Peer Community In Archaeology

A practical computational approach to stratigraphic analysis using conjoinable material culture.

Hector A. Orengo based on reviews by Robert Bischoff, Matthew Peeples and 1 anonymous reviewer

A recommendation of:

The strength of parthood ties. Modelling spatial units and fragmented objects with the TSAR method – Topological Study of Archaeological Refitting

Sébastien Plutniak *(2021), OSF Preprints, q2e69, ver. 3 peer-reviewed and recommended by PCI Archaeo.* https://osf.io/q2e69

Open Access

Submitted: 14 January 2021, Recommended: 24 June 2021

Published: 24 June 2021

Recommendation

The paper by Plutniak [1] presents a new method that uses refitting to help interpret stratigraphy using the topological distribution of conjoinable material culture. This new method opens up new avenues to the archaeological use of network analysis but also to assess the integrity of interpreted excavation layers. Beyond its evident applicability to standard excavation practice, the paper presents a series of characteristics that exemplify archaeological publication best practices and, as someone more versed in computational than in refitting studies I would like to comment upon.

It was no easy task to find adequate reviewers for this paper as it combines techniques and expertise that are not commonly found together in individual researchers. However, Plutniak, with help from three reviewers, particularly M. Peeples, a leading figure in archaeological applications of network science, makes a considerable effort to be accessible to non-specialist archaeologists. The core Topological Study of Archaeological Refitting (TSAR) method is freely accessible as the R package archeofrag, which is available at the Comprehensive R Archive Network (https://CRAN.R-project.org/package=archeofrag) that can be applied without the need to understand all its mathematical, graph theory and coding aspects. Beside these, an online interface including test data has been provided (https://analytics.huma-num.fr/Sebastien.Plutniak/archeofrag/), which aims to ease access to the method to those archaeologists inexperienced with R. Finally, supplementary material showing how to use the package and evaluating its potential through excellent examples is provided as both pdf and Rmw (Sweave) files. This is

an important companion for the paper as it allows a better understanding of the methods presented in the paper and its practical application.

The author shows particular care in testing the potential and capabilities of the method. For example, a function is provided “frag.observer.failure” to test the robustness of the edge count method against the TSAR method, which is able to prove that TSAR can deal well with incomplete information. As a further step in this direction both simulated and real field-acquired data are used to test the method which further proves that archeofrag is not only able to quantitatively assess the mixture of excavated layers but to propose meaningful alternatives, which no doubt will add an increased methodological consistency and thoroughness to previous quantitative approaches to material refitting work, even when dealing with very complex stratigraphies.

All in all, this paper makes an important contribution to core archaeological practice through the use of innovative, reproducible and accessible computational methods. I fully endorse it for the conscious and solid methods it presents but also for its adherence to open publication practices. I hope that it can become of standard use in the reconstruction of excavated stratigraphical layers through conjoinable material culture.

[1] Plutniak, S. 2021. The Strength of Parthood Ties. Modelling Spatial Units and Fragmented Objects with the TSAR Method – Topological Study of Archaeological Refitting. OSF Preprints, q2e69, ver. 3 Peer-reviewed and recommended by PCI Archaeology. https://doi.org/10.31219/osf.io/q2e69.

Cite this recommendation as:
Hector A. Orengo (2021) A practical computational approach to stratigraphic analysis using conjoinable material culture.. *Peer Community in Archaeology, 100009.* 10.24072/pci.archaeo.100009

Reviews

Revision round #1

2021-03-04

Author's Reply

I thank the three reviewers for their valuable comments and suggestions, which helped a lot to improve this paper. In addition to the change suggested by the reviewers (which are detailed below), this second version comes with all the required suppelmentary material, including a new version of the archeofrag package (v0.7).

I would also appreciate the Rmd file as an additional supplemental document for easier reproducibility, but the code can be copied from the pdf

The Rmd file has been added to the supplementary materials. I also checked and ensured that all the materials needed to reproduce the content of this paper are provided.

