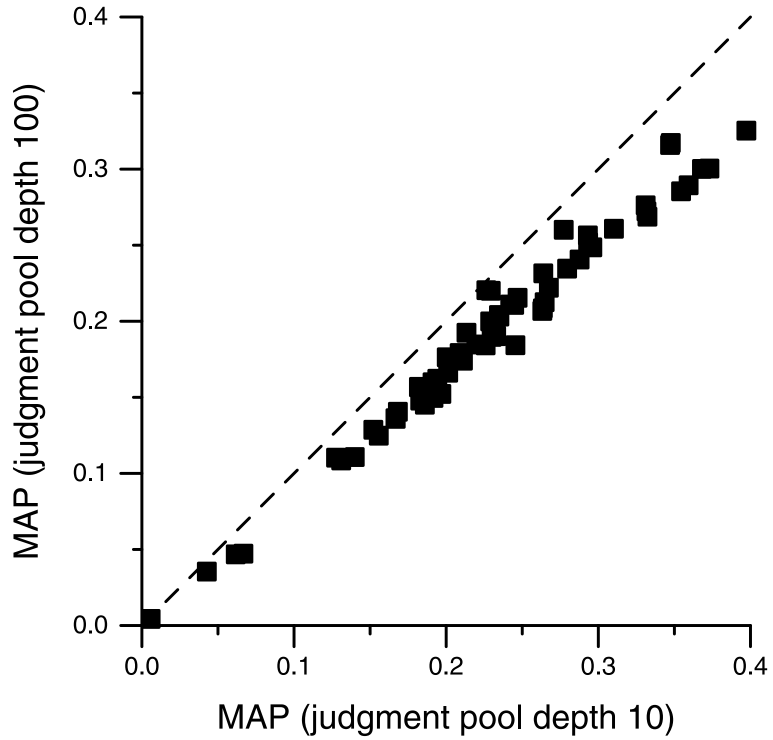


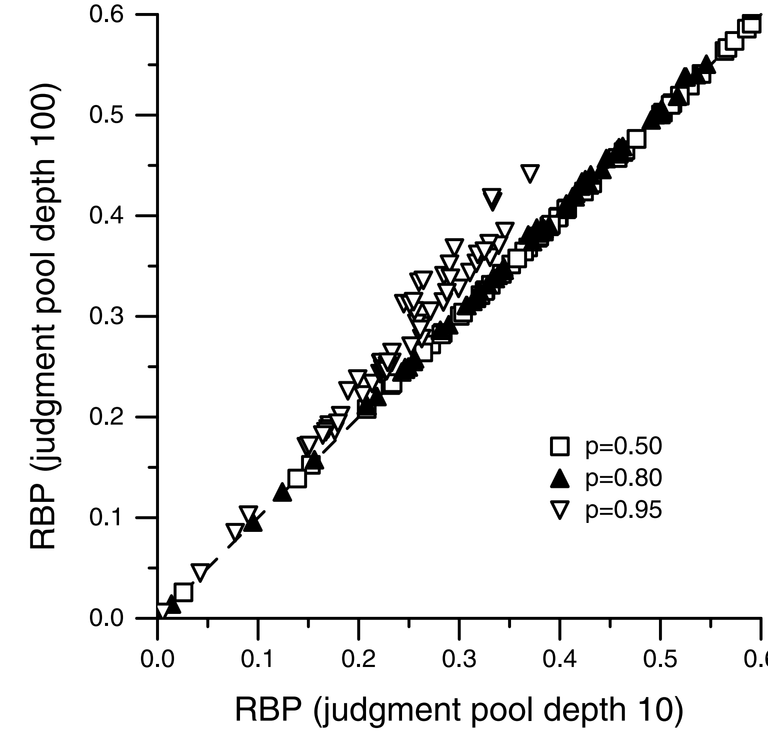
Reproduced Results



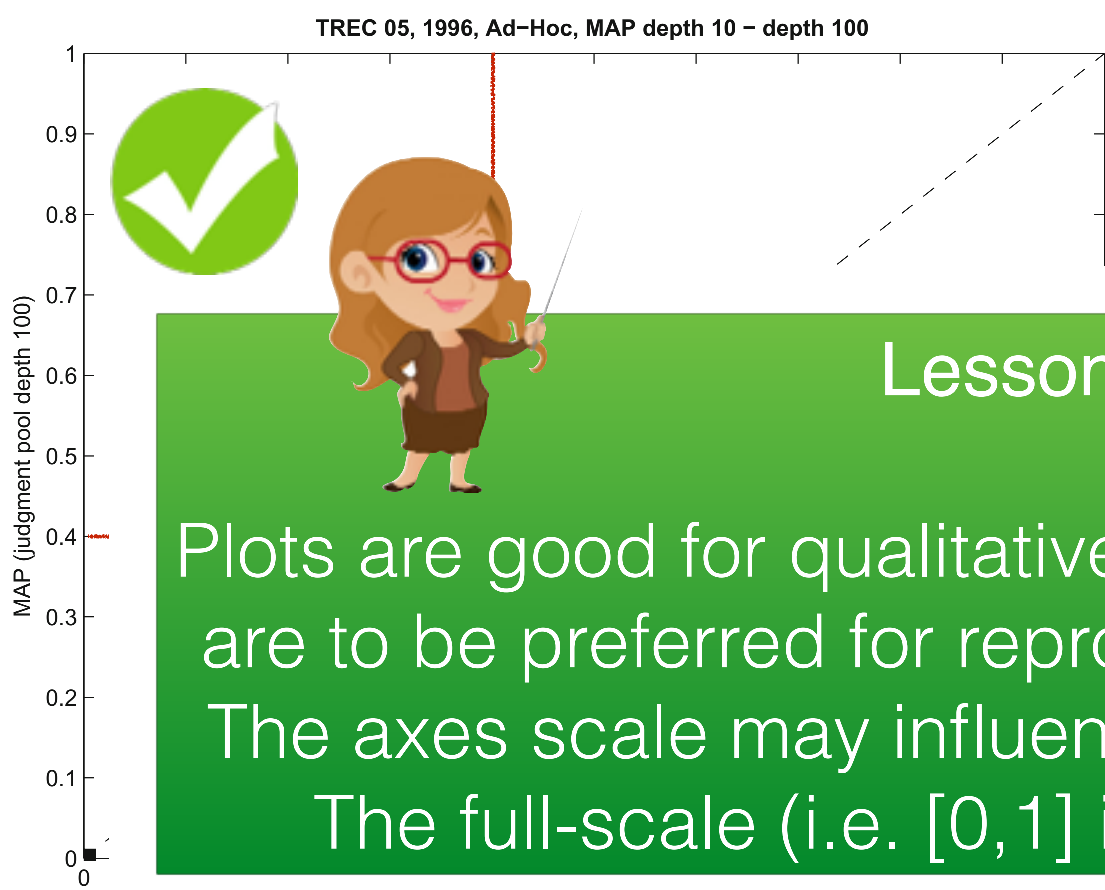


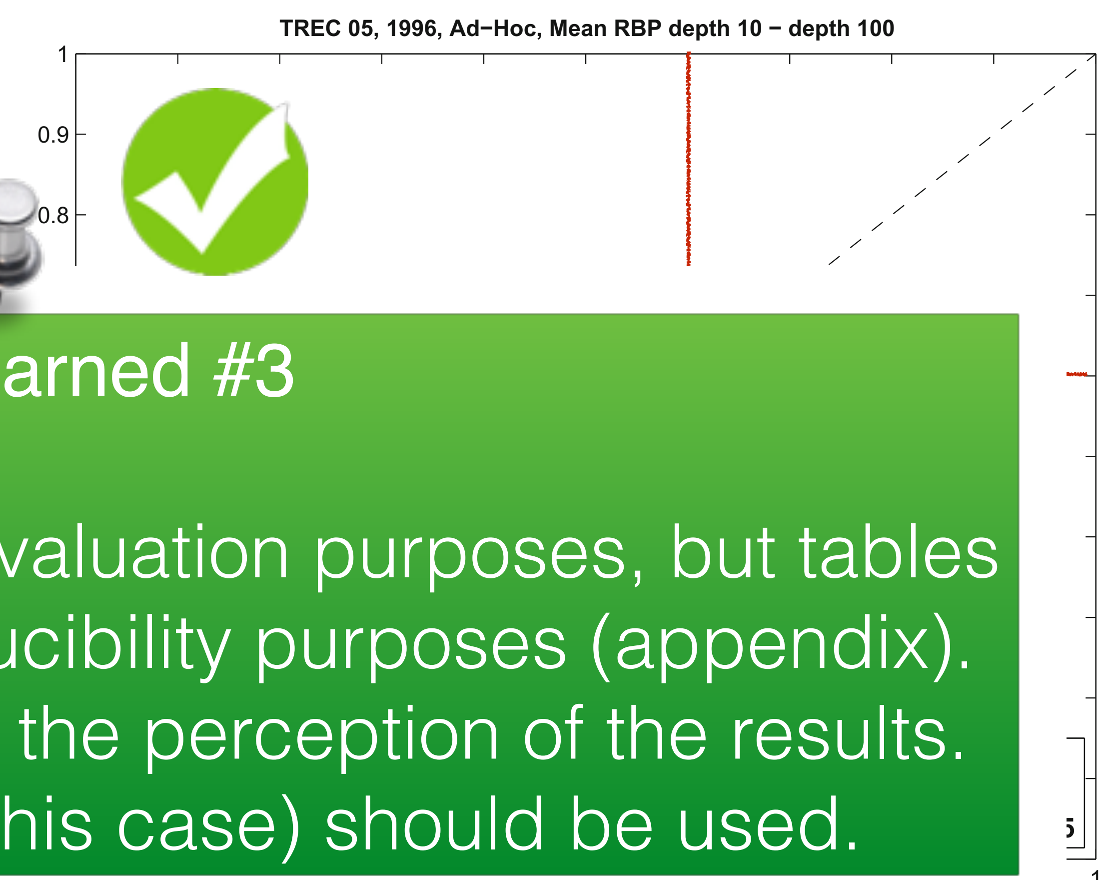
Pool Downsampling





Reproduced Results





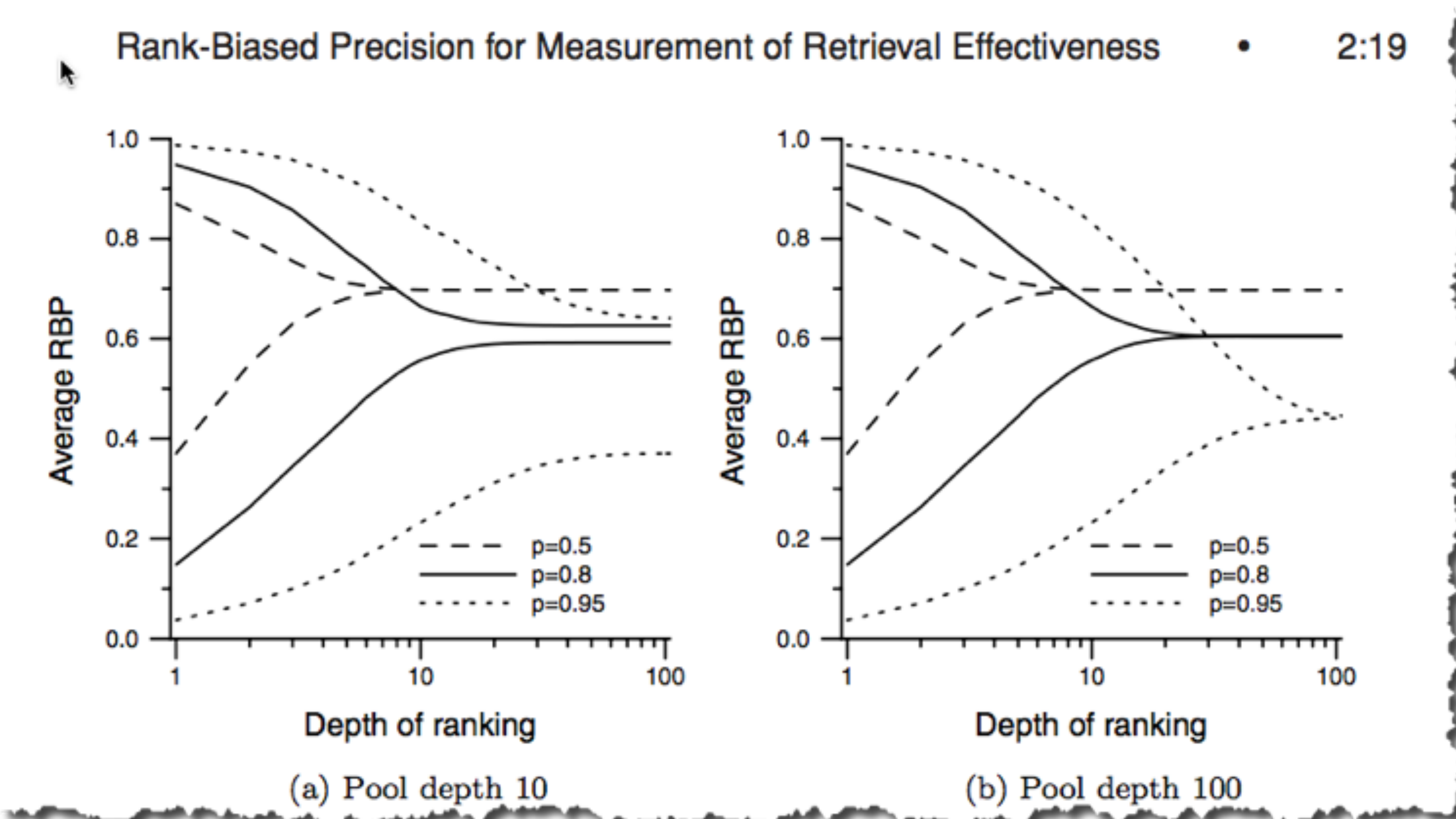
Lesson learned #3

Plots are good for qualitative evaluation purposes, but tables are to be preferred for reproducibility purposes (appendix).
The axes scale may influence the perception of the results.
The full-scale (i.e. [0,1] in this case) should be used.

