Calibrating with Multiple Criteria: A Demonstration of Dominance

Jennifer Badham[1], **Chipp Jansen**[2], **Nigel Shardlow**[3], **Thomas French**[3]

[1]*Centre for Public Health, Queen's University, Belfast BT12 6BA, United Kingdom*
[2]*Department of Informatics, King's College, London, United Kingdom*
[3]*Sandtable Ltd, 69 Old Street, London, EC1V 9HX, United Kingdom*
Correspondence should be addressed to *research@criticalconnections.com.au*

Journal of Artificial Societies and Social Simulation 20(2) 11, 2017
Doi: 10.18564/jasss.3212 Url: http://jasss.soc.surrey.ac.uk/20/2/11.html

Received: 02-05-2016 Accepted: 12-09-2016 Published: 31-03-2017

Abstract: Pattern oriented modelling (POM) is an approach to calibration or validation that assesses a model using multiple weak patterns. We extend the concept of POM, using dominance to objectively identify the best parameter candidates. The TELL ME agent-based model is used to demonstrate the approach. This model simulates personal decisions to adopt protective behaviour during an influenza epidemic. The model fit is assessed by the size and timing of maximum behaviour adoption, as well as the more usual criterion of minimising mean squared error between actual and estimated behaviour. The rigorous approach to calibration supported explicit trading off between these criteria, and ultimately demonstrated that there were significant flaws in the model structure.

Keywords: Multi-Criteria Decision Making, Calibration, Pattern-Oriented Modelling, Dominance, Behaviour Modelling

Introduction

1.1 Agent-Based Models (ABMs) simulate "unique and autonomous entities that usually interact with each other and their environment locally" (Railsback & Grimm 2012, p. 10). Such models are therefore designed at the micro-scale, with rules to guide the actions of the simulated individuals based on their specific characteristics and situation. In contrast, much of the interesting behaviour of the model occurs at the macro-level.

1.2 This scale mismatch complicates model calibration. Parameters for those micro-scale rules may be unmeasurable, but the aggregated effect of the decisions is routinely collected in data about the operation of the system being modelled. With a large number of parameters, it may be relatively easy to obtain an apparently good fit overall that is nevertheless hiding structural invalidity or other problems. One way to make the calibration more robust is by assessing model output against multiple criteria selected for their diversity, referred to as pattern-oriented modelling (Wiegand et al. 2004; Railsback & Grimm 2012). Doing so, however, introduces the problem of defining an overall 'best fit' since different sets of parameter values may generate model output that meet different criteria.

1.3 One approach is to establish an overall objective function that combines each of the criteria in some way. For example, the criteria could be weighted and the model calibrated to best fit the weighted combination. However, this approach introduces an arbitrary function to combine the criteria (such as additional parameters in the form of criteria weights), typically with only limited knowledge of what is being traded away. Another method uses stakeholder or other experts to assess the reasonableness of the model's behaviour (Moss 2007).

1.4 Categorical calibration or filtering (Wiegand et al. 2004; Railsback & Grimm 2012) uses acceptance thresholds for each criterion and retains all parameter sets that meet all the thresholds for further consideration. However, this is inefficient. If any threshold is set too high, a parameter set could be rejected that is an excellent fit on all other criteria. On the other hand, setting a lower threshold passes too many potential solutions to be easily compared.

