Received: 3 August 2018 | Revised: 7 December 2018 | Accepted: 13 December 2018

DOI: 10.1111/psyp.13335

ORIGINAL ARTICLE

WILEY PSYCHOPHYSIOLOGY SPR

Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location

Jona Sassenhagen | **Dejan Draschkow**

Department of Psychology, University of Frankfurt, Frankfurt am Main, Germany

Correspondence
Jona Sassenhagen, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, 60323 Frankfurt am Main, Germany.
Email: jona.sassenhagen@gmail.com

Funding information
SFB/TRR, Grant/Award Number: SFB/TRR 135 project C7; ERC Consolidator grant, Grant/Award Number: 617891; Deutsche Forschungsgemeinschaft, Grant/Award Number: VO 1683/2-1

Abstract

Cluster-based permutation tests are gaining an almost universal acceptance as inferential procedures in cognitive neuroscience. They elegantly handle the multiple comparisons problem in high-dimensional magnetoencephalographic and EEG data. Unfortunately, the power of this procedure comes hand in hand with the allure for unwarranted interpretations of the inferential output, the most prominent of which is the overestimation of the temporal, spatial, and frequency precision of statistical claims. This leads researchers to statements about the onset or offset of a certain effect that is not supported by the permutation test. In this article, we outline problems and common pitfalls of using and interpreting cluster-based permutation tests. We illustrate these with simulated data in order to promote a more intuitive understanding of the method. We hope that raising awareness about these issues will be beneficial to common scientific practices, while at the same time increasing the popularity of cluster-based permutation procedures.

KEYWORDS

cluster-based permutation test, EEG, MEG, statistics

1 | INTRODUCTION

Advances in cognitive neuroscience rely on powerful non-invasive methods to corroborate and extend behavioral findings. Magnetoencephalographic or EEG (hereafter referred to as MEEG) experiments typically result in high-dimensional outcomes with a spatiotemporal structure: hundreds of time points are sampled at multiple sensors. A popular approach to analyzing ERPs is to submit activity averaged over multiple time points and sensors to parametric tests of significance (e.g., analysis of variance or *t* tests). However, the electrodes and time points to average over cannot be selected contingent on having seen the data (unless, e.g., proper cross-validation is employed; Kilner, 2013), as this completely invalidates the resulting *p* values. Specific analysis parameters can be set a priori (e.g., in the form of a preregistration; Chambers, 2012), but if this was not possible (e.g., because a novel phenomenon is investigated), a more explorative approach is required. One solution is to conduct massively univariate tests (Groppe, Urbach, & Kutas, 2011); that is, conduct the same test as above at every time point and sensor. However, this results in an outcome as high dimensional as the original data, leading to a multiple comparison problem. When hundreds or thousands of individual tests with the usual significance thresholds (e.g., $p < 0.05$) are conducted, the actual error rate greatly exceeds the nominal rate (5%). Correction for multiple contrasts must be applied (Groppe et al., 2011); but while these can provide nominal error rate controls, many of these methods reduce power and cripple the chances of the researcher to observe a true effect given that it is there (Button et al., 2013).

Cluster-based permutation tests are an approach to address this problem, providing both high Type II (power) and nominal Type I (false positive) error rates. For this reason, they have become highly popular, with over 2,000 citations of the seminal paper by Maris and Oostenveld (2007) on Google Scholar. Their effectiveness in controlling Type I error rate has been convincingly established (e.g., Pernet, Latinus, Nichols,

& Rousselet, 2015). The cluster-based permutation approach (introduced into neuroimaging by Bullmore et al., 1996) considers the specific structure of MEEG data to maximize power. It builds on the assumption that effects are clustered along the dimensions of interests: (a) time, and (b) space. On one hand, it assumes that a true signal has a temporal extension reflected at multiple adjacent time points, and, on the other hand, adjacent sensors show similar patterns. Spatial clustering at the sensor level is motivated by the fact that the sensor level signals are the result of volume-conducted, source-level currents (i.e., one cortical source projects to all of the surface-based sensors), which leads to correlated activity of adjacent sensors. Thus, MEEG data—and presumably true MEEG effects (i.e., MEEG events reflecting neurocognitive events)—show a characteristic correlation structure.

