CENTRE FOR ADVANCED SPATIAL ANALYSIS
Working Paper Series

Paper 15

THE DYNAMICS OF URBAN SPRAWL

Michael Batty
Yichun Xie
Zhanli Sun

Centre for Advanced Spatial Analysis
University College London
1-19 Torrington Place
Gower Street
London WC1E 6BT

Tel: +44 (0) 171 391 1782
Fax: +44 (0) 171 813 2843
Email: casa@ucl.ac.uk
http://www.casa.ucl.ac.uk

http://www.casa.ucl.ac.uk/sprawl.pdf

Date: November 1999

ISSN: 1467-1298

ABSTRACT

This paper introduces a framework for understanding the dynamics of urban growth, particularly the continuing problem of urban sprawl. The models we present are based on transitions from vacant land to established development. We propose that the essential mechanism of transition is analogous to the way an epidemic is generated within a susceptible population, with waves of development being generated from the conversion of available land to new development and redevelopment through the aging process. We first outline the standard aggregate model in differential equation form, showing how different variants (including logistic, exponential, predator-prey models) can be derived for various urban growth situations. We then generalize the model to a spatial system and show how sprawl can be conceived as a process of both interaction/reaction and diffusion. We operationalize the model as a cellular automata (CA) which implies that diffusion is entirely local, and we then illustrate how waves of development and redevelopment characterizing both sprawl and aging of the existing urban stock, can be simulated. Finally we show how the model can be adapted to a real urban situation - the Ann Arbor area in Eastern Michigan - where we demonstrate how waves of development are absorbed and modified by particular historical contingencies associated with the pre-existing urban structure.

Acknowledgments

This project is part funded by the US National Science Foundation (NSF Grant 9752086) and by the UK Office of Science and Technology Foresight Challenge (EPSRC GR/L54950).

Affiliations

Batty: Centre for Advanced Spatial Analysis (CASA), University College London, 1-19 Torrington Place, London WC1E 6BT, UK (m.batty@ucl.ac.uk); *Xie*: Center for Environmental Information Technology and Application (CEITA), Eastern Michigan University, Ypsilanti, Michigan MI 48197, USA (gis_xie@online.emich.edu): *Sun*: Key Laboratory for Resources & Environment Information Systems (LREIS), Institute of Geography, Chinese Academy of Sciences, Building 917, Datun Road, Anwai, Beijing 100101, PRC (sunzl@lreis.ac.cn)

1 Defining Sprawl

Contemporary urban growth consists of three interrelated problems of spatial dynamics: the decline of central or core cities which usually mark the historical origins of growth, the emergence of edge cities which both compete with and complement the functions of the core, and the rapid suburbanization of the periphery of cities - core and edge - which represent the spatially most extensive indicator of such growth. Our understanding of these growth processes is rudimentary, notwithstanding at least 50 years of sustained effort in their analysis. Our abilities to "control and manage" such growth or "sprawl" as it is colloquially and often pejoratively referred to, is virtually non-existent despite occasional but short lived successes through planning instruments such as green belts. The suburbanization of cities and methods for the control of such growth go back to the origins of cities themselves. Urban history reveals a succession of instruments used to separate the growing city from its suburbs. Documented examples refer to Ur in Sumeria, ancient Rome, to Elizabethan London, where edicts were in place to ensure the quality of life in the core city by restricting access and overbuilding (Morris, 1979). However the concept of suburb has changed through history. Jackson (1985) sums this up quite cogently when he says: “... the suburb as a residential place is as old as civilization ... ... However, suburbanization as a process involving the systematic growth of fringe areas at a pace more rapid than that of core cities ... ... occurred first in the United States and Britain, where it can be dated from about 1815” (page 130).

As Jackson (1985) implies, until the mid-20th century the problem of sprawl was coincident with the growth of the industrial city. This provided a focus for the transition from comparatively low density agricultural society to one dominated by much more intensive production and consumption in cities where economies of scale were achieved through very rapid population growth. The problem in the early 21st century is somewhat different. Production and consumption can now be spread out less intensively, although very specialized cores continue to grow at all levels of the urban hierarchy, and developments in transport and communications technologies enable populations to gain much wider access to facilities both physically and remotely. This image of a world composed entirely of cities and their suburbs has been anticipated for a long time. Almost one hundred years ago, H. G. Wells (1901, quoted in Hall, 1988) wrote about a completely urbanized Britain loosely cemented by new forms of communication technologies, a prospect which is well on the way to realization, despite relatively modest population growth. In the United States, where population growth has been much greater, where there are far fewer constraints posed by the land supply, and where the incentives for suburban development are much clearer, the problem of sprawl is different again although no less significant (Nivola, 1999).

Most of the focus of physical planning in western countries during the 20th century has been on ways of controlling urban growth but the recent wave of economic and related forms of institutional deregulation have given the problem a new urgency. Traditional solutions such as reducing our ability to locate in suburban locations through selective taxation of travel, combined with incentives to develop residential and other activities nearer the core of old cities are being suggested once again. But such policies are doomed in that they ignore completely the structure of the modern spatial economy where the central city is now just one of many nodes within a complex sea of urbanization whose pricing and market structure almost defies understanding (Krugman, 1993). More sensitive policies, particularly those being canvassed in North America, admit that such growth is going to take place and that it will be suburban, but more selective ways of letting it take place are being proposed. This is the idea of "smart growth" which basically involves controlled or managed sprawl where balance is the watchword for developing communities in which there is much the same variety of opportunity in travel and recreation as there is in less controlled growth (Nivola, 1999).

This brief commentary reveals our woefully inadequate understanding of urban growth processes but it also suggests that there are many different types of process and thus many different varieties of urban sprawl. To develop a better understanding of such processes as a prelude to better classification, it is hard to escape the conclusion that we must return to first principles, and examine the growth process in terms of its fundamentals. In this paper, we will begin this quest, abstracting urban growth to such a degree that we will only concentrate the geometric properties posed by such growth in space and time. We will assume away the practical issues posed by preferences for travel, space and other amenities as well as policy issues reflected through taxation and the market. We will also assume away demand and simply concentrate on how cities grow through the addition of space to their periphery and through aging processes which determine the condition of their physical stock. As such our models focus on generic processes which we assume are fundamental to growth and location; but these do not provide models that can be directly applied to real situations unless fiercely adapted to local conditions. In short, the tradition that we will adopt in this foray into urban growth modelling is aggregate and geometric, with little appeal to the kind of individual behaviours that define how populations react to prices and markets. Despite such simplicity and parsimony, we will argue that our models do provide insights into the dynamics of urban sprawl which are useful for intelligent discussion of the problem and its potential solution.

The key elements in the spatial dynamics of growth involve the available space within which growth takes place, and the aging process of development associated with that growth. A force or momentum for growth is required or growth will stop but this usually relates to wider considerations which pertain to issues beyond the immediate spatial confines of the city in question. The overall level of growth will depend on the local demographic factors as well as the attraction of the city to new growth from outside. Changing preferences for the consumption of space will affect growth but all these can be linked to the amount of space available. Our simplest model of urban growth can be conceived quite coherently without recourse to any formal or symbolic apparatus: imagine a city growing around a seed or core where the amount of new growth is proportional to the space available on its periphery, and that this space is unlimited in time, that is the city can keep on growing in this fashion in perpetuity. Then a wave of growth will move out from the core as the city grows in such a way that space and time are coordinated. However, development has a limited life span. Buildings do not last forever and thus as the city grows, past waves of growth will decline as buildings are demolished. If we assume that new growth occurs immediately following this regular decline, new waves of growth appear and as the city grows larger, its growth dynamic becomes dominated by successive waves on a spectrum from new growth at its edge to redevelopment phased according to the life cycle of the city's historic development.

In so simple a form, this model does not exactly mirror reality but already we have an image of how some cities develop. Often when development comes to the end of its life, it is abandoned and waves of new growth, building on the old, do not take place. Central cities are depopulated and abandoned while the suburbs keep on growing. This is a process of succession without the kind of invasion that some cities experience when older neighborhoods are rehabilitated and reoccupied. Of course, cities do not grow in perpetuity, and another element of our analysis will be to examine the effect of limiting the growth process on the waves of development that characterize this process. One element of our analysis does change the conventional analytic approach to suburbanization and sprawl. Often suburbanization and sprawl is assumed to embrace all the processes of change that are taking place outside the core of the city at any one time. In fact, this analysis will show that suburbanization only occupies a very small fraction of what is happening in the city at any one time, and that in mature cities, most change potentially comes from redevelopment and movement within the existing stock rather than new development. We will argue that herein lie one of the keys to solutions to the problems of sprawl.

The framework we will adopt is based on an analogy between the process of converting land from non-urban to urban and the idea that an individual becomes infected with a disease and then recovers. In short, we will articulate land conversion as a process of infection and then recovery, following quite well-established ideas in the simulation of epidemics. We first introduce this method of modelling sprawl, showing its key elements for the non-spatial case, and how it is linked to conventional growth processes based on exponential and logistic growth. We then generalize the model to a spatial context in which waves of growth are modelled as spatial diffusion, and we show how redevelopment is intrinsic to this process. We implement the model as a cellular automata (CA) and illustrate its working for significantly different examples. This model enables us to vary key parameters controlling the spread of development, its timing, and the degree to which the growth process is subject to random influences. This provides a basis for us to classify different outcomes and although our model is quite limited in import, we are able to show significantly different growth outcomes which provide a basis for classification. To progress this analysis to real world applications, we conclude by showing this logic can be embedded within a wider CA framework based on the **DUEM** model (Batty, Xie, and Sun, 1999; Xie, 1996), illustrating an idealized simulation of suburban growth using data for Ann Arbor, MI.

2 Sprawl as an Epidemic: Spatially Aggregate Models

Nivola (1999) makes the obvious point that cities "… can only grow in four directions, in, up, down and out … " with growth " … likely to follow the last of these paths overwhelmingly, particularly in advanced countries endowed with abundant usable territory" (page 2). Such growth reflects a combination of influences and preferences: population growth which must meet the geometrical and resource constraints posed by the shape and technology of the city, the preference for newness which is illustrated in higher demand and higher prices for new in contrast to old housing, the preference for lower densities which can best be met by development on the fringe of the city where more land is available, access to better environmental amenities which again suggest newer rather than older locations, and the desire to avoid traffic congestion. These elements are all reinforced in western countries by policies that provide a cornucopia of tax advantages to location in the suburbs. Empirically, the way western cities have grown over the last 200 years is entirely consistent with these forces. The notion that urban growth has advanced in waves of development outward from the central city or historic core is well established, as in Blumenfeld's (1954) characterization of the growth of Philadelphia, the same phenomena being implicit in much of the work on urban densities, as for example in Bussiere's (1972) work on Paris.

In treating the process of suburbanization as the cumulative growth of the city through additions to its periphery, we need to divide development at any time t into three constituent parts: development which is established (in this context surrounded by other development) which we call $P(t)$, new development $N(t)$ (which has just made the transition from undeveloped land), and available land $A(t)$ which drives the process of development in the first place. This process is one in which new development first makes the transition from undeveloped or available land and once this development is no longer adjacent to any available land, it passes into a mature state as established development. When the land is first developed, this depletes the stock of available land and if this stock is limited, then eventually growth will cease as the limit C of what can be brought onto the market is reached. In fact we will make the assumption in treating the problem as a spatial aggregate that the city and its hinterland are capacitated so that

$$P(t) + N(t) + A(t) = C, \quad \forall t, \quad t = 0,1,2,\ldots,T \qquad . \tag{1}$$

From equation (1), it is clear that any changes in each of the categories of development must also meet the constraint

$$\frac{dP(t)}{dt} = \frac{dN(t)}{dt} = \frac{dA(t)}{dt} = 0 \qquad . \tag{2}$$

The mechanism in the model is one-directional: available land passes to new development which then passes to established development as $A(t) \to N(t) \to P(t)$. Usually the process begins with all land $A(0)$ set equal to the capacity $C - e$, where the fraction e is the seed of new development that starts the process of growth and transition, that is $N(0) = e$. The pivotal change is in newly developed land from the addition of new development $dn(t)/dt$ and the transfer of such development to its established state, $dP(t)/dt$. Then

$$\frac{dN(t)}{dt} = \frac{dn(t)}{dt} - \frac{dP(t)}{dt} \qquad . \tag{3}$$

We hypothesize, as in similar models of change, that new development associated with available land is a proportion a of the product of $N(t)A(t)$. We can consider $a\,N(t)$ to be the proportion of new development that generates a unit of new development and the number of such units is defined by the available land $A(t)$. The transfer of new development to its mature state is also a proportion of g of each unit of new development $N(t)$ and using these definitions equation (3) becomes

$$\frac{dN(t)}{dt} = a\,N(t)A(t) - g\,N(t) \qquad . \tag{4}$$

Note that the term $a\,N(t)A(t)$ is an interaction term that represents the essential mechanism of conversion in the model. The two other components of change follow directly as

$$\frac{dP(t)}{dt} = g\,N(t) \text{ , and} \qquad (5)$$

$$\frac{dA(t)}{dt} = -\frac{dn(t)}{dt} = -a\,N(t)A(t) \quad . \qquad (6)$$

It is quite clear that these definitions meet the constraints posed by equations (1) and (2). A simple block diagram of the relations is provided in Figure 1 which shows that new development is an essential filter in the growth process which is articulated as the transition between undeveloped or available land and established development.