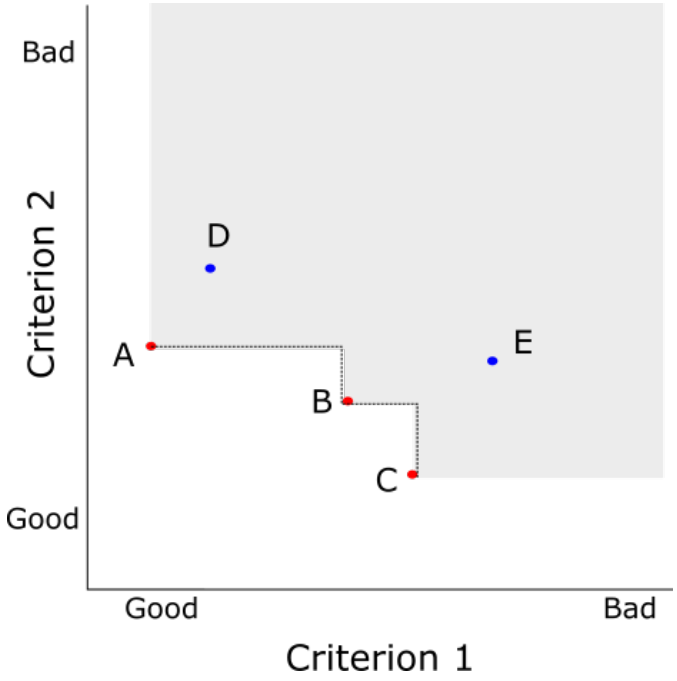

Figure 1: Definition of dominance (two dimensions). Point D is dominated by point A because point A is better against all criteria than point D. That is, regardless of the relative importance of the two criteria, point A is always preferred over point D. Similarly, point E is dominated by both point B and point C. But point E is not dominated by point A; if criterion 2 was much more important than criterion 1, it may be appropriate to select E for the small improvement in criterion 2 at the expense of the loss in criterion 1. The shaded area indicates the parameter space that is dominated by any of the three points A, B or C. The Pareto efficient front is the set of points that are not dominated by any other, in this case any points on the dashed line specified by the points A, B and C. Along this line, improvements in one dimension can only be achieved at the expense of at least one other criterion; for example, moving from A to B improves criterion 2 but worsens criterion 1. With more dimensions, the Pareto front is given by a piecewise hyperplane, but is also the set of points that appear on the front of any pair of dimensions, regardless of whether they are dominated in other pairs of dimensions.

1.5 This paper instead presents the dominance approach, which does not arbitrarily prioritise criteria or set subjective thresholds. Instead, dominance is used to identify all the parameter sets that are on the Pareto efficient frontier. These are the parameter sets that are objectively best, where an improvement in one criterion can only be made by reducing the fit for another criterion (see Figure 1). While this approach is well established in operations research for multi-criteria decision making or optimisation (Müssel et al. 2012), it is less well known in social simulation (with some exceptions, such as Schmitt et al. 2015).

1.6 The method is described using a case study: calibrating the TELL ME model concerning protective behaviour in response to an influenza epidemic. This paper first presents the model structure and the parameters required to operationalise the links between attitude, behaviour and epidemic spread. The description focuses on the necessary background to understand the calibration process presented in the following sections. The approach to setting parameter values is then described, with the results of that process and conclusions following.

Case Study Description: TELL ME Model

2.1 The European funded TELL ME project [1] concerned communication before, during and after an influenza pandemic. Ending in January 2015, it was intended to assist health agencies to develop communication plans that encourage people to adopt appropriate behaviour to reduce influenza transmission. One project output was a prototype ABM, to explore the potential of such models to assist communication planning. The agents in that model represent people making decisions about protective behaviour (such as vaccination or hand hygiene) in light of personal attitudes, norms and epidemic risk.

2.2 The core of the TELL ME model is individual agents making decisions about whether to adopt behaviour to reduce their chance of becoming infected with influenza. Protective behaviour is adopted (or dropped) by an

agent if the weighted average of attitude, subjective norms and perception of threat exceeds (or falls below) some threshold.

2.3 Each agent is attached to a patch (a location defined by a grid) overlaid on a map of the country in which the epidemic is being simulated. The epidemic is mathematically modelled by the patches; there is no transmission between individual agents. The infectivity at any patch is adjusted for the proportion of local agents who have adopted protective behaviour and the efficacy of that behaviour. In addition, the number of new infections in nearby patches is a key input to each agent's perception of threat. Thus, the agent protective behaviour decisions and the transmission of the epidemic are mutually dependent.

2.4 The operationalisation of this model design is described briefly below. This description focuses on those elements of the model that were calibrated using dominance. The behaviour of the agents is also affected by communication plans, which are input to the model as sets of messages. The communication elements were disabled for calibration purposes due to lack of data, and are therefore not described here. The model was implemented in NetLogo (Wilensky 1999), with the full code, complete model design (Badham & Gilbert 2015) and other documentation available online.[2]

Epidemic transmission

2.5 The epidemic is modelled by updating counts for each disease state of the population at each patch. For influenza, a suitable epidemic model is the SEIR model, in which people conceptually start in the susceptible (S) state, become exposed (E) but not yet infectious, then become infectious (I) and are eventually removed from calculations (R) because they either recover and become immune or they die. The model represents this process mathematically (Diekmann & Heesterbeek 2000), governed by transition rate parameters (β for $S \rightarrow E$, λ for $E \rightarrow I$, and γ for $I \rightarrow R$).

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dE}{dt} &= \beta SI - \lambda E \\ \frac{dI}{dt} &= \lambda E - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1}$$

2.6 In each patch or region (r), the value of the transition rate parameter from S to E (β) is reduced in accordance with the behaviour decisions taken by individuals at that patch and the efficacy(E) of the behaviour. The reduced infectivity rate (calculated with Equation 2) is used in the transmission equations (Equation 1), leading to a lower local incidence. To support a mix of behaviour (and hence different reductions in infectivity between patches), each patch is home to at least ten agents, with greater numbers in those patches that correspond to high population density real world locations.

$$\beta_r = \beta \left(1 - P_r E\right) \tag{2}$$

2.7 To allow the epidemic to spread, a proportion of estimated new exposures for a region are actually created in neighbouring patches to simulate travel. This requires two additional parameters, the proportion of new infections created at other locations, and the split between neighbouring or longer distance patches.

