System Effect Estimation by Sharding: A Comparison between ANOVA Approaches to Detect Significant Differences

Guglielmo Faggioli[1] and Nicola Ferro[1]

Department of Information Engineering, University of Padua, Italy

Abstract. The ultimate goal of the evaluation is to understand when two IR systems are (significantly) different. To this end, many comparison procedures have been developed over time. However, to date, most reproducibility efforts focused just on reproducing systems and algorithms, almost fully neglecting to investigate the reproducibility of the methods we use to compare our systems. In this paper, we focus on methods based on ANalysis Of VAriance (ANOVA), which explicitly model the data in terms of different contributing effects, allowing us to obtain a more accurate estimate of significant differences. In this context, recent studies have shown how sharding the corpus can further improve the estimation of the system effect. We replicate and compare methods based on "traditional" ANOVA (tANOVA) to those based on a bootstrapped version of ANOVA (bANOVA) and those performing multiple comparisons relying on a more conservative Family-wise Error Rate (FWER) controlling approach to those relying on a more lenient False Discovery Rate (FDR) controlling approach. We found that bANOVA shows overall a good degree of reproducibility, with some limitations for what concerns the confidence intervals. Besides, compared to the tANOVA approaches, bANOVA presents greater statistical power, at the cost of lower stability. Overall, with this work, we aim at shifting the focus of reproducibility from systems alone to the methods we use to compare and analyze their performance.

1 Introduction

Comparing IR systems and identifying when they are significantly different is a critical task for both industry and academia [4, 15, 23]. In recent years, many fields have devoted a lot of effort to reproducing and generalizing their systems and algorithms [5, 7, 9, 17]. Yet, the literature still lacks reproducibility studies on the statistical tools used to compare the performance of such systems and algorithms. Using reproducible – and thus trustworthy – statistical tools is crucial to drawing robust inferences and conclusions. In this respect, our work makes a first step toward the study of the reproducibility of evaluation methodologies themselves. In this context, ANalysis Of VAriance (ANOVA) [21] is a widely used technique, where we model performance as a linear combination of factors, such as topic and system effects, and, by developing more and more sophisticated models, we accrue higher sensitivity in determining significant differences among systems. We focus on two recently developed ANOVA models. Voorhees et al. [27] used sharding of the document corpus to obtain the replicates of the performance score for every (topic, system) pairs needed to develop a model accounting not only for the main effects, but also for the interaction between topics and systems; Voorhees et al. also used an ANOVA version based on residuals bootstrapping [6], which we call bANOVA. Given the absence, at the current time, of publicly available code, we are interested in replicating some of the results presented by Voorhees et al.. Ferro and Sanderson [11] used document sharding as well but they developed a more comprehensive model, based on traditional ANOVA, which also accounts for the shard factor, the shard*system interaction, and the topic*shard interaction; we call this approach tANOVA. Another fundamental aspect to consider when comparing several IR systems is the need to adjust for *multiple comparisons* [12, 22]. Indeed, when comparing just two systems, significance tests control the *Type-I error* at the significance level α. The Type-I error is the possibility to find a statistically significant difference between a pair of systems when they are not (also called *false positive*). However, when c simultaneous tests are carried out, the probability of committing at least one Type-I error increases up to $1 - (1 - \alpha)^c$. Several procedures have been developed for controlling Type-I errors when multiple comparisons are performed [14]. Voorhees et al. adopted a lenient False Discovery Rate (FDR) correction by Benjamini and Hochberg [2]; Ferro and Sanderson used a conservative Family-wise Error Rate (FWER) correction, using the Honestly Significant Difference (HSD) method by Tukey [25]. In conclusion, we identified three aspects that can impact the reproducibility of the above-mentioned ANOVA approaches: *i)* the strategy used to obtain replicates, *ii)* the kind of ANOVA used, and *iii)* the control procedure for the pairwise comparisons problem.

Our work is articulated in two research questions:

- **RQ1:** Given the absence of publicly available code, we are interested in determining the degree of replicability of the evaluation methodology proposed in Voorhees et al. [27][1];
- **RQ2:** We are interested in studying the behaviour of tANOVA and bANOVA under different experimental settings – with respect to the above-mentioned focal points – and the generalizability of their results.

The paper is organized as follows: Section 2 discusses the related works; Section 3 details on the replicated approach (i.e. Voorhees et al. [27]) and the experimental setup; Sections 4 and 5 describe our efforts in generalizing the results by Voorhees et al. and Ferro and Sanderson; finally, Section 6 draws some conclusions and outlooks for future work.

