PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Towards victim-oriented crime modelling in a social science e-infrastructure

Nick Malleson and Mark Birkin

Phil. Trans. R. Soc. A 2011 **369**, 3353-3371
doi: 10.1098/rsta.2011.0142

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2011) **369**, 3353–3371
doi:10.1098/rsta.2011.0142

Towards victim-oriented crime modelling in a social science e-infrastructure

By Nick Malleson* and Mark Birkin

School of Geography, University of Leeds, Leeds LS2 9JT, UK

The National e-Infrastructure for Social Simulation (NeISS) is a multi-disciplinary collaboration between computation and social science within the UK Digital Social Research programme. The project aims to develop new tools and services for social scientists and planners to assist in performing 'what-if' scenario predictions in a variety of policy contexts. A key part of the NeISS remit is to explore real-world scenarios and evaluate real policy applications. Research into the processes and drivers behind crime is an important application area that has major implications for both improving crime-related policy and developing effective crime prevention strategies. This paper will discuss how the current e-infrastructure and available microsimulation tools can be used to improve an existing agent-based burglary simulation (BurgdSIM) by including a more realistic representation of the victims of crime. Results show that the model produces different spatial patterns when individual-level victim data are used and a risk profile of the synthetic victims suggests which types of people have the largest burglary risk.

Keywords: e-infrastructure; crime simulation; agent-based modelling; microsimulation; synthetic data

1. Introduction

The National e-Infrastructure for Social Simulation (NeISS) [1] is a multi-disciplinary collaboration between computation and social science within the UK Digital Social Research programme. The project aims to develop new tools and services for social scientists and planners to assist in performing 'what-if' scenario predictions in a variety of policy contexts. The tools that are already available include:

— a census-extraction tool that can be used to extract census variables for a given region [2];
— a population reconstruction model (PRM) that disaggregates census data to create a simulated population of individuals and households [3];
— a dynamic simulation tool that performs population projections over a number of years; and

*Author for correspondence (n.malleson06@leeds.ac.uk).

One contribution of 12 to a Theme Issue 'e-Science: novel research, new science and enduring impact'.

— a traffic simulator that predicts ward-level traffic congestion and is capable of performing 'what-if' scenarios (such as forecasting the potential effects of the recent congestion charge in Manchester).

A key part of the NeISS remit is to explore real-world scenarios and evaluate real policy applications. Research into the processes and drivers behind crime is an important application area that has major implications for both improving crime-related policy and developing effective crime prevention strategies. This paper will discuss how the current e-infrastructure and the above tools can be used to improve an existing crime model (BurgdSIM) [4,5] by including a more realistic representation of crime victims.

The layout of the paper is as follows. Section 2 will introduce the NeISS project in detail and §3 will provide information about the tools that are currently available. Following this, §4 will discuss the burglary model and indicate how the currently available tools can be used to enhance it, aligning the model with modern criminology theories regarding the impacts of victim behaviour on crime. Following this, §5 will cover the technical details regarding how the model is being integrated into the infrastructure and provide preliminary results. Finally, §6 will draw conclusions.

2. The NeISS project

(*a*) *NeISS aims and scope*

Social simulation is a relatively new and rapidly expanding field with applications across the academic research community. The aim of the NeISS project is to build a social simulation e-infrastructure covering the whole social simulation life cycle by providing new tools and services to support research communities and policy makers in the public and private sectors. The tools will enable users to create workflows to run their own simulations, visualize and analyse results, as well as publish them for future discovery, sharing and re-use. It is hoped that this will facilitate development and sharing of social simulation resources within the social science community, encourage cooperation between model developers and researchers and help foster the adoption of simulation as a research method in the social sciences. User requirements for the project are being captured through a series of semi-structured interviews with domain experts. Later stages will include workshops with experts and non-experts as well as follow-up interviews.

(*b*) *Components of the infrastructure*

The social simulation modelling process can be seen as a 'life cycle' with four essential components: data assembly and integration; model implementation and deployment; interpretation and evaluation of the model results; and publication and preservation of each of these analytical components [6]. The overall objective of the infrastructure is to support this life cycle in either research or policy applications. To this end, table 1 summarizes the current components that make up the NeISS infrastructure, which can be used to enhance the burglary simulation.

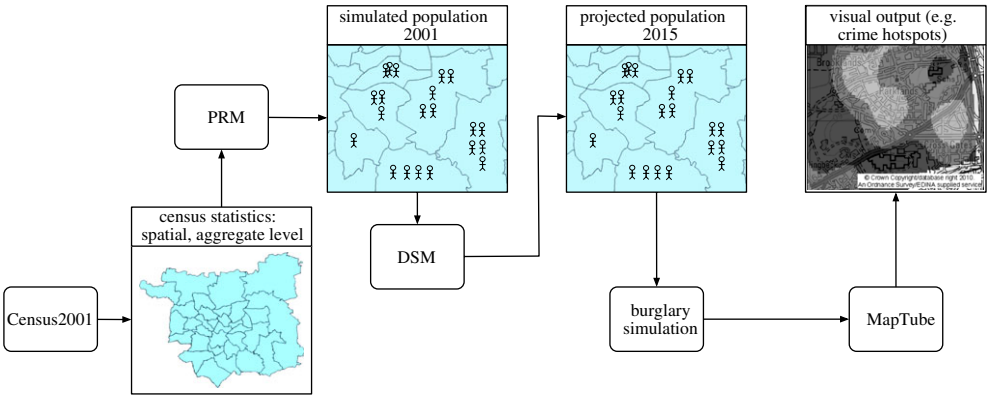

Figure 1. An example workflow that uses existing NeISS components to enhance a burglary simulation. (Online version in colour.)

Table 1. Relevant components of the infrastructure.

component	type	description
Census2001	data	data service providing access to census data
PRM	model	the 'population reconstruction model'—used to synthesize individual-level datasets from aggregate-level census data
DSM	model	the 'dynamic simulation model'—capable of simulating the processes of birth, death, marriage, migration, etc. to project a population through time
MapTube	interpretation	a service that provides the creation of maps

All components are implemented as Web services, which allows them to be executed individually or chained together using a scientific workflow management system such as Taverna [7]. Figure 1 provides an example workflow, which can be used to generate a synthetic population of individuals that can be used as an input into the burglary simulation. Each of the individual components will be discussed in greater detail later. This example is a general illustration of the types of procedures that will be possible using NeISS.