Operationalising decisions about protective behavior

2.8 The agents' behaviour decisions are based on three psychological models: the Theory of Planned Behavior (Ajzen 1991), Health Belief Model (Rosenstock 1974), and Protection Motivation Theory (Maddux & Rogers 1983). The key factors of attitude, norms and threat from these models were used as the inputs for agent behaviour. The agent compares the weighted average of the three inputs to a threshold (Equation 3) for each type of behaviour (vaccination or other protective). If the value is higher, the agent adopts the non-vaccination behaviour or seeks vaccination, and non-vaccination behaviour ceases once the value falls below the relevant threshold. Vaccination cannot be dropped. Threat has the same value for both types of behaviour, but attitude, norms, weights and thresholds may be different.

$$\begin{aligned} \omega_A A + \omega_N N + (1 - \omega_A - \omega_N) T_t \geq B \quad &\text{adopt behavior} \\ \omega_A A + \omega_N N + (1 - \omega_A - \omega_N) T_t < B \quad &\text{abandon behavior} \end{aligned} \tag{3}$$

2.9 Attitude is operationalised as a value in the range [0,1], initially selected from a distribution that reflects the broad attitude range of the population. Subjective norms describe how a person believes family, friends and other personally significant people expect them to behave and the extent to which they feel compelled to conform. The norm is operationalised as the proportion of nearby agents who have adopted the behaviour.

2.10 Perceived threat (T_t) reflects both susceptibility and severity (Equation 4). Following the method of Durham & Casman (2012), susceptibility is modelled with a discounted (δ) cumulative incidence time series. This means that perceived susceptibility will increase as the epidemic spreads but recent new cases (c_t) will impact more strongly than older cases. In contrast to the cited paper, only nearby cases are included in the time series for the TELL ME model, so perceived susceptibility will be higher for the simulated individuals that are close to the new cases than for those further away. Severity is included as a simple 'worry' multiplier (W), and can be interpreted as subjective severity relative to some reference epidemic.

2.11

$$\begin{aligned} s_t &= \delta s_{t-1} + c_{t-1} \\ T_t &= W s_t \end{aligned} \tag{4}$$

Calibration Process

3.1 From the model structure discussion, it is clear that there are many parameters to be determined. Some may be estimated directly from measurable values in the real world, such as population counts. Ideally, unmeasurable values should be calibrated to optimise some measure of goodness of fit between model results and real world data.

3.2 The first phase simplified the model to reduce the number of parameters influencing results. This was done by excluding the communication component and fixing protective behaviour to have no effect. Other values were fixed at values drawn from literature, specifically those that affected the distribution of attitudes and the transmission of the epidemic. The exclusion of some components and setting of other parameters to fixed estimates can be interpreted as reduction in the dimensions of the parameter space, reducing the scope of the calibration task.

3.3 The second phase calibrated the parameters that are central to the model results; those that govern the agents' decisions to adopt or drop protective behaviour as an epidemic progresses (weights in Equation 3 and discount in Equation 4). This phase is where dominance was used, to assess parameter sets against three criteria: size and timing of maximum behaviour adoption, as well as the more usual criterion of minimising mean squared error between actual and estimated behaviour.

3.4 The model parameters are summarised in Table 1, together with how they were used in the calibration process. While the TELL ME model included both vaccination and non-vaccination behaviour, only the latter is reported here because the process was identical. Non-vaccination behaviour was calibrated with various datasets collected during the 2009 H1N1 epidemic in Hong Kong. The calibration process is described in more detail in the remainder of this section.

Dimension reduction: protective behaviour

3.5 Attitude distribution was based on a study of behaviour during the 2009 H1N1 epidemic in Hong Kong (Cowling et al. 2010), which included four questions about hand hygiene: covering mouth when coughing or sneezing, washing hands, using liquid soap, and avoiding directly touching common objects such as door knobs. A triangular distribution over the interval [0,1] with mode of 0.75 was used to allocate attitude scores in the model as an approximation to these data.

3.6 The efficacy of protective behaviour (E) was set to zero (ineffective) during calibration. That is, agents respond to the changing epidemic situation in their decision processes, but do not influence that epidemic. This ensures simulations using the same random seeds will generate an identical epidemic regardless of behaviour adoption, allowing simulated behaviour to respond to the relevant incidence levels.

Dimension reduction: epidemic transmission

3.7 Several parameters that influence epidemic spread were estimated from data. These are the various transition rates between epidemic states, the structure of the population in which the epidemic is occurring, and the

Symbol	Description	Value
A_0	Base attitudes	Fixed distribution
P_r	Population by region	GIS data
R_0	Basic reproduction ratio (gives β)	Fixed at 1.5
$1/\lambda$	Latency period $E \rightarrow I$	Fixed at 2
$1/\gamma$	Recovery period $I \rightarrow R$	Fixed at 6
	Population traveling	Fixed at 0.30
	Long distance traveling	Fixed at 0.85
W	Severity relative to H1N1	Fixed at 1
δ	Incidence discount	To be calibrated
ω_A	Attitude weight (x2)	To be calibrated
ω_N	Norms weight (x2)	To be calibrated
B	behavior threshold (x2)	To be calibrated
E	behavior efficacy (x2)	Fixed at 0
	Meaning of 'nearby'	Fixed at 3 patches

Note: Parameters associated with the effect of communication are not listed as communication was removed from the calibration process.

