Elsevier Editorial System(tm) for Ecological Modelling

Manuscript Draft

Manuscript Number: ECOMOD1606R1

Title: A comparative analysis of parallel processing and super-individual methods for improving the computational performance of a large individual-based model

Article Type: Research Paper

Section/Category:

Keywords: Agent-based modelling; Individual-based modelling; Parallel computing; Super-individuals

Corresponding Author: Dr Hazel Ruth Parry, Ph.D.

Corresponding Author's Institution: Central Science Laboratory

First Author: Hazel Ruth Parry, Ph.D.

Order of Authors: Hazel Ruth Parry, Ph.D.; Andrew J Evans, Ph. D.

Manuscript Region of Origin: UNITED KINGDOM

A comparative analysis of parallel processing and super-individual methods for improving the computational performance of a large individual-based model

Hazel R. Parry [a,b,*], Andrew J. Evans [b]

[a] *Central Science Laboratory, Sand Hutton, York, YO41 1LZ, England*

[b]*School of Geography, University of Leeds, LS2 9JT, England*

Abstract

Individual-based modelling approaches are being used to simulate larger complex spatial systems in ecology and in other fields of research. Several novel model development issues now face researchers: in particular how to simulate large numbers of individuals with high levels of complexity, given finite computing resources. A case study of a spatially-explicit simulation of aphid population dynamics was used to assess two strategies for coping with a large number of individuals: the use of 'super-individuals' and parallel computing. Parallelisation of the model maintained the model structure and thus the simulation results were comparable to the original model. However, the super-individual implementation of the model caused significant changes to the model dynamics, both spatially and temporally. When super-individuals represented more than around 10 individuals it became evident that aggregate statistics generated from a super-individual model can hide more detailed deviations from an individual-level model. Improvements in memory use and model speed were perceived with both approaches. For the parallel approach, significant speed-up was only achieved when more than five processors were used and memory availability was only increased once five or more processors were used. The super-individual approach has potential to improve model speed and memory use dramatically, however this paper cautions the use of this approach for a density-dependent spatially-explicit model, unless individual variability is better taken into account.

Key words: Agent-based modelling, Individual-based modelling, Parallel computing, Super-individuals

* Corresponding author. Address: Central Science Laboratory, Sand Hutton, York, YO41 1LZ, England. Tel.: +44 1904 462724; Fax: +44 1904 462111.
Email address: h.parry@csl.gov.uk (Hazel R. Parry).

1 Introduction

A desire to better understand and inter-link the complex dynamic structures of ecosystems, along with self-organisation, emergence of spatial and temporal patterns and apparent unpredictability, has prompted a shift in the general approach to ecological modelling today (Grimm and Railsback, 2005; Parrott and Kok, 2000). Following trends in other fields of research, from social science (Gilbert and Troitzsch, 1999) to fluvial sediment transport (Schmeeckle and Nelson, 2003), there has been a shift away from procedural, equation-based models to object-based simulations. These include individual-based models (IBMs), cellular automata and multi-agent simulation (MAS). Such models are concerned with modelling variation among individuals in a population, and the interaction between individuals (DeAngelis and Gross, 1992; Grimm, 1999; Grimm and Railsback, 2005; Grimm et al., 1999; Huston et al., 1988; Judson, 1994; Uchmański and Grimm, 1996). IBM is closely related to multi-agent simulation. MAS has arisen from artificial intelligence (AI) research and is used widely in other fields such as social science and computing (Gilbert and Troitzsch, 1999).

Object-based approaches have been successfully implemented to model a range of ecological systems (for a review see Grimm, 1999; Grimm and Railsback, 2005). They have the potential to further understanding of the local processes that influence regional species population dynamics spatially and temporally, enabling better understanding of how individual local-level and field-scale interactions result in larger scale population distributions. However, some of the potential of MAS and IBM methods is constrained by the demands that may be placed on computing power. For realistic scenarios, it may be nec-

essary to simulate large numbers of individuals. There may also be added complexity, such as in models where interactions or agent-density are important, but populations are sparse (for example insect populations, Parry et al., 2006b, 2004), or where agents are memory-heavy because they are complex (e.g. forest dynamics, Verzelen et al., 2006), or multiple types of agent are used, such as in models of competition or predator-prey models (e.g. Hosseini, 2006). Haefner (1992: pp.156-157), with some foresight, identified future developments in ecological individual-based models that would benefit from advanced computing as: multi-species models; models of large numbers of individuals within a population; models with greater realism in the behavioural and physiological mechanisms of movement; and models of individuals with 'additional individual states' (e.g. genetic variation).

The key limitations imposed by computer hardware are: (1) the number of calculations that can be performed in a reasonable time (controlled by processing power); (2) the number of agents that can be modelled (controlled by memory). Relationships were determined between increasing numbers of initial agents and the memory and simulation speed of a simple agent model (described in section 2) run on a single 2.80 GHz Intel Xeon processor 2097 MB RAM machine. Once the model is running, processor use of memory is nearly linear and using an equation derived from the curve we can predict that at a maximum available memory capacity of 1.5GB RAM on the single machine, the theoretical limit to the initial number of agents is approximately 7,500,000. However, at this limit, the simulation is calculated to take approximately 1 million seconds (12 days) to run (calculated from the simulation speed curve using a quadratic function). This may be an under-estimate, as there is a slight processing overhead for dealing with additional memory blocks, which

would result in a less linear relationship over time. The potential number of replicates of a stochastic simulation are affected by this, so for example the use of Monte Carlo techniques would no longer be possible.

There are a number of solutions to the problem of large numbers of individuals in an individual- or agent-based simulation (table 1). Solutions may range from hardware investment (such as obtaining a more powerful computer) to computational solutions, such as changes in the software design (e.g. parallelisation) or changes in the the model structure. This paper evaluates and compares two such solutions to this problem. The first is a computational solution that requires access to networked hardware: to parallel program the model software to work across a network of powerful computers, so splitting the processing/data load. The second is a mathematical solution, where the model itself is altered so that individuals are aggregated into 'super-individuals' (after Scheffer et al., 1995). The two methodologies are applied to a case study of a spatially-explicit, individual-based simulation of aphid population dynamics in agricultural landscapes (Parry et al., 2006b). The comparability of the model results with the original model are first determined, then the two approaches are evaluated in terms of improved model efficiency (memory use and speed).

A key advantage of parallel programming is that it maintains the strengths of an individual-based approach whilst potentially increasing the number of agents that can be simulated, as opposed to the super-individual approach where the key interactions in the model are altered. However, parallelisation is a complex solution, and although the agent interactions are unchanged significant restructuring of the model software is needed. Haefner (1992) outlined the potential applications of parallel computing to individual-based simula-

tions in ecology, but also pointed out the need for ecological modellers to improve their technical knowledge. Few examples of parallel simulations exist in the ecological literature to date. Some examples of note are a parallel simulation of a school of fish by Lorek and Sonnenschein (1995) and several in relation to the ATLSS project (`http://atlss.org/`), which include a parallel individual-based model of Everglades deer ecology by Abbott et al. (1997) and a parallel spatially-explicit fish model (ALFISH) by Wang et al. (2004). Other agent simulation examples can be found outside ecology in the use of parallel agents for reducing genetic algorithm search times (Lefley and McKew, 2004) and performing large scale traffic simulations (Dupuis and Chopard, 2001).

The simplicity of the super-individual approach makes it attractive, particularly as it does not require complex programming and powerful computer systems to implement. It maintains the philosophy and integrity of an individual-based approach without reverting to a population model to deal with large numbers of individuals. However, implementations of this approach to date are primarily not spatially-explicit.