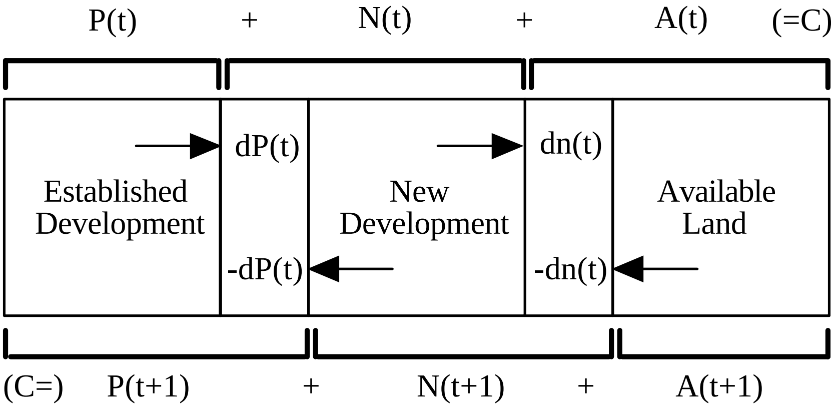


Fig 1: An Aggregate Model of Land Development

Equations (4) to (6) are identical to those used to model epidemics. If we define new development $N(t)$ as infectives, available land $A(t)$ as susceptibles, and established development $P(t)$ as removals or infectives who have recovered from a disease, then the process can be likened to the infection of a susceptible population with the gradual removal of those infected from the population either through recovery or death. In this interpretation, the central mechanism of the model is the interaction between infectives and susceptibles and in this case, $a\,N(t)A(t)$ has the immediate interpretation as the number of new infectives $dn(t)$ which is a proportion a of all the potential contacts between infectives and susceptibles. It is easy to guess what happens in this model for as the number of susceptibles from the fixed pool declines, the number of new infectives also falls and ultimately the epidemic dies out. The model has been quite widely applied to various kinds of epidemic (see Murray, 1993 for a good discussion) but it also has relevance to any problem which involves the transfer of resources from one state to another, a process that is widely used to model technological as well as political change (Banks, 1994; Epstein, 1997).

Closed forms for the three differential equations (4) to (6) that define the process cannot be derived although considerable algebraic analysis of the system is possible and it is quite easy to show how the model behaves. In particular, it is clear that the initial amount of available land $A(0)$ must be greater than some threshold for growth in the developed land categories to take place at all. If we write equation (4) as

$$\frac{dN(t)}{dt} = a\, N(t)\left[A(t) - r\right] \quad , \qquad (7)$$

where $r = g \,/\, a$, and noting that available land $A(t)$ must always reduce with time (as equation (6) implies), for the number of infectives to increase, equations (4) or (7) must always be positive. This implies that $A(0) > r$. r can thus be interpreted as a threshold which is the minimum amount of available land that is necessary to get an epidemic going. It is the ratio of the rate of transition of new to established development and the contact rate which is central to the generation of new development from available land. In epidemic models, it is referred to as the 'relative removal rate', and such models are often called 'threshold-epidemic' models accordingly.

This point is made clear another way. If we take the ratio of equations (6) and (5) which we write as

$$\frac{dA(t)}{dP(t)} = -\frac{a}{g}\, A(t) = -\frac{A(t)}{r} \quad , \qquad (8)$$

we can write an equation for land available in the steady state at $t = \infty$ as

$$A(\infty) = A(0)\exp\left[-\,P(\infty)\,/\,r\right] = A(0)\exp\left[-\frac{C - A(\infty)}{r}\right] \quad . \qquad (9)$$

Equations such as (9) can be used to compute the steady state values of the three key components for different parameter values as well as the parameter values which are able to reproduce a steady state from fixed initial conditions (Murray, 1993).

In Figure 2, we show typical trajectories generated by this growth process where the capacity limit C is set as 1000. We start the process with $A(0) = C - 1$, $N(0) = 1$, and $P(0) = 0$. The parameters $a = 0.000146$ and $g = 0.03199$ give a relative removal rate $r = 219.12$ which implies that $A(0)$ must be greater than this value for growth (an epidemic) to take place. Solving equation (9) iteratively gives the steady state value for available land $A(\infty) = 10.94$ which means that not all available land is utilized for development. The fact that growth stops a little short of the capacity limit is simply a consequence of the parameter values a and g . Nevertheless, what we see here is classic capacitated growth - logistic growth - but through a conversion process - a filter. Mature development grows according to the classic S-shaped curve while available land declines as a mirror image of this. The structural diagram which illustrates the way the three components of the model are related in Figure 1 suggests as much.

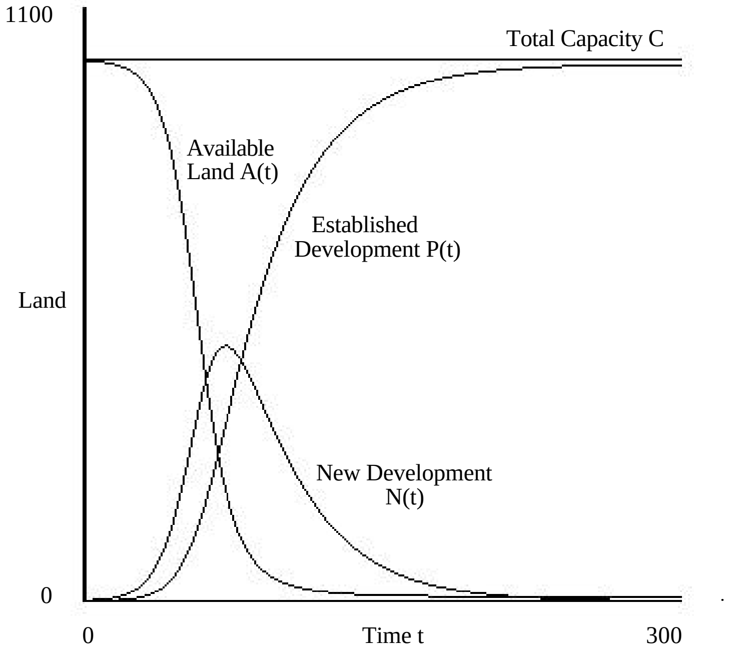


Fig 2: Trajectories of Development from the Basic Model

3 Simplifications and Extensions to the Aggregate Model

Some obvious simplifications need to be noted which ground the model in more conventional exponential and logistic growth processes. First assume that there is no transition from new to established development, that is $g = 0$, meaning that new development becomes established immediately. In other words, the filter of new development no longer exists and we write the overall capacity constraint as $N(t) + A(t) = C$. Then the system of equations in (4) to (6) collapses to

$$\frac{dN(t)}{dt} = \frac{dn(t)}{dt} = aN(t)A(t), \quad \frac{dP(t)}{dt} = 0, \text{ and } \frac{dA(t)}{dt} = -\frac{dN(t)}{dt} \quad .$$

As the terms for $dN(t)$ and $dA(t)$ are symmetric, then we need only deal with one of these. Noting that $A(t) = C - N(t)$, we can write $dN(t)$ as

$$\frac{dN(t)}{dt} = a\,N(t)\big[C - N(t)\big] = a\,C\,N(t)\left[1 - \frac{N(t)}{C}\right] \quad . \tag{10}$$

Equation (9) is one form of the logistic equation which on integration yields

$$N(t) = C \,/ \left\{1 + \left[\frac{C}{N(0)} - 1\right] \exp\left(-\,a\,C t\right)\right\} \quad , \tag{11}$$

while the equation for $A(t)$ is the reverse of (10) which we illustrate in Figure 3(a) for this growth process. This logistic is sometimes called the ‘simple’ epidemic model in contrast the ‘general’ model already introduced. The key insight here is that both the epidemic and logistic processes are capacitated but with the epidemic process involving a more complex transition from the undeveloped to the developed state. We can relax this model even further if we take off the capacity constraint for it might be argued that there are many growth situations in cities that are not limited by available land. We will also relax this assumption later in our more realistic spatial models but when we take the capacity constraint off here, a little algebraic manipulation of equation (11) produces the exponential growth model: $N(t) = N(0)\exp(a\ t)$.

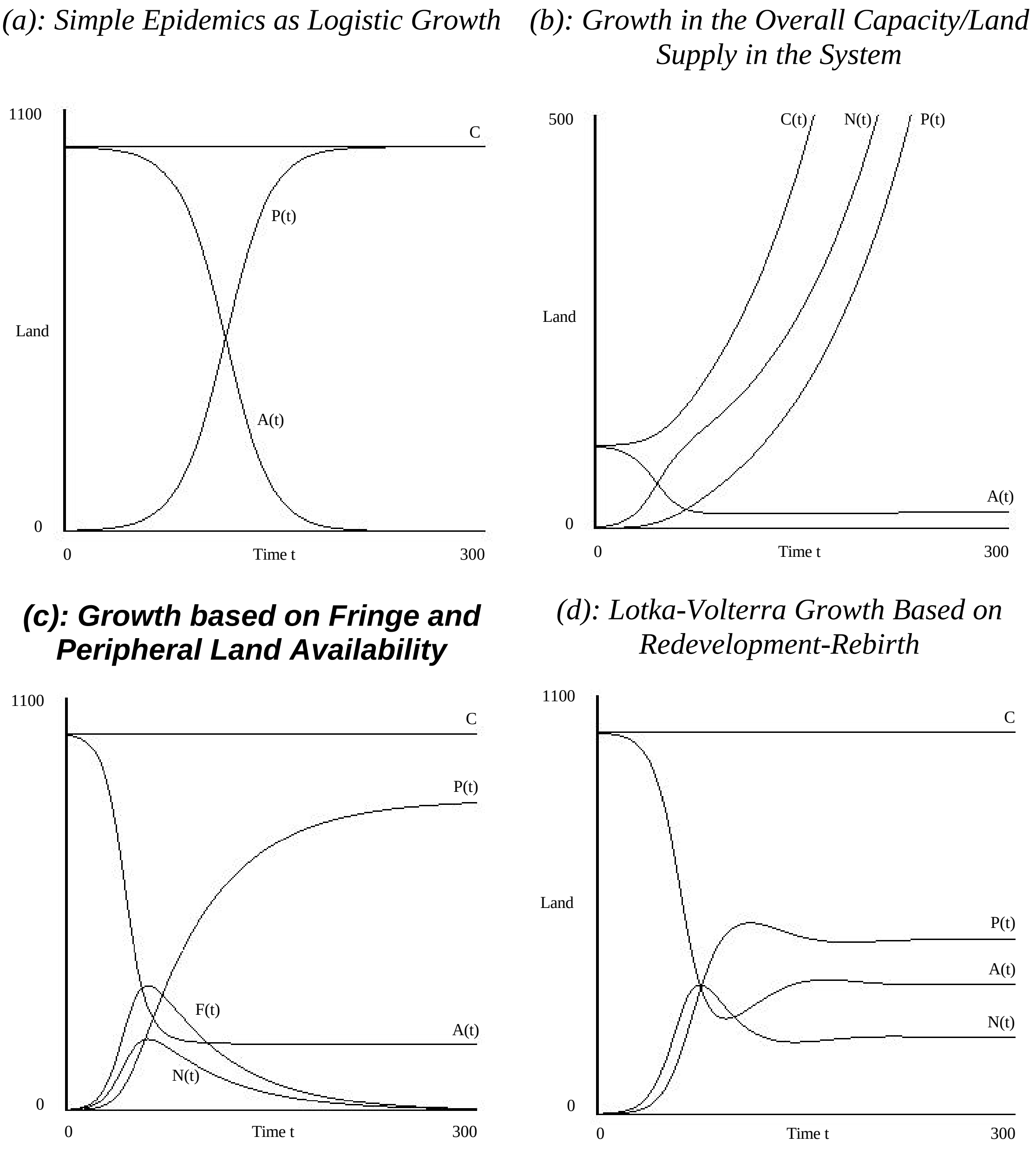
There are other ways of adapting the general epidemic model to deal with the more realistic conditions of urban growth. If the capacity of the system $C(t)$ is made a function of developed land, that is

$$\frac{dC(t)}{dt} = b\,P(t) \quad , \tag{12}$$

where b is the rate of growth in capacity, then the change in available land - equation (6) in the general aggregate model - becomes

$$\frac{dA(t)}{dt} = -\frac{dn(t)}{dt} + \frac{dC(t)}{dt} = -a\,N(t)A(t) + b\,P(t) \quad . \tag{13}$$

The trajectories generated by this model are illustrated in Figure 3(b), and as in many such models, after some initial oscillation, the system produces growth at a constant rate. This is indicated by the convergence of $A(t)$ to a constant level implying that all growth in available land is transferred directly into growth in new and thence established development.


