1
Spatial Simulation Models: What? Why? How?

It is easy to see building and using models as a rather specialised process, but models are *not* mysterious or unusual things. We routinely use models in everyday life without giving them much thought, if any at all. Consider, for example, the word 'tree'. We may not exactly have a 'picture in our heads' when we use the word, but we could certainly oblige if we were asked to draw a 'tree'. The word is associated with some particular characteristics, and we all have some notion of the intended meaning when it is used. In effect, everyday language models the world, using concrete nouns, as a wide variety of categories of thing: cats, dogs, buses, trains, chairs, toothbrushes and so on. We do this because if we did not, the world would become an unfathomable mess of sensory inputs that would have to be continually and constantly untangled in order to accomplish even the most trivial tasks.

If you are reading this book, then you are already well-versed in using models in the language that you use everyday. We define scientific models as simplified representations of the world that are deliberately developed with the particular purpose of exploring aspects of the world around us. We are particularly concerned with *spatial simulation models* of real world systems and phenomena. Our aim in this book is to help you become as comfortable with *consciously* building and using such models as you are with the models you use in everyday language and life.

This aim requires us to address some basic questions about simulation models:

- What are they?
- Why do we need them and use them?
- How can (or should) we use them?

Spatial Simulation: Exploring Pattern and Process, First Edition.
David O'Sullivan and George L.W. Perry.

It is clearly important in a book about simulation models and modelling to address these questions at the outset, and that is the purpose of this chapter.

The views we espouse are not held by every scientist or researcher who uses models in their work. In particular, we see models as primarily exploratory or heuristic learning tools, which we can use to clarify our thinking about the world, and to prompt further questions and further exploration. This view is somewhat removed from a more traditional perspective that has tended to see models as primarily predictive tools, although there is increasing realisation of the power of models as heuristic devices. As we will explain, our view is in large measure a product of the types of system and types of problem encountered in the social and environmental sciences. Nevertheless, as should become clear, this perspective is one that has relevance to simulation models as they are used across all the sciences, and becomes especially important when scientific models are used, as increasingly they are, to inform critical decisions in the policy arena.

After dealing with these foundational issues, we briefly introduce probability distributions. Our goal is to show that highly abstract models, *which make no claim to realism*, may nevertheless still be useful. It is also instructive to realise that probability distributions are actually models of a specific kind. Understanding the strengths and weaknesses of such models makes it easier to appreciate the role of more detailed models that take realism seriously and also the costs borne by this increased realism. Finally, we end the chapter by making a case for the more complicated dynamic, spatial simulation models that are the primary focus of this book.

1.1 What are simulation models?

You may already have noticed that we are using the word ‘model’ a great deal more than the word ‘simulation’. The reason for this will become clear shortly, but in essence it is because models are a more generic concept than simulations. We consider the specific notion of a simulation model in Section 1.1.5, but focus for now on what models are.

The term *model* is a difficult one to pin down. For many, the most familiar use of the word is probably with reference to architectural or engineering models of a new building or product design. Until relatively recently, most such models were three-dimensional representations constructed from paper, wood, clay or some other material, and they allowed the designer to explore the possibilities of a proposed new building or product before the expensive business of creating the real thing began. Such ‘design models’ are often built to scale, necessitating simplification of the full-size object so that the overall effect can be appreciated without the finer details becoming too distracting. Contemporary designers of all kinds generally build not only *physical models*

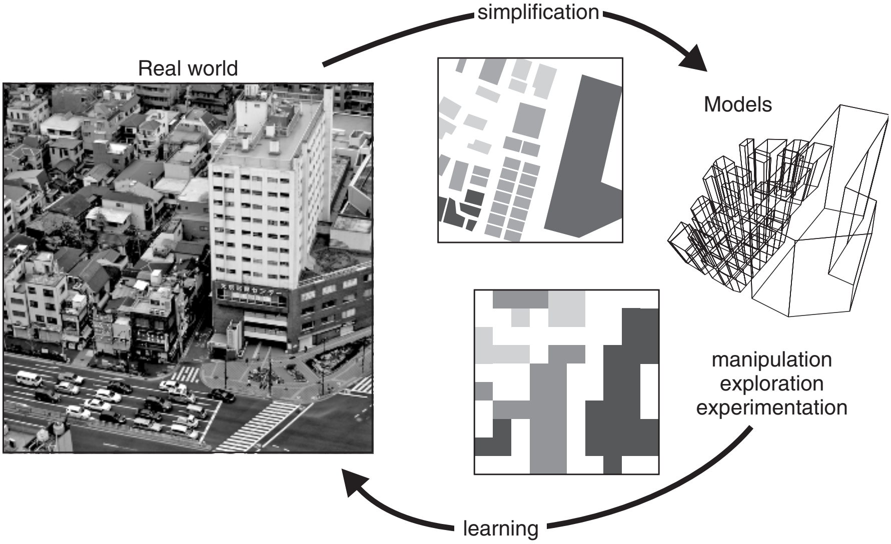

Figure 1.1 Schematic illustration of the concept of models. Models simplify the real world, enabling manipulation, exploration and experimentation, from which we aim to learn about the real world. Photograph from authors' collection.

but *computer models*, using computer-aided design (CAD) software to create virtual models that can be manipulated and explored interactively on screen. Design models then, are simplified representations of real objects that are used to improve our understanding of the things they represent. The underlying idea of model building of this kind is shown in Figure 1.1. An important idea is that more than one model is likely to be useful.

Scientific models perform a similar function—and follow the same general logic of Figure 1.1. Therefore, for our purposes, we define a scientific model as

> *a simplified representation of a system under study, which can be used to explore, to understand better or to predict the behaviour of the system it represents*

The key term in this definition is 'simplified'. In most scientific studies there are many details of the phenomena at hand that are irrelevant from the particular perspective under consideration. When we are tackling the transport problems of a city, we focus on aspects that matter, such as the relative allocation of resources for building roads relative to those for public transport infrastructure, the connectivity of the network and how to convince more people to car-pool. We do not concern ourselves with the colours of the cars, the logos on the buses or the upholstery on the subway seats. At the level at which we are approaching the system under study some components matter and others are irrelevant and may be safely ignored. The process of model development demands that we simplify from the often bewildering complexity of the real world by deciding what matters (and what does not) in the context of the current investigation. An important consequence of this simplification process, as George Box succinctly points out, is that, '[m]odels, of course, are never true' (Box, 1979, page 2). Luckily, as Box goes on to say, 'it is only necessary that they be useful'.

1.1.1 Conceptual models

The first step in any modelling exercise is the development of a *conceptual model*. *All* scientific models are conceptual models, and a particular conceptual model can be given more concrete expression as any of the distinct types discussed below. Thus, developing a conceptual model is fundamental to the development of any scientific model. Approaching the phenomenon under study from a particular theoretical perspective will bring a variety of abstract concepts into play, and these will inform how the system is broken down into its constituent elements in systems analysis.

In simple cases, a conceptual model might be expressible in words ('if parking costs more, fewer people will drive'), but usually things are more complicated and we need to consider breaking the phenomenon down into simpler elements. *Systems analysis* is a method by which we simplify a phenomenon of interest by systematically breaking it down into more manageable elements to develop a conceptual model (see Figure 1.2). A critical issue is the desired level of detail. In the case shown, depending on our interests, a forest might be simplified or abstracted to a single value, its total biomass. A more detailed model might break this down into the biomass stored in trees and other plant species, with a single submodel representing how both categories function, the difference between trees and other plants being represented by differences in attribute values. A still more detailed analysis might consider individual species and develop submodels for each of them. The most appropriate model representation is not predetermined and will depend on the goals of the model-building exercise. In this case, a focus on carbon budgets may mean that the high-level 'biomass only' model is most appropriate. On the other hand, if we are concerned about the fate of a particular plant species faced with competition from invasive weeds, then a more detailed model may be required.