2 Application

The results presented in this paper relate to a simplified version of a spatially explicit individual-based simulation model of aphid population dynamics in agricultural landscapes (Parry, 2006; Parry et al., 2006b, 2004). A spatially-explicit IBM of aphid populations was constructed to assess the impact of variation in agronomic practices in time and space. These practices included crop introduction and configuration, pesticide spray application, matrix habitat availability and fragmentation. The impacts that these have upon aphid

populations were observed, including both regional and local population dynamics as well as individual movement paths. A key limitation of the model was the restriction on the number of insect agents that could be modelled. The simplified model was used to explore and evaluate the options for coping with large numbers and complexity in the full model. The simple model moves aphid agents (*Rhopalosiphum padi*) randomly from cell to cell around a uniform landscape and local agent density is recorded. Aphids reproduce parthogenetically with winged and non-winged morphs produced. Density determines the proportion of alate (winged) morphs that are born at each iteration. The simulation is begun with a population of alate agents originating from a central cell in a 50×50 cell landscape, where each cell is 25×25 m. The wind is set to a constant speed of $8kmh^{-1}$ and a constant westerly direction. In the full version of the model there are a number of variables (some of which are density dependent), realistic immigration across a region and a more complex environment. The complexity of the full version of the model increases computational demands beyond those demonstrated here and a large number of agents (several million) were required for the simulation to be realistic at the landscape scale.

Initial populations of 10, 100, 1,000, 10,000, 100,000 and 500,000 aphids were used, originating from a single central cell. Each simulation was run thirty times and an average taken to represent the total population trend over time (as several parameters in the model are stochastic). While the model was allowed to run for 120 days, spatial comparisons were made after 2, 20 and 40 days by creating surfaces that show the mean density in each cell over the thirty runs. Temporal comparisons were made of the population dynamics at the central cell.

3 Parallel computing

3.1 Implementation

In order to address computing problems where a model is hindered by data requirements far larger than can be accommodated at any individual processing element, parallel solutions are often implemented. The combined or 'virtual shared' RAM of several computers is used to cope with the amount of data and processing needed, using a Sequential-Algorithm Multiple-Data approach (SAMD), where the same algorithm is applied to different data items on different processors ("nodes"). The scale of the problem for each individual computer is therefore reduced (often speeding the model up), or more resources are made available (allowing for larger models). To parallelise the model software, sections are run on each node and then the nodes periodically communicate together to share results. This requires somewhat complex communication strategies to make the physically distributed systems act as a single unit.

Key to an efficient parallel model is minimising the inter-processor communications; it is these that take valuable time (Pacheco, 1997). Because of this, in models with static agents that only interact locally it makes sense to divide the environment and agents between processors. Conversely, in models with roaming agents, it makes sense to divide up the agents and have a copy of the whole environment on each processor. Hardest to deal with are the situations where agents roam the environment, but also interact with each other. In such cases dividing up either the agents or the environment results in an increase in inter-processor messages or agent transfer.

In this case study, parallelisation essentially splits the agents in the simulation between a number of processors (nodes), each containing information on the environment and total agent densities. Direct agent interaction (to determine morphology) is mediated by density in the model. Thus, aphids can be split between processors because it is not necessary for each processor to know the exact position of all the aphids, just the density within each section of the environment on the other machines. This information can be collated at a single 'control' node and the total densities broadcast to each 'worker' node.

The initial model was created using the agent-based simulation toolkit Repast (`http://repast.sourceforge.net`). The toolkit was implemented in parallel by running the Repast interface on the control node (including the GUI etc.), while the rest of the model code is run independently on worker-nodes, synchronised by the control node using message passing (Parry, 2006; Parry et al., 2006a). Agents are established on the worker nodes, coordinated by the control node. In coding the parallel version of the model software, the same code is placed on each processor but different sections of code are run dependent on the node ID. The code on the control node controls the model input, output and program flow, using the standard Repast methods of 'setup', 'buildModel', 'preStep', 'step' and 'postStep' to structure the code and to initiate the simulation steps (figure 1). For example, when the method preStep is run, the control node (node zero) is programmed to send out messages to the other nodes to invoke agent methods associated with preStep (the model is strongly synchronised). The updated agents pass density information back to the control node when needed (figure 1). It was expected that the speed of the simulation would increase with the number of nodes used, compensating for any minor time delay caused by the timing control of the simulation from

the control node. A similar strategy was employed by Lorek and Sonnenschein (1995) for a non-Repast simulation, which was found to increase simulation speed as well as enable the size of the simulation to increase.

3.1.1 Message passing

The agent model was parallelised using a Message-Passing Interface (MPI) for Java, MPIJava (`http://www.hpjava.org`), run on a 30-node distributed memory parallel computer known as a Beowulf cluster. Message passing (MP) is the principle manner by which Beowulf clusters are linked. MPIJava uses the open-source native MPI 'LAM' (`http://www.lam-mpi.org/`). Further details on the methods used to incorporate the MPI into the model are given in Parry (2006); Parry et al. (2006a). The particular Beowulf cluster used for the simulations presented here was a dedicated cluster of thirty machines (nodes), where each node has dual 2.66 GHz Intel Xeon processors with 1280 MB of DDR memory and 40GB 7200rpm internal IDE disks running over a switched GB network. Although the results presented in this paper refer to simulations conducted on a dedicated Beowulf cluster, the principles of parallelising the model for a multi-core machine or a non-dedicated cluster would be very similar. The model presented here has since been adapted to run on an Intranet cluster of non-dedicated PCs and on a multi-core processor system without re-coding the parallelisation, only altering the MPI commands in the code to work with a customised MPI. Non-dedicated clusters of machines with mixed specifications may however introduce problems of network unreliability and performance bottlenecking on the slowest machines. Such issues are explored further in Parry (in press).

3.1.2 Data mapping and load balancing

Data must be evenly distributed between the nodes, known as 'load balancing' (Pacheco, 1997). In this model, a balanced load was calculated using a form of 'block mapping' (Pacheco, 1997). Agents were split evenly across the system, whilst each node contained full environmental information. The agents remained on the same node throughout the simulation, thus maintaining a balanced load while being able to roam the environment. For all the parallel simulations, it was found that the maximum memory used by each node was so similar that 95% confidence limits derived from the standard error evaluated to ± 0.00 in all cases. This shows that the distribution of individuals across the worker nodes was highly efficient, and the load very well balanced.

4 Super-individuals

4.1 Implementation

The super-individual approach to modelling large populations on an individual basis was proposed by Scheffer et al. (1995), comparable to the earlier 'generalised individuals' of Metz and de Roos (1992). A super-individual approach 'allows zooming from a real individual-by-individual model to a cohort representation or ultimately an all-animals-are-equal view without changing the model formulation' (Scheffer et al., 1995: pp. 161). The simple idea is that individuals in a population can be grouped together into 'super-individuals', thus reducing the number of objects to simulate and therefore reducing the memory and processing power required (figure 2). For populations such as aphids where there are high reproductive and mortality rates leading to large

juvenile populations, this approach can be very useful (Grimm and Railsback, 2005). It is possible to use the approach to test the effects of grouping individuals, and also to examine the degree to which individual behaviour explains the observed phenomena. A similar approach is used in physical models such as the lattice models of fluid dynamics, particle modelling and Lagrangian modelling (e.g. Woods and Barkmann, 1994).