(a): Simple Epidemics as Logistic Growth

(b): Growth in the Overall Capacity/Land Supply in the System

(c): Growth based on Fringe and Peripheral Land Availability

(d): Lotka-Volterra Growth Based on Redevelopment-Rebirth

Fig 3: Variants of the Basic Model

A third variant of the basic model involves introducing a structural distinction between two varieties of available land: a suburban fringe defined by $F(t)$ which we can associate directly with new development, and a wider suburban belt or periphery which makes up the rest of the stock of undeveloped land. We will define this periphery as $A(t)$ and the capacity of the system is now composed of

$$P(t) + N(t) + F(t) + A(t) = C \qquad . \qquad (14)$$

The difference between this and the general model is that the land in the fringe is made a function of change in new development, thus incorporating the idea that the fringe depends on where new development is located just as new development depends upon the fringe. This relationship is written as

$$\frac{dF(t)}{dt} = m\frac{dN(t)}{dt} \quad , \tag{15}$$

where m is the transition parameter. Equations for new and established development are the same as previously - equations (4) and (5) - but the change in available land is now set as

$$\frac{dA(t)}{dt} = \frac{dn(t)}{dt} - \frac{dF(t)}{dt} \quad . \tag{16}$$

If we now add the change equations (4), (5), (15) and (16), these sum to 0, thus ensuring that the total development in the system including fringe and peripheral land always equals the capacity of the system. The trajectories of this model are shown in Figure 3(c).

This variant shows some of the difficulties in making *ad hoc* adaptations to this framework. Figure 3 (c) illustrates that the problem with the model converging to a stable level of unused land is caused by the way the fringe interacts with this land and with new development. It is possible to predict the values of a ,g , and m which lead to different steady states as we did with the general model but the algebraic analysis becomes increasingly convoluted and it is beyond the scope of this paper. What this model does introduce is the notion that interaction between development and available land must be more sophisticated to ensure that the capacity constraint is only exercised when the system approaches this limit spatially. These models are not spatial and to anticipate those we develop below, we will resolve the problem by letting available land be generated locally rather than imposing any global limit. Such a limit will in fact still operate but only when development comes within ‘sight’ of this limit. This can only be accomplished when the model is specified spatially.

In all these models it is assumed that once development has been established, this state is irreversible. However development (like many diseases) has a life cycle which needs to be accounted for. Development ages physically and has an average life span before it is removed or demolished and new development takes it place. We model this as a process in which a proportion of established development comes up for redevelopment each time period, and this then adds to the available land supply to be converted to new development and back into the stock of established development. This process is analogous to an infected individual recovering and being removed from the infected population but after a time, becoming susceptible again to the disease. To equations (4) to (6) we add a new component reflecting development $l P(t)$ which re-enters the development process. l is the rate at which established development $P(t)$ leaves the stock of development and is converted back to available land. Equations (5) and (6) now become

$$\frac{dP(t)}{dt} = g\, N(t) - l\, P(t) \text{ , and} \tag{17}$$

$$\frac{dA(t)}{dt} = l\, P(t) - a\, N(t) A(t) \quad , \tag{18}$$

and it is clear that the capacity constraint in (14) is still preserved by these re-definitions.

When this model is run, various types of steady state eventually emerge whereas in the previous model, a certain proportion of new development is always passing through the process of development from available land to the established state. In one sense, redevelopment into available land simply cancels out the creation of new development but in the limit, the process repeats itself at a stable level. With the parameter values we have chosen, there is some mild oscillation before stability is attained as Figure 3(d) illustrates. To generate trajectories which would continually oscillate would require breaking the one-directional link between available land, new and established development completely. In this model the link is modified to pass back development to the available land pool but continued oscillations would require the introduction of a competition or interaction term with a different weight on $a\, N(t) A(t)$. In fact this model is a variant of the Lotka-Volterra form in which new development is akin to the predator and available land the prey. There are well-established solutions to such models for different parameter regimes which are illustrated for biological models by Murray (1993) and for highly aggregate urban evolution by Dendrinos and Mullaly (1985). Our purpose in breaking the one-directional chain in development at this point to include renewal and redevelopment simply anticipates the way we will deal with these issues in the spatial models that are introduced in the next section.

Our last variant extends the direct logic of the epidemic model structure. As we have emphasized, the general model simulates a transfer of resources - available land - through a filter called new development - to an ultimate state - established development, in such a way that the available land supply always decreases, and development always increases. More than one filter can of course be placed between these two states of non-development and development if there is good reason. For example, the process might be conceived as one in which peripheral land $A(t)$ is first transferred to a semi-prepared state $F(t)$ (on the urban fringe perhaps), which is then newly developed as $N(t)$ and then passed to its final state as established development $P(t)$. This in fact is more characteristic of land development in that fringe land often remains for some years in this intermediate state before it is developed and it seldom reverts back to available land first. The block diagram in Figure 4 illustrates this logic which is exactly the same as in Figure 1 but with $N(t)$ and $F(t)$ now both acting as filters. As in the general model, peripheral available land gradually declines but fringe land first increases (and then declines) with new development following the same process a little later, finally adding to the stock of established development.

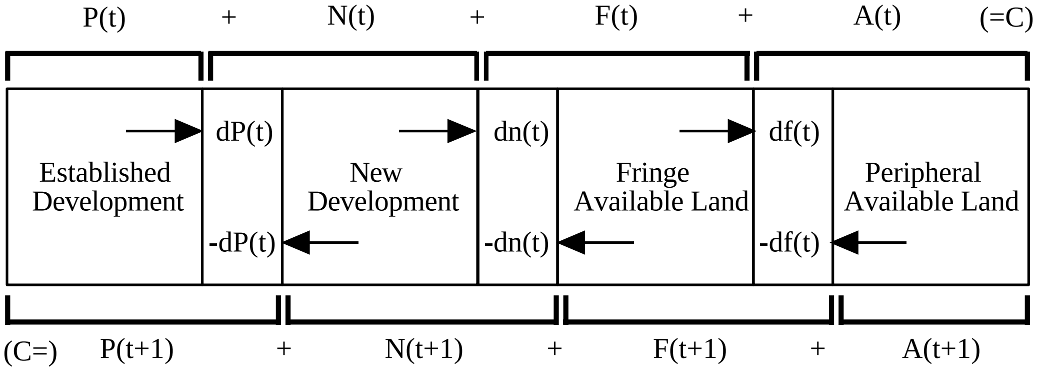


Fig 4: Extending the Aggregate Model

The equations describing this process need to be made explicit and following earlier definitions, we will now present the entire set as

$$\frac{dP(t)}{dt} = g\,N(t)\,; \quad \frac{dn(t)}{dt} = a\,N(t)A(t)\,; \quad \frac{dN(t)}{dt} = \frac{dn(t)}{dt} - \frac{dP(t)}{dt}$$

$$\frac{df(t)}{dt} = s\,F(t)A(t)\,; \quad \frac{dF(t)}{dt} = \frac{df(t)}{dt} - \frac{dn(t)}{dt}\,; \quad \frac{dA(t)}{dt} = -\frac{df(t)}{dt}$$

where s is the proportion of fringe land $F(t)$ interacting with available land on the periphery. The two key relations $dn(t)/dt$ and $df(t)/dt$ reflect the linking between the filters, newly developed land and fringe land determining new development but fringe land and the periphery determining the amount of land that passes into the fringe. A typical example is illustrated in Figure 5 which has the same steady state properties as the general model. Murray (1993) demonstrates the logic of how this variant can be further developed. More filters could be added in the same way and with the same recurrence of patterns while it is possible to extend this more general model to incorporate the other variants that have been developed in this section, particularly that involving redevelopment. However, at this point the basic logic of showing how land can be transferred through a process similar to the way epidemic waves sweep through a population has been illustrated and we will now generalize the model to a spatial context, first by considering how such development might diffuse across space.

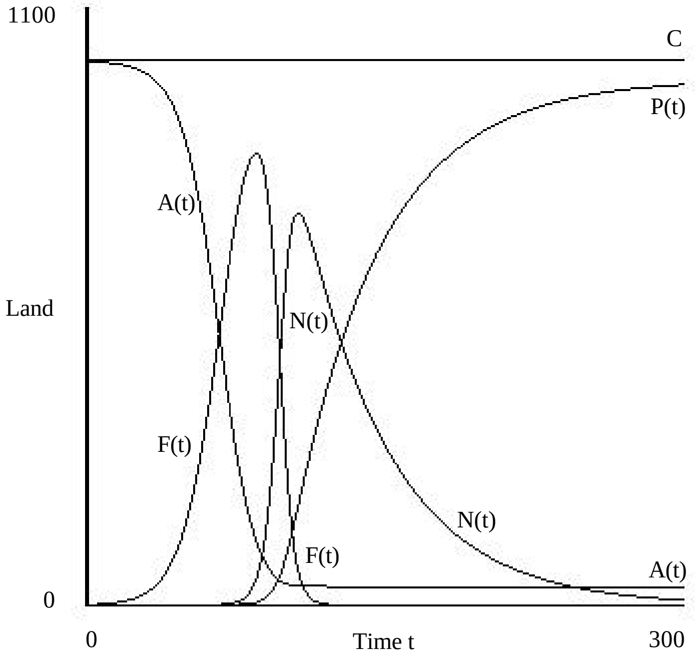


Fig 5: Trajectories of Mature, New, Fringe and Peripheral Land Development

4 Sprawl as Spatial Diffusion: Spatially Disaggregate Models

The translation of the general model to spatial coordinates x,y involves several changes in interpretation in that each cell or coordinate location has its own model while the interactions between locations are usually modelled through diffusion. First, land available across all locations does not formally enter the model framework as a constraint. Second, available land and new development are generated locally through diffusion to adjacent or nearby locations; that is land which is available for development at location x,y generates more land for development at locations adjacent to x,y. The same process can be articulated for new development which generates further new development around it. We now define the three land components as $A(x,y,t)$ - available land, $N(x,y,t)$ - new development, and $P(x,y,t)$ - established development and state the model as follows. Change in new development is specified as

$$\frac{\P N(x,y,t)}{\P t} = a\, N(x,y,t) A(x,y,t) - g\, N(x,y,t) + D_N \nabla^2 N(x,y,t) \qquad (19)$$

where D_N is a diffusion coefficient associated with new development. Change in available land is defined as

$$\frac{\partial A(x,y,t)}{\partial t} = -\alpha N(x,y,t)A(x,y,t) + D_A \nabla^2 A(x,y,t) \qquad (20)$$

where D_A is a diffusion coefficient associated with available land. New development passes to established development in a similar manner to that used in the aggregate model as

$$\frac{\partial P(x,y,t)}{\partial t} = \gamma N(x,y,t) \qquad . \qquad (21)$$

It is not meaningful to sum these change equations across x,y for all these show is the fact that available land and new development diffuse across the space according to how much land there is locally around new development. Any global constraints on the capacity of the whole system are thus operated externally, as we will see in the discrete version of this model which we operationalize below.

This model has been less widely applied than might be imagined, notwithstanding some seminal contributions to the spatial diffusion of epidemics. It is mathematically complex in that it generates waves of development that diffuse across space but are bounded by the model's parameter values and diffusion coefficients. Richardson (1941) appears to have been the first to propose a similar structure for the movement of population but applications to the urban domain have been restricted to theoretical illustrations based on spatial generalizations of the Lotka-Volterra framework with diffusion (Bracken and Tuckwell, 1992; Zhang, 1988). Most applications however have been to epidemics in human and animal populations, particularly measles, rabies, and the plague (Cliff, Haggett, Ord and Versey, 1981; Cliff, Haggett, and Smallman-Raynor, 1998; Murray, 1987; Noble, 1974; and Raggett, 1982). There are also suggestions that the model might be applicable to social phenomena such as revolutions (Batty, 1999; Epstein, 1997). There is however a large literature on the geographical diffusion of settlements (Morrill, 1968) but none of these examples have been embedded in the transition processes that we examine here.