— The process starts by using the Census2001 service to access data from the UK 2001 census. These data are aggregated to the output area geography and, as such, it is not possible to isolate individual people using the data directly. However, social simulations frequently simulate at the level of the individual person and subsequently require individual-level data.
— The PRM service is a microsimulation procedure that is able to simulate a synthetic population of individuals from aggregate census data. This individual-level population can then be used as an input for other models.
— The dynamic simulation model (DSM) service is a population projection model that simulates the processes of birth, death, marriage and migration to estimate how the demographics of the population will change over time.

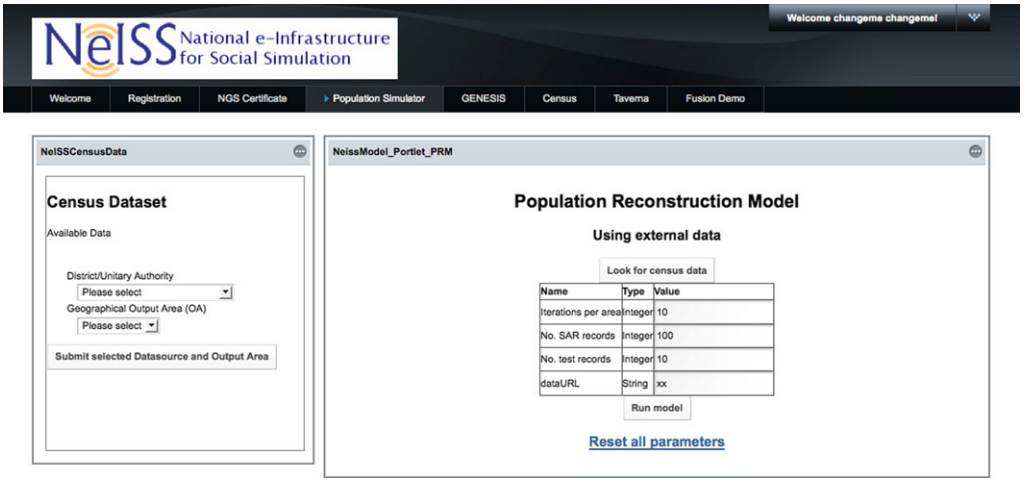

Figure 2. The census and PRM portlets hosted in Liferay portal. (Online version in colour.)

The resulting data can also be used as an input into other models and can provide a means of temporally extending a model that uses census data beyond 2001.

— The projected synthetic population can then be used as an input to any social science model and in this case the population is input into a residential burglary simulation.

— The final stage in the procedure is to export the simulation results to the MapTube service and generate a map to visualize the results (e.g. simulated crime 'hotspots').

(*c*) *Implementation details*

The components of the infrastructure exist as independent Web services. This approach was chosen because it allows for the services to be accessed using a number of mechanisms. Currently, interfaces to the Web services have been implemented as independent Web components called *portlets*, which are hosted in Liferay and Sakai portal servers. Figure 2 illustrates the census and PRM portlets hosted in an instance of the Liferay portal. The NeISS services can be accessed either individually through these portlets, through any client software capable of accessing Web services or through a scientific workflow that users can design themselves. The infrastructure will encourage users to design their own services, which could, if required, interact with the existing ones. This mechanism is being achieved by specifying service inputs and outputs as XML (eXtensible Markup Language) using a common schema. The simulation outlined in §4 is an example of such a tool, which makes use of the outputs from other services.

Scientific workflow management systems, such as Taverna [7], provide a mechanism to automatically orchestrate the execution of services such as those produced for NeISS. Rather than executing services manually, the user can chain them together in a workflow, which simplifies the process of running an experiment, enhances repeatability and also encourages sharing through the use of workflow publishing services such as myExperiment [8].

There are a number of issues that need to be addressed in the proposed architecture, including the definition of interfaces for user-defined services (so that they can interact with new and existing services), data transfer mechanisms between services, data persistence and security/authentication controls. These activities form the majority of the current design and implementation work.

3. Existing NeISS simulation tools

(*a*) *Background: microsimulation*

Microsimulation is a modelling technique in which the members of a population are represented as individual entities rather than as aggregate groups sharing similar characteristics. Although the individual-level data that underpin the models may be generated synthetically, more often they are derived from real records, which can be sourced from either surveys or real administrative data. The method is particularly useful in situations where an individual state can be predicted from specific rules. For example, knowing the income of an individual allows us to infer how much tax they will pay. Microsimulation was popularized by economists, who have argued that it is particularly suited to understanding the detailed consequences of specific policy decisions, and particularly in understanding their distributional consequences across heterogeneous populations [9]. Geographers and other social scientists have also become interested in microsimulation as a means for representing spatial or social disaggregation, and example applications are widespread in healthcare planning, transport research and demographic analysis. Typically, microsimulation is seen as a means for applying well-defined rules to a wide variety of individual circumstances in order to achieve insights with real predictive or applied value, in contrast to agent-based modelling approaches, which focus on a more theoretically rich understanding of the interaction between individuals and their environment [10].

(*b*) *The Census2001 service*

The Census2001 service provides a means to exploit the considerable resources of the UK census by providing access to key data in a number of tables for over 220 000 *output areas* (an output area is the smallest geographical area at which census data are released). The tool consists of a Web service and an associated portlet front-end (as depicted in figure 2). Using the tool, it is possible to specify the individual columns that the user is interested in for a particular geographic region—specified as unitary authority (e.g. 'South Yorkshire') and district (e.g. 'Sheffield'). The data are extracted from a locally hosted database and made available as a comma-separated-values (CSV) file addressed by a randomly generated Uniform Resource Locator (URL). The data can then be downloaded by a user directly (if using the portlet front-end) or passed to another module that requires census data, such as a microsimulation tool.

There are considerable security and authentication considerations that must be taken into account when providing access to this type of data. For example, it must be established whether or not the user is authorized to access census data in the first instance. These issues are being taken into account as part of the security infrastructure of NeISS as a whole. For more information about this aspect of the framework, see Watt *et al.* [2].

(c) *Disaggregating the census using the population reconstruction model*

The UK census provides a source of information about the population that is unparalleled in its robustness and scope. However, what the census fails to provide is a comprehensive and spatially disaggregate representation of individual people and household units. The PRM uses a combination of small-area statistics and anonymized individual records to provide synthetic lists of the entire population of any city or region in the country. The aggregate characteristics of the synthetic population are known to show an extremely close match to the small-area distributions from which they are derived [11], while also providing valuable individual-level detail for applications such as atomic simulation and forecasting.