In the systems analysis process, we typically break a real-world phenomenon or system down into general elements, as follows:

Components are the distinct parts or entities that make up the system. While it is easy to say that a phenomenon can be broken down into components, this step is critical and typically difficult. The components chosen will have an impact on the resulting model's behaviour so that these basic decisions

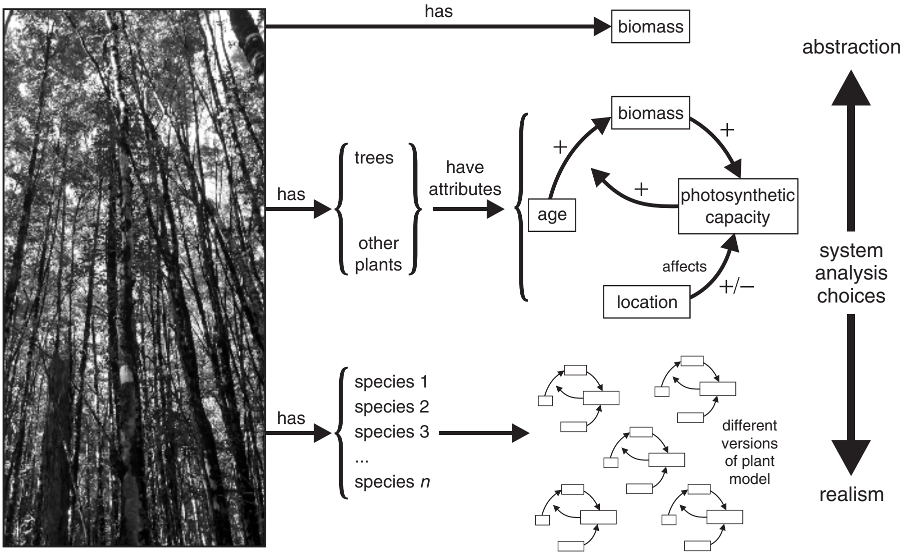

Figure 1.2 The systems analysis process. A real-world phenomenon is broken down into components, their attributes, how they interact with one another and how they change via process relationships. A particular phenomenon might be represented and analysed in a variety of ways, with the desired level of realism or, conversely, abstraction a key issue.

are of fundamental importance to how adequate a given representation will be. An important assumption of the systems approach is that the behaviour of the components in isolation is easier to understand than that of the system as a whole.

State variables are individual component or whole-system level measures or attributes that enable us to describe the overall condition of the system at a particular point in space or moment in time. Total forest biomass might be such a variable in Figure 1.2.

Processes are the mechanisms by which the system and its components make the transition from one state to another over time. Processes dictate how the values of the involved components' state variables change over time.

Interactions between the system components. In most systems not all components interact with each other, and how component interactions are organised is an important aspect of a system's structure. In many of the systems which interest us in this book, interactions are more likely and/or stronger between components that are near one another in space, making a spatially explicit model desirable.

Thus, a conceptual model of a system will consist of components, state variables, processes and interactions, and taken together these provide a simplified description of the phenomenon that the model represents.

A common way to represent a model is using 'box and arrow' diagrams, a format that will be familiar from numerous textbooks. Interactions between system components are represented as arrows linking boxes. We might add + or − signs to arrows to show whether the relationship between two components is positive or negative (as has been done in Figure 1.2). It is then only a short step from the conceptual model to a *mathematical model* where component states are represented by *variables* and the relationships between them by mathematical equations. As with design models, the advent of widely available and cheap computing has seen the migration of such mathematical models onto computers. The relationships and equations governing the behaviour of a model are captured in a *computer model* or *simulation model*. The simulation model can then be used to explore how the system changes when assumptions about how it works or the conditions in which it is set are altered. In practice the most appropriate way to represent a system will depend on the purpose of the modelling activity, and so there are many different model types. Although this book's focus is on spatially explicit mathematical and simulation models, it is important to recognise that many different approaches to modelling are possible. We consider a few of these in more detail in the sections that follow.

As we have already suggested, conceptual models can be represented in a variety of ways. Simple models may be described in words perfectly satisfactorily. Newton's three laws of motion provide a good example. The third law, for example, can be stated as: 'for every action there is an equal and opposite reaction'. Even very complicated models *can* be described in words, although their interpretation may then become problematic, and it is debatable how sensible it is for such models to remain as *solely* verbal descriptions. Nevertheless, in many areas of the social sciences, verbal descriptions remain the primary mode of representation, and extremely elaborate conceptual models are routinely presented in words alone (give or take the occasional diagram), at book length. The work of political economists such as Karl Marx and Adam Smith provides good examples. Partly as a consequence, the interpretation of these theories remains in dispute, although it is important to acknowledge that the subsequent mathematisation of economic theory through the twentieth century has not greatly reduced the interpretative difficulties: however we represent them, complicated models remain complicated and open to interpretation.

Even so, verbal descriptions of complicated conceptual models have evident limitations. As a result many conceptual models are presented in graphical form, with accompanying explanation. As we have noted, it is a short step from graphical representation to the development of mathematical models, and the success since Newton's time of those physical sciences which adopted mathematical modelling as a tool has been a persuasive argument in other disciplines for the adoption of the approach.

1.1.2 Physical models

We began by briefly touching on the three-dimensional scale models often used by design professionals. *Physical* or *hardware models* are also occasionally used in the sciences, and in engineering. Wave tanks can be used to simulate coastal erosion and wind tunnels to investigate turbulence around aerofoils or the aerodynamics of vehicles. Hardware models can provide guidance about a system's behaviour when that system is not sufficiently understood for mathematical or computational models to provide reliable guidance. Careful scaling of the model system's properties or extrapolation from the model to the target system is necessary for this approach to work well. For example, the grain size of sand in a hardware model of a beach must be carefully considered in combination with the wave heights and speeds if the results from a flume are to be applicable to real beaches. Rice et al. (2010) provide a good summary of the potential value of this approach in the specific context of river science, where they show how it can be used to relate stream hydraulics to stream ecology.

A highly specific kind of 'hardware' model is the use of laboratory animals in medical research, where mice or rats or fruit flies are considered 'model' organisms for aspects of animal biology in general. Similar to flumes or wind tunnels, this use of models recognises that our ability to mathematically model whole organisms remains limited at present, and for the foreseeable future, and these models provide an alternative. Even setting ethical issues to one side, the difficulties of generalising from findings based on such models are apparent.

An interesting hardware model of a different kind is provided by the hydraulic model of a national economy built by the economist Bill Phillips while he was a student at the London School of Economics (Fortune Staff, 1952). The so-called MONIAC (MOnetary National Income Analogue Computer) used a system of reservoirs and pipes to represent flows of money circulating in a national economy. Flow rates in different pipes in the system were adjusted to represent factors such as the rate of income tax. Several MONIAC machines were built and working examples are maintained by the Science Museum in London and by the Reserve Bank of New Zealand (Phillips was a New Zealander). While this may seem a strange way to model an economy, Phillips would have been well aware that his hydraulic model was a representation of an underlying conceptual and mathematical model.

1.1.3 Mathematical models

The MONIAC example points us towards *mathematical models*, the most widely adopted approach to scientific modelling. In a mathematical model, the state of the system components is represented by the values of *state variables* with equations describing how the state variables change over time. Newton's Second Law of Motion is a straightforward example, where the equation

$$F = m\frac{\mathrm{d}v}{\mathrm{d}t} \tag{1.1}$$

tells us that the velocity v of an object changes at a rate proportional to the net force on it, F, and in inverse proportion to its mass, m. Since Newton's time, many laws of physics governing the basic behaviour and structure of matter have been found to be well approximated by relatively simple mathematical equations not much more complicated than Newton's Second Law. Perhaps the best-known example is Einstein's celebrated $E = mc^2$ concerning the relationship between energy, E, mass, m, and the speed of light, c.