The model in equations (19) to (21) is not in a suitable form for application to urban sprawl, for the processes of diffusion that characterize development are hardly likely to involve the kind of diffusion implied above. However, the way this model emphasizes local growth through diffusion is characteristic of the way an urban fringe develops and thus the framework can be easily adapted to suburbanization. Urban areas grow around their edges primarily due the demand for new space which is translated through the usual process into new development. New land becomes available adjacent to new development largely because of the existence of that development and thus the relevant diffusion is of new development determining additional land supply in its local neighbourhood. A more appropriate model form can be based on

$$\frac{\P N(x,y,t)}{\P t} = a\, N(x,y,t)A(x,y,t) - g\, N(x,y,t)\,, \text{ and} \qquad (22)$$

$$\frac{\P A(x,y,t)}{\P t} = -a\, N(x,y,t)A(x,y,t) + D_A \nabla^2 N(x,y,t)\,, \qquad (23)$$

where the spatial coupling is now through the diffusion term in equation (23). Equation (21) is unchanged. In this model, land is made available through the existence of new development although new development still depends functionally upon the existence of prior new development and available land.

In the model we will build, there is still a global capacity constraint but as land and development are local operations, then the global constraint only becomes relevant at the point in time when the system approaches this limit. Available land is not affected by this limit until the diffusion reaches the local neighbourhood of the limit and then growth stops rather quickly. Therefore most development simulated by this model is locally motivated and occurs without reference to the overall capacity of the system. It is this that makes the spatial model considerably more appropriate even in theoretical terms than the aggregate models of the previous sections. Furthermore, in the model based on equations (21) to (23), we will also incorporate the development cycle initiating redevelopment and renewal in the same way we introduced earlier. In the continuous version of the spatial model, such time delays have been studied in some depth and it is possible to simulate waves of development across space and time which are set off by such processes (Murray, 1993; Zhang, 1988). However in the computable form we will develop here, our analysis is more empirical and exploratory and to take this further, we must now move from a continuous to discrete structure in formulating operations analogous to those in equations (21) to (23).

5 A Computable Structure Based on Cellular Automata

Let us assume a cellular space based on a regular tessellation of grid squares whose referent is given by the coordinate pair x,y. In analogy to the continuous model in (21) to (23) we define available land as A_{xyt}, new development as N_{xyt}, and established development as P_{xyt}. Each cell can be either filled or empty with respect to the activity in question implying equal densities of development per cell. To indicate cell occupancy, we set the relevant activity equal to 1, that is $A_{xyt} = 1$ or $N_{xyt} = 1$ or $P_{xyt} = 1$ such that no cell can have more than one activity, that is $A_{xyt} + N_{xyt} + P_{xyt} = 1$. We now consider t a time period, and in any such interval, the development process consists of three key transitions based on: (1) the diffusion effect which adds to the available land supply in the neighbourhood of each unit of new development equivalent to the term $D_A \nabla^2 N(x, y, t)$, (2) the transition from available land to new development equivalent to $a\, N(x, y, t) A(x, y, t)$, and (3) the transition of new development to established development equivalent to $g\, N(x, y, t)$. Each of these three changes form the components of the continuous model equations (21) to (23) and we will deal with each in turn.

The diffusion effect is computed around each cell of new development by making adjacent cells available for development:

$$\left.\begin{array}{l} \text{if} \quad N_{xyt} = 1 \quad \text{then} \quad A_{ijt+1} = 1 \quad \text{where } i = x \pm 1,\, j = y \pm 1 \\ \qquad\qquad\qquad\quad \text{unless} \quad N_{ijt} = 1 \;\text{ or }\; P_{ijt} = 1 \end{array}\right\} \quad . \qquad (24)$$

Available land can thus come onto the market in any of the eight cells surrounding the new development in question, unless the cell is already developed. These cells lie at the eight points of the compass in a regular grid and define the so-called ‘Moore’ neighbourhood which is the most widely used in cellular automata modelling (Toffoli and Margolus, 1987). In this sense therefore we can consider the model to be a variant of CA although this is entirely dependent upon defining the neighbourhood in this local way. The second transition is from available land to new development:

$$\text{if} \quad A_{xyt} = 1 \quad \text{then} \quad N_{xyt} = 1 \;\text{ and }\; A_{xyt+1} = 0 \qquad , \qquad (25)$$

where the term $a\, N(x, y, t) A(x, y, t)$ is translated into this simple unweighted transfer. The last transition is from new to established development and in essence, we assume that when a cell of new development is no longer adjacent to any available land, then it is regarded as having entered the mature cycle. This is encoded in the model as:

$$\text{if} \quad \sum_{\text{i}=-1,+1} \sum_{j=-1,+1} \left[N_{x+i,y+j,t} + P_{x+i,y+j,t} \right] = 8 \quad \text{then} \quad P_{xyt} = 1 \;\text{ and }\; N_{xyt} = 0 \quad . \qquad (26)$$

This is a very simple process for in this form, available land and new development are separated by one time period. From a single seed, the process produces a wave of available land advancing one time period ahead of new development while behind the wave, new development becomes established. On a grid, this can be visualized as a square band of land one cell wide advancing in front of a band of new development one cell wide which is then converted to established development, the lag separating each component being one time period. When the edge of the system is approached, growth immediately stops for the capacity limit is only recognized locally through equation (24). Of course is it possible to vary this process by changing the size of the neighbourhoods and incorporating differential time lags, but a more important issue is to break the spatial symmetry by introducing noise into the system.

It is much more realistic to assume that new development is generated from available land according to a certain probability, thus reflecting the fact that different developers of different cells vary in the way they finance development and react to the market. Such an extension would clearly break spatial symmetry, producing development clusters with irregular edges much more characteristic of real cities. A straightforward modification to equation (25) achieves this:

$$\text{if} \quad A_{xyt} = 1 \ \text{ and } \ random(\Lambda_{xy}) > \Phi \quad \text{then} \quad N_{xyt} = 1 \ \text{ and } \ A_{xyt+1} = 0 \ , \quad (27)$$

where $random(\Lambda_{xy})$ is a random number from a predefined range and Φ is a threshold above which land is converted to new development. If for example, $\Phi = 950$ and the range is from 0 to 999, then this implies that if a random number drawn from this range is less than 950, then development would not take place. In turn, this implies that about 95 times out of 100, new development would not take place on the set of cells which constitute available land (in that time period in question).

Aggregate growth trajectories equivalent to those in the aggregate models presented previously can be computed as

$$A_t = \sum_{x,y} A_{xyt}, \quad N_t = \sum_{x,y} N_{xyt}, \quad \text{and} \quad P_t = \sum_{x,y} P_{xyt} \qquad .$$

When these are plotted for development around a single seed placed at the centre of a square bounded grid, A_t and N_t produce waves which are coordinated in space and time, rising exponentially in size at first and then decreasing rapidly as the boundary of the system is reached. This is also reflected in the path of established development P_t which reflects logistic growth although the logistic effect only kicks in when new development reaches the nearest boundary cells in the N-S-E-W directions, illustrating that capacitated growth is only a feature of the simulation once the boundary has been sensed. This is somewhat different from the aggregate model above where this capacity is always accounted for through the interaction between development and available land.

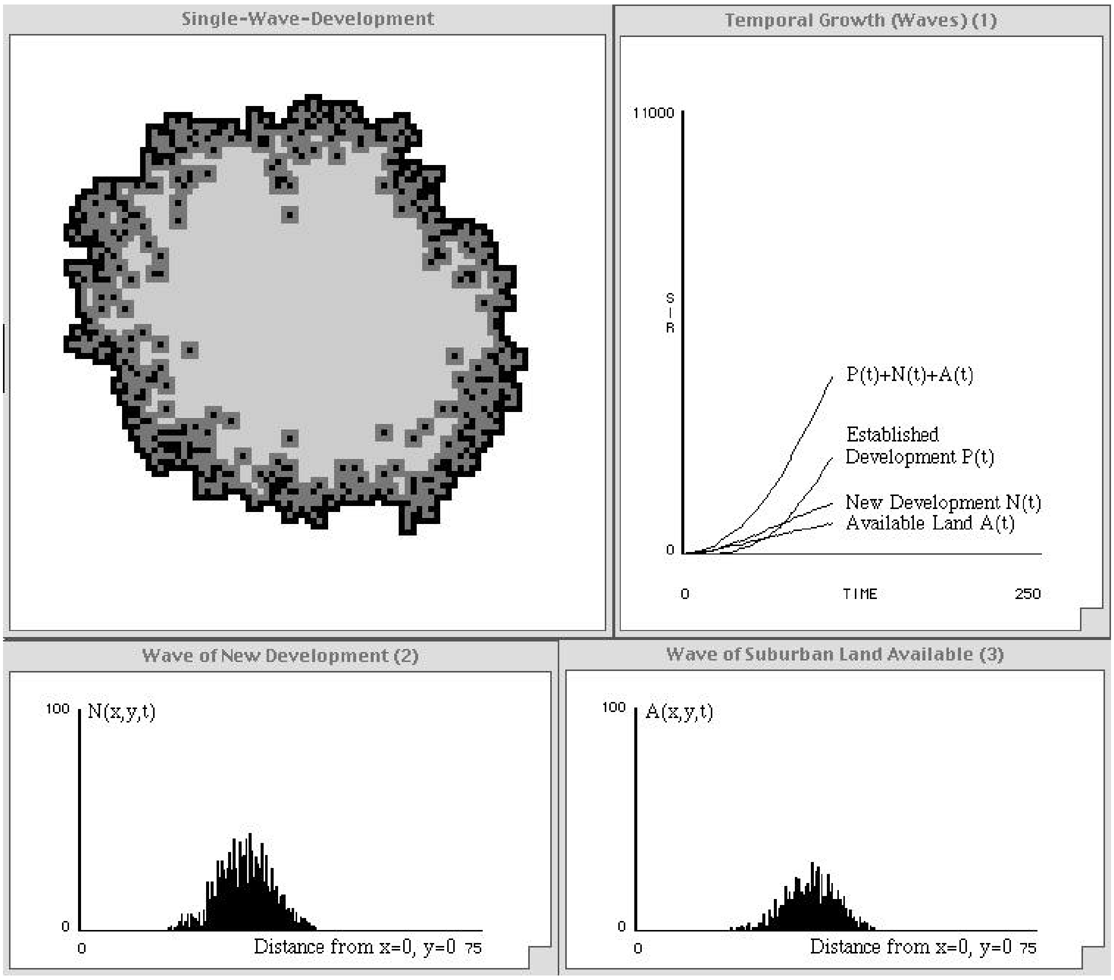


Fig 6: Spatial Diffusion from a Single Seed: Growth and Morphology

Our first example plants a seed of new development at the centre $x = 0, y = 0$ of a grid whose dimensions are 101 x 101. We have set the development threshold at $\Phi = 900$ which implies on average only 10 percent of our transfers from available land to new development are successful. However, this simply delays the development process and it does not widen the fringe of land available for development very much. The key feature of this probabilistic process is that it breaks spatial symmetry. There are no other parameters in this model and in Figure 6 we show the development of the growing cluster at $t = 105$. Light gray is established development, mid gray new development, and black available land. The cluster is characteristic, in terms of its boundary at least, of compact urban development but this level of compactness - development without holes - is unusual in western cities and in a later section, we will relax this assumption. The three other windows alongside the cluster show the distribution of new development and available land over space (measured from the centre

$x = 0, y = 0$ up to a distance of 75 units), and the total activities A_t, $N_{t,}$ and P_t together with the total land affected by the development process $C_t = A_t + N_t + P_t$. The two windows which are associated with the spatial distribution of new development and available land show waves of land development. These waves move outward from the centre as the simulation proceeds and in the software used, these are updated throughout the simulation. These waves grow in size as the cluster grows and as the cluster approaches the capacity of the system these waves gradually decline.