1) On Page 3, the first Figure and the discussion related to it are somewhat abstract. I would suggest to add a real-world archaeological example to explain to the dumb folks like me what this actually means in a concrete archaeological example. Otherwise the abstraction might scare people off.

A concrete example is given in Figure 2. It might be better to have Figure 2 before Figure 1, but this is not optimal for the sake of the presentation. I assume that if things are unclear to the reader in Figure 1, it should become clear few paragraphs later with Figure 2.

2) It would be nice, although not absolutely necessary, to see how this method could be deployed in conjunction with other methods of stratigraphic analysis, as well as Bayesian modelling of radiocarbon dates. I would think that this method would be extremely useful in fine-tuning the modelling of C14 dates. It would be useful if the author could maybe comment on that and give an idea how this method might be used in combination with others?

This is an interesting suggestion, but far beyond the scope of this paper. Development of the TSAR method (or based on this method) are ongoing and not presented in this first paper. For example, the development version of the archeofrag package includes (from version 0.6.6, see https://github.com/sebastien-plutniak/archeofrag) new methods to weight the edges based on morphometric values of the objects and on the spatial distance between them.

3) I guess this method is still under development, but I wonder if the author might be able to comment a little more on how we could make this method something more commonplace? At the moment, it seems fairly inaccessible, so what will be the pathway to making it more accessible to a wide range of archaeologists (especially those like me who are a bit challenged by the maths)?

This paper is intended to give an in-depth presentation of a method that can be used without a complete understanding of all the aspects addressed in this paper. The "archeofrag" R package offers a set of functions to apply the method without going into its mathematical aspects. Given that not all archaeologists manage the R language, I also developed an online interface from which all users can experiment (https://analytics.huma-num.fr/Sebastien.Plutniak/archeofrag). This online application, with few options for the moment, will be further developed in the future.

The "preliminary definitions" section includes some useful information, but I also think this could be reduced not just to save space but also to add to the clarity of the argument.

The reference to connections relying on "solid-state physics" and then the next sub-section with the discussion of Queen and Rook contiguity and the analogy to spatial analysis made a pretty clear distinction a bit more muddled in my reading.
I think a very simple definition of connection and similarity refits and Figure 2 is plenty and felt that the Queen/Rook example did not add clarity and was not referred to again so could be removed.

section 2.1.3 Topological properties could be basically eliminated

I added a figure and revised the text for clarity. It seems important to keep the paragraphs on contiguity, because it is an important aspect when recording the data for refitting analysis. Consequently, I think it must be addressed in the main body of the text (and not in the supplementary material), because the user of the TSAR method will face this problem.

section 2.1.2 could be simplified to make the two basic points that are necessary for the rest of the paper

1) archaeologists identify potential refits by either identifying pieces that fit together across substantial surfaces in 3D space or alternatively by looking for similarities in the objects themselves

2) you are only focused on the former in this paper.

I simplified this paragraph by removing the part on the epistemological aspects of the two types of relationships.

It wasn't clear to me how the "merge graphs" step in the two initial layers works. I can think of several ways you might have chosen to implement this, but I do not think the choices you made are completely clear in this draft or the supplement. I think a short paragraph would be needed to address this.

I rephrased this part, hoping that it is clear now. About the way the algorithm works, I also added a description in pseudo-code, which complements the flowchart diagram. The reader interested in learning more about it can read the code of the function, included in the archeofrag package and written with abundant comments.

the degree to which missing connections would hamper our ability to connect empirical patterns to generative processes. Given the nature of fragmentation and post-depositional processes, I expect that it would be common for archaeologists to miss refits due to damage to potential conjoining edges of pieces or all manner of other processes. How would this influence your results? I think this issue merits some discussion at least and perhaps some small analyses.

I had one suggestion: missing connections could potentially be evaluated with an additional global parameter such as a probability that a connection is removed once the graphs have been formed to essentially simulate "missed" or unidentifiable connections. Especially with the focus on "connections" rather than similarity, I think it's likely that missed connections would be present in any empirical refit study and this simulation model provides a setup that could help evaluate the impacts.