(d) *Population projections* (*the dynamic simulation model*)

One of the major strengths of microsimulation is in providing a context for the application of rules to individual members of a population (see §3*a*). In the context of the PRM, it is possible to define a series of transition rules that govern the evolution of those individuals through time. Some of these rules are quite simple—for example, that an individual aged x at time t becomes an individual aged $x+1$ at time $t+1$ (other things being equal). Other rules—for example, involving the combination of individuals to form new households—are somewhat more complex. The DSM uses a series of transition rules in order to provide updates (to bring legacy census data up to the present day) and projections of the future population. The DSM has been validated by considering how well it aligns with aggregate forecasts of demographic change [12] and by a process of backcasting to compare the model against historical patterns of evolution [13].

(e) *The MapTube visualization service*

Originally conceived as a map repository, the MapTube website (http://www.maptube.org) was designed for disseminating geographic information to the general public. However, in recent years the technology has been extended to allow end-users to dynamically combine spatially referenced data with boundary data (stored on the MapTube server) to create thematic overlays on Google Maps or OpenStreetMap. Furthermore, a Web service has been implemented as an alternative to the standard Web interface, which allows maps to be automatically generated as part of a workflow. For the data to be mapped using the service, data must be available via a URL (so that the service can access it) and contain at least one column that can be used to link to geographic boundary data (a unique output area code, for example).

4. An agent-based model of burglary (BurgdSIM)

(a) *Background: environmental criminology and mathematical models*

Since the pioneering geography of crime research of the nineteenth century (e.g. Glyde [14]), the field been moving towards analysis using smaller and smaller geographical units. However, modern environmental criminology theories and

recent empirical research suggest that even the smallest areal units of analysis (such as census output areas) hide important intra-area crime patterns [15]. With the crime of residential burglary, for example, it cannot be assumed that all houses in an area are homogeneous with respect to burglary risk because burglars choose individual homes based on their individual characteristics [16]. Researchers are starting to focus on the micro-level factors that make up the 'environmental backcloth' [17] of crime, including the local spatio-temporal behaviour of those involved in a crime (offenders, victims and other people), the effects of different land-use types and other environmental characteristics (levels of street lighting, the number of passers-by, etc.)—all of which cause local effects on criminal behaviour and are lost when data are aggregated.

Despite the movements towards an individual-level perspective on crime, 'traditional' crime models still generally use aggregate crime rates as the dependent variable in a statistic such as a regression equation. This type of model fails to capture the dynamics of the *crime system*, where a crime is an individual incident located in a specific time and space involving individual people. Agent-based modelling, on the other hand, is a methodology that attempts to represent the dynamics of an underlying system more accurately by simulating the behaviour of the individual units ('agents') from which aggregate-level outcomes arise. This is a more natural means of describing a system than by trying to control behaviour from the 'top down' [18]. The methodology has been applied to a wide range of subject areas, including models of crowd behaviour [19], insects [20], land use [21] and retail [22]. However, only in recent years has the methodology been applied to crime, and the model presented here is one of the most advanced in the field [23]. For examples of other agent-based crime models, the interested reader could refer to [24].

Models such as this provide a unique platform for evaluating policy before being implemented in the real world. This is possible because, unlike aggregate mathematical models, the burglary simulation includes a realistic behavioural model and a detailed virtual environment, which, in combination, lead to a comprehensive model that can directly account for the interactions and dynamics that drive the system. Therefore, it is all the more important that models such as this can be integrated into an e-infrastructure in order for them to be available to policy makers and other researchers as a part of a planning support system for new or existing initiatives.

(*b*) *Structure of the agent-based model*

The agent-based model presented here is designed to realistically model the behaviour of individual virtual burglars (the 'agents') as they navigate an accurate urban environment behaving as they would do in the real world (socializing, using substances, sleeping, etc.). Artificial intelligence techniques are used to equip the agents with realistic, dynamic behaviour. In this case, the PECS (physical conditions, emotional state, cognitive capabilities and social status) [25] behavioural framework is used, which provides the agents with competing *needs* and *motives*; it is the strongest of these that determines their current behaviour at any point in time. Some activities that the agents can perform require money (socializing or purchasing drugs, for example) and, therefore, the agents must

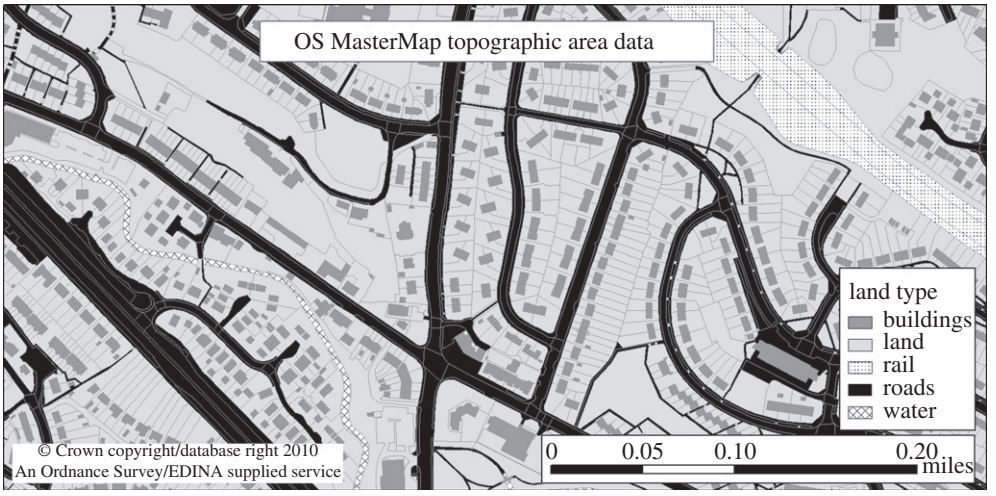

Figure 3. An example of Ordnance Survey MasterMap data used to create the virtual environment.

commit burglary occasionally. In this manner, it is possible to build up city-wide burglary patterns by simulating the behaviour of the individuals who are ultimately responsible for the individual crimes.

Following a movement in environmental criminology towards analysing crime at very small areas [26], the virtual environment in the model (the space that the agents inhabit) has been made as realistic as possible with the available data. The environment itself consists of three layers:

— The *buildings* layer contains the physical buildings that are the potential victims of burglary. Each house in the study area is a unique object with different physical attributes (accessibility, visibility, etc.) reflecting current theoretical understanding of the crime system.
— The *community* layer is used to account for the effects that other people will have on a potential crime occurrence. For example, high levels of community cohesion have been linked to low levels of violent crime because local people are more likely to intervene to prevent a crime from occurring. This layer effectively acts as a proxy for the effects that *non-burglars* will have on crime occurrences, without actually simulating every individual citizen directly.
— The *transport* layer consists of roads and other networks (i.e. rail and bus routes) that the agents can use to navigate the virtual environment.