2^{nd} set of experiments to be reproduced

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

2. Upper and lower bounds for RBP varying the p parameter and increasing number of documents are considered (from 1 to 100)

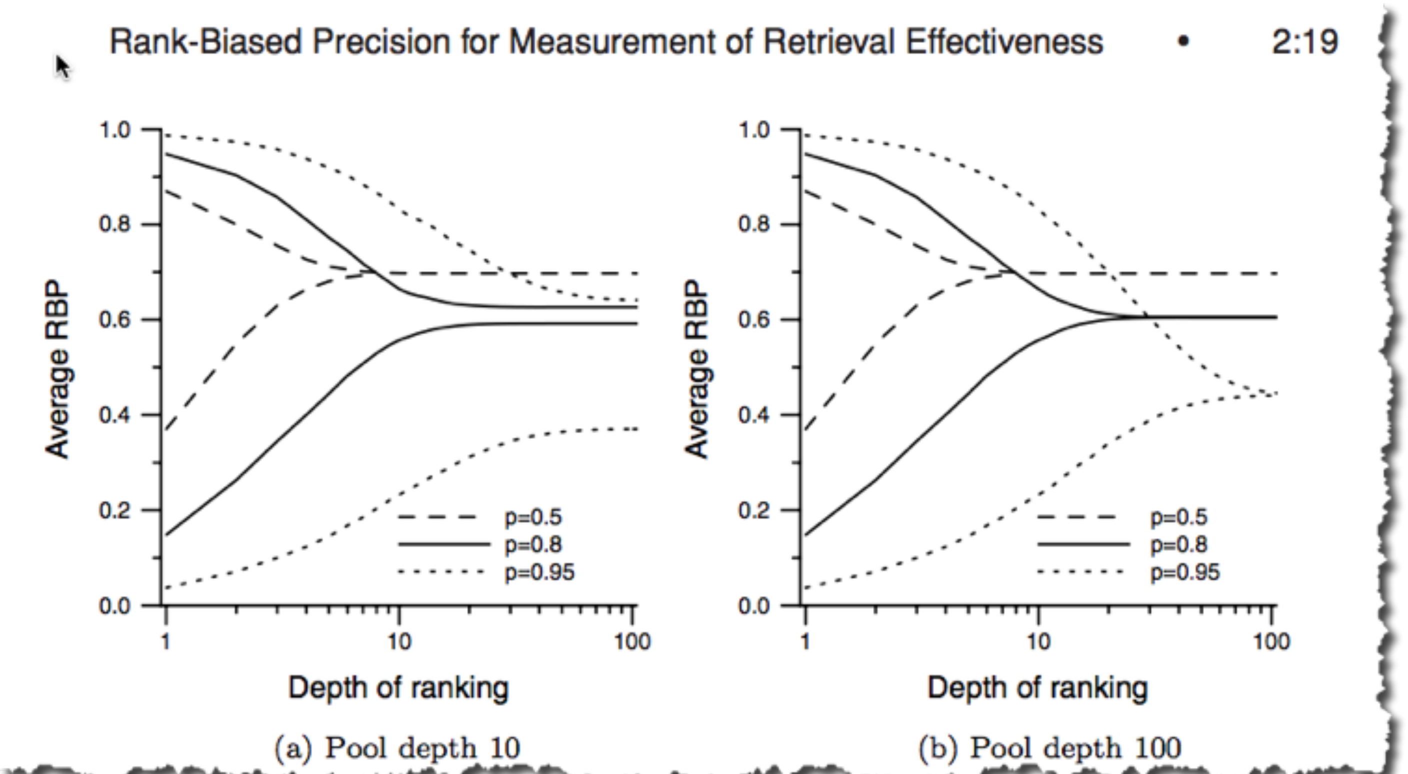


Upper and Lower Bounds

- RBP lower bounds are defined by calculating RBP in a normal setting where unjudged docs are considered as not relevant (pessimistic assumption)

- RBP upper bounds are calculated by summing the lower bounds with the *residuals* (optimistic assumption)

  - *Residuals* are calculated on an item-by-item basis by summing the weight that the docs would have had if they were relevant

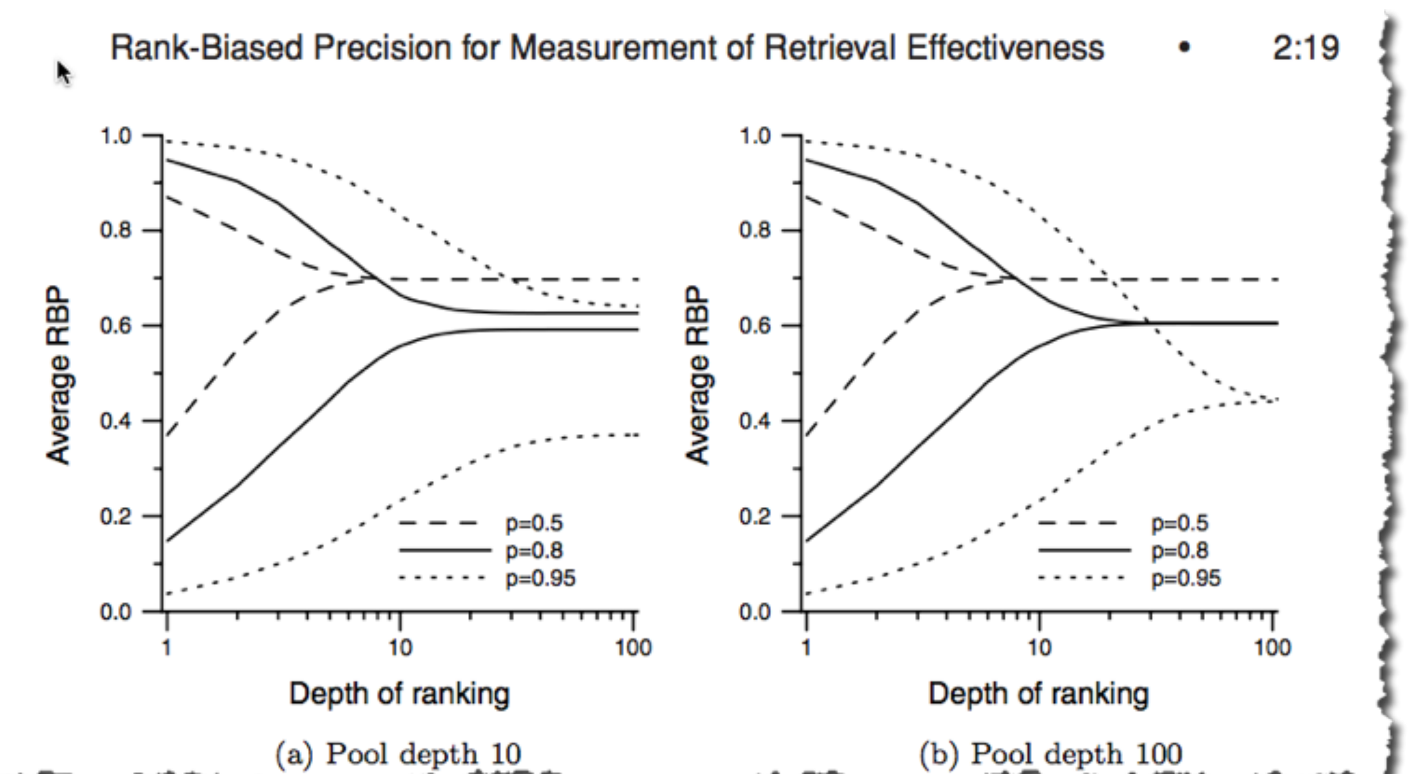


Upper and Lower Bounds

- RBP lower bounds are defined by calculating RBP in a normal setting where unjudged docs are considered as not relevant (pessimistic assumption)

- RBP upper bounds are calculated by umming the lower bounds with the

The goal is to show that the bounds stabilize as the depth of evaluation is increased