Thank you for this very inspiring suggestion. I developed a new function in the package to simulate missed connections (the "frag.observer.failure" function), and used it to test the robustness of the edge count method versus the TSAR method. This gave a strong and additional argument supporting the TSAR method, demonstrating that it is more robust and less sensible to the "lack" of information. I shortly reported these results in section 3.2.2 (to not overload the main text) and referred to the supplementary materials for details.

section 3.1.1 you compare the admixture model and others to the archaeological intuition. Given the variability in the archaeological intuitive coding, I don't know that the methods used here fitting in the same cluster is a particularly convincing argument in favor of these new methods (and that's okay).

To me, the results you present suggest that people are generally not great at evaluating the relative order of topological patterns and that, in and of itself, argues for some sort of automated approach like you've developed here.

The clustering diagram and the discussion in the text is a bit confusing, however, as it stands as not every bar in the cluster diagram is labeled and it's unclear which "secondary clusters" you are referring to in the text.

Beyond this, the Figure 8 caption says you are evaluating the four methods but it's not entirely clear to me which is represented by which label in the existing cluster tree diagram. I think this figure will need to be modified.

I modified the figure and its caption, making them more clear, and revised the text as suggested.

The results discussed and shown in Figure 9 are helpful for thinking about how different values for these variables change with variation in balance and disturbance. I think this figure could be improved with a bit more guiding text. The x-axis is labeled Cohesion but Admixture is shown on this same axis with the same 0-1 scale so I would suggest changing that. It would also help to label the rows and columns with balance and disturbance respectively. Finally, I would suggest just increasing the border for the Admixture color as it isn't really visible in the plots further to the left.

Also, I think this is an excellent place for you to return to your Table 1 insights and discuss how you might interpret various of these plots in light of that table.

I improved the readability of the figure as suggested (I modified the labels and increased the border of the red shades representing the admixture IQR). The caption of this figure and the text have been re-written, with reference to Table 1 (which has been modified, limiting its use for direct interpretation in terms of post-depositional processes). The correspondence between the Figure with simulated results and the Table has also been enhanced by reordering their rows and colums in similar fashion.

Figure 10: It would be interesting if these were also associated with a non-parametric correlation test like Spearman's rho just to get a sense of how they differ in terms of rank order correlations.

Consider splitting Figure 11 into two figures so that the labels in the boxplot are clearer.

I modified the figure but kept the two plot alongside (to avoid too many figures and, above all, because they represent the same data and it makes sense to have them next to each other — I would have preferred to use margin boxplots, but there is a limitation in ggplot2, it is not possible to use facets and margins plots at the same time). In addition, as suggested, I computed the Spearman coefficients on the ranks of the 10 theoretical graphs computed by the four methods. A table reports the results, which are also discussed in the text.

The empirical example from Liang Abu is informative for thinking about how this all would actually work in practice. Notably, the values for cohesion differ in a pretty obvious way but the other three measures (excluding Modularity) show pretty minor differences (and all show 0 for the 0&1 comparison which makes sense). How might you think about these results in relation to your schema in table 1? Is the difference in Cohesion substantial? What would you interpret as a meaningful difference here? Discussing this a bit in section 3.3.1 would be helpful in walking a reader through what you've done and provide a justification for the hypothesis testing and simulation you do in the next section.

The discussion in this paragraph has been revised and completed accordingly, with reference to Table 1 as suggested and strengthening the relation with the use of simulation in the next section.

