# 2nd Interactive comment on "Improving the understanding of flood risk in the Alsatian region by knowledge capitalization: the ORRION participative observatory" by Florie Giacona et al.

June 5, 2019

## 1 General comments

I am happy to see that the authors took many of the comments made by the reviewers to improve clarity and reproducibility of the paper. I have, however, still three main points to suggest, see below.

My suggestion is minor revisions with the following points to reconsider

1. conditional means/probabilities instead of biserial point correlation,
2. keep the full categorical variables in the data set published instead of their representation by a set of binaries (which you might do anyway),
3. rescale the moving averages to the same units as the observations.

These suggestions are argumented for below.

## 2 Detailed comments

### 2.1 Pearson's correlation

I suggested to the authors not to use Pearson's correlation for dichotomous and categorical variables. My line of argument was two-fold a) not straight-forward to interpret, as the generic reader is probably used to use this coefficient for continuous variables, b) significance test questionable. Instead, I suggested a simpler and more robust method: obtain mean values conditioned on a dichotomous variable taking the value 1 and another mean value for this variable taking the value 0; compare these mean values by means of the $t$-test. As this involves only first moments (mean), I consider it as simpler than correlation (second moment) and easier to interpret/understand. If appropriate for the analysis, the conditional means can be exchanged for conditional probabilities. This can be very easily extended to more than two categories.

Now the authors argue that they want to remain with Pearson because a) they want to do the same analysis for all variables, b) alternative metrics like biserial point, Kendall or Spearman lead to the "same measure of association" (p.12, l.28), and c) they want to keep the analysis simple and easy to understand.

a) I can see their point for equal treatment of variables as it is simple. But with the same line of argument you could reduce your cutlery to having only a fork: equal treatment for all food. It will fail for a soup, however.

b) Alternative metrics do lead to alternative measures of association (unlike your statement p.12, l.28), that is why they are not the same metric. Here, the *numerical value* obtained by calculating these metrics for a given series happens to be the same. For example, Pearson describes the linear association, Spearman the monotonic association. If you happen to obtain a numerical value of 0.3 for both, not the value, but the interpretation differs.

c) In their reply to my comments on Pearson's correlation, the authors explain that an appropriate association for a continuous and a dichotomous variable is the biserial point correlation and, doing the math, the formula is the same as the one for Pearson. So why do these coefficients have different names? I think the reason is that it is a priori not clear that using Pearson for the binary case makes sense. And that is probably also a question the generic reader might have. To be consise, the authors need to explain this in the paper. I did not know that there is a dedicated name for this concept. However, if you look at the derivation of the biserial coefficient, you can see that it reduces to a comparison of conditional means, which was in fact my earlier suggestion. Calling for an easy analysis which is understandable w/o introducing new concepts (like the biserial point correlation), I still think the comparision of conditional means would do the job here.

## 2.2 Categorical vs. binary

I don't think that there is something fundamentally wrong with encoding a categorical variable as a set of binary variables. I found it just an awkward choice in cases, where only one out of the set of binary variables can take on the value 1 while the others then must be 0. It is an unnecessary increase in complexity of the problem. Consider, for example, a categorical variable with 3 states A, B and C; there are 3 possible values for this variable. If you encode it with three binary variables A, B and C wich can all three take on the values 0 and 1, you have a system with $2^3 = 8$ possible values from which only 3 are taken. There is nothing wrong with it but I find it a strange choice. But as I understand, the original categorical variabele is still part of the data set to be published? That would be very useful for all those people who do not want to reduce their analysis from a categorical variable to a set of binaries!

## 2.3 Standardized moving averages

From my point of view, moving averages are used to smooth a series, particularly small and large values do not appear. That is why the moving average has by definition a smaller variance. If you want the moving average to be visually comparable to your count observations, just plot them using the same units, e.g. counts/year. I expect your 31-year moving average might have the unit counts/31 years? Does a multiplication with 31 bring it to similar values as the observed counts? It might, as the ratio of variances should be 31. I also wonder, that you have values for the moving averages also at the borders of your series. Is the window than reduced?