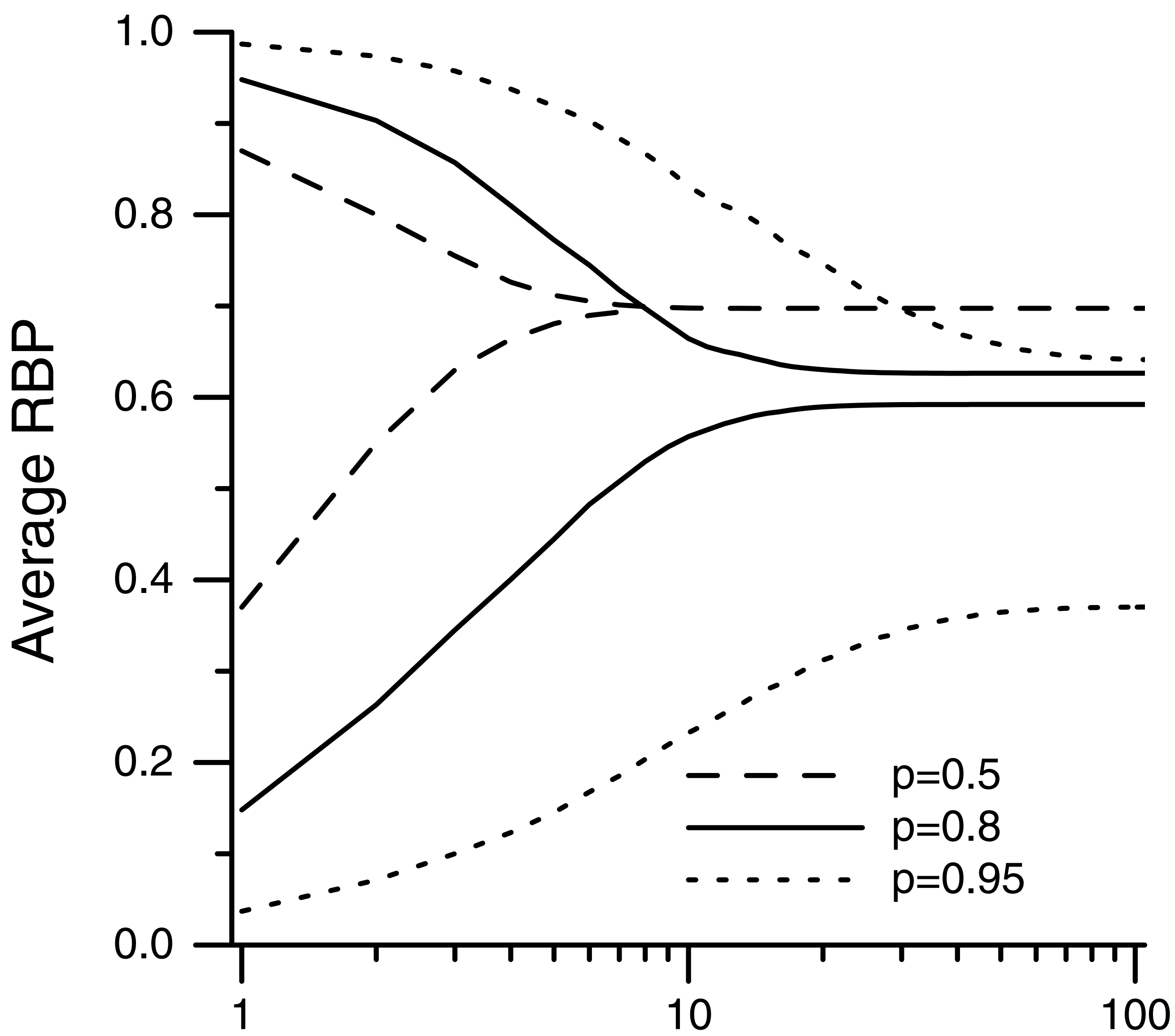


DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

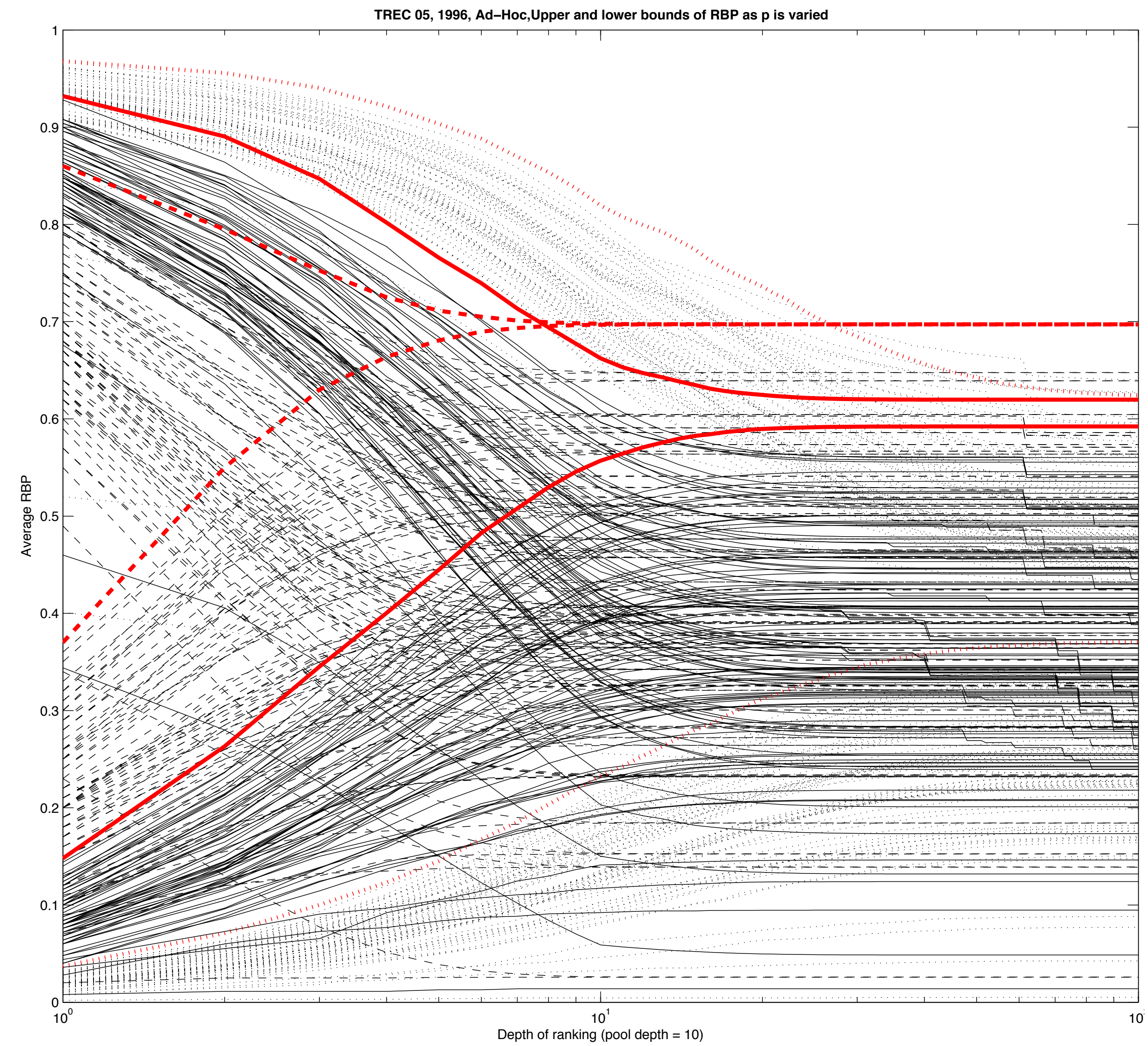
Upper and Lower Bounds

- In the original RBP paper there is no indication about which run has been used to produce the bounds plots

Original plot for one run (pool depth 10)



Reproduced plot for all the runs (pool depth 10)

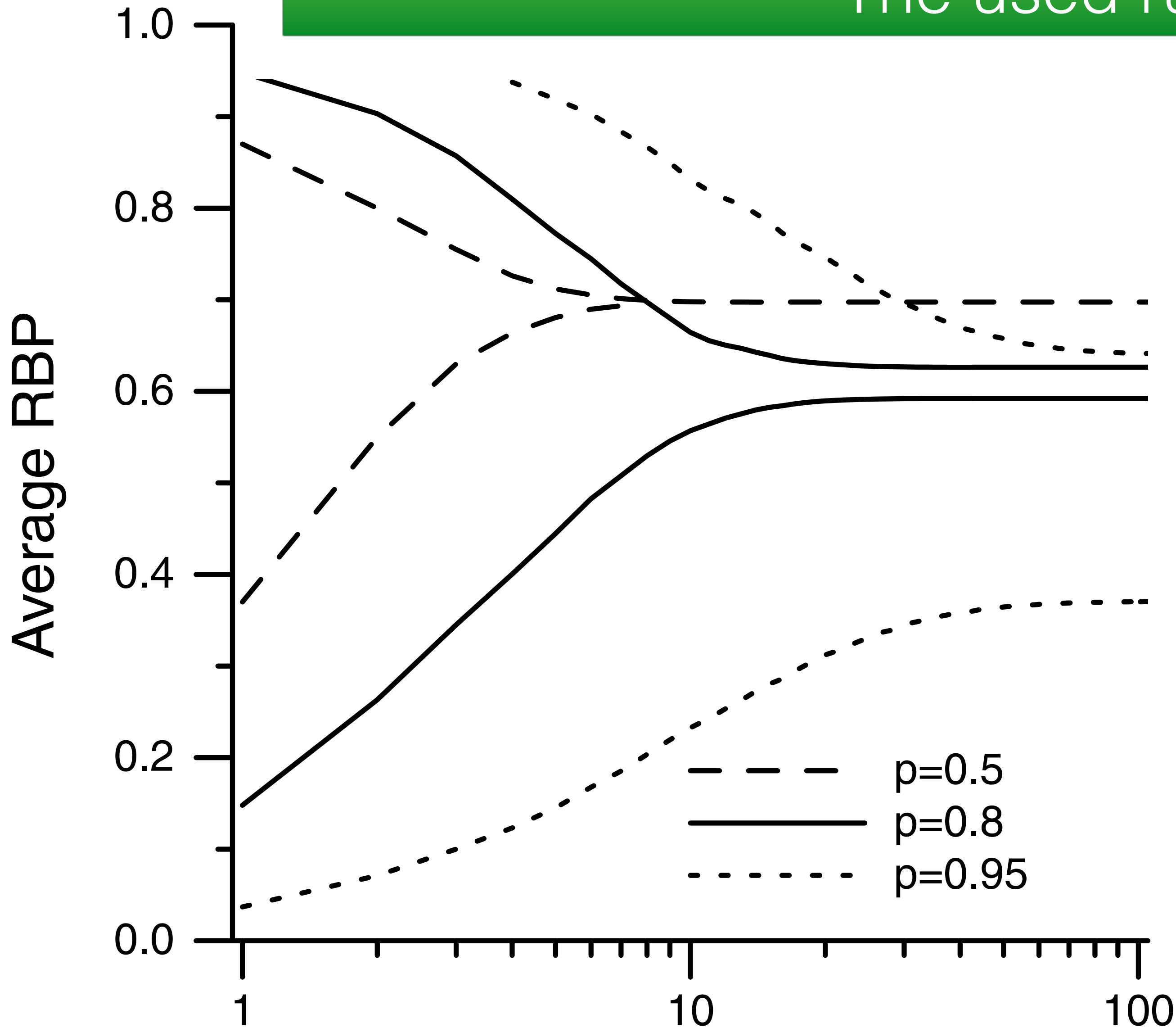


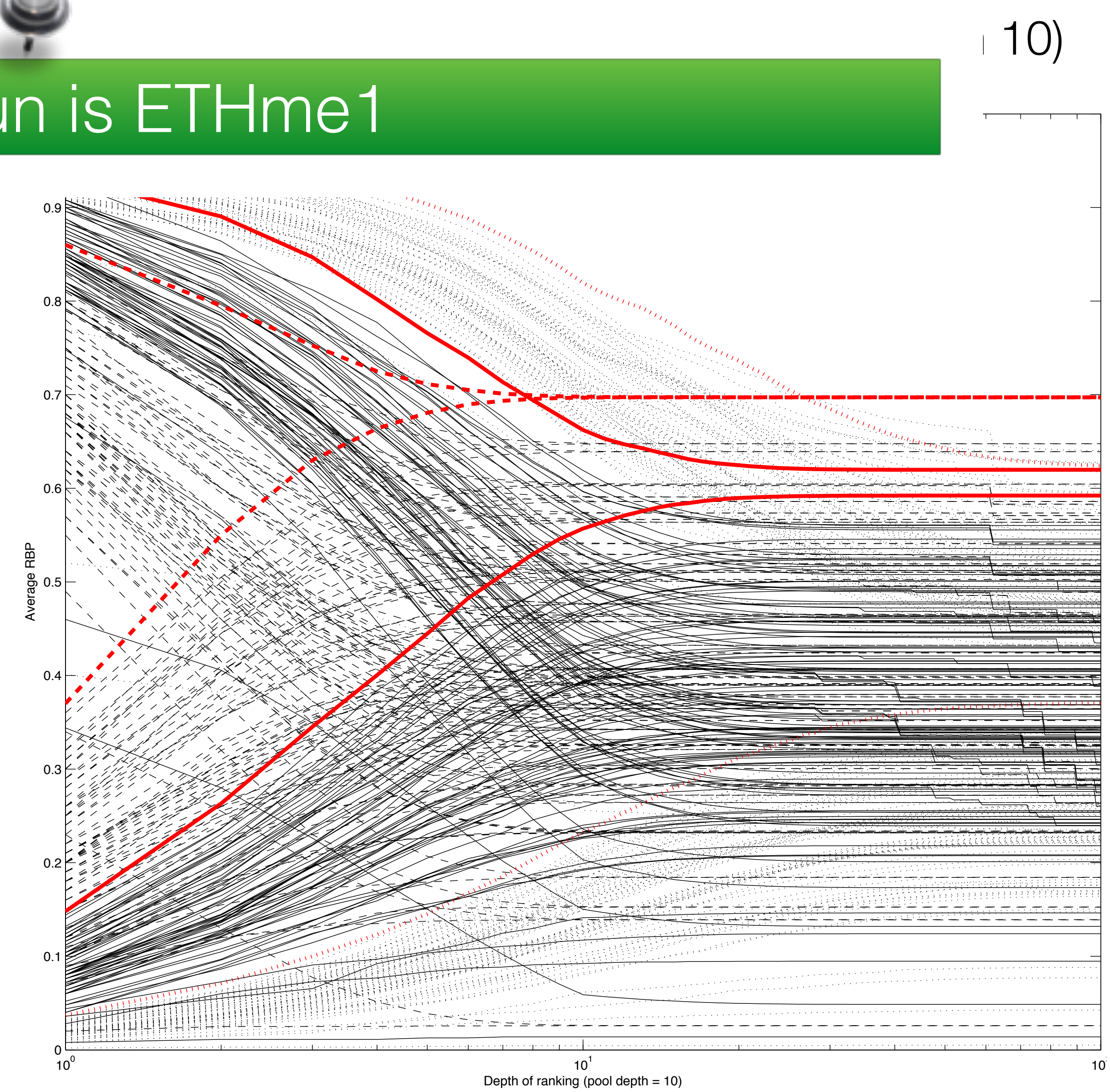
Upper and Lower Bounds

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

- In the original RBP paper there is no indication about <u>which run</u> has been used to produce the bounds plots