Table 1: TELL ME model parameter settings for calibration.

mobility of that population. The multiplier in Equation 4 was set at $W = 1$, establishing H1N1 as the reference epidemic.

3.8 The basic reproductive ratio (denoted R_0) is related to the parameters in Equation 1 with $R_0 = \beta/\gamma$ (Diekmann & Heesterbeek 2000). R_0 for the 2009 H1N1 epidemic was estimated as 1.1-1.4 (European Centre for Disease Prevention and Control 2010). Calibration experiments were run with $R_0 = 1.5$ (the lowest value for which an epidemic could be reliably initiated), latency period of 2 days (European Centre for Disease Prevention and Control 2010), and infectious period of 6 days (Fielding et al. 2014).

3.9 The population at each patch was calculated from population densities taken from GIS datasets of projected population density for 2015 (obtained from Population Density Grid Future collection held by Center for International Earth Science Information Network - CIESIN - Columbia University & Centro Internacional de Agricultura Tropical - CIAT 2013). These densities were adjusted to match the raster resolution to the NetLogo patch size and then total population normalised to the forecast national population for 2015 (United Nations, Department of Economic and Social Affairs, Population Division 2013).

3.10 As epidemic processes (Equation 1) occur independently within each patch, the model explicitly allocates a proportion of the new infections created by a patch to other patches to represent spreading of the epidemic due to travel. The proportion of new infections allocated to other patches was set at 0.3, with 0.85 allocated to immediate neighbours and 0.15 allocated randomly to patches weighted by population counts. These values provide a qualitatively reasonable pattern of epidemic spread.

Dominance analysis of behaviour parameters

3.11 Four parameters are directly involved in agent adoption of protective behaviour: weights for attitude and norms, the discount applied for the cumulative incidence, and the threshold score for adoption (ω_A, ω_N, δ, and B in Equations 3 and 4). Briefly, multiple simulations were run while systematically varying these parameters to generate a behaviour adoption curve. That curve was assessed against empirical data on three criteria, and dominance analysis was used to identify the best fit candidates.

3.12 Broadly, the empirical behaviour data has an initial population proportion of approximately 65%, which rises to 70% and then falls below the starting level. This rise and fall was considered the key qualitative feature of the data and two aspects were included: timing and size of the bump. The three criteria to select the best fit parameter sets were:

- mean squared error between prediction and actual over all points in the data series (MSE);
- the difference in values between the maximum predicted adoption proportion and maximum actual adoption proportion (ΔMax); and

Parameter	Range
Attitude weight (ω_A)	0.2 by 0.05 to 0.7
Norms weight (ω_N)	0.1 by 0.05 to 0.5
Incidence discount (δ)	0.02 by 0.02 to 0.2
Behaviour threshold (B)	0.2 by 0.05 to 0.7

Table 2: Parameter values tested in the calibration process.

- the number of ticks (days) between the timing of the maximum predicted adoption and maximum actual adoption (ΔWhen).

3.13 Experiments and dominance analysis were performed with the Sandtable Model Foundry (Sandtable 2015). This proprietary system was used to manage several aspects of the simulation in a single pass: sampling the parameter space, submitting the simulations in a distributed computing environment via the NetLogo API, comparing the result to the specified criteria, and calculating the dominance fronts. As each run takes several minutes, the sampling and distributed computing environment made it feasible to comprehensively explore the parameter space in a reasonable time and the within-system dominance calculation simplified analysis.[3]

3.14 Simulations were run with parameter values selected from the ranges at Table 2, chosen so as to require a contribution by attitude ($\omega_A \geq 0.2$) to support heterogeneity of behaviour between agents on a single patch. Parameter combinations were excluded if they did not include contributions by all three influencing factors of attitude, norms ($\omega_N \geq 0.1$), and threat ($\omega_A + \omega_N \leq 0.9$). The parameter space was sampled using the Latin Hypercube method, with 813 combinations selected.

3.15 Ten simulations were run for each parameter combination. Preliminary testing with 30 repetitions indicated that simulations using the same parameters could generate epidemics that differ substantially on when they 'take off', but they had similar shapes once started, and hence similar behaviour adoption curves (not specifically shown, but visible in Figure 5). Ten of the seeds were retained for use with the calibration simulations. These random seeds generated epidemics with known peaks regardless of the behaviour parameter combination as the generated epidemic was not affected by protective behaviour (since efficacy is set to 0).

3.16 The behaviour curves from the 10 simulations were centred on the timestep of the epidemic peak and averaged. The average curve was compared to the (centred) 13 data points of the Hong Kong hand washing dataset (Cowling et al. 2010, supplementary information) for calculation of the three fit criteria.

3.17 Parallel plot analysis was used as an exploratory tool. This is an interactive technique using parallel coordinates (Inselberg 1997; Chang 2015) to simultaneously show the full set of model parameters and the criteria metrics. That is, simulation runs can be filtered with specific values or ranges of one or more of the input parameters or difference from criteria.