The hypothesis testing is an important part of the overall argument and I have some questions and need a bit of clarification in a couple of places here. I followed along with your results and also replicated them in R using the code in the supplement. I found the presentation of these results to be fairly brief given that some major interpretations hinge on them. The set up makes sense as your question is basically whether you could consider this context as having one or two layers initially. You use Wilcox test to assess rank order differences in several variables, which seems like a sensible approach and find evidence for differences in the simulated distributions and not others. Given that you have a low but non-null admixture you are really conducting these simulations to assess differences in cohesion in terms of the schema you laid out in table 1. Since scenario 3 and scenario 4 have the same interpretation (and admixture doesn't differ for either scenario) how do the results you present in this section and Figures 12 and 13 related to the setup you initially provided in table 1? I feel like this needs more discussion. The results for specific variables differ in terms of which hypothesis is more or less supported. The results seem equivocal for several variables and looking at the supplement although there are statistically significant rank order shifts the distributions are quite similar for some variables suggesting the actual effect is small. You land on an interpretation of two independent layers but I think this needs to be better justified. Given the low admixture value and the setup in table 1, is there a scenario that would have generated different results?

Thank you for pointing this. I added a table providing a "grid" for interpretation (Table 7). The plots showing the simulated and empirical results by parameter now also include a boxplot, and the Interquartile range is used to get a better reading of the relation between the empirical value and the simulated values, and finally a more objective interpretation. However, as you suggest in your comment, the interpretation is (must be) left to the archaeologist, since the final assertion is qualitative. I stress this point in the text and present a more prudent interpretation for the Liang Abu layers 1 and 2 case study.

I would also be interested in seeing an exploration of the relationship between structural admixture and the absolute difference in cohesion for a range of scenarios.

this empirical example is only able to capture a small range of the possible scenarios that might be encountered in real settings. In general, I would like to see what kind of numbers would generate higher admixture values and how those might compare to what is shown here.

For sure, the Liang Abu case study is limited. To present a wider range of possibilities is the role of Figure 9 (now Figure 10), unless I do not understand exactly what do you mean here.

In general, the set up in the beginning of the paper outlines the relationship between admixture and cohesion in table 1 but this doesn't come in to play in the interpretation of the empirical example and I think this makes the final interpretation of the hypothesis test less impactful. I would suggest returning to the table 1 set up in the hypothesis testing section and perhaps even discussing in more detail other realistic simulated datasets that would land you somewhere else on table 1.

I think the major thing that could be improved is the connection between the initial discussion of the method (section 2.2 in particular) and the hypothesis testing and simulation in the end.

I tried to satisfy this general demand (see answers to the previous points).

In addition to the text and supplement, I also installed and reviewed the R package. I was able to install this but I had to roll back to and earlier (3.6) version of R as one dependency was not available for the most recent version (4.0) of R. If that is an easy fix (looking at the BiocManager packages) it would be helpful to potential future users. As for the supplement, I was not able to fully replicate those results because there are a couple of files that were called in the code that weren't provided on the OSF link as the markdown document in pdf was the only document posted. I think it would be useful to post the supplement as an Rmd file along with all of the required files to completely replicate the results. Once I was able to get the package installed in R 3.6, I was able to run all of the functions without any problems.

The compatibility of the archeofrag package is checked with continuous integration procedues (on github for macOS, Windows, and linux (Ubuntu 20.04); here: https://github.com/sebastien-plutniak/archeofrag/actions/runs/785741281, and on Travis CI for Linux Ubuntu (18.04.5 LTS), here: https://travis-ci.org/github/sebastien-plutniak/archeofrag). I hope that other users will not encounter troubles and will continue to maintain the package in the future.

Decision round #1

Dear Dr Plutniak,

Thank you very much for your submission to PCI Archaeology. Apologies for the amount of time it took to contact you after your initial submission. It has been very difficult to find adequate reviewers for your paper as it combines techniques and expertise that are not commonly found together in individual researchers.

I personally find the paper innovative and of interest to PCI Archaeology readers and users. After reading the positive reviews I would like to recommend it for publication after some minor issues, as detailed by reviewers 2 and 3, have been addressed. These changes should not take much of your time as they seem relatively minor.

I hope you decide to provide a corrected version of your manuscript.