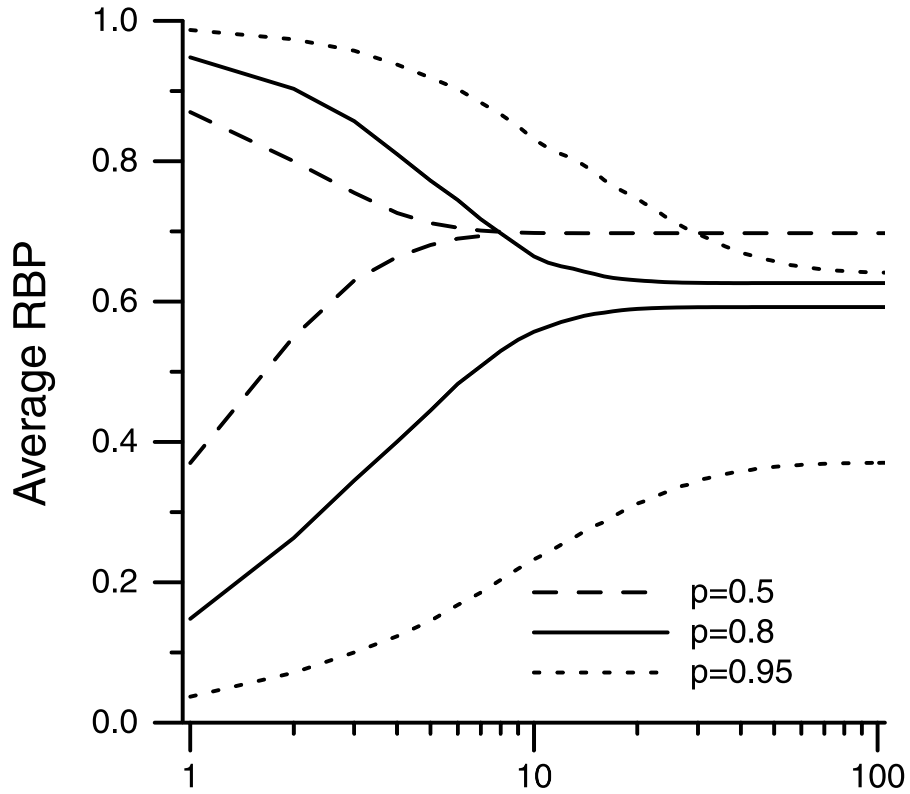
The used run is ETHme1

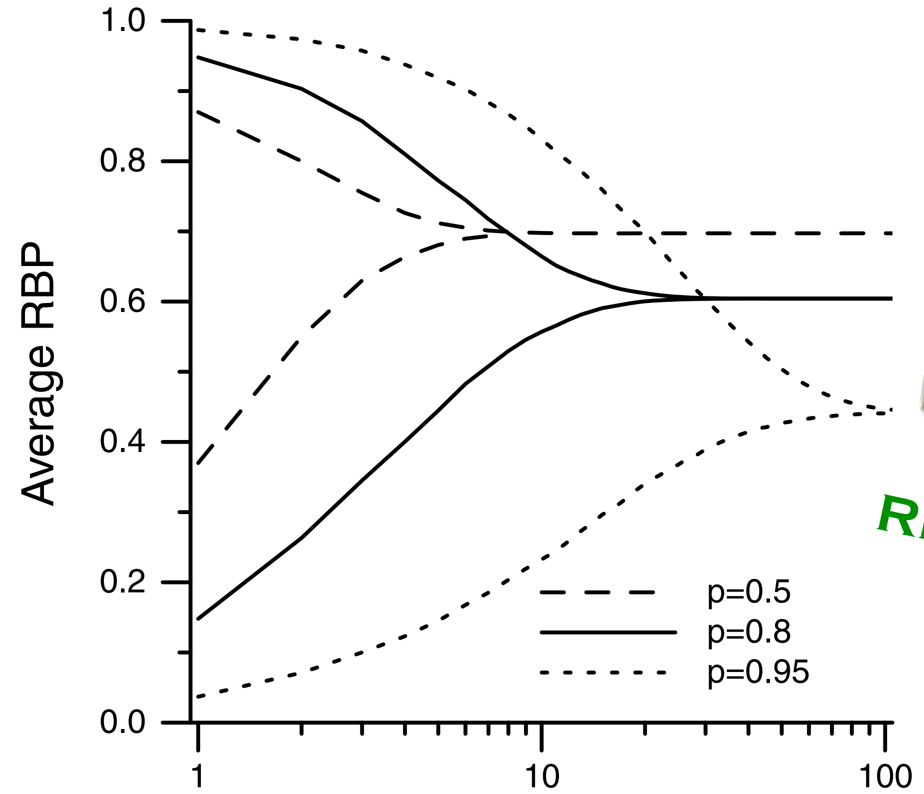




Upper and Lower Bounds

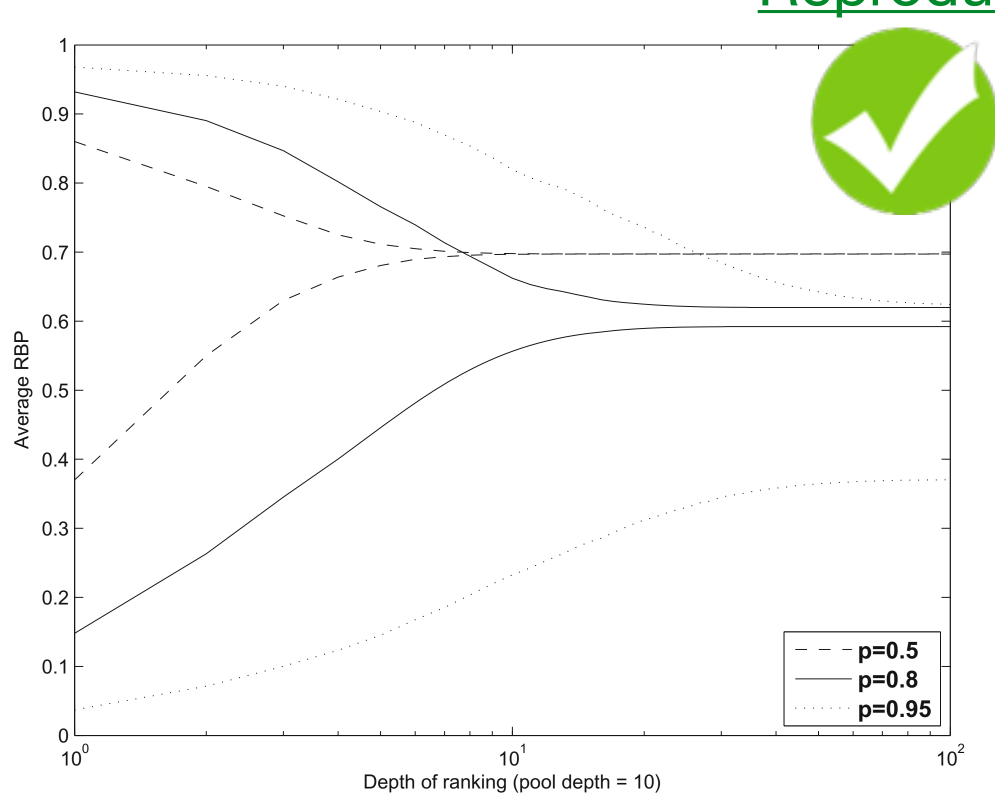
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

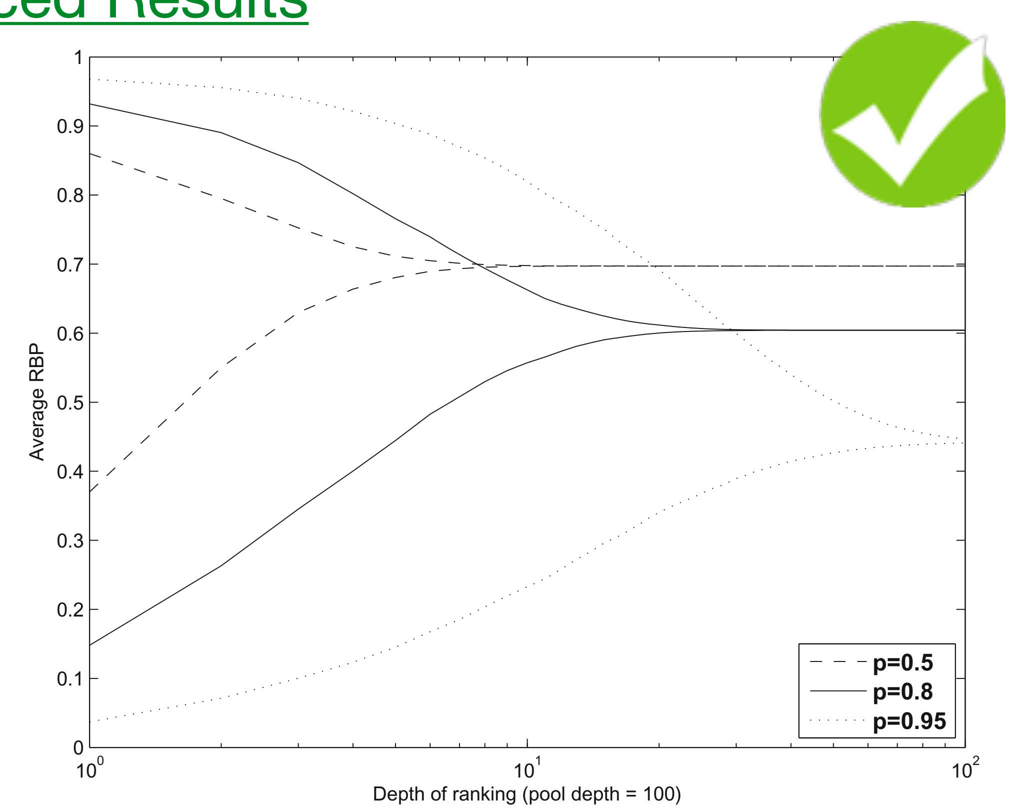




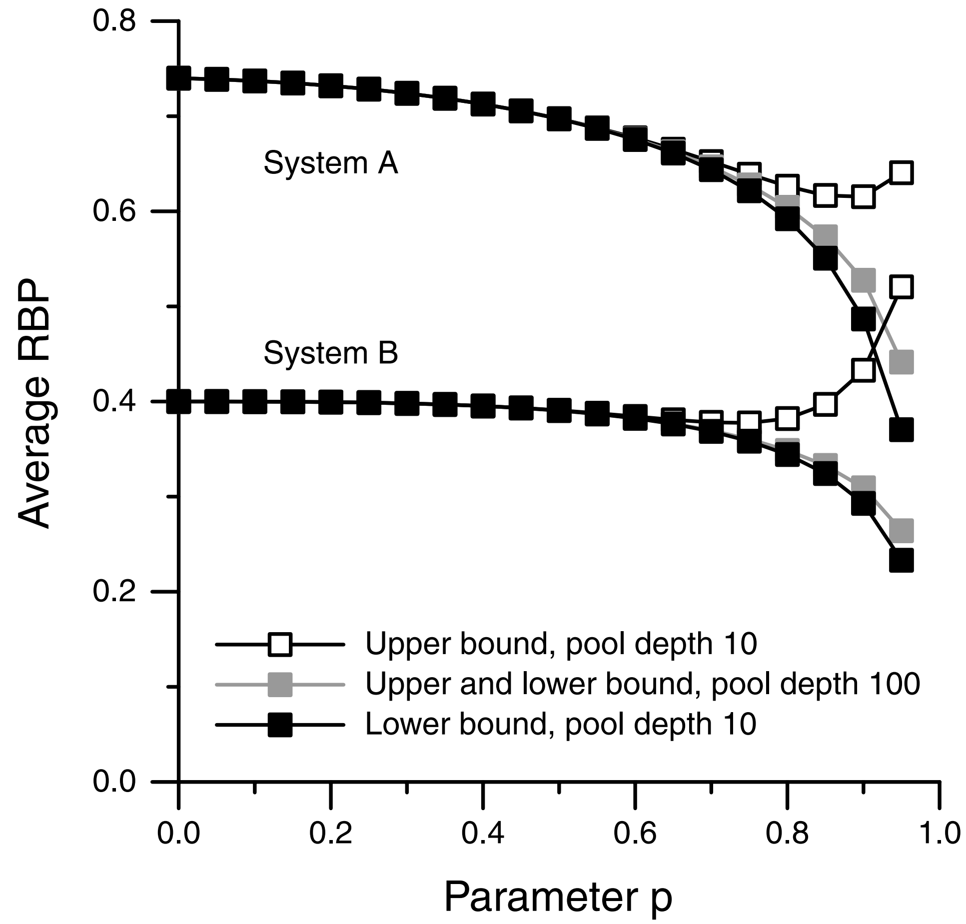
REPRODUCED

Reproduced Results





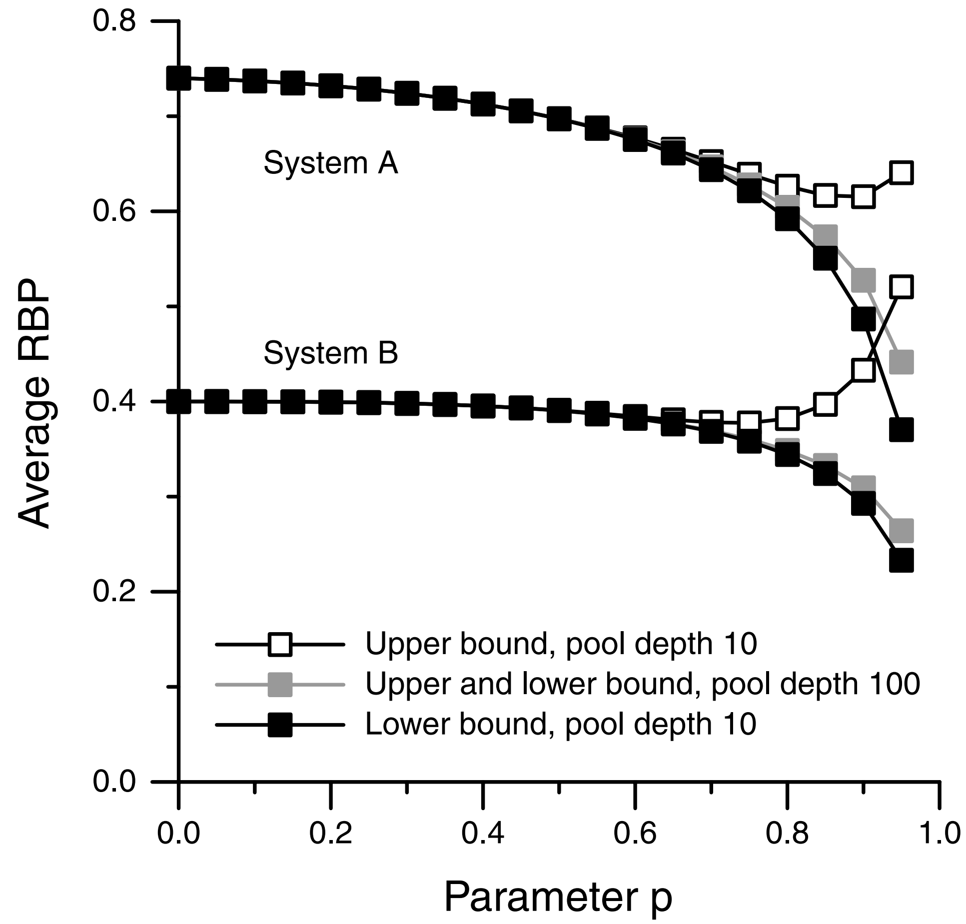
Upper and Lower Bounds



- The previous problem affects the experiment reported in Figure 6 on page 20: the names of the system A and system B are not reported

- There are 1830 possible pairs of system in TREC-5

Upper and Lower Bounds



- The p us problem affect the experiment reported in

Lesson learned #4

The need for confidentiality may make more difficult to reproduce results. If explicitly mentioning system ids is not detrimental, they should be reported.

-

3rd set of experiments to be reproduced

3. *Discriminative power*: *t* test and Wilcoxon test for determining the rate at which different effectiveness metrics allow significant distinctions to be made between systems

2:24 • A. Moffat and J. Zobel

Table IV.
The rate at which different effectiveness metrics allow significant distinctions to be made between retrieval methods. A total of 61 system runs were pairwise compared using the TREC-5 queries, making a total of $61 \times 60/2 = 1830$ system comparisons. The four columns show the number of those tests that were judged to be significant using the indicated statistical comparison. Of the traditional metrics, AP is the most consistent, in terms of allowing systems to be experimentally separated; of the RBP variants, that with $p = 0.95$ is the most consistent. The NDCG measure is a little better than both RBP and AP. In all cases the test undertaken was a two-tailed one, to answer the question "Are the two systems significantly different?"