2 Related Work

Tague-Sutcliffe and Blustein [24] used ANOVA to decompose performance into a topic and a system factor and adopted the Scheffe tests to compensate for

[1] We already have access to the code and data used by Ferro and Sanderson, so we are not interested in their replicability.

multiple comparisons. Tague-Sutcliffe and Blustein were not able to model the topic*system interaction factor due to the lack of replicates for each (topic, system) pair but, later on, Banks et al. [1] suggested that the topic*system interaction should have been a large size effect. Bodoff and Li [3] used multiple relevance judgements to obtain replicates. Ferro et al. [8], Ferro and Sanderson [10, 11], Voorhees et al. [27] investigated document shards as a mean to obtain replicates and develop more sophisticated ANOVA models. One problem when using document shards is that some topics may not have any relevant document in a shard and this prevents the computation of any performance measure on that shard. Voorhees et al. [27] solved this issue by resampling shards until all the topics have relevant documents on all the shards; they developed an ANOVA model consisting of a topic and system factors plus the topic*system interaction. Ferro et al. [8], Ferro and Sanderson [11] substituted missing values with an interpolated value. They developed models accounting for the topic, system, and shard factors as well as all their interactions. Ferro and Sanderson [11] (mathematically) proved that the system effect estimation is independent from the used interpolation value, when adopting the most accurate ANOVA model. Also Robertson and Kanoulas [20] explored the bootstrap usage to investigate the inter-topic variability and to obtain the replicates necessary to compute the interaction between topics and systems, while Robertson [19] investigated the usage of document sampling to estimate the stability of traditional IR evaluation. Multiple comparisons procedures aim at controlling either Family-wise Error Rate (FWER) [16] or False Discovery Rate (FDR) [2]. FWER is the probability of having at most one false positive among all rejected null hypoteses, and FWER-controlling procedures aim at keeping it equal to $1 - \alpha$. One of the most popular FWER correction approaches is the Honestly Significant Difference (HSD) by Tukey [25]. Given $\hat{\mu}_{.u.}$ and $\hat{\mu}_{.v.}$ the marginal means for two different systems, the test value for the HSD is computed as:

$$|tk| = \frac{|\hat{\mu}_{.u.} - \hat{\mu}_{.v.}|}{\sqrt{\frac{MS_{error}}{T \cdot S}}}$$

where: MS_{error} is the mean square error according to the ANOVA model and T and S are respectively the number of topics and shards. This test value is then compared against the critical value, obtained from $Q^{\alpha}_{R,df_{error}}$, the studentized range distribution, where R is the number of systems. Conversely, FDR-controlling procedures aim at keeping the false discovery rate (the number of false findings over all findings) at level α: this corresponds to allowing the number of false positives to increase, as long as the number of true discoveries increases. One of the most important FDR-controlling procedures is the Benjamini-Hochberg (BH) [2] procedure. It sorts in ascending order the p-values associated with N tested hypotheses. The greatest value of k for which $p_{(k)} \leq \alpha \frac{k}{N}$ is then found: null hypotheses associated to p-values in ranks from 0 to k are rejected.

3 Approach

3.1 ANOVA Models

We consider the following ANOVA models:

$$y_{ijk} = \mu_{\cdots} + \tau_i + \alpha_j + \varepsilon_{ijk} \tag{MD1}$$

$$y_{ijk} = \mu_{\cdots} + \tau_i + \alpha_j + (\tau\alpha)_{ij} + \varepsilon_{ijk} \tag{MD2}$$

$$y_{ijk} = \mu_{\cdots} + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk} + \varepsilon_{ijk} \tag{MD3}$$

where: $\mu_{\cdots}$ is the grand mean; τ_i is the effect of the i-th topic; α_j is the effect of the j-th system; β_k is the effect of the k-th shard; $(\tau\alpha)_{ij}$, $(\tau\beta)_{ik}$, and $(\alpha\beta)_{jk}$ are respectively interactions between topics and systems, topics and shards, and systems and shards; ε is the error committed by the model in predicting y. Our **(MD1)** is the model originally used by Tague-Sutcliffe and Blustein [24], it corresponds to the model in equation (2) of Voorhees et al. [27] and to (MD2) of Ferro and Sanderson [11]. Our **(MD2)** corresponds to the model in equation (3) of Voorhees et al. [27] and to (MD3) of Ferro and Sanderson [11]. Finally, our **(MD3)** corresponds to the model (MD6) of Ferro and Sanderson [11]. Voorhees et al. did not experimented with the latter model; so, its usage represents an aspect of generalizability.