All best wishes,

Hector

Reviewed by Robert Bischoff, 2021-02-09 16:53

This paper describes a novel method to assess the integrity and amount of mixing between excavation layers. The method, termed the TSAR approach (Topological Study of Archaeological Refitting), uses network typologies and edge metrics to assess the level of cohesion and admixture between two layers. The nodes (artifacts) are connected based on object refitting and adjusts for sample size. The utility of this method compared to other approaches is tested using simulated data and excavation data. The author created a package in the R programming language to facilitate the method, which is

freely available from R's package archive and includes some of the data used in the article. Furthermore, the R package allows users to generate their own test scenarios. Assessing stratigraphic integrity is often a critical task for interpretation. Refitting alone is enough to determine whether some mixing occurred but a quantitative assessment of the mixing requires a more complex methodology. This is the problem addressed by the author. I found the overview of current methods and the comparison to the proposed TSAR method convincing and reasonable. The application of network methods is, to my knowledge, a novel approach that solves a challenging problem and reveals interpretable results. The use of both simulated data and field data to test the method strengthens the author's argument. As stated in the article, testing this method in additional scenarios is necessary but not required to justify recommending this paper. I applaud the inclusion of a pdf containing the code used for analysis and explaining the steps. The development and publishing of an R package based on this method is a significant contribution. I would also appreciate the Rmd file as an additional supplemental document for easier reproducibility, but the code can be copied from the pdf. I recommend endorsing this paper and found no major concerns that should be addressed. There are a few minor corrections I have suggested included in the pdf of the article that I have submitted.

Download the review (PDF file)

Reviewed by anonymous reviewer, 2021-02-08 09:00

I think this is a valuable paper that should be published. The paper proposes a novel method to evaluate the stratigraphic integrity of archaeological deposits by combining refitting studies with graph theory. Usually, refitting work has been an ad hoc process providing some clues for excavators on how archaeological deposits might be related. This paper add significant methodological rigor to realise the full potential of refitting studies to assess the stratigraphy at complex sites. The most immediate relevance is probably for Palaeolithic cave sites and the like, but the method is potentially relevant to a wide range of situations.

I have to point out though I am not a statistician or particularly well versed in mathematics or graph theory. The calculations in the paper are therefore a bit beyond my abilities. I therefore hope that another reviewer was assigned to this paper that has this expertise and can comment on the mathematical aspects in more detail.

Three points I would like to raise are these: 1) On Page 3, the first Figure and the discussion related to it are somewhat abstract. I would suggest to add a real-world archaeological example to explain to the dumb folks like me what this actually means in a concrete archaeological example. Otherwise the abstraction might scare people off. 2) It would be nice, although not absolutely necessary, to see how this method could be deployed in conjunction with other methods of stratigraphic analysis, as well as Bayesian modelling of radiocarbon dates. I would think that this method would be extremely useful in fine-tuning the modelling of C14 dates. It would be useful if the author could maybe comment on that and give an idea how this method might be used in combination with others? 3) I guess this method is still under development, but I wonder if the author might be able to comment a little more on how we could make this method something more commonplace? At the moment, it seems fairly inaccessible, so what will be the pathway to making it more accessible to a wide range of archaeologists (especially those like me who are a bit challenged by the maths)?

In conclusion, I think this is a solid paper and I would recommend publishing it. It can be published as it is, but perhaps the author can take some of the comments I made above into consideration.

Reviewed by Matthew Peeples, 2021-02-19 22:43

Review of "The strength of parthood ties. Evaluating archaeological layers using graph theory to model objects refitting" by Sébastien Plutniak

This paper presents a novel approach to assessing the mixing of archaeological layers using refitting analysis and the topology of refits. This is certainly different than most previous approaches which, as the author notes, rely largely on assessing counts of refits but not the structure connections among them. I think this approach is interesting and has the potential to be useful in helping researchers think thorough and simulate different scenarios involving fragmentation, deposition, and post-depositional processes to aid in the interpretation of real archaeological data. There are some changes I would suggest that I think could help to improve the clarity and usability of this manuscript and I outline those in detail here. I go through each section of the paper and then return to some overarching comments on the method and presentation at the end. The introductory section to the paper is clear and presents a good justification for why this approach could be useful and how it expands upon the existing literature. I am not particularly familiar with the body of literature cited on refits and disturbance processes, but the article cites several classic papers and a range of newer studies that capture much of the literature I have seen.

