

Understanding Input Data Requirements and Quantifying Uncertainty for Successfully Modelling ‘Smart’ Cities ^{*}

Nick Malleson^{1,3}[0000–0002–6977–0615], Jonathan A. Ward²[0000–0003–3726–9217],
Alison Heppenstall^{1,3}[0000–0002–0663–3437], Michael Adcock³, Daniel Tang⁴,
Jonathan Coello⁴, and Tomas Crols^{1,3}[0000–0002–9379–7770]

¹ School of Geography, University of Leeds, LS2 9JT, UK
<http://geog.leeds.ac.uk/>
n.s.malleson@leeds.ac.uk

² School of Mathematics, University of Leeds, LS2 9JT, UK
<http://maths.leeds.ac.uk>

³ Leeds Institute for Data Analytics (LIDA), University of Leeds, LS2 9JT, UK
<http://lida.leeds.ac.uk>

⁴ Improbable, 30 Farringdon Road, London, EC1M 3HE, UK
<http://www.improbable.io>

Abstract. Agent-based modelling (ABM) is ideally suited to modelling the behaviour and evolution of social systems. However, there is inevitably a high degree of uncertainty in projections of social systems – input data are noisy and sparse, and human behaviour is itself extremely uncertain – so one of the key challenges facing the discipline is the quantification of uncertainty within the outputs of models. Without an adequate understanding of model uncertainty, nor a means to better constrain models to reality, simulations will naturally diverge from the target system which limits the value of their predictions. This paper presents ongoing work towards methods that will (i) allow real-time data to be assimilated into models to reduce the uncertainty of their predictions and to (ii) quantify the *amount* of data (including overall volume as well as spatio-temporal granularity and regularity) that are required for successful assimilation (i.e. to model the system within an acceptable level of uncertainty). Specifically, this project emulates a simple system of pedestrians who all move towards an exit. The paper reports on initial experiments to constrain the range of possible model outcomes using Bayesian inference techniques as implemented in a new probabilistic programming library. Ultimately the project aims to provide valuable information about the number and type of sensors that might be required to model movements of humans around real urban systems.

Keywords: Agent-based modelling · Uncertainty · Data assimilation · Bayesian inference

^{*} This work was supported by a European Research Council (ERC) Starting Grant [number 757455], a UK Economic and Social Research Council (ESRC) Future Research Leaders grant [number ES/L009900/1], and through an internship funded by the UK Leeds Institute for Data Analytics (LIDA).

1 Introduction and Background

Individual-level modelling approaches, such as agent-based modelling (ABM), are ideally suited to modelling the behaviour and evolution of social systems. This is especially true in the context of modern ‘smart’ cities, where large volumes of data, supported by innovative ‘big’ data analytics, can be leveraged to better understand and capture the characteristics of the underlying systems. However, there is inevitably a high degree of uncertainty in projections of social systems – input data are noisy and sparse, and human behaviour is itself extremely uncertain – so one of the key challenges facing the discipline is the quantification of uncertainty within the outputs of these models.

This work focusses on the simulation of *urban flows*, i.e. the movement of people around urban areas over relatively short time scales (minutes and hours). It paper presents ongoing work towards methods that will (i) allow real-time data to be assimilated into models to reduce the uncertainty of their predictions and (ii) to quantify the *amount* of data (including overall volume as well as spatio-temporal granularity and regularity) that are required for successful assimilation. In other words, given some human system and a simulation of that system (we study the movement of pedestrians in this case), how much data are required from the real system in order to prevent the model uncertainties from causing the simulation to rapidly diverge from reality? With too little data it will be impossible to reliably constrain the model to reality, but how much is too little? Is one well-placed footfall counter sufficient to capture the dynamics of the system, or in reality would it be necessary to track a large proportion of the individual people? The hypothetical model used here is a simple system of pedestrians, each of whom move from an entrance towards one of two exits. This is analogous to a train arriving at a train station and passengers moving across the concourse to leave. The model environment is illustrated in Figure 1.

