Improving the Accuracy of System Performance Estimation by Using Shards

Nicola Ferro
ferro@dei.unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Mark Sanderson
mark.sanderson@rmit.edu.au
Computer Science, School of Science, RMIT University
Melbourne, Australia

ABSTRACT

We improve the measurement accuracy of retrieval system performance by better modeling the noise present in test collection scores. Our technique draws its inspiration from two approaches: one, which exploits the variable measurement accuracy of topics; the other, which randomly splits document collections into shards. We describe and theoretically analyze an *ANalysis Of VAriance (ANOVA)* model able to capture the effects of topics, systems, and document shards as well as their interactions. Using multiple TREC collections, we empirically confirm theoretical results in terms of improved estimation accuracy and robustness of found significant differences. The improvements compared to widely used test collection measurement techniques are substantial. We speculate that our technique works because we do *not* assume that the topics of a test collection measure performance equally.

CCS CONCEPTS

• **Information systems** → **Test collections**; **Retrieval effectiveness**;

KEYWORDS

effectiveness model; ANOVA; multiple comparison

ACM Reference Format:
Nicola Ferro and Mark Sanderson. 2019. Improving the Accuracy of System Performance Estimation by Using Shards. In *Proceedings of SIGIR'19*. ACM, New York, NY, USA, 10 pages. https://doi.org/http://dx.doi.org/10.475/123_4

1 INTRODUCTION

Measuring the difference in performance between two *Information Retrieval (IR)* systems using an offline test collection has long been recognized as noisy. Attempts to improve the accuracy of such measurement are extensive and diverse. Techniques explored include multiple evaluation measures; different significance tests; alternate acquisitions of relevance judgments; and determining the ideal number of topics. Surveys [30] and descriptions of best practice [29] detail such attempts. There are, however, less explored approaches to improving performance measurement accuracy.

SIGIR'19, July 21–25, 2019, Paris, France.

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/http://dx.doi.org/10.475/123_4

Robertson and Kanoulas [26] pointed out a common assumption in the use of test collections namely "*all topics are considered equally valuable*". They examined this assumption by measuring (via bootstrapping) the confidence intervals of each topic score and of each system. The intervals were found to be variable across topics but largely independent of system. The researchers concluded that some topics measure performance more accurately than others.

Ferro and Sanderson [11] examined splitting the documents of a test collection into *shards*, measuring the performance of systems on each shard. They used an ANOVA model to understand if system performance changed across shards. The authors mentioned that significant differences between systems on sharded collections were more common than on unsharded. However, the reasons for the result was not explored as the experiment was designed to address a different research question. Voorhees et al. [39] randomly split a collection in half. The authors stated that the two resulting shards allowed more accurate performance measurement. However, it was reported that splitting the collection further did not improve accuracy; reasons for no improvement were not examined in detail.

We describe research that takes the Robertson and Kanoulas view that topics have unequal value and combines it with the ANOVA approach of Ferro and Sanderson [11] and the sharding method of Voorhees et al. [39]. We ask: *Can the unequal value of topics be exploited to improve measurement of system performance accuracy on a test collection?* We make the following contributions:

- We validate an ANOVA model via a theoretical examination, showing why explicitly accounting for differences across topics yields accuracy improvements.
- We experimentally show that the model identifies notable numbers of significant differences between systems.
- We experimentally show that the differences are not due to measurement error of the significance formulas.

Next, related work is described followed by ANOVA models and their properties. The setup and report of experimental findings are described before conclusions and future work are detailed.

2 RELATED WORK

We review three research areas: topics with few relevance judgements, ANOVA modeling, and the sharding of collections.

2.1 Topics with few relevance judgments

There is an assumption, in test collection based evaluation, that all topics are valuable *equally*. Performance is measured by taking the arithmetic mean of topics scores. Swanson [33] described such a process in 1960. When the mean is taken, each topic score contributes equally regardless of the accuracy of that measure. The potential for error was described by Voorhees [36]: "*When [topics] have very few relevant documents (fewer than five or so), summary evaluation measures such as average precision are themselves unstable; tests that include many such queries are more variable*". Soboroff [32] pointed out that rank cut off evaluation measures (e.g. precision at 10) will have an upper bound < 1 for topics with few relevant documents.

The notion that not all topics have equal value was implicitly exploited in work identifying a subset of test collection topics that rank systems similarly to a full topic set [13, 19]. To the best of our knowledge, however, Cormack and Lynam [9] were the first to incorporate an unequal view of topics into test collection measurement. They treated each topic as a "*separate test*", calculating topic confidence intervals using a bootstrap approach. Topics with ≤ 5 relevant documents were subject to a "*Small-R Correction*" to overcome measurement instability.

Robertson [25] considered the broader question of what is the "*per-topic noise or error*" present in the topics of a test collection. The paper considered if evaluation measures could be adapted to cope with an unequal view of topics. Later, Robertson and Kanoulas [26] measured the variance of topic scores by bootstrapping from the document collection. The researchers found that topics showed different levels of variance, but the variance was relatively consistent across systems. The researchers described a significance test that incorporated topic score variation. Comparisons between the new test and the commonly used t-test showed some differences in the conclusions one might draw when comparing systems.

More recently, Yang et al. [41] examined how much rankings of systems were affected by per-topic score variance and if there was any impact on significance tests. They found that the variance did not affect overall rankings notably, but that the number of significant differences observed between systems dropped.

Note, there is much research on subjects such as query difficulty prediction [42], topic score normalization [40], average average precision [20], GMAP [24], etc. Such work focuses on so-called difficult topics, we focus on topics for which measurement is variable.

2.2 ANOVA modeling

ANOVA can decompose the data of an IR experiment into a model of factors, into interactions between those factors, and into a level of unmodeled error. Tague-Sutcliffe and Blustein [34] described an example of this approach by comparing the variation in performance across two factors: topics and systems. The former was found to be larger than the latter. Measurement of interaction between topics and systems was not possible owing to a lack of *replicates* of topic*system measures. Banks et al. [2] approximated such an interaction, suggesting it would be strong and significant. Later, Bodoff and Li [3] used a test collection with multiple relevance assessments to obtain the required replicates. The authors reported that the magnitude of the topic*system interaction factor was less than the topic factor, but greater than the system factor.

Both Ferro and Sanderson [11] and Voorhees et al. [39] generated replicates by sharding a document collection. This enabled them to measure the topic*system effect. We describe that work next.

2.3 Sharding

Voorhees et al. [39] used a bootstrap ANOVA approach that drew on a sample of the scores of topics measured across different systems and shards. The researchers tested on the TREC-3, TREC-8, and 2006 Terabyte track collections. Success of the approach was measured by counting the number of significant differences found between systems submitted to TREC tracks. The researchers found substantially more such differences were measured than with conventional approaches. Two shards were used. When three or five shards were tried, the researchers found the number of significant differences dropped, the reasons for which were not examined in detail. The relative impact of each component of the technique – bootstrap ANOVA, the approach to multiple comparisons, and sharding method – was not described.

As part of a study on the interaction between different types of shards and system scores, Ferro and Sanderson [11] described a series of ANOVA models tested on the TREC-7 and TREC-8 adhoc test collections. Like the previous research, the value of these models was quantified by the number of significant differences measured between systems. The researchers showed that a more sophisticated ANOVA model produced the highest number of significant differences measured between systems. However, the shards were very skewed in size.

The research described shows that the topics of test collections can produce scores of different variance, which can impact the measurement of significance between systems. There is, as yet, not an extensive body of research examining such topic variability. Most work has explored bootstrap approaches from document collections to assess the variance. The recent examination of sharding has not been explored in conjunction with the work on topic variability. We explore the connection between these two lines of inquiry examining the style of ANOVA modeling used by Ferro and Sanderson [11]. We also measure the accuracy of the model across a range of sharding configurations that have not been examined before.

3 METHODOLOGY

Suppose we have T topics, R systems, and S shards and thus $N = T \cdot R \cdot S$ total samples. We can form the following six ANOVA models:

$$y_{ij} = \mu_{\cdot\cdot} + \underbrace{\tau_i + \alpha_j}_{\text{Main Effects}} + \varepsilon_{ij} \tag{MD1}$$

$$y_{ijk} = \mu_{\cdots} + \underbrace{\tau_i + \alpha_j}_{\text{Main Effects}} + \varepsilon_{ijk} \tag{MD2}$$

$$y_{ijk} = \mu_{\cdots} + \underbrace{\tau_i + \alpha_j}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD3}$$

$$y_{ijk} = \mu_{\cdots} + \underbrace{\tau_i + \alpha_j + \beta_k}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD4}$$

$$y_{ijk} = \mu_{\cdots} + \underbrace{\tau_i + \alpha_j + \beta_k}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij} + (\alpha\beta)_{jk}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD5}$$

$$y_{ijk} = \mu_{\cdots} + \underbrace{\tau_i + \alpha_j + \beta_k}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD6}$$

Where:

- y_{ijk} is the performance score of three factors, the i-th topic ($i = 1, \ldots, T$) retrieving on the j-th system ($j = 1, \ldots, R$) from the k-th shard ($k = 1, \ldots, S$);
- $\mu_{\cdots}$ is the grand mean;
- $\tau_i = \mu_{i\cdot\cdot} - \mu_{\cdots}$ is the effect of the i-th topic, where $\mu_{i\cdot\cdot}$ is the marginal mean of the topic;
- $\alpha_j = \mu_{\cdot j\cdot} - \mu_{\cdots}$ is the effect of the j-th system, where $\mu_{\cdot j\cdot}$ is the marginal mean of the system;
- $\beta_k = \mu_{\cdot\cdot k} - \mu_{\cdots}$ is the effect of the k-th shard, where $\mu_{\cdot\cdot k}$ is the marginal mean of the shard;
- $(\tau\alpha)_{ij} = \mu_{ij\cdot} - \mu_{i\cdot\cdot} - \mu_{\cdot j\cdot} + \mu_{\cdots}$ is the interaction between topics and systems, where $\mu_{ij\cdot}$ is the marginal mean of the interaction between the i-th topic and j-th system;
- $(\tau\beta)_{ik} = \mu_{i\cdot k} - \mu_{i\cdot\cdot} - \mu_{\cdot\cdot k} + \mu_{\cdots}$ is the interaction between topics and shards, where $\mu_{i\cdot k}$ is the marginal mean of the interaction between the i-th topic and k-th shard;
- $(\alpha\beta)_{jk} = \mu_{\cdot jk} - \mu_{\cdot j\cdot} - \mu_{\cdot\cdot k} + \mu_{\cdots}$ is the interaction between systems and shards, where $\mu_{\cdot jk}$ is the marginal mean of the interaction between the j-th system and k-th shard; and
- ε_{ijk} is the error of the model in predicting y_{ijk}.