Such equations are simple mathematical models but taken in isolation they do not tell us much about the dynamic behaviour of systems. If we decompose a system into a number of interacting components, and equations representing the interactions can be established, then we have a mathematical model in the form of a system of simultaneous equations describing the system's state variables. This mathematical model can be used to explore how the state variables will change over time and through space, and how different factors affect overall system behaviour. How the necessary equations are determined depends on the nature of the system. The most productive approach over time has usually been the *experimental method*, where the relationships between different variables are determined by setting up laboratory experiments in which key system variables are carefully manipulated while others are held fixed, so that the independent effects of each can be determined. Experiments can be deliberately constructed to explore expected or hypothesised relationships between system variables. When the expected relationships are found not to hold, hypotheses are adjusted and new experiments designed, and over time a clear picture of the relationships between system variables emerges.

1.1.4 Empirical models

In an experiment, we artificially close the system under study so that only the single effect of interest is 'in play'. However, many systems, such as ecosystems or social systems, cannot be experimentally closed. In these situations, classic experimental methods are problematic. An alternative approach is to make empirical observations of such systems and to use quantitative methods to construct an *empirical model* of the relationships among the system variables. Such models are generally statistical in nature, with regression models the most widely used technique. An example of this approach is hedonic price modelling of real estate markets, where the sale price of housing is modelled as an outcome of a collection of other variables, such as floor area, date of construction, number of bedrooms and so on (Malpezzi 2008, provides a review of the approach). A statistical empirical model might show that other things being equal the sale price of larger houses is higher than that of smaller houses.

In empirical models, meaningful interpretation of observed statistical relationships can be challenging. In some cases, interpretations will be obvious and well-founded (for example larger houses attract higher prices because people are happy to pay more for extra space) but in others the mechanisms will be less obvious and interpretation may be controversial. It is particularly challenging to handle *interaction effects* among variables. In the case of house prices, an interaction effect would be an association between proximity to the central city and land parcel sizes. This might make it appear that small parcels are more expensive than larger parcels, when it is actually the more central location of smaller parcels that leads to the observed relationship between parcel size and price.

There are many examples of empirical models where interpretation is problematic, particularly in the social sciences. For example, what are we to make of findings relating different rates of gun-ownership in various US states to crime rates? Many interpretations are possible (and are offered), and the observed empirical relationships can only be linked back to underlying processes via long chains of argument and inference, which are not as readily tested as is possible in the experimental sciences. Such concerns have led to considerable dissatisfaction with empirical models (see, for example, Sayer, 1992, pages 175–203). The root of the problem is that empirical models do not consider the mechanisms that give rise to the observed relationships.

Nevertheless, in cases where the causal mechanisms remain poorly understood, empirical models can provide useful starting points for the development of ideas. We have a more or less stable empirical relationship: the question is why? As a result, and also due to the numerous statistical tools available for their development, empirical models are the most common type of model across the sciences, other than the ubiquitous conceptual model.

1.1.5 Simulation models

Whether experimentally, empirically or theoretically derived, quantitative relationships among the state variables in a system can be used to build a *simulation model*. In a simulation model, a computer is programmed to iteratively recalculate the modelled system state as it changes over time in accordance with the relationships represented by the mathematical and other relationships that describe the system.

The most widely used simulation models are *systems dynamics models*. Systems dynamics models consist of essentially three types of variable: stocks, flows and parameters:

Stocks are system variables that represent quantities stored in the system over time. Examples are the energy in a system, carbon in storage pools in the carbon cycle, money in the economy, population in a demographic model (as in Figure 1.3, which consists of linked population models) or workers in an international labour market.

Flows represent the movement of stock variables among different parts of the system. In a model of the carbon cycle, carbon moves from being stored in the ground into the atmosphere dependent on the rate at which fossil fuels are being used, among other pathways. In a population model, flows represent births 'flowing' into the population or deaths leaving the population.

Parameters (also called auxiliary variables) describe the overall state of the system and govern the relationships among stocks and flows, sometimes via several intervening parameters. In a carbon cycle model, the price of petrol, the propensity of consumers to drive and the rate of carbon taxation might be model parameters.

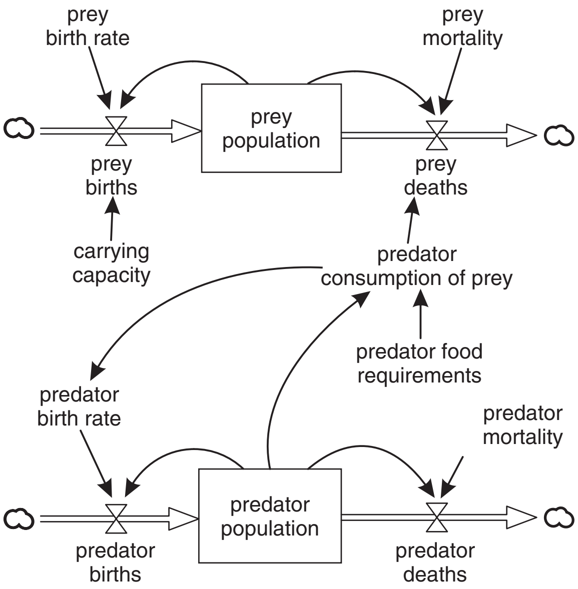

Figure 1.3 A simple systems dynamics model of the interaction between predator and prey species. Boxed variables are stocks, double-width arrows with 'valve' symbols are flows, clouds represent everything outside the system, and other named items are parameters. Arrows indicate where a relationship exists between variables.

Systems dynamics models are effectively computer simulations of mathematical models consisting of interrelated differential equations, although they often also include empirically derived relationships and semi-quantitative 'rules'. The predator–prey model in Figure 1.3 is a typical, albeit simple, example. Note that many of the probable relationships in a real predator–prey system are *deliberately* omitted from the model, for example the predator might have other potential food sources whose availability affects the level of predation. Phillips's MONIAC model referred to in Section 1.1.3 was a systems dynamics model of a national economy simulated using reservoirs and plumbing rather than a computer. The diagrammatic representation of systems dynamics models (see Figure 1.3) closely reflects the basis of such models in mathematical models of physical systems.

While systems dynamic models are among the most widely used type of simulation model, many other model types and variations are possible. Of particular interest in this book are *spatially explicit* models, where we decide that representing distinct spatial elements and their relationships is important for a full understanding of the system. Thus, a spatially explicit model has elements representing individual regions of geographical space or individual enities located in space (such as trees, animals, humans, cities and so on). Each region and entity has its own set of state variables and interacts continuously through time with other regions and entities, typically interacting more strongly with those objects that are nearest to it. These interactions may be governed by mathematical equations, by logical rules or by other mechanisms, and they directly represent mechanisms that are thought to operate in the real-world system represented by the simulation.

The simplest way to think of the relationship between models and simulations is that simulations are implementations, usually in a computational setting, of an underlying conceptual or mathematical model. Simulation is often necessary because detailed analysis of the model is difficult or impossible, or because simulation allows convenient exploration of the implications of the model. Because any simulation implements an underlying model, most of the time in this book we discuss models rather than simulations. In each case discussion of the models and understanding of their properties is only possible because we have built a simulation of the model.[*] Winsberg (2009a) provides a useful account of the role of simulation in science that clarifies both the distinction and the close relationship between models and simulations.

[*] Where the marginal turtle icon appears, as shown here, we have provided simulations of most of the models discussed in the text for readers to explore at http://patternandprocess.org

1.2 How do we use simulation models?

Some of the reasons for using simulation models have already been mentioned in passing. Here we expand on these ideas and present the perspective on the role of simulation models in research in which this book is grounded. The critical issue is that not all models are created equal. Figure 1.4 shows in schematic form how two critical aspects of any phenomenon, the data we have available and how well we understand it, affect our ability to use simulation models for different purposes.