The "preliminary definitions" section includes some useful information, but I also think this could be reduced not just to save space but also to add to the clarity of the argument. The distinction between "connection" and "similarity" is useful but I found the discussion of this distinction a bit confusing without a second glance. The reference to connections relying on "solid-state physics" and then the next sub-section with the discussion of Queen and Rook contiguity and the analogy to spatial analysis made a pretty clear distinction a bit more muddled in my reading. I section 2.1.3 could be basically eliminated and section 2.1.2 could be simplified to make the two basic points that are necessary for the rest of the paper 1) archaeologists identify potential refits by either identifying pieces that fit together across substantial surfaces in 3D space or alternatively by looking for similarities in the objects themselves and 2) you are only focused on the former in this paper. I think a very simple definition of connection and similarity refits and Figure 2 is plenty and felt that the Queen/Rook example did not add clarity and was not referred to again so could be removed.

Section 2.2 was straight forward and Table 1/Figure 4 are helpful. The method chosen for edge weighting makes sense and is clearly described with the equation as are the definitions for cohesion and admixture. The measures selected for assessing the topological properties of refits make intuitive sense as do the alternative methods and the criticisms of them (such as the issue of sample size and modularity).

The TSAR simulator is a sensible approach to evaluating the potential generative processes creating empirical patterns. There are several extensions that I could think of to this procedure (and some of those are mentioned in the conclusions) but this model is a good beginning. The parameter descriptions in table 2 and the flowcharts in Figure 6 make sense and help me understand what is happening under the hood with one exception. Reading this text, it wasn't clear to me how the "merge graphs" step in the two initial layers works. I can think of several ways you might have chosen to implement this, but I do not think the choices you made are completely clear in this draft or the supplement. I think a short paragraph would be needed to address this.

One major question I have about this approach that is not discussed here or elsewhere is the degree to which missing connections would hamper our ability to connect empirical patterns to generative processes. Given the nature of fragmentation and post-depositional processes I expect that it would be common for archaeologists to miss refits due to damage to potential conjoining edges of pieces or all manner of other processes. How would this influence your results? I think this issue merits some discussion at least and perhaps some small analyses. I had one suggestion the author could consider here. Missing connections could potentially be evaluated with an additional global parameter such as a probability that a connection is removed once the graphs have been formed to essentially simulate "missed" or unidentifiable connections. Especially with the focus on "connections" rather than similarity, I think it's likely that missed connections would be present in any empirical refit study and this simulation model provides a setup that could help evaluate the impacts. I'd be interested in how this influences the shape of distributions in Figure 9, for example.

Moving on to the examples, I think the theoretical graphs, the Liang Abu empirical example, and simulation based on the empirical example provide a good three-pronged approach to evaluating the technique. I have some questions about the results that I think could require some clarification in the text.

For the first analysis where archaeologists were asked to sort the theoretical graphs you got results that were extremely variable which is not surprising to me because it seems a pretty difficult task to evaluate these kinds of topological patterns on sight. In section 3.1.1 you compare the admixture model and others to the archaeological intuition. Given the variability in the archaeological intuitive coding, I don't know that the methods used here fitting in the same cluster is a particularly convincing argument in favor of these new methods (and that's okay). To me, the results you present suggest that people are generally not great at evaluating the relative order of topological patterns and that, in and of itself, argues for some sort of automated approach like you've developed here. The clustering diagram and the discussion in the text is a bit confusing, however, as it stands as not every bar in the cluster diagram is labeled and it's unclear which "secondary clusters" you are referring to in the text. Beyond this, the Figure 8 caption says you are evaluating the four methods but it's not entirely clear to me which is represented by which label in the existing cluster tree diagram. I think this figure will need to be modified.

