Detecting Significant Differences Between Information Retrieval Systems via Generalized Linear Models

Guglielmo Faggioli
University of Padova
guglielmo.faggioli@phd.unipd.it

Nicola Ferro
University of Padova
ferro@dei.unipd.it

Norbert Fuhr
University of Duisburg-Essen
norbert.fuhr@uni-due.de

ABSTRACT

Being able to compare *Information Retrieval (IR)* systems correctly is pivotal to improving their quality. Among the most popular tools for statistical significance testing, we list t-test and ANOVA that belong to the linear models family. Therefore, given the relevance of linear models for IR evaluation, a great effort has been devoted to studying how to improve them to better compare IR systems.

Linear models rely on assumptions that IR experimental observations rarely meet, e.g. about the normality of the data or the linearity itself. Even though linear models are, in general, resilient to violations of their assumptions, departing from them might reduce the effectiveness of the tests. Hence, we investigate the use of the *Generalized Linear Model (GLM)* framework, a generalization of the traditional linear modelling that relaxes assumptions about the distribution and the shape of the models. To the best of our knowledge, there has been little or no investigation on the use of GLMs for comparing IR system performance. We discuss how GLMs work and how they can be applied in the context of IR evaluation. In particular, we focus on the *link function* used to build GLMs, which allows for the model to have non-linear shapes.

We conduct a thorough experimentation using two TREC collections and several evaluation measures. Overall, we show how the log and logit links are able to identify more and more consistent significant differences (up to 25% more with 50 topics) than the identity link used today and with a comparable, or slightly better, risk of publication bias.

ACM Reference Format:
Guglielmo Faggioli, Nicola Ferro, and Norbert Fuhr. 2022. Detecting Significant Differences Between Information Retrieval Systems via Generalized Linear Models. In *Proceedings of 31st ACM International Conference on Information and Knowledge Management (CIKM '22).* ACM, New York, NY, USA, 11 pages. https://doi.org/11.111/111111.1111

1 INTRODUCTION

Evaluation in *Information Retrieval (IR)* allows researchers and practitioners to study and compare their systems in order to understand how to improve them. To this end, sound statistical inference methods are needed to obtain robust and generalizable insights and to predict what happens when systems run in a real-world scenario. Therefore, statistical analyses, such as bootstrap, randomization tests [37, 44], t-tests, and *ANalysis Of VAriance (ANOVA)* [3, 36, 45] have been widely studied and successfully employed in IR evaluation.

In particular, t-test and ANOVA belong to the family of statistical methods called *General Linear Models (GLiMs)*, a generalization of the multiple linear regression, which is based on the following assumptions: *i)* independence of the observations, *ii)* constant variance of the data, i.e. *homoscedasticity iii)* normal distribution of the data, i.e., *normality*; last but not least and too often overlooked: *iv)* linear correlation between experimental conditions and the expectation of the response i.e., *linearity*. These assumptions allow for an analytical solution of the model and its practical computation. Furthermore, the more such assumptions are satisfied, the more accurate is the estimation of the model and the inferences drawn from it. Previous literature showed great interest in studying the empirical consequences of using data violating such assumptions, both from a theoretical standpoint [22, 43], and also considering empirical IR data [7, 10, 21, 45]. Such works show that, in general, linear models are resilient to the violation of their assumptions. At the same time, several works have explored how to make IR data closer to the GLiM assumptions, e.g. by transforming the data [10, 45]. In all the cases, the ultimate goal is to obtain models which are capable to better and more reliably distinguish among IR systems.

The *Generalized Linear Model (GLM)* framework is a generalization of the GLiMs that relaxes some of the underlying assumptions in order to increase models' applicability. In particular, the data is no longer required to follow a normal distribution or to have constant variance. Moreover, GLMs also relax the fourth assumption, allowing the relationship between the expectation of the response and the experimental conditions – called *link* – to have different forms besides the linear one.

In this work, we investigate the application of Generalized Linear Models to IR evaluation and show how they can help us in better comparing and distinguishing among systems. More precisely, we focus on the *link function* used in the GLMs framework, investigating the impact of different links on the modelling of the IR performance. The main contributions of this work are:

- We propose a new visualization of IR data that highlights linear models' assumptions. We then illustrate the behaviour of the IR data using such visualization;
- We instantiate the GLMs framework in the case of IR experimental evaluation, illustrating how to apply it;
- We experimentally compare different links to determine the most suited to the IR scenario, showing which links improve the ability of the models to better and more reliably distinguish among IR systems.

CIKM '22, October 17–22, 2022, Atlanta, USA

ACM ISBN 111-1-1111-1111-1/11/11…$15.00
https://doi.org/11.111/111111.1111

The remainder is organized as follows: Section 2 reports the main related works. Section 3 describes GLMs and what are the challenges to applying them in the IR scenario. Section 4 reports the methodology applied to assess the behaviour of GLMs for IR evaluation tasks, while Section 5 illustrates the empirical findings. Finally, Section 6 draws our conclusions and future works.

2 RELATED WORK

GLMs were proposed by Nelder and Wedderburn [30] in 1972 and, since then, have been studied and documented extensively [25], becoming a widely accepted standard in the statistics community. GLMs have been applied successfully to several scenarios, such as engineering, biology and medicine [29]. We describe this method in more detail in Section 3.1. To the best of our knowledge, they have found very limited application to modelling and comparing performance of IR systems.

Note that, when it comes to compare system performance, there are also completely alternative approaches to linear modelling, such as Carterette [6, 8] who proposes to develop a Bayesian framework for hypothesis testing. In our work, instead of adopting a completely different modelling strategy, we remain in the linear modelling framework but expand it via GLMs.

GLiMs are a specific type of GLMs and they include t-test and ANOVA which, according to both Sakai [38] and Carterette [9], are the most widely adopted statistical significance tests in IR evaluation.

2.1 Violation of Assumptions and Response Transformation

Early work by Saracevic [42] pointed out that IR data do not comply with the assumptions of significance tests and this was stated again by van Rijsbergen [51, chap. 7], who suggested to use non parametric tests, such as the sign test, due to the smaller number of violations to assumptions. On the other hand, Hull [21] studies several tests, among which t-test and ANOVA, and concluded that, despite the violation of assumptions, it was of practical importance to adopt such tests in order to properly validate experimental conclusions. More recently, Carterette [7] studied, mostly by using simulations, the impact of deviations from the assumptions and concluded that homoscedasticity and linearity matter more than normality, even if their impact is smaller compared to not adjusting for multiple comparisons, which is known to be a severe flaw in many experiments [19, 40].

One of the most widely adopted strategies to bring IR data closer to the assumptions of linear models is to transform the data by applying a function to them and, consequently, change their distribution. Transformations help to obtain more normal and homoscedastic distributions of performance.