3.18 Dominance analysis was used to identify the best fit candidate parameter sets. This technique assigns each parameter set to a dominance front (using the algorithm of Deb et al. 2002). Front 0 is the Pareto efficient frontier, where any improvement in the fit for one criterion would decrease the fit against at least one of the other criteria (Figure 1). Front 1 would be the Pareto efficient frontier if all the front 0 parameter sets were removed from the comparison, and so on for higher front values until all parameter sets are allocated a front number.

Results

4.1 The parameter sets that are not dominated are those on the Pareto efficient frontier (front 0). These are described at Table 3 with their performance against the three criteria. By definition, for all other parameter sets, there is at least one on the frontier that is a better fit on at least one criterion and at least as good a fit on all others. Thus, these are the objectively best candidates.

4.2 The choice between these for the best fit overall is subjective, trading performance in one criterion against performance in the others and also adding other factors not captured in the criteria. Two methods were used to assist with that choice, quantitative distance from best fit criteria and qualitative fit of behaviour curves.

4.3 The fit for all tested parameter sets is displayed at Figure 2, with the non-dominated (front 0) candidates marked in red and labelled with the set number from Table 3. Each appears in the lower left corner of at least one of the sub-figures. From (a) and (b), a small error in the timing of the maximum adoption cannot be combined with a small error in either of the other properties. Focussing only on those other properties (sub-figure (d), the

Set	Parameter values ω_A	ω_N	δ	B	Criteria MSE	Δ Max	ΔWhen
1	0.70	0.20	0.18	0.50	0.00	0.09	78
2	0.70	0.20	0.10	0.50	0.00	0.08	78
3	0.65	0.10	0.20	0.40	0.00	0.05	71
4	0.60	0.20	0.06	0.45	0.00	0.05	84
5	0.55	0.10	0.10	0.35	0.00	0.01	76
6	0.35	0.10	0.18	0.25	0.01	0.01	73
7	0.65	0.10	0.00	0.50	0.07	0.01	206
8	0.55	0.10	0.18	0.25	0.08	0.19	69
9	0.70	0.20	0.14	0.30	0.12	0.26	62
10	0.55	0.35	0.16	0.40	0.13	0.27	61
11	0.20	0.30	0.04	0.20	0.19	0.29	36
12	0.30	0.35	0.00	0.30	0.19	0.29	19
13	0.25	0.30	0.10	0.25	0.23	0.29	9
14	0.25	0.50	0.14	0.25	0.29	0.29	33
15	0.25	0.25	0.02	0.35	0.34	0.01	104

Table 3: Best fit parameter sets and their assessment.

relevant section of (c) expanded), parameter sets 5 and 6 achieve a much closer maximum adoption compared to sets 3 and 4, with only a small loss in the mean squared error. While the same analysis could have been performed by examining Table 3 directly, the visualisation allows fast comparison, even with a larger number of criteria.

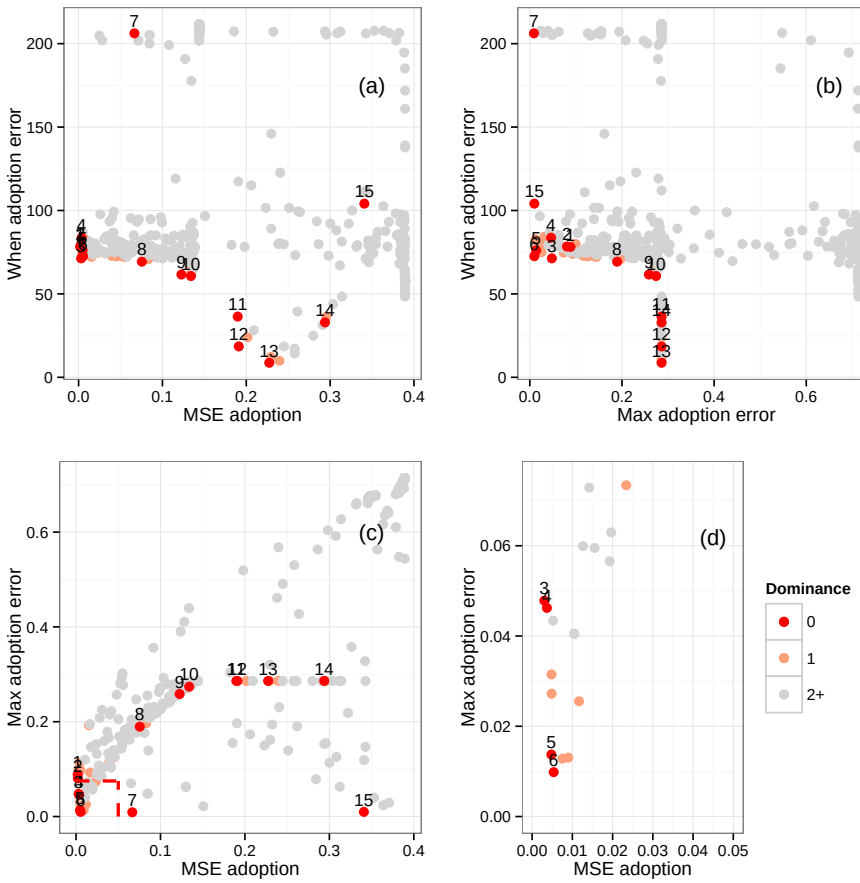

Figure 2: Average outcome over 10 simulations for each of 813 parameter sets. Subfigures (a), (b) and (c) display the outcome against different pairs of criteria, with subfigure (d) focussing on the best fit section of (c). Those on the Pareto efficient frontier are coloured red and numbered according to Table 3.

