MASCOT-NUM 2022
Clermont Ferrand, France
7-9 June 2022

When Global Sensitivity Analysis provides insights into Group Fairness

C. Bénesse
Institut de Mathématiques de Toulouse (IMT), France

Supervisor(s): Prof. F. Gamboa (IMT) and Prof. J-M. Loubes (IMT)

PhD expected duration: Oct. 2019 – Dec. 2022

Address: 114 Route de Narbonne, 31400 Toulouse

E-mail: clement.benesse@math.univ-toulouse.fr

Abstract: The European Commission has recently published drafts for a new wave of laws concerning AI systems – namely the *White Paper on Artificial Intelligence* [1] and the *Artificial Intelligence Act* [2]. In these documents, lawmakers put emphasis on the need to "correctly interpret the [...] AI system's output, taking into account the characteristics of the system and the interpretation tools", by using "preliminary defined metrics and probabilistic thresholds". These provisions, outlined for *interpretable and fair AIs* echo classical sentences and expressions found in the Uncertainty Quantification (UQ) literature. In fact, Global Sensitivity Analysis (GSA) tools were first developed, and are still used, as a means for interpretability; and it has been shown that these techniques can be used for the quantification of Fairness, especially Group Fairness [3, 5].

In this presentation, we begin by presenting new results explicited below in the GSA framework. We build upon the link mentioned above to translate them in a Fairness framework. This allows us to show how developments proposed by the UQ community can have direct fallouts in lawmaking.

We expend here on two recent and impactful methods of interest.

- Surrogate models or meta-models – denoted by $\hat{f}$ – recently gained traction as a means to obtain a low-resource approximation of expensive computer codes – denoted by f – so as to obtain critical information and to infer characteristics of a system. GSA indices are usually computed to know which variables are the main drivers behind the computer code output. It is therefore necessary to know if the indices of the surrogate model are close to the indices of the true model. We extend a partial result proposed for the special case of Sobol' indices with independent inputs [9]. We provide upper bounds and rates of convergence on the error for various extensions of Sobol' indices – namely Extended Sobol' indices, Extended Cramér-von-Mises indices and Shapley indices – in a framework where inputs are no longer independent. The bounds we obtain are driven by the classical quadratic risk $||f - \hat{f}||_2$. For instance, for Extended Sobol' indices, we have

$$|GSA(f) - GSA(\hat{f})| \leq \frac{||f - \hat{f}||_2}{\mathrm{Var}(f)}, \tag{1}$$

  when we denote by $GSA(\varphi)$ one of the Extended Sobol' index computed for a given algorithm φ. Because of this fact, classical literature on non-parametric regression yields rates of convergence for data-driven metamodels minimizing an empirical loss.

  In a Fairness framework, these results quantify how fair an algorithm is, even if access is impossible in itself, by only using an approximation or surrogate model. These techniques enable lawmakers for instance to audit industrial algorithms.

- A classical assumption of GSA is that the inputs' distribution is perfectly known – that is $\mathbf{X} = (X_1, \cdots, X_p) \sim \mathbb{P}_{\mathbf{X}}$ with $\mathbb{P}_{\mathbf{X}}$ fixed. If we remove this assumption, another level of uncertainty arises, as distributional change can modify the assumed influence of an output. This can happen for instance if we change the underlying distribution of our data. As often in UQ, solutions for

this issue can be found on a local level [7], or a global level [8]. The latter is called "Second-Level GSA". In Second-Level GSA, for instance, we want to assess how much the allocated influnce of X_i on the output $f(\mathbf{X})$ is subject to distributional changes of $\mathbb{P}_{\mathbf{X}}$, along a range of possible distributions. When working in a parametric setting – that is when X_i is distributed according to a law $\mathbb{P}_{\theta_i}, \theta_i \in \Theta_i$ – we obtain this information by computing quantities of the form

$$GSA_{\theta_i}(GSA_{X_i}(f)). \tag{2}$$

We provide two different estimators for these quantities, one based on the now-classical Pick'n'Freeze method and the other on the Chatterjee correlation estimator [4, 6], when the GSA index is Sobol'. Converting this idea in a Fairness framework, these indices tell if an algorithm remains fair after distributional changes in the data, for instance when applied to a different population.

References

[1] White Paper on Artificial Intelligence: a European approach to excellence and trust.

[2] Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021.

[3] Clément Bénesse, Fabrice Gamboa, Jean-Michel Loubes, and Thibaut Boissin. Fairness seen as Global Sensitivity Analysis. *arXiv:2103.04613 [math, stat]*, September 2021. arXiv: 2103.04613.

[4] Sourav Chatterjee. A New Coefficient of Correlation. *Journal of the American Statistical Association*, 0(0):1–21, April 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1758115.

[5] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1229–1239. Curran Associates, Inc., 2020.

[6] Fabrice Gamboa, Pierre Gremaud, Thierry Klein, and Agnès Lagnoux. Global Sensitivity Analysis: a new generation of mighty estimators based on rank statistics. *arXiv:2003.01772 [math, stat]*, March 2020. arXiv: 2003.01772.

[7] Joseph Hart and Pierre Gremaud. Robustness of the Sobol' indices to distributional uncertainty. *arXiv:1803.11249 [math, stat]*, November 2018. arXiv: 1803.11249.

[8] Anouar Meynaoui, Amandine Marrel, and Béatrice Laurent. New statistical methodology for second level global sensitivity analysis. *arXiv:1902.07030 [math, stat]*, February 2019. arXiv: 1902.07030.

[9] Ivan Panin. Risk of estimators for Sobol' sensitivity indices based on metamodels. *Electronic Journal of Statistics*, 15(1):235–281, January 2021. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Short biography – Clément Bénesse is currently a third-year PhD student at the *Institut de Mathématiques de Toulouse*. His research interests revolve around Global Sensitivity Analysis and Algorithmic Fairness, and more precisely on how to merge these two literatures. His academical works are co-funded by a *Contrat Doctoral Spécifique Normalien* from *ENS Lyon* and by the 3IA *ANITI*, based in Toulouse.