3.2 Bootstrap ANOVA (bANOVA)

The bootstrap based version of ANOVA is the focus of our reproducibility study. It relies on bootstrap sampling of the residuals produced by a tradional ANOVA linear model. The use of bootstrap is motivated by the fact that, since it does not rely on the traditional F statistics, it allows for minimizing the assumptions imposed on the distribution of the data. To compute the bootstrap ANOVA, it is necessary to fit a traditional ANOVA linear model. Once the model is estimated, we can use it to compute the estimated performance $\hat{y}_{ijk}$, for the i-th topic, using the j-th system on the k-th shard. Note that estimated performance values can be organized in an estimated performance tensor $\hat{\mathbf{Y}}$, where $\hat{Y}_{ijk} = \hat{y}_{ijk}$. Afterwards, residuals are computed as $r_{ijk} = y_{ijk} - \hat{y}_{ijk}$, where y_{ijk} is the observed performance value. Called $\mathcal{R}$ the set of all residuals, B different perturbation tensors $\mathbf{R}^{(b)}$ are sampled, with $b \in \{0, ..., B-1\}$. In particular, $R^{(b)}_{ijk} = r^{(b)}_{ijk}$ where $r^{(b)}_{ijk}$ is sampled uniformly with replacement among all possible original ANOVA residuals $\mathcal{R}$. These perturbation tensors are then added to $\hat{\mathbf{Y}}$, producing B perturbed observation tensors $\tilde{\mathbf{Y}}^{(b)}$. Each perturbed observation tensor is then used to fit an ANOVA model, providing B new bootstrap sampled estimations for the effect of each system. Using these estimations, it is possible to fit a Probability Density Function (PDF) of the effect of the system. Note that, Voorhees et al. do not specify the approach to fit the PDF, and thus we used the Kernel Density Estimation (KDE) technique [28], using a Maximum Likelihood Estimation (MLE) approach. The average MLE bandwidth is 0.0016 and ranges between 0.0005 and 0.0033, according to the system, the number of shards, and model considered. Such distribution is used to compute the p-value associated with the null hypothesis that the system with greater effect is not statistically significantly better then the other (one-tail hypothesis). Once a p-value for each pairwise comparison is available, Voorhees et al. propose to apply Benjamini-Hochberg correction procedure to correct for multiple comparisons. Finally, using the information on the number of significant differences found, Voorhees et al. propose a strategy to compute an interval of confidence around the system effect, by trimming the vector of the bootstrap sampled estimations of the system effects. In particular, the proportion of samples removed from each side is $\alpha\frac{k}{2N}$, where N is the total number of pairwise comparisons between systems and k is the number of pairs of systems for which one of the two system has statistically larger effect size, according to the Benjamini-Hochberg procedure.

3.3 Experimental Setup

Akin Voorhees et al., we used two collections: the TREC-3 Adhoc track [13] and TREC-8 Adhoc track [26]. TREC-3 contains 50 topics and 40 runs for a total of 820 pairwise run comparisons. TREC-8 consists of 50 topics and 129 runs for a total of 8,256 pairwise run comparisons.

We conducted all the experiments on both collections and we observed very similar behaviours. However, due to space constraints, the replicability results in Section 4 are reported on TREC-3, since Voorhees et al. provide more details on this collection; the generalizability results in Section 5 are reported on TREC-8, since it contains more runs. Note that the replicability experiments concern only b`ANOVA` by Voorhees et al. and not also t`ANOVA` by Ferro and Sanderson, since the latter is our own code. We use Average Precision (AP) and Precision (P) with the cutoff at 10 documents (P@10) as performance measure. The document corpus has been split in $2, 3, 5, 10$ even-sized random shards and we repeated the sampling 5 times. For replicability in Section 4, we repeated the sampling until all the shards contain at least one relevant document for each topic; for generalizability in Section 5, if a shard does not contain any relevant document for a topic, we interpolate the missing value using 4 possible strategies: `zero`; `lq`, the value of the lower quartile of the measure scores; `mean`, the average value of the measure scores; and, `one`. Note that, for generalizability in Section 5, due to space constraints, we report only the case of 5-shards, being the others very similar. To ease the reproducibility of our experiments, the source code is publicly available at `https://github.com/guglielmof/replicate_URIIRE`.

4 Replicability of b`ANOVA`

We tried to replicate the widths of the confidence intervals of the system effect and the number of s.s.d. pairs, i.e. systems for which one is significantly better than the other. Table 1 reports the results of our replicability analysis. Confidence intervals are much smaller, approximately halved, than those reported

Table 1. Confidence interval widths on systems effects and number of s.s.d. system pairs using one-tailed bANOVA on TREC-3. Between parentheses, values originally reported by Voorhees et al.; dashed values were not reported in the original paper.

