Downloaded from www.sacemaquarterly.com | The quarterly update on epidemiology from the South African Centre for Epidemiological Modelling and Analysis (SACEMA).

Published: September 2014

Challenges of Agent-based Modelling of HIV Transmission

Lucio Tolentino - former visiting researcher at SACEMA

Epidemiological models for describing how a disease spreads through a population have been extremely useful to reduce the number of individuals who get sick or even die from illness. John Snow's now famous spatial model of how Cholera was spreading through 19th century London saved thousands of lives. Developing meaningful and useful models is not easy however. A careful balance must be made between making a model *meaningful*, having sufficient resemblance to the real-life system it claims to represent, and making a model *useful*, being able to explain or forecast beyond the data that informs it. As Albert Einstein famously put it: “A [model] should be as simple as possible, but no simpler”.

In this paper, we first motivate the use of agent-based modelling by introducing what it is and discussing our reasons for using it. Secondly, we present common challenges associated with agent-based modelling of HIV and our approaches to dealing with them. In particular, we discuss the need for simulating large populations and our parallel and distributed computing approaches to address this need. We argue the difficulty of validating agent-based models and present the approximate Bayesian computation method we used to produce sexual networks with summary statistics that are similar to those from real-world sexual networks. We talk about challenges of simulating dynamic processes that last several decades. Our goal is to give an overview of some of the key challenges associated with agent-based modelling of HIV and present our solutions to these challenges.

What is agent-based modelling and why do we use it?

The first key consideration is choosing the type of model to use. Generally speaking, models try to explain and give understanding to processes or phenomena seen in the world. Agent-based models attempt to understand these processes by simulating individuals and the individuals’ behaviours from which the process emerges. This is in comparison with compartmental models that aggregate individuals into groups (or compartments) and use more coarse-grain view of a system to describe a process.

For example, an agent-based model might simulate the behaviour of 100 individual wolves and 10,000 individual sheep, each with unique location in the simulated world, to explain how a predatory-prey system works. A compartmental model on the other hand might aggregate the wolves and sheep into two compartments and use the total number of animals in each to explain the same system. Choosing a model type then depends on the level of detail desired: If the starting location of the animals is thought to be important (e.g. if animals are so far apart that wolves have difficulty finding sheep) then an agent-based model is a good option. However, if location is not thought to be important (e.g. all animals are randomly intermixing) then the extra granularity gained by simulating the actions of individuals is likely unnecessary and a compartmental model might be a better choice.

In our recent work, we chose to use agent-based models to simulate HIV transmission because we are interested in modelling fine-grain processes that may otherwise be lost in a compartmental model. For example, we are interested in simulating HIV transmission in a highly heterogeneous population – i.e. a population where all the individuals have characteristics and behaviours that are unique. To do this we simulate individuals in a given population with individual agents, and assign them characteristics, like gender, age and an intrinsic sexual activity drive. These agents move around in a simulated world and form and dissolve sexual relationships with other agents based on the assigned characteristics. In this way the agents produce a dynamic (i.e. changing over time) sexual network through which HIV is able to spread. In this way agent-based models are intuitively similar to how the real world operates: HIV diffusing through a population is the result of discrete events (like forming a relationship or becoming infected with HIV) happening to distinct individuals. These discrete events contain randomness, but are informed by individual characteristics – their individual sexual drive or a preference for older partners. This means that, like in real life, different agents experience different events at different times.

Simulating large populations

While an agent-based model is somewhat intuitive, a modeller faces many questions while developing the model. One is answering the question of how many agents are needed to adequately simulate the underlying processes of HIV propagation? A tempting solution is to simply use the largest population size possible. However, as the number of agents in the simulation increases so does the amount of time required to run the model – and a model that takes months or years to run is not very useful. Large population sizes are necessary though to avoid "small world" phenomena: processes that emerge purely from having unrealistically few agents being modelled. For example, consider a purely heterosexual agent-based model of HIV transmission. If we use a population with 4 agents whose sex is randomly assigned then our model will fail to see any transmission in approximately 12.5% of simulations. This is because in approximately an eighth of those simulations all the agents will be the same sex. It's for this reason that larger population sizes are necessary to create robust and reliable results from simulations.

In an attempt to simulate very large populations (hundreds of thousands of agents), we've developed parallel algorithms that distribute the model's workload among multiple processors on a single computer and among multiple computers on a cluster of machines. Running the agent-based model in a high performance setting enables us to significantly speed up the simulation of large population sizes. With these new algorithms, simulations with large population sizes that used to take months now only take hours.

Validating agent-based models

Unfortunately, a model with a large population size is, by itself, insufficient to be a useful model. Once a model is "complete" (i.e. decisions have been made as to how many agents will be in the simulation, the events that can happen to the agents, the laws that govern these events, and the time horizon over which we want to simulate) we need to show that it is valid. This is done through a process that is aptly named *validation*. This can be hard because validation, in part, means showing that any change in the model world, and the consequences of those changes, would play out in the real world system that the model is supposed to represent. It is also the other way around where real-world changes should be seen in the model world. The conundrum is that the real world system is often too complex to test changes and their consequences – if it weren't too difficult we likely wouldn't spend time trying to model it!