Importantly, the cluster-based permutation approach suggested by Maris and Oostenveld (2007) realizes its control of the multiple comparison problem while maximizing power by employing the cluster structure of the data as its sole test statistic. That means that no inference is made over individual voxels (i.e., one specific time point at one specific sensor). Instead, first, clusters are identified in the data via an algorithm (see below), and second, the cluster structure of the observed data is compared to the pattern of cluster sizes from data constructed under the null hypothesis. Conceptually, this means that individual voxels are never visible to the second statistical inference stage, and no statistical inference is made about individual points. Thus, no statements about the onset or offset of a significant difference with millisecond precision will be accurate. Furthermore, the spatial location of the cluster is similarly invisible at this stage, meaning that it is not the location (e.g., temporal and spatial extent) of the cluster to which, for example, resulting p values refer, but only their size.

Nevertheless, researchers consistently employ cluster-based permutation test results to support claims not only about the existence of a significant difference, but also about the (spatial or temporal) extent or location of effects. This procedure, while common, is inapplicable. Cluster-based permutation tests as discussed by Maris and Oostenveld (2007) do not provide statistical inference for the location of effects. In the following, we discuss the nature of cluster-based permutation tests in more detail in order to show why this inference is not justified, and then present what inference is justified, as well as how inference about timing or spatial extent of effects can be made.

The relevance of elaborating on these issues stems from the high prevalence of the method. It is integrated in the major MEEG analysis suites (e.g., Delorme & Makeig, 2004; Gramfort et al., 2013; Oostenveld, Fries, Maris, & Schoffelen, 2011) and has become standard practice in MEEG analysis. Misuses of the method are ubiquitous in the literature; while we will abstain from identifying individual "offenders," the authors themselves must admit to having inappropriately utilized the method in the way here described, and further examples can be abundantly found in the published literature.

1.1 | How cluster-based permutation tests work

Cluster-based permutation tests have two components. One is the cluster-forming algorithm, which converts one high-dimensional observation into a quantifiable summary regarding its cluster structure. The other creates a surrogate null distribution, against which the observed data is compared to obtain p values. In the following, we present only a conceptual overview; for a more formal treatment, see Maris and Oostenveld (2007).

1.2 | Cluster formation stage

In the following, we refer to the case where the data (i.e., each observation) has the shape (Space × Time), although essentially the same considerations apply to one-dimensional data (e.g., only time), or 3D data (e.g., Frequency × Space × Time). For simplicity, we consider the case of inference regarding a binary condition contrast (i.e., the contrast between the ERP in two conditions) across subjects. The same concerns apply for more complex designs, with multiple levels and with reference to a null hypothesis of exchangeability of all levels of a factor. In the two-condition case, at each coordinate, a first-order test statistic is calculated (i.e., a t test is conducted, contrasting the values at this time point, for each subject, of Condition 1 versus Condition 2). The H_0 of this test is that, in an unobserved population of subjects, exposed to the same experimental manipulation, the difference between the two conditions would equal precisely 0. Repeating this procedure for the entire data set results in a (Sensor × Time) t score map. Comparing these t values to the t distribution yields p values, but, of course, due to the number of scores, many are expected to exceed a critical value (resulting in, e.g., $p < 0.05$) even if the true effect is zero everywhere (i.e., the multiple comparison problem). In principle, a range of approaches is available to account for this (discussed by, e.g., Groppe et al., 2011; Maris & Oostenveld, 2007), but in the case of cluster-based permutation tests, clusters in the data are identified and utilized. For this, voxels are thresholded according to an a priori defined criterion (e.g., $t = 1.96$), and adjacent voxels with t scores exceeding this value are grouped together. Temporal adjacency is trivially defined, but spacial adjacency for MEEG sensors requires a priori established neighborhood definitions. Finally, groups are summarized into a single number by, for example, taking the sum of the t values—yielding the cluster size(s).

Critically, the extent of the cluster(s) in the data—where in time they start, their topography, and if applicable their

frequency boundaries—is entirely fixed at this stage and is a purely descriptive function of the data at hand. Moreover, it depends strongly on the specific cluster-forming algorithm employed and the chosen parameters (e.g., What is the first-stage test statistics—*t* scores? Raised to the first or second power? A threshold corresponding to $p < 0.05$ or $p < 0.01$? Is the cluster sum or the count of the included values considered? How many voxels must a voxel be adjacent to in order to count for the same cluster?) and on preprocessing choices (filter settings? spatial neighborhood templates?).