		no interactions (MD1)				**interactions** (MD2)			
sample	**measure**	**mean**	**min**	**max**	**s.s.d.**	**mean**	**min**	**max**	**s.s.d.**
2 shards	AP	0.045	0.044	0.045	683.80	0.016	0.016	0.017	749.00
		(0.075)	(0.071)	(0.082)	(—)	(0.029)	(0.026)	(0.031)	(743)
	P@10	0.078	0.076	0.080	666.00	0.038	0.037	0.039	728.00
		(0.130)	(0.122)	(0.140)	(—)	(0.065)	(0.061)	(0.069)	(712)
3 shards	AP	0.038	0.037	0.039	699.40	0.018	0.018	0.019	746.20
		(0.064)	(0.060)	(0.069)	(—)	(0.032)	(0.030)	(0.034)	(741)
	P@10	0.062	0.061	0.063	682.20	0.037	0.036	0.037	727.00
		(0.106)	(0.099)	(0.112)	(—)	(0.065)	(0.061)	(0.071)	(712)
5 shards	AP	0.033	0.032	0.033	714.40	0.020	0.020	0.021	742.20
		(0.055)	(0.052)	(0.058)	(—)	(0.033)	(0.031)	(0.034)	(—)
	P@10	0.046	0.045	0.047	697.00	0.031	0.030	0.032	723.00
		(0.081)	(0.076)	(0.086)	(—)	(0.055)	(0.052)	(0.060)	(—)

Table 2. Confidence intervals width on systems effects and number of s.s.d. system pairs using two-tailed bANOVA on TREC-3.

		no interactions (MD1)				**interactions** (MD2)			
sample	**measure**	**mean**	**min**	**max**	**s.s.d.**	**mean**	**min**	**max**	**s.s.d.**
2 shards	AP	0.045	0.044	0.046	661.40	0.016	0.016	0.017	743.20
	P@10	0.078	0.076	0.080	639.60	0.038	0.037	0.039	717.40
3 shards	AP	0.038	0.038	0.039	678.80	0.019	0.018	0.019	739.60
	P@10	0.062	0.061	0.064	662.40	0.037	0.036	0.038	717.80
5 shards	AP	0.033	0.032	0.034	696.00	0.020	0.020	0.021	734.80
	P@10	0.047	0.046	0.048	677.60	0.031	0.030	0.032	712.00

in the original paper. On the other hand, the number of s.s.d. pairs is slightly higher for both AP and P@10; however, this could be still considered within the bounds of the variability due to the random sharding, observed also by Voorhees et al.. To further investigate the interval size, we hypothesized that, even if the original paper describes a single-tailed test, its implementation might have used a more-strict two-tailed one, which is often the default in many statistical software libraries. Table 2 shows the results when using such a two-tailed test. We can note that the confidence intervals are still very similar to the case of Table 1 and, thus, the difference between one-tailed and two-tailed test is not the cause of the observed discrepancy. On the other hand, the number of s.s.d. pairs is getting even closer to those of Voorhees et al.; a little bit less close in the case of P@10 but, as also observed by Voorhees et al., it is a less stable measure. To understand the issue with confidence interval sizes, we modified how they are computed. Instead of removing a percentage of the total number of samples, as described by Voorhees et al., we treated that number as an integer value, representing the actual number of samples to discard. Basically, this milder cut-off

Table 3. Mean, Min and Max modified confidence intervals widths of systems effects on TREC-3, using 3 shards. Highlighted values are the closest to the original ones by Voorhees et al. (* for AP and ‡ for P@10).

		no interactions (MD1)			**interactions**(MD2)		
sample	**measure**	**mean**	**min**	**max**	**mean**	**min**	**max**
original	AP	0.064	0.060	0.069	0.032	0.030	0.034
	P@10	0.106	0.099	0.112	0.065	0.061	0.071
1	AP	0.065^*	0.061	0.071	0.033	0.030^*	0.035
	P@10	$0.106^\ddagger$	0.100	0.113	0.063	0.058	$0.069^\ddagger$
2	AP	0.065^*	0.061	0.072	0.032^*	0.030^*	0.034^*
	P@10	0.105	$0.099^\ddagger$	$0.112^\ddagger$	0.063	$0.060^\ddagger$	0.068
3	AP	0.068	0.065	0.073	0.037	0.034	0.041
	P@10	0.107	0.101	0.113	$0.066^\ddagger$	0.062	0.074
4	AP	0.065^*	0.060^*	0.070	0.030	0.028	0.033
	P@10	0.105	0.098	$0.112^\ddagger$	0.061	0.057	0.064
5	AP	0.065^*	0.059	0.069^*	0.030	0.026	0.032
	P@10	0.105	$0.099^\ddagger$	0.114	0.063	0.059	0.068
avg	AP	0.066	0.061	0.071	0.032	0.030	0.035
	P@10	0.106	0.099	0.113	0.063	0.059	0.069