The results discussed and shown in Figure 9 are helpful for thinking about how different values for these variables change with variation in balance and disturbance. I think this figure could be improved with a bit more guiding text. The x-axis is labeled Cohesion but Admixture is shown on this same axis with the same 0-1 scale so I would suggest changing that. It would also help to label the rows and columns with balance and disturbance respectively. Also, I think this is an excellent place for you to return to your Table 1 insights and discuss how you might interpret various of these plots in light of that table. Finally, I would suggest just increasing the border for the Admixture color as it isn't really visible in the plots further to the left.

Figure 10 works and makes the point discussed in the text well. It would be interesting if these were also associated with a non-parametric correlation test like Spearman's rho just to get a sense of how they differ in terms of rank order correlations. Consider splitting Figure 11 into two figures so that the labels in the boxplot are clearer. The differences in the stabilization of disturbance values for edge count and admixture (edge betweenness) are interesting.

The empirical example from Liang Abu is informative for thinking about how this all would actually work in practice. Notably, the values for cohesion differ in a pretty obvious way but the other three measures (excluding Modularity) show pretty minor differences (and all show 0 for the 0&1 comparison which makes sense). How might you think about these results in relation to your schema in table 1? Is the difference in Cohesion substantial? What would you interpret as a meaningful difference here? Discussing this a bit in section 3.3.1 would be helpful in walking a reader through what you've done and provide a justification for the hypothesis testing and simulation you do in the next section.

The hypothesis testing is an important part of the overall argument and I have some questions and need a bit of clarification in a couple of places here. I followed along with your results and also replicated them in R using the code in the supplement. I found the presentation of these results to be fairly brief given that some major interpretations hinge on them. The set up makes sense as your question is basically whether you could consider this context as having one or two layers initially. You use Wilcox test to assess rank order differences in several variables, which seems like a sensible approach and find evidence for differences in the simulated distributions and not others. Given that you have a low but non-null admixture you are really conducting these simulations to assess differences in cohesion in terms of the schema you laid out in table 1. Since scenario 3 and scenario 4 have the same interpretation (and admixture doesn't differ for either scenario) how do the results you present in this section and Figures 12 and 13 related to the setup you initially provided in table 1? I feel like this needs more discussion. The results for specific variables differ in terms of which hypothesis is more or less supported. The results seem equivocal for several variables and looking at the supplement the although there are statistically significant rank order shifts the distributions are quite similar for some variables suggesting the actual effect is small. You land on an interpretation of two independent layers but I think this needs to be better justified. Given the low admixture value and the setup in table 1, is there a scenario that would have generated different results? I would also be interested in seeing an exploration of the relationship between structural admixture and the absolute difference in cohesion for a range of scenarios.

In general, the results outlined here make sense for this particular example, but I think this empirical example is only able to capture a small range of the possible scenarios that might be encountered in real settings. In general, I would like to see what kind of numbers would generate higher admixture values and how those might compare to what is shown here. In general, the set up in the beginning of the paper outlines the relationship between admixture and cohesion in table 1 but this doesn't come in to play in the interpretation of the empirical example and I think this makes the final interpretation of the hypothesis test less impactful. I would suggest returning to the table 1 set up in the hypothesis testing section and perhaps even discussing in more detail other realistic simulated datasets that would land you somewhere else on table 1.

Overall, I think this method is potentially useful and most of the implementation was described in an understandable way. I think the major thing that could be improved is the connection between the initial discussion of the method (section 2.2 in particular) and the hypothesis testing and simulation in the end.

In addition to the text and supplement, I also installed and reviewed the R package. I was able to install this but I had to roll back to and earlier (3.6) version of R as one dependency was not available for the most recent version (4.0) of R. If that is an easy fix (looking at the BiocManager packages) it would be helpful to potential future users. As for the supplement, I was not able to fully replicate those results because there are a couple of files that were called in the code that weren't provided on the OSF link

as the markdown document in pdf was the only document posted. I think it would be useful to post the supplement as an Rmd file along with all of the required files to completely replicate the results. Once I was able to get the package installed in R 3.6, I was able to run all of the functions without any problems.