Tague-Sutcliffe and Blustein [45] proposed to use the *arcsin of the square root* or the *rank* of the performance scores to make them closer to a normal distribution. *Logit* is another widely used transformation, which maps performance scores in the (0, 1) range to $\mathbb{R}$, making them closer to a normal distribution. This transformation was originally studied by Cormack and Lynam [10] and later on employed by Robertson and Kanoulas [35] and Berto et al. [4]. Robertson [34] explores a further smoothed version of the logit transformation of the *Average Precision (AP)*, which exhibits higher normality, dubbed "yaAP". The logit transformation has the limitation of not being defined for values equal to 0 and 1; therefore, it requires either to ignore them or to resort to some smoothing-based solution, as proposed by [10, 34]. Robertson [33] compares the advantages - and disadvantages - of several transformation strategies, including log and logit ones, including use of *Geometric Mean Average Precision (GMAP)* by Voorhees [52].

While the efforts mentioned above focus primarily on the non-normality of the performance scores, Sakai [39] and, later, Urbano et al. [49] explore a standardization-based approach aimed at increasing the homoscedasticity of the data. They propose to transform the performance scores into z-scores, as also proposed by Webber et al. [55] for other purposes, and apply a linear transformation to such scores to reduce the inter-topic heteroscedasticity.

Both GLMs and response transformation exploit non-linear functions to improve the fitness of the models to the data. Nevertheless, they represent two completely different approaches regarding assumptions, computation, and interpretation of the predictions, as we discuss in detail in Section 3.1. This markedly differentiates our approach based on GLM from previous works in the field based on transformations. Moreover, while Carterette [7] pointed out the potential impact of departing from linearity, to the best of our knowledge, less (or no) attention has been paid to how to address this aspect, which instead is the focus of our work, by exploring the *link* function. Finally, most of the approaches mentioned above focused just on AP, being the most popular measure in IR evaluation, while in our work we also study other measures, namely Precision, Recall, *Normalized Discounted Cumulated Gain (nDCG)* [23] and *Rank-Biased Precision (RBP)* [27].

2.2 Model Factors

Another relevant strand of research about GLiMs, not related to compliance with assumptions, concerns the factors in the models and how to increase their ability to distinguish among IR systems.

Tague-Sutcliffe and Blustein [45], first, and Banks et al. [3], later, compared IR systems using a two-way ANOVA where factors were topics and systems. Robertson and Kanoulas [35] use simulated data to add a further factor, i.e. the interaction between topics and systems. Ferro et al. [18], Voorhees et al. [54] and Ferro and Sanderson [14] "sharded" a collection by randomly partitioning the documents and this allowed them to expand the models by including the shard factor as well as all the the interaction factors between topics, systems, and shards. Bodoff and Li [5] considered multiple assessors while Bailey et al. [2] exploited multiple formulations for the same topic; in both cases, the replicates coming from either multiple assessors or multiple formulations allowed for estimating the interaction between topics and systems. Recently, Culpepper et al. [11] used topic reformulations and different corpora (not shards of the same corpus) in order to study the difficulty of topics with respect to corpora. On a slightly different note, Ferro and Silvello [16, 17], instead of adding more factors to the models, decomposed the system factor into its constituting components (stop list, stemmer, IR model) in order to study the contribution (and interaction) of these components to the overall performance.

Following Tague-Sutcliffe and Blustein [45], we consider GLMs constituted by the topic and system factors, leaving multi-factor analyses for future work, since our focus in this work is to take the first step and understand how GLMs behave with IR data rather than making GLMs able to distinguish among more and more systems by adding more and more factors.

2.3 Assessing a Test

Over the years, several approaches have been developed to assess significance tests and determine which are most effective.

2.3.1 Number of significantly different pairs. Faggioli and Ferro [12], Ferro and Sanderson [13, 14], Ferro et al. [18], Voorhees et al. [54] compared several alternative ANOVA models and considered superior those able to identify an higher number of *statistically significantly different (ssd)* system pairs, relating this improvement to smaller confidence intervals [14, 54] and to smaller un-modelled error levels [18]. We will rely also on this approach in our experiments.

2.3.2 Topic splitting. Zobel [57] randomly partitioned topics into two sets and compared systems across both sets using different significance tests (ANOVA, Wilcoxon, and t-test). The *level of agreement* across the two sets is used as an indicator of the superiority of a test over another one and as a proxy of test errors. This methodology was adopted by others: Voorhees and Buckley [53] examined the impact of topic set size on evaluation consistency; Sanderson and Zobel [41] studied the sign, Wilcoxon, and t-test; Faggioli and Ferro [12] compared different approaches for estimating ANOVA, with/without bootstrapping and different approaches to adjusting for multiple comparisons.

The definition of *agreement* between topic sets evolved over time. Zobel considered agreement when, on both topic sets, a system was significantly better than another and the sign of the difference was preserved. Moffat et al. [28], comparing different evaluation measures (not significance tests), identified five categories of agreement or disagreement that such a comparison could result in. One form, called SSA, required the same significant improvement to be found in both sets, recalling Zobel's definition; Moffat et al.'s categories were later on adopted by Faggioli and Ferro [12]. Urbano et al. [48] created five categories of agreement and the union of two categories – "Success" and "Lack of power" – aligns with Zobel's definition. Recently, Ferro and Sanderson [15] proposed a set of six categories which consider all the possible cases of agreement/disagreement and sign difference across the two topics sets and include all the previous definitions, somehow systematizing them. We will rely also on this approach in our experiments.

2.3.3 Simulation. Wilbur [56] exploits simulations of IR data to compare non-parametric and parametric tests, finding the former to be superior. Robertson and Kanoulas [35] develop a bootstrap based simulation approach to model the intra-topic variance, showing its effect when modelling IR performance. Urbano et al. [50] adopt a simulation process, previously defined by Urbano and Nagler [47], capable of jointly modelling both the system's internal variance and its covariance with another system via copulas. More recently, Parapar et al. [31, 32] model new runs as a stochastic process capable of simulating significantly different runs. Nevertheless, the proposed simulation process does not include the topic-system interaction.

Every simulation above produces IR-like data, more or less approximated depending on the underlying assumptions and the generation procedure. We do not use simulated data in our study to avoid possible biases due to either the approximation of the simulated data or inter-dependencies between the assumptions underlying simulations, on the one side, and GLMs, on the other side.

3 METHODOLOGY

3.1 Generalized Linear Models

Parametric statistical tests, such as t-tests or ANOVA, rely on the assumption that data can be modelled using a linear model. Focusing on the IR scenario, we typically have a set of m systems applied to a set of n topics. Given a system s and a topic t, we can compute a measure, e.g. AP, that quantifies how well s performs on t. To align with previous work in the GLMs domain, we refer to such a score as y_{ts} and call it *response*. The response y_{ts} is a realization of a random variable Y that represents the score achieved by a system on a topic. The experimental conditions – the topic and the system used in our case – are somehow correlated with the response, and therefore we refer to them as *covariates*. Using traditional linear models, the expectation of the response $E[Y]$, is modeled as a linear combination η of the covariates as follows:

$$E[Y] = \eta = \mu + \tau_1 t_1 + ... + \tau_n t_n + \alpha_1 s_1 + ... + \alpha_m s_m \quad (1)$$

where t_i and s_j are respectively the dummy coding variables for the topic and systems considered, τ_i is the effect due to the i-th topic, α_j is the effect due to the j-th system. The intercept μ represents the grand mean of our data.