1.3 | Inference stage

The cluster formation stage described in the previous section often contains a nominally inferential stage—thresholding voxels by, for example, if their *t* scores exceed a certain value. However, the results of this procedure are not interpreted as inferential claims because they fall prey to repeated tests. Instead, a second-order inference stage is employed. The null hypothesis of this stage is that the cluster structure of the data identified in the first stage is exchangeable between conditions (i.e., clusters simply reflect the inherent correlation of MEG data in time, space, and frequency). The cluster test statistic is a complex, nonlinear function of the data at hand. Different from, for example, the mean, whose distribution is often well understood, it is not a priori clear what distribution would result if we were to repeatedly take samples from data where the null is true, establish cluster sizes as in the first stage, and consider their distribution. Thus, an analytic approximation of the null distribution is currently not practically feasible for MEEG data. Instead, the probability of the data under a null hypothesis of exchangeability can be established with permutation tests.

Note the specific form of the null hypothesis: samples are exchangeable with regard to the manipulation. Simplified, it makes no difference if we consider data from Condition 1 or Condition 2, they were drawn from the same probability distribution (with regard to the cluster-forming procedure). The data at hand are used to simulate a null hypothesis of exchangeability. A permutation test simply realizes this null hypothesis by enumerating all possible assignments of data points to the conditions. Full permutation tests are usually computationally intractable; however, a special class of approximations—Monte-Carlo sampling—are, and they yield satisfactory results if > 800 (Pernet et al., 2015) permutations are conducted. That is, for each iteration, assignment to conditions is randomized for each data set—that is, for the data from Condition 1 and Condition 2 of Subject 1, it is randomly chosen if Condition 1 is subtracted from 2 or the reverse, and so on. Then, within this iteration, the first stage process is repeated to establish its cluster structure, and the value stored. Subsequently, a new iteration is conducted and so on, until a large number of samples under the null hypothesis of exchangeability has been obtained. If, for example, the sum of the largest cluster is taken for each iteration, the result is one value for each.

Finally, the empirical cumulative density function of these surrogate-null values is computed as an approximation of the distribution of the test statistic under the null (paralleling the well-known normal and *t* distributions for the distribution of a much simpler test statistic in the case of simply considering, e.g., means). The percentage of surrogate-null values that the actual observed data exceed corresponds to the *p* value under the null of exchangeability.

Importantly, while the original data is high dimensional and thus prone to multiple comparison issues, the first stage reduces it to a single number, and it is this number whose probability under the null is established.

1.4 | What to do and not to do with cluster-based permutation test results once you have them

Note that, in our description of the procedure, the extent and dimensions of the cluster(s) were fixed at the first stage. The inferential second stage does not ever "see" first stage coordinates, only the cluster size(s).

In practice, this means that there is no guarantee of the false-positive rate on any of the points included in the cluster. That is, statistical claims regarding differences between the two conditions are not justified for any of the points included. Thus, there is no statistical certainty or confidence regarding claims about a difference at, for example, the earliest time point in a cluster. However, such claims are routinely made in the literature, taking a form such as (following is a quote from a recently published article, changed to prevent identification): "A cluster-based permutation test was conducted to identify the time point at which the conditions differ."

Such a claim is not justified due to a multitude of reasons.

First, the test did not evaluate if the inclusion of an earlier time point in the cluster, or the omission of the earliest time point, would not also have resulted in a significant result. The cluster extent was not established by the inferential stage (i.e., by the permutation and inference stage). The cluster extent depends on the sample at hand as it is a descriptive statistic, and on the specific workings and parameter settings of the cluster-forming algorithm. Importantly, this does not mean it is inherently wrong to report cluster extents. It is perfectly appropriate to describe that, in the observed data, the cluster-forming algorithm established a cluster of a certain extent. It is simply that the specific shape of this cluster—its spatial, temporal, frequency … extent—has not been the subject of an inferential test with guaranteed error rates; only that its size occurs in the outmost tails of the surrogate null data (i.e., that it is improbable under the specific null).