Metric	Wilcoxon		*t* test	
	95%	99%	95%	99%
RR	1020	759	1000	752
P@10	1141	897	1150	915
P@R	1209	989	1142	931
AP	1259	1077	1164	969
RBP, $p = 0.5$	1067	834	1050	810
RBP, $p = 0.8$	1164	919	1166	917
RBP, $p = 0.95$	1231	1006	1209	987
NDCG	1291	1092	1269	1101

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

t test and Wilcoxon test

Metric	Wilcoxon		t test	
	95%	99%	95%	99%
RR	1020	759	1000	752
P@10	$\boxed{1141}$	897	1150	915
P@R	1209	989	1142	931
AP	1259	1077	1164	969
RBP, $p = 0.5$	1067	$\boxed{834}$	1050	810
RBP, $p = 0.8$	1164	919	1166	917
RBP, $p = 0.95$	1231	1006	1209	987
NDCG	1291	$\boxed{1092}$	1269	$\boxed{1101}$

Reproduced results

Numbers in bold are those which are at least 1% different from those in the original RBP paper

	Wilcoxon		t test	
Metric	99%	95%	99%	95%
RR	1030	763	1000	752
P@10	**1153**	904	1150	915
P@R	1211	994	1142	931
AP	1260	1077	1164	969
RBP.5	1077	**845**	1052	812
RBP.8	1163	921	1167	918
RBP.95	1232	1009	1209	987
nDCG	1289	**1104**	1267	**1089**

t test and Wilcoxon test

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

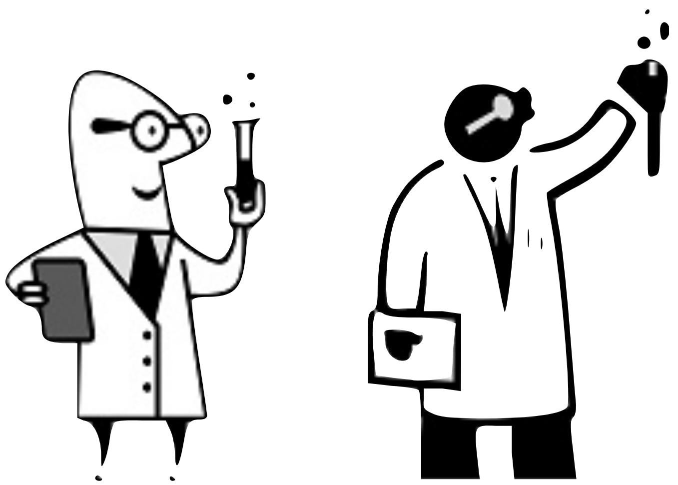
Metric	Wilcoxon		*t* test	
	95%	99%	95%	99%
RR	1020	759	1000	752
P@10	1141	897	1150	915
P@R	1209	989	1142	931
AP	1259	1077	1164	969
RBP, $p = 0.5$	1067	834	1050	810
RBP, $p = 0.8$	1164	919	1166	917
RBP, $p = 0.95$	1231	1006	1209	987
NDCG	1291	1092	1269	1101

Reproduced results

Numbers in bold are those which are at least 1% different from those in the original RBP paper

Metric	Wilcoxon		*t* test	
	99%	95%	99%	95%
RR	1030	763	1000	752
P@10	**1153**	904	1150	915
P@R	1211	994	1142	931
AP	1260	1077	1164	969
RBP.5	1077	**845**	1052	812
RBP.8	1163	921	1167	918
RBP.95	1232	1009	1209	987
nDCG	1289	**1104**	1267	**1089**

Reproducibility and Generalization



Reproducibility and Generalization

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

1. Same experiments employing the same methods but in a different context ➞ change the experimental collection

2. Same experiment employing different (but similar) methods in a different context ➞ change pool downsampling technique and experimental collection

Reproducibility and Generalization

- We investigated three main aspects:

    A. stability to deterministic downsampling at depth 10 by using two TREC and two CLEF collections

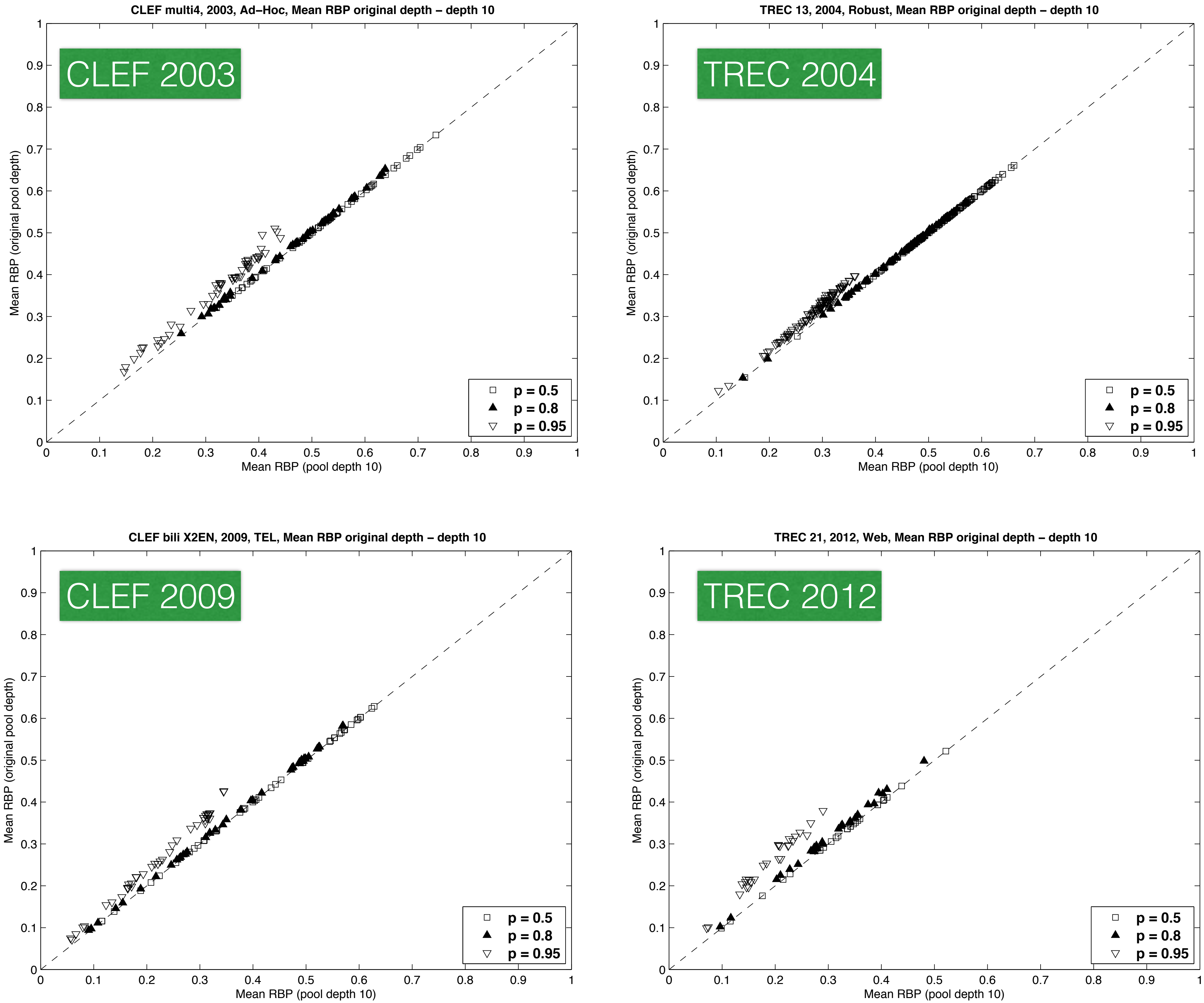
    B. robustness to downsampling according to the stratified random sampling technique (SRS)

    C. behavior of upper and lower bound in the average case

Experimental collections

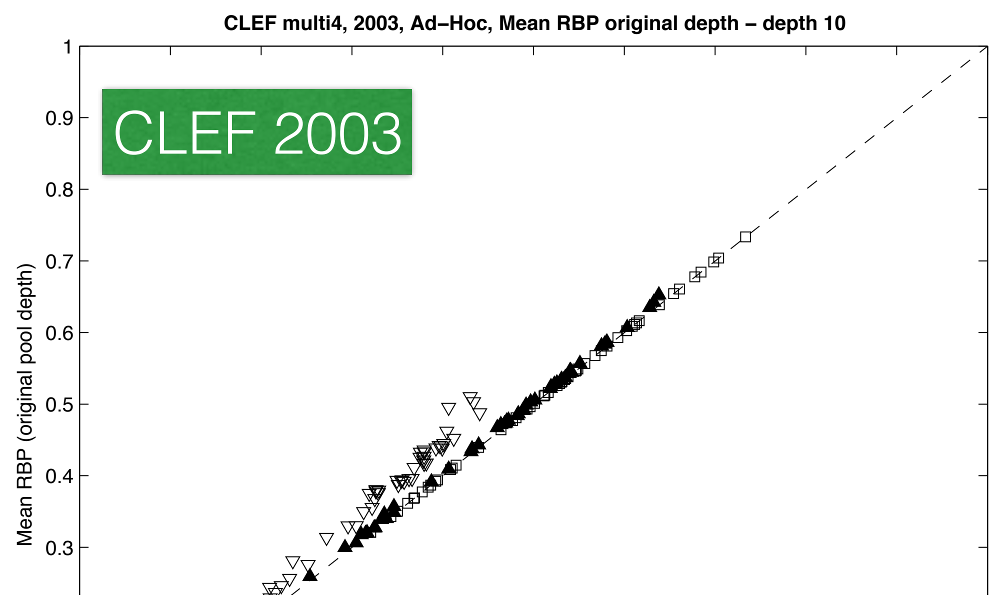
Collection	**CLEF 2003**	**TREC 13**	**CLEF 2009**	**TREC 21**
Year	2003	2004	2009	2012
Track	Ad-Hoc	Robust	TEL	Web
# Documents	1M	528K	2.1M	1B
# Topics	50	250	50	50
# Runs	52	110	43	27
Run Length	1,000	1,000	1,000	10,000
Relevance Degrees	2	3	2	4
Pool Depth	60	100 and 125	60	30 and 25
Languages	EN, FR, DE, ES	EN	DE, EL, FR, IT, ZH	EN

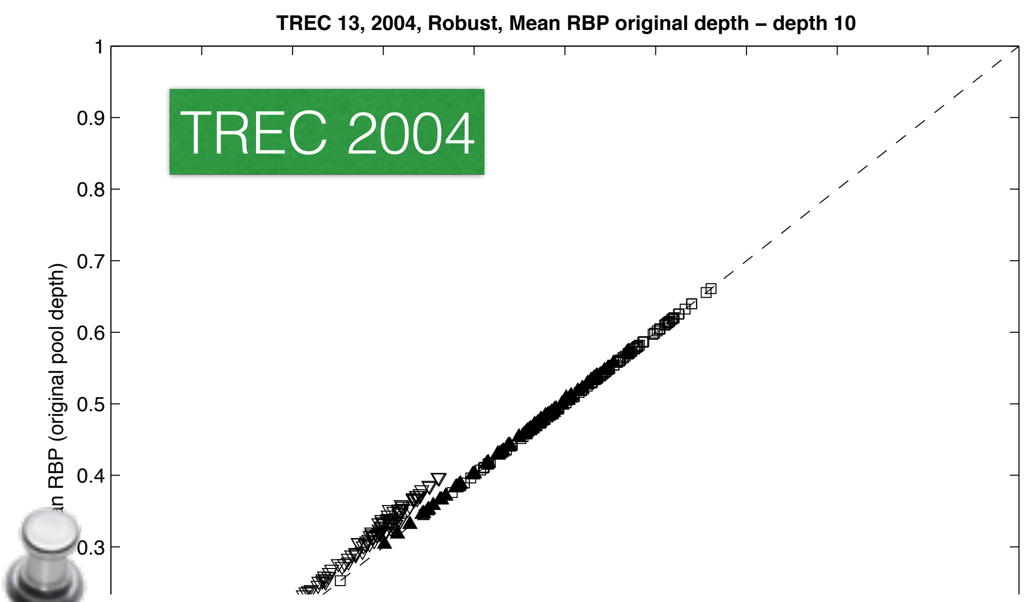
A. Stability to Deterministic Downsampling



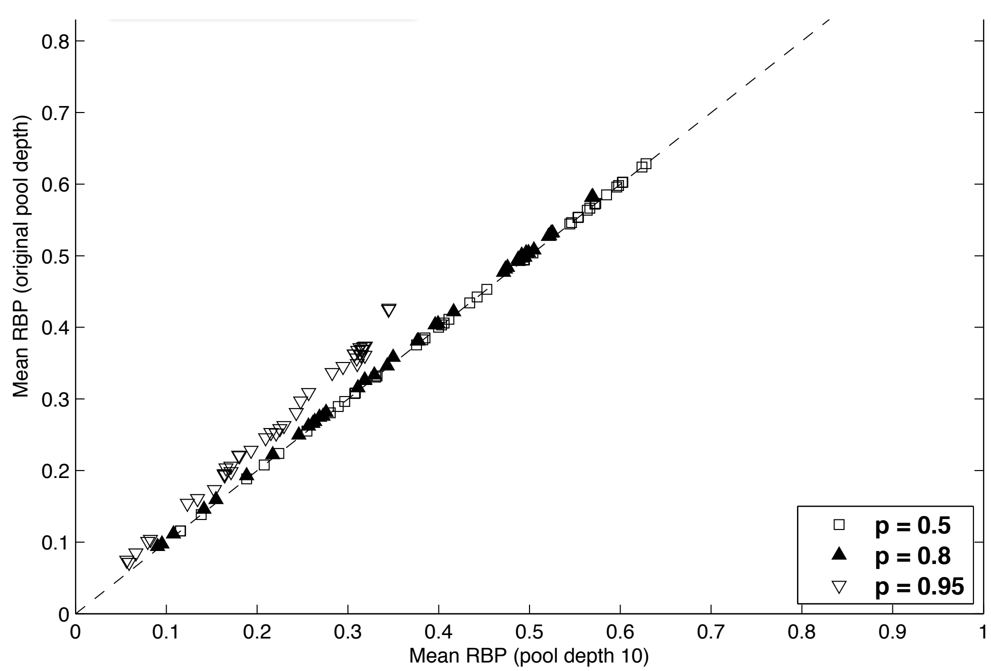
A. Stability to Deterministic Downsampling