Very limited prior work has been conducted in this area. Some authors have attempted to conduct data assimilation, but use agent-based models that are simple in the extreme [4, 6]. Here, a more advanced model will be used in order to test how well the various methods handle complex features such as emergence and feedback. Others have developed more complex agent-based models [1] but those must remain mathematically coupled to an aggregate proxy model which limits the flexibility of the underlying agent-based model. The most similar work is that of [5], who attempt to assimilate data into a model of peoples’ movement in buildings. They do this by running an ensemble of models and re-starting each one using new input conditions that are created each time new data become available. The main difference between the approach in [5] and this paper is that here the aim is to eventually develop methods that are able to assimilate new data that automatically *alter the state* of the simulation while it is running. This is more analogous with data assimilation techniques in other fields [3].

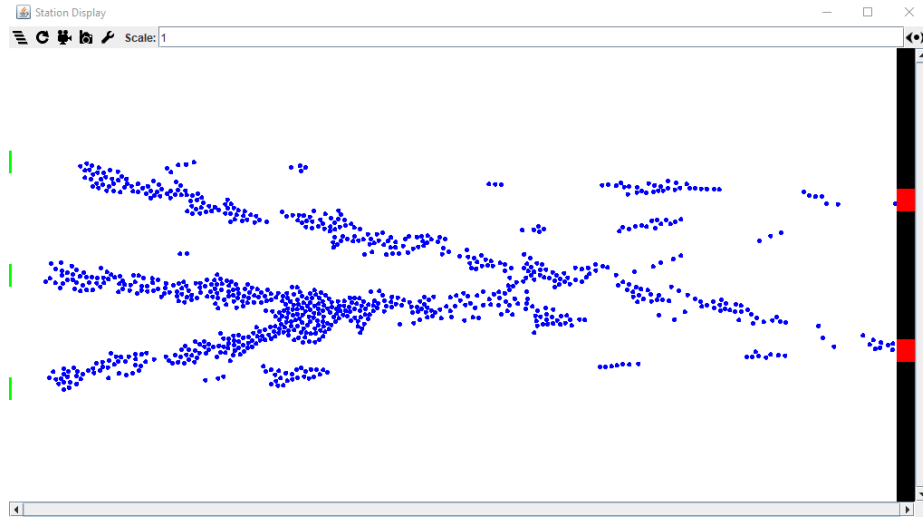


Fig. 1: A snapshot of the simple, hypothetical model that is used here. Agents arrive from the green entrances on the left and move towards the red exits on the right.

2 Methods

2.1 Overall Approach

This paper presents a work in progress, whose overall proposed approach is to:

1. Develop a agent-based simulation of pedestrian movements from an entrance to an exit using a simple social forces framework (similar to [2]) to control the behaviour of the agents (see Figure 1);
2. Run the simulation to generate hypothetical ‘truth’ data – we pretend that these data represent the real underlying system, and can then sample from them in a manner analogous to that of using sensors of peoples’ movements to sample from the real world;
3. Re-run the simulation with a different random seed and use samples of varying resolution from the ‘truth’ data to reduce the uncertainty of the new simulation;
4. Quantify the volume (total number of samples), granularity (amount of aggregation) and regularity (the number of samples per time period) that are required to successfully model the hypothetical ‘real’ data.

This paper presents the preliminary results up to stage 3, with the ultimate aim of quantifying the amount of data required to successfully constrain the simulation being immediate future work.

2.2 Preliminary Results: Constraining Model Uncertainty

Initial experiments consisted of using a prototype probabilistic programming library to attempt simple parameter estimation using Bayesian inference. This is a precursor to performing data assimilation (i.e. step 4). Parameter estimation was accomplished by configuring the model such that all parameters were fixed, with the only input being a list of random numbers that the model uses each time a probabilistic decision needs to be made. If the model uses all of these random numbers and requires another, it recycles from the beginning of the list. (Note that if only one random number to be passed to the model then every run would be identical as all probabilistic decisions would become stochastic). The output of the model is a list that specifies the number of agents in the simulation at each time step. This varies with each simulation because crowding can occur at random depending on the speed that different agents travel in the simulation and the exits that they are moving towards. In the future, different types of output that describe the hypothetical 'real' world in more or less detail will be experimented with to ascertain which have enough information to constrain the simulation adequately.

XXXX NICK TO INSERT DIAGRAM ILLUSTRATING THIS METHOD

A list of 100 random numbers (i.e. the model input) were created and these were passed to the model to create the 'truth' data. Following this, the model was integrated into a Bayesian network and the posterior distribution (the simulated number of agents in the system at each time step) was sampled. Importantly, two sampling routines were executed: one that used the truth data to attempt to constrain the posterior distribution by accurately estimating the values of the original random numbers, and another that did not include this observation step.