Second, the preinference thresholding of individual data points reflects the power of the test. That is, with more trials or less noise, earlier points will pass the threshold at this stage (and sensors at the borders of topographical pattern will turn out significant). Lower measurement noise or larger samples shift observed effects forward in time/shrink topographical patterns/make effects appear over a narrower frequency range. Inversely, high noise/small samples shift effects backward in time/widen the spatial extent/make effects show up in wider frequency bands. For the temporal dimension, extent also strongly depends on filter settings (Tanner, Morgan-Short, & Luck, 2015); for the frequency domain, on, for example, wavelet cycles.

Finally, cluster-based methods can underestimate the latency, spatial, or frequency extent of effects because the power of the whole cluster may carry forward points at its margins through the inference stage. As introduced by Groppe et al. (2011, p. 1718), this in fact is more likely for the earliest and latest time points (and for the lowest and highest frequencies), as the cluster margin points gain power from the cluster peaks. Often, at the first stage, points are thresholded on p value cutoffs (e.g., only points where $p < 0.05$ or $|t| > 1.96$ are included). However, these p values are not corrected for multiple tests, and in fact many false positives are expected. Thus, in many contexts, significantly large clusters will include points where the null hypothesis is true (see, e.g., the simulation below), and there is no control (e.g., a 5% level) on the rates of such errors. This is why cluster-based permutation tests are not equipped to provide precise estimation of spatial distributions, temporal onsets, or frequency bands. The locations (latencies, topographies …) of clusters identified by cluster-based permutation tests will be strongly correlated with the true extent of effects, but there is no error rate control on these locations, as would be expected for scientific estimation.

For specific recommendations on how to test for MEG effect latencies, we recommend the works of Kiesel, Miller, Jolicœur, and Brisson (2008), Luck (2005), Miller, Ulrich, and Schwarz (2009), Piai, Dahlslätt, and Maris (2015), and Rousselet (2012). These methods include, for example, jackknife estimates of fractional area latency or (properly calibrated) counting of successive positive tests. For contrasting spatial patterns, we refer to, for example, King and Dehaene (2014), Tian and Huber (2008), and Urbach and Kutas (2002).

1.5 | Cluster-based permutation tests for post hoc comparisons

Having observed a significant cluster, researchers (perhaps, in particular, researchers who are in principle aware of these issues) sometimes may feel inclined to instead use cluster-based tests to establish effect masks, that is, to conduct follow-up analyses, such as investigate the time course of an effect at those sensors identified as showing significant effects by the test and correlating cluster strength with a variable of interest. Here, two important caveats must be considered. First, as the test thresholds points by effect strength, it brings the danger of circular analysis (Vul, Harris, Winkielman, & Pashler, 2009). If a cluster-based permutation test indicating a difference between subjects scoring high versus low on a test is followed up by a correlation of subject score on cluster activity, this is circular. To conduct such an analysis, independence must be established via, for example, cross-validation.

Second, while using cluster extents as masks can be justified if the masks stem from an independent data set, the cluster-based test (i.e., the inference stage) may not do any meaningful work. Remember again that the extent of the cluster (i.e., the shape of the mask) is established at the first stage. Thus, the costly calculation of the surrogate null data is not inherently useful.

1.6 | How to report cluster-test results

As noted, reports of inapplicable usage of cluster-based permutation tests in the literature are abundant. Such misuses can hardly be blamed on the original presenters of the method. Maris and Oostenveld (2007, p. 187) make this point explicitly:

> There is a conflict between this interest in localized effects and our choice for a global null hypothesis: by controlling the FA [false alarm] rate under this global null hypothesis, one cannot quantify the uncertainty in the spatiotemporal localization of the effect.

It is also stated emphatically on the Fieldtrip website: "Here is what NOT to write: 'We found a significant cluster in area X, between time point A and B.'"