Model (MD1) was used by Tague-Sutcliffe and Blustein [34] and Banks et al. [2]. It can be viewed as a classic approach to measuring significance on a test collection, as in this form, it is operationally similar to a t-test. The model components have two subscripts (i, j) because the collection does not have shards. Model (MD2) is model (MD1) but with shards. While model (MD1) has only one performance score for each (topic, system) pair, in model (MD2) the shards provide replicates scores for the pairs when estimating the model parameters.

The presence of replicates is exploited in model (MD3) by adding a topic*system interaction factor. Model (MD3) was used by Robertson and Kanoulas [26] and Voorhees et al. [39], though Voorhees et al. did not rely on classical ANOVA, instead adopting a bootstrap approach [10]. Model (MD4) explicitly accounts for a shard factor and model (MD5) adds the interaction between systems and shards. Both models are close to models proposed by Ferro and Sanderson [11], but they omitted the topic*system interaction in their models.

Model (MD6) adds a topic*shard interaction, by leveraging the presence of more replicates for each (topics, shard) pair – there are as many replicates as the number of used systems R. It is the focus on our work here[1].

3.1 Exploiting topic variability with the model

How does improved measurement accuracy arise from a more sophisticated ANOVA model applied over a test collection whose documents are randomly split into shards? Models add more factors with the goal of better fitting the data. Since the total *Sum of Squares (SS)* is the same for all models, each new factor should explain a further part of the total SS. As a consequence, there is a reduction of the error SS, i.e. the leftover unexplained by the model, and, broadly speaking, this leads to a more accurate estimate.

[1]We have also examined different sharding approaches and how they impact the effect size of ANOVA model factors [12]. That paper does not examine in detail the impact of the model on significance tests.

How does model (MD6) exploit the variable measurement of topics? With random even sized shards, the probability of having relevant documents in a shard is uniform across the shards. This probability is smaller for topics with fewer relevant documents and greater for topics with more relevant documents. Therefore, for each topic, the number of shards without any relevant documents is proportional to the number of relevant documents for that topic. Model (MD6) accounts for this by explicitly considering $(\tau\beta)_{ik}$, i.e. the topic*shard interaction effect. When there are no relevant documents for a topic on a given shard, we set the score to `undefined` for all the systems with respect to that topic on that shard. The more shards without relevant documents for a topic, the more `undefined` values there are, which is reflected in the estimation of the $(\tau\beta)_{ik}$ factor. Therefore, the estimation of the SS of the topic*shard interaction factor directly removes from the total SS the variability due to these intrinsic differences among topics, reducing the error SS and giving us the possibility of a more accurate estimation of the differences among systems. Instead of seeing shards as a mere "technical trick" to obtain replicates, we can look at them as a form of "diagnostic tool", which allows us to systematically probe measurement differences across topics and to account for the differences in a model.

We next consider a series of questions about model (MD6):

- How do different models affect the significant differences among systems, accounting for multiple comparisons?
- How do we compute confidence intervals from the model?
- How do we estimate effect size?
- Is it legitimate to use `undefined` values?

The following sections will answer the questions by showing that model (MD6) provides benefits in all these areas and, most importantly, makes estimations concerning the system factor independent of `undefined` values due to the sharding process.

4 MULTIPLE COMPARISONS

If one simultaneously compares multiple system pairs, the probability of committing a *Type I* error increases and the *Family-wise Error Rate (FWER)* (the probability of committing at least one Type I error) is $\text{FWER} = 1 - (1 - \alpha)^c$, where c is the total number of comparisons to be performed [15, pp. 7–8]. It is crucial to control Type I errors when performing multiple comparisons [4, 6, 29].

Tukey [35] proposed the *Honestly Significant Difference (HSD)* test, which creates confidence intervals for all pairwise differences between factor levels, while controlling the FWER. Two systems u and v are considered significantly different when:

$$|tk| = \frac{|\hat{\mu}_{\cdot u\cdot} - \hat{\mu}_{\cdot v\cdot}|}{\sqrt{\frac{MS_{error}}{T \cdot S}}} > Q^{\alpha}_{R,df_{error}} \qquad (1)$$

where: $\hat{\mu}_{\cdot u\cdot}$ and $\hat{\mu}_{\cdot v\cdot}$ are the marginal means of the systems u and v as estimated from the actual data; df_{error} are the *Degrees of Freedom (DF)* of the error; MS_{error} is the *Mean Squares (MS)* of the error, i.e. an estimation of the variance left unexplained; and $Q^{\alpha}_{R,df_{error}}$ is the upper $100 * (1 - \alpha)$-th percentile of the studentized range distribution [22]. Note, that in the case of the model (MD1) the denominator of eq. (1) becomes just T, since the whole corpus is constituted by a single shard and thus $S = 1$.

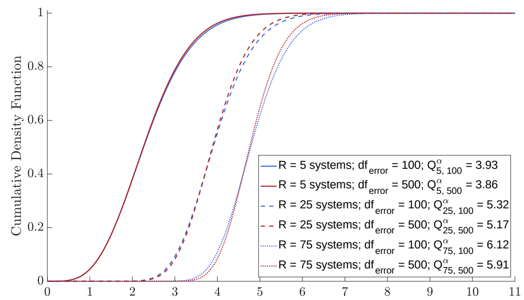


Figure 1: Studentized range distribution $Q_{R,df_{error}}$ for different numbers of systems to be compared and different degrees of freedom of the error. The lines in each plot corresponds to different DF of the error $Q^{\alpha}_{R,df_{error}}$ for $\alpha = 0.05$: red lines are for 100 DF, blu lines are for 500 DF. Solid lines are for $R = 5$ systems; dashed lines are for $R = 25$ systems; and, dotted lines are for $R = 75$ systems.

Figure 1 shows the *Cumulative Density Function (CDF)* of the Studentized range distribution for different numbers of compared systems and different values of the DF of the error. The DF lines are almost superimposed on each other. The values of $Q^{\alpha}_{R,df_{error}}$ are equal, apart from the lower values of DF where they are marginally different. The main difference across plots is that increasing the number of systems to be compared shifts the CDF to the right. In a typical IR setting where R systems are compared, the factor $Q^{\alpha}_{R,df_{error}}$ in eq. (1) is practically constant. As a consequence, even if models from (MD1) to (MD6) lead to different values df_{error}, the models "see" the same value of $Q^{\alpha}_{R,df_{error}}$ and, therefore, the size of the interval needed to consider two systems as significantly different mostly depends on the factor $\sqrt{\frac{MS_{error}}{T \cdot S}}$.

In models (MD2) to (MD6), the marginal means $\hat{\mu}_{\cdot u \cdot}$ and $\hat{\mu}_{\cdot v \cdot}$ of the compared systems are the same as well as the $T \cdot S$ factor; therefore, differences in the size of the intervals are due only to the $\sqrt{MS_{error}}$ factor. Since the typical benefit of having richer models is to reduce the size of the error, we expect MS_{error} to decrease[2] and, consequently, the test statistic $|tk|$ increases, allowing us to detect more significant differences. The increasingly richer models lead to a more accurate estimate of the actual differences among systems. Moreover, the MS_{error} is further divided by $T \cdot S$, which suggests that, for a given number of topics T, increasing the number of shards S should provide further benefits.

The test statistic $|tk|$ allows us to compute the p-value

$$p = \mathbb{P}\left[Q_{R,df_{error}} \geq |tk|\right] \tag{2}$$

of observing a more extreme value of the Studentized range distribution. We can then compare this p-value to the desired significance level α and, if it is $\leq \alpha$, the two systems u and v are significantly different, still controlling the FWER. Eqs. (1) and (2) are two equivalent ways to perform multiple comparisons controlling the FWER.

[2]Strictly, the SS of the error decreases because the additional factors in a model explain more of the total SS, leaving less to the SS of the error. However, $MS_{error} = \frac{SS_{error}}{df_{error}}$, if a richer model causes a drop in df_{error}, this decreased denominator may lead to a greater MS_{error}, even if SS_{error} is decreased. However, as a first approximation, it is enough to consider both quantities as decreasing as we add factors to a model.

5 CONFIDENCE INTERVALS

We consider three types of confidence interval.

5.1 Tukey

The Tukey HSD test of eq. (1) allows us to define exact confidence intervals for the system main effects, still controlling the FWER. Hochberg and Tamhane [15] suggest creating a half-width confidence interval around the marginal mean of a system u

$$\hat{\mu}_{\cdot u \cdot} \pm \frac{1}{2} Q^{\alpha}_{R,df_{error}} \sqrt{\frac{MS_{error}}{T \cdot S}} \tag{3}$$

Systems u and v are significantly different, according to the Tukey HSD test of eq. (1), if and only if their confidence intervals of eq. (3) do not overlap [15, p. 116]. From model (MD2) to (MD6), we expect that confidence intervals will reduce as MS_{error} decreases.

5.2 Standard Error of the Mean

The confidence interval of eq. (3) differs from the typical confidence interval based on the *Standard Error of the Mean (SEM)*:

$$\hat{\mu}_{\cdot u \cdot} \pm t^{\alpha/2}_{T \cdot S-1} \sqrt{\frac{\hat{\sigma}^2_u}{T \cdot S}} \tag{4}$$

where $\hat{\sigma}^2_u = \frac{1}{T \cdot S-1} \sum_{i=1}^{T} \sum_{k=1}^{S} (y_{iuk} - \hat{\mu}_{\cdot u \cdot})^2$ is the sample variance of the u-th system and $t^{\alpha/2}_{T \cdot S-1}$ is the upper $100*(1-\alpha/2)$-th percentile of the Student's t distribution with $T \cdot S - 1$ degrees of freedom. Note, these are the confidence intervals used by Ferro and Sanderson [11] when showing the improved accuracy due to the use of shards.

Differently from the confidence interval of eq. (3), those of eq. (4) do not depend on any of the more accurate ANOVA models, they just depend on the underlying data. Moreover, they do not account for any multiple comparison adjustment since they consider each system in isolation. While the confidence intervals of eq. (3) have the same size for all systems as they need to control for FWER, the confidence intervals of eq. (4) change size from system to system as they depend on the sample variance of each system.

5.3 ANOVA

We can define the following confidence interval [29, p. 57], which falls between those of eq. (3) and those of eq. (4)

$$\hat{\mu}_{\cdot u \cdot} \pm t^{\alpha/2}_{df_{error}} \sqrt{\frac{MS_{error}}{T \cdot S}} \tag{5}$$

As with eq. (3), the interval depends on the ANOVA model and its ability to explain the data. As with eq. (4), the interval does not adjust for multiple comparisons. Different from eq. (4) but similar to eq. (3), the interval has the same size for all systems. As above, the term $t^{\alpha/2}_{df_{error}}$ is practically constant, following the discussion about eq. (3), we expect the confidence interval of eq. (5) to reduce either as the ANOVA models become richer or if we use more shards.