When we instantiate Y to a real observation y_{ts}, we must include the error ε_{ts}, i.e. what the model is not able to explain:

$$y_{ts} = \mu + \tau_1 t_1 + ... + \tau_n t_n + \alpha_1 s_1 + ... + \alpha_m s_m + \varepsilon_{ts}$$

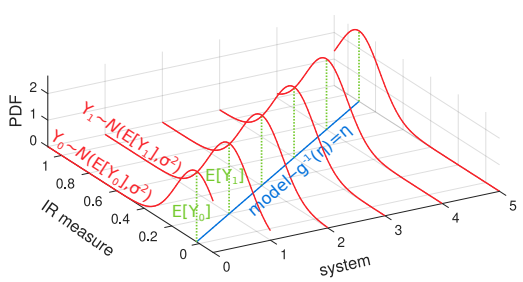
Under this framework, a t-test used to determine if system i is better than system j corresponds to verifying that the coefficient α_i is statistically significantly greater than α_j. Similarly, ANOVA is equivalent to check that at least one among the α coefficients is statistically significantly different from the others. In both cases, to compute the linear model and grant its inferences, we assume $Y \sim \mathcal{N}(\eta, \sigma^2)$ and thus $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ – Y distributes normally and has the same variance σ^2 everywhere (homoscedasticity).

Note that, without losing generality, we can say that we model $g(E[Y]) = \eta$, where g is the identity function $g(x) = x$. In this sense, g is the function that *links* $E[Y]$ to η. Summing up, fitting a linear model requires to define the following elements:

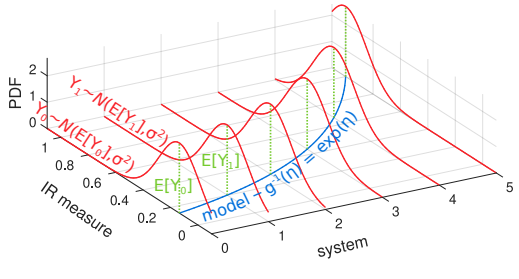
(1) a linear combination η of the different explanatory variables;
(2) a *link* function g to connect $E[Y]$ to η;
(3) a distribution for Y.

Under this framework, the traditional linear model GLiM is a particular case where, as aforementioned, g is the identity function, and Y distributes following a Gaussian distribution with expectation η and constant variance σ^2.

A visualization of an ideal linear model is depicted in Figure 1a. We assume to have a set of systems, each with a set of performance scores. For each system, the distribution of the scores is depicted in



(a) The "traditional" linear model: an identity link and a Gaussian distribution of the response. Red lines represent the distribution of the explained variable Y, while the blue line, the model, tries to describe how $E[Y]$ (green lines) changes.



(b) $E[Y]$ is not directly proportional to the system, thus a different link should be used: instead of modeling $E[Y]$ we model $g(E[y])$, where g is the log function. Y distributes still normally, only $E[Y]$ has changed.

Figure 1: Visual description of what changes when we change the link function in a GLM.

red. All the distributions are normal and homoscedastic. According to eq. (1) the focus of the linear modelling is the expectation of the explained variable, represented in Figure 1a by the green lines. Since all the expectations follow a straight line, we model the data using a traditional linear model (depicted in blue).

Compared to a traditional linear model, a GLM relaxes the assumptions for item 2 and 3. First, it models $g(E[Y])$, where g, the *link*, can be any monotonic continuous function. Secondly, the response Y can follow a distribution $f(\theta)$ that is not necessarily Gaussian. As a consequence, the *homoscedasticity* assumption is relaxed as well, since the variance can change with the expected mean. Thus, a GLM can be expressed in the following form:

$$g(E[Y]) = \eta, \text{ with } Y \sim f(\theta)$$

The chosen probability distribution $f(\theta)$ must be a member of the exponential distributions family. A location parameter θ characterizes distributions belonging to the exponential family - e.g., the normal distribution's mean. If we observe that $g(E[Y]) = \theta$ for a given distribution of Y, then we say that g is the *canonical link* of such a distribution. The canonical link has some advantages related to the optimization and the speed of convergence of the model parameters. Nevertheless, choosing which link function to use depends just on the data and their characteristics, often relying on empirical observation.

Figure 1b shows a scenario where a traditional linear model is not suited anymore, since $E[Y]$ does not follow a straight line We therefore resort to use GLMs. As shown in Figure 1a, the performance distributes normally with equal variance for all the systems, but the expectation $E[Y]$ appears to follow an exponential line. Therefore, to bring it back to a linear space, we can transform $E[Y]$ using the log link. Thus, our model becomes $\log(E[Y]) = \eta$ or, equivalently $E[Y] = \exp(\eta)$.

Notice that, in Figure 1b, data are not transformed, only the *link* between the expectation of the response and the model is. If data were transformed, the Y axis, as well as the response y_{ts} predicted by the model, would have changed. On the contrary, the response remains on the same scale: what changes is just the model's shape. Therefore, fitting a GLM is substantially different from transforming the response in a non-linear space, as the approaches discussed in Section 2.1 do instead. Indeed, when we apply a non-linear transformation g directly to the response Y, we assume that, in the new space, $E[g(Y)]$ is linearly correlated with the covariates and $g(Y)$ follows a normal homoscedastic distribution. In this sense, $g(Y)$ should comply with the linear modelling assumptions, as well. When instead, as in the case of our work, we choose to use a GLM, we believe that the linear correlation is between the predictors and the transformation $g(E[(Y)])$. Transforming the response also means that predictions y_{ts} – and errors ε_{ts} – are in the transformed space and no more directly comparable; on the other hand, when using a GLM, predictions and errors remain in the original scale. As a final observation on the difference between transforming the response and using a GLM, we can note that, in general:

$$\underbrace{g(E[Y])}_{\text{use of a GLM}} < \underbrace{E[g(Y)]}_{\text{response transformation}}$$

To carry out statistical inference and test whether a system is significantly different from another, we need two elements: *i)* the difference between the effects of the systems *ii)* and the *Standard Error (SE)* associated with their comparison. Both elements rely on the concept of *contrasts* [20]. A contrast is a linear combination of the coefficients of a linear model using a vector $\mathbf{c}$, where $\mathbf{c} \in \mathbb{R}^{k\times 1}$, with k the number of possible coefficients and $\sum_j \mathbf{c}(j) = 0$. Contrasts allow to model comparisons between (groups of) factors. Each contrast corresponds to a specific hypothesis that we are interested in testing. For example, in IR we are usually interested in carrying out pairwise comparisons between systems. In such a case, the contrast vector to compare i-th and j-th systems is:

$$\mathbf{c}_{ij}(h) = \begin{cases} 1, & \text{if } h = i \\ -1, & \text{if } h = j \\ 0, & \text{otherwise} \end{cases}$$

Then, calling $\boldsymbol{\alpha}$ the systems coefficients vector, we can compute the pairwise difference between effects as $\Delta_{ij} = \mathbf{c}_{ij} \cdot \boldsymbol{\alpha}$. Using the procedure mentioned above, we can define all the pairwise contrasts and obtain all the differences between pairs of systems.