4.4 These best fit candidates are also coloured red in the parallel coordinate analysis (see Figure 3). This revealed that good fit parameter sets existed throughout the tested parameter space for the weights and discount, but that the threshold should not exceed 0.5. The main benefit of this analysis, however, is interactive. For example, it can provide a visual method of pattern-oriented modelling filtering, by adjusting ranges on the criteria results and displaying the parameter values of the simulations that survive.

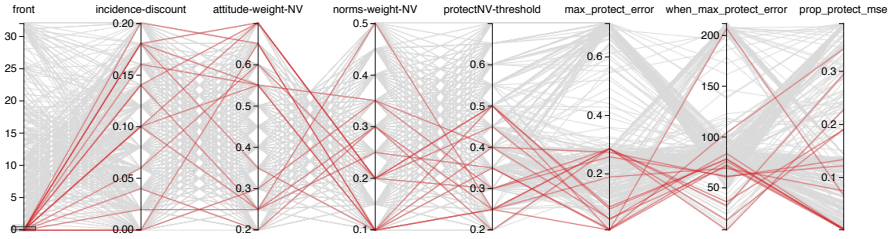

Figure 3: Interactive analysis of simulation experiments. The input parameter values appear in the left section of screen, and the fit against each criteria on the right. Simulation runs can be highlighted in groups (such as all those on the Pareto efficient frontier as displayed) or individually to explore the effect of different combinations of parameter values.

4.5 For the qualitative visualisation, fifty simulations were run using the NetLogo BehaviorSpace tool (Wilensky 1999) for each of the non-dominated parameter sets. The average adoption curve is shown in Figure 4. Only sets 1 to 6 display the appropriate pattern of behaviour, with approximately two thirds of the population adopting the behaviour before the start of the epidemic followed by an increase and then return to a similar level once the epidemic has passed. An inspection of Table 3 shows that the mean squared error is similar for all six, but parameter sets 5 and 6 also have a good match in the estimated maximum adoption level, supporting the selection of either of these as the best fit.

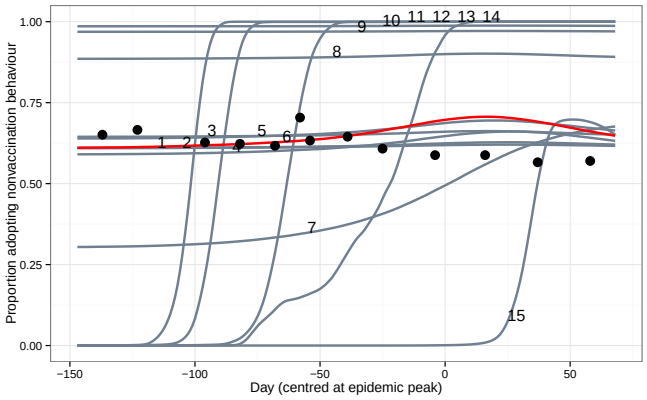

Figure 4: The average of 50 simulation runs for each of the non-dominated candidate parameter sets. The selected best fit parameter set (set 6) is drawn in red. Empirical behavior values (extracted from Cowling et al. 2010, supplementary information) are shown with dots.

4.6 Ultimately, parameter set 6 was selected as the best fit and used as the TELL ME non-vaccination behaviour default values. The individual runs for the model with these default parameter values are shown in Figure 5, together with the average behaviour curve.

