Review of manuscript “A sensitivity analysis of the PAWN sensitivity index”

In response to the issues raised in the previous round of reviews, the authors made some changes to their manuscript. Unfortunately, in our opinion these changes are not fully satisfactory and some crucial points have remained unaddressed. Indeed new questions have arisen now that a clearer visualisation in Fig. 2 has enabled us to better inspect the results.

The authors’ key highlights are that “*PAWN is more sensitive (than Sobol’) to its design parameters*” and that “*This sensitivity increases the risk of PAWN yielding biased results*”. These conclusions are essentially based on the results in Figure 2, and particularly the comparison of the PAWN sensitivity indices in the second panel (PAWN in its most obvious set-up) and Sobol’ ones (bottom planel). However, the distributions in the second-row **shows that the confidence intervals of PAWN indices are essentially as small as Sobol’ (if not smaller). The statement that PAWN is sensitive to design parameters is solely supported by the fact that PAWN consistently shows many more outliers** than Sobol’, particularly on the high end of the distribution.

We therefore analysed these outliers, using the dataset *AB.Pawn* generated by the R code made available by the authors (our own code to perform the further analysis is attached). In particular, we investigated whether the outliers correspond to any particular combination of the tuning parameters N (sample size) and n (number of conditioning intervals). We found this is indeed the case. Figure B below shows the (n,N) combinations that generate the outliers in the PAWN index of X1 for the four benchmark functions. (outliers are defined as exceeding the threshold of S1=0.41 for the Liu function, 0.53 for Homma, 0.57 for Sobol’, and 0.21 for Morris). Figure B shows these combinations are all concentrated in the bottom-right corner of the (n,N) space, meaning that they correspond to experiments where “many” conditional distributions (high n) are estimated from “few” datapoints (low N, and thus low N/n). In other words, **outliers (= badly approximated PAWN indices) are obtained when using too few datapoints to estimate the conditional distributions. This is not surprising and seems to point to a badly designed experiment rather than a fundamental problem of the PAWN method**.

To confirm our guess, we did some more analysis, shown in Figure C. Here we report the boxplots of the PAWN indices after discarding “badly designed” experiments – defined in two possible ways: experiments when the ratio N/n is lower than 50 (less stringent criterion; top panel) or lower than 80 (more stringent; middle panel). The results show that, indeed, **as we discard experiments with a low N/n ratio, the outliers reduce in both their number and their distance from the mean values**, and the overall distributions of PAWN indices gets narrower and narrower – actually way narrower than the Sobol’ ones (for the sake of comparison, we have also copied below a screenshot of the bottom panel of Figure 2 in the manu script, showing Sobol’ results).

In summary, it seems to us that **a more in-depth analysis of the results does not really support the claim that PAWN is more sensitive than Sobol’ to its tuning parameters (if anything, Figure C below seems to suggest the opposite). At most, the results show that a sensible application of PAWN requires the user to pay attention when selecting n and N, so to ensure a “sufficiently high” N/n ratio**. In real world applications, this will likely be achieved by adjusting n, given that the user has less flexibility in choosing N. Also, notice that a condition such as “N/n>80” is not particularly stringent: for example if N=1000 (a relatively small sample size for GSA standards) then n=10 (the suggested “default” for PAWN) would comfortably satisfy it.

Based on the above discussion, we also raise the issue whether the overall manuscript provides enough novel contribution for publication in EMS. As we have shown, once the variability space of the PAWN parameters (n,N) is defined reasonably (i.e. avoiding combinations of low N and high n that a user should not try, and that would anyway be singled out as critical when using a dummy parameter or bootstrapping as we propose as standard in our implementation within the SAFE toolbox), the PAWN indices are indeed quite well approximated – despite the residual uncertainties in n, N, epsilon and theta.

Francesca Pianosi & Thorsten Wagener – 2 February 2020

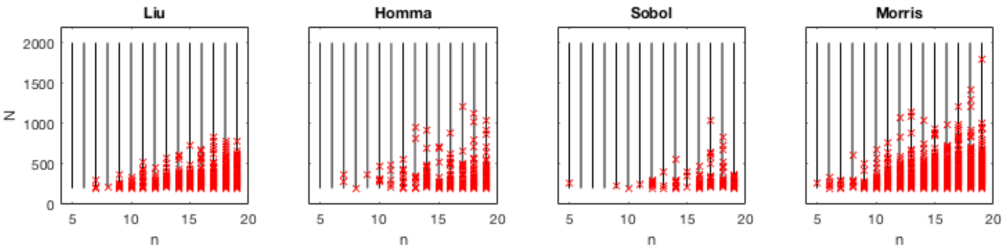


Figure B: Combinations of tuning parameters (n,N) that generated the outliers of the PAWN index of input X1 for the four benchmark functions (outliers are defined as exceeding the threshold of S1=0.41 for the Liu function, 0.53 for Homma, 0.57 for Sobol', and 0.21 for Morris).

