

Multiple Testing in Remote Sensing: Addressing the Elephant in the Room

Oliver Gutiérrez-Hernández^{1*}, Luis V. García²

¹ Department of Geography, University of Málaga. Bulevar Louis Pasteur 27, 29010 Málaga, Spain. Mail: olivergh@uma.es. Tel.: +34 951953293. * Corresponding author.

² Institute of Natural Resources and Agrobiology of Seville (IRNAS-CSIC), Spanish National Research Council (CSIC). Av. Reina Mercedes 10, 41012 Seville, Spain. Mail: lv.garcia@csic.es Tel.: +34954624711.

1. Introduction.

Supplementary Material II includes R code for applying the Benjamini–Hochberg (FDR-BH) and Benjamini–Krieger–Yekutieli (FDR-BKY) procedures to raster-based datasets, using a single raster layer of p -values as input. To support understanding of how these corrections are computed, this document provides a step-by-step explanation designed to clarify the logic of both procedures transparently and intuitively.

2. From the multiple testing problem to false discovery rate (FDR) control

Consider a multiple testing situation in which many statistical tests are performed (Table 1), as commonly occurs in remote sensing, where thousands or even millions of tests are conducted simultaneously, typically one per pixel across the image. Traditional procedures usually aim to control the family-wise error rate (FWER), which is defined as the probability of making at least one false rejection. While effective at limiting false positives, this criterion becomes too strict when the number of tests is large, often resulting in very few rejections and many missed detections (García 2004). To address this limitation, Benjamini and Hochberg (1995) proposed the false discovery rate, which instead focuses on controlling the expected proportion of false positives among all rejected hypotheses. This approach provides a balanced compromise between avoiding false alarms and retaining sensitivity to real effects (García 2003), making it especially suitable for large-scale spatial multiple testing (Sun et al. 2015).

Table 1. Cross-classification of hypothesis testing outcomes.

	H_0 not rejected	H_0 rejected	Total
H_0 True	$N_{0 0}$ (True negative)	$N_{1 0}$ (False positive)	m_0 (True null hypotheses)
H_0 False	$N_{0 1}$ (False negative)	$N_{1 1}$ (True positive)	m_1 (False null hypotheses)
Total	$m - R$ (Non-rejections)	R (Rejections)	m (Total tests)

Benjamini and Hochberg (1995) yet powerful algorithm that relies solely on the list of p -values resulting from the tests to operationalize false discovery rate control. Rather than applying a fixed threshold across all hypotheses, their method adapts the rejection criteria according to the distribution of observed p -values (Benjamini 2010). This approach increases the potential for true discoveries while still controlling the inflation of false positives and forms the basis of the *step-up* procedures described in the next section (Korthauer et al. 2019).

3. Step-up Procedures for FDR Control

Early formulations by Simes (1986) and Hochberg (1988) laid the groundwork for step-up procedures, which were later adapted to control the false discovery rate (FDR) by Benjamini and Hochberg (1995). The step-up procedure refers to how these methods sequentially compare ordered p -values to increasingly permissive thresholds. Starting from the smallest p -value, each is tested against a progressively higher critical value. The procedure identifies the largest rank for which the p -value remains below its corresponding threshold, and all hypotheses up to that point are rejected. This ascending structure—both in p -value rank and threshold level—strikes a flexible balance between statistical power and false discovery control, thus justifying the “step-up” designation.

3.1. The Benjamini–Hochberg procedure (FDR-BH).

The Benjamini–Hochberg (BH) procedure assumes that individual tests are either independent or positively dependent (Benjamini 2010; Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001). While the method does not require all null hypotheses to be true, it computes rejection thresholds as if all hypotheses were null, using the total number of hypotheses m rather than estimating the number of true nulls m_0 . This conservative design guarantees strict control of the false discovery rate (FDR), even under the global null $m_0=m$. Under these conditions, the expected proportion of false discoveries remains below the pre-specified level, typically $q=0.05$.

The Benjamini and Hochberg (BH) procedure is applied as follows:

1. Order the p -values of all the hypothesis tests in ascending order [1]:

$$p_{(1)}, p_{(2)}, \dots, p_{(m)} \quad [1]$$

where:

$p_{(1)}$ represents the smallest p -value from the set of hypothesis tests; $p_{(2)}$ represents the second smallest p -value; $p_{(m)}$ represents the largest p -value, and m is the total number of hypothesis tests performed; and m is the total number of hypothesis tests performed.

2. Determine the critical value k as the largest i such that [2]:

$$p_{(i)} \leq \frac{i}{m} \cdot q \quad [2]$$

where:

$p_{(i)}$ denotes the p -value with rank i in ascending order among the m hypothesis tests; i is the rank of the p -value when ordered from smallest to largest; m is the total number of hypothesis tests performed; q is the desired false discovery rate (FDR) level.

3. Reject all null hypotheses [3]:

$$H_{(1)}, H_{(2)}, \dots, H_{(k)} \quad [3]$$

where:

$H_{(1)}$ represents the hypothesis corresponding to the smallest p -value; $H_{(2)}$ represents the hypothesis corresponding to the second smallest p -value; $H_{(k)}$ represents the hypothesis corresponding to the k -smallest p -value; and k is the largest rank for which the p -value meets the BH criterion.