allows for removing just the most extreme values. Table 3 reports the result for such modification and we can now see that these modified confidence intervals are closer to those of Voorhees et al.. To double-check the confidence intervals, we also tried the vice-versa, i.e. we used the intervals reported in Voorhees et al. to determine the number of s.s.d. pairs. Note that Voorhees et al. use the BH correction to determine the s.s.d. pairs and not the confidence intervals; in their case, they estimate confidence intervals in such a way that they should be consistent with the number of s.s.d. pairs obtained by the BH correction. Since we do not have the sizes of the original intervals, we use, for all the systems, in turn, the mean, minimum, and maximum interval widths reported by Voorhees et al.. Table 4 reports the results of such analysis. The number of s.s.d. pairs is still lower compared to the expected one, in the range of 30 to 70 less, on average (cf. Tab. 2). This suggests that the original intervals are still a bit large to obtain the reported number of s.s.d. pairs; this might be due to the intrinsic accuracy of the estimation procedure or to some differences in the implementation, as we hypothesized in Table 3. Overall, we can conclude that it is possible to fully replicate the bANOVA with BH correction and the resulting number of s.s.d. system pairs which, to us, is the core contribution of the paper and what is used in actual analyses. On the other hand, we were not able to replicate the derived estimation of the confidence intervals and remains an open issue.

5 Generalizability of tANOVA and bANOVA

5.1 Impact of the multiple comparison strategies and bootstrapping

To investigate the differences between ANOVA approaches, our first analysis compares the number of s.s.d. system pairs found by them. We consider the

Table 4. s.s.d. system pairs as obtained by using the confidence intervals widths reported by Voorhees et al.. Compare them with the ones reported in Table 1

		no interactions (MD1)			**interactions** (MD2)		
sample	**measure**	**mean**	**min**	**max**	**mean**	**min**	**max**
2 shards	AP	577.20	590.00	563.20	711.00	721.60	706.00
	P@10	544.60	558.20	528.80	670.40	678.80	661.40
3 shards	AP	608.80	622.80	592.00	702.80	708.60	695.00
	P@10	573.80	583.20	562.00	659.80	667.60	638.60
5 shards	AP	638.80	645.60	629.00	697.40	704.80	695.00
	P@10	597.00	608.20	586.40	656.80	663.60	644.00

Table 5. s.s.d. pairs of systems for different ANOVA approaches, using AP.

Model	**Approach**	bANOVA(BH)	tANOVA(BH)	tANOVA(HSD)
MD1	bANOVA(BH)	6866.60 ± 36.965	329.20 ± 22.027	2275.80 ± 39.844
	tANOVA(BH)	-	6537.40 ± 57.107	1946.60 ± 23.190
	tANOVA(HSD)	-	-	4590.80 ± 75.850
MD2	bANOVA(BH)	7231.80 ± 51.085	375.20 ± 17.436	2133.40 ± 70.456
	tANOVA(BH)	-	6856.60 ± 65.859	1758.20 ± 54.580
	tANOVA(HSD)	-	-	5098.40 ± 113.429
MD3	bANOVA(BH)	7563.40 ± 15.273	262.00 ± 11.681	1655.80 ± 25.377
	tANOVA(BH)	-	7301.40 ± 11.734	1393.80 ± 32.585
	tANOVA(HSD)	-	-	5907.60 ± 37.359

following multiple comparison procedures: HSD for tANOVA, as originally proposed by Ferro and Sanderson, indicated with tANOVA(HSD); BH for bANOVA, as originally proposed by Voorhees et al., indicated with bANOVA(BH); and, BH for tANOVA, indicated with tANOVA(BH). tANOVA with Benjamini-Hochberg correction is here employed and analyzed for the first time, representing a generalizability aspect. It takes the p-values on the difference between levels of the factors produced by the traditional ANOVA, but corrects them using the BH correction. The rationale behind it is that it enjoys the statistical properties provided by the ANOVA while granting a higher discriminative powerf, due to the BH correction procedure. Finally, in this specific setting, such correction procedure allows us to investigate whether the differences between the bANOVA and tANOVA are due to the different ANOVA computation (bootstrap vs direct computation of F-statistics), or are due to the correction procedure applied (BH vs HSD) correction. zero has been used as interpolation strategy; in Section 5.3 we empirically show that the interpolation strategy has a negligible effect on the results. Finally, we experiment all the models from (MD1) to (MD3) with all the ANOVA approaches; note that (MD3) has not been studied before for bANOVA and this represents another generalizability aspect.

Table 5 reports the results averaged over the five samples of shards together with their confidence interval. Numbers on the diagonal of Table 5 describe how many pairs of systems are considered s.s.d. by a given approach; numbers above the diagonal are the additional s.s.d. pairs found by one method with respect to the other. Table 5 shows that, as the complexity of the model increases from (MD1) to (MD3), the pairs of systems deemed significantly different increase as well, confirming previous findings in the literature. tANOVA(HSD) controls tANOVA(BH) since all the s.s.d. pairs for tANOVA(HSD) are significant also for tANOVA(BH); this was expected since FWER controls FDR [14]. It is possible see this by considering the differences between approaches (above diagonal): by summing the difference between tANOVA(HSD) and tANOVA(BH) to the tANOVA(HSD) you obtain back the number of s.s.d. pairs identified by tANOVA(BH). However, this pattern holds also for bANOVA(BH) and tANOVA(BH), i.e. all the s.s.d. pairs of tANOVA(BH) are s.s.d. pairs for bANOVA(BH) too. While the relation between BH and HSD was expected, this finding sheds some light on the difference between using a traditional or a bootstrapped version of ANOVA. In summary, most of the increase in the s.s.d. pairs is due to the correction procedure rather than the use of bootstrap or not. Since bANOVA is more computationally demanding than tANOVA, due to its iterative nature, its use may be not worth if not when you really need to squeeze out all the possible s.s.d. pairs.