Further, Groppe et al. (2011, p. 1718) write:

> It is important to note that because p values are derived from cluster level statistics, the p value of a cluster may not be representative of any single member of that cluster. For example, if the p value for a cluster is 5%, one cannot be 95% certain that any single member of that cluster is itself significant […]. One is only 95% certain that there is some effect in the data. Technically, this means that cluster-based tests provide only weak [family-wise error rate] control …

It is discussed in detail by Maris (2011, p. 8):

> Part of the appeal of cluster-based permutation tests comes from the fact that they localize effects in space, frequency, and time. This feature

> should not be overvalued. In fact, cluster-based permutation tests do not control the false alarm rate at the level of the (channel, frequency, time)-triplets, the (channel, frequency)-pairs, or any other pairs or singletons. Therefore, they do not allow probabilistic statements about effects at the level of these triplets, pairs, and singletons. For instance, on the basis of a cluster-based permutation test, one cannot say that the probability of including a particular (channel, frequency)-pair in a significant cluster is controlled at the nominal alpha level under the null hypothesis of no effect at that particular pair regardless of any effects at the other pairs. Such a null hypothesis will be called a specific null hypothesis: equality of the probability distributions in the two conditions at a particular (channel, frequency, time) triplet (or any other pair or singleton), while there may be inequality at all other triplets (pairs or singletons). This specific null-hypothesis must be contrasted with the nonspecific null hypothesis that is tested in a permutation-based test: equality of the probability distributions of the complete multivariate data sets in the two conditions. Thus, in a permutation-based test, one controls the false alarm rate under this nonspecific null hypothesis, and not under a null hypothesis that is specific for a particular triplet, pair, or singleton in the hyperspace.

And it is also noted by Piai et al. (2015):

> [The] false alarm rate [for effect, e.g., latencies] is not controlled […] by cluster-based permutation tests (Maris & Oostenveld, 2007). In fact, these approaches only control the false-alarm rate under the omnibus null hypothesis involving no effect for none of the time points.

To summarize, it is not wrong to speak of a cluster being significant per se, if what is meant by this is to say that the null hypothesis is false with respect to the cluster structure of the data and that some specific cluster(s) exceed a critical value (see, e.g., Maris & Oostenveld, 2007, p. 187, sections 4.4.3 and 4.5). After all, each cluster is associated with a proper permutation-based p value corresponding to its position in the surrogate-null histogram. That is, the cluster(s) can be "significant" with respect to their size, as this is what the inference stage is concerned with. They are not significant with respect to their extent, which was established entirely at the first stage. Thus, when $p < 0.05$ after a cluster-based permutation test, while the existence of the cluster is significant, its precise shape is not—or: the significance of the cluster refers only to the location of its cluster-level statistic in the distribution of such statistics, not to its location in space, time, or frequency.

The cluster extent may still be an informative description of the data—much as, for example, its standard deviation—but it is not an inferential claim. Thus, it should not be reported as such: "A nonparametric cluster-based permutation analysis assessed the moment in time where conditions differed, which was 180 ms."

It could, however, be justified to say: "A nonparametric cluster-based permutation analysis indicated an effect of condition ($p < 0.05$). This corresponded to a cluster in the observed data beginning at 180 ms."

This statement is justified; however, it might not be considered best practice. The categorical distinction between the first and second sentence—one is inferential, the other descriptive—is very easily glossed over. The inferential reading is, in turn, very seductive. We think the abundance of misuses in the literature demonstrate this. Perhaps it is preferable to either prefer vague statements (e.g., "corresponding to a cluster beginning around 150–200 ms"), so as to not imply statistical precision where none is given or to inverse the order, so as to indicate the categorical difference between inferential and descriptive claims (e.g., "A cluster in the observed data extended from 180 to 250 ms. The cluster-based permutation test indicated that there was a significant effect of condition").

The same goes for spatial and frequency extents. Cluster tests do not statistically justify the claim that the effect occurs between X and Y Hz. While it can be correct to say, "A cluster in the observed data was found in the theta band. Cluster-based permutation testing indicated this cluster to be significant," it would not be appropriate to understand this to mean it was statistically established (in the sense of false-positive control) that the effect occurs only, or primarily, in the theta band.

1.7 | How bad can things be? A simulation

We claim above that it would be grossly misleading to assume that cluster-based permutation tests, which imply statistical guarantees on error rates, can be meaningfully employed to estimate the extent of an effect with statistical certainty. For demonstration purposes, we conduct a simulation study to exemplify realistic real-world failure rates.

Consider a researcher who wishes to estimate how early an effect appears and wishes to use a cluster-based permutation test for this purpose. We have found multiple counts in the literature where researchers have indeed done this—for example, conducting the test and assuming that the earliest time point included in any cluster significant at some alpha level is a reliable indication of the true onset of the effect. Again, note that this time point does not result from the "test"

part of the cluster-based permutation test, as cluster extents are established during the preinferential cluster formation stage; thus, it can be assumed that the test is performed partially to provide statistical guarantees on the effect.