The difference between eq. (3) and eq. (5) is the replacement of $\frac{1}{2} Q^{\alpha}_{R,df_{error}}$ with $t^{\alpha/2}_{df_{error}}$. The former is typically 2-3 times bigger than the latter. The bigger the difference, the bigger the number of systems R to be compared. This lets us understand the magnitude of adjustment needed to keep the FWER controlled. Consequently, the confidence intervals of eq. (3) are bigger than those of eq. (5).

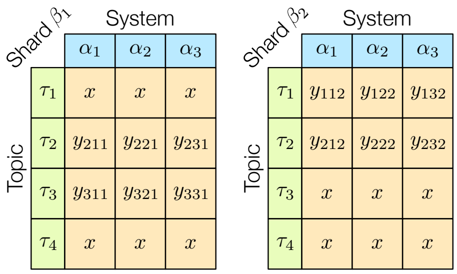


Figure 2: Example of $T = 4$ topics, $R = 3$ systems, $S = 2$ shards

6 EFFECT SIZE

We also consider the *effect size* of a factor, which accounts for the amount of variance explained by the model, by means of an unbiased estimator [23, 28]:

$$\hat{\omega}^2_{\langle fact \rangle} = \frac{\mathrm{df}_{fact}(\mathrm{F}_{fact} - 1)}{\mathrm{df}_{fact}(\mathrm{F}_{fact} - 1) + N} \tag{6}$$

where F_{fact} is the F-statistic and df_{fact} are the degrees of freedom for the factor while N is the total number of samples. The common rule of thumb [27] when classifying $\hat{\omega}^2_{\langle fact \rangle}$ effect size is: 0.14 and above is a *large size effect*; 0.06–0.14 is a *medium size effect*; and 0.01–0.06 is a *small size effect*. Note, $\hat{\omega}^2_{\langle fact \rangle}$ can be negative, in such cases it is considered as zero.

7 EFFECT OF UNDEFINED VALUES

A notable challenge with sharding a document collection is topics may not have any relevant documents in a shard. Ferro and Sanderson [11] dealt with this by keeping only the topics for which there was at least one relevant document in each shard, thus reducing the number of usable topics. Voorhees et al. [39] resampled shards until all shards contained relevant documents for all the topics. However, this introduces bias since the shards stop being random. Both approaches fail as the number of shards increase.

As described above, we substitute an `undefined` value. We demonstrate that we can substitute `undefined` values with any value x and these values do not affect the identification of significant differences, the calculation of confidence intervals, and the effect size of the system factor. We report here the main propositions but, for space reasons, we cannot report the corresponding proofs. Detailed proofs are reported in the electronic appendix available online as supplementary material to the paper.

In the example of Figure 2 we have $T = 4$ topics, $R = 3$ systems and $S = 2$ shards. Topic τ_1 has no relevant documents in shard β_1 and, therefore, all the systems have the `undefined` value x for that topic. Similarly, topic τ_3 has no relevant documents in shard β_2 and topic τ_4 has no relevant in β_1 and β_2. Note, when relevant documents are missing, a whole "row" is filled in with x. This regularity, allows us to achieve a balanced design where the comparison of systems is independent of the `undefined` values.

Definition 7.1. Given a shard $k \in [1, S]$, X_k is the set of the indexes i of the topics that have no relevant documents on that shard:

$$X_k = \left\{ i \in [1, T] \,\middle|\, y_{ijk} = x \ \forall j \in [1, R] \right\} \tag{7}$$

In Figure 2, we have $X_1 = \{1, 4\}$ and $X_2 = \{3, 4\}$. Note that, for any shard k, there are $|X_k| \cdot R$ `undefined` values and, in total, there are $R \sum_{k=1}^{S} |X_k|$ `undefined` values.

Proposition 7.2. *Given models from* (MD2) *to* (MD6) *and a system $j \in [1, R]$, its estimated marginal mean is given by:*

$$\hat{\mu}_{\cdot j \cdot} = \underbrace{\frac{1}{T \cdot S} \sum_{k=1}^{S} \sum_{\substack{i=1 \\ i \notin X_k}}^{T} y_{ijk}}_{\hat{\mu}'_{\cdot j \cdot}} + \frac{x}{T \cdot S} \sum_{k=1}^{S} |X_k| \tag{8}$$

Therefore, for any pair of systems $u \in [1, R]$ and $v \in [1, R]$, $u \neq v$, the difference of their estimated marginal means $\hat{\mu}_{\cdot u \cdot} - \hat{\mu}_{\cdot v \cdot}$ is independent of the `undefined` *values.*

Note, that the first element $\hat{\mu}'_{\cdot j \cdot}$ of eq. (8) is the estimated marginal mean of the system factor ignoring the `undefined` values. This is not the estimated marginal mean removing `undefined` values, since the denominator $T \cdot S$ still accounts for all the values, both defined and `undefined`. The second element is the contribution to estimated marginal mean due only to the `undefined` values. It is constant and equal for all the systems. Therefore, the regularity in the pattern of `undefined` values allows us to separate the contributions due to the systems from those due to `undefined` values, which are the same for all the systems. Proposition 7.2 has three consequences:

1. The numerator of eq. (1), i.e. the multiple comparison among systems, is not affected by the `undefined` values.
2. Eq. (8) shows that the shift due to `undefined` values is the same for all the systems and, therefore, does not affect the *Rankings of Systems (RoS)*, i.e. the ordering of the system by their estimated marginal mean. If a Kendall's τ correlation [17] was measured between the RoS on the whole corpus and the RoS when using shards, τ is not affected by the `undefined` values.
3. For each shard we could have at worst $|X_k| = T$, i.e. a shard for which no topic has relevant documents. However, test collections generally have at least one relevant document for each topic and, since shards are a partition of the whole corpus, it follows that $|X_k| < T$. Therefore $\frac{1}{T \cdot S} \sum_{k=1}^{S} |X_k|$ is always strictly < 1. The effect of the `undefined` value is to shift the estimated marginal mean of the system factor by a fraction of that `undefined` value. From this perspective, setting $x = 0$, our choice in the experimentation, is not lowering the mean system performance but just leaving them at their level.

Proposition 7.3. *Given models from* (MD2) *to* (MD6)*, the SS of the system factor and, as a consequence, the MS of the system factor are independent of the* `undefined` *values.*

Proposition 7.4. *Given model* (MD6)*, the residuals ε_{ijk} are independent from the* `undefined` *values. Therefore, the SS of the error and, as a consequence, the MS of the error are independent of the* `undefined` *values.*

Note that Proposition 7.4 holds only in the case of model (MD6) and only thanks to the topic*shard interaction $(\tau\beta)_{ik}$ factor.

Indeed, as shown in the appendix, all the estimated marginal means have a form similar to eq. (8), i.e. a mean contribution due to defined values plus a mean contribution due to undefined values. However, only the topic*shard interaction $(\tau\beta)_{ik}$ has the form

$$\hat{\mu}_{i \cdot k} = \begin{cases} \hat{\mu}'_{i \cdot k} & \text{if } i \notin X_k \\ x & \text{if } i \in X_k \end{cases}$$

which cancels out the undefined values when $y_{ijk} = x$ and makes the residuals ε_{ijk} independent from them. In this sense, in Section 3.1, we said that the topic*shard interaction $(\tau\beta)_{ik}$ is the factor dealing with the intrinsic differences among topics, since it is able to separate defined from undefined values. As we discussed, the number of undefined values is proportional to the number of relevant documents for a topic and, therefore, the topic*shard interaction $(\tau\beta)_{ik}$ factor accounts for the *unequal value of topics*.

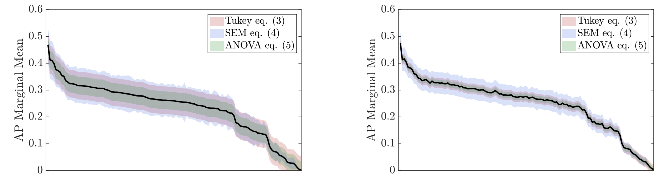
Therefore, model (MD6) is not only a more precise model because, thanks to the more factors it considers, it is able to explain more variance than all the other models, leading to more accurate estimations of the differences among systems. But, especially, it is also the model with the most desirable properties, thanks to the presence of the topic*shard interaction $(\tau\beta)_{ik}$ factor. Indeed, proposition 7.4 has two consequences:

(1) The denominator of eq. (1) is independent of the undefined values. This, jointly with Proposition 7.2, means that undefined values do not affect the identification of significantly different systems. Consequently, the confidence intervals of eq. (3) are independent from the undefined values. The same holds for the confidence intervals of eq. (5).
(2) Recall that the F-statistic of the system factor is given by $F_{system} = \frac{MS_{system}}{MS_{error}}$ where $MS_{system} = \frac{SS_{system}}{df_{system}}$ and $MS_{error} = \frac{SS_{error}}{df_{error}}$. Since both SS_{system} (Proposition 7.3) and SS_{error} (Proposition 7.4) are independent from the undefined values, it follows that the F-statistic of the system factor is also independent from undefined values. They do not affect the significance of this factor. Moreover, it follows that the effect size of the system factor $\hat{\omega}^2_{\langle system \rangle}$ of eq. (6) is independent of the undefined values.

8 EXPERIMENTAL SETUP

To empirically test the analyses above, we experimented on the collections, topics, and system runs of the following datasets:

- **Adhoc track T08** [38]: 528,155 documents of the TIPSTER disks 4-5 corpus minus congressional record (TIP); 50 topics, each with binary relevance judgments drawn from a pool depth of 100; 129 system runs retrieving 1,000 documents for each topic.
- **Web track T09** [14]: 1,692,096 documents of the WT10g Web corpus; 50 topics, each with multi-graded relevance judgments and a pool depth of 100; 104 system runs retrieving 1,000 documents for each topic.
- **Common Core track T27** [1]: 595,037 documents of the Washington Post corpus (WAPO); 50 topics, each with multi-graded relevance judgments; relevance judgments were obtained mixing depth-10 pools with multi-armed bandit [18, 37], stratified sampling [7] and move-to-front [8] approaches; 72 system runs retrieving 10,000 documents for each topic.



(a) Model (MD1). **(b) Model** (MD6).

Figure 3: Comparison of different types of confidence intervals on T08 for AP on the whole corpus (left) and on TIP_RNDE_03 shards (right). On the x-axis there are the systems ordered by descending performance.

We any mapped multi-graded relevance judgments to binary by treating everything above not relevant as relevant.