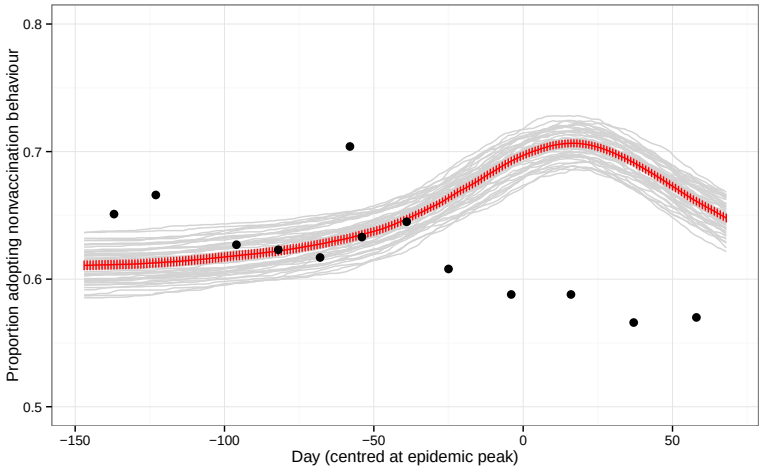

Figure 5: 50 simulation runs for proportion adopting hand washing behaviour during the 2009 H1N1 epidemic in Hong Kong, together with the average behaviour curve (and 95% confidence interval). Parameter values are: 0.35 for attitude weight, 0.1 for norms weight, 0.55 for threat weight, 0.18 discount in cumulative incidence, and 0.25 for threshold. Empirical behaviour values (extracted from Cowling et al. 2010, supplementary information) are shown with dots.

Discussion

5.1 This paper describes a detailed calibration process using the prototype TELL ME model as a case study. The model is complicated, with many components and parameters to reflect policy makers' understanding of their planning environment. It is also complex, with model behaviour shaped by two types of interactions. Personal decisions about protective behaviour affect epidemic progress, which influences perceptions of threat and hence personal decisions. Behaviour decisions of agents are also directly influenced by the decisions of nearby agents, through their perception of norms.

5.2 The calibration process first reduced the dimensions of the parameter space by setting epidemic parameters, population density and attitude distribution to values drawn from the literature. Some other parameters were set to values that removed their influence in the model (notably behaviour efficacy and those associated with communication).

5.3 This reduced the parameters required to calibrate the model to only four: attitude weight, norms weight, incidence discount and adoption threshold. These parameters control the central process of the model - adoption of protective behaviour in response to an epidemic. With only limited empirical information about behaviour throughout an epidemic, we used pattern-oriented modelling and attempted to calibrate against three weak signals: timing of the behaviour peak (compared to the epidemic peak), maximum level of protective behaviour, and minimising the mean square difference between the simulation estimate and measured behaviour level.

5.4 Having three assessment criteria opens the question as to how to compare the runs where they have different rankings across criteria. The standard approach is to set acceptance thresholds for each criterion (Railsback & Grimm 2012) and then select from only those that pass all. However, this is inefficient: if thresholds are set low enough to pass simulations that are generally excellent but are slightly less fit on one criterion, then the thresholds also allow through any simulation that is slightly less fit on all criteria. Instead, we have used the concept of dominance to identify the objectively best parameter sets; for any excluded simulation, there is at least one member of the dominant candidates that is better on at least one criterion and no worse on all others. Additional criteria were used to choose between these objectively good candidates, determining what to give up in order to achieve the best overall fit.

5.5 There is little similarity in the non-dominated parameter sets. Very different parameters can achieve similar outcomes (for example, sets 5 and 6), and parameter values in the best fit sets covered a broad range of values. This reflects the interdependence between the parameters and emphasises the difficulties in calibrating the TELL ME model, it would not have been possible to identify these candidates by tuning parameters individually.

5.6 The rigorous calibration process was instrumental in detecting structural problems with the model. In particular, the prototype was unable to generate results with a behaviour peak earlier than the epidemic peak, in

conflict with the empirical results for hand hygiene during the Hong Kong 2009 H1N1 epidemic. A reasonable fit could have been achieved against a minimum mean squared error single criterion, but assessing against multiple criteria highlighted the timing weakness.

5.7 Further consideration of the model rules makes it clear that this is a structural or theoretical gap rather than a failure in calibration. As attitude, weights and the threshold are fixed, change in behaviour arises from changes in the norms or perceived threat. The attitude weight is instrumental in setting the proportion adopted in the absence of an epidemic, but plays no part in behaviour change as the attitudes of agents are constant. As the epidemic nears an agent, incidence increases near the agent, which also increases perceived threat and may trigger adoption. This may also trigger a cascade through the norms (proportion of visible agents who are protecting themselves) component. However, the threat component of the behaviour decision (Equation 3) can only respond to an epidemic, not anticipate it, and the norms component can only accelerate adoption or delay abandoning it. Therefore, regardless of parameter values, the simulation is unable to generate a pattern with a behaviour curve peak before the epidemic peak.

Conclusion

6.1 Ultimately, the TELL ME ABM was unable to be calibrated adequately for policy assessment. That is, the best fit parameter set was used as the model default values, but the simulation did not produce realistic model behaviour. For the purposes of the TELL ME project, this outcome was disappointing but not unexpected. The ABM was a prototype intended to identify the extent to which such a model could be developed for planning purposes. The attempt highlighted both the limited empirical information about behaviour during an epidemic and the absence of information about the effect of communication. Relevant behavioural information must be collected if a full planning model is to be developed in the future.

6.2 In contrast, the use of dominance was successful in identifying candidate parameter sets that are objectively best against several competing criteria. Selection between these candidates was then relatively simple as only a limited number needed to be considered. Further, the rigorous process highlighted structural problems in the model as the desired timing of the behaviour peak could not be achieved while also achieving good performance in other criteria.