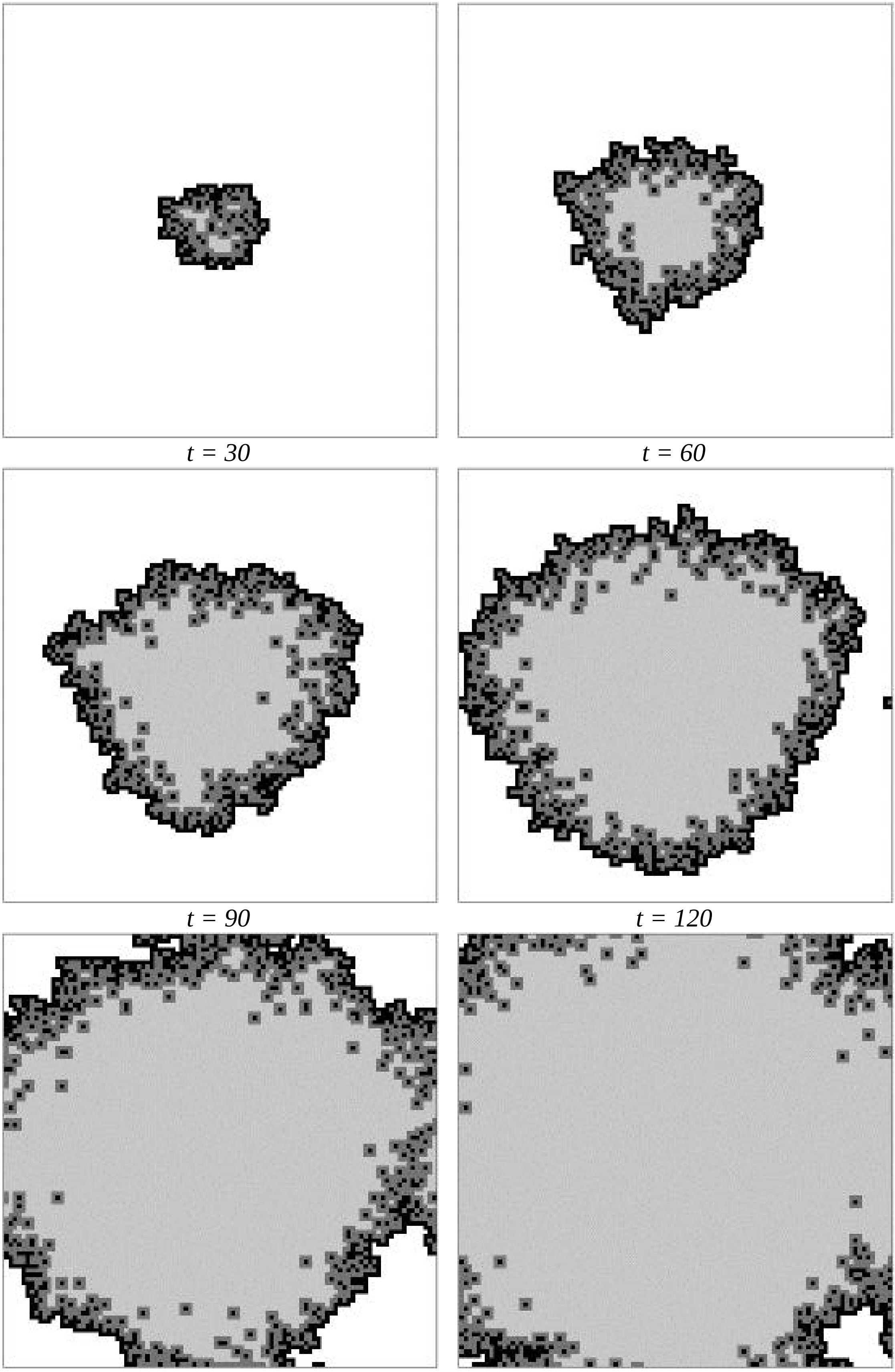
Each of the windows is coordinated within the software used and Figure 6 is thus a snapshot from an animation across 250 time periods (in this case), illustrating the way the system builds up and eventually reaches a steady state in which all development is established and growth ceases. Six snapshots from this process are shown in Figure 7 where it is clear that ultimately the cluster is composed only of established development. The overall growth trajectories are shown in Figure 8 after the simulation reaches $t = 250$. In one sense, the waves mirror those of the aggregate model in Figure 2; we also show the space remaining which is $R_t = C - C_t$ which mirrors the distorted S-shaped curves associated with P_t and C_t. Note that if required, we can also compute all the change variables associated with these stocks, but these do not show any surprises and can easily be visualized by considering the first derivatives associated with the growth paths in Figure 8. We should also note that spatial epidemic models have been conceived in this simple way before. The models presented by Durrett (1995) have many similar characteristics to those of this section although their purpose is for laying bare statistical assumptions than for actual simulation. Finally, it should be noted that some fractal models of urban growth have been based on diffusion and these employ the same kinds of local cellular operators (Frankhauser, 1994).

6 The Dynamics of Urban Regeneration

In the aggregate model, we indicated how established development might revert to available land thus initiating the development process once again for the locations in question. Putting established development back into the development process is in fact one of the key processes of urban redevelopment or regeneration. To derive a continuous spatial equivalent, we modify equations (21) and (23) as

$$\frac{\partial P(x,y,t)}{\partial t} = \gamma\, N(x,y,t) - \lambda\, P(x,y,t) \ , \qquad (28)$$

$$\frac{\partial A(x,y,t)}{\partial t} = -\alpha\, N(x,y,t) A(x,y,t) + D_A \nabla^2 N(x,y,t) + \lambda\, P(x,y,t) \, , \qquad (29)$$



t = 30

t = 60

t = 90

t = 120

t = 150

t = 180

Fig 7: Snapshots of Growth

where $l\,P(x, y, t)$ is the proportion l of established development $P(x, y, t)$ which comes back onto the market as available land. This model does not however take account of the life-cycle of development. To do this we need to introduce a lag consistent with the age of the development. This could be done by replacing $P(x, y, t)$ with $P(x, y, t-t)$ where t is an age limit after which development must be redeveloped. Models such as these have been developed in the continuous case for the study of the spatial spread of disease such as rabies (see Murray, 1993) and the form in equations (22), (28), and (29) is similar to that developed by Zhang (1988) for urban development. As previously, it is preferable for us to deal with the discrete CA form and extend equations (24) to (26).

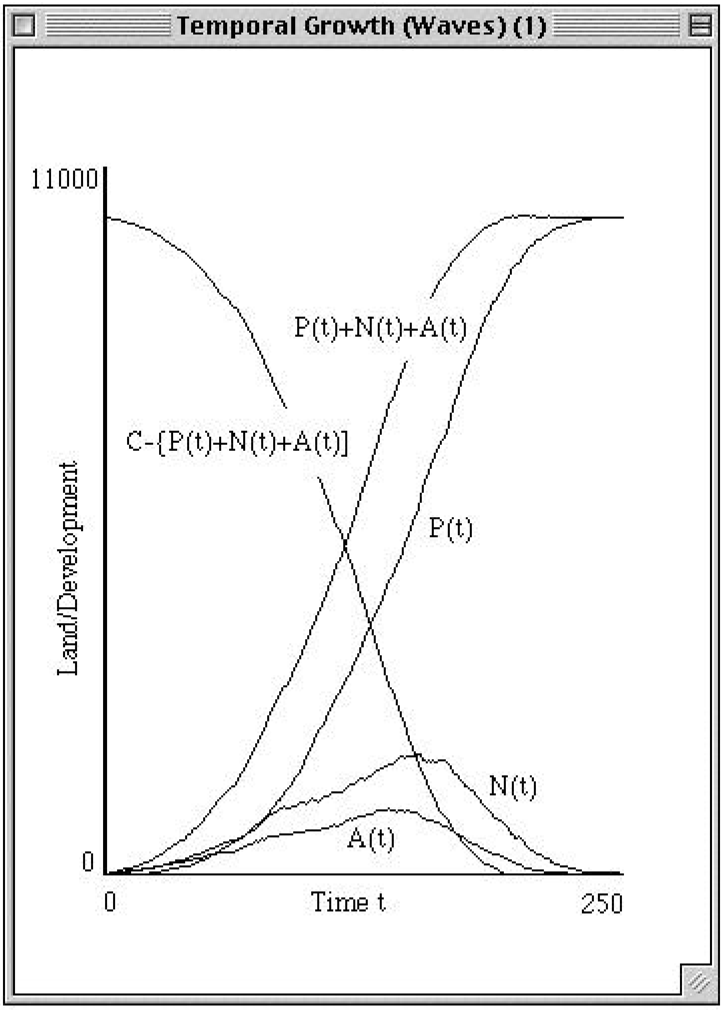


Fig 8: Waves of Development within Logistic Growth

We need to consider how established development reverts back to available land and thence once again enters the development process as new development. To do this at each time period, we need to define the age of a unit of development - the age of each cell x, y - as B_{xy}

which is defined whenever a cell becomes newly developed. To put development back into the development process, we test the following:

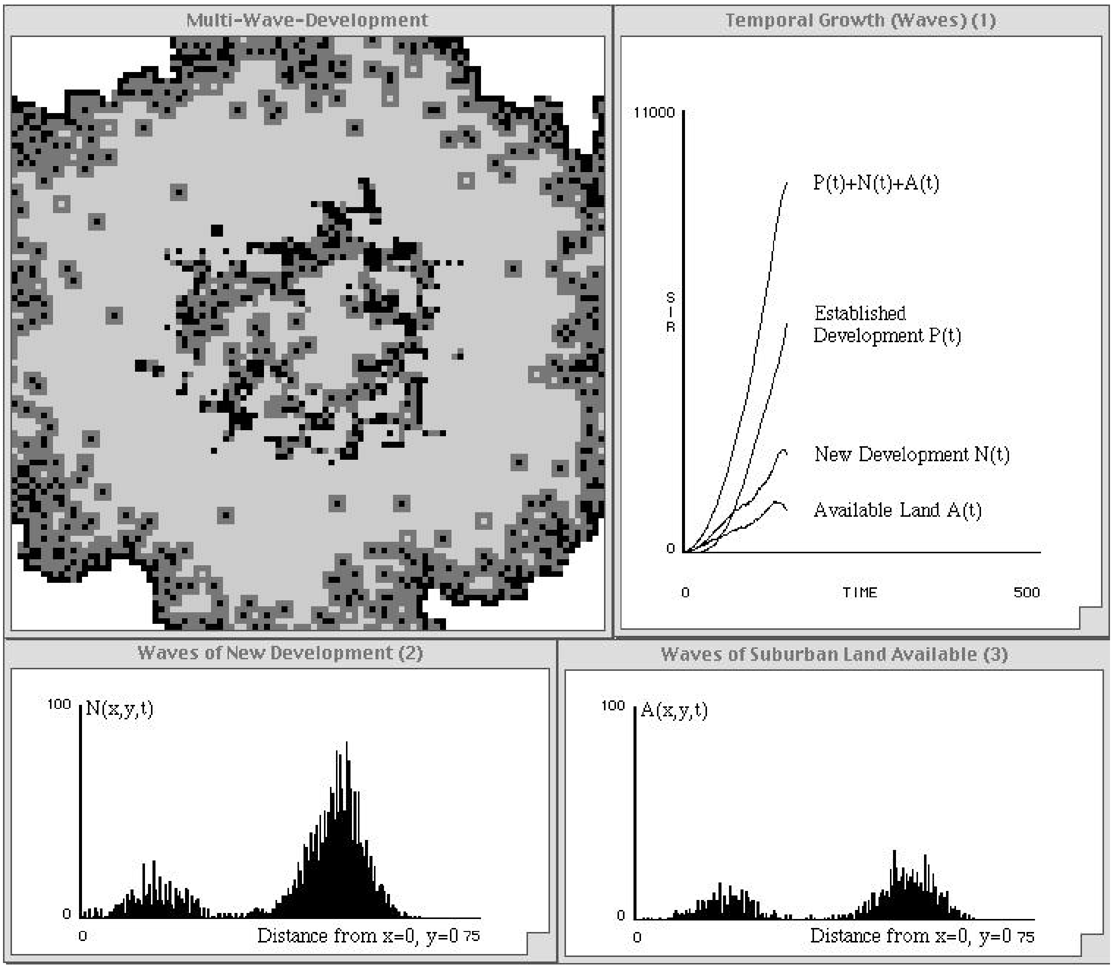
$$\text{if} \quad t - B_{xy} = t \quad \text{then} \quad A_{xyt} = 1 \ \text{ and } \ P_{xyt+1} = 0 \quad , \qquad (30)$$

where t is the age threshold of the development at which it must be redeveloped or renewed through demolishing the structure, clearing the land, and placing the land in question back into the pool of available land to be considered for new development during the next time period. This rule could be randomized on the assumption that there is variation in this decision but we have not done this here as it only adds to the number of possibilities which might be simulated.