For each corpus, we created S randomly formed even sized shards, where $S \in \{2, 3, 4, 5, 10, 25, 50\}$. We label the shards of a corpus as <corpus>_RNDE_S; e.g., the WAPO corpus split into 5 shards is labeled WAPO_RNDE_05. For each shard size, we re-sampled 10 times; i.e., in the case of WAPO_RNDE_05 we have 10 independent sets of 5 random even size shards on the WAPO corpus. For space reasons, we report only some combinations of measures and tracks but the observed trends hold also for the other results.

For each corpus split into shards, system runs retrieving from the corpus were also sharded. A run was split into the same number of shards as the corresponding corpus. The random document split used to shard a corpus was the same split used to shard a run. Such splitting is a simulation of how a system would retrieve documents on each shard. Past empirical work showed the simulation to work well Sanderson et al. [31].

We consider the following evaluation measures: *Average Precision (AP)* [5], Precision at ten retrieved documents (P@10), *Rank-Biased Precision (RBP)* [21], and *Normalized Discounted Cumulated Gain (nDCG)* [16]. We calculated RBP by setting $p = 0.8$ as persistence parameter while we use a log_{10} discounting function in nDCG, to consider not too impatient users. We considered $\alpha = 0.05$ to determine if a factor is statistically significant. Our experimental source code is at: https://bitbucket.org/frrncl/sigir2019-fs-code/.

9 EXPERIMENTS

We conduct three experiments.

9.1 Confidence Intervals

We study the three types of confidence intervals under different ANOVA models. Figure 3 compares the intervals on the whole corpus using model (MD1) and on three shards using model (MD6).

In the case of the whole corpus and model (MD1) in Figure 3a, we see that, as expected, the Tukey confidence intervals (eq. (3)) are larger than the ANOVA ones (eq. (5)) since the latter do not account for multiple comparisons. We also see that the Tukey intervals of eq. (3) are similar to the SEM intervals (eq. (4)), which are independent from any model of the data and just consider each system in isolation. The fact that model-dependent confidence intervals

(Tukey ones) look close to model-independent ones (SEM ones) suggests that the topic and system factors of model (MD1) are not enough to accurately explain the data.

When using the shards (Figure 3b), we note that both the Tukey and ANOVA confidence intervals are smaller than SEM, suggesting that model (MD6) better explains the underlying data thanks to the additional factors it considers.

Note, that this difference between model (MD1) and (MD6) is not due to the increased number of samples passing from the whole corpus to shards but to the better ability of model (MD6) to explain the data. Indeed, the additional beneficial effect of increasing the number of samples is apparent in Figure 3b from the fact that all the confidence intervals get smaller when using shards, but this would happen for whatever model.

Figure 4 shows how the Tukey confidence intervals change across different models. The black dotted line is the system performance (marginal mean of the system α_j factor) on the whole corpus, i.e. the same line shown in Figure 3a in the case of AP. The continuous line is the system performance (marginal mean of the system α_j factor) on shards, i.e. the same line shown in Figure 3b in the case of AP; note that the green line for model (MD2), the orange one for model (MD3), and the red one for model (MD6) are superimposed since the marginal mean of the system α_j factor is the same in all these models. The shaded areas in the color of the line of each model represent the Tukey confidence interval for the corresponding model; for example, gray shaded area is for model (MD1) while the red shaded area is for model (MD6).

For all measures, the confidence interval using model (MD1) on the whole corpus is bigger than the confidence interval when using the other models. In particular, comparing the confidence intervals of models (MD1) and (MD2), which are computed without and with shards respectively. Comparing models (MD2), (MD3), and (MD6), we see the increasingly complex models improve the accuracy by shrinking the confidence interval. Moreover, comparing model (MD3) to model (MD6) we see that adding shard*system and topic*shard factors substantially reduce the intervals.

We report the Kendall's τ correlation between the RoS on the whole corpus and on shards in the title of the plots in Figure 4. We can see that in three of the four plots, $\tau > 0.9$, the empirical threshold used to consider to ranking equivalence [36]. This suggests that we are not only improving accuracy but also maintaining coherence with what happens in traditional analyses.

Figure 5 compares the Tukey confidence intervals of eq. (3) for different shard numbers using model (MD6) in the case of AP on `T08`. As expected, the confidence intervals tend to reduce as the number of shards increases, due to the increased number of measurements on the shards. Kendall's τ remains > 0.9, suggesting that the increased number of shards does not substantially deteriorate the agreement of the RoS on the whole corpus.

9.2 Multiple Comparisons

Table 1 reports summary statistics for multiple comparison analyses on `T08` using different splits for AP. We observe a large system effect size ($\hat{\omega}^2_{\langle sys \rangle}$). We also can see a drop in $\hat{\omega}^2_{\langle sys \rangle}$ passing from model (MD1), i.e. the whole corpus, to model (MD2), i.e. the same model but using shards. The shards appear to introduce a new factor, which interacts with the other factors and thus the size of $\hat{\omega}^2_{\langle sys \rangle}$ reduces. However, as the models account for more factors ((MD2)-(MD6)), $\hat{\omega}^2_{\langle sys \rangle}$ increases, suggesting that the more a model explains the data, the more prominent $\hat{\omega}^2_{\langle sys \rangle}$ becomes. In the case of model (MD6) and for fewer shards, $\hat{\omega}^2_{\langle sys \rangle}$ can be notably bigger than on the whole corpus.

Considering the number of significantly different pairs (columns Sig and NotSig), we see how moving from (MD1) – a classic significance testing approach – to any shard-based model always increases the number of pairs. More shards also means more significantly different pairs. However, there is a limited gain in using more shards: in the case of model (MD2) passing from two to five shards gives a 13.28% increase in the number of pairs but passing from five to ten produces only a 0.15% gain. More complex models are less sensitive to the increase in the number of shards, since they detect almost all the significantly different pairs already at a low number of shards. For example, in the case of model (MD6) passing from two to five shards gives just a 0.78% increase in the number of significantly different pairs while passing from five to ten produces a 0.20% increase.

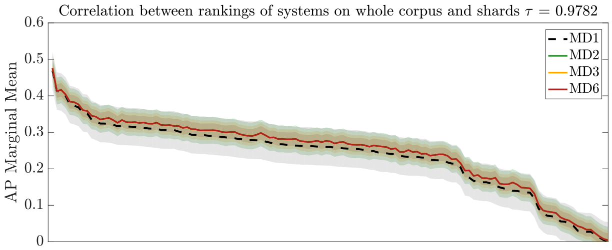
The more sophisticated a model, the more significant differences are detected. However, not all models are equally impactful. From model (MD2) to (MD3), i.e. adding the topic*system interaction, produces notable increases while passing from model (MD3) to (MD4) and (MD5), do not provide substantial benefits. However, model (MD6), i.e. adding the topic*shard interaction, makes another substantial increase in the number of significant differences, confirming the importance of this factor.

If we consider the group of the systems insignificantly different from the top performing system (column TopG), we can appreciate another benefit of using shards. The number of systems in the top group drops from 7 when using the whole corpus to 1 when using shards and the more descriptive models, suggesting that the increased accuracy in estimating differences among systems allows us to detect that the top system is actually different from others.

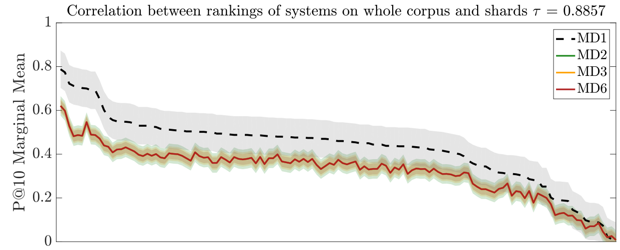
9.3 Robustness to Shard Sampling

Table 2 show the summary of the analyses for AP across different shard sizes when using ten samples for each shard size. The Kendall's τ column reports the average value of τ over the samples and its 95% confidence interval. For all the tracks, the τ values are quite high with small confidence intervals. This suggest that the RoS is quite stable and does not depend much on the specific random shards. Similar considerations hold also in the case of the Tukey confidence interval, which gets smaller has the shard size increases and whose values are similar across shard samples. This suggests that the detection of significantly different systems is not affected much by the specific random shards at hand.

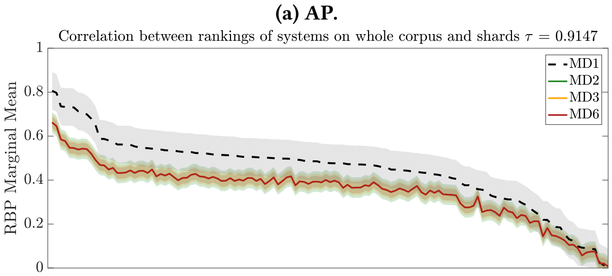
The total number of significantly different pairs support this hypothesis since we can see how the confidence interval around this value is small, indicating that their number does not change much when the shard sample changes. The final column reports the fraction of significant pairs found in common across all 10 samples. Here, there is a notable level of consistency across the samples.



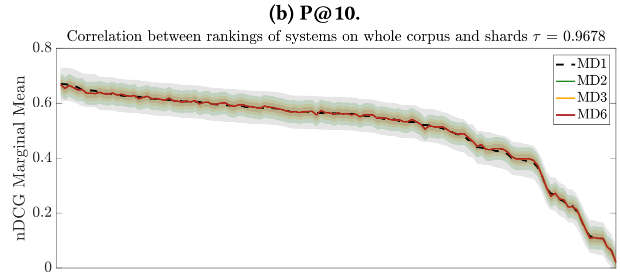
(a) AP.



(b) P@10.

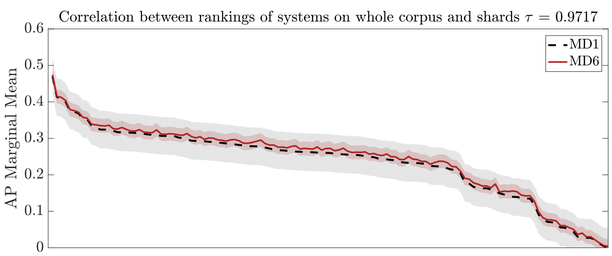


(c) nDCG.

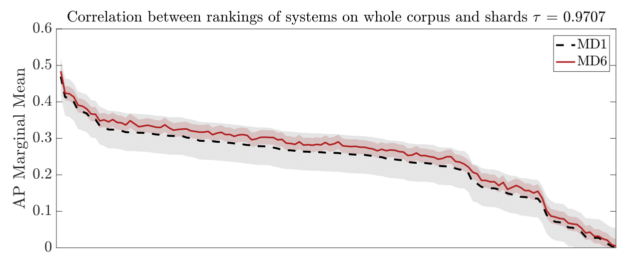


(d) RBP.