4.1.1 Combining individuals into a single super-individual

Although Scheffer et al. (1995) state that no changes to the model formulation are required for the super-individual approach, there are some significant changes to the model structure that potentially influence the model results. To convert the individual-based model to super-individuals, individuals originating at the same spatial location (cell) were split by initial age and morphology (whether they have wings or not) into super-individuals. Each super-individual represented a fixed number of individuals throughout the course of the simulation.

4.1.2 Adding individual immigrants to super-individuals

Initial immigrants were added as super-individuals of the same scale factor and, as in the unmodified model, these were of uniform age and morphology (adult alates (i.e. winged)).

4.1.3 Mortality of individuals/super-individuals

Estimating the mortality of super-individuals can be done in a number of ways, all of which are prone to error. The three main approaches are given by

Grimm and Railsback (2005: pp. 267) (figure 3):

N = the number of individuals represented by the super-individual (i.e. the scale factor). N_0 = the number of individuals represented by the super-individual at the start of the simulation.

(1) The number of super-individuals remains constant, and mortality reduces N.
(2) N is kept relatively constant, by mortality reducing N until super-individuals are recombined when N falls below $N_0/2$.
(3) Assume that an entire super-individual dies when subject to mortality.

Both approaches 1 and 2 require dynamic updating of the number of individuals represented by the super-individual, but in this way they do maintain more of the original variability of the model. However, significant errors, particularly spatial errors, would be introduced as individuals are re-grouped, and the process would be computationally intensive. Reducing the number of super-individuals in approaches 2 and 3 has computational advantages (the number of super-individuals to iterate is minimised and individual variability is less important so calculations are less complex).

Approach 3 was chosen: super-individuals are subject to the same probability of mortality as individuals and when the super-individual dies all individuals represented by the super-individual die. This approach was chosen because the variability between individuals of the model (particularly age) meant that approach 2 (recombining individuals) was problematic. Approach 1 (maintaining a constant number of super-individuals) would also be problematic to implement as the constant updating and variability of N would be computationally intensive, particularly as the density of individuals is important to a

number of model processes. Approach 3 was therefore considered to be the most computationally efficient, although the least biologically realistic as it suggests that mortality affects equally a group of aphids of uniform age and morph in a particular cell (discretization of mortality). Another potential issue with approach 3 is that it may require a lower value of N than the other approaches to avoid excessive discretization of mortality. This paper assesses whether this is the case.

4.1.4 Changes to the model structure

The construction of a super-individual simulation involved very little alteration of the model structure (for details of this structure see Parry et al., 2006b, 2004). A variable was added to record the number of individuals all super-individuals actually represent. Equations that were dependent on density (such as morphology determination) were altered so that the density values were related to the real number of individuals in the simulation, not the number of super-individuals (see equation 1). This was because the proportion of alates produced is in relation to the density of individuals.

Morph determination is represented by the equation:

$$ALPROP = \frac{0.002 + 0.991}{(1 + EXP(-0.076 \times (DENSITY - 67.416)))} \quad (1)$$

where ALPROP = the proportion of newly laid nymphs that will become alate and DENSITY = the total number of individual aphids per plant.

5 Evaluation

The parallel version of the model produced extremely similar results to the non-parallel model, as expected (no changes were made to the model structure, only to the software). Variability between the original model and the parallel model was only due to the model's stochasticity. However, the super-individual model did alter the model structure, therefore some variation was expected in the output between the super-individual model and the original model. This variability is presented first below, then a comparison is made of the improvement in performance in terms of model speed and memory use for both the parallel and super-individual approach in relation to the original model.

5.1 Super-individual temporal and spatial replication of the individual-based simulation

Movement of super-individuals followed the same rules as that of individuals, however this produced spatial clustering of the populations. To test the super-individual model, populations of 100, 1,000, 10,000 and 100,000 and 500,000 individuals were represented by varying numbers of super-individuals (Table 2). Results are compared to the original individual-based model, both temporally and spatially, in the following sections. Results from simulations with 10,000 individuals are given in more detail as an example to demonstrate the effects of combining individuals.

5.1.1 Temporal

Overall, for simulations of fewer than 10,000 individuals the super-individual simulations produced population densities that were much lower than the individual-based model equivalent (figure 4). For 10,000 individuals, densities only become significantly lower at the second population peak, and the super-individual simulations also reach this peak earlier. This can be related to the spatial results (below), where it is only after this point in time that it is evident that differing spatial distributions and densities are beginning to emerge. The only case where the super-individual simulation falls within the 95% confidence limits of the original model for the duration of the simulation period is the simulation of 10,000 individuals with 1,000 super-individuals (scale factor 10), figure 4. The percentage error between the temporal results for all the super-individual simulations and the individual-based simulations is shown graphically in figure 5. This confirms that super-individual simulations of 10,000 aphids and above with low scale factors may be acceptable. This also shows that when a large number of individuals are represented by very few super-individuals (in this case 10 super-individuals) the error is greatest. Excessive discretization of mortality is therefore evident (suggested in section 4.1), resulting in a need to reduce the scale factor for results to better represent the individual-based model.

5.1.2 Spatial

Clustering is evident in the spatial distribution. The super-individuals are contained in fewer cells, closer to the origin, than the individual-based simulation. This is illustrated for 10,000 individuals by figure 6. The distribution

better replicates the unmodified model when the number of super-individuals is maximised and the individuals they represent minimised, due to the assumption that when mortality occurs, the whole super-individual dies. Only when the number of individuals within the super-individual (N) is minimised in a large population of super-individuals can this be overcome (Grimm and Railsback, 2005). However, even when this is the case, for 10,000 individuals with 1,000 super-individuals (scale factor 10) (figure 4) this still does not produce a similar spatial distribution pattern, despite giving a satisfactory temporal result. This suggests that errors in spatial distribution may be hidden in super-individual models validated temporally. The super-individual patterns are in fact most comparable to the patterns of the individuals for the same number, e.g. 10 super-individuals compares well with the distribution of 10 individuals, the difference is the density at each cell. This is the expected result when the local redistribution of (super)individuals is the main process determining the spatial distribution, despite density affecting morphology.

5.2 Speed

Super-individuals always improve the model speed (figure 7). The speed improvement is enormous for the largest simulations, where 500,000 individuals simulated with super-individuals using a scale factor of 100,000 increases the model speed by over 500 times the original speed. However, it was shown above that only large simulations with a low scale factor (10-100) may benefit from the super-individual approach, thus for these scale factors an improvement in model speed of approximately 10,000-30,000% (100-300 times) the original speed would result for simulations of 100,000 to 500,000 individuals.

Adding more processors does not necessarily increase the model speed. Figure 7 shows that for simulations run on two nodes (one control node, one worker node) the simulation takes longer to run in parallel compared to the non-parallel model. Message passing time delay and the modified structure of the code are responsible. As the number of nodes used increases, the speed improvement depends on the number of agents simulated. The largest improvement in comparison to the non-parallel model is when more than 500,000 agents are run across twenty-five nodes, although the parallel model is slower by comparison for lower numbers of individuals. However, when only five nodes are used the relationship is more complex: for 100,000 agents five nodes are faster than the non-parallel model, but for 500,000 the non-parallel model is faster. This is perhaps due to the balance between communication time increasing as the number of nodes increases versus the decrease in time expected by increasing the number of nodes. Overall, these results seem to suggest that when memory is sufficient on a single processor, it is unlikely to ever be efficient to parallelise the code.