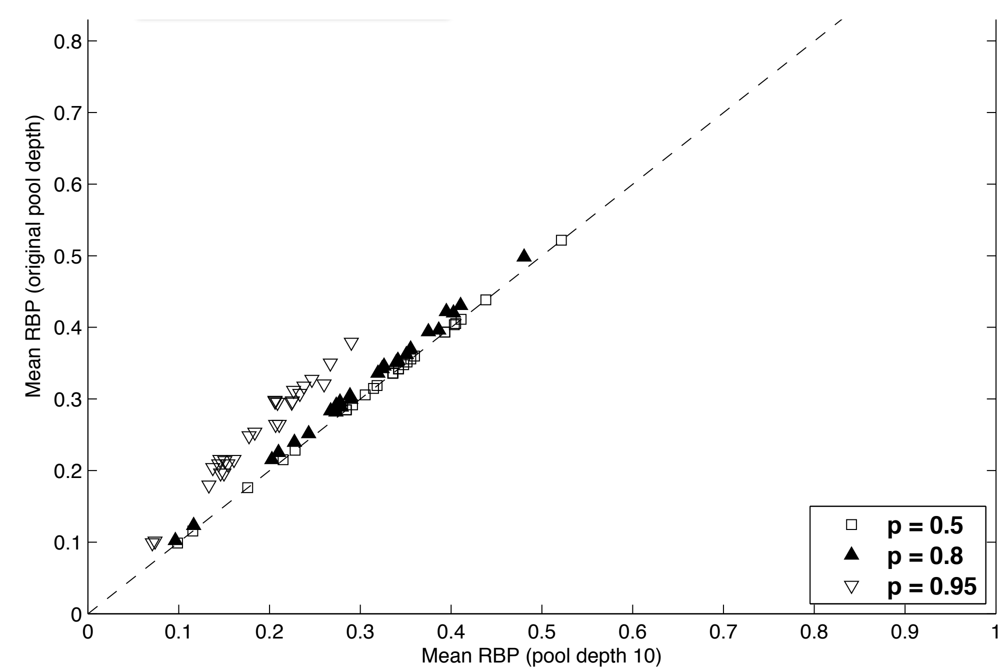
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE





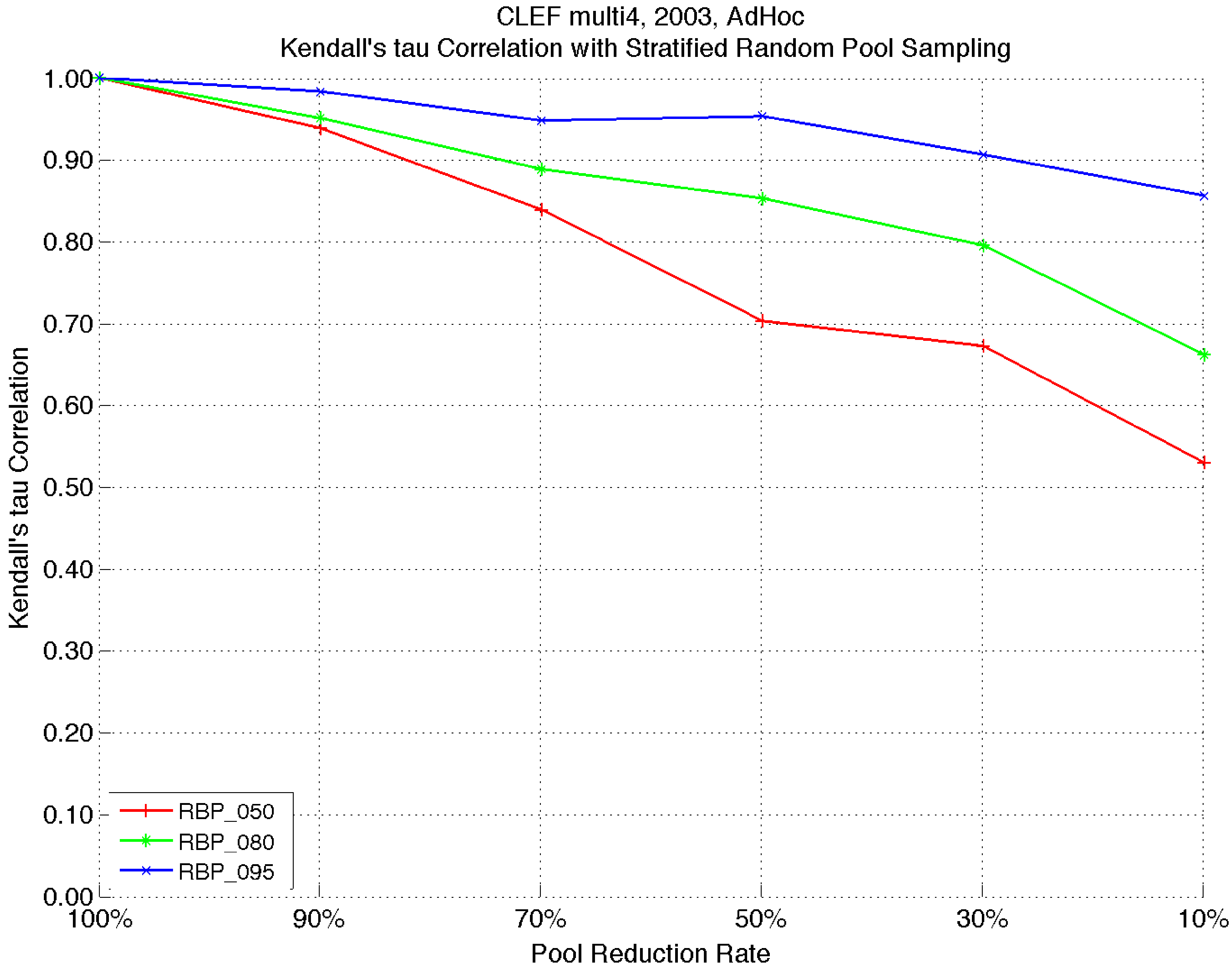
The results presented for TREC-05 are confirmed also with these collections, showing that RBP.5 and RBP.8 are robust to downsampling while RBP.95 tends to underestimate the effectiveness of the runs when using pool depth 10





