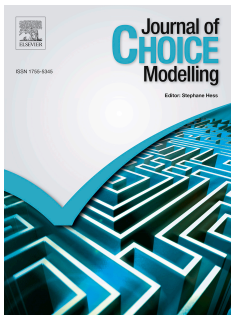
Accepted Manuscript



Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis

Ahmad Alwosheel, Sander van Cranenburgh, Caspar G. Chorus

PII: S1755-5345(18)30005-8

DOI: 10.1016/j.jocm.2018.07.002

Reference: JOCM 149

To appear in: *Journal of Choice Modelling*

Received Date: 15 January 2018

Revised Date: 10 July 2018

Accepted Date: 11 July 2018

Please cite this article as: Alwosheel, A., van Cranenburgh, S., Chorus, C.G., Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis, *Journal of Choice Modelling* (2018), doi: 10.1016/j.jocm.2018.07.002.

Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis

Ahmad Alwosheel[a1], Sander van Cranenburgh[a], Caspar G. Chorus[a]

[a]*Transport and Logistics Group, Department of Engineering Systems and Services, Delft University of Technology*

Abstract:

Artificial Neural Networks (ANNs) are increasingly used for discrete choice analysis. But, at present, it is unknown what sample size requirements are appropriate when using ANNs in this particular context. This paper fills this knowledge gap: we empirically establish a rule-of-thumb for ANN-based discrete choice analysis based on analyses of synthetic and real data. To investigate the effect of complexity of the data generating process on the minimum required sample size, we conduct extensive Monte Carlo analyses using a series of different model specifications with different levels of model complexity, including RUM and RRM models, with and without random taste parameters. Based on our analyses we advise to use a minimum sample size of fifty times the number of weights in the ANN; it should be noted, that the number of weights is generally much larger than the number of parameters in a discrete choice model. This rule-of-thumb is considerably more conservative than the rule-of-thumb that is most often used in the ANN community, which advises to use at least ten times the number of weights.

1. Introduction

Artificial Neural Networks (ANNs) are receiving an increasing interest from the choice modelling community to analyse choice behaviour in a variety of contexts (e.g., Hagenauer & Helbich, 2017; Hensher & Ton, 2000; Mohammadian & Miller, 2002; Van Cranenburgh & Alwosheel, 2017). This recent and profound increase in interest is due to 1) a range of recent innovations in ANN research – leading to improved performance; 2) the availability of "click-n'play" software to work with ANNs; 3) a rapid increase in computational resources, and 4) the increasing volumes and diversity of data which is at the disposal of choice modellers; this latter aspect being the core focus of the current special issue in the Journal of Choice Modelling.

To successfully train ('estimate' in choice modellers' parlance) and use ANNs, the dataset (on which the ANN is trained) needs to be sufficiently large (i.e., consist of a sufficient number of observations). In the ANNs literature such data requirements have extensively been studied (Anthony & Bartlett, 2009; Bartlett & Maass, 2003; Haussler, 1992a), leading to a series of theoretical results regarding lower bounds in terms of data size for a variety of ANNs

[1] a.s.alwosheel@tudelft.nl | +31152783420