Systems for which we have reliable, detailed data and whose underlying mechanisms we understand well are the ones most suitable for predictive modelling. Engineered systems such as bridges, cars, telephones and so on are good examples. We have a thorough understanding of how they work, and we have abundant and reliable data based on previous experience with using such systems. In this context, using models for prediction is worthwhile. However, in many cases we lack reliable, detailed observational data, even if we have a reasonable grasp of the underlying processes. For example, animal population dynamics are relatively well understood in general terms, but detailed rich datasets are scarce, especially for long-lived organisms. In this context, prediction is more problematic, and the best use for models may be to inform us about where critical data shortages lie. On the other hand, there are an increasing number of areas where we now have access to detailed data, but still have a poor understanding of how the system works, global financial markets being a prime example (May et al., 2008). Here, using models to help in developing theories may be appropriate. Finally, there are many fields where both the available data and current levels of understanding are poor. Here, the best prospect may be to use models as devices for learning, which

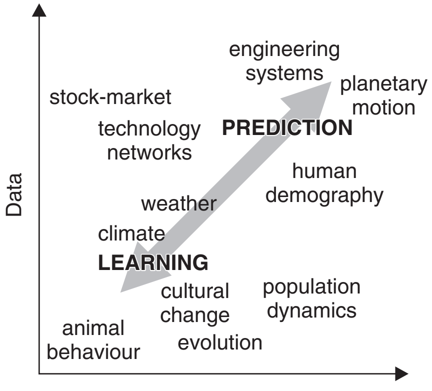

Figure 1.4 How data and understanding affect the use of models, with indicative fields of study (after Starfield and Bleloch, 1986). Note that the locations of different fields in the figure are suggestive only. See text for details.

may inform the development of theory and also steer us towards the areas of most pressing need for better data.

1.2.1 Using models for prediction

It is generally not long after a simulation model is presented to an interested audience before someone asks something like: ‘So what is Auckland going to look like 25 years from now?’ or ‘So are polar bears going to survive climate change?’ or any number of other questions about the future that are of interest. It is understandable that researchers, policy-makers and the public in general are interested in using simulation models to predict the future. After all, ‘real’ sciences make reliable predictions. Arguably, it was the advent of reliable predictions of the tides, seasons and planetary motion that began the human species’ rise to dominance of life on Earth. Added to this, the ability of engineers to confidently predict the behaviour of the systems they design has enabled rapid technological progress over the last few centuries. Both kinds of prediction ultimately rely on scientific models. But there are good reasons to be wary of seeing models primarily as predictive tools.

We have already suggested that lack of data or of good understanding of the field in question are reasons for being cautious about prediction. If there are deficits in either dimension, it is quite likely that prediction will do more harm than good, although in such cases, it may not be prediction *per se* that is the problem. Placing unwarranted faith in inherently uncertain predictions is as likely to lead to ill-founded decision-making as having no predictive tools available at all. More circumspect approaches that consider alternative scenarios and possible responses to allow better-informed decision-making are usually a more appropriate way forward than blind faith in predictive models (see Coreau et al., 2009, Thompson et al., 2012). We consider aspects of the uncertainty around model outcomes in more detail in Chapter 7.

It is also important to be wary of the prospects for prediction in contexts where although individual elements are well understood, the aggregation of many interacting elements can lead to unexpected nonlinear effects. Perhaps the most important examples in this context are technological networks of various kinds. Here, individual system elements are designed objects, whose individual behaviour is well known. In many such networks, we also have good data on overall system usage and patterns. Nevertheless the nonlinear systems effects discussed in Section 1.3.2 have the capacity to surprise us, and there is an important role for modelling in improving our understanding of such systems (see, for example, O’Kelly et al., 2006).

1.2.2 Models as guides to data collection

Where the most pressing deficits in our knowledge of systems are empirical, so that we lack good observational data about actual behaviour, the most important role for models may be to help us decide which data are most critical to improving our understanding. Assuming that adequate models of the underlying processes in such situations can be developed, then we can perform *sensitivity analysis* (see Section 7.3.2) on the models to find out which parameters have the strongest effect on the likely outcomes. Once we know which data are most essential to improving our ability to forecast outcomes more reliably, we can direct our data collection energies more effectively.

Good examples of this sort of application of models are provided by animal conservation efforts. For endangered animals we may not have reliable data on their populations, but we do have good general frameworks for modelling population dynamics (for example Lande et al., 2003). Applying such models to endangered species with best guess estimates of critical parameters and the associated uncertainties may help conservation workers to target future data collection on the critical features of the system, and also to design improved management schemes. Unfortunately, population models applied to fisheries over recent decades provide depressing evidence of how important this is: arguably too much faith in the predictive capacity of models, and ignoring uncertainties in the models and the data used to inform them, has left many fisheries in a worse state than they may otherwise have been (see Clover, 2004, Pilkey and Pilkey-Jarvis, 2007).

1.2.3 Models as 'tools to think with'

In some cases there is no shortage of empirical data, but we have limited understanding of how a system works. The advent of sensor technologies of various kinds—remote-sensing platforms being the most obvious, but more recently including phenomena such as volunteered geographic information (Haklay et al., 2008)—has created a situation where we may have vast amounts of empirical data (Hilbert and López, 2011), but no clear idea of what it tells us. While data-mining methods may offer some relief in these situations (Hilbert and López, 2011) and provide at least candidate ideas about what is going on, science without theory is an unsatisfactory approach.

In this situation, simulation models can become 'tools for thought' (Waddington, 1977). In this approach we use simulation models that represent whatever working hypotheses we have about the target system and then *experiment on the model* to see what the implications of those theories and hypotheses might be in the real world (see Dowling, 1999, Edmonds and Hales, 2005 and Morrison, 2009). This turns the traditional hypothetico-deductive approach to science on its head. In deductive science, we make observations of the world and then evaluate hypotheses that could explain those observations. We then make controlled experimental observations to decide if our hypotheses provide adequate explanations of the observations. When we experiment on models by simulation, we effectively explore what the implications of various possible hypotheses would be. Observation of model behaviours and comparison of those behaviours with real-world observations of the target system may then allow us to dismiss some hypotheses as implausible and so refine our theories. Some authors refer to this approach as science *in silico* (see, for example, Epstein and Axtell, 1996, Casti, 1997).

This is not an easy approach. In particular, it is beset by the *equifinality* problem, which refers to the fact that large numbers (in fact an infinite number) of possible models, including the same model with different input data, could produce behaviour consistent with observational data. In other words identifying the processes responsible for a given pattern, on the basis of that pattern alone, is problematic. This makes problems of model selection and inference acute for anyone adopting this approach. We consider some of these issues in more detail in Section 7.6, where we consider *pattern-oriented modelling*, which may be particularly suited to making inferences with spatially explicit models (Grimm et al., 2005, Grimm and Railsback, 2012).

In sum, the proper role of simulation models in advancing our understanding of the world and the consequent variety of their uses is complicated and heavily dependent on the nature of the systems under study and on the current state of empirical and theoretical knowledge about them. As is often the case, it is sensible to be flexible and pluralistic in outlook, using models as one tool among many in a mixed methods approach.

1.3 Why do we use simulation models?

Simulation models are a relatively recent addition to the scientific toolbox, and the reasons for their widespread adoption call for some explanation. Broadly speaking, we believe there are three reasons that scientific research based on simulation modelling has become so prevalent in recent decades. The first is obvious and quickly dealt with: simply that this approach has now become possible. Cheap, widely available computing, although it appears commonplace from the vantage point of the early twenty-first century, is a historically recent development. The electronic computer was invented towards the end of the Second World War, and the desktop personal computer only appeared towards the end of the 1970s, and was not widely available until a decade after that. In the same way that widely available computing power has changed society generally, we might also expect it to have changed aspects of scientific practice, and simulation modelling is an example of such change.