While the precise bias of this procedure is hard to estimate and highly context dependent, there are multiple factors at play establishing such a bias. As noted above, time points at the core of a cluster carry forward points at the margins—that is, those supporting claims about effect onsets—through the inference stage. This is how underestimations of latencies can come into existence—for example, the (temporal) extent of the cluster is overestimated. On the other hand, the cluster-formation stage does not benefit from the sensitivity-enhancing clustering itself (although see Mensen & Khatami, 2013, for an alternative algorithm). That means that it can also in a sense be conservative if employed for selecting points at the margin: if the power of a study is low, the probability of detecting points at the margins will be low. Thus, while the existence of any effect at all is detected with high power, points at the margin will only be included in a fraction of cases. Their inclusion will depend on various factors, in particular, cluster-forming thresholds and how broadly distributed versus how focal the effect is. Of course, effects have spatial extent, too, so there will be multiple sensors at which the preselection test at this stage can succeed. However, (a) this, of course, means false positives become likely, and (b) as tests across multiple sensors are not independent (due to volume conduction and spatial correlation), a true increase in sensitivity is hard to estimate. Either way, the following inference stage (i.e., construction of the surrogate null and comparing observed clusters with the resulting critical value) does not provide any guarantees of nominal Type I error rates (i.e., upper bounds for false positives) for a claim about cluster extent. Thus, while it provides Type I controls on reporting any significant differences at all, it does not provide Type I controls on claims such as "the effect onset is as early as X ms." As our simulation below shows, Type I error rates can far exceed nominal alpha levels.

2 | METHOD

Note that an OSF (Open Science Framework) repository containing the exact code required to replicate this simulation is provided at https://osf.io/5cw7n/. The EEGLAB example data set (Delorme & Makeig, 2004), a 32-channel continuous recording, was processed in MNE-Python (Gramfort et al., 2013). Continuous activity was cleaned of eye movement artifacts with independent component analysis (Jung et al., 2000), downsampled to 100 Hz, and low-pass filtered at 30 Hz.

Across 10,000 simulation runs, on each run, 700-ms segments at random time points were extracted as epochs. The prestimulus period mean was subtracted as a baseline and then dropped, leaving 500-ms long epochs. On each run, 100 such nonoverlapping epochs were created. Then, to one half of these epochs, a simulated monophasic "ERP" effect was added. For this, a normal distribution probability density function, evaluated from −1.5 to 1.5 over 21 data points (210 ms) and scaled to range from 0 to 15 uV, was added to each data segment, beginning at 150 ms (so that the first sample to exceed zero—and thus the ground truth effect onset—was at 160 ms). A topography of the effect was simulated by multiplying it with the topography of the first independent component of neural origin (resulting in a frontocentral topography). The result of this was an ERP-like perturbation embedded in real EEG background noise on every trial, beginning at 160 ms.

Then, a cluster-based permutation test, as implemented in MNE-Python, was conducted with the null hypothesis that the epochs to which the ERP was added are exchangeable with the ones where it was not added. That is, we implemented a test just as those realistically employed to establish the generalizability of differences between conditions. Spatial adjacency was computed via Delaunay triangulation. A one-tailed F test was conducted, and initial cluster forming thresholded at a value corresponding to $p < 0.05$ (uncorrected).

As noted, this procedure was repeated 10,000 times. On simulation runs with a positive result, the earliest time point in any of the significant clusters was extracted. The difference between the actual effect onset and this number was noted, and its distribution plotted over all runs, as well as the degree of underestimation of effect onset latency.

3 | RESULTS

As shown in Figure 1, in general, the test tended to overestimate the latency of the effect. Underestimations of 40 ms and longer occurred in many more than 5% of runs (see Figure 1, right). Specifically, on >20% of runs, the effect onset was estimated too early; divergences of 40 ms or more were found at >10% of runs.

4 | DISCUSSION

These findings show that cluster-based permutation tests hardly provide statistical guarantees for claims such as "the manipulation induced an ERP no later than X ms." Divergences can be large and occur in many more cases than the nominal error rate of the test. This is not unexpected, as the test's guarantee of error rates concerns the reporting of a significant cluster when the null hypothesis of exchangeability is true, not the extent of the cluster.