The SE for a pairwise contrast between coefficients α_i and α_j is computed as:

$$SE_{ij} = \hat{\sigma}^2(\alpha_i) + \hat{\sigma}^2(\alpha_j) - 2\hat{\rho}(\alpha_i, \alpha_j) \tag{2}$$

where $\hat{\sigma}^2(\alpha_i)$ is the variance associated with the coefficient α_i (which should not be confused with the sample variance of the scores observed for system i). Similarly, $\hat{\rho}(\alpha_i, \alpha_j)$ is the covariance

Table 1: Link functions considered. Φ and Cauchy are the *Cumulative Density Function (CDF)* of a Standard Normal and Cauchy Distribution respectively.

name	function	inverse
identity	$g(x) = x$	$g^{-1}(x) = x$
log	$g(x) = \log(x)$	$g^{-1}(x) = e^x$
exp	$g(x) = e^x$	$g^{-1}(x) = \log(x)$
tanh	$g(x) = \tanh(x)$	$g^{-1}(x) = \text{arctanh}(x)$
logit	$g(x) = \log \frac{x}{1-x}$	$g^{-1}(x) = \frac{1}{1+e^{-x}}$
probit	$g(x) = \Phi^{-1}(x)$	$g^{-1}(x) = \Phi(x)$
cauchit	$g(x) = \text{Cauchy}^{-1}(x)$	$g^{-}1(x) = \text{Cauchy}(x)$

between the coefficients α_i and α_j. These values can be obtained from the covariance matrix. The (asymptotic) covariance matrix for a GLM is the inverse of the negative of the matrix of the second derivative of the log-likelihood function.

Once we have the SE for each contrast, to test if the performance of systems i and j are different, we compute our test statistics as:

$$t_{ij} = \frac{\Delta_{ij}}{SE_{ij}} \tag{3}$$

t_{ij} can be compared to the proper critical value according to the chosen distribution or used to obtain the p-value. This is a generalization of the traditional t statistics and can be used to carry out several inferential tests, including t-test, ANOVA, and F-test. By comparing t_{ij} with the proper value Q from the Studentized range distribution, we can carry out the Tukey's *Honestly Significant Difference (HSD)* [46] test, correcting for the multiple comparisons problem.

3.2 Using GLM in IR scenarios

As pointed out in the previous section, to fit GLMs, we need to select the link function and the response distribution. Any possible monotonic continuous function can be a suitable link. The choice of which link to use depends on the shape of the data. In our analyses, we focus on the most popular link functions, reported in Table 1. Notice that the inverse of the link describes how the expectation of the response changes. For example, as shown in Figure 1b, the log link suits scenarios where $E[Y]$ appears to follow an exponential pattern. We include the log, exponential and hyperbolic tangent (tanh) functions in our experiments. We also experiment with a series of sigmoidal functions: logit, probit and cauchit. Such functions have similar shapes with different steepness and are typically used for data that can be interpreted as probabilities.

Previous works on transforming AP [4, 10, 34, 35] observed that the logit transformation renders the score distribution more normal but it has the drawback of making observations for which AP is zero or one unusable. GLMs based on the logit link, on the other hand, by transforming the expectation of the response rather than the response itself, avoid such corner cases. However, when even the expectation of a system's performance is close to zero, using log-based links – e.g., log and logit – determines a high variance of the coefficients associated with such a system. As a consequence, according to eq. (2), a larger variance increases the standard error which, in turn, decreases the test statistics of eq. (3) used for comparing systems; overall this causes us to detect fewer significantly different pairs. In the literature, it is therefore suggested to remove outliers with close-to-zero expected performance.

Concerning the distribution, as aforementioned, we are limited to distributions drawn from the exponential family which includes Gaussian, Bernoulli and Binomial, Poisson, Gamma, and Inverse Gaussian. Except for the Gaussian, which has $\mathbb{R}$ as domain, all the other distributions have a domain which differs from the one of IR measures. The Binomial is defined over the natural numbers, up to a given threshold – the Bernoulli is a special case, where the threshold is 1. The Poisson distribution is defined over $\mathbb{N}$. Finally, both the Gamma and Inverse Gaussian are defined on $\mathbb{R}^+$, therefore excluding 0, which is a possible value for most of the IR measures. By adequately changing the IR measure, it might be possible to use different distributions besides the Gaussian. Nevertheless, as observed by [7, 22, 43, 45], most of the tests are typically resilient to the violation of the normality assumption. Furthermore, we are interested in investigating the impact of the links alone. Therefore, we focus on the Gaussian distribution, leaving the study of other distributions to future work.

Finally, to investigate the GLMs, we use a series of models always consisting of the system and topic factors but changing the links. We leave models comprising more factors, e.g. shards, as future work.

4 EXPERIMENAL APPROACH

4.1 Deviance

GLiMs are commonly fit using the *Ordinary Least Squares (OLS)* approach that minimizes the *sum of squares of residuals (RSS)* and this allows for comparing different models by their RSS. On the other hand, GLMs are fit using maximum likelihood instead of OLS and, therefore, comparing the RSS is not suitable.

The most common goodness-of-fit statistics under the GLM framework is the *deviance* [26], which is analogous to RSS under the GLiM framework. Deviance is defined as:

$$D = 2 * (LL_s - LL_m)$$

where LL_s is the log-likelihood of the saturated model – a model with a parameter for each observation – and LL_m is the log-likelihood for the fitted model. Similarly to the RSS, the lower the deviance, the better a model fits the data.

4.2 Number of Significantly Different Pairs

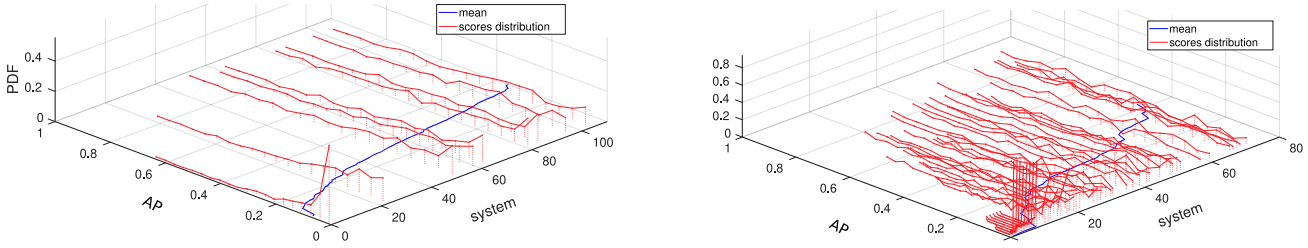
As done in previous work (see Section 2.3.1), we consider the total number of ssd pairs as a first indicator of the ability of a model to distinguish among systems. In general, the higher, the better.

From a practitioner perspective, being able to identify more ssd pairs is essential: correctly individuating which system performs better allows us to invest on more promising solutions that might have been discarded otherwise.

4.3 Topic Splitting

We consider the agreement measures defined by Ferro and Sanderson [15], following also previous works [12, 28, 48] (see Section 2.3.2).