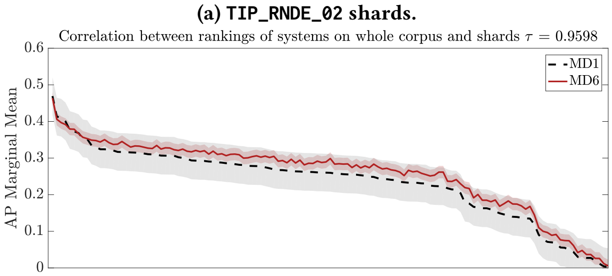
Figure 4: The Tukey confidence intervals (eq. (3)**) of four measures across four models. On `T08` with `TIP_RNDE_10` shards. On the x-axis there are the systems ordered by descending performance.**



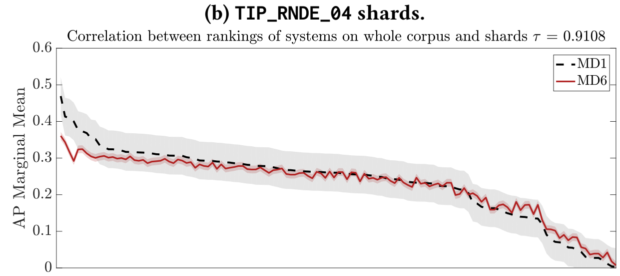
(a) `TIP_RNDE_02` shards.



(b) `TIP_RNDE_04` shards.



(c) `TIP_RNDE_10` shards.



(d) `TIP_RNDE_50` shards.

Figure 5: Comparing confidence intervals of eq. (3) **using models** (MD1) **and** (MD6) **for AP on `T08` with different shard numbers. On the x-axis there are the systems ordered by descending performance.**

10 CONCLUSIONS AND FUTURE WORK

At the start of the paper, we asked: *can an unequal value of topics be exploited to improve measurement of system performance accuracy on a test collection?*

We described and validated, theoretically and empirically, an ANOVA model combined with a random sharding technique. We showed that the model (MD6) measures substantially more significant differences between IR systems than conventional approaches, as represented by model (MD1). While it is true that a more sophisticated ANOVA model is expected to reduce measurement error, the scale of improvement seen with (MD6) is perhaps less expected. We showed that model (MD6) agrees well with the RoS taken from conventional test collection measurement and that the increased significance is not due to measurement error.

Past work has examined the question of whether the variability of topic measurement can be exploited to improve the accuracy of IR system measurement, we contend that our research shows that

Table 1: Comparing models for three shard sizes across 8256 system pairs, AP, track T08.

Model	vs Model	TIP_RNDE_02, $\tau = 0.9717$				TIP_RNDE_05, $\tau = 0.9707$				TIP_RNDE_10, $\tau = 0.9598$			
		$\hat{\omega}^2_{\langle sys \rangle}$	Sig	NotSig	TopG	$\hat{\omega}^2_{\langle sys \rangle}$	Sig	NotSig	TopG	$\hat{\omega}^2_{\langle sys \rangle}$	Sig	NotSig	TopG
MD1	–	0.3991	3423	4833	7	0.3991	3423	4833	7	0.3991	3423	4833	7
MD2	–	0.3500	4067	4189	4	0.2556	4607	3649	2	0.1595	4614	3642	2
	MD1	-12.29%	+18.81%	-13.33%	-42.86%	-35.95%	+34.59%	-24.50%	-71.43%	-60.02%	+34.79%	-24.64%	-71.43%
MD3	–	0.5678	5175	3081	1	0.3495	5133	3123	1	0.1840	4831	3425	1
	MD1	+42.28%	+51.18%	-36.25%	-85.71%	-12.42%	+49.96%	-35.38%	-85.71%	-53.90%	+41.13%	-29.13%	-85.71%
	MD2	+62.22%	+27.24%	-26.45%	-75.00%	+36.74%	+11.42%	-14.41%	-50.00%	+15.31%	+4.70%	-5.96%	-50.00%
MD4	–	0.5693	5180	3076	1	0.3511	5140	3116	1	0.1849	4833	3423	1
	MD1	+42.67%	+51.33%	-36.35%	-85.71%	-12.03%	+50.16%	-35.53%	-85.71%	-53.66%	+41.19%	-29.17%	-85.71%
	MD2	+62.66%	+27.37%	-26.57%	-75.00%	+37.34%	+11.57%	-14.61%	-50.00%	+15.92%	+4.75%	-6.01%	-50.00%
	MD3	+0.27%	+0.10%	-0.16%	–	+0.44%	+0.14%	-0.22%	–	+0.53%	+0.04%	-0.06%	–
MD5	–	0.5675	5173	3083	1	0.3486	5129	3127	1	0.1829	4818	3438	1
	MD1	+42.22%	+51.12%	-36.21%	-85.71%	-12.65%	+49.84%	-35.30%	-85.71%	-54.18%	+40.75%	-28.86%	-85.71%
	MD2	+62.15%	+27.19%	-26.40%	-75.00%	+36.38%	+11.33%	-14.31%	-50.00%	+14.62%	+4.42%	-5.60%	-50.00%
	MD3	-0.05%	-0.04%	+0.06%	-0.26%	–	-0.08%	+0.13%	-0.59%	–	-0.27%	+0.38%	–
	MD4	-0.32%	-0.14%	+0.23%	-0.70%	–	-0.21%	+0.35%	-1.12%	–	-0.31%	+0.44%	–
MD6	–	0.7143	5889	2367	1	0.5235	5935	2321	1	0.3777	5947	2309	1
	MD1	+78.99%	+72.04%	-51.02%	-85.71%	+31.19%	+73.39%	-51.98%	-85.71%	-5.36%	+73.74%	-52.22%	-85.71%
	MD2	+104.07%	+44.80%	-43.49%	-75.00%	+104.82%	+28.83%	-36.39%	-50.00%	+136.71%	+28.89%	-36.60%	-50.00%
	MD3	+25.80%	+13.80%	-23.17%	+49.79%	–	+15.62%	-25.68%	–	+105.29%	+23.10%	-32.58%	–
	MD4	+25.46%	+13.69%	-23.05%	+49.14%	–	+15.47%	-25.51%	–	+104.20%	+23.05%	-32.54%	–
	MD5	+25.86%	+13.84%	-23.22%	+50.18%	–	+15.71%	-25.78%	–	+106.52%	+23.43%	-32.84%	–

Table 2: Summary of analyses for AP using 10 samples of each random split and model (MD6).

T08 – 8256 system pairs compared				
Split	τ	CI Width	Sig. Pairs	Frac. Sig Pairs
TIP_RNDE_02	0.9803	0.0540	5142.20	0.6228
TIP_RNDE_03	0.9745	0.0551	5085.90	0.6160
TIP_RNDE_04	0.9680	0.0546	5104.10	0.6182
TIP_RNDE_05	0.9689	0.0549	5051.20	0.6118
TIP_RNDE_10	0.9613	0.0538	5008.70	0.6067
TIP_RNDE_25	0.9418	0.0445	5242.80	0.6350
TIP_RNDE_50	0.9189	0.0351	5462.40	0.6616
T09 – 5356 system pairs compared				
Split	τ	CI Width	Sig. Pairs	Frac. Sig Pairs
WT10g_RNDE_02	0.9609	0.0732	2808.30	0.5243
WT10g_RNDE_03	0.9453	0.0717	2874.00	0.5366
WT10g_RNDE_04	0.9380	0.0683	2947.70	0.5504
WT10g_RNDE_05	0.9275	0.0657	3034.50	0.5666
WT10g_RNDE_10	0.9037	0.0530	3426.80	0.6398
WT10g_RNDE_25	0.8813	0.0389	3748.00	0.6998
WT10g_RNDE_50	0.8675	0.0288	3893.60	0.7270
T27 – 2556 system pairs compared				
Split	τ	CI Width	Sig. Pairs	Frac. Sig Pairs
WAPO_RNDE_02	0.9764	0.0460	1821.50	0.7126
WAPO_RNDE_03	0.9634	0.0495	1791.20	0.7008
WAPO_RNDE_04	0.9617	0.0485	1800.10	0.7043
WAPO_RNDE_05	0.9583	0.0480	1802.70	0.7053
WAPO_RNDE_10	0.9470	0.0460	1822.80	0.7131
WAPO_RNDE_25	0.9219	0.0410	1848.30	0.7231
WAPO_RNDE_50	0.8812	0.0337	1853.30	0.7251

this is an approach with great promise. Model (MD6) allows us to make better use of existing test collections.

Our work in this particular direction of research is relatively new. Consequently, there are a number of avenues of future work:

- We want to compare our method with the recently published work of Voorhees et al. [39]. Their method also produces a substantial increase in the number of significant differences measured. However, their method of controlling for multiple significance test comparisons is more liberal than the method we use. There is the potential for combining our approaches, their technique uses a bootstrapping technique new to IR research, our technique uses a new ANOVA model.
- The metric of success, number significant differences, could be replaced by comparing the predictive power of our method with conventional methods. We could measure which of two systems is better on one test collection and see if those systems are similarly ordered on another test collection.
- What happens if we consider topics as random factors and/or heteroskedastic data, following the approach adopted by Robertson and Kanoulas [26]?
- Can we turn model (MD6) into a tool for designing better offline test collections, since it provides us with means for coping with differences across topics?
- Can model (MD6) also allow us to build test collections with fewer relevance judgments or topics while maintaining currently attainable measurement accuracies?
- Does this approach for offline testing tell us anything about online testing? Like the topics of test collections, online topics will have high and low numbers of relevant; do we need to think about how averaging works there too?
- How much of a benefit will model (MD6) bring to performance measurement on test collections where topics with very few relevant documents are rare or non-existent?

11 ACKNOWLEDGMENTS

This research is supported in part by the Australian Research Council's Discovery Projects Scheme (DP180102687).

The work is also 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] J. Allan, D. K. Harman, E. Kanoulas, and E. M. Voorhees. TREC 2018 Common Core Track Overview. In E. M. Voorhees and A. Ellis, editors, *The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018)*. National Institute of Standards and Technology (NIST), Special Publication, Washington, USA, 2019.

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

[3] D. Bodoff and P. Li. Test theory for assessing ir test collections. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, pages 367–374. ACM Press, New York, USA, 2007.

[4] L. Boytsov, A. Belova, and P. Westfall. Deciding on an Adjustment for Multiplicity in IR Experiments. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai, editors, *Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013)*, pages 403–412. ACM Press, New York, USA, 2013.

[5] C. Buckley and E. M. Voorhees. Retrieval System Evaluation. In D. K. Harman and E. M. Voorhees, editors, *TREC. Experiment and Evaluation in Information Retrieval*, pages 53–78. MIT Press, Cambridge (MA), USA, 2005.

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

[7] B. A. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over Thousands of Queries. In T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 651–658. ACM Press, New York, USA, 2008.

[8] G. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient Construction of Large Test Collections. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 282–289. ACM Press, New York, USA, 1998.

