

Chapter 33

Integrating Systems Modelling and Data Science: The Joint Future of Simulation and ‘Big Data’ Science

Erik Pruyt

Delft University of Technology, The Netherlands

ABSTRACT

Although System Dynamics modelling is sometimes referred to as data-poor modelling, it often is –or could be– applied in a data-rich manner. However, more can be done in the era of ‘big data’. Big data refers here to situations with much more available data than was until recently manageable. The field of data science makes big(ger) data manageable. This paper provides a perspective on the future of System Dynamics with a prominent place for bigger data and data science. It discusses different approaches for dealing with bigger data. It reviews methods, techniques and tools for dealing with bigger data in System Dynamics, and sheds light on the modelling phases for which data science is most useful. Finally, it provides several examples of current applications in which big data, data science, and System Dynamics modelling and simulation are being merged.

1. INTRODUCTION

Today, ‘big data’ and ‘big data analytics’ are the talk of the day. Are they just the latest hypes? Or will they fundamentally change the analytic world? Do they matter for the field of System Dynamics (SD) (Forrester, 1961; Sterman, 2000; Azar, 2012; Pruyt, 2013), and by extension, the broader field of modelling and simulation? Will they affect the field of modelling and simulation? And if so, how? Or do they distract from dynamic complexity and should they therefore be ignored or kept down? These are some of the questions addressed in this paper. However, before answering any of these questions, I need to shed light on what I mean by big data in the context of SD modelling and simulation. Big data simply refers here to situations in which more data is available than was until recently manageable. Real big data often requires data science techniques to make it manageable and useful. In this paper, it is probably better to

DOI: 10.4018/978-1-5225-1759-7.ch033

refer to data-rich situations and ‘bigger data’ since the data referred to here is often much smaller than in real big data applications. Many data science techniques used for big data are nevertheless also useful for the type of data-rich situations referred to here.

So far, the worlds of data science and SD modelling and simulation have hardly met. Although SD modelling is sometimes referred to as data-poor modelling, it does not mean that SD modelling is per definition, or ought to be, data-poor. To the contrary. SD software packages enable one to get real data from and write model-generated data to databases. Moreover, real data is often used in SD to calibrate parameters (Oliva, 2003) or bootstrap parameter ranges (Dogan, 2007), and sampling techniques are often used to generate ensembles of many simulation runs (Ford, 1990; Clemson *et al.* 1995). These ensembles of runs are in fact model generated bigger data.

In an era of ‘big data’, there may be interesting opportunities for SD. There are at least three ways in which bigger data and data science may play a role in SD: (i) to obtain useful inputs and information from bigger real data, (ii) to infer plausible theories and model structures from bigger real data, and (iii) to analyse and interpret large ensembles of simulation runs (i.e., bigger model-generated data). Interestingly, data science techniques that are useful for obtaining useful inputs and information from bigger real data can also be made useful for analysing and interpreting large ensembles of model-generated bigger data, and vice versa. Hence, both are dealt with here. There are nevertheless also fundamental differences between data science for model-generated data and data science for real data. The most important differences are that: (i) in modelling, the underlying causes are –or could be– known, while that is often not true for real data, (ii) the causes of model-generated data may not relate to the real causes, (iii) in modelling, there is no missing output data, while real data is often missing or of poor quality, and (iv) in modelling, it is possible to generate more data if more data would be needed, which for real data requires a pro-active data gathering strategy or luck.

Adapting and adopting data science techniques may be more than just an opportunity for SD. It may be a necessity, simply because more and bigger real data becomes available, and because some evolutions/innovations in the SD field require more inputs and result in ever bigger sets of intermediate data or model-generated outputs. Examples of such evolutions/innovations within the SD field include: spatially specific SD modelling (Ruth & Pieper, 1994; Struben, 2005) especially if combined with Geographic Information Systems; individual agent-based SD modelling (Castillo & Saysal, 2005; Osgood, 2009; Feola *et al.*, 2012), hybrid Agent-based-SD modelling, and Entity Based SD (Yeager *et al.*, 2014); ‘brute force’ simulation approaches using common sampling techniques (Ford, 1990; Fiddaman, 2002); simulation under deep uncertainty (Lempert *et al.*, 2003; Pruyt & Hamarat, 2010; Auping *et al.*, 2015; Auping *et al.*, 2016), multi-model simulation under deep uncertainty (Pruyt & Kwakkel, 2014; Auping *et al.*, 2012; Kovári & Pruyt, 2014; Pruyt *et al.*, 2015), and multi-method simulation under deep uncertainty (Moorlag *et al.*, 2015); as well as the combination of (multi-objective) robust optimization for adaptive robust planning (Hamarat *et al.* 2013, 2014).

There are many application domains for which the combination of modelling and simulation and big data or data science would be beneficial. Examples, some of which are discussed in this paper, include monitoring and policy making with regard to epidemics of infectious diseases, crime fighting, cyber-security, national safety and security, financial stress testing, market assessment, asset management, future-oriented technology assessment, and even traffic and congestion management.

The remainder of the paper is structured as follows. First, a picture of the joint future of SD and (bigger) data science is briefly presented in section 2. Alternative approaches for dealing with bigger data are presented in section 3. Different data science methods, techniques and tools for dealing with bigger data

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/integrating-systems-modelling-and-data-science/173361

Related Content

Some Linear Fuzzy Collaborative Forecasting Models for Semiconductor Unit Cost Forecasting

Toly Chen (2013). *International Journal of Fuzzy System Applications* (pp. 25-39).

www.irma-international.org/article/some-linear-fuzzy-collaborative-forecasting/76297

Deep Learning Approaches for Earth Science

Deepali Virmani, Ananda Kumar K. S., Sajitha L. P. and Nikhat Parveen (2024). *Novel AI Applications for Advancing Earth Sciences* (pp. 186-203).

www.irma-international.org/chapter/deep-learning-approaches-for-earth-science/336220

Transformative Power of Artificial Intelligence in Decision-Making, Automation, and Customer Engagement

R. Nalini (2024). *Complex AI Dynamics and Interactions in Management* (pp. 189-208).

www.irma-international.org/chapter/transformative-power-of-artificial-intelligence-in-decision-making-automation-and-customer-engagement/339748

Relative Superiority of Key Centrality Measures for Identifying Influencers on Social Media

Yifeng Zhang and Xiaoqing Li (2014). *International Journal of Intelligent Information Technologies* (pp. 1-23).

www.irma-international.org/article/relative-superiority-of-key-centrality-measures-for-identifying-influencers-on-social-media/123941

A Two-Tuple Linguistic Model for the Smart Scenic Spots Evaluation

Li Tang (2023). *International Journal of Fuzzy System Applications* (pp. 1-20).

www.irma-international.org/article/a-two-tuple-linguistic-model-for-the-smart-scenic-spots-evaluation/329959