We assume to have two non-overlapping splits of topics, a statistical test based on a given link, and a pair of systems - $S1$ and $S2$. According to the decisions taken on each topic subset, we have the following possibilities: Active (A-) decisions – the test considers the difference between $S1$ and $S2$ statistically significant



(a) robust 04 - Systems **(b) core 18 - Systems**

Figure 2: Figures show that, for both the Robust 04 and the core 18 collection *i)* the distributions of the AP scores (red lines) are not normal for a given system (and they hardly have the same variance) *ii)*, the expectation is not linear.

on both topic sets; Passive (`P-`) decisions – none of the topic sets provides enough evidence to determine if S_1 is statistically better than S_2; Mixed (`M-`) decisions – only one of the topic sets allows to say that S_1 is statistically better than S_2. Furthermore, it might be possible that the topic sets agree (Agreement `-A`) or disagree (Disagreement `-D`) on considering S_1 to have a greater effect than S_2. This classification determines six possible scenarios: *Active Agreements (AA)*; *Active Disagreements (AD)*; *Passive Agreements (PA)*; *Passive Disagreements (PD)*; *Mixed Agreements (MA)*; *Mixed Disagreements (MD)*. The more AA decisions, the more a test is able to distinguish among systems and the more consistent is its outcome about what is significantly different. Conversely, AD indicates opposite inference on what is significantly different and represents the worst outcome. Both MA and MD indicate that a test is not able to confirm its conclusions and, therefore, the smaller the better; MD is a bit more severe than MA, since there is also swap in the order of the two systems, but not as severe as AD. PA and PD both indicate that a test confirms its outcomes on what is not significantly different, even if in the latter case there is also a swap between the two systems.

Following Ferro and Sanderson [15], we also consider *Bias* as the likelihood of a researcher publishing a significant result when in fact a significance test on a different topic set would have produced either no significance (MA, MD) or a significant result in the opposite direction (AD):

$$Bias = 1 - \frac{AA}{AA + AD + \frac{MA}{2} + \frac{MD}{2}}$$

Note that *Bias* represents an error, thus we subtract the fraction from 1.

5 EXPERIMENTAL ANALYSIS

5.1 Experimental Setup

We consider two collections for *ad-hoc* retrieval: TREC 13 Robust 04 [52] and TREC 27 Core 18 [1]. Robust 04 relies on disks 4 and 5 of the TIPSTER corpus minus the Congressional Records, has 249 topics and 110 runs, amounting to 5,995 pairwise comparisons. Core 18 relies on the Washington Post document collection, has 50 topics and 72 runs, amounting to 2556 pairwise comparisons.

As discussed in Section 3.2, systems whose mean performance is very close to 0 can challenge the logit link. Since this happens in the case of Core 18, we also consider a second version of it, where we remove eight outlier runs, performing extremely low in terms of their *Mean Average Precision (MAP)*. Following Laurikkala et al. [24], we define "outliers" those runs having MAP 1.5 times the inter-quartile range lower than the lowest quartile. We dub the reduced version of Core 18 *without outliers* "Core 18-wo". It has 64 runs that lead to 2016 pairwise comparisons.

All the collections have ternary relevance judgements with possible values {0, 1, 2}, that indicate respectively, not relevant, partially relevant and highly relevant documents. As performance measures, we use *Average Precision (AP)*, Precision with cutoff 10 (P@10), Recall (R), *Normalized Discounted Cumulated Gain (nDCG)* [23], and RBP with persistence of 0.95 [27]. In Section 5.5, we repeat the topic sampling 1000 times. The code is publicly available to allow for reproducibility[1].

5.2 Fitting Linearly IR Data

Figure 2 illustrates what happens when we plot the data from both the Robust 04 (Figure 2a on the left) and the Core 18 (Figure 2b on the right) collection. We plot the MAP, i.e. the expectation of the response, for each system using a blue line while the red lines represent the distribution of the AP scores of a system; we display just a subset of available systems for the sake of visualization.

By looking at Figure 2, we can note that, for both Robust 04 and Core 18, the blue line representing the expected performance is not straight. As a consequence, the identity link may not be the most appropriate to describe the underlying the IR data. Moreover, as also noted in the literature, the distributions of the observations (red lines) are far from being normal. Therefore, a GLM, thanks to its relaxed assumptions, might better fit IR data[2].

When it come to Core 18 in Figure 2b, besides the general behaviour already discussed, we can note how the eight outliers, all submitted by the same group, have a completely different distribution from the other runs, extremely skewed towards low performance, and this explains why they have been removed in Core 18-wo version.

[1] To be released upon acceptance

[2] For the sake of completeness, the actual linear model fitted on the IR data is more complex. There is a dimension for each system and thus we have a hyper-plane instead of a single "blue line". Nevertheless, this representation gives the idea of how far we are from the ideal scenario to apply a linear model.

Table 2: Deviance. In bold minimum value, i.e. best fit. The color indicates optimal (green), average (white) or low (red) results.

	Robust 04					Core 18					Core 18-wo				
link	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP
identity	356.79	901.51	622.67	467.98	381.62	52.48	156.32	99.05	69.21	75.90	44.93	135.61	74.46	59.89	63.23
log	334.06	886.42	646.34	471.87	368.30	40.90	**132.00**	91.57	61.73	**59.00**	40.76	**128.12**	75.36	60.18	**57.76**
exp	387.52	955.56	719.93	505.65	400.98	61.31	219.01	159.33	90.92	98.80	49.19	160.79	79.90	64.29	72.55
tanh	348.91	894.19	640.07	467.54	376.76	50.30	147.13	95.89	66.00	71.15	43.81	132.13	75.84	59.80	61.00
logit	**329.46**	**882.14**	593.82	**458.04**	**366.89**	**40.50**	132.77	**85.47**	**60.51**	59.44	**40.36**	128.52	72.21	**58.93**	58.12
probit	330.36	882.66	**590.87**	458.05	367.65	40.71	133.26	85.52	60.63	59.81	40.57	128.84	**72.13**	58.99	58.42
cauchit	332.49	884.67	627.34	463.81	367.40	40.87	132.24	86.80	60.63	59.03	40.73	128.42	74.05	59.05	57.81

5.3 Deviance

Table 2 illustrates the deviance measured for different GLMs using several link functions, IR measures, and experimental collections.

The traditional GLiM approach based on the *identity link*, (corresponding to the current evaluation methodology) presents a low goodness-of-fit, given its high deviance compared to other links. This evidence supports the idea of investigating and using GLMs instead.

The *exponential link* is the worst, systematically underperforming on all experimental conditions and its high deviance indicates poor goodness-of-fit compared to all the other links. Its poor capability in fitting IR data leads to overall instability, especially concerning shallow performing systems, and to convergence problems when fitting the model – highlighted by the increased iterations to reach convergence, not reported here due to space constraints. The degraded performance seems to be correlated with the presence of low performing systems; in fact, when we consider Core 18-wo, the degradation in terms of goodness-of-fit due to the exponential link, still being the worst one, is lower than the Core 18 where the eight exceedingly low performing runs are present instead. Given the unsuitable behaviour of the exponential link, we exclude it from further experimentation.

The *log link* shows improved goodness-of-fit compared to identity one, especially for the Core 18 collection (both with and without outliers) when adopted with measures not depending on the recall base – P@10 and RBP –, where it appears to be the most suited model, given its minimum deviance. The *tanh link* exhibits an intermediate behaviour in all scenarios: it appears slightly better than the identity without providing substantial improvements. The *logit link* has the best goodness-of-fit in most cases, achieving the lowest deviance. Logit, *probit* and *cauchit links* tend to perform quite similarly. This behaviour is somehow expected since their shapes are overall very similar.