5.2 Effect of the Random Shards on the Stability of the Approaches

To assess the stability of different approaches against random resharding, we fix the number of shards (5 in the following analysis).We resampled the shards 5 times and we considered all the possible pairs of shard samples – i.e. 10 possible pairs of shards. To assess the stability with respect to random resharding, we consider the following counting measures proposed in [18]:

- Active Agreements (AA), i.e. the number of pairs of systems A and B for which an approach considers A to be significantly better than B on both samples of shards;
- Active Disagreements (AD), i.e. the number of pairs of systems A and B for which an approach considers A to be significantly better than B on a sample but B is significantly better than A on the other sample;
- Passive Agreements (PA), i.e. the number of pairs of systems A and B for which an approach considers A to not be significantly better than B on both samples of shards;
- Passive Disagreements (PD), i.e. the number of pairs of systems A and B for which an approach considers A to be significantly better than B on a sample but A is not significantly better than B on the other sample.

We did not find any occurrence of AD in any of our experiments, which would indicate a dependency of an approach on a specific random shard, raising some concerns about its stability . AA, PA, and PD are aggregated as follows:

- The Proportion of Active Agreements (PAA), given by $PAA = 2AA/(2AA+PD)$, represents how many times an approach agrees on two systems being s.s.d. concerning the total number of times two systems are claimed s.s.d.;

Table 6. Average PAA and PPA.

Model	**Approach**	**Average PAA**	**Average PPA**
MD1	bANOVA(BH)	0.979 ± 0.001	0.903 ± 0.005
	tANOVA(BH)	0.980 ± 0.001	0.924 ± 0.004
	tANOVA(HSD)	0.979 ± 0.002	0.973 ± 0.003
MD2	bANOVA(BH)	0.980 ± 0.001	0.866 ± 0.007
	tANOVA(BH)	0.979 ± 0.001	0.896 ± 0.006
	tANOVA(HSD)	0.977 ± 0.002	0.963 ± 0.004
MD3	bANOVA(BH)	0.982 ± 0.001	0.802 ± 0.012
	tANOVA(BH)	0.980 ± 0.001	0.850 ± 0.006
	tANOVA(HSD)	0.981 ± 0.001	0.953 ± 0.003

– The Proportion of Passive Agreements (PPA), given by $PPA = 2PA/(2PA+PD)$, shows how often an approach agrees on two systems not being s.s.d. compared to the total number of times two systems are not claimed s.s.d..

PAA and PPA indicate, respectively, the stability of the decisions about which systems are and are not s.s.d., independently from the shard samples. Overall, these two proportions indicate how much you would not change your mind when changing the random shard sample at hand.

Table 6 shows the PAA and PPA averaged over every possible pair of shards together with their confidence intervals. All the approaches have a very high PAA, suggesting that the conclusion about which systems are to be considered s.s.d. is quite stable. The PAA is also very close for all the approaches, slightly increasing as we adopt the more sophisticated (MD3) model but without notable differences between bootstrap and traditional ANOVA or between HSD and BH correction. On the other hand, tANOVA approaches lead to higher PPA than bANOVA ones. The HSD correction produces notably higher PPA than the BH one. We hypothesize that the additional s.s.d. pairs brought in by bootstrap and BH are "corner cases" and the decision about them depends more on the actual shards at hand. We can also observe as the PPA tends to decrease as the models get more sophisticated from (MD1) to (MD3); also, in this case, a more complex model can identify more s.s.d. pairs, but some of them are "corner" cases subject to change from a random shard to another. Overall, the findings concerning PAA and PPA suggest that tANOVA with HSD correction is the most stable approach against different random shards. It should therefore be used when the goal is not the absolute number of s.s.d. pairs, but the accuracy of the decisions.

5.3 Stability of ANOVA Models with respect to Different Interpolation Values

We study the impact of the interpolation strategy, i.e. how to substitute missing values for topics without any relevant document on a given shard, for the different approaches. Here, for space reasons, we report only the results for tANOVA(HSD)

Table 7. Average number of PD for ANOVA model MD2.