Ordnance Survey (OS) MasterMap data [27] were used to create the building and transport layers. As figure 3 illustrates, the data are highly detailed and ideally suited to creating an accurate virtual environment containing individual houses. To distinguish residential properties from other types of buildings, the National Land Use Database (NLUD) code was used. This gives information about the type of the building (e.g. code U071 describes residential properties). A substantial amount of information about the physical vulnerability of a house can be gained by analysing the data geographically. Detached houses, for example,

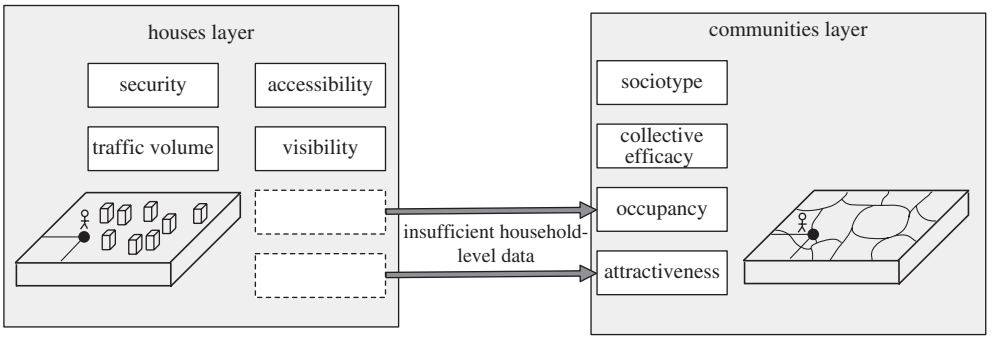

Figure 4. The variables that are used to construct the different layers of the virtual environment. The *occupancy* and *attractiveness* variables should be unique to each individual house, but because they have been derived from the census they have been modelled at an aggregate level.

have been found to be more vulnerable to burglary due to the number of possible entry and exit points, and these types of buildings can be established using simple spatial routines. Unfortunately, no data regarding security precautions (such as burglar alarms or door/window gates) were available, so, although a security variable is included in the model, at this stage it does not influence the results.

However, the geographic data contain no information about the people who live in the houses; i.e. the potential victims of burglary. Environmental criminology research has shown that victim behaviour is extremely important in determining household burglary risk, so this is a feature that should be included in a model. To circumvent this problem, certain variables that would ideally be heterogeneous across a population of households were aggregated so that census data could be used to find values for them. As illustrated in figure 4, these variables relate to the predicted *occupancy* of a house at the given simulation time of day (research has found that burglars are much less likely to attempt a burglary if the house is occupied) and the *attractiveness* of the house (affluent properties are generally seen as more attractive than others).

It is unfortunate that a model with such an accurate representation of the physical environment must aggregate certain key variables due to a lack of individual-level demographic data. Furthermore, even when data aggregation is not necessarily a drawback—factors such as community cohesion are more suitably modelled at the aggregate level—they still rely on census data that are, at the time of writing, extremely outdated.

However, the microsimulation models described in §3 are ideally suited to address these drawbacks. The PRM is able to create a population of individual people and the DSM is able to project this population forward in time. Therefore, integration of the burglary simulation into NeISS holds obvious advantages in terms of bringing the existing model in line with current criminological thinking (that crime should be analysed in terms of the micro-units that influence it [26]). Another advantage is that, once the tool is integrated into the infrastructure, it will be available for others to use, including researchers, policy makers and the police.

It is also important to note that, although we focus specifically on a burglary simulation, obtaining useful individual-level demographic data is a major stumbling block for many agent-based models in the social sciences.

(c) Model implementation: SIMPHONY and the Grid

This section will briefly outline the technical implementation details so that it is clear what the challenges are for including this type of model in a general infrastructure. The model itself is implemented in Java using the Repast SIMPHONY [28] agent-based modelling library. The library comes equipped with an interface, which provides a graphical display of a running model and run-time controls, but it is also possible to execute a model remotely or by using a console. A model can be packaged with the required Repast libraries to run as a stand-alone Java program on any computer that has a Java VIRTUAL MACHINE installed. The library is able to read and write geographical data in the form of ESRI (Environmental Systems Research Institute) Shapefiles, and therefore does not need to be tightly coupled to a geographical information system (GIS), unlike other systems such as Agent Analyst [29]. This has benefits in terms of efficiency (there are no costly model–GIS interactions) and flexibility (the system that executes a Repast model does not need to have a particular GIS installed).

At first sight, incorporating the burglary model into an infrastructure seems trivial, as it can run as a simple Java program and requires only limited extra resources such as input spatial data. It is, however, extremely computationally expensive. The experiments discussed in §5 run for 30 simulated days (approx. 40 000 iterations) with 250 agents in an area of approximately 3 km^2 and require approximately 20 h on a typical desktop personal computer. This alone is not prohibitively expensive, but the probabilistic nature of the model means that it requires as many as 100 separate runs to ensure that results are reliable. The resources required to generate model results, therefore, can only be found using a high-performance computer system, and in this case the UK e-Science National Grid Service (NGS) [30] was used. This adds an additional layer of complexity with regard to incorporating the model into the NeISS infrastructure, as discussed in §5.

Parallelizing the simulation to run on NGS resources required relatively few changes to the underlying code base. As the main purpose was to execute a large number of independent models, 'lazy parallelization' could be used whereby each separate compute node was responsible for running a single model in isolation. The parallelized model works as follows. On initialization, a single node is allocated as a 'master' and the remaining are allocated as 'slaves'. The master is responsible for instructing slaves to run jobs; it maintains a list of completed and pending jobs and allocates them to free slaves as appropriate. Job descriptions are passed to slaves that contain information about a particular model run, so it is possible for the master to implement a parameter sweep in this manner. The MPJ Express software library—which is based on the popular Message Passing Interface (MPI) specification—was used for message passing between nodes. Model results are stored directly by each running instance in a shared database (Oracle in this case), so it is not necessary for the master node to collate results.

(d) Model validation

The evaluation of agent-based models is a divisive subject as there is no established method that can be used across different research projects. Evaluation of this model was divided into three separate tasks: verification, calibration and validation (see [23] for more details).