5.4 Number of Significantly Different Pairs

Table 3 contains the number of *statistically significantly different (ssd)* pairs detected by the GLMs based on different links, using Tukey's HSD [46] test.

On Robust 04 collection, all the links outperform the traditional modelling strategy – i.e., identity link – using AP, P@10, and RBP as performance measures. Logit is the best-performing link: it detects 7.9%, 8.9%, and 7.0% more ssd pairs compared to identity, when regarding AP, P@10, and RBP, respectively. On the other hand, considering Recall and nDCG, log and tanh fail to identify more pairs than identity, while logit, probit and cauchit increase the number of ssd pairs found. nDCG tends to be the measure that benefits the least from the new links, with the logit link providing only 1.9% more pairs. As a general consideration, the fact that logit, probit, and cauchit obtain good results suggests that their sigmoidal shape is well suited to model IR data.

Concerning the Core 18 collection, Table 3 confirms what we pointed out in Section 3.2 about log-based links: when used in presence of low-performing outliers, they tend to underperform compared to the identity link. In particular, we observe that log, logit and cauchit almost always fail to outperform the identity baseline. Indeed, in our specific case almost all the ssd pairs lost with respect to the identity link correspond to the eight outlier runs[3] submitted by a single group. As shown also in Figure 2b, these runs have extremely low mean performance, being their MAP between 0.003 and 0.007. As discussed in Section 3.2, the lower the expectation of the response, the higher is the variance of the coefficients, and

[3] 8 runs appear in 540 pairwise comparisons.

Table 3: Number of statistically significantly different pair. In bold maximum value. The color indicates optimal (green), average (white) or low (red) results.

	robust 04 - (5995 systems pairs)					core 18 - (2556 systems pairs)					core 18-wo - (2016 systems pairs)				
link	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP
identity	3427	2347	3848	3704	2837	1210	1054	1115	1270	1247	789	596	427	786	803
log	3556	2383	3622	3550	2946	925	934	672	1097	1220	878	635	384	748	941
tanh	3509	2354	3639	3641	2905	**1301**	**1130**	1086	1283	**1361**	843	633	380	766	892
logit	**3700**	**2557**	4018	**3773**	**3035**	976	1034	1267	1251	1257	**926**	**713**	594	818	929
probit	3693	2541	**4027**	3766	3034	974	1079	**1304**	**1340**	1341	**926**	710	**597**	815	928
cauchit	3682	2552	3929	3764	3016	848	739	877	872	968	796	**713**	**597**	**823**	**937**

Table 4: Agreement over two topic sets for different links.

		125			50			25			10		
collection		identity	log	logit	identity	log	logit	identity	log	logit	identity	log	logit
robust 04	AA	2332.45	2491.36	2616.35	1229.23	1415.57	1542.27	489.06	533.35	705.29	62.22	12.49	65.08
	MA	633.34	675.22	703.57	665.41	766.20	797.01	616.21	787.56	836.24	280.01	214.32	454.29
	PA	2593.33	2345.11	2215.39	3399.18	3027.13	2907.45	3909.39	3594.09	3415.37	4236.23	4246.35	3993.31
	PD	435.20	479.27	455.52	700.32	780.13	741.49	979.24	1073.27	1028.08	1415.34	1517.12	1470.24
	MD	1.07	4.44	5.36	1.26	6.38	7.17	2.29	7.53	10.42	2.39	5.12	13.28
	AD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Bias	0.120	0.120	0.120	0.213	0.214	0.207	0.387	0.427	0.375	0.694	0.898	0.782
core18-wo	AA	—	—	—	—	—	—	300.12	247.56	318.50	66.04	11.57	43.23
	MA	—	—	—	—	—	—	268.29	362.37	410.22	163.35	116.39	207.22
	PA	—	—	—	—	—	—	1161.31	1138.24	1013.13	1386.25	1481.43	1360.08
	PD	—	—	—	—	—	—	286.04	266.10	271.49	398.21	404.59	401.16
	MD	—	—	—	—	—	—	1.04	2.12	3.05	2.15	2.03	5.51
	AD	—	—	—	—	—	—	0.00	0.00	0.00	0.00	0.00	0.00
	Bias	—	—	—	—	—	—	0.310	0.424	0.393	0.556	0.837	0.711

thus the standard error, causing a reduction in the number of ssd pairs detected. Concerning Recall, both logit and probit yield more ssd pairs than the identity. The mean Recall for the eight outliers ranges between 0.01 and 0.02, one order of magnitude greater than in the previous case of AP. These runs are still outliers compared to the rest of the distribution, but they are not so close to zero to cause issues for the logit and probit links. Finally, probit, even though similar to logit and cauchit, outperforms the identity on all the measures except AP. This might be due to the shape of the link functions. The cauchit function is the steepest and thus the most vulnerable to outliers. Logit function has intermediate steepness, exhibiting medium vulnerability to outliers. Finally, probit is the least steep and the more resilient to outliers.

Finally, we consider Core 18-wo collection, where we removed the eight outlier runs: all the new links obtain a consistent improvement over the identity baseline for what concerns AP, P@10, and RBP. Logit and probit links are the best, gaining 17.4% new pairs on the AP. Logit, probit and cauchit perform well also with Recall and nDCG. Similarly to Robust 04, both tanh and log links lose several ssd pairs with respect to identity when using Recall and nDCG.

Overall, nDCG is the measure that benefits the least from the new links, gaining only 5% of pairs at most. This lower increase in ssd pairs found is likely due to the distribution of the nDCG scores. The plots for the nDCG like those of Figure 2, omitted for space reasons, tend to be closer to the assumptions of the linear model compared to other measures. The lower increase in the number of ssd pairs thus highlights two insights: i) nDCG violates less the assumptions underlying linear models; ii) other measures, by departing more from the assumptions, produce worse comparisons and thus GLMs help in mitigating this phenomenon.

In the next two sections, for space reasons, we focus our analyses just on AP and the log and logit links, being their behaviour consistent with the other cases.