3.2. The Benjamini-Krieger-Yekutieli (BKY).

The adaptive procedure proposed by Benjamini, Krieger, and Yekutieli (2006) extends the original BH method by introducing a data-driven estimate \hat{m}_0 of the number of true null hypotheses, out of the total number of tests m . Rather than computing thresholds under the conservative assumption that $m_0 = m$, the FDR-BKY method estimates m_0 from the data and adjusts the rejection threshold accordingly. When $\hat{m}_0 < m$ — that is, when some of the hypotheses are truly non-null — the procedure relaxes the rejection criterion, thereby increasing statistical power while still controlling the false discovery rate under independence and certain forms of positive dependence. This makes the method particularly advantageous in settings where the proportion of false nulls is high, offering a more adaptive and less stringent alternative to the classical BH approach.

The Benjamini-Krieger-Yekutieli (BKY) procedure is applied as follows.

In the first stage, apply the original FDR-BH procedure to the data. This step identifies an initial set of rejected hypotheses and prepares the ground for estimating the number of true null hypotheses, \hat{m}_0 , using the following formula using the following formula [4]:

$$\hat{m}_0(k) = \frac{m+1-k}{1-p_{(k)}} \quad [4]$$

where:

$\hat{m}_0(k)$ is the estimate of the number of true null hypotheses; m is the total number of hypotheses; k is the rank of the p -value used in the estimation; and $p_{(k)}$ is the k -th largest p -value.

This process is repeated for different values of k , and the final value of \hat{m}_0 is obtained by selecting the minimum value across the different k values, ensuring that the estimate is non-decreasing in k . With the estimate of \hat{m}_0 obtained in the first stage, the adjusted critical value q^* to be used in the second stage is calculated as follows [5]:

$$q^* = \frac{q \cdot m}{\hat{m}_0} \quad [5]$$

where:

q^* is the adjusted false discovery rate threshold; q is the desired false discovery rate (e.g., 0.05); m is the total number of hypotheses; \hat{m}_0 is the estimated number of true null hypotheses obtained in the previous step.

This adjusted value q^* is used instead of q in the modified BH procedure, thus making the threshold for rejecting hypotheses less stringent, particularly when \hat{m}_0 is less than m .

Finally, perform the BH procedure again using the adjusted critical threshold from the second stage [6].

$$p_{(i)} \leq \frac{i}{m \cdot \hat{m}_0} \cdot q^* \quad [6]$$

where:

$p_{(i)}$ is the i -th ordered p -value; i is the rank of the p -value; m is the total number of hypotheses; \hat{m}_0 is the estimated number of true nulls; q^* is the adjusted FDR level.

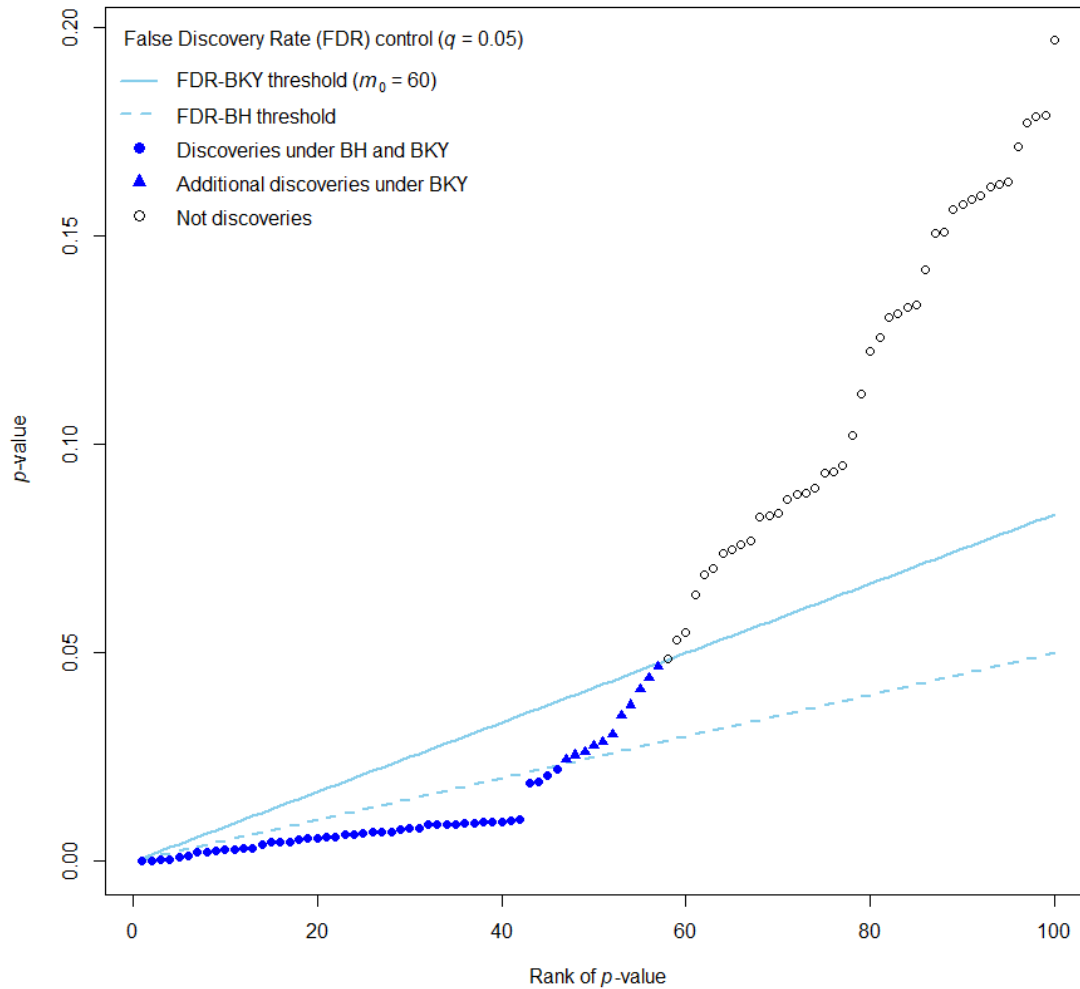
3.3. Graphical Comparison of Thresholds in FDR Step-up Procedures