5.3 Memory usage

Super-individuals always reduce the memory requirements of the simulation (figure 8). The memory requirements for a simulation of super-individuals has a similar memory requirement to that of an individual-based simulation with the same number of agents. For simulations of 100,000 agents this can reduce the memory requirement to less than 10% of the memory required for the individual-based simulation with a scale factor of 10,000, and for simulations of 500,000 agents this may be reduced to around 1% with the same scale

factor.

The mean maximum memory usage by each worker node in the parallel simulations is significantly lower than the non-parallel model, for simulations using more than two nodes (figure 8). The two node simulation used more memory on the worker node than the non-parallel model when the simulation had 100,000 agents or above. This is probably due to the memory saved due to the separation of the GUI onto the control node being over-ridden by the slight additional memory requirements introduced by the density calculations. However, when 5 and 25 nodes are used, the memory requirements on each node are very much reduced, below that of the super-individual approach in some cases. The super-individual approach uses the least memory for 500,000 individuals, apart from when only a scale factor of 10 is used (then the 25 node parallel simulation is more memory efficient).

6 Discussion

The parallel model produced identical results to the initial model, as this modifies only the model software and not the model itself. However, the super-individual approach did not produce identical results to the initial model, especially when assessed spatially. The similarity between the super-individual results and the initial, unmodified model varied according the number of real individuals that the super-individual was representing, and the number of individuals simulated. The super-individual approach can only be considered in situations where the number of individuals is high and the number of real individuals represented by each super-individual is low (i.e. a low scale factor).

For the super-individual approach, within-cell density peaks vary temporally between simulations run with different super-individual sizes. This is due to the differences in emigration and movement patterns as a result of the size of the super-individuals, as well as the method used to represent mortality. Excessive discretization of mortality is evident as it is assumed that an entire super-individual dies when subject to mortality. Further assessment of the model (Parry, 2006) shows that regionally, the total population density is similar between the different super-individual configurations and the unmodified model, but as shown in figure 6 there is a clear difference in the dispersal patterns. Overall, the evidence indicates that the variability is such that the super-individual approach is not suitable for the spatially-explicit simulation of the aphid model, as presented here. Indeed, although the aphid model is more strongly density dependent than most ecological models, most are to some degree density dependent, rendering super-individual models problematic for spatially informative work. Modifications to the approach could make it a possibility for future work. Experimentation with the other rules for super-individual mortality suggested in section 4.1 would be a first step. Other possible modifications include:

(1) Weighted kernels around a central 'super-individual', so that a more realistic dispersal pattern is achieved.
(2) Relocation of a percentage of the super-individuals from a cell, without actual population redistribution.
(3) Cell population model with individual migration.

However, re-distribution of individuals could significantly increase run-time, adds complexity to the simulations and may take more memory than the individual-based approach. This would also rely on a non-naturalistic model

of dispersion. Most movement in the model is a short distance each day, so there will be constant shifting from super-individual to individual or creation of dispersal kernels.

Further investigation may also indicate that spatial heterogeneity may have a strong impact on the accuracy of the super-individual approach. The simulations presented here were conducted in a neutral landscape, but if the model were run in a heterogeneous landscape the interactions of the individuals with the landscape may create model feedback that might further affect the accuracy of the super-individual results, both spatially and temporally.

Although the parallel solution appears to be more appropriate, in order to ensure it is optimised for agent simulations the balance between the advantage of increasing the memory availability and the cost of communication between nodes must be assessed in relation to the number of individuals simulated. When the number of individuals is low, parallel simulations take longer (figure 7) and are less efficient (figure 8) than a non-parallel model run on a single node. Increasing the number of nodes can reduce the demands on each individual node, but time to communicate between processors may also be increased (depending on the way in which the model is parallelised).

For the model presented here, estimates of the maximum number of agents that can be simulated for varying numbers of nodes (table 3) and the maximum number of agents for a given super-individual scale factor (table 3) were calculated with 1GB RAM, based upon information in figure 8. For the parallel version, when only two nodes are used the non-parallel simulation is estimated to have a higher maximum agent capacity per worker node, because space is not being used by the message passing code. However, from five nodes and

up there is a higher maximum agent capacity for the parallel version than the non-parallel model. The maximum agent capacity of 25 nodes is very high, at nearly 100 million. This is approximately ten times the number of agents that can be run within a reasonable time in the individual-based model. For low numbers of nodes run times are huge: for two and five nodes the run times are estimated to be 13 days and 47 days respectively. This would be expected to increase with the complexity of the simulation. For any given model there will be a threshold below which parallelisation is not efficient. Investigating this threshold is likely to be a matter of iterative development as demonstrated here, starting with a stripped-down model containing just the basic message-passing elements. Passing of agents between nodes is processor intensive, and therefore should be minimised. In this model, only the environment object and information on the number of agents to create on each node are passed from the control node to each of the nodes, and only density information is returned to the control node for redistribution and display (see figure 1).

For the super-individuals (table 4) the relationship between the maximum number of individuals that can be simulated and the scale factor is very simple. The run time remains the same, as the maximum number of super-individuals or individuals is constant. The maximum number of individuals that can be simulated by this approach therefore depends purely on the scale factor used. It would therefore not be unrealistic to assume this approach may potentially enable the simulation of very large numbers of individuals indeed, in excess of $7E^{11}$ if a scale factor of 1,000,000 is used, for example (assuming this scale factor may be acceptable).

7 Conclusion

In order to address the limitations on the number of agents imposed by the processing power and memory available, two solutions have been tested: parallel processing and super-individuals. The parallel approach involved significant recoding of the model software, but no changes were made to the model structure and performance increased significantly, enabling the simulation of at least ten times more agents. The parallel model produced results that are comparable to the initial, non-parallel model, leading to the conclusion that for the simulation of very large populations the parallel model is a good solution.

Although initially far simpler to implement, the super-individual approach is inappropriate for spatial simulations in the form presented here. However, it may be possible to use this approach if the model were to be significantly altered, by using another approach to simulate super-individual mortality, or a super-individual model merged with an individual-based model, where dispersal can be simulated by switching from a super-individual to an individual-based model when necessary. There is a high risk that the complexity of switching between model or implementing retrospective re-distribution of agents could introduce significant error and put high demands on the processor and/or memory, which are already limited. Overall, the results presented here indicate that the super-individual approach is inappropriate to the spatially-explicit simulation of populations with density-dependent functions or interactive agents, unless individual variability is better taken into account. If this can be achieved satisfactorily, it has been demonstrated that the super-individual approach may lead to very large reductions in computational demands.

8 Acknowledgements

This work is part of a PhD funded by the Central Science Laboratory (Defra Seedcorn). Thanks to Daniel Parry for his advice on Java programming and Phil Northing and two anonymous reviewers for their comments on the manuscript. The authors are members of the Multi-Agent Systems and Simulation Research Group, School of Geography, University of Leeds.

References

Abbott, C. A., Berry, M. W., Comiskey, E. J., Gross, L. J., Luh, H.-K., 1997. Parallel individual-based modeling of everglades deer ecology. IEEE Computational Science and Engineering 4 (4), 60–78.

DeAngelis, D., Gross, L. (Eds.), 1992. Individual-based models and approaches in ecology: populations, communities and ecosystems. Routledge, Chapman and Hall, New York, 544 pp.

Dupuis, A., Chopard, B., 2001. Parallel simulation of traffic in Geneva using cellular automata. Nova Science Publishers, Inc., Commack, NY, USA.

Gilbert, N., Troitzsch, K. G., 1999. Simulation for the Social Scientist. Open University Press, Buckingham, 288 pp.

Grimm, V., 1999. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? Ecological Modelling 115, 129–148.