5.5 Topic Splitting

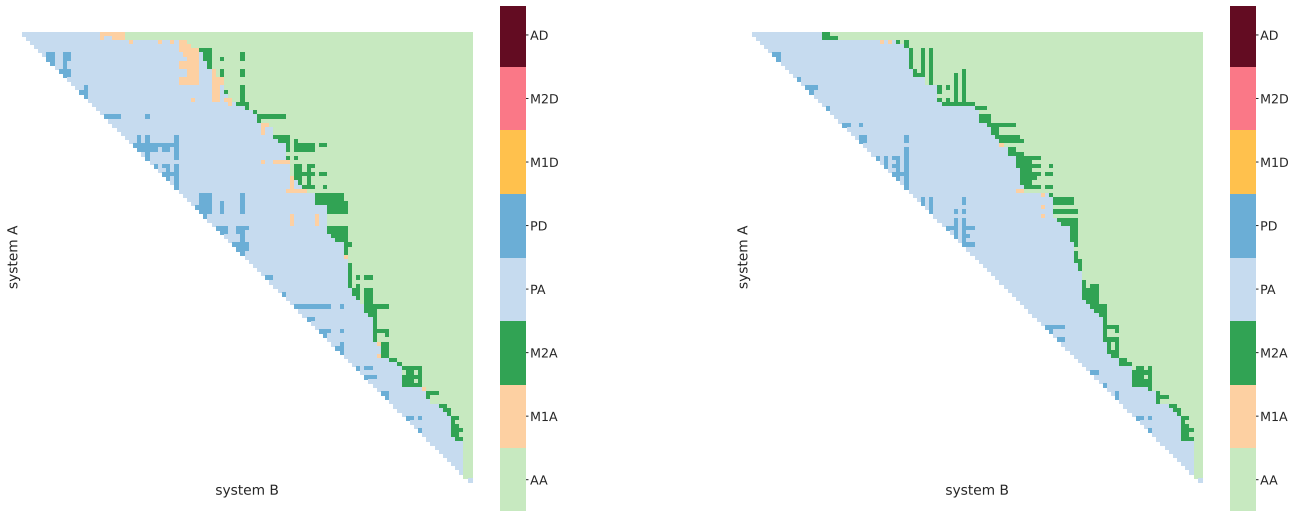
We consider the following topic set sizes {125, 50, 25, 10}: all of them in the case of Robust 04; only {25, 10} for Core 18, since it consists of 50 topics only. For each size, we re-sample the topic sets 1000 times. In Table 4, we report average performance over these samples.

For all the links, the number of AA decreases as the topic set size decreases, since the less evidence available causes the total number of ssd pairs to decrease; for example, for logit on Robust 04, it drops from 43.64% of the total ssd pairs at 125 topics to 1.09% at 10 topics. Logit always has the highest AA, with the exception of Core 18-wo when using 10 topics. The increments range between 4.60% and 44.21% with respect to the identity link; in particular, with 50 topics, the typical size adopted in experimentation, logit gains 25.47% more pairs than identity and AA is 25.73% of the total ssd pairs. The behaviour of the log link is mixed: on Robust 04 it gains with respect to the identity between 125 and 25 topics, achieving 15.16% more pairs at 50 topics; on the other hand, at 10 topics it loses 79.93% of the pairs with respect to identity (same behaviour on Core 18-wo). The logit always performs better than the log, gaining between 5% and 32% more pairs for 125 and 25 topics on Robust 04, and an astonishing 421% more pairs at 10 topics.

For all the links, AD is zero under all circumstances, indicating that all the models never reach severely inconsistent conclusions. Both the logit and the log links have more MA pairs than the identity one, the logit being a bit higher than the log. For example, the logit achieves between 11.09% and 62.24% more pairs than identity on Robust 04, ranging between 7.58% and 13.95% of the total ssd pairs. This indicates that they need less evidence to consider two systems as ssd but that this may also lead to inconsistencies. When it comes to MD, the number of pairs is generally very low, between 0.02% and 0.22% of the total ssd pairs, but logit and log have between 2 and 4 times more than identity, suggesting possible inconsistencies.

"Not significant" decisions (PA and PD) tend to be more frequent for the identity link in most cases, being it the most conservative. In this regard, by making fewer decisions, the identity link is the most likely to prevent false positives, but also the one with the most considerable risk of incurring false negatives.

Let us now consider the significance indicators all together and examine the risk of publication bias when using the different models. We can observe that on Robust 04 all the links behave very similarly,



(a) Identity vs log link. **(b) Identity vs logit link.**

Figure 3: Comparison between decisions taken by different links. Each square represents a pair of systems. System A (y axis) has a higher MAP than System B (x axis). M1* and M2* indicate a significant decision taken only by the identity link and log or the logit links respectively.

the logit being just a bit better (-2.82% bias) than identity at 50 topics and 25 topics (-3.10% bias). On the other hand, at 10 topics, as well as on Core 18-wo, identity performs better than logit, achieving between -12.62% and -27.88% bias. It should be noted, however, that small topic set sizes (25 and 10) incur in quite high bias for all the models, between 0.31 and 0.90, confirming, as known from the literature, that they are insufficient sizes.

Overall, we can conclude that at typical topic set sizes, i.e. 50 (or more) topics, the logit and log links provide a sizeable improvement in the number of AA pairs with a comparable risk of publication bias with respect to the identity link.

5.6 Visual Comparison of the Decisions Taken

Figures 3a and 3b compare the identity link and log and logit ones respectively on Robust 04 using all the topics. Note that while in the topic splitting case we compare the same model on two different topics sets and breakdown its "agreement" into the different counts, here we compare two different models on the same topic set (the whole corpus) and we use the previous count to analyze their "agreement". The axes represent all the possible systems sorted by MAP, comparing each system against all the systems that performed worse; therefore, only the upper triangular area of the matrix is coloured. The cell color indicates the kind of agreement/disagreement in the decisions taken. We do not observe any MD or AD, being the majority of the observations either "not significant" decisions (PD and PA) or AA.

Looking at Figure 3a, the identity link identifies some more ssd pairs in the upper part of the system ranking (M1A, light orange). On the other hand, the log link gains several ssd pairs for middle and low performing systems (M2A, dark green). This behaviour of the log link might be useful to identify which system performs better on a set of particularly hard queries.

When it comes to logit in Figure 3b, it identifies more ssd pairs (M2A dark green) across all tiers of system performance and it also distinguishes better among the top-tier systems.

6 CONCLUSIONS AND FUTURE WORK

We studied *Generalized Linear Models (GLMs)*, an extension of the traditional linear models typically used in IR evaluation to compare systems. GLMs overcome the main reasons of departure of IR data from assumptions underlying linear models: non-normality and heteroscedasticity of the data and non-linearity of the empirical mean. In this work, we focused on the latter and studied how to address it using different *link functions*, since the former two tend to have less severe impact.

We proposed a new visualization of the IR data, capable of highlighting the empirical link and the performance distribution. When it comes to the links, using different evaluation measures and collections, our experiments indicate that the exp link should be avoided, because of its worse fitting to the data while the log, logit, tanh, prob, and probit provide general improvements with respect to the identity link used today. We then dug into the log and logit links, which were the most promising ones, and we found out that they are able to detect a sizeably greater number of consistent ssd pairs than the identity link with a comparable, or slightly better, risk of publication bias. In particular, the logit link delivers these improvements all over the range of system performance while the log link is a bit more focused on the middle and lower range. On the other hand, some care has to be put in using these new links, since systems with average performance exceedingly close to zero hamper their functioning and should be removed.

As future work, we plan to regard distributions other than the Gaussian one used in this work, in order to deal with both the non-normality and the heteroscedasticity of the data.

ACKNOWLEDGMENTS

The work was partially supported by University of Padova Strategic Research Infrastructure Grant 2017: "CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione".

REFERENCES

[1] J. Allan, D. K. Harman, E. Kanoulas, and E. M. Voorhees. TREC 2018 Common Core Track Overview. In *The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018)*, 2019.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User Variability and IR System Evaluation. In *Proceedings of the 38th ACM SIGIR Conference on Research and Development on Information Retrieval*, pages 625–634, 2015.

