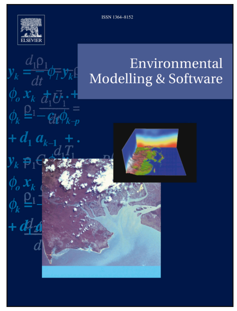
Accepted Manuscript



Tree-based ensemble methods for sensitivity analysis of environmental models: A performance comparison with Sobol and Morris techniques

Marc Jaxa-Rozen, Jan Kwakkel

PII: S1364-8152(17)31158-1
DOI: 10.1016/j.envsoft.2018.06.011
Reference: ENSO 4241

To appear in: *Environmental Modelling and Software*

Received Date: 31 October 2017

Revised Date: 26 May 2018

Accepted Date: 1 June 2018

Please cite this article as: Jaxa-Rozen, M., Kwakkel, J., Tree-based ensemble methods for sensitivity analysis of environmental models: A performance comparison with Sobol and Morris techniques, *Environmental Modelling and Software* (2018), doi: 10.1016/j.envsoft.2018.06.011.

Tree-based ensemble methods for sensitivity analysis of environmental models: a performance comparison with Sobol and Morris techniques

Marc Jaxa-Rozen[a,*], Jan Kwakkel[a]

[a] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Delft 2628 BX, The Netherlands

Abstract

Complex environmental models typically require global sensitivity analysis (GSA) to account for non-linearities and parametric interactions. However, variance-based GSA is highly computationally expensive. While different screening methods can estimate GSA results, these techniques typically impose restrictions on sampling methods and input types. As an alternative, this work evaluates two decision tree-based methods to approximate GSA results: random forests, and Extra-Trees. These techniques are applicable with common sampling methods, and continuous or categorical inputs. The tree-based methods are compared to reference Sobol GSA and Morris screening techniques, for three cases: an Ishigami-Homma function, a H1N1 pandemic model, and the CDICE integrated assessment model. The Extra-Trees algorithm performs favorably compared to Morris elementary effects, accurately approximating the relative importance of Sobol total effect indices. Furthermore, Extra-Trees can estimate variable interaction importances using a pairwise permutation measure. As such, this approach could offer a user-friendly option for screening in models with inputs of mixed types.

Keywords: Global sensitivity analysis, factor screening, decision tree methods, ensemble learning methods

1. Introduction

Sensitivity analysis (SA) is recognized as a key step for analyses which involve the assessment and propagation of uncertainty in mathematical models (Frey and Patil, 2002; Helton and Oberkampf, 2004). In particular, techniques for global sensitivity analysis (GSA) have become an accepted standard for the evaluation of the impact and interactions of uncertain inputs in complex environmental models (as described in this journal by e.g. Saltelli and Annoni 2010; Nossent et al. 2011; Pianosi and Wagener 2015). These techniques consider the output behaviour of the model over the full domain of uncertain inputs; specifically, this implies that the full distribution of each input parameter should be evaluated, and that the importance of each input should be evaluated across the domain of all other parameters (Liu and Homma, 2009). This is in contrast to "one-at-a-time" (OAT) sensitivity analysis – which focuses on response to changes in individual inputs around an initial baseline value, and which for instance inadequately captures non-additive responses caused by interactions between input parameters. These properties make GSA particularly relevant for applications such as integrated assessment models, which frequently combine a large number of highly uncertain inputs with a non-linear, non-additive structure. In these conditions, a OAT analysis can lead to an incomplete or misleading interpretation of model uncertainty. As such, GSA can help analysts and decision-makers better understand and communicate the results of complex models, and ultimately make these models more credible in a decision support context. However, the computational cost of existing GSA methods can quickly become prohibitive with complex simulation models.

This paper therefore draws on the statistical learning literature to evaluate the performance of decision tree-based ensemble methods, when applied to typical sensitivity analysis problems. These methods rely on ensembles of decision trees which match partitions of the input space with a predicted output, and are commonly implemented using the random forests and Extra-Trees algorithms (Breiman, 2001; Geurts et al., 2006). These techniques perform well at relatively small sample sizes for non-linear regression or classification problems in which input interactions are significant; they are also able to handle both numerical and categorical inputs (Louppe, 2014). Building on previous investigations of decision tree methods for sensitivity analysis (e.g. Harper et al., 2011), this paper will show that these methods can replicate some of the key insights of GSA by estimating relative variable importances and interactions, at a much smaller computational cost.

In the context of GSA, Saltelli et al. (2008) summarize four analysis objectives, or "settings": i) factor prioritiza-

*Corresponding author
Email addresses: m.jaxa-rozen@tudelft.nl (Marc Jaxa-Rozen), j.kwakkel@tudelft.nl (Jan Kwakkel)