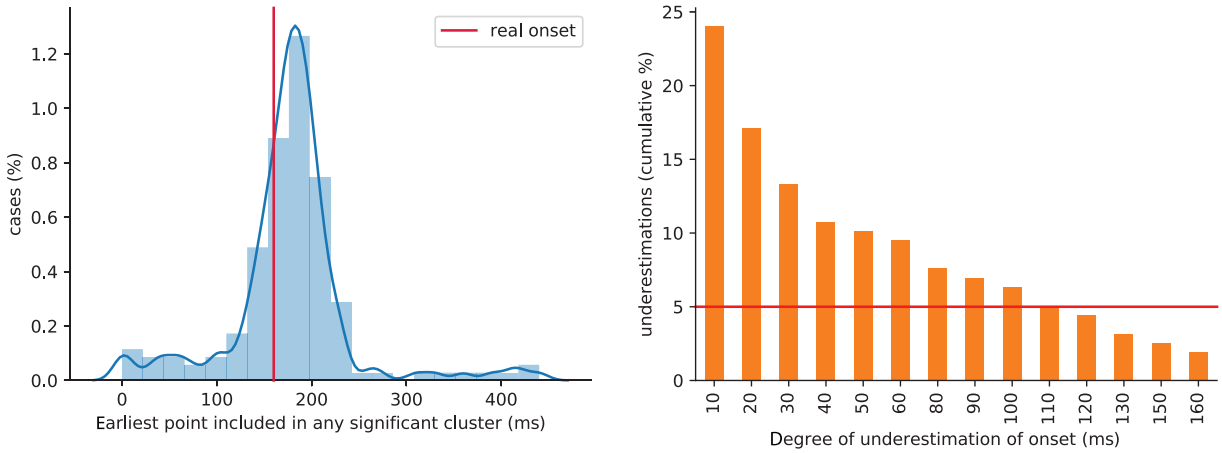

FIGURE 1 Left: Distribution of earliest time points included in any significant clusters, and ground truths (red line). Kernel density estimate and histogram are shown. Significant-cluster margins do not track real onsets well. Right: Cumulative error rates, per degree of divergence, for each level of underestimation of effect latency. Red line indicates 5% level

4.1 | Limitations

The precise outcomes of this simulation are very dependent on various choices. As noted, the specific cluster-forming algorithm, filtering, and signal-to-noise level greatly influence the precision of the procedure. For example, raising the cluster-forming threshold and (most) low-pass filtering will increase estimated onset latencies; increasing the signal-to-noise ratio will improve the accuracy of misusing cluster-based permutation tests for estimating effect onsets. (Trivially: consider the case where the null hypothesis is false and the noise level approaches zero—i.e., approaching an effect size of 1. Now, the test approaches 100% power and alpha can be lowered arbitrarily. That is, during cluster formation, almost all selected points will be true positives, and almost all deselected points will be true negatives. At the second stage, in the limit, the only observed clusters will be the true effect; that is, as effect size approaches infinite, the empirical cluster approaches the true effect and the empirical cluster increasingly is the only cluster ever observed.) However, note that this is very different from normal error rate control. A *t*-test—or any other proper hypothesis test, such as cluster-based permutation tests—guarantees that the null hypothesis it actually tests is only reported to be false, given that it is true, at nominal levels (i.e., in 5% of cases). This is true by the test's construction and will be completely unaffected by, for example, noise levels; even if the signal-to-noise ratio is extremely poor, while the test's power will be low, its false-positive ratio will be nominal. Nonetheless, the specific outcomes of this simulation are dependent on various ad hoc parameter choices.

It cannot be stressed enough that this finding does not reveal an error in the cluster-based permutation test procedure. The procedure is doing exactly what it was designed, and promised, to do: control false-positive rates for the null hypothesis of exchangeability with regard to the cluster structure of the data, while maximizing sensitivity. It is only the shortcomings of a misuse of this procedure that are highlighted here.