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

[4] A. Berto, S. Mizzaro, and S. Robertson. On Using Fewer Topics in Information Retrieval Evaluations. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, page 30–37, 2013.

[5] D. Bodoff and P. Li. Test Theory for Assessing IR Test Collections. In *Proceedings of the 30th the ACM SIGIR International Conference on Theory of Information Retrieval*, SIGIR '07, page 367–374, 2007.

[6] B. Carterette. Model-Based Inference about IR Systems. In *Proceedings of the Third International Conference on Advances in Information Retrieval Theory*, ICTIR'11, page 101–112, 2011.

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

[8] B. Carterette. Bayesian Inference for Information Retrieval Evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, page 31–40, 2015.

[9] B. Carterette. But Is It Statistically Significant? Statistical Significance in IR Research, 1995-2014. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1125–1128, 2017.

[10] G. V. Cormack and T. R. Lynam. Statistical Precision of Information Retrieval Evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 533–540, 2006.

[11] J. S. Culpepper, G. Faggioli, N. Ferro, and O. Kurland. Topic difficulty: Collection and Query Formulation Effects. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–36, 2021.

[12] G. Faggioli and N. Ferro. System Effect Estimation by Sharding: A Comparison between ANOVA Approaches to Detect Significant Differences. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, editors, *Advances in Information Retrieval. Proc. 43rd European Conference on IR Research (ECIR 2021) – Part II*, pages 33–46. Lecture Notes in Computer Science (LNCS) 12657, Springer, Heidelberg, Germany, 2021.

[13] 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.

[14] N. Ferro and M. Sanderson. Improving the Accuracy of System Performance Estimation by Using Shards. In *Proceedings of the 42nd ACM SIGIR Conference on Research and Development on Information Retrieval*, SIGIR '19, pages 805–814, 2019.

[15] N. Ferro and M. Sanderson. How do you Test a Test? A Multifaceted Examination of Significance Tests. In K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, and J. Tang, editors, *Proc. 15th ACM International Conference on Web Searching and Data Mining (WSDM 2022)*, pages 280–288. ACM Press, New York, USA, 2022.

[16] N. Ferro and G. Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In *Proceedings of the 39th ACM SIGIR Conference on Research and Development on Information Retrieval*, SIGIR '16, pages 25–34, 2016.

[17] N. Ferro and G. Silvello. Toward an anatomy of IR system component performances. *jasist*, 69(2):187–200, 2018.

[18] 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.

[19] N. Fuhr. Some common mistakes in ir evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, February 2018.

[20] J. Hsu. *Multiple comparisons: theory and methods.* CRC Press, 1996.

[21] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, page 329–338, 1993.

[22] P.K. Ito. 7 robustness of anova and manova test procedures. In *Analysis of Variance*, volume 1 of *Handbook of Statistics*, pages 199–236. 1980.

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

[24] J. Laurikkala, M. Juhola, E. Kentala, N. Lavrac, S Miksch, and B. Kavsek. Informal Identification of Outliers in Medical Data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, volume 1, pages 20–24, 2000.

[25] H. Madsen and P. Thyregod. *Introduction to General and Generalized Linear Models.* CRC Press, 2010.

[26] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Springer, 1989. ISBN 978-1-4899-3242-6. doi: 10.1007/978-1-4899-3242-6. URL https://doi.org/10.1007/978-1-4899-3242-6.

[27] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27, 2008.

[28] A. Moffat, F. Scholer, and P. Thomas. Models and Metrics: IR Evaluation as a User Process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, ADCS '12, page 47–54, 2012.

[29] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson. *Generalized Linear Models: with Applications in Engineering and the Sciences*, volume 791. John Wiley & Sons, 2012.

[30] J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238.

[31] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, and A. Barreiro. Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 71 (1):98–113, 2020.

[32] J. Parapar, D. E Losada, and Á. Barreiro. Testing the tests: simulation of rankings to compare statistical significance tests in information retrieval evaluation. In C.-H. Hung, J. Hong, A. Bechini, and E. Song, editors, *Proc. 36th ACM/SIGAPP Symposium On Applied Computing (SAC 2021)*, pages 655–664. ACM Press, New York, USA, 2021.

[33] S. Robertson. On GMAP: And Other Transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, page 78–83, 2006.

[34] S. Robertson. On Smoothing Average Precision. In *Advances in Information Retrieval*, pages 158–169, 2012.

[35] S. E. Robertson and E. Kanoulas. On Per-Topic Variance in IR Evaluation. In *Proceedings of the 33rd ACM SIGIR Conference on Research and Development on Information Retrieval*, SIGIR '12, page 891–900, 2012.

[36] A. Rutherford. *Introducing ANOVA and ANCOVA: a GLM approach.* Sage, 2001.

[37] T. Sakai. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 525–532, 2006.

[38] T. Sakai. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 5–14, 2016.

[39] T. Sakai. A Simple and Effective Approach to Score Standardisation. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 95–104, 2016.

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

[41] M. Sanderson and J. Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. 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 162–169. ACM Press, New York, USA, 2005.

[42] T. Saracevic. Comparative Systems Laboratory Final Technical Report, An Inquiry into Testing of Information Retrieval Systems Part II: Analysis of Results. Technical report, Case Western Reserve University, USA, 1968.

[43] S. M. Scariano and J. M. Davenport. The Effects of Violations of Independence Assumptions in the One-Way ANOVA. *The American Statistician*, 41(2):123–129, 1987.

[44] M. D. Smucker, J. Allan, and B. Carterette. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings of the 16th ACM Conference on Information and knowledge Management*, CIKM '07, pages 623–632, 2007.

[45] J. M. Tague-Sutcliffe and J. Blustein. A Statistical Analysis of the TREC-3 Data. In *Proceedings of The 3rd Text REtrieval Conference*, TREC '94, pages 385–398, 1994.

[46] J. W. Tukey. Comparing Individual Means in the Analysis of Variance. *Biometrics*, pages 99–114, 1949.

[47] J. Urbano and T. Nagler. Stochastic Simulation of Test Collections: Evaluation Scores. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 695–704, 2018.

[48] J. Urbano, M. Marrero, and D. Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *Proc. SIGIR*, page 925–928, 2013.

[49] J. Urbano, H. Lima, and A. Hanjalic. A New Perspective on Score Standardization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1061–1064, 2019.

[50] J. Urbano, H. Lima, and A. Hanjalic. Statistical significance testing in information retrieval: An empirical analysis of type i, type ii and type iii errors. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 505–514, 2019.

[51] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, England, 2nd edition, 1979.

[52] E. M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. In *Proceedings of The 13th Text REtrieval Conference*, TREC '13, 2004.

[53] E. M. Voorhees and C. Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. Hyon Myaeng, editors, *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 316–323. ACM Press, New York, USA, 2002.

[54] 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, 2017.

[55] W. Webber, A. Moffat, and J. Zobel. Score Standardization for Inter-Collection Comparison of Retrieval Systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 51–58, 2008.

[56] W. John Wilbur. Non-parametric Significance Tests of Retrieval Performance Comparisons. *Journal of Information Science*, 20(4):270–284, 1994.

[57] J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments. 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 307–314. ACM Press, New York, USA, 1998.