Grimm, V., Railsback, S. F., 2005. Individual-based Modeling and Ecology. Princeton Series in Theoretical and Computational Biology. Princeton University Press, Princeton, 480 pp.

Grimm, V., Wyszomirski, T., Aikman, D., Uchmański, J., 1999. Individual-

based modelling and ecological theory: synthesis of a workshop. Ecological Modelling 115, 275–282.

Haefner, J. W., 1992. Parallel computers and individual-based models: An overview. In: DeAngelis, D. L., Gross, L. J. (Eds.), Individual-based models and approaches in ecology: populations, communities and ecosystems. Routledge, Chapman and Hall, New York, Ch. 7, pp. 126–164.

Hosseini, P. R., 2006. Pattern formation and individual-based models: The importance of understanding individual-based movement. Ecological Modelling 194 (4), 357–371.

Huston, M., DeAngelis, D., Post, W., 1988. New computer models unify ecological theory. BioScience 38 (10), 682–691.

Judson, O., 1994. The rise of the individual-based model in ecology. Trends in Ecology and Evolution 9, 9–14.

Lefley, M., McKew, I. D., 2004. Can a parallel agent approach to genetic algorithms reduce search times. In: Loffi, A., Garobaldi, J. (Eds.), Applications and Science in Soft Computing. New York: Cambridge University Press, pp. 69–74.

Lorek, H., Sonnenschein, M., 1995. Using parallel computers to simulate individual-oriented models in ecology: a case study. In: Proceedings: ESM '95 European Simulation Multiconference, Prague, June 1995.

Metz, J. A. J., de Roos, A. M., 1992. The role of physiologically structured popualation models within a general individual based model perspective. In: DeAngelis, D. L., Gross, L. J. (Eds.), Individual Based Models and Approaches in Ecology: Concepts and Models. Routledge, Chapman and Hall, New York, pp. 88–111.

Pacheco, P. S., 1997. Parallel Programming with MPI. Morgan Kauffman Pub-

lishers, San Francisco, CA.

Parrott, L., Kok, R., 2000. Incorporating complexity in ecosystem modelling. Complexity International 7, 1–19.
URL http://www.csu.edu.au/ci/

Parry, H. R., in press. Agent Based Modelling, Large Scale Simulations. In: Meyers, R. A. (Ed.) Encyclopedia of Complexity and System Science. Springer-Verlag, GmbH Berlin Heidelberg.

Parry, H. R., October 2006. Effects of land management upon species population dynamics: A spatially explicit, individual-based model. Ph.D. thesis, University of Leeds.

Parry, H. R., Evans, A. J., Heppenstall, A. J., 2006a. Millions of agents: Parallel simulations with the Repast agent-based toolkit. In: Trappl, R. (Ed.), Cybernetics and Systems 2006, Proceedings of the 18th European Meeting on Cybernetics and Systems Research. Austrian Society for Cybernetic Studies, Vienna.
URL http://www.lintar.disco.unimib.it/ABModSim/

Parry, H. R., Evans, A. J., Morgan, D., 2006b. Aphid population response to agricultural landscape change: A spatially explicit, individual-based model. Ecological Modelling 199, 451–463.

Parry, H. R., Evans, A. J., Morgan, D., June 2004. Aphid population dynamics in agricultural landscapes: An agent-based simulation model. In: Pahl-Wostl, C., Schmidt, S., Jakeman, T. (Eds.), iEMSs 2004 International Congress:"Complexity and Integrated Resources Management". International Environmental Modelling and Software Society, Osnabrueck, Germany.

Scheffer, M., Baveco, J. M., DeAngelis, D. L., Rose, K. A., van Nes, E. H., 1995. Super-individuals: a simple solution for modelling large populations

on an individual basis. Ecological Modelling 80, 161–170.

Schmeeckle, M. W., Nelson, J. M., 2003. Direct numerical simulation of bed-load transport using a local, dynamic boundary condition. Sedimentology 50, 279–301.

Uchmański, J., Grimm, V., 1996. Individual-based modelling in ecology: what makes the difference? Trends in Ecology and Evolution 11 (10), 437–441.

Verzelen, N., Picard, N., Gourlet-Fleury, S., 2006. Approximating spatial interactions in a model of forest dynamics as a means of understanding spatial patterns. Ecological Complexity 3 (3), 209–218.

Wang, D., Gross, L., Carr, E., Berry, M., 2004. Design and implementation of a parallel fish model for South Florida. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04).
URL `http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/09/205690282c.pdf`

Woods, J. D., Barkmann, W., 1994. Simulating plankton ecosystem by the Lagrangian Ensemble Method. Philosophical Transactions of the Royal Society London B 343, 27–31.

List of Tables

Solution	**Pros**	**Cons**
Invest in an extremely powerful computer.	No changes to model code or structure.	High cost.
Invest in an extremely powerful computer network and reprogram the model in parallel.	Makes available high levels of memory and processing power. Model remains the same.	High cost. Advanced computing skills required for restructuring of model software.
Super-individuals	Relatively simple solution. Little change to model formulation.	Reprogramming of model and altered structure and interactions. Untested in spatial context.
Reduce the number of individuals in order for model to run.	No reprogramming of model.	Unrealistic population. Alters model behaviour.
Revert to a population based modelling approach.	Could potentially handle any number of individuals.	Lose insights from IBM. Potentially unsuitable for the particular research questions of the study. Construction of entirely new model.

Table 1
Possible solutions when faced with a large number of individuals to model.

Number of individuals	**Number of super-individuals**	**Number of individuals represented by each super-individual ('scale factor')**
100	10	10
1,000	10	100
	100	10
10,000	10	1,000
	100	100
	1,000	10
100,000	10	10,000
	100	1,000
	1,000	100
	10,000	10
500,000	50	10,000
	500	1,000
	5,000	100
	50,000	10

Table 2
Table to show the construction of the tested super-individuals: individuals, super-individuals and the number of individuals each super-individual represents

Number of Nodes	**Maximum number of agents**	**Estimated run time of simulation (seconds)**
IBM	$7.49E^6$	$1.12E^6$
2	$3.33E^6$	$1.11E^6$
5	$1.43E^7$	$4.06E^6$
25	$1.00E^8$	$2.20E^4$

Table 3
The maximum number of agents that can be simulated for 2, 5 and 25 processors, and the associated estimated run time

Scale factor	**Maximum number of individuals**	**Maximum number of super-individuals (agents)**	**Estimated run time of simulation (seconds)**
IBM	$7.49E^6$	-	$1.12E^6$
10	$7.49E^7$	$7.49E^6$	$1.12E^6$
100	$7.49E^8$	$7.49E^6$	$1.12E^6$
1,000	$7.49E^9$	$7.49E^6$	$1.12E^6$
10,000	$7.49E^{10}$	$7.49E^6$	$1.12E^6$
100,000	$7.49E^{11}$	$7.49E^6$	$1.12E^6$

Table 4
The maximum number of agents that can be simulated when the super-individual scale factor (number of individuals represented by each super-individual) is 10, 100, 1,000, 10,000 and 100,000 and the associated estimated run time.