Using the same cellular space as previously (in Figure 6), and now setting the threshold for redevelopment as $t = 90$, we show the morphology and waves of development at $t = 146$ in Figure 9. It is immediately apparent from each of the four windows, that a new travelling wave of development from redevelopment, begins at $t = 90$. The initial wave moves outward from the seed of the cluster and this can be identified as the first - most right-hand - wave in the new development and available land windows, followed by the second wave some 90 time periods behind. This second wave also embodies the randomness of new development produced by the probabilistic land conversion process and although this second wave is associated with the exact times at which the first wave reaches the redevelopment threshold, this is further randomized by the conversion once again to new development. The second waves are thus less sharp than the first. Were we to randomize the time at which the age threshold for redevelopment were activated, this would spread these waves out even further.

We now show what happens when this process is simulated over 500, then 2000 time periods. Every time a site is (re)-developed, more noise is introduced into the actual time it passes from available land to new development. Over time, this means that two sites which were originally developed at the same time, might find themselves being redeveloped at times more and more distant from one another. In short although initially the two sites were part of the same wave of growth, as they are redeveloped they are more likely to be in different waves; or in fact, the waves themselves are more likely to become less distinct. More and more noise is thus introduced into the spatial cluster each time sites are redeveloped and thus the initial spatial-temporal structure which might be very distinct, begins to break down. In Figure 10, we show the pattern of development of the growing cluster after 500 time periods during which redevelopment of all sites will have taken place at least 4, probably 5 times. Because the age threshold is 90 and given the speed of growth in this model, it is only likely that two waves of development are distinct at any one time in a space of 101 x 101 cells. In

Figure 10, these two waves are much less sharp for new development and available land than after 146 time periods as in Figure 9. Figure 10 also shows how these waves are reflected in aggregate trajectories. Because the system is capacitated, available land and new development as well as established development oscillate on a cycle of 90 years, reflecting the fact that a fully developed system never quite results due to the fact that a proportion of land remains undeveloped in the redevelopment process throughout time. Although we cannot demonstrate this here, the simulation is highly sensitive to the age threshold and the system becomes much more volatile as this is decreased.


Multi-Wave-Development

Temporal Growth (Waves) (1)

Waves of New Development (2)

Waves of Suburban Land Available (3)

Fig 9: Spatial Diffusion with Waves of Redevelopment in Every 90th Time Period

We continue the simulation to 2000 time periods in Figure 11. By this time, the local waves across space and time and the spatial pattern itself have become quite unstructured. This is clearly evident in the aggregate trajectories where the system has more or less converged to a steady state in which the level of redevelopment is almost constant during each time period. Note that this situation has been reached after 20 or more cycles of redevelopment which only begin when the first development reaches $t = 90$. The oscillations in fact become more and more frequent and their amplitude dampens considerably. The travelling waves of new development and available land have become quite unstructured as they have spread out in space and time while the morphology of the cluster shows that the original concentric pattern of growth has all but disappeared. If we had added a random component to the time when the age threshold activated redevelopment, then these waves and patterns would be even less structured for another level of mixing would have occurred. This kind of randomness in fact is characteristic of real cities and the kind of mixing of new development and sites becoming available for redevelopment that marks Figure 11 virtually wipes out the original structure which initially emerges from development around a single site. In terms of this model, there is no clue to how the initial structure might be recovered from Figure 11 although in real cities despite such mixing, clues do remain as to the initial structures and rules that have generated the morphology.

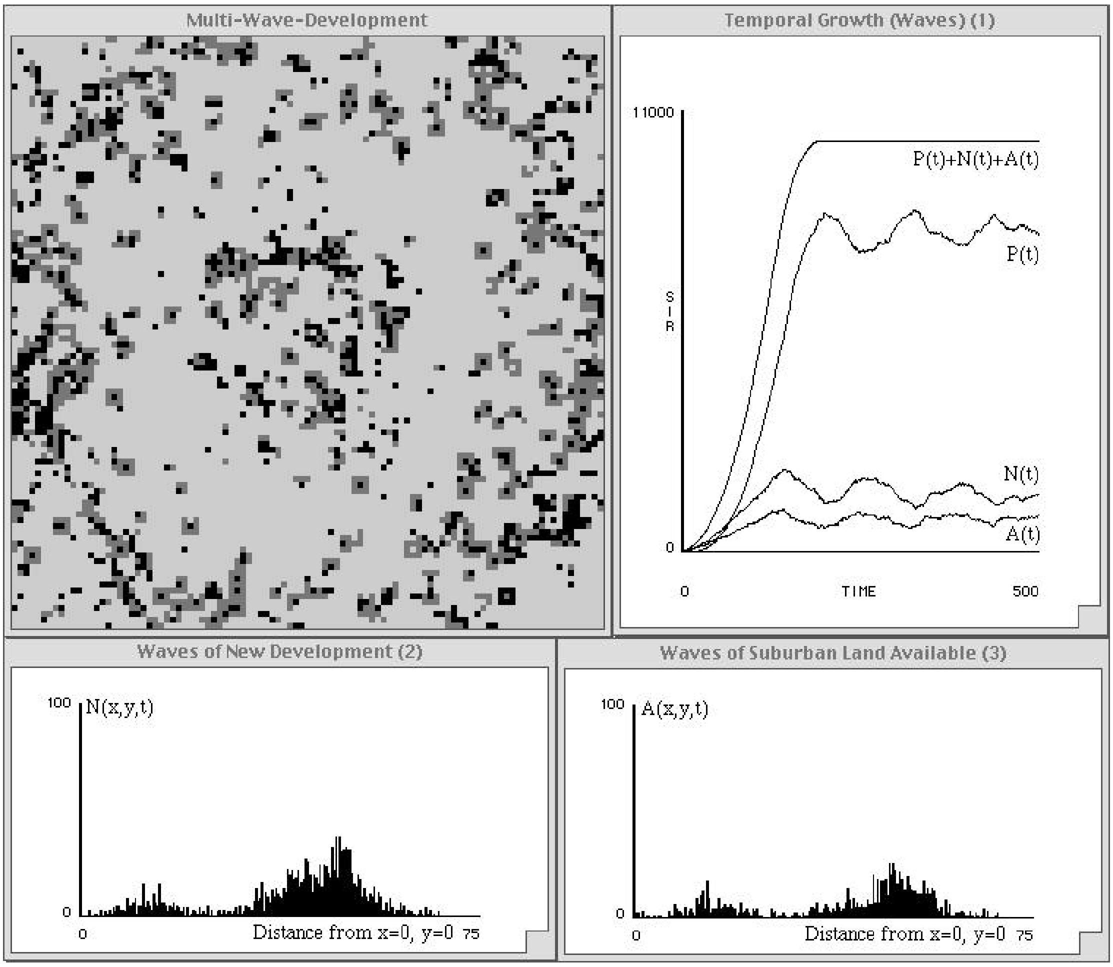


Fig 10: Waves of Development and Redevelopment over 500 Time Periods

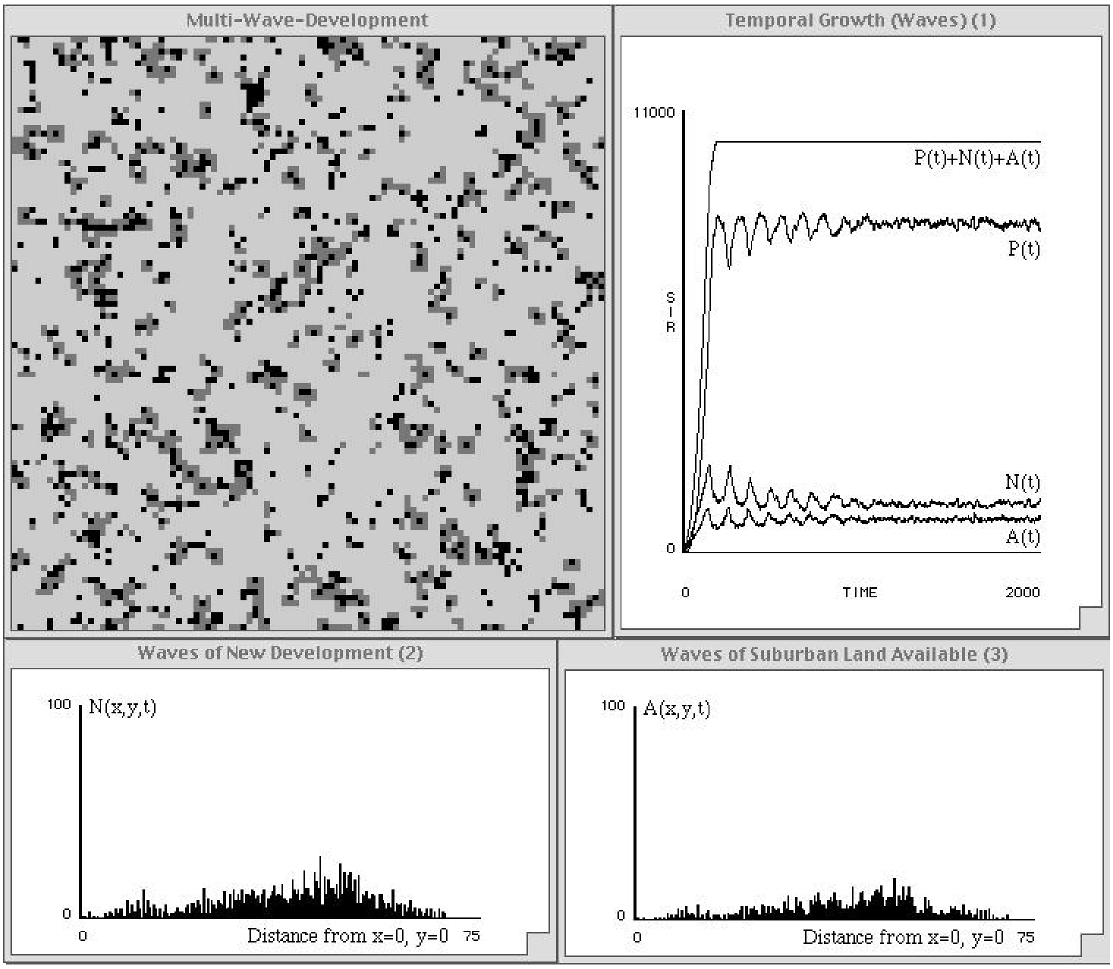


Fig 11: Disappearing Spatial and Temporal Structure through Successive Mixing

7 Classifying Sprawl through Morphology

In all our simulations so far, we have ignored the fact that urban morphology is peppered with undeveloped and semi-developed land which often remains in this state in the presence of rapid growth. In fact, just as urban form is fractal, then the 'holes' that comprise such structure often display an equivalent degree of fractality, with 'holes' distributed similarly in terms of size and shape across many scales (Frankhauser, 1994). We need to introduce a fourth type of land use to make our models morphologically more realistic in these terms and thus we define a state of vacancy for any cell as V_{xyt} . Vacant land is not land which cannot be developed because of physical constraints, nor is it land that has been withheld from the market by legal controls, for such land categories can only be accounted for by external inputs to our models. This land is that which remains vacant in the presence of growth due to behavioural, not physical decisions, involving the land development process. It may be the wrong shape, or be part of a longer term process of land assembly or it may simply be kept in a vacant state as an investment. Whatever the reasons, we need to introduce mechanisms in our model to ensure that land remains vacant and we will do this once again using simple probabilities.