An additional challenging aspect of validation is that the real world and the data derived from the real world are the result of many components and their subcomponents, and the interactions of all these components. The result is a complex system with many dimensions and begs several questions: How many of these components and interactions must be represented in the model? How "true" are the data collected for all of these dimensions? How does one test and confirm that the model is in line with the data across all these dimensions?

There are nonetheless a plethora of methods for validating models – and a large number of academic articles and books have been written describing how to do it. However, techniques like Cross Validation (the model is calibrated with a subset of the available data and the model is then tested on its ability to reproduce the remaining data) and Predictive Validity (the model makes a prediction about the future and is tested on whether the prediction comes to fruition) are often not applicable to complex long-term models like those studying HIV epidemiology. This does not mean that models of HIV cannot be validated – it means that the stamp of validation will likely be more subjective and not involve a formal p-value from a goodness-of-fit test. Modellers in effect must decide which dimensions are most likely driving the processes and determine the best way to show that their model captures those dimensions. For example, a model interested in the effect of age-mixing on HIV incidence will need to show that it is able to reasonably reproduce metrics like age-specific sexual activity and HIV prevalence. However, it would not be unreasonable to omit processes related to random biological variation in HIV infectiousness that is not associated with age or gender. This means that it's important to clearly link research question, model design, and validity checks to achieve high quality, meaningful models.

In our dynamic sexual network models we claim validity by showing that they can produce a sexual network that is approximately similar to the real-world sexual network: We compare prevalence of age-disparate relationships across different age groups and sexes; we compare the frequency with which individuals form multiple concurrent relationships; we compare the duration of relationships and the time between relationships. In short, we compare our simulated sexual network to a real world sexual network with statistics that are known to be important in the epidemiology of HIV. Hence our simulation is able to produce a facsimile of a real world sexual network.

The Demographic Challenge

Questions of population size and validity challenge the agent-based modelling field in general. However there are challenges that are specific to modelling of HIV transmission. The first is a problem of time scales: the biological nature of the disease means that epidemiological models for the spread of HIV need to describe a very long time window. Most HIV models consider at least a 30-year time period (the approximate beginning of the epidemic till now) and many consider more (projecting further into the future). This long time horizon is challenging: simulations of HIV can take relatively long to run (e.g. influenza models typically model daily time steps for six months, our HIV models use weekly time steps for 30 years – 180 time steps versus 1560). Simulating such a long time window also mandates that we simulate demographic processes in the model population. How individuals age, reproduce and die may seem like simple processes, but can be very complex. In fact, demographic processes often warrant their own models. The South African Actuarial Society (ASSA) has produced several such HIV models that consider demographic effects in great details. Their models are compartmental and hence very good at describing overall population effects, but aren't flexible to the individual level processes which we mentioned earlier.

We have employed a few different strategies to include demographics in our models. The simplest solution merely has a young agent replace an agent that has become "too old" to form sexual relationships – when an agent turns 65-years-old he or she is replaced by a 15-year-old agent. This strategy forces the number of agents in the simulation to remain constant over time, but can be useful when we are primarily interested in understanding processes like basic age-mixing. A more complex solution allows sexual relationships to produce offspring that then become part of the simulation. This allows the population size to grow over time, but enforces a closed world that is not very realistic. The most comprehensive strategy we have used, which comes at the cost of run time and model complexity, uses age-specific fertility and mortality rates to inject and remove agents into and out of the simulation system. This is the most satisfying solution in terms of creating realistic demographics, but the increased complexity can be cumbersome.

Conclusions

All these challenges are computational in nature. We can develop more efficient algorithms for simulating larger and more dynamic population. We can build more sophisticated models that more closely match sexual network and demographic data. However, these point to a larger challenge: How do we simulate a process that is governed by highly volatile rules that are constantly changing? We can collect more data and build more models, but the reality is that effectively simulating sexual networks means effectively simulating human behaviour – and effectively simulating human behaviour is a hard problem. This does not mean that modelling should not be done – modelling efforts have already saved lives. It means that all assumptions made when developing a model should be carefully documented, and the implications of these should be thoroughly investigated.

If we employ useful tools like sensitivity analysis and approximate Bayesian inference to explore the range of answers that models produce, given the data and additional assumptions; if we explicitly acknowledge the gaps in our knowledge and our suspicions of biased data; if we clearly state the intentions and limitations of our models; then the use of models will no longer be a straw man treasure hunt for the fountain of truth or unscientific attempt at predicting the future. Models can be what they are: A systematic exploration of plausible trends and phenomena in a stylized model world; a representation of a system that helps us to understand the findings of previous empirical studies; an aid in narrowing our focus for follow-up empirical experiments.

Lucio Tolentino - *former visiting researcher at SACEMA.sean-tolentino@uiowa.edu*