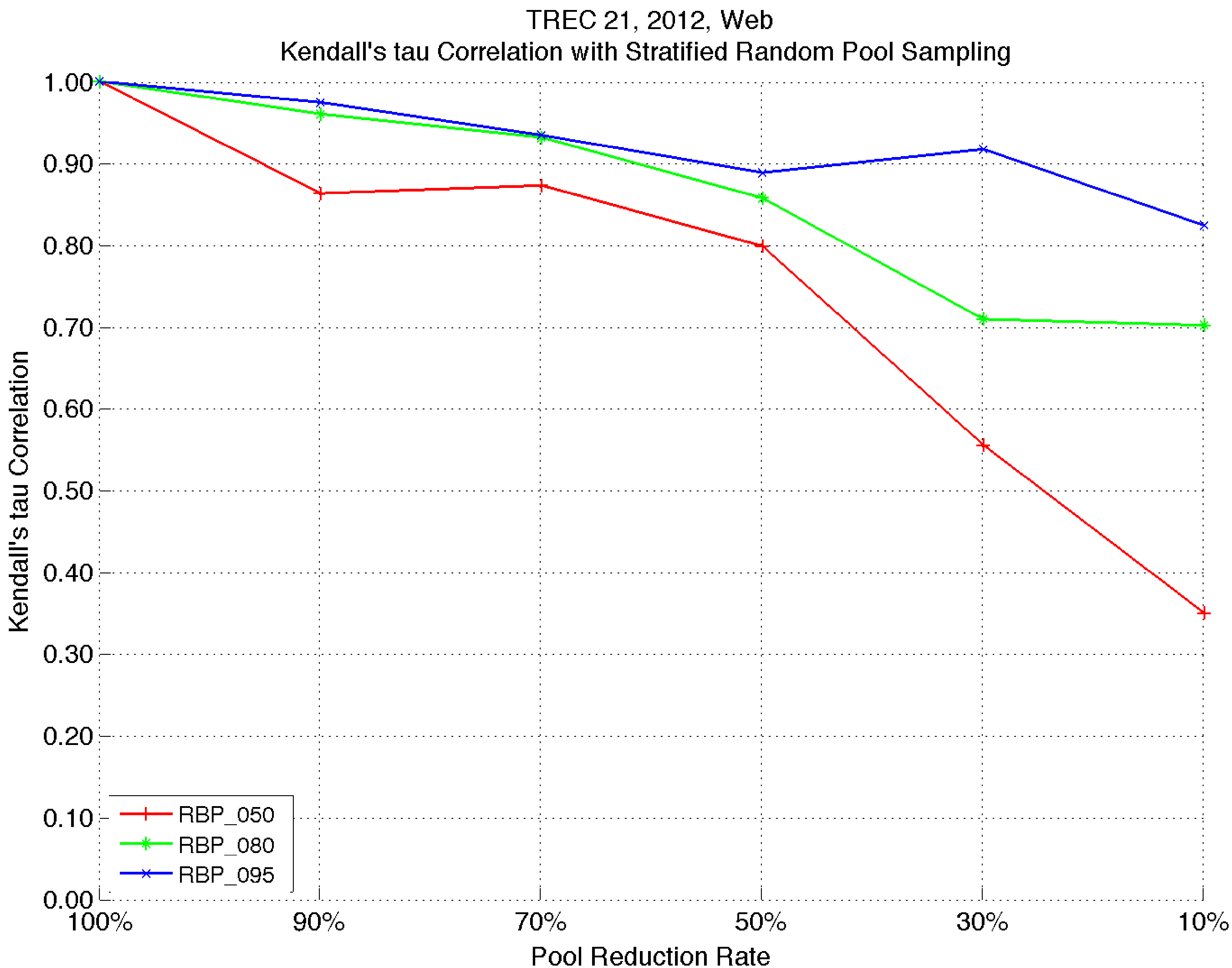
B. Robustness to SRS downsampling

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

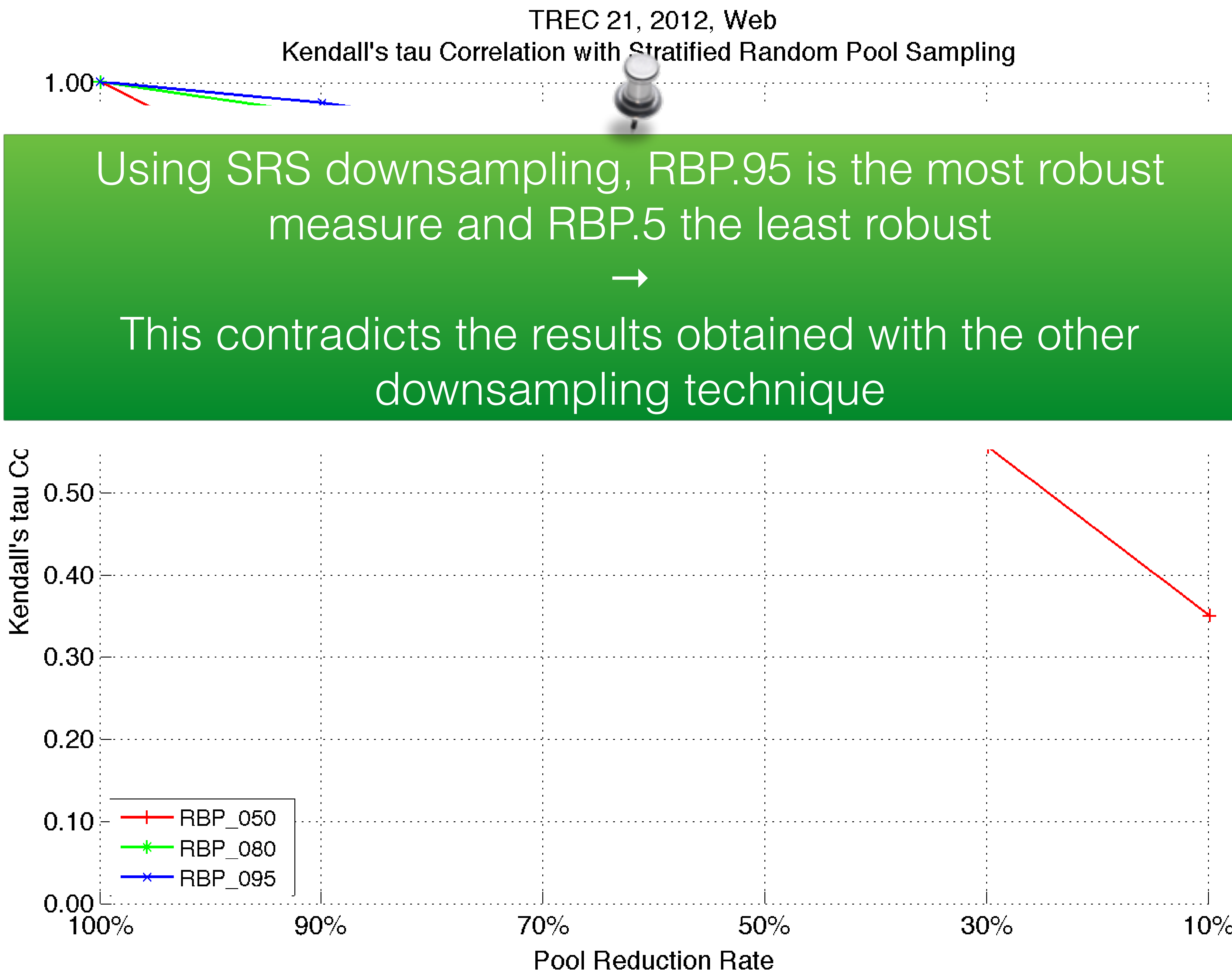


DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

B. Robustness to SRS downsampling



B. Robustness to SRS downsampling



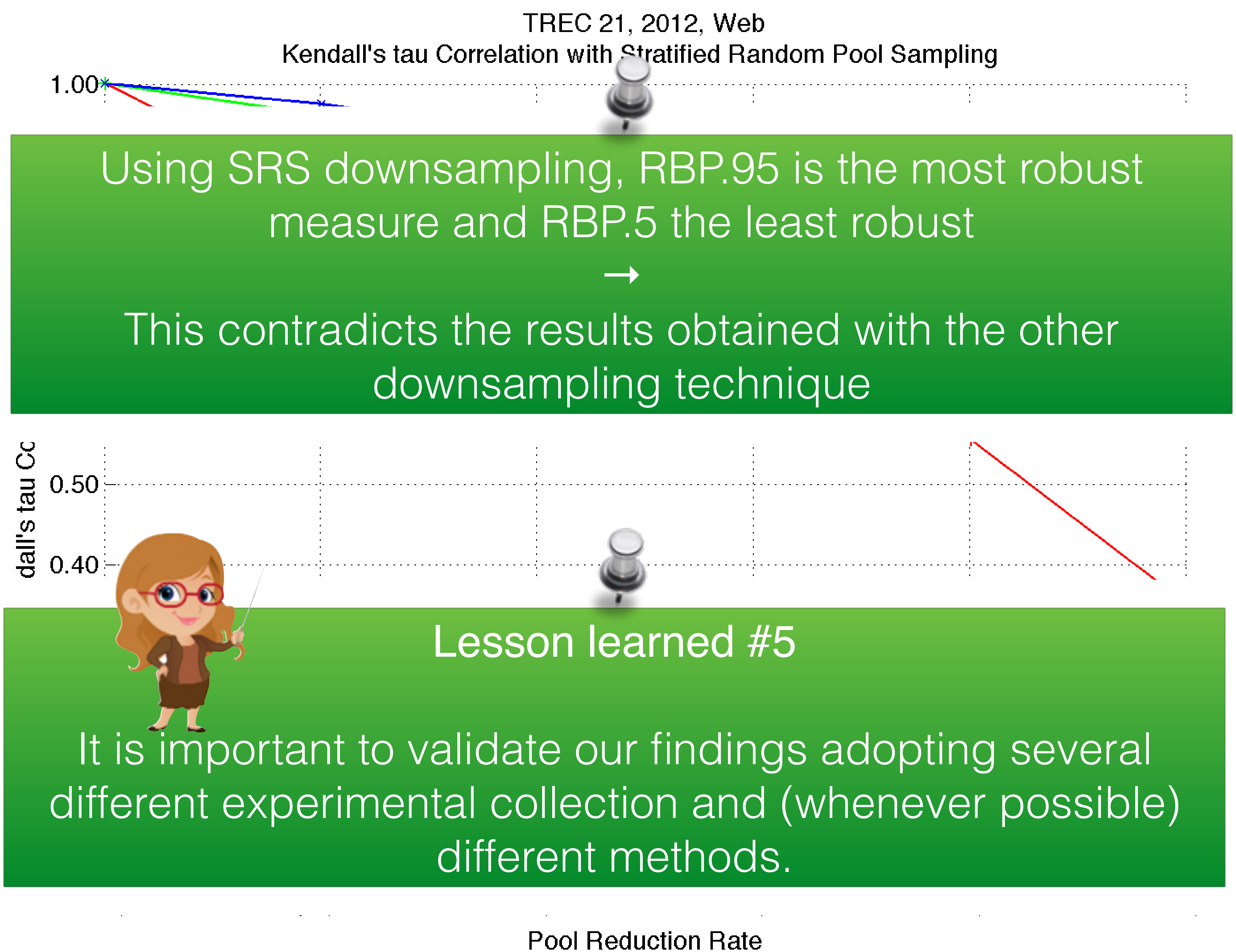
Using SRS downsampling, RBP.95 is the most robust measure and RBP.5 the least robust

→

This contradicts the results obtained with the other downsampling technique

B. Robustness to SRS downsampling

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



Using SRS downsampling, RBP.95 is the most robust measure and RBP.5 the least robust

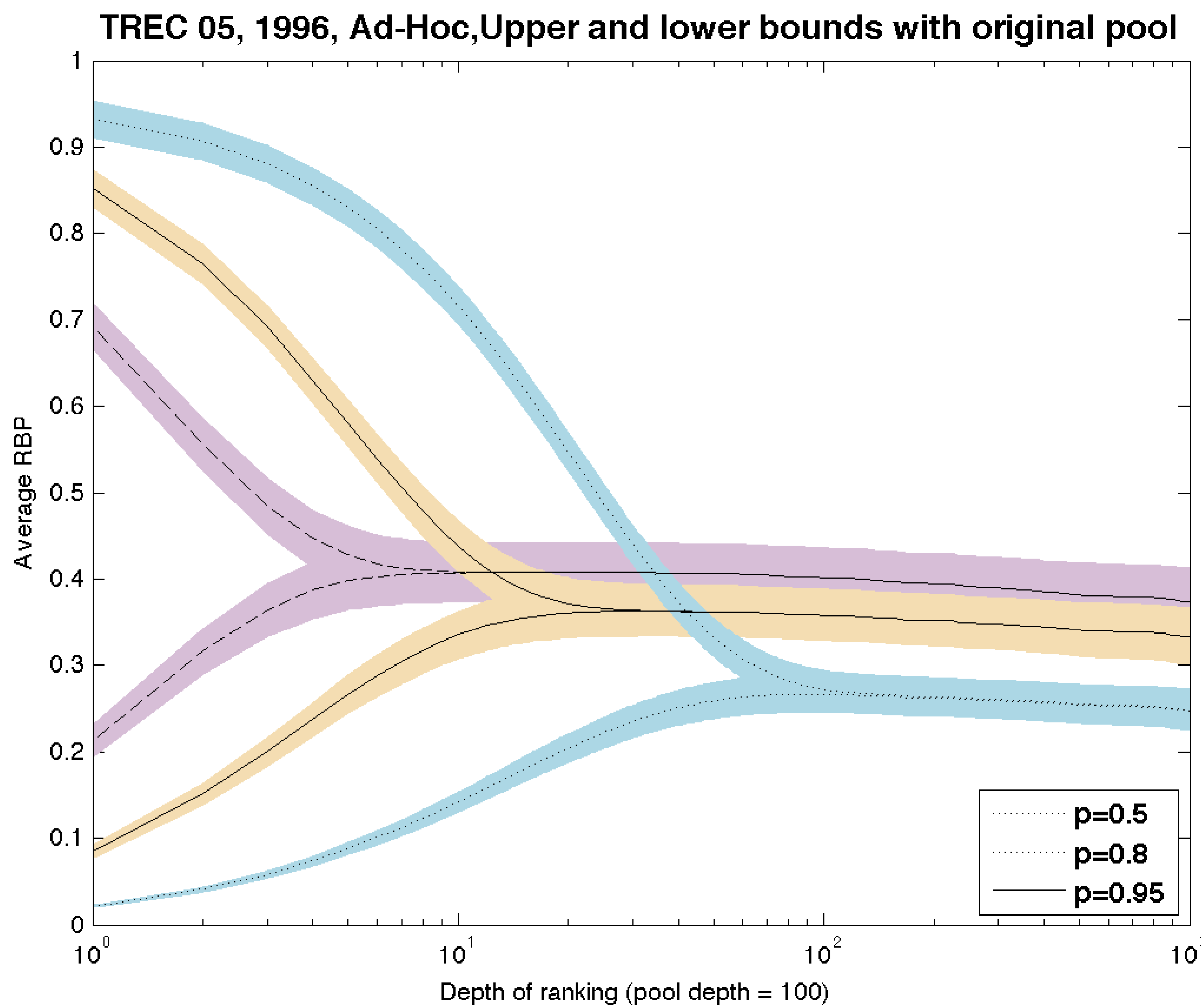
→

This contradicts the results obtained with the other downsampling technique

Lesson learned #5

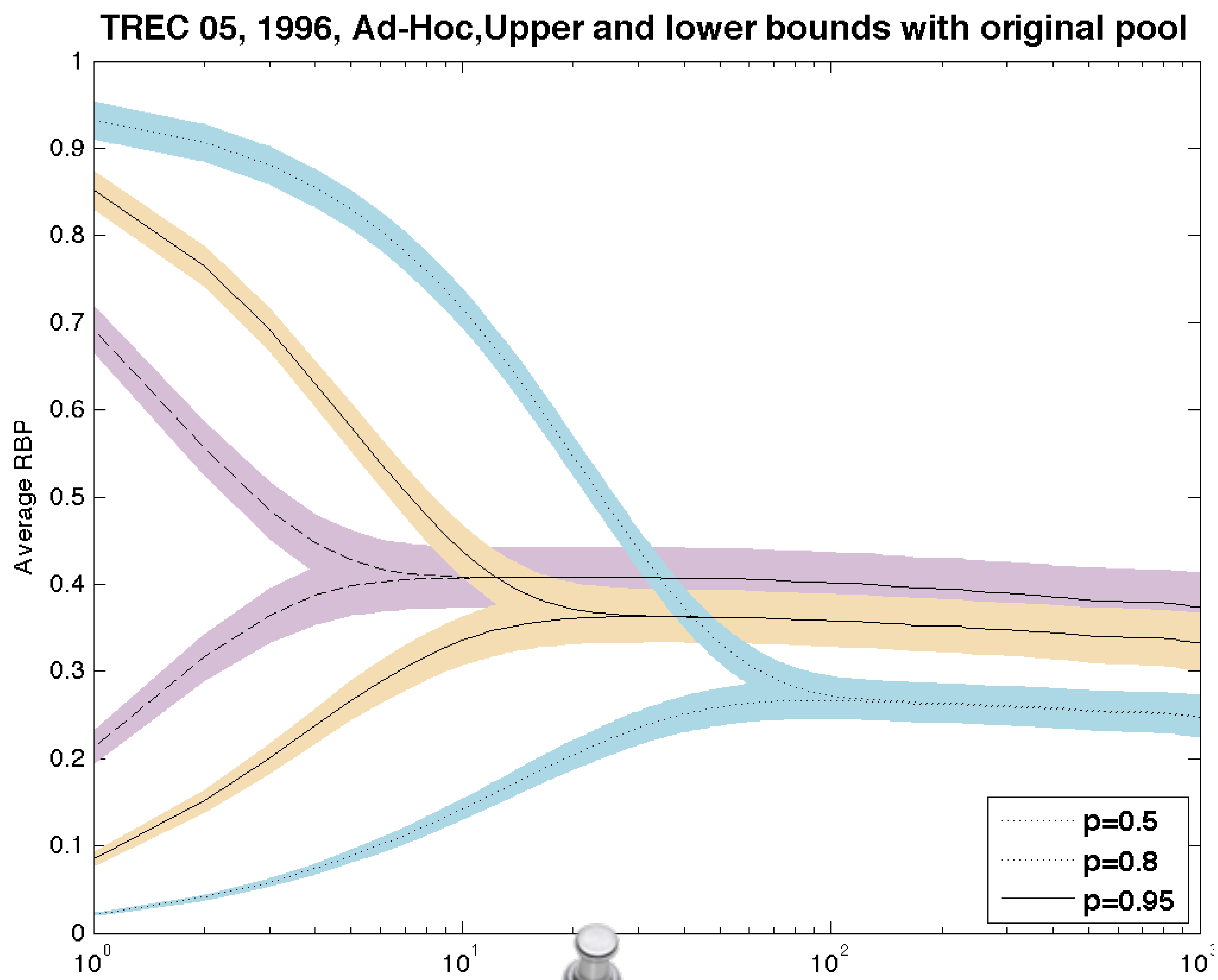
It is important to validate our findings adopting several different experimental collection and (whenever possible) different methods.