[9] G. V. Cormack and T. R. Lynam. Statistical Precision of Information Retrieval Evaluation. In E. N. Efthimiadis, S. Dumais, D. Hawking, and K. Järvelin, editors, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 533–540. ACM Press, New York, USA, 2006.

[10] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, USA, 1994.

[11] N. Ferro and M. Sanderson. Sub-corpora Impact on System Effectiveness. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White, editors, *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 901–904. ACM Press, New York, USA, 2017.

[12] N. Ferro, Y. Kim, and M. Sanderson. Using collection shards to study retrieval performance effect sizes. *ACM Transactions on Information Systems (TOIS)*, 37(3): 30:1–30:40, May 2019. ISSN 1046-8188. doi: 10.1145/3310364.

[13] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, 27(4):21, 2009.

[14] D. Hawking. Overview of the TREC-9 Web Track. In E. M. Voorhees and D. K. Harman, editors, *The Ninth Text REtrieval Conference (TREC-9)*, pages 87–103. National Institute of Standards and Technology (NIST), Special Publication 500-249, Washington, USA, 2000.

[15] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, USA, 1987.

[16] K. Järvelin and J. Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, October 2002.

[17] M. G. Kendall. *Rank correlation methods*. Griffin, Oxford, England, 1948.

[18] D. E. Losada, J. Parapar, and A. Barreiro. Feeling Lucky? Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation. In S. Ossowski, editor, *Proc. 2016 ACM Symposium on Applied Computing (SAC 2016)*, pages 1027–1034. ACM Press, New York, USA, 2016.

[19] R. Mehrotra and E. Yilmaz. Representative & informative query selection for learning to rank using submodular functions. In *Proceedings of the 38th international ACM sigir conference on research and development in information retrieval*, pages 545–554. ACM, 2015.

[20] S. Mizzaro and S. Robertson. Hits hits trec: exploring ir evaluation results with network analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 479–486. ACM, 2007.

[21] A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2:1–2:27, 2008.

[22] D. Newman. The Distribution of Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. *Biometrika*, 31(2):20–30, July 1939.

[23] S. Olejnik and J. Algina. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, 8(4): 434–447, December 2003.

[24] S. Robertson. On GMAP: and other transformations. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 78–83. ACM, 2006.

[25] S. Robertson. 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*, pages 9–22, 2007.

[26] S. E. Robertson and E. Kanoulas. On Per-topic Variance in IR Evaluation. In W. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 891–900. ACM Press, New York, USA, 2012.

[27] A. Rutherford. *ANOVA and ANCOVA. A GLM Approach*. John Wiley & Sons, New York, USA, 2nd edition, 2011.

[28] T. Sakai. Metrics, Statistics, Tests. In N. Ferro, editor, *Bridging Between Information Retrieval and Databases - PROMISE Winter School 2013, Revised Tutorial Lectures*, pages 116–163. Lecture Notes in Computer Science (LNCS) 8173, Springer, Heidelberg, Germany, 2014.

[29] T. Sakai. *Laboratory Experiments in Information Retrieval*, volume 40 of *The Information Retrieval Series*. Springer Singapore, 2018.

[30] M. Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR)*, 4(4):247–375, 2010.

[31] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. Differences in Effectiveness Across Sub-collections. In X. Chen, G. Lebanon, H. Wang, and M. J. Zaki, editors, *Proc. 21st International Conference on Information and Knowledge Management (CIKM 2012)*, pages 1965–1969. ACM Press, New York, USA, 2012.

[32] I. Soboroff. On evaluating web search with very few relevant documents. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 530–531. ACM Press, New York, USA, 2004.

[33] D.R. Swanson. Searching Natural Language Text by Computer. *Science*, 132(3434): 1099–1104, 1960. ISSN 0036-8075. URL https://www.jstor.org/stable/1706747.

[34] J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 1994.

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

[36] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, September 2000.

[37] E. M. Voorhees. On Building Fair and Reusable Test Collections using Bandit Techniques. In A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Broder, M. J. Zaki, S. Candan, A. Labrinidis, A. Schuster, and H. Wang, editors, *Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018)*, pages 407–416. ACM Press, New York, USA, 2018.

[38] E. M. Voorhees and D. K. Harman. Overview of the Eigth Text REtrieval Conference (TREC-8). In E. M. Voorhees and D. K. Harman, editors, *The Eighth Text REtrieval Conference (TREC-8)*, pages 1–24. National Institute of Standards and Technology (NIST), Special Publication 500-246, Washington, USA, 1999.

[39] E. M. Voorhees, D. Samarov, and I. Soboroff. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)*, 36(2): 12:1–12:21, September 2017.

[40] W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 51–58. ACM Press, New York, USA, 2008.

[41] M. Yang, P. Zhang, and D. Song. A study of per-topic variance on system comparison. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1181–1184. ACM, 2018. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3210122. URL http://doi.acm.org/10.1145/3209978.3210122.

[42] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, editors, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 512–519. ACM Press, New York, USA, 2005.

Improving the Accuracy of System Performance Estimation by Using Shards

Electronic Appendix

Nicola Ferro
ferro@dei.unipd.it
Dept. of Information Engineering, University of Padua
Padua, Italy

Mark Sanderson
mark.sanderson@rmit.edu.au
Computer Science, School of Science, RMIT University
Melbourne, Australia

CCS CONCEPTS

• **Information systems** → **Test collections**; **Retrieval effectiveness**;

KEYWORDS

effectiveness model; ANOVA; multiple comparison

A BACKGROUND ON ANOVA

ANalysis Of VAriance (ANOVA) [2, 3] attempts to explain data (the dependent variable scores) in terms of the experimental conditions (the model) and an error component. Typically, ANOVA is used to determine under which experimental condition do dependent variable score means differ and what proportion of variation in the dependent variable can be attributed to differences between specific experimental groups or conditions, as defined by the independent variable(s).

A typical ANOVA model is $y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, where y_{ij} is the i-th subject's dependent variable score in the j-th experimental condition, the parameter μ is the grand mean of the experimental condition population means that underlies all subjects' dependent variable scores, the parameter α_j is the effect of the j-th experimental condition and the random variable ε_{ij} is the error term, which reflects variation due to any uncontrolled source.

For a given model, the ANOVA table summarizes the outcomes of the ANOVA test indicating, for each factor, the *Sum of Squares (SS)*, the *Degrees of Freedom (DF)*, the *Mean Squares (MS)*, the F statistics, and the p-value of that factor, which allows us to determine the significance of that factor.

When it comes to independent variables they can be either *fixed effects* – i.e., they have precisely defined levels, and inferences about its effect apply only to those levels – or *random effects* – i.e., they describe a randomly and independently drawn set of levels that represent variation in a clearly defined wider population. The latter case is a more sophisticated model which, in the estimation of the variance attributed to the different factors, also accounts for the additional randomness due sampling of effect levels.

SIGIR'19, July 21–25, 2019, Paris, France.

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/http://dx.doi.org/10.475/123_4

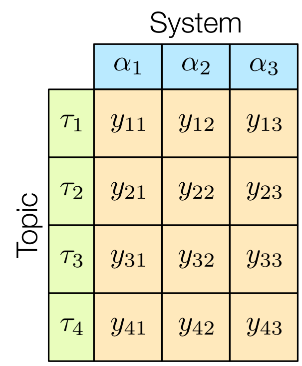


Figure 1: Example of layout of model (MD1) **for the topic and system effects, using $T = 4$ topics and $R = 3$ systems.**

The experimental design determines how you compute the model and how you estimate its parameters. In particular, it is possible to have an *independent measures* design where different subjects participate to different experimental conditions (factors) or a *repeated measures* design, where each subject participates to all the experimental conditions (factors).

A final distinction is between *crossed/factorial* designs, where every level of one factor is measured in combination with every level of the other factors, and *nested* designs, where levels of a factor are grouped within each level of another nesting factor.

A.1 Estimating the Model

Figure 1 shows the experimental layout for the model (MD1). This is the typical *Information Retrieval (IR)* setting where you have a set of topics and a set of systems which are run against those topics; in ANOVA terms this is a crossed/factorial repeated measures design. Note that this is the same model used by Banks et al. [1] and Tague-Sutcliffe and Blustein [4] to analyse TREC data.

$$y_{ij} = \mu_{\cdot\cdot} + \tau_i + \alpha_j + \varepsilon_{ij} \qquad \text{(MD1)}$$

where:

- $\mu_{\cdot\cdot}$ is the grand mean;
- τ_i with $i = 1, \ldots, T$ represents the effect of topics;
- α_j with $j = 1, \ldots, R$ represents the effect of systems;
- ε_{ij} is the residual error.

Model (MD1) has the following estimators:

- estimated grand mean

$$\hat{\mu}_{\cdot\cdot} = \frac{1}{T \cdot R} \sum_{i=1}^{T} \sum_{j=1}^{R} y_{ij}$$

- estimated topic marginal mean and topic effect

$$\hat{\mu}_{i\cdot} = \frac{1}{R} \sum_{j=1}^{R} y_{ij}$$

$$\hat{\tau}_i = \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}$$

- estimated system marginal mean and system effect

$$\hat{\mu}_{\cdot j} = \frac{1}{T} \sum_{i=1}^{T} y_{ij}$$

$$\hat{\alpha}_j = \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$$

Therefore, the score predicted by the model is

$$\hat{y}_{ij} = \hat{\mu}_{\cdot\cdot} + \hat{\tau}_i + \hat{\alpha}_j = \hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$$

and the estimated residuals are:

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})$$

A.2 Assessment of the Model

We can compute the *Sum of Squares (SS)*, *Degrees of Freedom (DF)*, *Mean Squares (MS)* and F statistics as follows:

- total effects

$$SS_{\text{total}} = \sum_{i=1}^{T} \sum_{j=1}^{R} (y_{ij} - \hat{\mu}_{\cdot\cdot})^2$$

$$df_{\text{total}} = T \cdot R - 1$$

$$MS_{\text{total}} = \frac{SS_{\text{total}}}{df_{\text{total}}}$$

- topic effects

$$SS_{\text{topic}} = \sum_{i=1}^{T} \sum_{j=1}^{R} \hat{\tau}_i^2 = R \sum_{i=1}^{T} \hat{\tau}_i^2 = R \sum_{i=1}^{T} (\hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot})^2$$

$$df_{\text{topic}} = T - 1$$

$$MS_{\text{topic}} = \frac{SS_{\text{topic}}}{df_{\text{topic}}}$$

$$F_{\text{topic}} = \frac{MS_{\text{topic}}}{MS_{\text{error}}}$$

- system effects

$$SS_{\text{system}} = \sum_{i=1}^{T} \sum_{j=1}^{R} \hat{\alpha}_j^2 = T \sum_{j=1}^{R} \hat{\alpha}_j^2 = T \sum_{j=1}^{R} (\hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})^2$$

$$df_{\text{system}} = R - 1$$

$$MS_{\text{system}} = \frac{SS_{\text{system}}}{df_{\text{system}}}$$

$$F_{\text{system}} = \frac{MS_{\text{system}}}{MS_{\text{error}}}$$

- error effects

$$SS_{\text{error}} = \sum_{i=1}^{T} \sum_{j=1}^{R} \hat{\varepsilon}_{ij}^2 = \sum_{i=1}^{T} \sum_{j=1}^{R} \big(y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})\big)^2$$

$$df_{\text{error}} = (T-1)(R-1)$$

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{df_{\text{error}}}$$

Note that:

$$SS_{\text{total}} = SS_{\text{topic}} + SS_{\text{system}} + SS_{\text{error}}$$

We can then compute the critical value for the F statistics of a factor, i.e. $F_{crit} = F_{(df_{fact}, df_{err})}$, and determine its significance if $F_{fact} > F_{crit}$; this allows us also to obtain the p-value for that factor.