Figure 2 compares the results of the sampling with and without the incorporation of the truth data. It is clear that when the Bayesian network makes use of the observations from the 'truth' model, the output of each model (i.e. sample of the posterior distribution) is much more similar truth data. In other words, the procedure is more accurately estimating the initial random numbers (the prior distribution). Although this result is not of value in isolation, it is very useful as a proof-of-concept. It shows that Bayesian inference on an agent-based model that has reasonably complex characteristics (namely the emergence of crowding due to agent interactions) is able to perform parameter estimation. More rigorous data assimilation is a relatively small step.

3 Link to ABMUS Workshop Themes

This paper addresses a core aspect of the ABMUS workshop theme; that of the *trust* that we can have in model outputs. The conference recognises challenges in “designing, developing and implementing **trusted** models that can be used by industry and governments to enhance decision-making”. By adapting existing methods that are aimed at reducing uncertainty in models through the incorporation of up-to-date data, this work advances the methodology towards more a

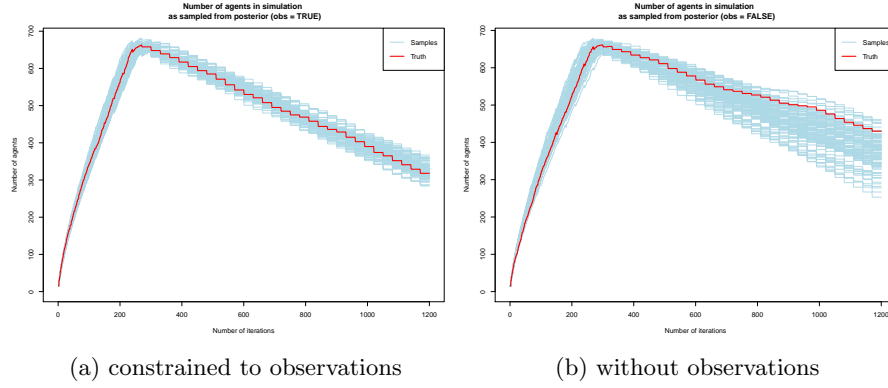


Fig. 2: Results of sampling the posterior with observations (2a) and without (2b). When the ‘truth’ data are used to constrain the posterior distribution, the sampling routine is much better able to estimate the input model state, so the outcomes of the samples are much closer to the ‘truth’ data.

rigorous representation of real urban systems that will be more acceptable for policy use.

Appendix: Model Code and Parameters

I think I'll drop this section

For reproducibility, the model code, in full, is available at XXXX. The parameters used to generate the results here are provided in Table 1.

Table 1: Model parameters used to generate the results outlined here

Parameter	Description	Value
numSamples	XX	500
numTimeSteps	XX	1200
numRandomDoubles	XX	100
totalNumPeople	XX	700
dropSamples	XX	200
downSample	XX	3
sigmaNoise	XX	0.1

References

1. Flötteröd, G., Bierlaire, M., Nagel, K.: Bayesian Demand Calibration for Dynamic Traffic Simulations. *Transportation Science* **45**(4), 541–561 (Nov 2011). <https://doi.org/10.1287/trsc.1100.0367>
2. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. *Physical Review E* **51**(5), 4282–4286 (1995). <https://doi.org/10.1103/PhysRevE.51.4282>
3. Kalnay, E.: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press (2003)
4. Malleson, N., Tapper, A., Ward, J., Evans, A.: Forecasting Short-Term Urban Dynamics: Data Assimilation for Agent-Based Modelling. In: *Proceedings of the European Social Simulation Conference, 25-29 September 2017, Dublin*. p. 13 (2017)
5. Wang, M., Hu, X.: Data assimilation in agent based simulation of smart environments using particle filters. *Simulation Modelling Practice and Theory* **56**, 36–54 (2015). <https://doi.org/10.1016/j.simpat.2015.05.001>
6. Ward, J.A., Evans, A.J., Malleson, N.S.: Dynamic calibration of agent-based models using data assimilation. *Royal Society Open Science* **3**(4) (2016). <https://doi.org/10.1098/rsos.150703>