(MD2)		**5 Shards**			
Approach	Interp.	zero	lq	mean	one
tANOVA(HSD)	zero	230.60± 21.55	23.00± 15.21	100.20± 74.45	89.80± 82.47
	lq	—	239.20± 22.56	77.20± 62.86	85.60± 96.98
	mean	—	—	253.20± 32.18	124.40± 92.81
	one	—	—	—	265.80± 53.21
bANOVA(BH)	zero	282.60± 13.70	5.80 ± 3.45	41.60 ± 24.44	33.20 ± 28.83
	lq	—	280.80± 12.99	35.80 ± 21.12	32.60 ± 30.75
	mean	—	—	285.00 ± 13.24	49.20 ± 40.73
	one	—	—	—	288.40 ± 18.59

and bANOVA(BH), being the tANOVA(BH) midway between these two.

Ferro and Sanderson [11] mathematically proved that model (MD3) is independent of the adopted interpolation values while Voorhees et al. [27] did not experiment with interpolation values and did not consider this model at all.

Tables 7 and 8 report the average PD counts together with their confidence interval (remember that AD turned out to be zero in our experiments), respectively for models MD2 and MD3. Values on the diagonal are the average PD observed using the same interpolation strategy, but over the pairs of shards samples. The upper triangle of the Table contains the average PD when using two different interpolation values. The PD counts on the diagonal are consistent with the findings of Table 6 in terms of PPA, confirming that bANOVA(BH) is more sensitive to the random sampling of shards than tANOVA(HSD). Table 7 shows what happens if, using model (MD2) by Voorhees et al., instead of re-sampling shards we use an interpolation value. We can note that the PD count on the diagonal, compared to the one of Table 8, slightly increases for both bANOVA(BH) and tANOVA(HSD). On the other hand, the values are in the same confidence interval, and thus are not significantly different.We can also note that, as the interpolation value increases, the PD count on the diagonal tends to increase too. When it comes to the upper triangles, we interestingly find that bANOVA(BH) is much less sensitive to the interpolation values than tANOVA(HSD), being the PD counts substantially lower. Thus, Voorhees et al. could have used an interpolation value instead of re-sampling, without drastically changing the conclusions. The bootstrapped version of ANOVA (bANOVA) appears to be less stable with respect to the resharding. This phenomenon is likely due to its greater discriminative power: since a small evidence for bANOVA is enough to assess when two systems are different, the random resharding might produce spurious evidence and thus large variation among different samples. In Table 8, as expected from [11], the upper triangle for tANOVA(HSD) is zero, since tANOVA(HSD) with (MD3) is independent from the interpolation values. The most interesting finding is that also bANOVA(BH) with (MD3) is independent of the interpolation values. Indeed, the bANOVA approach samples the residuals and Ferro and Sanderson proved that they are independent of the interpolation value for (MD3). Therefore, using (MD3) also the bootstrap approach by Voorhees et al. does not need to re-sample shards.

Table 8. Average number of PD for ANOVA model MD3.

(MD3)		**5 Shards**			
Approach	Interp.	`zero`	`lq`	`mean`	`one`
`tANOVA`(HSD)	`zero`	222.60± 15.392	0.00± 0.000	0.00± 0.000	0.00± 0.000
	`lq`	—	222.60± 15.392	0.00± 0.000	0.00± 0.000
	`mean`	—	—	222.60± 15.392	0.00± 0.000
	`one`	—	—	—	222.60± 15.392
`bANOVA`(BH)	`zero`	279.20± 16.60	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	`lq`	-	279.20± 16.60	0.00 ± 0.00	0.00 ± 0.00
	`mean`	-	-	279.20± 16.60	0.00 ± 0.00
	`one`	-	-	-	279.20± 16.60

6 Conclusions and Future Work

The aim of this paper is multi-folded: we wanted to replicate results by Voorhees et al., generalize the proposed method and compare it with other ANOVA approaches. We were able to replicate the number of s.s.d. found by b`ANOVA`, i.e. the main contribution of the paper, but not the size of the confidence interval. Furthermore, we compared the t`ANOVA` and b`ANOVA` approaches under different conditions. We found out that t`ANOVA` tends to be more robust than b`ANOVA` with respect to the actual random shards used, suggesting more reliability in drawing the same conclusions. On the other hand, when using partial ANOVA models like (MD2) which are not able to deal with shards without relevant documents, b`ANOVA` is more robust than t`ANOVA` to the chosen interpolation value. Regarding the multiple comparison strategy, we have found that t`ANOVA` with HSD is more restrictive than b`ANOVA` but t`ANOVA` with BH correction behaves similarly to b`ANOVA`. Overall, we can conclude that, the decision of the model and the correction technique depends on the final aim of the researcher. If you prioritize the stability of the results over the number of s.s.d. pairs found and you plan to use a full model like (MD3), it is preferable to use t`ANOVA`(HSD), since it is more stable with respect to random shards and less computationally expensive. If instead, your focus is on the number of pairs, b`ANOVA`(BH) gives you the maximum boost but at the price of less stability for random shards. If you plan to use a partial model, like (MD2), which is less expensive from the computational point of view, b`ANOVA`(BH) frees you more from the dependency on topics without relevant documents on some shards. Future work will investigate the use of uneven-size random shards, instead of the even-size ones used in the literature so far.