Because our model has slowly become more and more complex, we will now restate all the model equations which now involve vacant land, noting that any cell can only be in one use; this is equivalent to ensuring that $V_{xyt} + A_{xyt} + N_{xyt} + P_{xyt} = 1$, where $V_{xyt} = 1$ when it is vacant, 0 otherwise. Let us assume again a simulation which begins with the central cell in the system as new development, that is $N_{000} = 1$, and $N_{xy0} = 0, x \neq 0, y \neq 0$ with $A_{xy0} = 0$, $V_{xy0} = 0$, $P_{xy0} = 0, \forall xy$. We will present a typical simulation of the model in the order in which the various model operations take place for any time period t. First the age threshold is checked and if the development has reached the age at which it must be renewed it is put back into the pool of available land

$$\text{if} \quad t - B_{xy} = t \quad \text{then} \quad A_{xyt+1} = 1 \text{ and } P_{xyt+1} = 0 \quad . \qquad [(30)]$$

Then the diffusion effect which determines whether or not available land (around new development) comes onto the market is effected

$$\left.\begin{array}{l} \text{if} \quad N_{xyt} = 1 \quad \text{then} \quad A_{ijt+1} = 1 \quad \text{where } i = x \pm 1, j = y \pm 1 \\ \qquad\qquad \text{unless} \quad N_{ijt} = 1 \text{ or } P_{ijt} = 1 \end{array}\right\} \quad . \qquad [(24)]$$

The new vacancy criterion is then checked. If land is available and or already vacant, it becomes or remains vacant if a probability threshold Γ is exceeded:

$$\text{if} \quad A_{xyt} = 1 \text{ or } V_{xyt} = 1 \text{ and } random(\Lambda_{xy}) > \Gamma \quad \text{then} \quad V_{xyt+1} = 1 \quad . \qquad (31)$$

The neighbourhood constraint which determines whether new development is transferred to mature development is now augmented with the vacant state as

$$\begin{array}{l} \text{if} \quad \sum_{i=-1,+1} \sum_{j=-1,+1} \left[N_{x+i,y+j,t} + P_{x+i,y+j,t} + V_{x+i,y+j,t} \right] = 8 \\ \qquad\qquad\qquad \text{then} \quad P_{xyt} = 1 \text{ and } N_{xyt} = 0 \end{array} \quad . \qquad (32)$$

Finally the transition from available to new development is made, noting that at this stage the age of the development is determined:

$$\begin{array}{l} \text{if} \quad A_{xyt} = 1 \ \text{ and } \ random(\Lambda_{xy}) > \Phi \\ \qquad\qquad \text{then} \quad N_{xyt} = 1,\ B_{xy} = t,\ V_{xyt} = 0, \text{ and } \ A_{xyt+1} = 0 \end{array} . \tag{33}$$

In this model, there are only three parameters t, Γ, and Φ and one of these t controls the redevelopment process. In the experiments we report in this section, we will set t to a large number to ensure that redevelopment has no effect and we will concentrate on the morphologies that result when we vary the values of the vacancy and spread parameters, Γ and Φ respectively.

The effect of the spread parameter Φ is to vary the speed of development. In the basic model with redevelopment and vacant land, all that Φ does is to slow down development as the value of the parameter is increased. Because there is no vacancy possible in this model and all land is ultimately developed, when Φ is large and near 1000, sites in the neighbourhood of new development have a low chance of being developed but as the growth process continues indefinitely until the capacity of the system is reached, at some point they will get developed. This does make a difference to the edge morphology of the system. When Φ is small, the system develops rapidly with little irregularity to the urban fringe. However when we introduce the vacancy parameter Γ, if development is proceeding slowly with a high value of Φ, then land which is not developed at any time period but is within the available land pool, continually has a chance of falling vacant whereas with faster development, this chance is lower. In short, it would appear *a priori* that development is increasingly unlikely as both the vacancy and spread thresholds are increased.

We have modelled the impact of a range of values for the vacancy and spread parameters and for each of the clusters which develop, we have measured the final proportion of vacant land when the growth process is terminated. This occurs when no more available land is generated from new development. In Figure 12 we plot changes in the proportion of vacant land associated with a range of vacancy parameters from 0 to 700 for two growth situations: very slow growth with $\Phi = 975$ and very fast growth with $\Phi = 50$. The functions generated are very similar in shape but displaced from one another with respect to the values of the vacancy parameter. What this graph shows is that for small values of Γ and Φ, almost complete clusters are generated with a low proportion of vacant sites, less than 10 percent. However for the large value of Φ - slow speed of growth - the vacant sites increase very quickly and once the vacancy parameter reaches around 400, then the growth of the cluster is halted around its seed. This is in marked contrast to the low value of Φ - fast speed - where vacant sites in the growing cluster are much harder to generate and retain, although by the time the vacancy parameter has reached 700 or so, the combination of this with the growth speed halts development before it gets a chance to start.

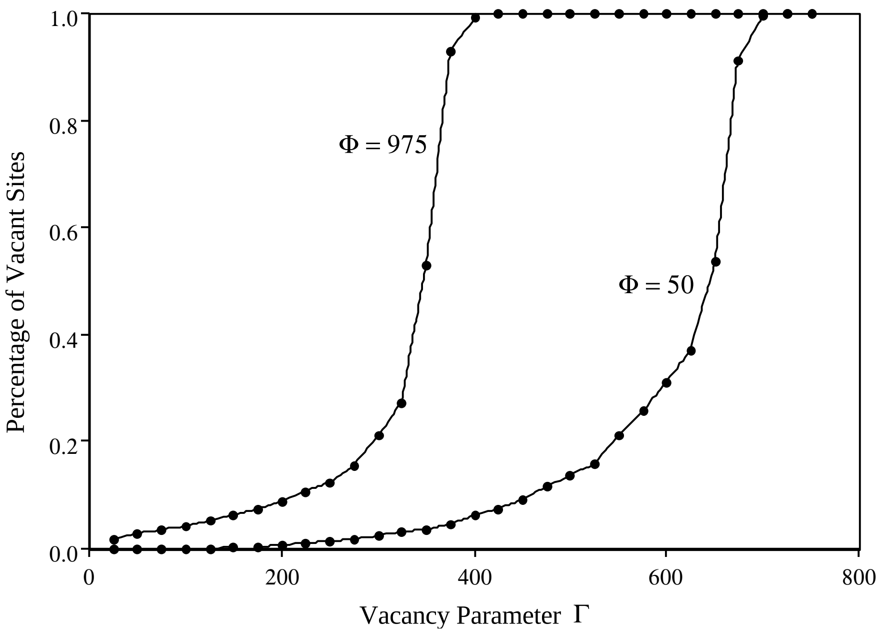


Fig 12: The Parameter Space { Φ, Γ *} Associated with the Proportion of Vacant Land*

It is however the morphologies that are of real interest here for the combination of the two parameters does give rise to fractal growth under most regimes. In Figure 13, we have graphed a series of relationships between the percentage of vacant sites in the steady state and the vacancy parameter, the family of curves being associated with different speeds of growth. To evaluate the resultant morphologies, we have sampled the space showing various forms for 10 sets of parameter values. In fact the shapes that we show are not those of the steady state for we need to give an impression of how the clusters develop. In the steady state, as the system is capacitated, spatial irregularity is harder to observe; the system is bounded by the square grid although it would be possible to colour code the development to provide a general impression of this irregularity. What we have done is to stop the growth as the cluster begins to approach the boundaries of the grid. As the grid is a torus, the usual default of graphic computer screens, development does wrap around but the clusters shown give an immediate sense of the size and distribution of vacant sites under different combinations of threshold values.

From the static snapshots of growth that we provide in Figure 13, we are not able to provide a sense of how these clusters develop. In the sparser examples with higher vacancy rates and faster speeds, development often proceeds dendritically with the tips of the dendrites being the focus of development although because the development wraps around in the closed spaces we have used, these dendrites soon disappear. In some measure, the pattern of development is not unlike that of a forest fire with tips of the growing cluster being ignited as available land comes onto the market and is then turned into new development. With high levels of vacancy, many of these tips do not develop. If you look closely at the pictures in Figure 13, you will see the tips which are composed of available land and new development which drive the process either to extinction or to complete occupation of the space. These examples show the power of this framework to simulate relatively realistic forms and this is without using varying neighbourhood sizes, multiple seeds, transportation networks, and all the other features that clearly have an influence on urban form.

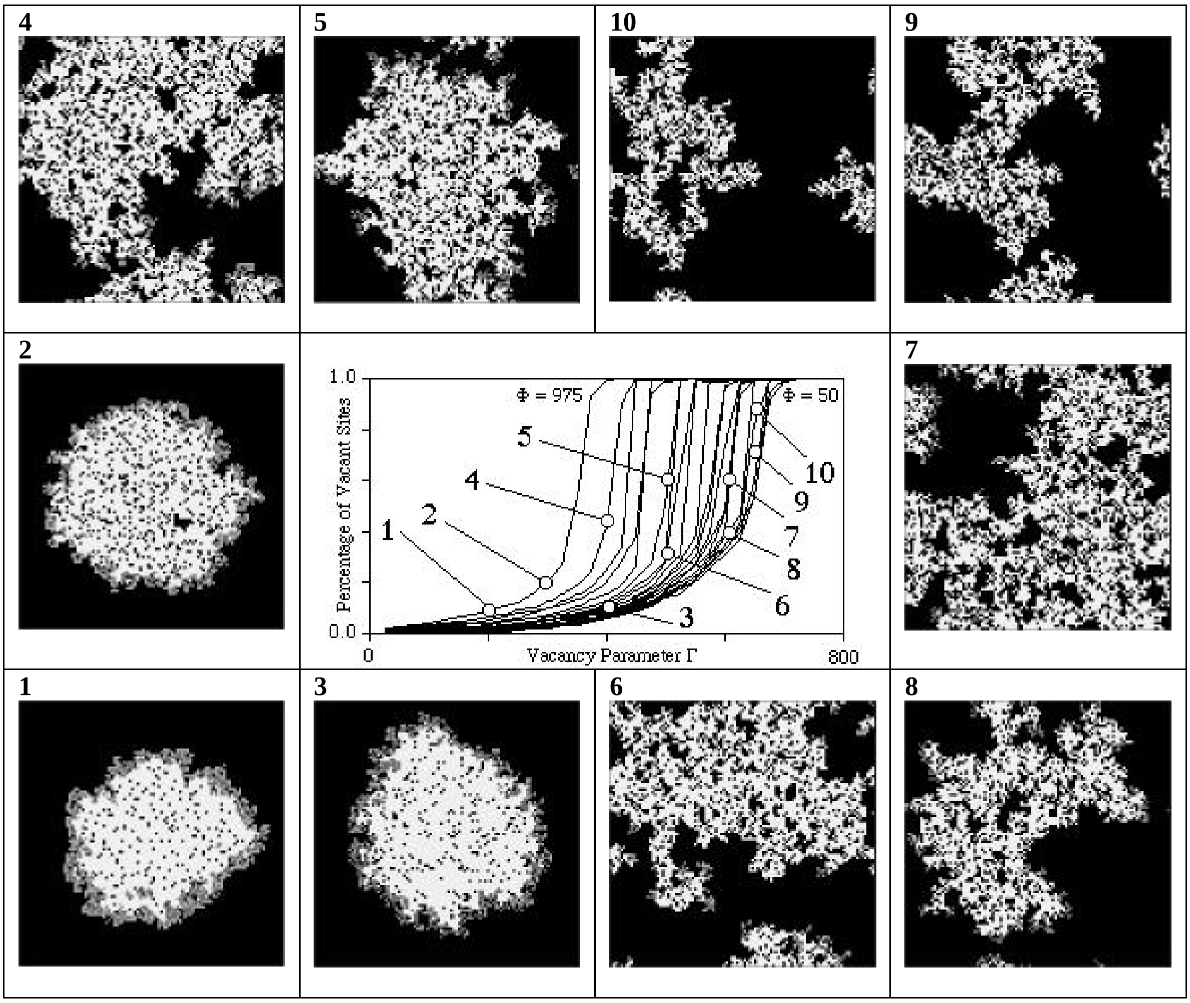


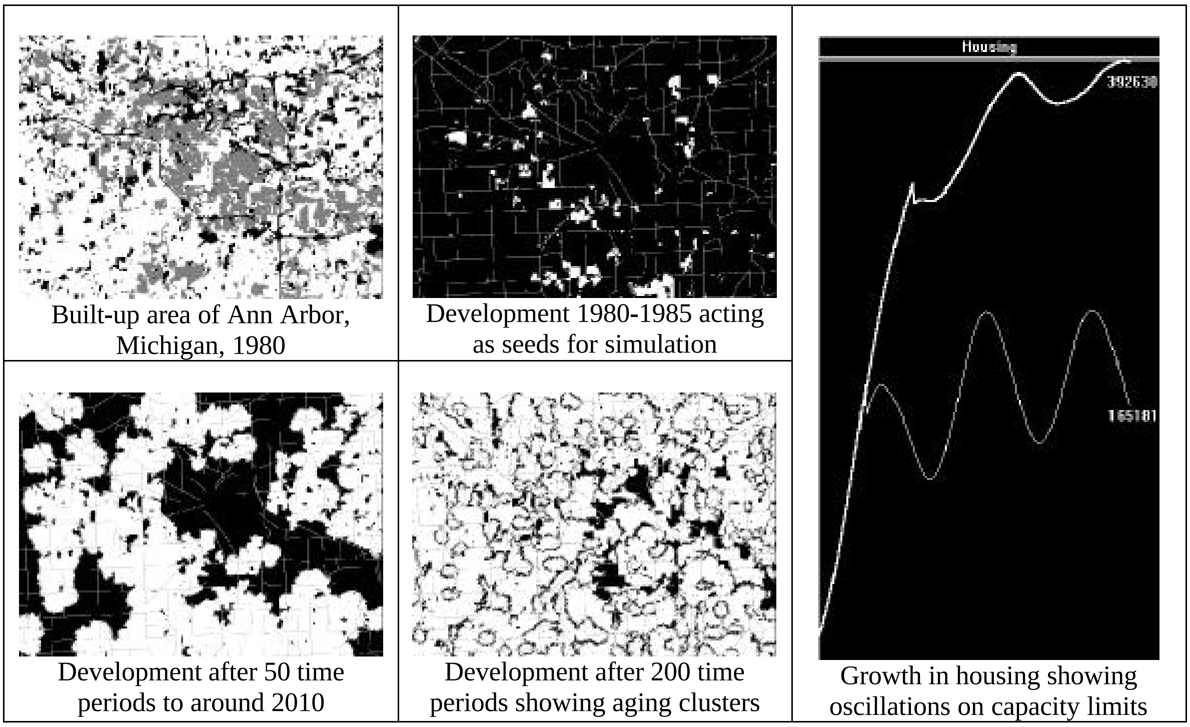
Fig 13: Samples of Growth from the Parameter Space { Φ, Γ *}*