However, given the success of science *before* simulation models, a better answer to the question ‘Why use simulation models?’ than ‘Because we can!’ is surely necessary. This leads us to two further reasons: the difficulty of experimental science in many fields of enquiry and the realisation that many systems are nonlinear.

1.3.1 When experimental science is difficult (or impossible)

The experimental method of science has proved to be extremely successful at systematically extending our understanding of the world. However, it is not always recognised that the success of the experimental method relies on certain features of the systems under study.

Some systems of interest to social and environmental scientists are not amenable to experiment for rather prosaic reasons. Archaeologists, anthropologists and palaeontologists are interested in systems that no longer exist. Climate scientists are interested in a system that is effectively the whole of the terrestrial, atmospheric and oceanic system, and hence not controllable in any meaningful sense. In many fields, such as forest ecology or geology, the time horizons of interest are too long for experiments to be practical.

Other limitations to experiment relate to the nature of the systems themselves. The most obvious consideration is that experiments only produce conclusive evidence when they are *closed systems*. Laboratory experiments go to great lengths to control all the factors that might have an effect on the phenomenon under study. For example, chemistry experiments must strictly control the temperature and pressure at which reactions are carried out, as either is likely to affect the outcome of any chemical reaction. In specific cases, many other factors will have to be controlled, such as the concentrations of the various reagents and the presence of even trace amounts of any impurities. The benefit of these efforts in experimental set-up is that it is clear when the experimenter varies a particular system input of interest that any response from the system is a result of that change. A laboratory experiment is thus a tightly controlled way to investigate the mechanisms in a system.

Such tight control is only possible for certain kinds of science. In many systems of interest in the environmental and social sciences system closure is impossible, impractical or unethical. A fragile ecosystem cannot easily be sealed off from external influences and have its climate controlled so that we can investigate the effects of a pesticide. It is generally not considered ethical to experiment with different education policies in different parts of a city to see which works, because the impact of potentially poor policy will have to be borne by some of the children going through the school system at that time. Even if such experiments were considered ethical, it is impossible to close off the education system from external variables such as the degree of home involvement in education or differences in the income, culture, language and ethnicity of the households in each school, any of which might have an equal or greater effect on outcomes than the policies being tested.

Drug trials are among the most carefully designed and constructed observational studies in the non-experimental sciences. Properly conducted trials involve an elaborate double-blind case-control methodology where a control group of patients are administered with placebo or 'sugar pill' treatment rather than the treatment under investigation. Neither the patients nor those administering the treatment know if they are part of the control group or the treatment group. Control and treatment groups are carefully selected to balance their characteristics in terms of the many demographic and physiological factors that might affect the outcome. Even so, interpretation of the results of such trials remains difficult and controversial.

It is rarely possible for field-based scientists to control the observational setting as closely as in medical trials. A fire ecologist may be able to set up (at great expense and with considerable difficulty) a study where control and treatment sites are matched and subjected to prescribed burns in order to investigate the effects of wildland fires (one example is the Kapalga savanna fire experiment in northern Australia described by Andersen et al., 2003). However, the number of observations that can be made in this way is severely limited, and the control and treatment sites are not identical—they are, after all different sites!—so that determining which differences in outcome are attributable to the treatment and which are variation in the outcomes that might have occurred anyway is extremely difficult. This is not to say that such experiments are not valuable, but that they are limited in fundamental ways that model-based approaches can possibly circumvent.

Occasionally *natural experiments* arise that provide brief windows when something like experimental results can be obtained for environmental systems. For example, when all air traffic over the United States was suspended for a period after the events of 11 September 2001, it provided an opportunity for climate scientists to collect data on the effect of aircraft contrails on the daily temperature ranges (Travis et al., 2002). But such opportunities are rare, they are not truly controlled and the observational sciences would make slow progress indeed if they relied solely on this approach.

Another possibility, rather than hoping to get lucky or trying to control the field *in situ*, is to deliberately set up laboratory experiments in the social or environmental sciences. The difficulty then becomes deciding how relevant to real systems are the findings from such artificial settings. Indeed, it can be readily argued that laboratory experiments in such cases are best understood as simplified models of the real world (see, for example, Diamond, 1983, Carpenter, 1996), so that the findings must be treated with caution.

In short, the inferential lot of observational field scientists compared with their laboratory colleagues is not a happy one. One option that simulation modelling opens up is to experiment not on real systems, but on virtual ones. If an adequate simulation model of the system of interest can be developed, then hypotheses about how the system behaves in different circumstances can potentially be investigated in a controlled manner. This approach is sometimes referred to as *surrogative reasoning* because the model is a surrogate for the real-world system. Of course, there are serious difficulties with this approach.

How can we know if the model is an adequate representation of the system? How can we map findings from experimenting with a model to conclusions about the real world? These are certainly not trivial considerations, and we consider them in more detail in Chapters 2 and 7. For now, it is sufficient to note that the potential of this approach is a major impetus for the growth of simulation modelling across various fields.

1.3.2 Complexity and nonlinear dynamics

The 'game of life' automaton An informative way to approach complex systems and complexity is by means of one of the most celebrated examples: John Conway's 'game of life' cellular automaton. *Cellular automata* (CA) are a standard type of spatially explicit simulation model, and we will encounter many examples in this book (see particularly Chapter 3). A CA consists of a *lattice* of *cells*, which defines for each cell those other cells that are its *neighbours*. The most obvious spatial lattice is a two-dimensional grid of square cells, with neighbours defined either as the four immediately adjacent orthogonal neighbours or as the eight immediately adjacent neighbours, including the diagonals. Given the lattice structure, the full CA definition includes a set of *states* for each cell—most simply 'on' or 'off'—and a set of *transition rules* that determine how the state of each cell changes from one time step to the next, based on the current state of a cell and those of its neighbours.

A brief history of cellular automata is presented by Chopard and Droz (1998). CAs were proposed in the 1940s by John von Neumann in an attempt to devise a self-replicating machine—that is, a machine capable of creating copies of itself. Von Neumann's solution (described by Burks, 1970) was an elaborate CA with a lattice of around 200 000 cells, and 29 distinct cell states. Although von Neumann proved that a self-replicating machine was possible, his solution is so complicated that even up to the present it remains a historical curiosity that has not been implemented (although Pesavento, 1995, describes a partial implementation).

These unpromising beginnings changed dramatically with the advent of John Conway's game of life (as described in Gardner, 1970). Conway was intrigued by the question of how simple he could make the rules of a CA and still get 'interesting' behaviour. Thus, unlike the examples we consider in later chapters, the game of life is not a model of any specific system or entity. Rather, it is a mathematical system whose purely theoretical interest lies in the relationship between the intricacy of the rules that define a system's behaviour and the richness of that behaviour. The game of life CA is defined as follows:

- The lattice is a two-dimensional grid, theoretically infinite but in practice as large as needed.

- Cell neighbours are the eight immediately adjacent orthogonal and diagonal grid cells.
- Cell states are ‘alive’ or ‘dead’.
- There are two transition rules:
  birth a dead cell is born if it has three live neighbours, otherwise it remains dead
  survival a live cell survives if it has two or three live neighbours, otherwise it dies.