List of Figures

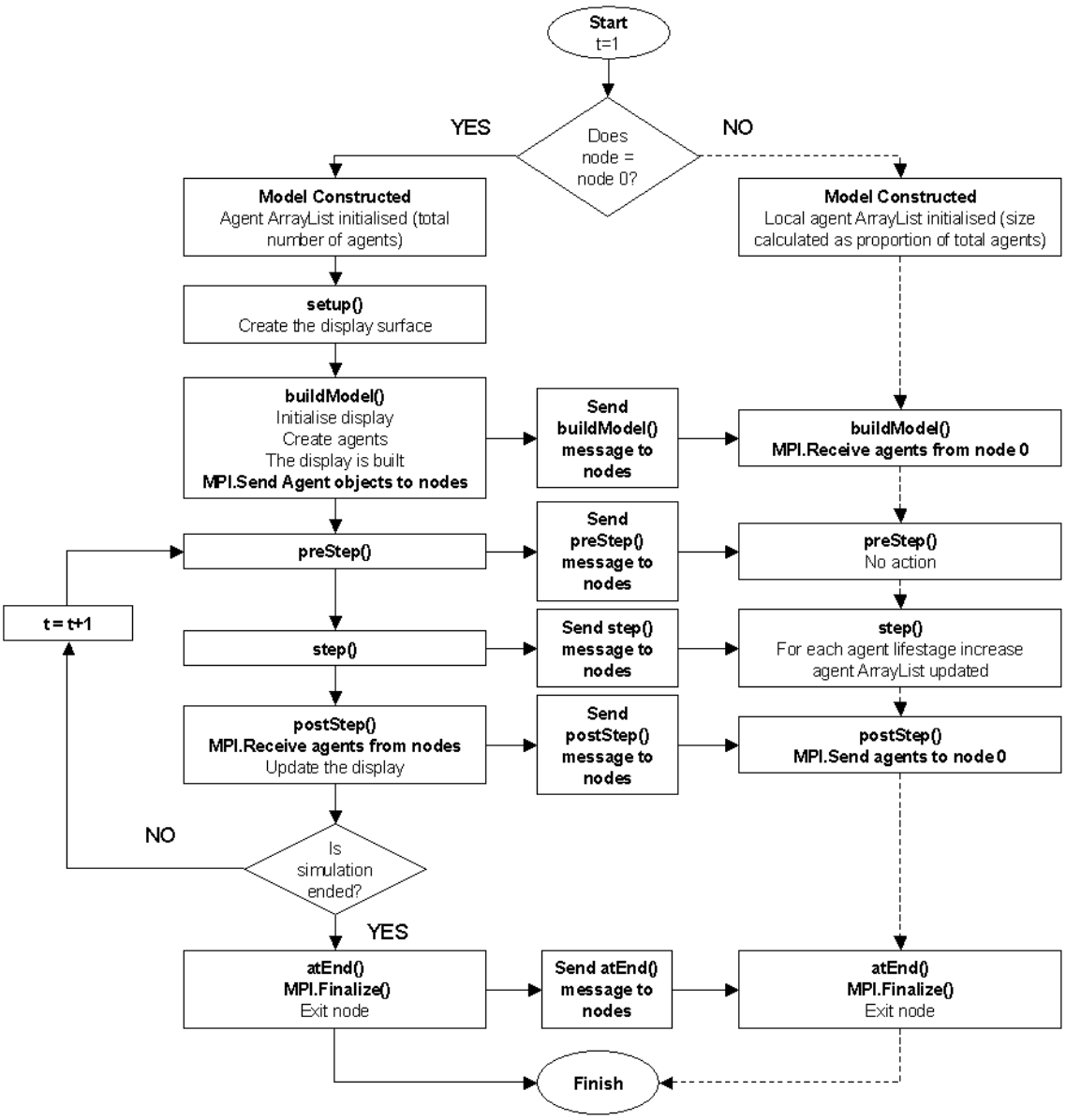


Fig. 1. Flow chart illustrating the operation of rules at each stage of a model run for a simple Repast model, and the role of message passing to control the program flow between node 0 and the other nodes.

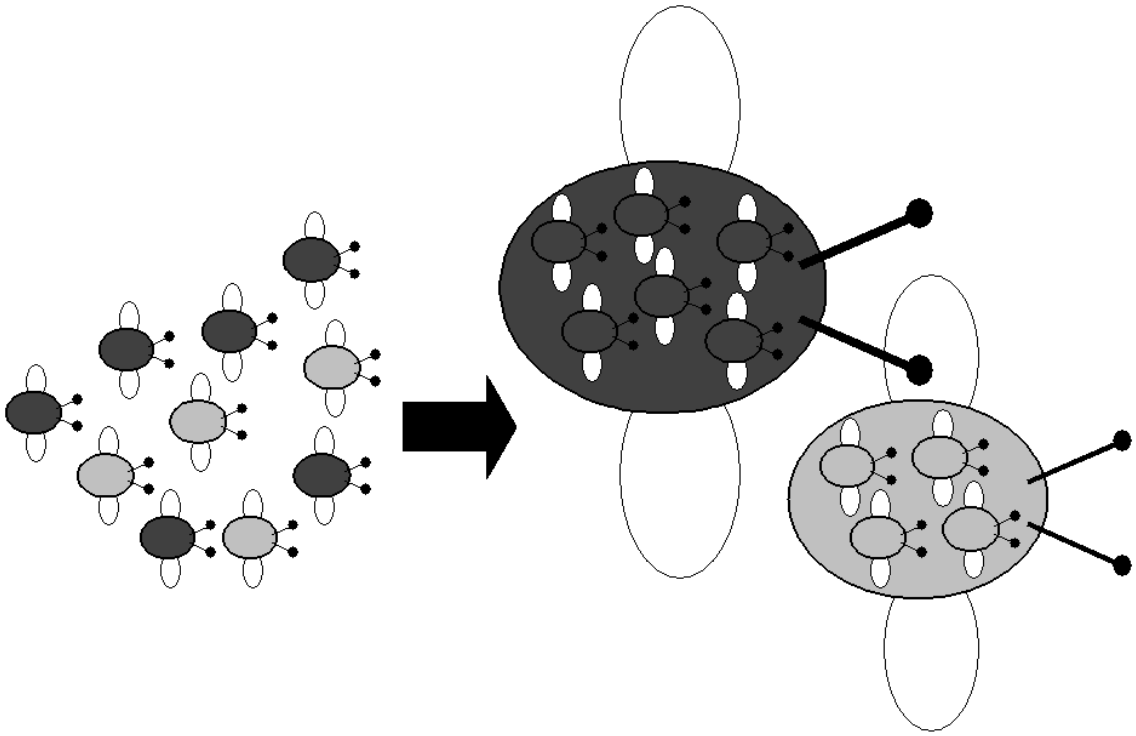


Fig. 2. Super-individuals: Grouping of individuals into single objects that represent the collective

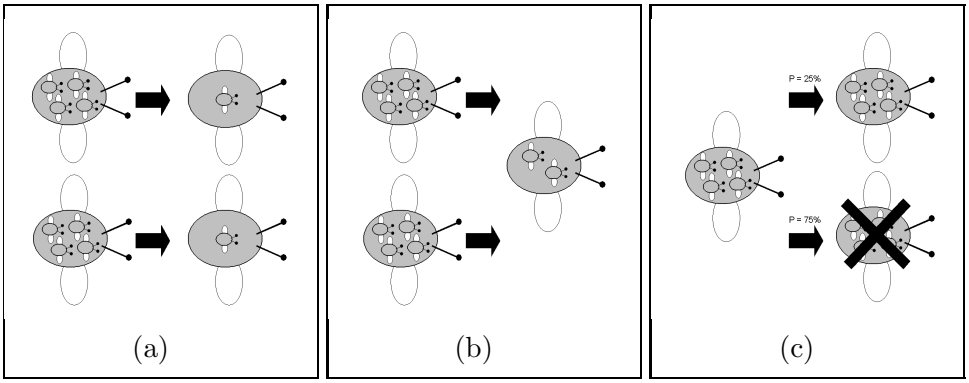


Fig. 3. The three main approaches to estimating the mortality of super-individuals: (a) The number of super-individuals remains constant, and mortality reduces the number of individuals (N) represented by the super-individual. (b) N is kept relatively constant, by mortality reducing N then super-individuals are recombined when N falls below $N_0/2$. (c) Assume that an entire super-individual dies when subject to mortality.