C. Bounds in the Average case



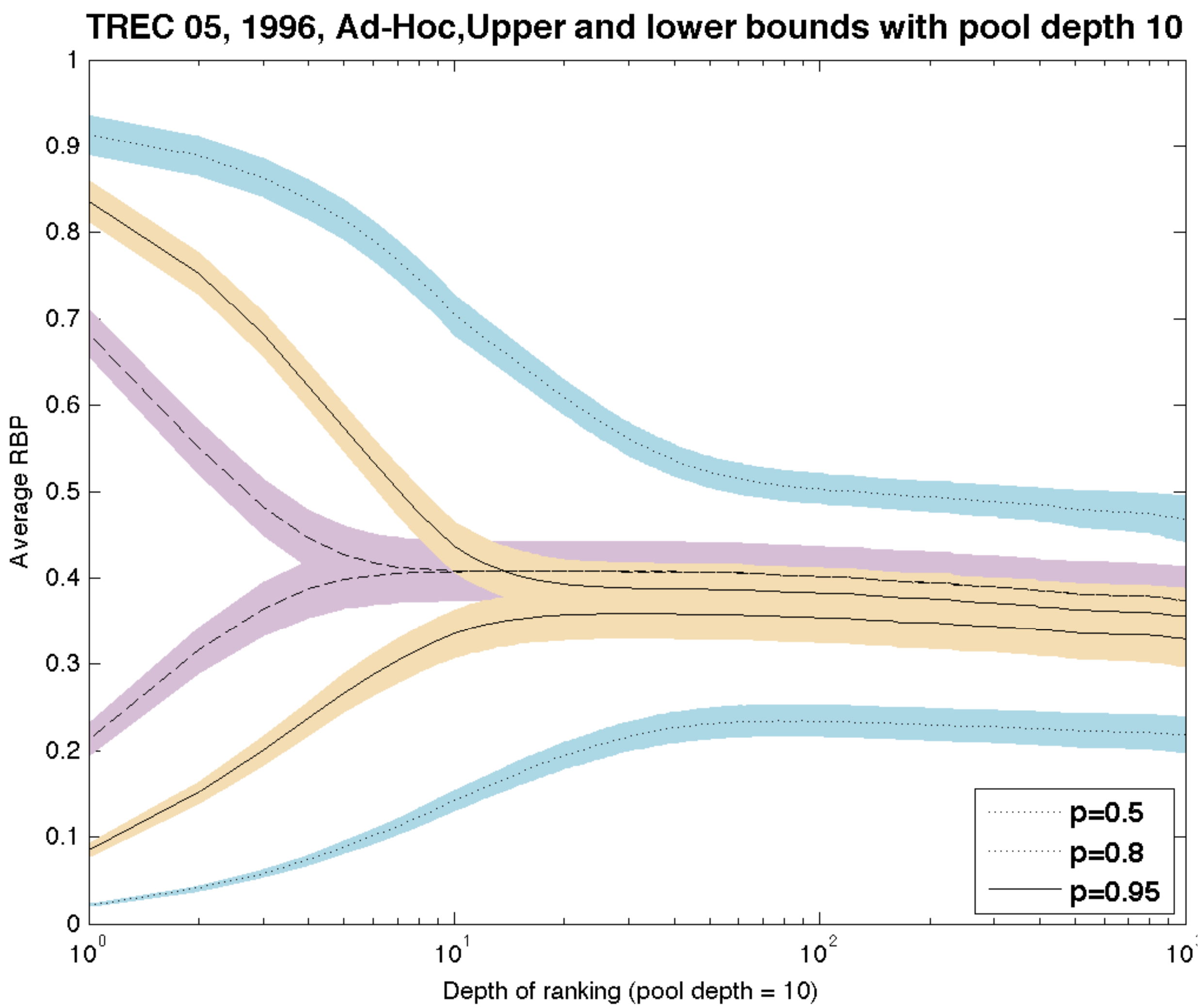
C. Bounds in the Average case

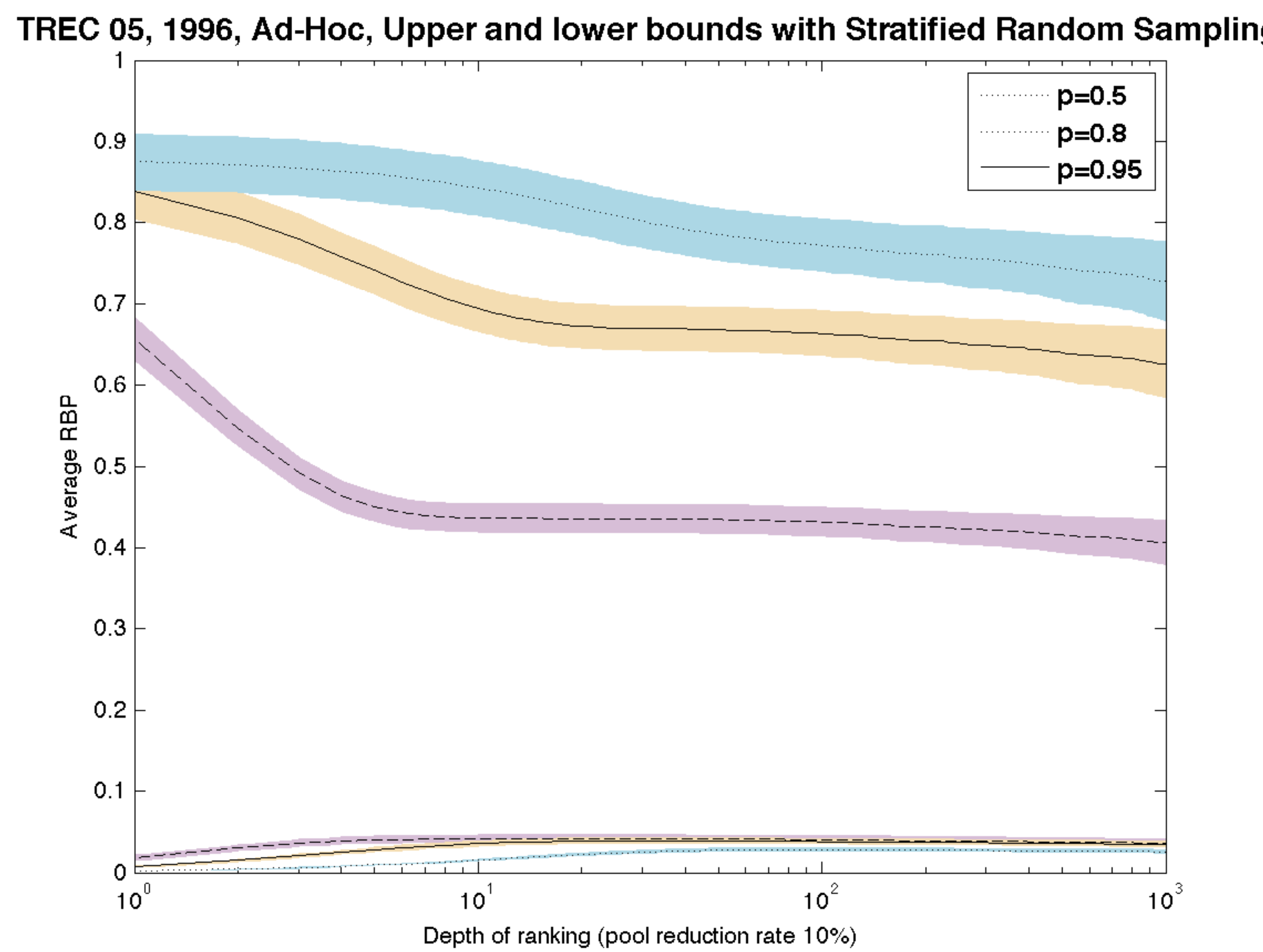
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE



The average case is better for reproducibility purposes
and more general w.r.t.
to show the behavior of only one selected run

C. Bounds in the Average case





Reproduce the reproduction: MATTERS

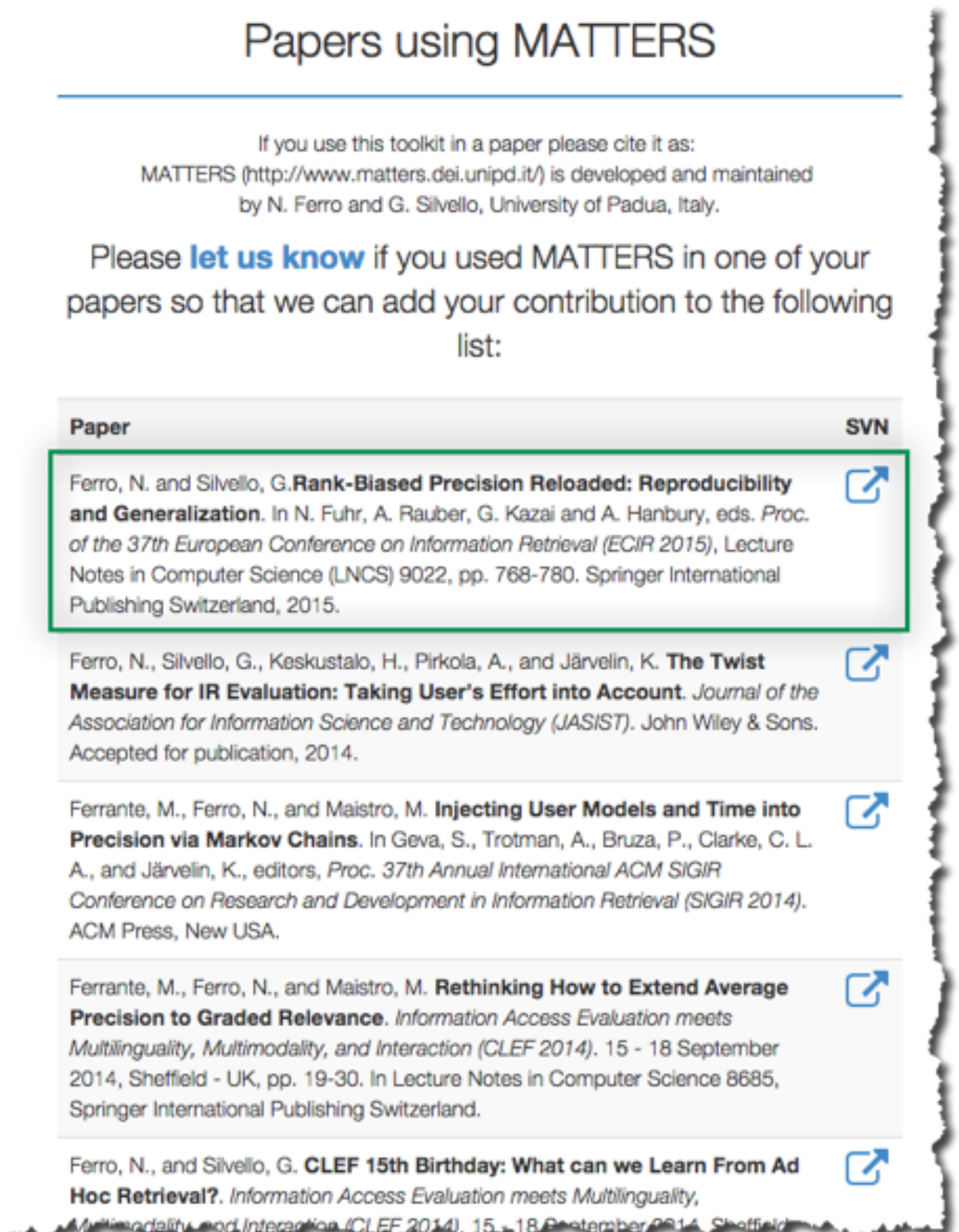
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Reproduce the reproduction: MATTERS

- Open source library written in MATLAB
- MATLAB was chosen mainly because of its
  - widely tested and robust to numerical approximations implementations of statistical methods:
    - Kendall's Tau
    - Student's *t* test
    - Wilcoxon signed rank test
    - …

http://matters.dei.unipd.it/

Reproduce the reproduction: MATTERS



Papers using MATTERS

If you use this toolkit in a paper please cite it as:
MATTERS (http://www.matters.dei.unipd.it/) is developed and maintained by N. Ferro and G. Silvello, University of Padua, Italy.

Please **let us know** if you used MATTERS in one of your papers so that we can add your contribution to the following list:

Paper	SVN
Ferro, N. and Silvello, G.**Rank-Biased Precision Reloaded: Reproducibility and Generalization**. In N. Fuhr, A. Rauber, G. Kazai and A. Hanbury, eds. *Proc. of the 37th European Conference on Information Retrieval (ECIR 2015)*, Lecture Notes in Computer Science (LNCS) 9022, pp. 768-780. Springer International Publishing Switzerland, 2015.	
Ferro, N., Silvello, G., Keskustalo, H., Pirkola, A., and Järvelin, K. **The Twist Measure for IR Evaluation: Taking User's Effort into Account**. *Journal of the Association for Information Science and Technology (JASIST)*. John Wiley & Sons. Accepted for publication, 2014.	
Ferrante, M., Ferro, N., and Maistro, M. **Injecting User Models and Time into Precision via Markov Chains**. In Geva, S., Trotman, A., Bruza, P., Clarke, C. L. A., and Järvelin, K., editors, *Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*. ACM Press, New USA.	
Ferrante, M., Ferro, N., and Maistro, M. **Rethinking How to Extend Average Precision to Graded Relevance**. *Information Access Evaluation meets Multilinguality, Multimodality, and Interaction (CLEF 2014)*. 15 - 18 September 2014, Sheffield - UK, pp. 19-30. In Lecture Notes in Computer Science 8685, Springer International Publishing Switzerland.	
Ferro, N., and Silvello, G. **CLEF 15th Birthday: What can we Learn From Ad Hoc Retrieval?**. *Information Access Evaluation meets Multilinguality, Multimodality, and Interaction (CLEF 2014)*. 15 - 18 September 2014, Sheffield	

http://matters.dei.unipd.it/

The code for reproducing this work is available at:

http://ims-svn.dei.unipd.it/repos/matters/trunk/src/scripts/papers/2014/ECIR2015-FS/

Wrapping up

- The use of public and shared experimental collections enhances reproducibility of results and eases generalization
- Data (pre-)processing choices should be explicitly reported
- Whenever possible a finding should be validated adopting different methods
- For reproducibility purposes tables are better than plots (put them in an appendix or on-line)
- Share all the code for the experiments

That's it: RBP **REPRODUCED**

Questions?