B EFFECT OF `UNDEFINED` VALUES ON ESTIMATION

Suppose we are dealing with T topics, R system (runs), and S shards and thus $N = T \cdot R \cdot S$ total samples.

We consider the following ANOVA models:

$$y_{ijk} = \mu_{\cdot\cdot\cdot} + \underbrace{\tau_i + \alpha_j}_{\text{Main Effects}} + \varepsilon_{ijk} \tag{MD2}$$

$$y_{ijk} = \mu_{\cdot\cdot\cdot} + \underbrace{\tau_i + \alpha_j}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD3}$$

$$y_{ijk} = \mu_{\cdot\cdot\cdot} + \underbrace{\tau_i + \alpha_j + \beta_k}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD4}$$

$$y_{ijk} = \mu_{\cdot\cdot\cdot} + \underbrace{\tau_i + \alpha_j + \beta_k}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij} + (\alpha\beta)_{jk}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD5}$$

$$y_{ijk} = \mu_{\cdot\cdot\cdot} + \underbrace{\tau_i + \alpha_j + \beta_k}_{\text{Main Effects}} + \underbrace{(\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk}}_{\text{Interaction Effects}} + \varepsilon_{ijk} \tag{MD6}$$

where:

- y_{ijk} is the effectiveness score of the i-th subject (topic, $i = 1, \dots, T$) in the j-th (system, $j = 1, \dots, R$) and k-th (shard, $k = 1, \dots, S$) factors;
- $\mu_{\cdot\cdot\cdot}$ is the grand mean;
- $\tau_i = \mu_{i\cdot\cdot} - \mu_{\cdot\cdot\cdot}$ is the effect of the i-th subject, i.e. a topic, where $\mu_{i\cdot\cdot}$ is the marginal mean of the i-th subject;
- $\alpha_j = \mu_{\cdot j\cdot} - \mu_{\cdot\cdot\cdot}$ is the effect of the j-th factor, i.e. a system, where $\mu_{\cdot j\cdot}$ is the marginal mean of the j-th factor;
- $\beta_k = \mu_{\cdot\cdot k} - \mu_{\cdot\cdot\cdot}$ is the effect of the k-th factor, i.e. a shard, where $\mu_{\cdot\cdot k}$ is the marginal mean of the k-th factor;
- $(\tau\alpha)_{ij} = \mu_{ij\cdot} - \mu_{i\cdot\cdot} - \mu_{\cdot j\cdot} + \mu_{\cdot\cdot\cdot}$ is the interaction between topics and systems, where $\mu_{ij\cdot}$ is the marginal mean of the interaction between the i-th topic and j-th system;
- $(\tau\beta)_{ik} = \mu_{i\cdot k} - \mu_{i\cdot\cdot} - \mu_{\cdot\cdot k} + \mu_{\cdot\cdot\cdot}$ is the interaction between topics and shards, where $\mu_{i\cdot k}$ is the marginal mean of the interaction between the i-th topic and k-th shard;
- $(\alpha\beta)_{jk} = \mu_{\cdot jk} - \mu_{\cdot j\cdot} - \mu_{\cdot\cdot k} + \mu_{\cdot\cdot\cdot}$ is the interaction between systems and shards, where $\mu_{\cdot jk}$ is the marginal mean of the interaction between the j-th system and k-th shard;

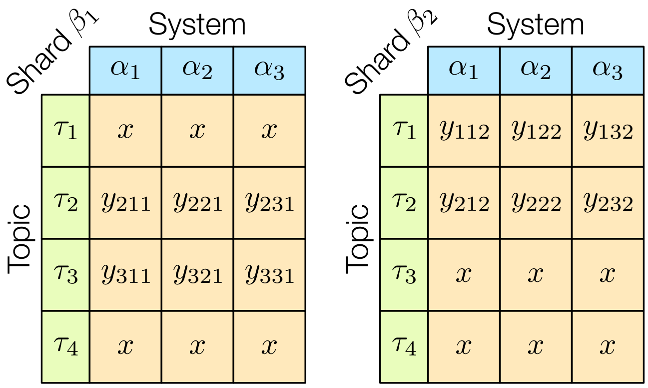


Figure 2: Example of ANOVA with $T = 4$ topics, $R = 3$ systems and $S = 2$ shards.

- finally, ε_{ijk} is the error committed by the model in predicting y_{ijk}.

Consider the example of Figure 2 where we have $T = 4$ topics, $R = 3$ systems and $S = 2$ shards. Moreover, topic τ_1 does not have any relevant document on shard β_1 and, therefore, all the systems have the `undefined` value x for that topic; similarly, topic τ_3 does not have any relevant document on shard β_2; finally topic τ_4 does not have any relevant document on both shard β_1 and β_2.

Definition B.1. Given a shard $k \in [1, S]$, X_k is the set of the indexes i of the topics which do not have any relevant document on that shard:

$$X_k = \Big\{ i \in [1, T] \,\Big|\, y_{ijk} = x \;\forall j \in [1, R] \Big\} \tag{1}$$

In the example of Figure 2, we have $X_1 = \{1, 4\}$ and $X_2 = \{3, 4\}$.

Note that, for any shard k, there are $|X_k| \cdot R$ `undefined` values and, in total, there are $R \sum_{k=1}^{S} |X_k|$ `undefined` values.

Definition B.2. Given a topic $i \in [1, T]$, X^i is the set of the indexes k of the shards for which the topic i does not have any relevant document:

$$X^i = \Big\{ k \in [1, S] \,\Big|\, y_{ijk} = x \;\forall j \in [1, R] \Big\} \tag{2}$$

In the example of Figure 2, $X^1 = \{1\}$, $X^2 = \varnothing$, $X^3 = \{2\}$, and $X^4 = \{1, 2\}$.

Note that, for any topic i, there are $|X^i| \cdot R$ `undefined` values and, in total, there are $R \sum_{i=1}^{T} |X^i|$ `undefined` values.

Proposition B.3. *Given models from* (MD2) *to* (MD6) *and a system $j \in [1, R]$, its estimated marginal mean is given by:*

$$\hat{\mu}_{\cdot j \cdot} = \underbrace{\frac{1}{T \cdot S} \sum_{k=1}^{S} \sum_{\substack{i=1 \\ i \notin X_k}}^{T} y_{ijk}}_{\hat{\mu}'_{\cdot j \cdot}} + \frac{x}{T \cdot S} \sum_{k=1}^{S} |X_k| \tag{3}$$

Therefore, for any pair of systems $u \in [1, R]$ and $v \in [1, R]$, $u \neq v$, the difference of their estimated marginal means $\hat{\mu}_{\cdot u \cdot} - \hat{\mu}_{\cdot v \cdot}$ is independent of the `undefined` *values.*

Proof. The estimated marginal mean of the system factor is given by:

$$\begin{aligned}
\hat{\mu}_{\cdot j \cdot} &= \frac{1}{T \cdot S} \sum_{i=1}^{T} \sum_{k=1}^{S} y_{ijk} = \frac{1}{T \cdot S} \sum_{k=1}^{S} \sum_{i=1}^{T} y_{ijk} = \\
&= \frac{1}{T \cdot S} \sum_{k=1}^{S} \left(\sum_{\substack{i=1 \\ i \notin X_k}}^{T} y_{ijk} + \sum_{i \in X_k} x \right) = \\
&= \frac{1}{T \cdot S} \sum_{k=1}^{S} \left(\sum_{\substack{i=1 \\ i \notin X_k}}^{T} y_{ijk} + x \cdot |X_k| \right) = \\
&= \underbrace{\frac{1}{T \cdot S} \sum_{k=1}^{S} \sum_{\substack{i=1 \\ i \notin X_k}}^{T} y_{ijk}}_{\hat{\mu}'_{\cdot j \cdot}} + \frac{x}{T \cdot S} \sum_{k=1}^{S} |X_k|
\end{aligned}$$

Therefore, for any two systems u and v, we have that:

$$\begin{aligned}
\hat{\mu}_{\cdot u \cdot} - \hat{\mu}_{\cdot v \cdot} &= \left(\hat{\mu}'_{\cdot u \cdot} + \frac{x}{T \cdot S} \sum_{k=1}^{S} |X_k| \right) - \left(\hat{\mu}'_{\cdot v \cdot} + \frac{x}{T \cdot S} \sum_{k=1}^{S} |X_k| \right) = \\
&= \hat{\mu}'_{\cdot u \cdot} - \hat{\mu}'_{\cdot v \cdot}.
\end{aligned}$$

Therefore, the difference of the estimated marginal means of the system factor is independent of the `undefined` values.