Verification consisted of ensuring that the model was logically valid, i.e. it had been correctly programmed. Lacking sufficient resources to re-implement the model using a different programming language, a novel approach was employed whereby the type of virtual environment was varied without changing the underlying model. Three types of environment were used: a 'null' environment that contained no virtual space (all 'journeys' took the same amount of time), a 'grid' environment that consisted of a regular grid of cells, and a 'GIS' environment that accurately reflected a real city. By running the model in each type of environment it was possible to fully understand the dynamics of the model before increasing geographical complexity; any unexpected results were caused by logical errors and were not artefacts of the intricacies of the virtual environment.

The process of calibration involved optimizing the model parameters so that the results reflected field data. In this case, the field data were the times and locations of all burglaries that were reported to the police in the study area for the period 2001–2002. Although an artificial intelligence optimization routine such as a genetic algorithm would have been preferable, the computation time required to execute such a procedure was prohibitive (it would have required thousands of model runs) so the model was calibrated manually. The final procedure, validation, involved comparing the model to data other than that on which it was calibrated. In this case, burglary rates from a different time period were used.

Although the evaluation *procedure* is straightforward, the data used are made up of points in space, and therefore actually comparing two datasets is non-trivial. Frequently, point data are first aggregated to some administrative boundary so that traditional statistics (such as R^2) can be used, but this approach is highly susceptible to the modifiable areal unit problem [31] and provides no information about the spatial accuracy at which the model is able to predict. To avoid this pitfall here, a new method of comparing point datasets (based on [32]) was developed. The method works by first placing a number of cellular grids of varying resolutions over the study area, aggregating points to cells in the grids and then calculating error using a traditional goodness-of-fit statistic—here R^2 and the standardized root mean squared error (s.r.m.s.e.) were used to provide a comprehensive error assessment. By using various different sized grids, the method is able to indicate at what resolution a model is performing well and also minimize the effects of the modifiable areal unit problem.

This type of model evaluation is part of the 'interpretation' phase of the social simulation life cycle (discussed in §2). At present, NeISS contains the MapTube tool, which supports interpretation of data through mapping. An obvious avenue for future development, therefore, is to add interpretation tools that are able to comprehensively compare spatial datasets, and point patterns in particular. These will then be linked to visualization tools, completing the life cycle in an e-infrastructure.

5. Integrating BurgdSIM into NeISS

Having outlined the NeISS project and an agent-based model of burglary (BurgdSIM) in §4, this section will discuss how the model can be integrated into NeISS. The advantages of integrating the simulation include being able to provide

it with individual-level demographic data (to model the 'victims' of burglary more realistically) and also allowing others to use the model to predict burglary rates in any area that they are interested in. Although this work is ongoing, partial integration has been completed, and preliminary results comparing the model before and after the new data are used will be discussed in §5*c*.

(*a*) *Preparing the data*

The process of extracting, preparing and synthesizing the data for the burglary model is non-trivial. In summary, the process is as follows:

1. Run the PRM to disaggregate the UK census, creating a population of synthetic individuals.
2. Optionally run the DSM to advance the population through time.
3. Link the synthetic population to real house objects (as established from OS MasterMap data).
4. Execute the simulation in a grid (or other high-performance) environment.

The population modelling tools at stages 1 and 2 are already available as part of the NeISS infrastructure, as discussed in §3. However, at this stage of the infrastructure development, linking the synthetic data to geographical data (stage 3) must be completed using bespoke routines that are not yet part of the infrastructure. The output population from microsimulation models consists of a list of people with gender, age, social/employment group, marital status and ethnicity. Each person is part of a household and each household has a particular type (one of detached, semi-detached, terraced or flat). It is known in which output area a household is situated (the output area is the smallest scale administrative area at which census data are released) but not which *individual house* they actually live in. To assign synthetic households to actual buildings, the geographical data (released as part of the OS MasterMap product) were analysed spatially to determine building type and then synthetic households were assigned randomly to a house of the correct type within an output area. Clearly, there are a number of ways in which error can enter the final output when this method is used, so immediate future work will explore how this process can be improved. One way could be to base the allocation on more than simply house type. For example, the affluence or income of the household could be used to assign richer families to physically larger houses. If house price data are available, then these could also be used. In the next phase of development of the infrastructure, the linkage process will be automated through the deployment of a data fusion tool [33], for which this crime model will provide a useful testbed.

Once the households have been assigned to buildings, it is necessary to adapt them so that we can establish how affluent they are and at what times their house is likely to be unoccupied—these are the 'key' variables required for the burglary model, as discussed in §4. This is another process that should be conducted with the data fusion tool but at present must be completed using routines that are not available as part of the infrastructure. The employment group of the synthetic individuals was used to estimate both occupancy and affluence and can be one of four types: *managerial*, *intermediate*, *manual* and *other* (including unemployed, retired and students). Here, overall household affluence is estimated directly from

Table 2. The different household occupancy behaviour as implemented in the model.

group	description	occupancy
family	the house has young children and someone will be at home during the day to look after them	house occupied during the day and in the evenings
students	the household is made up of university students	house occupied during the day but not in the evenings
unemployed	no one in the household is employed	house occupied at all times

the employment type of the head of household (*managerial* types being the most affluent and *other* the least). Occupancy likelihood can be estimated by placing people into one of the groups illustrated in table 2 (note that if a household does not fall into one of the specified groups then they are assumed to have typical '9 to 5' jobs).

It is important to note that we treat occupancy not as a binary value but rather as a probability, e.g. it is more probable that a house containing unemployed synthetic people will be occupied during the day than one where all residents work. When burglar agents make their decision about whether or not to burgle, this probability is considered along with other variables such as the apparent security of the house, the volume of pedestrian/vehicle traffic on the adjacent road, the visibility of the house to surrounding neighbours, etc.