We have not yet undertaken a systematic analysis of the fractal dimensions of the vacant sites in the samples in Figure 13 but casual calculations suggest that the distribution of ‘holes’ follows a scaling relation between frequency and size and that the boundaries of the clusters accord with space-filling lines. Within this parameter space, it is clear that there are forms that accord with many types of urban settlement and as such, this would appear a useful basis on which to begin some classification. What we propose is a thorough study of different growth situations in real cities with an emphasis on measuring speed of growth, average times of land development, and vacancy rates, thus linking our theoretical analysis to actual examples. We have not yet begun this for the purpose of this paper is to direct ideas in the study of sprawl, and to propose a framework rich enough to capture the key elements of the development process. To begin our classification, Figure 13 suggests at least a 2 x 2 set of types which define the ends of the spectrum across the two parameters: cities with low growth and low vacancy, probably characteristic of those which have grown slowly maximizing their densities and having enough time to ensure that land is used efficiently - such cities might be those that exist in some parts of Europe; cities with high growth rates and low vacancy which are even more compact which might be characteristic of those in places like China; those with high vacancy rates and low growth rates which lead to forms that are relatively compact but full of holes like cities in western Europe which have undergone some form of de-industrialization; and cities with high growth rates and high vacancy rates which sprawl more like those cities of the American west.

8 Conclusions: Application and Policy

Although this paper has been largely theoretical in tenor, we will conclude with some ideas about how these models might be focussed on more realistic applications. Moreover, there are many variants of the framework that we have barely hinted at, in particular the generalization of these models to multiple seeds and the consequent merger of urban development as land around these seeds passes through the growth process. Thus we will use the example of sprawl here to show the effects of multiple seeds. We are currently developing an operational variant of this model for the town of Ann Arbor, MI for which we have a rich database on how the region has developed over the last 20 years. To illustrate how the model might be applied, we take residential development from 1980 to 1985 and use this as the seeds to the development process which we simulate from 1985 until 1990 for which we have observed data. In our empirical work which will be reported in due course, our focus will be on “calibrating” the model over this time period and then making conditional predictions for the period 1990 to 1995 which will form an independent test of the model’s ability to predict the dynamics of urban sprawl. Here however we will simply illustrate the model’s ability to simulate the oscillations caused by the aging process of housing, with an emphasis on how "initial conditions" – historical contingency – dictates the ultimate form of any and every urban system.

We report elsewhere a detailed CA model framework – **DUEM** (**D**ynamic **U**rban **E**volutionary **M**odelling) – which we are using for simulating sprawl (Batty, Xie and Sun, 1999). This model has many more functions than the theoretical structure presented here in that several land uses are simulated, neighbourhood sizes can be varied, and the transport network is predicted as a function of land use. In general terms every element can influence every other but in this demonstration, we will restrict our focus exclusively to residential development, thus faithfully implementing the CA given in equations (23) to (26) and (29) above. In Figure 14(a), we show the urban extent of Ann Arbor in 1980 with the main highway structure, and in 14(b), residential development from 1980 until 1985, the seeds for our simulation. The size of the grid on which this simulation takes place is 595 x 824 with each pixel about 30 square meters on the ground. We have used the 3 x 3 Moore neighbourhood with no constraints on existing development, with the result that ultimately the entire space is converted to new, then established development. Prior development is part of the development process which we assume is aged at t = 0 at the outset of the simulation.



Built-up area of Ann Arbor, Michigan, 1980

Development 1980-1985 acting as seeds for simulation

Development after 50 time periods to around 2010

Development after 200 time periods showing aging clusters

Growth in housing showing oscillations on capacity limits

Fig 14: Long Term Simulation of Sprawl in the Ann Arbor Region

The aging process assumed in DUEM involves each seed or site of new development being active or new for T_N time periods and then entering a mature phase in which it no longer determines new development for T_M time periods. After the period T_N+T_M, activity is redeveloped and renters the development process immediately. There is no vacancy within this model currently (although it can be simulated through a convolution of the aging process with variations in neighbourhood size) and thus the development that results is compact. In short, all land is developed in the long term except that reserved for the street network that acts a constraint. In the same way we illustrated for the CA models in Figures 9 to 11, waves of new and established development as well as redevelopment move out concentrically around the 200 or so distinct seeds that form the initiators of the development process (Figure 14(b)). The 200 units of development which were created between 1980 and 1985 when used in the model process imply that each time unit in the model is about 6 months in length in real time. In terms of the aging process, we assume that both T_N and T_M are set as 30 time units each and this implies that sites move through the development process to redevelopment in 30 years. This might be a little short for full-fledged redevelopment although after this time, there is a high probability of some form of renovation for all residential property.

Our simulation is thus one of very long term growth, two snapshots of which we show in Figures 14(c) and (d). The first is growth after 50 time periods (about 25 years) which might reflect the situation in 2010 while the second is the simulation after 200 time periods (100 years) where the entire region has been developed and redeveloped several times. In both these scenarios, the narrow concentric bands of vacant sites show the paths of redevelopment around the initial seeds. The sole distinguishing morphological characteristic of the latter simulation is marked by these bands, thus reinforcing the notion that urban morphology is a combination of particular historical antecedents or accidents – the initial conditions – combined with the generalities of the development process. The oscillating behaviour of the development trajectories are shown in Figure 14(e) where it is clear that the oscillations are muted until the system reaches its capacity and then the waves of redevelopment become quite distinct with respect to both new and established development. The phase and amplitude of these waves of course is a function of T_N and T_M.

In our quest to simulate the dynamics of the development process in the Ann Arbor region, we are introducing many more factors into our model which will constrain its output to be more realistic as well as enable more realism to be simulated. For example, CA models do not usually contain constraints on the total activity generated for it is assumed the attraction of sites and the probabilities of occupancy combine to determine the rate of growth. In real situations however, some limits on growth must be established and we are achieving this by ensuring that the resulting distribution of development sites by size mirrors the actual scaling distribution which we observe to be stable in time.

In developing these models, we are ultimately interested in policy applications. For the set of models from which we have drawn inspiration in this paper, there are some quite well developed policy controls, the effects of which are well-known for epidemic-threshold models. For example, in countering the spatial diffusion of epidemics particularly amongst animal population, spatial controls on spread through traps of various sorts can be determined from the models. For example, in stopping the spread of rabies in foxes, Murray (1987) has shown how these models can be used to predict a belt of sufficient width across which the transmission of the disease cannot cross. Exactly the same logic can be developed in these urban models with respect to the imposition of green belts. In a sense, this is already clear from the simulations here where redevelopment causes waves of vacancy to be established. In this sense, it is clear that if the development process is to be constrained through green belts, then ‘artificial waves’ might be introduced along the lines in which ‘natural waves’ occur within such models. We have laid out the principles by which aging in the development process might be simulated and we have moved theoretical models towards the kinds of practical applications which are urgently required in an increased understanding of the problems of urban sprawl. In future research, we will take both the theoretical and practical consequences of such models further, particularly with respect to the detailed mechanisms of the development process and the kinds of urban morphology that these are able to reproduce.

References

Banks, R. B. (1994**) Growth and Diffusion Phenomena: Mathematical Frameworks and Applications**, Springer-Verlag, Berlin.

Batty, M., Xie, Y., and Sun, Z. (1999) Modeling Urban Dynamics Through GIS-Based Cellular Automata, **Computers, Environments, and Urban Systems**, **23**, 205-233.

Batty, D. (1999) The Dynamics of Strike and Protest: France, May-June, 1968, unpublished Masters Dissertation, Department of Economic History, London School of Economics, London.

Blumenfeld, H. (1954) The Tidal Wave of Metropolitan Expansion, **Journal of the American Institute of Planners**, **20**, 3-14.

Bracken, A. J., and Tuckwell, H. C. (1992) Simple Mathematical Models for Urban Growth, **Proceedings of the Royal Society of London A**, **438**, 171-181.

Bussiere, R. (Ed.) (1972) **Modeles Mathematiques de Repartition des Populations Urbaines**, Centre de Recherche d'Urbanisme, Paris, France.

Clarke, K. C. and Gaydos, L. J. (1998) Loose-Coupling of a Cellular Automaton Model and GIS: Long-term Growth Prediction for the San Francisco and Washington/Baltimore Regions, **International Journal of Geographical Information Science**, **12**, 699-714.

Cliff, A. D., Haggett, P., Ord, J. K., and Versey, G. R. (1981) **Spatial Diffusion: An Historical Geography of Epidemics in an Island Community**, Cambridge University Press, Cambridge, UK.

Cliff, A. D., Haggett, P., and Smallman-Raynor, M. (1998) **Deciphering Global Epidemics: Analytical Approaches to the Disease Records of World Cities, 1888-1912**, Cambridge University Press, Cambridge, UK.

Dendrinos, D. S., and Mullaly, H. (1985) **Urban Evolution: Studies in the Mathematical Ecology of Cities**, Oxford University Press, Oxford, UK.

Durrett, R. (1995) Spatial Epidemic Models, in D. Mollison (Editor**) Epidemic Models: Their Structure and Relation to Data**, Cambridge University Press, Cambridge, UK, pp. 187-201.

Epstein, J. M. (1997) **Nonlinear Dynamics, Mathematical Biology, and Social Science**, Addison-Wesley, Reading, MA.

Frankhauser, P. (1994) **La Fractalite des Structures Urbaine**, Collections Villes, Anthropos, Paris, France.

Hall, P. (1988) **Cities of Tomorrow: An Intellectual History of Urban Planning and Design in the Twentieth Century**, Basil Blackwell, Oxford, UK.

Jackson, K. T. (1985) **Crabgrass Frontier: The Suburbanization of the United States**, Oxford University Press, New York.

Krugman, P. R. (1993) First Nature, Second Nature, and Metropolitan Growth, **Journal of Regional Science**, **33**, 129-144.

Morrill, R. L. (1968) Waves of Spatial Diffusion, **Journal of Regional Science**, **8**, 1-18.

Morris, A. J. (1979) **History of Urban Form: Before the Industrial Revolutions**, Longmans, London.

Murray, J. D. (1987) Modeling the Spread of Rabies, **American Scientist**, **75**, 280-284.

Murray, J. D. (1993) **Mathematical Biology**, Second Edition, Springer-Verlag, Berlin.

Nivola, P. S. (1999) **Laws of the Landscape: How Policies Shape Cities in Europe and America**, Brookings Institution Press, Washington, DC.

Noble, J. V. (1974) Geographic and Temporal Development of the Plague, **Nature**, **250**, 726-729.

Raggett, G. F. (1982) Modelling the Eyam Plague, **Bulletin of the Institute of Mathematics and Its Applications**, **18**, 221-226

Richardson, L. F. (1941) Mathematical Theory of Population Movement, **Nature**, **148**, 357.

Toffoli, T. and Margolus, N. (1987) **Cellular Automata Machines: A New Environment for Modeling**, MIT Press, Cambridge, MA.

Ward, D. P., Murray, A. T., and Phinn, S. R. (1999) A Stochastically Constrained Cellular Automata Model of Urban Growth, submitted to **Computers, Environments and Urban Systems**.

Xie, Y. (1996) A Generalized Model for Cellular Urban Dynamics, **Geographical Analysis**, **28**, 350-373.

Zhang, W. B. (1988) The Pattern Formation of an Urban System, **Geographical Analysis**, **20**, 75-84.