□

Proposition B.4. *Given models from* (MD2) *to* (MD6)*, the SS of the system factor and, as a consequence, the MS of the system factor are independent of the* `undefined` *values.*

Proof. The estimated grand mean is given by:

$$\begin{aligned}
\hat{\mu}_{\cdots} &= \frac{1}{T \cdot R \cdot S} \sum_{i=1}^{T} \sum_{j=1}^{R} \sum_{k=1}^{S} y_{ijk} = \frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \sum_{i=1}^{T} \sum_{j=1}^{R} y_{ijk} = \\
&= \frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \left(\sum_{\substack{i=1 \\ i \notin X_k}}^{T} \sum_{j=1}^{R} y_{ijk} + \sum_{i \in X_k} \sum_{j=1}^{R} x \right) = \\
&= \frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \left(\sum_{\substack{i=1 \\ i \notin X_k}}^{T} \sum_{j=1}^{R} y_{ijk} + \sum_{i \in X_k} R \cdot x \right) = \\
&= \frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \left(\sum_{\substack{i=1 \\ i \notin X_k}}^{T} \sum_{j=1}^{R} y_{ijk} + |X_k| \cdot R \cdot x \right) = \\
&= \underbrace{\frac{1}{T \cdot R \cdot S} \sum_{k=1}^{S} \sum_{\substack{i=1 \\ i \notin X_k}}^{T} \sum_{j=1}^{R} y_{ijk}}_{\hat{\mu}'_{\cdots}} + \frac{x}{T \cdot S} \sum_{k=1}^{S} |X_k|
\end{aligned}$$

The SS of the system factor is given by:

$$
\begin{aligned}
SS_{\text{system}} &= \sum_{i=1}^{T}\sum_{j=1}^{R}\sum_{k=1}^{S}\left(\hat{\mu}_{\cdot j\cdot} - \hat{\mu}_{\cdots}\right)^2 = T\cdot S\sum_{j=1}^{R}\left(\hat{\mu}_{\cdot j\cdot} - \hat{\mu}_{\cdots}\right)^2 = \\
&= T\cdot S\sum_{j=1}^{R}\left(\hat{\mu}'_{\cdot j\cdot} + \frac{x}{T\cdot S}\sum_{k=1}^{S}|X_k| - \hat{\mu}'_{\cdots} - \frac{x}{T\cdot S}\sum_{k=1}^{S}|X_k|\right)^2 = \\
&= T\cdot S\sum_{j=1}^{R}\left(\hat{\mu}'_{\cdot j\cdot} - \hat{\mu}'_{\cdots}\right)^2
\end{aligned}
$$

Therefore, the SS of the system factor is independent of the `undefined` values, i.e. they do not change the amount of variance attributed to the system factor. Since $MS_{\text{system}} = \frac{SS_{\text{system}}}{df_{\text{system}}}$, also the the MS of the system factor is independent of the `undefined` values. □

Proposition B.5. *Given model* (MD6), *the residuals ε_{ijk} are independent of the* `undefined` *values. Therefore, the SS of the error and, as a consequence, the MS of the error are independent of the* `undefined` *values.*

Proof. The estimated marginal mean of the topic factor is given by:

$$
\begin{aligned}
\hat{\mu}_{i\cdot\cdot} &= \frac{1}{R\cdot S}\sum_{j=1}^{R}\sum_{k=1}^{S} y_{ijk} = \frac{1}{R\cdot S}\sum_{j=i}^{R}\left(\sum_{\substack{k=1\\k\notin X^i}}^{S} y_{ijk} + \sum_{k\in X^i} x\right) = \\
&= \frac{1}{R\cdot S}\sum_{j=i}^{R}\sum_{\substack{k=1\\k\notin X^i}}^{S} y_{ijk} + \frac{1}{R\cdot S}\sum_{j=i}^{R}\left|X^i\right| x = \\
&= \underbrace{\frac{1}{R\cdot S}\sum_{j=i}^{R}\sum_{\substack{k=1\\k\notin X^i}}^{S} y_{ijk}}_{\hat{\mu}'_{i\cdot\cdot}} + \frac{\left|X^i\right| x}{S}
\end{aligned}
$$

The estimated marginal mean of the shard factor is given by:

$$
\begin{aligned}
\hat{\mu}_{\cdot\cdot k} &= \frac{1}{T\cdot R}\sum_{i=1}^{T}\sum_{j=1}^{R} y_{ijk} = \frac{1}{T\cdot R}\sum_{j=1}^{R}\sum_{i=1}^{T} y_{ijk} = \\
&= \frac{1}{T\cdot R}\sum_{j=i}^{R}\left(\sum_{\substack{i=1\\i\notin X_k}}^{T} y_{ijk} + \sum_{i\in X_k} x\right) = \\
&= \frac{1}{T\cdot R}\sum_{j=i}^{R}\sum_{\substack{i=1\\i\notin X_k}}^{T} y_{ijk} + \frac{1}{T\cdot R}\sum_{j=i}^{R}|X_k|\, x = \\
&= \underbrace{\frac{1}{T\cdot R}\sum_{j=i}^{R}\sum_{\substack{i=1\\i\notin X_k}}^{T} y_{ijk}}_{\hat{\mu}'_{\cdot\cdot k}} + \frac{|X_k|\, x}{T}
\end{aligned}
$$

The estimated marginal mean of the topic*system interaction factor is given by:

$$
\begin{aligned}
\hat{\mu}_{ij\cdot} &= \frac{1}{S}\sum_{k=1}^{S} y_{ijk} = \frac{1}{S}\sum_{\substack{k=1\\k\notin X^i}}^{S} y_{ijk} + \frac{1}{S}\sum_{k\in X^i} x = \\
&= \underbrace{\frac{1}{S}\sum_{\substack{k=1\\k\notin X^i}}^{S} y_{ijk}}_{\hat{\mu}'_{ij\cdot}} + \frac{\left|X^i\right| x}{S}
\end{aligned}
$$

The estimated marginal mean of the system*shard interaction factor is given by:

$$
\begin{aligned}
\hat{\mu}_{\cdot jk} &= \frac{1}{T}\sum_{i=1}^{T} y_{ijk} = \frac{1}{T}\sum_{\substack{i=1\\i\notin X_k}}^{T} y_{ijk} + \frac{1}{T}\sum_{i\in X_k} x = \\
&= \underbrace{\frac{1}{T}\sum_{\substack{i=1\\i\notin X_k}}^{T} y_{ijk}}_{\hat{\mu}'_{\cdot jk}} + \frac{|X_k|\, x}{T}
\end{aligned}
$$

The estimated marginal mean of the topic*shard interaction factor is given by:

$$
\begin{aligned}
\hat{\mu}_{i\cdot k} &= \frac{1}{R}\sum_{j=1}^{R} y_{ijk} = \begin{cases} \frac{1}{R}\sum_{j=1}^{R} y_{ijk} & \text{if } i\notin X_k \\ \frac{1}{R}\sum_{j=1}^{R} x & \text{if } i\in X_k \end{cases} \\
&= \begin{cases} \hat{\mu}'_{i\cdot k} & \text{if } i\notin X_k \\ x & \text{if } i\in X_k \end{cases}
\end{aligned}
$$

The predicted score for (MD6) is given by:

$$
\begin{aligned}
\hat{y}_{ijk} &= \hat{\mu}_{\cdots} + \left(\hat{\mu}_{i\cdot\cdot} - \hat{\mu}_{\cdots}\right) + \left(\hat{\mu}_{\cdot j\cdot} - \hat{\mu}_{\cdots}\right) + \left(\hat{\mu}_{\cdot\cdot k} - \hat{\mu}_{\cdots}\right) + \\
&\quad \left(\hat{\mu}_{ij\cdot} - \hat{\mu}_{i\cdot\cdot} - \hat{\mu}_{\cdot j\cdot} + \hat{\mu}_{\cdots}\right) + \left(\hat{\mu}_{\cdot jk} - \hat{\mu}_{\cdot j\cdot} - \hat{\mu}_{\cdot\cdot k} + \hat{\mu}_{\cdots}\right) + \\
&\quad \left(\hat{\mu}_{i\cdot k} - \hat{\mu}_{i\cdot\cdot} - \hat{\mu}_{\cdot\cdot k} + \hat{\mu}_{\cdots}\right) = \\
&= \hat{\mu}_{\cdots} - \hat{\mu}_{i\cdot\cdot} - \hat{\mu}_{\cdot j\cdot} - \hat{\mu}_{\cdot\cdot k} + \hat{\mu}_{ij\cdot} + \hat{\mu}_{\cdot jk} + \hat{\mu}_{i\cdot k} = \\
&= \hat{\mu}'_{\cdots} + \frac{x}{T\cdot S}\sum_{k=1}^{S}|X_k| - \hat{\mu}'_{i\cdot\cdot} - \frac{\left|X^i\right| x}{S} - \hat{\mu}'_{\cdot j\cdot} - \frac{x}{T\cdot S}\sum_{k=1}^{S}|X_k| - \\
&\quad \hat{\mu}'_{\cdot\cdot k} - \frac{|X_k|\, x}{T} + \hat{\mu}'_{ij\cdot} + \frac{\left|X^i\right| x}{S} + \hat{\mu}'_{\cdot jk} + \frac{|X_k|\, x}{T} + \hat{\mu}_{i\cdot k} = \\
&= \hat{\mu}'_{\cdots} - \hat{\mu}'_{i\cdot\cdot} - \hat{\mu}'_{\cdot j\cdot} - \hat{\mu}'_{\cdot\cdot k} + \hat{\mu}'_{ij\cdot} + \hat{\mu}'_{\cdot jk} + \hat{\mu}_{i\cdot k} = \\
&= \begin{cases} \hat{\mu}'_{\cdots} - \hat{\mu}'_{i\cdot\cdot} - \hat{\mu}'_{\cdot j\cdot} - \hat{\mu}'_{\cdot\cdot k} + \hat{\mu}'_{ij\cdot} + \hat{\mu}'_{\cdot jk} + \hat{\mu}'_{i\cdot k} & \text{if } i\notin X_k \\ \hat{\mu}'_{\cdots} - \hat{\mu}'_{i\cdot\cdot} - \hat{\mu}'_{\cdot j\cdot} - \hat{\mu}'_{\cdot\cdot k} + \hat{\mu}'_{ij\cdot} + \hat{\mu}'_{\cdot jk} + x & \text{if } i\in X_k \end{cases} = \\
&= \begin{cases} \hat{\mu}'_{ijk} + \hat{\mu}'_{i\cdot k} & \text{if } i\notin X_k \\ \hat{\mu}'_{ijk} + x & \text{if } i\in X_k \end{cases}
\end{aligned}
$$

Thus, the estimated residuals for (MD6) are given by:

$$\hat{\varepsilon}_{ijk} = y_{ijk} - \hat{y}_{ikj} = \begin{cases} y_{ijk} - \hat{\mu}'_{ijk} - \hat{\mu}'_{i\cdot k} & \text{if } i \notin X_k \\ x - \hat{\mu}'_{ijk} - x & \text{if } i \in X_k \end{cases} =$$

$$= \begin{cases} y_{ijk} - \hat{\mu}'_{ijk} - \hat{\mu}'_{i\cdot k} & \text{if } i \notin X_k \\ -\hat{\mu}'_{ijk} & \text{if } i \in X_k \end{cases}$$

Therefore, the estimated residuals $\hat{\varepsilon}_{ijk}$ are independent of the `undefined` values. As a consequence, the SS of the error and $MS_{\text{error}} = \frac{SS_{\text{error}}}{df_{\text{error}}}$ are independent of the `undefined` values as well.

□

REFERENCES

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

[2] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, New York, USA, 5th edition, 2005.

[3] S. Maxwell and H. D. Delaney. *Designing Experiments and Analyzing Data. A Model Comparison Perspective*. Lawrence Erlbaum Associates, Mahwah (NJ), USA, 2nd edition, 2004.

[4] J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In D. K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 1994.