(*b*) *Using grid infrastructure*

As illustrated in §2, the infrastructure is built largely from interconnected Web services with portlet interfaces. Although this design is suitable for services that are not overly computationally expensive, the burglary model requires substantial compute resources and it cannot be expected that the server hosting the simulation service will be able to meet these requirements. Therefore, hardware has been purchased specifically for the purposes of running computationally expensive NeISS models such as that described here. Although the results presented in this paper were generated using NGS resources, when the burglary system is fully integrated into the NeISS infrastructure it is expected to run on this new system instead of the NGS. The hardware is being added to a local grid system hosted at the University of Leeds called 'ARC1'. As the NeISS developers effectively own the hardware, there are fewer restrictions on the terms of use and immediate access is guaranteed (it will not be necessary to wait for other jobs in a queue to finish first). This is particularly important for the infrastructure to be able to deliver model results rapidly. Therefore to run a simulation the user accesses the burglary Web service through NeISS to initiate the model, which in turn uses SSH (Secure Shell) commands to communicate with the ARC1 grid. This has the added advantage that restricted or private data (such as OS MasterMap, the census and police data) that the model requires can be stored on ARC1 privately. As the data do not need to form part of the model output, a user who would otherwise not have the appropriate rights to use the data can still make use of the simulation. If users had to provide their own input data, it

would be extremely unlikely that policy makers or other interested parties (such as the police) from outside academia would be able to use the model. Two side effects of using privately owned hardware, rather than the NGS, are that users will only be able to access resources through the NeISS services (not directly as they could if they had access to the NGS) and NeISS becomes responsible for user management and accounting.

The other major difficulty that must be overcome relates to output data, as agent-based models can be extremely data intensive. The burglary simulation outlined here could potentially output the (x, y) coordinates of every agent at every time step: 250 agents $\times$ 40 000 iterations = 10 000 000 pieces of information. Also, this does not include any data regarding the 50 000 houses that make up the current study area and assumes the model will run only once (ideally it should run multiple times to take account of its probabilistic nature). Previously, results were stored in an NGS-provided Oracle database. This had the major advantage that SQL (Structured Query Language) could be used to explore the results in a manner that would not be possible if the results were stored in a different format such as plain text. For example, using SQL it is possible to create queries such as: find all the information about the burglars who burgled house a between the hours of 08.00 and 10.00 and then went on to purchase drugs from dealer x.

However, returning results data in their entirety is not feasible and two potential solutions will be investigated. The most commonly requested results, such as the final locations of all burglaries, can be provided in the form of URLs to plain text files that the user can download. In addition, to maintain the richness of the results data that are present in the results database, an interface to the database itself can be provided either directly—using WS-DAI (Web Services Data Access and Integration), for example—or through a bespoke Web service interface that allows the construction of detailed database queries.

(*c*) *Preliminary results*

Initial experiments have been conducted to explore the effects of replacing aggregate population groups with individual-level victims within the burglary simulation. The process discussed in §5*a* was followed manually and the model was executed on the NGS. The simulation configuration was identical, with the exception of the disaggregate victims. Figure 5 presents a comparison of the results before and after the changes using the cellular-grid aggregation method discussed in §4*d*.

To assess error between results we use the *relative percentage difference* between two cells, y_i and y'_i, defined as the difference between the proportions that the cells contribute to the total observation count:

$$100 \times \left(\frac{y_i}{\sum y}\right) - 100 \times \left(\frac{y'_i}{\sum y'}\right). \tag{5.1}$$

The spatial variations between the aggregate and disaggregate models are relatively modest, with the percentage difference in burglary rates for each cell ranging from a minimum of −0.1 per cent to a maximum of 0.3 per cent. This is not unexpected, as the changes only influence a small part of the burglars' decision process. However, it is clear that the differences between the two datasets are not random, with a strong overall pattern to the distribution of areas that experience

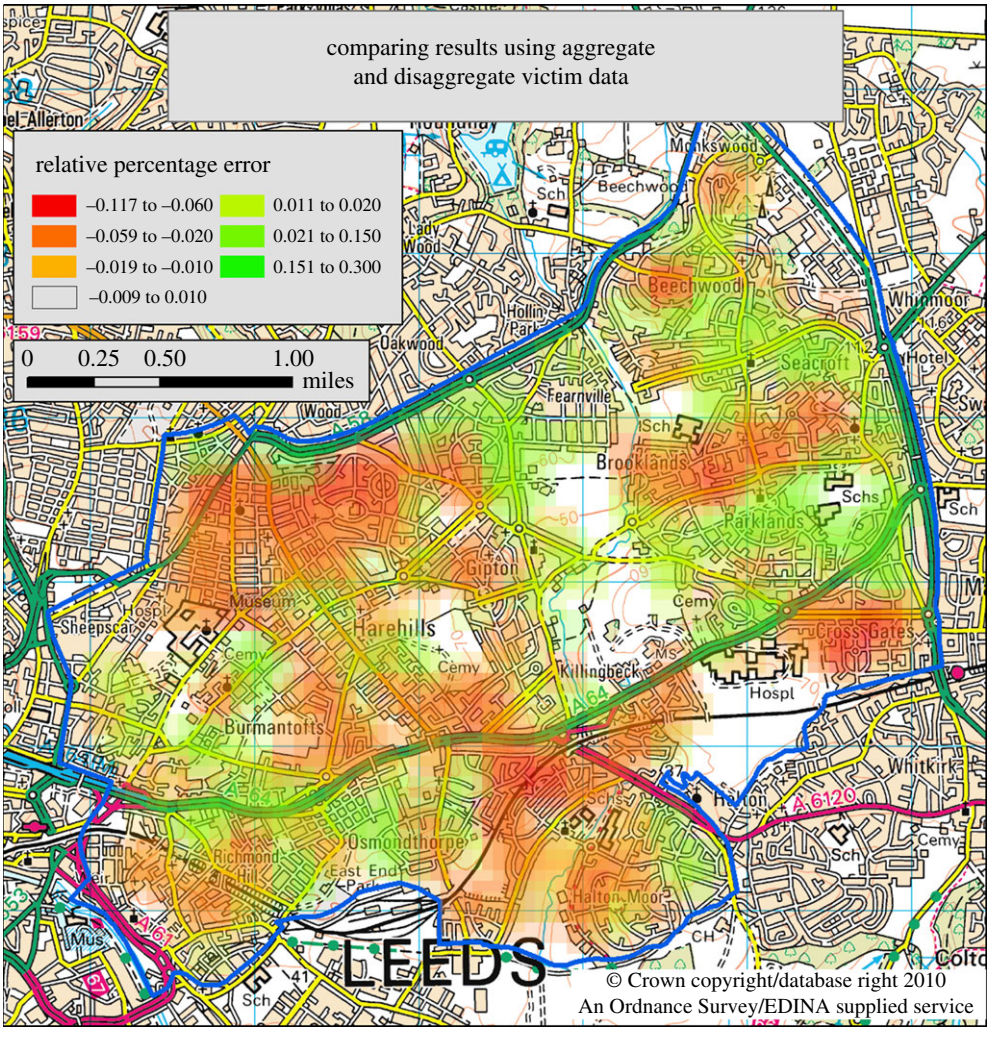

Figure 5. Comparing results before and after individual-level victims were added to the model. (Online version in colour.)

an increase in simulated burglary and others that see a decrease. An explanation for this effect is not immediately clear and further investigation is required. However, these preliminary results show that disaggregating the model to better represent the victims of crime *does* have an effect over the locations of burglary.