Acknowledgements

This paper benefited enormously from thoughtful comments and questions from several referees, and the authors would like to thank them for their care and effort. We would also like to thank Nigel Gilbert, Andrew Skates and other colleagues at the Centre for Research in Social Simulation (University of Surrey), and at Sandtable, and the TELL ME project partners for their input into the development of the TELL ME model.

Notes

[1]This research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013), Grant Agreement number 278723. The full project title is TELL ME: Transparent communication in Epidemics: Learning Lessons from experience, delivering effective Messages, providing Evidence, with details at `http://tellmeproject.eu/`.

[2]The model and supporting documentation are available from several online locations. The EU project site links to the model code and users' guide at `http://www.tellmeproject.eu/node/392`, together with reports concerning the project. The model and users' guide are also lodged with OpenABM at `https://www.openabm.org/model/4536/version/1`. The model and users' guide are also available from the CRESS website at `http://cress.soc.surrey.ac.uk/web/resources/models/tell-me-model`, as is the working paper with the detailed technical information. The calibration simulation dataset is available on request from the first author.

[3]Similar functionality could be achieved within an open source environment by combining tools: one for the parameter space sampling and simulation management (such as OpenMOLE, MEME or the lhs and RNetLogo packages in R), and another to analyse the results and calculate the dominance fronts (such as the tunePareto package in R).

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179 – 211. doi:http://dx.doi.org/10.1016/0749-5978(91)90020-T

Badham, J. M. & Gilbert, N. (2015). TELL ME design: Protective behaviour during an epidemic. CRESS Working Paper 2015:2, Centre for Research in Social Simulation, University of Surrey

Center for International Earth Science Information Network - CIESIN - Columbia University & Centro Internacional de Agricultura Tropical - CIAT (2013). Gridded population of the world, version 3 (GPWv3): Population density grid, future estimates (2015). *NASA Socioeconomic Data and Applications Center (SEDAC)*

Chang, K. (2015). Parallel Coordinates v0.5.0. Available from https://syntagmatic.github.io/parallel-coordinates/

Cowling, B. J., Ng, D. M., Ip, D. K., Liao, Q., Lam, W. W., Wu, J. T., Lau, J. T., Griffiths, S. M. & Fielding, R. (2010). Community psychological and behavioral responses through the first wave of the 2009 influenza A(H1N1) pandemic in Hong Kong. *Journal of Infectious Diseases*, *202*(6), 867–876

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on*, *6*(2), 182–197

Diekmann, O. & Heesterbeek, J. (2000). *Mathematical Epidemiology of Infectious Diseases*. Wiley Chichester

Durham, D. P. & Casman, E. A. (2012). Incorporating individual health-protective decisions into disease transmission models: A mathematical framework. *Journal of The Royal Society Interface*, *9*(68), 562–570

European Centre for Disease Prevention and Control (2010). The 2009 A(H1N1) pandemic in Europe. Tech. rep., ECDC, Stockholm

Fielding, J. E., Kelly, H. A., Mercer, G. N. & Glass, K. (2014). Systematic review of influenza A(H1N1)pdm09 virus shedding: Duration is affected by severity, but not age. *Influenza and Other Respiratory Viruses*, *8*(2), 142–150. doi:10.1111/irv.12216

Inselberg, A. (1997). Multidimensional detective. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, (pp. 100–107). IEEE

Maddux, J. E. & Rogers, R. W. (1983). Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology*, *19*(5), 469 – 479. doi:http://dx.doi.org/10.1016/0022-1031(83)90023-9

Moss, S. (2007). Alternative approaches to the empirical validation of agent-based models. *Journal of Artificial Societies and Social Simulation*, *11*(1), 5

Müssel, C., Lausser, L., Maucher, M. & Kestler, H. A. (2012). Multi-objective parameter selection for classifiers. *Journal of Statistical Software*, *46*(5)

Railsback, S. F. & Grimm, V. (2012). *Agent-Based and Individual-Based Modeling: A Practical Introduction*. Princeton University Press

Rosenstock, I. M. (1974). The health belief model and preventive health behavior. *Health Education & Behavior*, *2*(4), 354–386. doi:10.1177/109019817400200405

Sandtable (2015). White paper: The Sandtable model foundry

Schmitt, C., Rey-Coyrehourcq, S., Reuillon, R. & Pumain, D. (2015). Half a billion simulations: Evolutionary algorithms and distributed computing for calibrating the SimpopLocal geographical model. *Environment and Planning B*, *42*(2), 300–315. doi:10.1068/b130064p

United Nations, Department of Economic and Social Affairs, Population Division (2013). World population prospects: The 2012 revision (total population, 2015, medium variant)

Wiegand, T., Revilla, E. & Knauer, F. (2004). Dealing with uncertainty in spatially explicit population models. *Biodiversity & Conservation*, *13*(1), 53–78. doi:10.1023/B:BIOC.0000004313.86836.ab

Wilensky, U. (1999). *NetLogo*. Center for Connected Learning and Computer-Based Modeling, Northwestern University