The best way to appreciate the game of life CA is to experiment with a computer simulation. It does not take long to discover that the simplicity of the rules belies a rich array of dynamic behaviour. A flavour of this is provided by examining the small patterns in Figure 1.5 (see also Gardner, 1970, who presents a similar diagram). The smallest patterns, (a) and (b), unsurprisingly die immediately. Adding another live cell to produce the ‘corner’ pattern (c) results in a four cell block of live cells that is stable. Three live cells in a line (d) is a blinking pattern that switches each time step between a horizontal and vertical line. Adding one more cell to (d) to give the ‘T’-shaped pattern (e) produces a sequence of nine time steps resulting in four copies of the three cell blinker pattern (d). 1.1

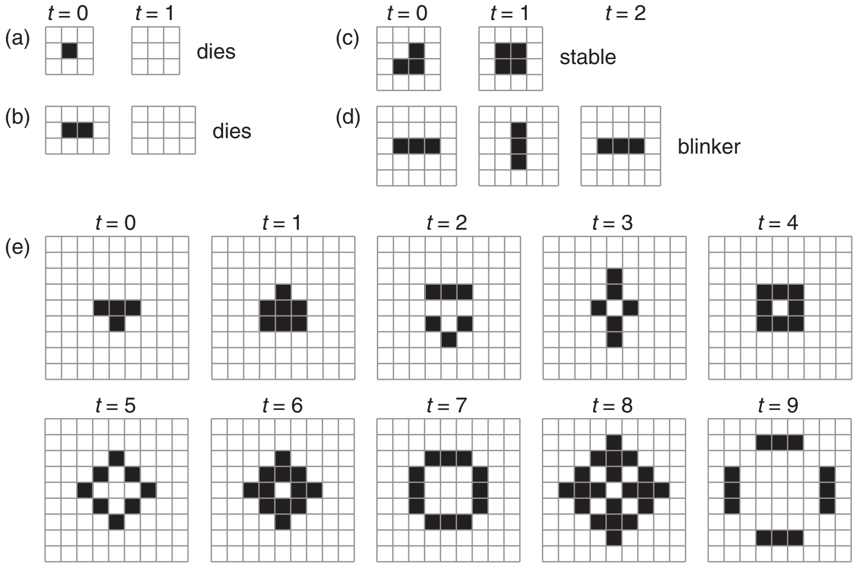

Figure 1.5 Some simple life patterns. Patterns (a) and (b) die immediately, pattern (c) produces a stable ‘block’, pattern (d) results in a ‘blinker’ that switches indefinitely every time step between two configurations and pattern (e) results in four copies of pattern (d) which blink in the same way.

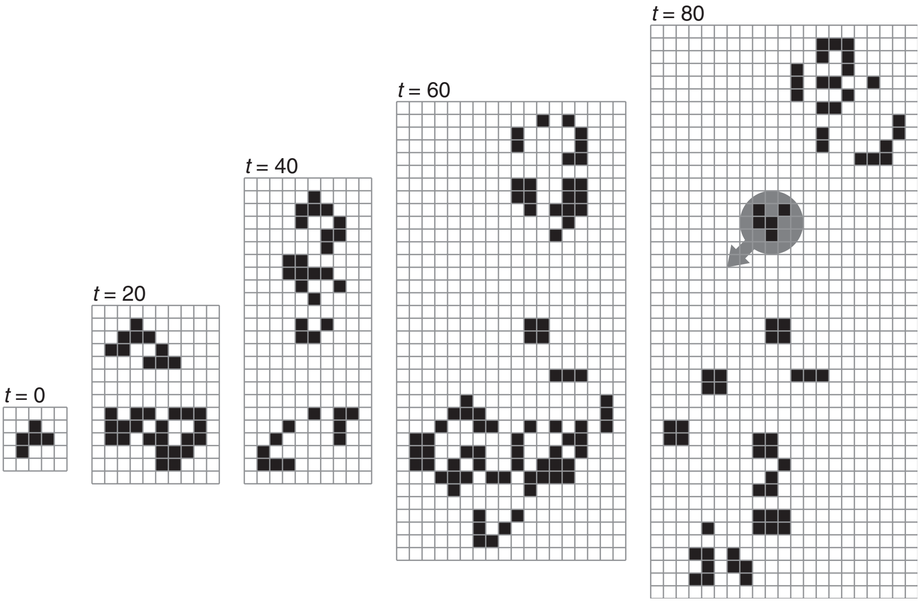

Figure 1.6 The first 80 time steps of the R pentonomino. The highlighted five cell pattern in the $t = 80$ snapshot is a *glider* which moves one grid cell diagonally in the direction shown every four time steps.

The addition of just one further live cell to pattern (e) produces the 'R pentomino' (see Figure 1.6), which has been shown to persist for 1103 time steps before the region near its origin stabilises. By that time, the *glider* released at $t = 69$ (highlighted in the image at $t = 80$), which travels one cell south and west every four time steps, will have travelled around 280 cells away from the origin of the pattern. Meanwhile the pattern has produced five other gliders. This means that the R pentomino effectively persists indefinitely and extends infinitely across space, since the gliders will continue to move away from the origin.

Conway's invention (or discovery, depending on your point of view) led to an explosion of interest in cellular automata. That such simple rules can yield such unexpectedly rich behaviour led to considerable interest in the life CA from the recreational mathematics and amateur programmer communities. From an academic perspective, renewed interest in CA in statistical physics soon led to the discovery of even simpler examples, which also yield rich dynamic behaviour, with the contributions of Stephen Wolfram especially significant (see Wolfram, 1986).

In Chapter 3 we describe other examples of the general category of *totalistic automata* of which Conway's life is an example. These examples further reinforce the important lesson of this case, that the behaviour of even simple *deterministic* systems can be unpredictable. In fact, it has been demonstrated that for life (see Berlekamp et al., 2004, pages 940–957) and other even simpler cases (see Cook, 2004) the only way to find out what state a particular system will be in after a specified number of time steps is to set up a simulation and run it!

Complexity science It is tempting to see the game of life as little more than an obscure amusement. In fact, it is an exemplar of the sorts of systems studied in *complexity science*. Complexity science is unfortunately rather ill-defined, but roughly speaking it encompasses the study of systems composed of many basic and interacting elements (see Coveney and Highfield, 1995, page 7). On this loose definition a gas might be a complex system, so we need something more distinctive than this. In an early paper, Warren Weaver (1948) argued that the systems successfully studied by classical science were composed of either very few or very many elements. The former are usually analytically tractable, while the latter can be successfully studied using statistical methods, the statistical mechanics derivation of the gas laws being the most compelling example. Unfortunately for classical science, many real-world phenomena are 'middle-numbered' systems somewhere between these two extremes. Weaver calls these 'systems of organised complexity' (1948, page 539) where interaction effects are important—they don't simply average out as happens to molecules in a gas—and individual interactions between particular elements in one part of the system can unexpectedly scale-up to cause system-wide transitions. Weaver perceptively (and prophetically) suggested that the study of these systems would involve the use of the then recently developed electronic computers:

> [it seems] likely that such devices will have a tremendous impact on science. They will make it possible to deal with problems which previously were too complicated, and, more importantly, they will justify and inspire the development of new methods of analysis applicable to these new problems of organised complexity. (Weaver 1948, page 541)

Complexity science is a research programme along the lines envisaged by Weaver—a diverse one extending across many disciplines—which, as he anticipated, makes extensive use of computer models to understand otherwise intractable systems. It is impossible here to give a thorough account of the disparate findings of complexity science. General features of the structure and dynamics of complex systems have been characterised, including *path dependence*, *positive feedback*, *self-organisation* and *emergence*. Readers interested in learning more about complexity science should consult any of the large numbers of general and popular introductions to the field (see, for example, Waldrop, 1992, Coveney and Highfield, 1995, Buchanan, 2000, Solé and Goodwin, 2000) or more formal treatments in the academic literature (Simon, 1996, Allen, 1997, Byrne, 1998, Cilliers, 1998, Holland, 1998, Manson, 2001, O'Sullivan, 2004, Batty, 2005, Solé and Bascompte, 2006, Érdi, 2008, and Mitchell, 2008).