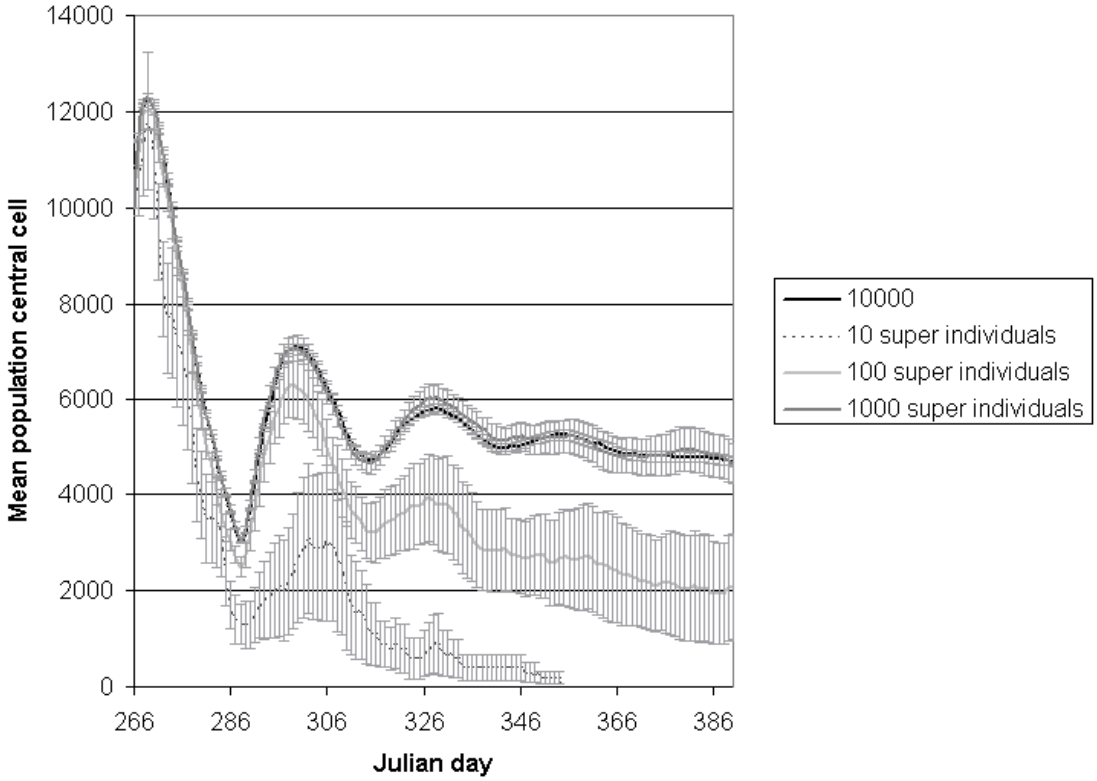


Fig. 4. 10,000 individuals: comparison between individual-based simulation, 1,000 super-individual simulation (each represents 10 individuals), 100 super-individual simulation (each represents 100 individuals) and 10 super-individual simulation (each represents 1,000 individuals), showing 95% confidence limits derived from the standard error.

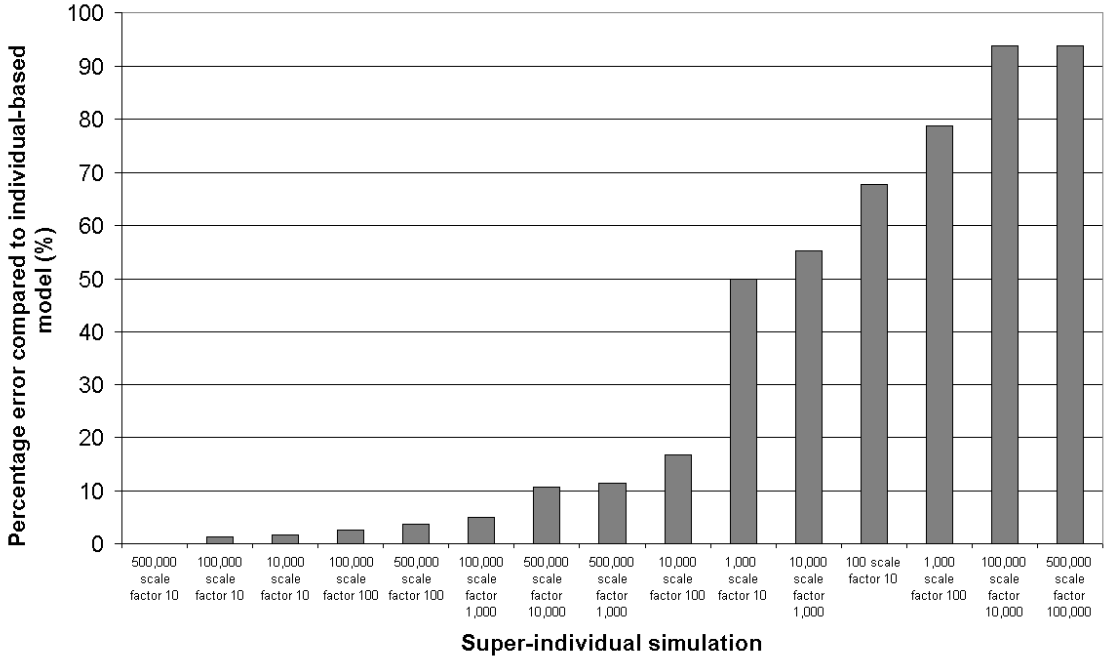
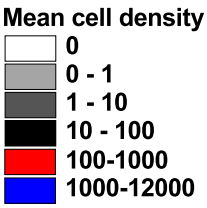
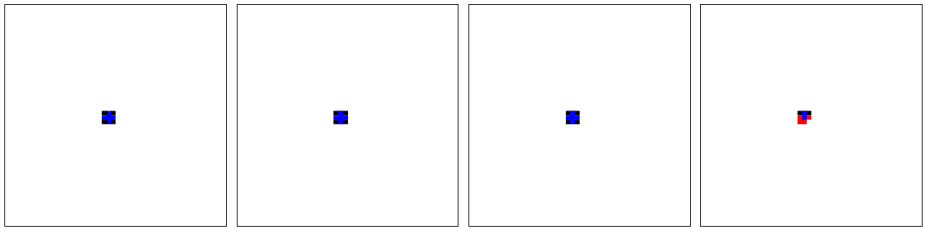
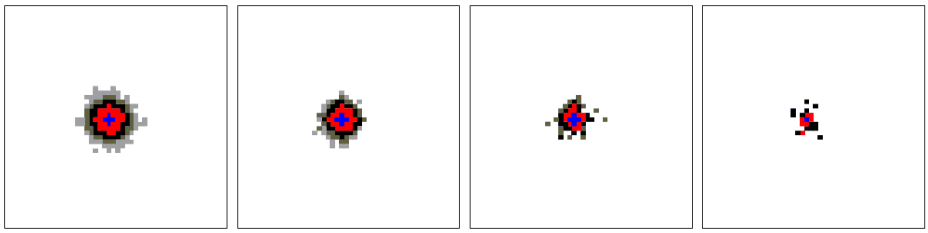


Fig. 5. Comparison of the mean (absolute) percentage error between the super-individual simulations and the individual-based simulation, at t = 40.