Another advantage with using disaggregate victims in the model is that it allows for investigation into *individual victims* of burglary. To illustrate this, figure 6 profiles the age, gender, social group and ethnicity of the burglary victims. Note that the classifications of social group and ethnicity are derived for use in the microsimulation; there are a larger number of groups in the original census data. With the exception of the social group, none of the attributes are taken into account by the virtual burglars when making their burglary decisions, so these trends are a result of where the individuals live—and the types of houses and neighbourhoods they inhabit—rather than an artefact of model rules (there

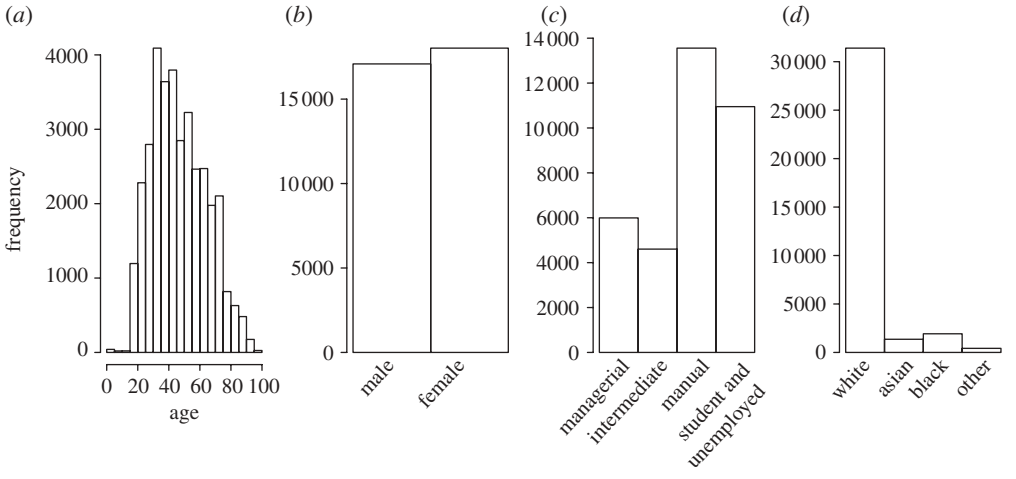

Figure 6. Profiles of the synthetic victims of burglary: (*a*) victim age distribution, (*b*) gender of victims, (*c*) victim social group, and (*d*) ethnicity of victims.

is nothing in the model to make burglars favour victimizing women over men, for example). Interestingly, unemployed groups appear to be burgled more than managerial/intermediate groups, even though the unemployed have a reduced risk because they spend more time at home. This is most probably a result of *where* these groups of people live. If victims live near the potential burglars, then their risk increases simply because the agents are aware of them as potential victims. The reverse is true for the more affluent groups who live farther from the burglar agents. Although more research is needed to explore these patterns in detail, this is valuable information that could not have been generated from the previous version of the model when victim demographics were aggregated to the neighbourhood.

6. Conclusion

This paper has discussed the ongoing work developing a National e-Infrastructure for Social Simulation (NeISS), which aims to support the social simulation life cycle. In its current form, the framework includes tools to extract data, generate synthetic populations, run simulations using the data and then visualize the outputs. In particular, we have discussed how these current tools and data can form an essential part of research in a particular social science field, that of crime simulation.

Current criminology research suggests that, for crime models to reliably capture the dynamics of the underlying system, they must take into account the individual-level dynamics that frame a crime event. These include the behaviour of offenders, victims and other people, the physical impacts of the surrounding environment and community-wide factors such as collective efficacy. Aggregating data hides many of these lower-level dynamics that are essential for

determining where and when crimes are likely to take place. Therefore, a synthetic population of individual households would be an extremely useful input for a crime model.

To illustrate these benefits, an advanced agent-based model was introduced, and this was seeded with individual-level synthetic data generated using a microsimulation tool. The results showed that, although there are only modest differences between the aggregate and non-aggregate versions of the model, there is a clear spatial change that occurs when individual victim demographics are taken into account. This is an encouraging result and paves the way for further investigation into why the patterns are different.

Immediate future work will focus on integrating the burglary simulation into the wider framework. This will not only allow other researchers to make use of the simulation—allowing criminologists to simulate burglary in their local area, for example—but also provide an opportunity for policy makers and other non-academics to test the impacts of new and existing policies on burglary before their implementation. Existing results in this area have shown promise [23].

There are a number of challenges that must be faced before the simulation (and others like it) can be fully integrated into NeISS. From an infrastructure perspective, the security/authorization model is still under development, so it is not possible to distinguish between users and identify which data permissions they have. Also, the data analysis and manipulation tools that are required to transform the outputs from one model into a form that is recognized by another are still under development (the simulations discussed here used manually transformed data). Wider questions still remain regarding how to manage the data-intensive model outputs, how to control the resources required to execute such a computationally intensive model and how to manipulate complex model parameters that are not expressed numerically. Owing to its data and computational requirements, the simulation discussed here is likely to be one of the most difficult that the infrastructure will have to cope with, so providing solutions to these problems will greatly enhance the flexibility of the infrastructure as a whole.

We would like to thank JISC for funding NeISS research and Ordnance Survey/EDINA for providing the necessary spatial data. We would also like to thank the NGS for providing the required compute and data resources as well as technical support.

References

1 NeISS. 2010 National e-Infrastructure for Social Simulation. See http://www.neiss.org.uk.

2 Watt, J., Sinnott, R., Jiang, J., Doherty, T., Higgins, C. & Koutroumpas, M. 2009 Tool support for security-oriented virtual research collaborations. In *2009 IEEE Int. Symp. Parallel and Distributed Processing with Applications* (ed. X. Liao), pp. 419–424. Silver Spring, MD: IEEE Computer Society.

3 Birkin, M., Turner, A. & Wu, B. 2006 A synthetic demographic model of the UK population: methods, progress and problems. In *Proc. Second Int. Conf. e-Social Science, Manchester, UK*. National Centre for e-Social Science. See http://www.ncess.ac.uk/events/conference/2006/papers.htm.