Acknowledgments. We thank the reviewers for their comments.
The work is partially funded by the DAta BenchmarK for Keyword-based Access and Retrieval (DAKKAR) Starting Grants project sponsored by University of Padua and Fondazione Cassa di Risparmio di Padova e di Rovigo.

References

[1] Banks, D., Over, P., Zhang, N.F.: Blind Men and Elephants: Six Approaches to TREC data. Information Retrieval **1**(1-2), 7–34 (1999)

[2] Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Royal Stat. Soc. **57**(1), 289–300 (1995)

[3] Bodoff, D., Li, P.: Test theory for assessing ir test collections. In: Proc. SIGIR, pp. 367–374 (2007)

[4] Carterette, B.A.: Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. ACM Trans. Inf. Syst **30**(1), 4:1–4:34 (2012)

[5] Clancy, R., Ferro, N., Hauff, C., Sakai, T., Wu, Z.Z.: The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). In: Proc. SIGIR, pp. 1432–1434 (2019)

[6] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall/CRC, USA (1994)

[7] Ferrari Dacrema, M., Boglio, S., Cremonesi, P., Jannach, D.: A troubling analysis of reproducibility and progress in recommender systems research. User Modeling and User-Adapted Interaction (2019)

[8] Ferro, N., Kim, Y., Sanderson, M.: Using Collection Shards to Study Retrieval Performance Effect Sizes. ACM Trans. Inf. Syst **37**(3), 30:1–30:40 (2019)

[9] Ferro, N., Maistro, M., Sakai, T., Soboroff, I.: Overview of centre@ clef 2018: a first tale in the systematic reproducibility realm. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 239–246, Springer (2018)

[10] Ferro, N., Sanderson, M.: Sub-corpora Impact on System Effectiveness. In: Proc. SIGIR, pp. 901–904 (2017)

[11] Ferro, N., Sanderson, M.: Improving the Accuracy of System Performance Estimation by Using Shards. In: Proc. SIGIR, pp. 805–814 (2019)

[12] Fuhr, N.: Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. SIGIR Forum **51**(3), 32–41 (2017)

[13] Harman, D.K.: Overview of the Third Text REtrieval Conference (TREC-3). In: Proc. TREC, pp. 1–19 (1994)

[14] Hsu, J.C.: Multiple Comparisons. Theory and methods. Chapman and Hall/CRC, USA (1996)

[15] Hull, D.A.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: Proc. SIGIR, pp. 329–338 (1993)

[16] Lehmann, E.L., Romano, J.P.: Generalizations of the Familywise Error Rate, pp. 719–735. Boston, MA (2012)

[17] Marchesin, S., Purpura, A., Silvello, G.: Focal elements of neural information retrieval models. an outlook through a reproducibility study. Information Processing & Management pp. 102–109 (2019)

[18] Moffat, A., Scholer, F., Thomas, P.: Models and Metrics: IR Evaluation as a User Process. In: Proc. ADCS, pp. 47–54 (2012)

[19] Robertson, S.: On document populations and measures of ir effectiveness. In: Proceedings of the 1st International Conference on the Theory of Information Retrieval (ICTIR'07)', Foundation for Information Society, pp. 9–22 (2007)

[20] Robertson, S.E., Kanoulas, E.: On Per-topic Variance in IR Evaluation. In: Proc. SIGIR, pp. 891–900 (2012)

[21] Rutherford, A.: ANOVA and ANCOVA. A GLM Approach. John Wiley & Sons, New York, USA, 2nd edn. (2011)

[22] Sakai, T.: On Fuhr's Guideline for IR Evaluation. SIGIR Forum **54**(1), p14:1–p14:8 (June 2020)

[23] Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management **33**(44), 495–512 (1997)

[24] Tague-Sutcliffe, J.M., Blustein, J.: A Statistical Analysis of the TREC-3 Data. In: Proc. TREC, pp. 385–398 (1994)

[25] Tukey, J.W.: Comparing Individual Means in the Analysis of Variance. Biometrics **5**(2), 99–114 (1949)

[26] Voorhees, E.M., Harman, D.K.: Overview of the Eigth Text REtrieval Conference (TREC-8). In: Proc. TREC, pp. 1–24 (1999)

[27] Voorhees, E.M., Samarov, D., Soboroff, I.: Using Replicates in Information Retrieval Evaluation. ACM Trans. Inf. Syst **36**(2), 12:1–12:21 (2017)

[28] Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall/CRC, USA (1995)