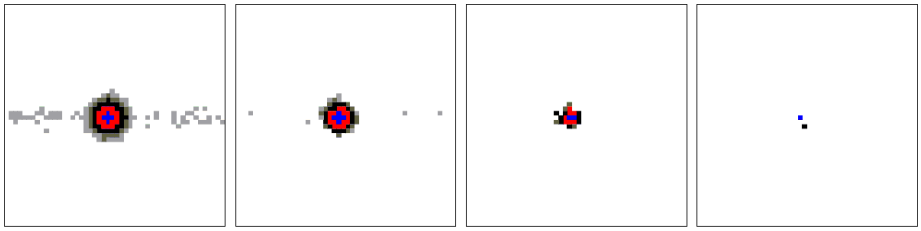




(a) 10,000 individuals, density at 2 days: (l-r) Individual-based simulation, super-individual simulation scale factor 10, 100 and 1,000



(b) 10,000 individuals, density at 20 days: (l-r) Individual-based simulation, super-individual simulation scale factor 10, 100 and 1,000



(c) 10,000 individuals, density at 40 days: (l-r) Individual-based simulation, super-individual simulation scale factor 10, 100 and 1,000

Fig. 6. Spatial density distributions for individual-based versus super-individual simulations (10,000 aphids) at (a) 2 days (b) 20 days and (c) 40 days. The distribution further from the central cell is influenced by the constant westerly wind direction to result in a linear movement pattern.

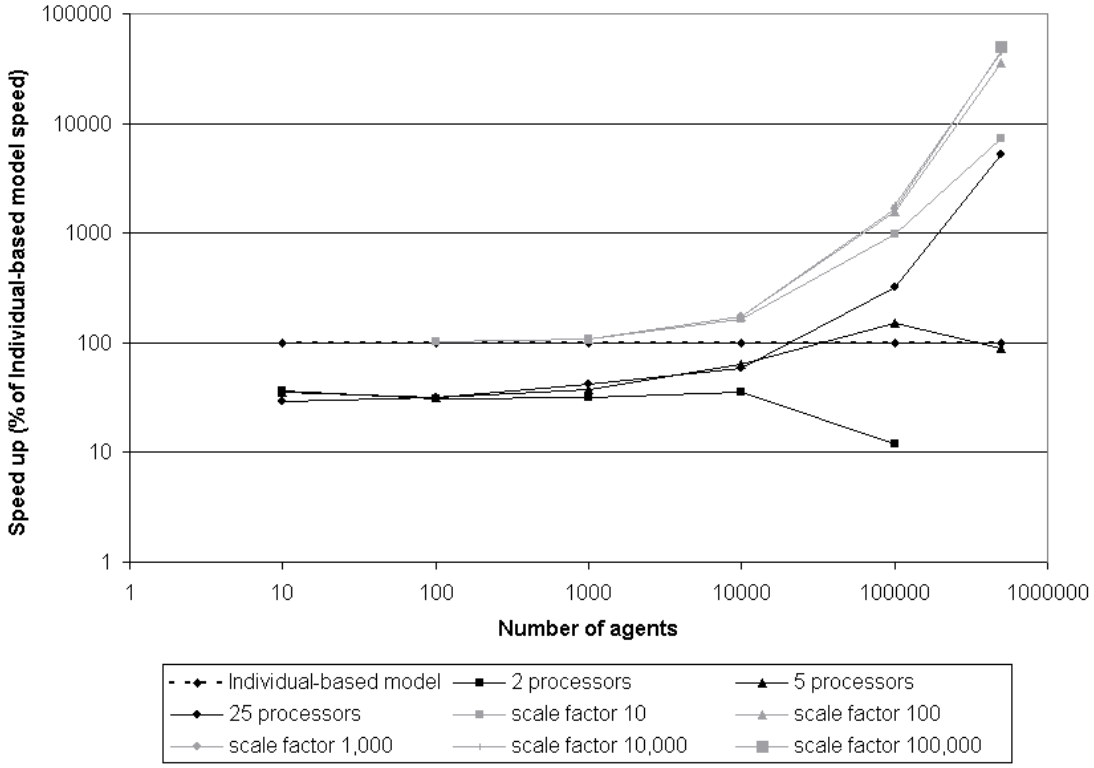


Fig. 7. Plot of the percentage speed up from the individual-based (non-parallel) model against number of agents modelled: comparison between parallel simulations using 2, 5 and 25 nodes and super-individuals of scale factor 10, 100, 1,000, 10,000, 100,000 and 500,000

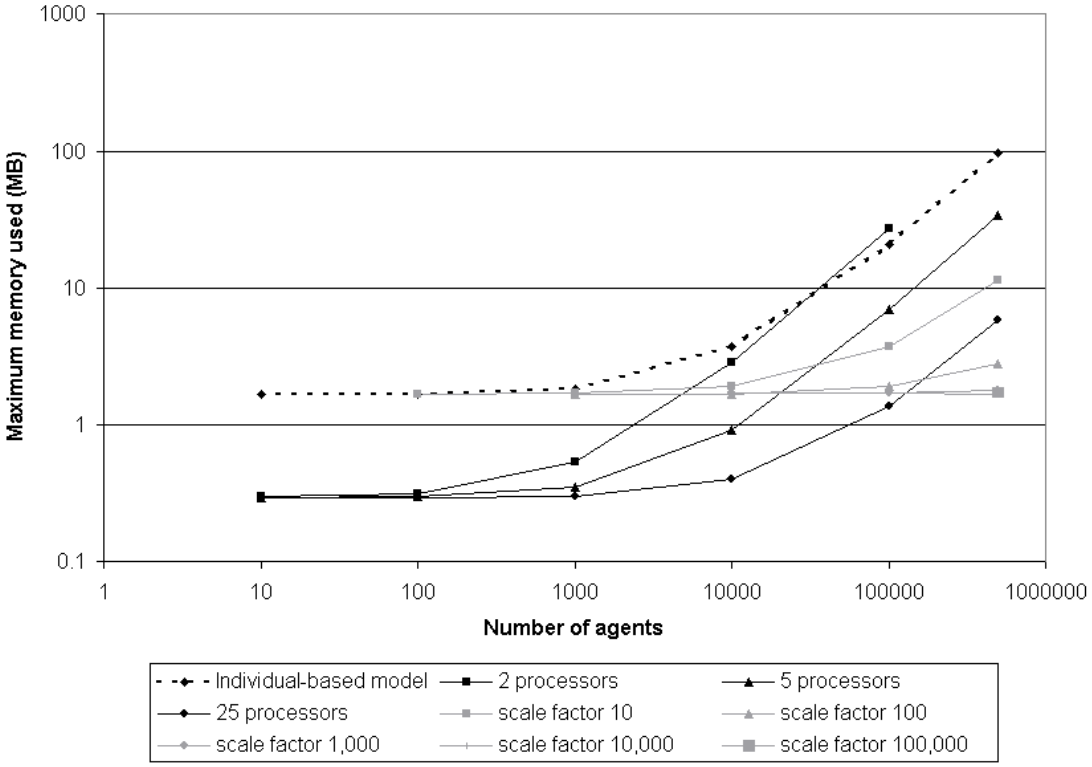


Fig. 8. Plot of the mean maximum memory used in a simulation run against number of agents for the model, for different numbers of nodes (memory per node) and scale factors for super-individuals