4 Malleson, N., See, L., Evans, A. & Heppenstall, A. In press. Implementing comprehensive offender behaviour in a realistic agent-based model of burglary. *Simulation: Trans. Soc. Model. Simul. Int.* (doi:10.1177/0037549710384124)

5 Malleson, N., Heppenstall, A. & See, L. 2010 Crime reduction through simulation: an agent-based model of burglary. *Comput. Environ. Urban Syst.* **34**, 236–250. (doi:10.1016/j.compenvurbsys.2009.10.005)
6 Townend, P., Xu, J., Birkin, M., Turner, A. & Wu, B. 2009 MoSeS: Modelling and simulation for e-social science. *Phil. Trans. R. Soc. A* **367**, 2781–2792. (doi:10.1098/rsta.2009.0041)
7 Oinn, T. *et al.* 2006 Taverna: lessons in creating a workflow environment for the life sciences. *Concurr. Comput.: Pract. Exper.* **18**, 1067–1100. (doi:10.1002/cpe.993)
8 De Roure, D., Goble, C. & Stevens, R. 2009 The design and realisation of the myExperiment virtual research environment for social sharing of workflows. *Future Generation Comput. Syst.* **25**, 561–567. (doi:10.1016/j.future.2008.06.010)
9 G. Orcutt. 1957 A new type of socio-economic system. *Int. J. Microsimulation* **1**, 3–9.
10 Wu, B., Birkin, M. & Rees, P. 2008 A spatial microsimulation model with student agents. *Comput. Environ. Urban Syst.* **32**, 440–453. (doi:10.1016/j.compenvurbsys.2008.09.013)
11 Harland, K., Heppenstall, A., Smith, D. & Birkin, M. In press. Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *J. Artif. Soc. Social Simul.*
12 Wu, B., Birkin, M. & Rees, P. 2011 A dynamic MSM with agent elements for spatial demographic forecasting. *Soc. Sci. Comput. Rev.* **29**, 145–160. (doi:10.1177/0894439310370113)
13 Birkin, M. & Malleson, N. 2010 An investigation of the robustness of a dynamic microsimulation model of urban neighbourhood dynamics. Paper presented to the North American Regional Science Council (NARSC), 10–13 November 2010, Denver, US. See http://www.geog.leeds.ac.uk/fileadmin/downloads/school/people/academic/m.birkin/backsim_paper_v2.pdf.
14 Glyde, J. 1856 Localities of crime in Suffolk. *J. Statist. Soc. London* **19**, 102–106. (doi:10.2307/2338263)
15 Andresen, M. A. & Malleson, N. 2011 Testing the stability of crime patterns: implications for theory and policy. *J. Res. Crime Delinquency* **48**, 58–82. (doi:10.1177/0022427810384136)
16 Rengert, G. & Wasilchick, J. 1985 *Suburban burglary: a time and a place for everything.* Springfield, IL: Charles Thomas.
17 Brantingham, P. L. & Brantingham, P. 1993 Environment, routine, and situation: toward a pattern theory of crime. In *Routine activity and rational choice* (eds R. Clarke & M. Felson), Advances in Criminological Theory, vol. 5. New Brunswick, NJ: Transaction Publishers.
18 Bonabeau, E. 2002 Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl Acad. Sci. USA* **99**, 7280–7287. (doi:10.1073/pnas.082080899)
19 Turner, A. & Penn, A. 2002 Encoding natural movement as an agent-based system: an investigation into human pedestrian behaviour in the built environment. *Environ. Plann. B* **29**, 473–490. (doi:10.1068/b12850)
20 Parry, H., Evans, A. J. & Morgan, D. 2006 Aphid population dynamics in agricultural landscapes: an agent-based simulation model. *Ecol. Model.* **199**, 451–463 (Special Issue on Pattern and Processes of Dynamic Landscapes).
21 Ngo, T. N., See, L. M. & Drake, F. 2009 An agent-based approach to simulating the dynamics of shifting cultivation in an upland village in Vietnam. *Rev. Int. Géomatique* (*Int. J. Geomatics Spatial Anal.*) **19**, 493–522. (doi:10.3166/geo.19.493-522).
22 Heppenstall, A. J., Evans, A. J. & Birkin, M. H. 2006 Using hybrid agent-based systems to model spatially-influenced retail markets. *J. Artif. Soc. Social Simul.* **9**, 2. See http://jasss.soc.surrey.ac.uk/9/3/2.html.
23 Malleson, N. 2010 Agent-based modelling of burglary. PhD thesis, School of Geography, University of Leeds, UK.
24 Liu, L. & Eck, J. (eds) 2008 *Artificial crime analysis systems: using computer simulations and geographic information systems.* Hershey, PA: Information Science Reference.
25 Schmidt, B. 2000 *The modelling of human behaviour.* Erlangen, Germany: SCS Publications.
26 Eck, J. E. & Weisburd, D. 1995 Crime places in crime theory. In *Crime and place* (eds J. E. Eck & D. Weisburd). Crime Prevention Studies, vol. 4, pp. 1–33. Monsey, NY: Criminal Justice Press.
27 Ordnance Survey. 2010 OS MasterMap—reliable spatial intelligence. See http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap [accessed August 2010].

28 North, M. J., Howe, T. R., Collier, N. T. & Vos, R. J. 2005 The Repast Simphony development environment. In *Agent 2005 Conf. on Generative Social Processes, Models, and Mechanisms, Argonne National Laboratory, Argonne, IL, USA, October*.
29 The Redlands Institute. 2009 Agent-based modelling extension for ArcGIS users. See http://www.spatial.redlands.edu/agentanalyst/.
30 Richards, A. & Sinclair, G. M. 2009 UK National Grid Service. In *Grid computing: infrastructure, service, and applications* (eds L. Wang, W. Jie & J. Chen), ch. 6, pp. 149–170. Boca Raton, FL: CRC Press.
31 Openshaw, S. 1984 *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography, no. 38. Norwich, UK: GeoBooks.
32 Costanza, R. 1989 Model goodness of fit: a multiple resolution procedure. *Ecol. Model.* **47**, 199–215. (doi:10.1016/0304-3800(89)90001-X)
33 Warner, G. C., Blum, J. M., Jones, S. B., Lambert, P. S., Turner, K. J., Tan, L., Dawson, A. S. F. & Bell, D. N. F. 2010 A social science data-fusion tool and the Data Management through e-Social Science (DAMES) infrastructure. *Phil. Trans. R. Soc. A* **368**, 3859–3873. (doi:10.1098/rsta.2010.0159)