Nonlinear dynamics The complexity of the life CA is an example of the broad category of *nonlinear dynamics*. Like the game of life, many nonlinear systems exhibit surprising shifts in behaviour in response to seemingly minor changes in their initial states. Linear systems, by contrast, behave more predictably, with small changes in inputs producing small changes in the response, and large changes in inputs required before large changes in the response will occur. While the systems studied in what Weaver (1948) dubbed classical science were linear, those that have increasingly come to preoccupy scientists are nonlinear. In fact, we now recognise that linear systems are a convenient idealisation of the real world that enables us to build mathematical models. In many situations, linearisation is a good approximation and a useful one (remember: *models are never true* . . .), but this is not always the case, and many systems require nonlinear models to represent them adequately. The key point to note is that nonlinear systems, because of their structure, are often more conveniently analysed by means of computer simulation models than by more traditional mathematical methods.

Chaotic systems provide another example of nonlinear behaviour and demonstrate that deterministic systems need not consist of large numbers of elements for unpredictability to emerge; deterministic systems with only a few elements can also be mathematically intractable and unpredictable (Gleick, 1987, Schroeder, 1991). Given *perfect* information about the current state of a deterministic system, *in principle* its state at any particular future time is calculable and knowable. In practice, however, chaos and complexity prevent such predictability. In chaotic systems the difficulty is that even minor perturbations—or minor inaccuracies in the data recording the initial state—can lead to arbitrarily large differences in the outcome at some later time, rendering them unpredictable for practical purposes. In fact a commonly used definition of a chaotic system is one that shows extreme sensitivity to initial conditions. In a complex system the lack of predictability arises from the unexpected ways in which, via positive feedbacks, self-organisation, lock-in and other mechanisms, local interactions among system elements scale up to cause system-wide outcomes and effects.

Beckage et al. (2011) expand on the implications of these characteristics in ecology, and similar conclusions are warranted for social, economic and cultural systems. The essence of their argument is that such systems are *irreducible*, that is, we cannot easily reduce their behaviour to aggregate rules of thumb or predict the precise outcome of a given starting configuration *even if the systems are completely deterministic*. The easiest way to think of this is that many nonlinear systems lack the quality of ‘predictability’. This makes simulation an essential tool for understanding and exploring their behaviour.

1.4 Why dynamic and spatial models?

1.4.1 The strengths and weaknesses of highly general models

Many typologies of models have been developed and these tend to emphasise dichotomies, such as ‘complicated’ *versus* ‘simple’, in model design. Although such black and white dichotomies are invariably simplistic, they provide a useful way of thinking about the trade-offs that must be made when developing appropriate and manageable abstractions or representations of real-world phenomena. In a classic paper on the strategy of model building, Levins (1966) argues that model building requires the modeller to make trade-offs between generality, precision and realism. A single model cannot have all three and so the modeller has to sacrifice generality for realism and precision, precision for generality and realism, or realism for generality and precision (Perry, 2009, provides a schematic illustration of these trade-offs). In other words, a highly detailed simulation model of a specific system at a specific place and time may be realistic and useful for understanding that particular system, but will be difficult to transfer to other systems and so it is not general. On the other hand, a general framework, such as von Thünen’s land-use model of urban spatial structure, may lack realism when applied to specific city systems because it glosses over the geohistorical details of the growth of particular cities, but it can inform us about the behaviour of cities in general terms.

To explore the idea of trade-offs between generality and realism more closely, in this section we discuss how we might model observational data using probability distributions such as the normal or lognormal and the highly general mechanisms that underpin them.

1.2 **An additive process: the normal (Gaussian) distribution** A familiar way to model observed data is to use the Gaussian (or normal) distribution, the ‘bell-shaped’ curve that underpins many data analysis methods (Figure 1.7). In many systems *approximately* Gaussian distributions arise as the result of a series of independent *additive* processes, as explained by the (additive) central limit theorem. The statistician Sir Francis Galton described a machine called a ‘Galton board’ or quincunx, which demonstrates how Gaussian distributions arise from repeated additive processes (see Figure 1.8). On the Galton board a ball is dropped from the top and on hitting a pin moves, with equal probability, either a unit to the left or a unit to the right before

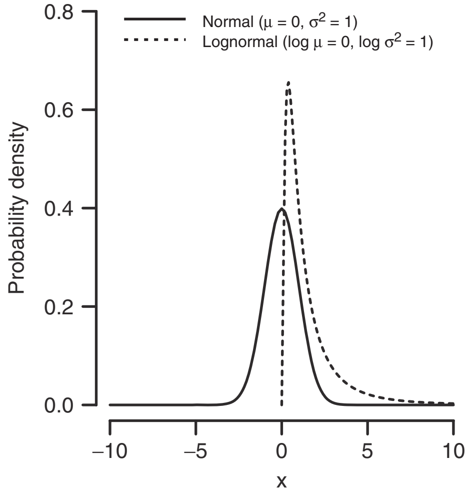

Figure 1.7 Normal and lognormal probability distributions. The symmetrical normal distribution arises from multiple independent *additive* processes, whereas the right-skewed lognormal distribution arises from multiple independent *multiplicative* processes. The normal (or Gaussian) distribution is defined by the mean (μ—the location parameter) and the standard deviation (σ—the scale parameter).

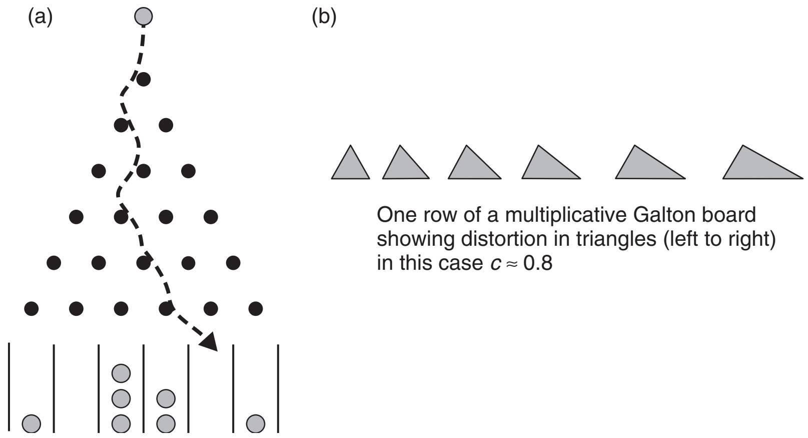
(a)
(b)
One row of a multiplicative Galton board showing distortion in triangles (left to right) in this case $c \approx 0.8$

Figure 1.8 (a) The Galton board described by Sir Francis Galton. Balls are dropped and move through the system additively, eventually being collected. The heights of the piles of balls in each unit at the bottom of the board follow a Gaussian distribution. (b) One row of the hypothetical multiplicative Galton board described by Limpert et al. (2001).

eventually being collected at the bottom. The sequence of movements that each ball makes constitutes an independent additive process (there is a direct relationship between this model and a random walk in one dimension—see Chapter 4). Our interest here is in the distribution of the final position of the balls on the board, that is, the relative sizes of the piles that collect below the board. After many balls have been dropped, the size of the piles under the board approximates a binomial distribution, which, according to the central limit theorem, will tend to a normal distribution if a sufficiently large number of balls are dropped, as shown in Figure 1.9(a).