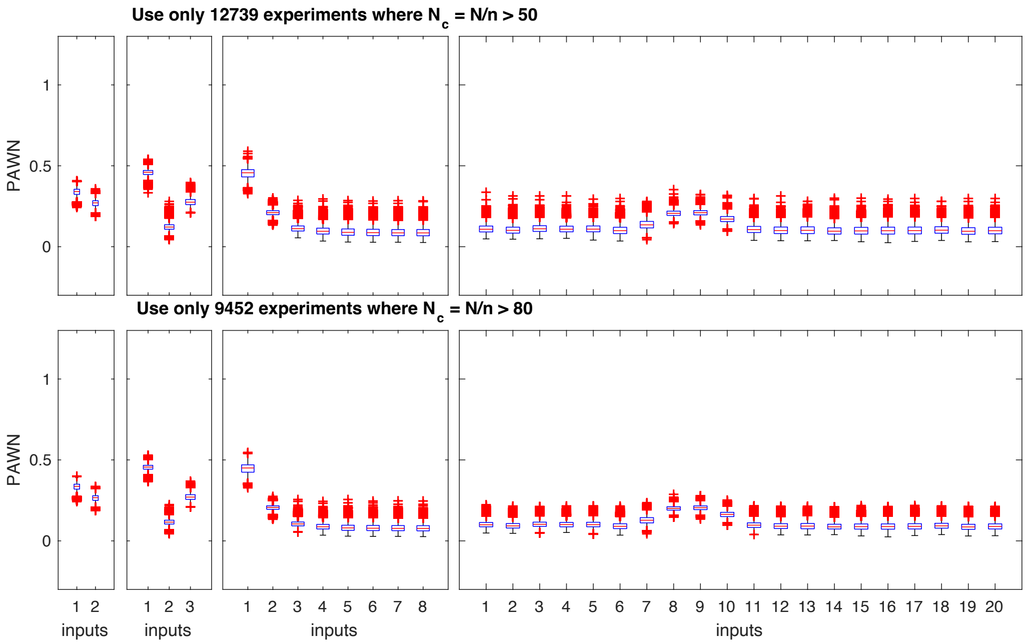
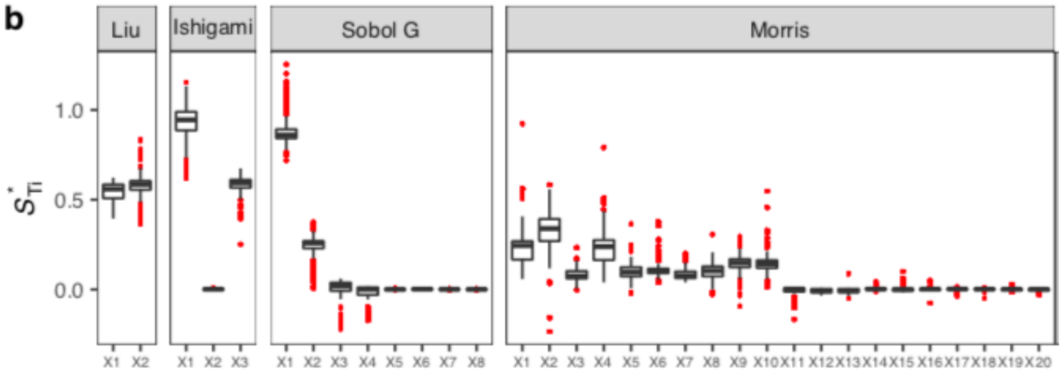


Figure C: Boxplots of PAWN indices when discarding "badly designed" experiments – defined as having a number of datapoints Nc in each conditional sub-sample lower than 50 (top) or 80 (bottom). As the criterion becomes more stringent, a lower number of samples is retained (12,739 out of the original 16,384 in the first case, 9,452 in the second).



Bottom panel of Figure 2 in the manuscript (screenshot): boxplots of Sobol' indices