4.2 | Conclusion

The strength of cluster-based permutation procedures in handling the multiple comparisons problem unfortunately can lead researchers on the slippery slope of misinterpretation. Common and problematic misuses are reporting the onset or offset of conditional differences on a millisecond scale or precise spatial extent of a cluster. This suggests unwarranted precision of the actual underlying test statistic and can lead to very strong but unsubstantiated claims. The aim of this article is to demonstrate and exemplify the problems and common pitfalls of using and interpreting cluster-based permutation tests. Our simulation reveals that individual time points at the beginning of a cluster are an unreliable estimate for the actual onset of differences between conditions. Additionally, in our case, >20% of the time the onset of the effect was estimated earlier then the true effect. This is corroborated by the architecture of this procedure: while there is multiple comparison control for establishing a significant difference between conditions, there is no such control for the individual time points included in a cluster. This means that any statement about a specific time point is misleading, and such statements should not be included when reporting the analysis. To foster accurate interpretation and unambiguous reporting of the outcome of cluster-based permutation tests, we suggest reporting an approximate but clearly descriptive time window of a cluster. A statement about significance can only be made for the overall statistical contrast (e.g., "The cluster-based permutation test indicated that there was a significant difference between conditions A and

B"). Reporting the extent of the cluster is recommended only if the descriptive nature of this information is made explicit (e.g., "A cluster in the observed data extended from approximately 180 to 250 ms"). We hope that accurate interpretations of cluster-based permutation tests will contribute to the adequate utilization of this powerful method.

ACKNOWLEDGEMENTS

We wish to thank Eric Maris and an anonymous reviewer for critical suggestions and discussions.

ORCID

Dejan Draschkow https://orcid.org/0000-0003-1354-4835

REFERENCES

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., … Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, *35*(2), 261–277. https://doi.org/10.1002/mrm.1910350219

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Chambers, C. D. (2012). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, https://doi.org/10.3389/fnins.2013.00267

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Jung, T.-P., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts from by blind source separation. *Psychophysiology*, *37*(2), 163–178. https://doi.org/10.1111/1469-8986.3720163

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*(2), 250–274. https://doi.org/10.1111/j.1469-8986.2007.00618.x

Kilner, J. M. (2013). Bias in a common EEG and MEG statistical analysis and how to avoid it. *Clinical Neurophysiology*, *124*(10), 2062–2063. https://doi.org/10.1016/j.clinph.2013.03.024

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Luck, S. J. (2005). *An introduction to the event-related potential technique* (p. 374).). Cambridge, MA: The MIT Press. https://doi.org/10.1118/1.4736938

Maris, E. (2011). Statistical testing in electrophysiological studies. *Psychophysiology*, *49*(4), 549–565. https://doi.org/10.1111/j.1469-8986.2011.01320.x

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mensen, A., & Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. *NeuroImage*, *67*(Suppl C), 111–118. https://doi.org/10.1016/j.neuroimage.2012.10.027

Miller, J., Ulrich, R., & Schwarz, W. (2009). Why jackknifing yields good latency estimates. *Psychophysiology*, *46*(2), 300–312. https://doi.org/10.1111/j.1469-8986.2008.00761.x

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 15869. https://doi.org/10.1155/2011/156869

Pernet, C., Latinus, M., Nichols, T., & Rousselet, G. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85–93. https://doi.org/10.1016/j.jneumeth.2014.08.003

Piai, V., Dahlslätt, K., & Maris, E. (2015). Statistically comparing EEG/MEG waveforms through successive significant univariate tests: How bad can it be? *Psychophysiology*, *52*(3), 440–443. https://doi.org/10.1111/psyp.12335

Rousselet, G. A. (2012). Does filtering preclude us from studying ERP time-courses? *Frontiers in Psychology*, *3*(May), https://doi.org/10.3389/fpsyg.2012.00131

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. https://doi.org/10.1111/psyp.12437

Tian, X., & Huber, D. E. (2008). Measures of spatial similarity and response magnitude in MEG and scalp EEG. *Brain Topography*, *20*(3), 131–141. https://doi.org/10.1007/s10548-007-0040-3

Urbach, T. P., & Kutas, M. (2002). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology*, *39*(6), 791–808. https://doi.org/10.1017/S0048577202010648

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Reply to comments on "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition". *Perspectives on Psychological Science*, *4*(3), 319–324. https://doi.org/10.1111/j.1745-6924.2009.01132.x

How to cite this article: Sassenhagen J, Draschkow D. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*. 2019;e13335. https://doi.org/10.1111/psyp.13335