Graphically (Figure 1), FDR step-up procedures unfold like a rising staircase: p -values are first ordered from smallest to largest and subsequently compared against increasingly permissive thresholds. Based on this comparison, the procedures identify the highest step at which a p -value remains below its corresponding cut-off and accordingly rejects all hypotheses up to that point.

This visual structure illustrates how the procedure balances discovery power with error control dynamically. The two-stage linear step-up procedure of Benjamini, Krieger, and Yekutieli (2006) yields more permissive thresholds, resulting in greater statistical power. The largest differences between the two methods occur at higher-ranked p -values (i.e., larger values), while both procedures yield similar thresholds for the smallest p -values.

While FDR procedures often reduce the number of significant findings compared to unadjusted p -values at a fixed α level, their key contribution lies in reinterpreting what significance means. Even when the total number of rejections remains similar, FDR control ensures that, among all discoveries, the expected proportion of false positives is $\leq q$ (e.g., 5%). This shifts the focus from per-test error control (α) to the reliability of the entire set of discoveries, embedding statistical rigour within the broader context of multiple testing (multiplicity). For instance, with $q = 0.05$, if 100 hypotheses are rejected, we expect ≤ 5 to be false positives —regardless of how many tests were performed.

Figure 1. Step-up rejection thresholds and discoveries under FDR control using BH and BKY procedures.



137

138 Note: This plot illustrates the rejection thresholds for analysing 100 p -values under a target
 139 false discovery rate of $q=0.05$. The estimated number of true null hypotheses ($m_0=60$) was
 140 computed from the data using the BKY method, and is only applicable to the BKY threshold
 141 line. The x -axis shows the rank of the ordered p -values, from smallest (left) to largest (right),
 142 while the y -axis displays the corresponding observed p -value at each rank. These values are
 143 compared against the procedure-specific thresholds (BH or BKY) to determine which
 144 hypotheses are rejected. The intersection point with each threshold line determines the cut-off
 145 rank for significance under that procedure.

146

References.

- Benjamini, Yoav. 2010. "Discovering the False Discovery Rate." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72 (4): 405–416. doi:10.1111/j.1467-9868.2010.00746.x.
- Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57: 89--300. doi:10.1111/j.2517-6161.1995.tb02031.x.
- Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." *Biometrika* 93 (3): 491–507. doi:10.1093/biomet/93.3.491.
- Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *The Annals of Statistics* 29 (4): 1165–1188. doi:10.1214/aos/1013699998.
- García, Luis V. 2003. "Controlling the False Discovery Rate in Ecological Research." *Trends in Ecology and Evolution* 18 (11): 553–554. doi:10.1016/j.tree.2003.08.011.
- García, Luis V. 2004. "Escaping the Bonferroni Iron Claw in Ecological Studies." *Oikos* 105 (3): 657–663. doi:10.1111/j.0030-1299.2004.13046.x.
- Hochberg, Yosef. 1988. "A Sharper Bonferroni Procedure for Multiple Tests of Significance." *Biometrika* 75 (4): 800–802. doi:10.1093/biomet/75.4.800.
- Korthauer, Keegan, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J. Alm, and Stephanie C. Hicks. 2019. "A Practical Guide to Methods Controlling False Discoveries in Computational Biology." *Genome Biology* 20 (1): 118. doi:10.1186/s13059-019-1716-1.
- Simes, R. J. 1986. "An Improved Bonferroni Procedure for Multiple Tests of Significance." *Biometrika* 73 (3): 751–754. doi:10.1093/biomet/73.3.751.
- Sun, Wenguang, Brian J. Reich, T. Tony Cai, Michele Guindani, and Armin Schwartzman. 2015. "False Discovery Control in Large-Scale Spatial Multiple Testing." *Journal of the*

175 *Royal Statistical Society Series B: Statistical Methodology* 77 (1): 59–83.
176 doi:10.1111/rssb.12064.

177