Multiplicative processes: various skewed distributions We might expect that a different distribution will arise in systems dominated by multiplicative 1.3 processes, and this is indeed the case. For processes comprising many independent but *multiplicative* events, a right-skewed approximately lognormal distribution can arise (also known as the Galton–McAlister, Kapteyn or

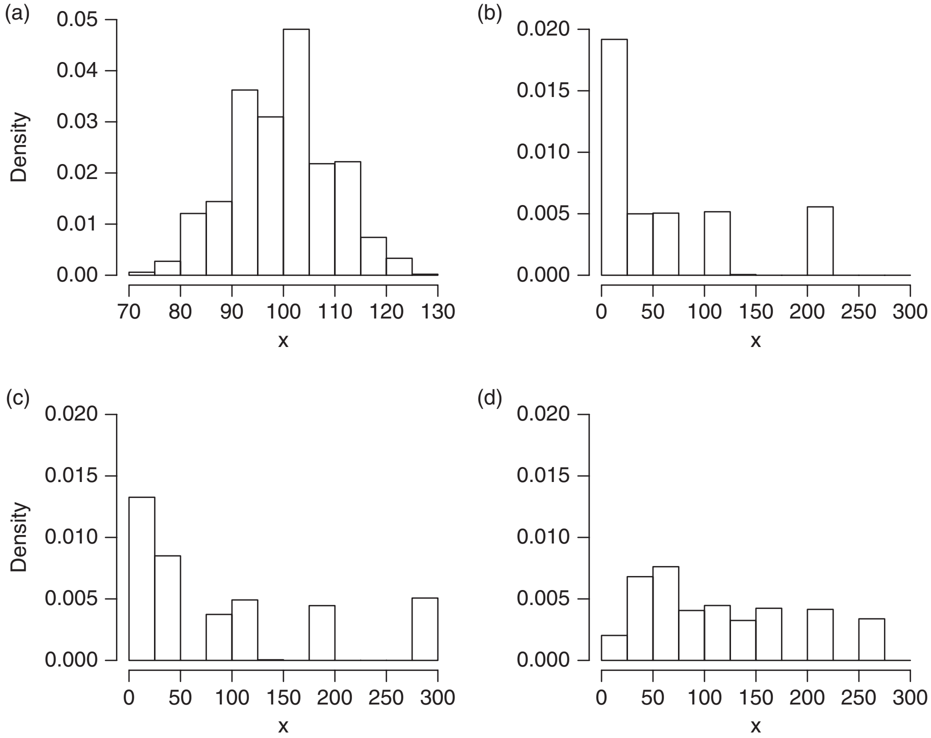

Figure 1.9 Outcomes of simulations of (a) additive and (b)–(d) multiplicative *in silico* Galton boards, with $c = 0.7$, 0.8 and 0.9 in (b)–(d), respectively. The right-skewed distributions produced by the multiplicative process become more obvious as c decreases and the strength of the multiplicative process increases. Note axes are scaled differently on (a) than on (b)–(d).

Gibrat distribution, see Koch, 1966). Thus, while the Gaussian distribution is widely used, many (perhaps most) processes in environmental and social systems are multiplicative rather than additive and so, as Heath (1967), Koch (1966), Limpert and Stahel (2011), Limpert et al. (2001), and others have argued, in many cases a right-skewed distribution, such as the lognormal, is a more appropriate model. For example, population growth is typically multiplicative (Limpert and Stahel, 2011), as is the growth of firms (Simon and Bonini, 1958) and many other things besides.

Limpert et al. (2001) describe a modification to the Galton board (Figure 1.8) that demonstrates the outcome of multiple independent multiplicative processes. In their modified Galton board, shown in Figure 1.8(b), the isosceles triangles are replaced by scalene triangles. Each triangle has a skew to the right such that at each successive level the ball moves to either $x \times c$ or x/c, where x is the current position on the horizontal axis and c is the skew to the right in the relevant triangle. This modified process is multiplicative rather than additive, but it is still independent from move to move. Under these conditions the distribution of the heights of the piles of balls follows a skewed distribution similar to the lognormal, as shown in Figure 1.9(b)–(d). This is an example of the *multiplicative* central limit theorem.

It is worth noting that the lognormal is not the only distribution generated by multiplicative processes. Mitzenmacher (2004) describes how Zipf, Pareto and power-law distributions, among others, can arise from multiplicative processes, and there has been considerable debate over which of these best describe various datasets. The point here—echoing the comments on equifinality on page 15—is that snapshot data, such as the frequency distribution of the heights of the piles of the balls under the Galton board, may be adequately described by more than one model, and the data themselves do not provide sufficient evidence to accept or reject a specific model (a general problem called *under-determination*, Oreskes et al., 1994, Stanford, 2009).

In spite of the high level of generality involved in both models, the contrast between the normal and lognormal distributions is sufficient that in most situations we can use observational data to at least say whether plausible generating mechanisms are broadly speaking additive or multiplicative in nature. However, there are limits to our ability to use distributional patterns in this way. As discussed, many different mechanisms can be shown to generate heavy-tailed distributions (see Mitzenmacher, 2004), so that distinguishing between particular processes based on observational data alone may be technically challenging and perhaps untenable (Clauset et al., 2009, Stumpf and Porter, 2012). Thus appeals to the highly generalised mechanisms that underlie probability distributions are problematic. Many different mechanisms yield similar distributional outcomes. Furthermore, the mechanisms themselves are so generalised that determining the most suitable distribution to describe a particular case can provide only limited information about systems and the processes driving them. This points to the potential value of more detailed models incorporating more specific mechanisms, although, keeping in mind Levins's (1966) trade-off, we might expect any advantages in specificity to come at a cost in generality and precision.

1.4.2 From abstract to more realistic models: controlling the cost

As we have seen, probability distribution functions are highly abstract and highly general—if we have right-skewed data, with all values greater than zero, then the lognormal distribution is (possibly) an adequate description of those data regardless of context, and we should consider how multiplicative processes might drive the system. If, however, we are considering specific patterns in specific environmental or social systems and want models that represent in more detail the processes that determine and are influenced by those patterns, such generality will be lost. Does this mean that more detailed models of particular systems cannot also teach us more general lessons about how the world works?

The key idea behind this book is that if we are alert to the similarities between patterns across diverse domains then perhaps the loss of generality as we move from abstract to more specific models can be mitigated to a degree. As an example, in many cities populations with different socio-economic characteristics are segregated with respect to each other, and conversely aggregated with respect to themselves. Likewise in temperate forests competitive interactions often result in individuals of different species being segregated from each other, and conversely they show strong conspecific aggregation. No one would suggest that the same causal mechanisms are driving these patterns, but is it really the case that there are no general ways of thinking about the processes that produce 'segregation' across the very different contexts in which it appears? Likewise, visitors to an art gallery looking for a celebrated work of art and caribou searching for food in the tundra are clearly very different, but is there any commonality in these 'foraging' processes that means we can, at least in the first instance, think about their representation in the same way?

We believe that many ostensibly different spatial patterns, arising in very different systems, can usefully be represented using similar frameworks (see also Ball, 2009). This does not mean that the processes are the same (of course they are not!) but rather that it may be useful to use similar models to understand the ways in which very different processes can produce similar patterns. Our goal is to make building and interpreting spatial models easier by showing how a relatively small number of models of spatial processes can generate many of the types of patterns seen across a wide range of social and environmental systems. Such models will not be as general as the independent additive and multiplicative processes that underlie the normal and lognormal distributions, but, with some care, we can identify more specific spatial processes that can account for diverse spatial patterns commonly observed in a wide variety of systems.

Central to our approach is the view that some broad categories of spatial outcomes, namely aggregation, movement and spread, can be accounted for by a relatively small range of relatively simple dynamic spatial models. These 'building-block' models, which are covered in Chapters 3, 4 and 5, can in turn be combined to create more complicated, detailed and realistic models of particular real-world systems (see Chapter 8).

Developing detailed, dynamic, spatial models comes at some cost in generality and interpretability, but buys us realism and the ability to represent specific processes in specific contexts. If we are willing to see the similarities in patterns and processes across many different domains, then some of those costs are offset by making the models themselves easier to understand and less daunting to work with. But before we tackle the building-block models themselves, we must first consider more closely two key concepts: *pattern* and *process*, and these form the subject matter of the next chapter.