Table of Contents

Go Back

Representing Heterogeneity in Complex Feedback System Modeling: Computational Resource and Error Scaling

Nathaniel Osgood

Department of Civil and Environmental Engineering, MIT[1]

1 Introduction

Heterogeneity often plays an important role in shaping complex system behavior. If the evolution of system components is characterized by non-linear relationships with respect to some factor that varies over those components, it is not in general possible to capture system reference modes or behavior of interest using simulations containing only aggregate representations of that factor.

The representation of heterogeneity is not only necessary for characterizing endogenous system behavior; it is often critical for assessing the relative effectiveness of policy decisions at even the most qualitative level. [Shephard Zeckhauser 1982] describes the essential role heterogeneity plays in determining impact of many health policies. Heterogeneity has a strong impact on decision making in domains as diverse as educational admissions, environmental management, electoral strategizing, insurance, energy regulation, and many other areas.

Within the context of this paper, we will focus on cases where a modeler seek to understand and better manage some system that involves a real-world population of system components (e.g. people, cars, towns) that is heterogeneous with respect to some attributes. For a health-oriented model to study the impact of tobacco policy, this might be a population's age and gender, weight, and smoking status. For a model aimed at understanding the environmental impact of emissions-control standards, the relevant attributes might be vehicle age and emissions profile. For students in a university, we might describe their length of time at matriculation, area of specialization and possible (for diversity concerns), gender and economic or ethnic background. In a model involving spatial effects, we may wish to characterize interacting regions by geographical location or using some other coordinate system (distance from a central city, etc.). Some of the attributes of interest – such as age, weight, family income level, vehicle age and emissions profile – may be continuous, while others are discrete (categorical), such as gender and area of academic specialization. In other cases, the appropriate designation of an attribute as discrete or continuous may not be clear (e.g. smoking status, ethnic background).

This paper briefly surveys three common representations of heterogeneity in complex systems modeling frameworks and analyzes the tradeoffs associated with these approaches on computational resources and system error. Because these approaches for modeling heterogeneity are applicable to a wide variety of modeling tools, the paper characterize the approaches in a generic manner.

The next section of the paper provides a brief motivation for capturing heterogeneity in complex system modeling, including discussion of both the mathematical need to capture heterogeneity in the context of non-linear system evolution and a simple example which demonstrates the importance representing

[1] Address correspondence and reprint requests to Nathaniel Osgood, 77 Massachusetts Avenue Rm 1-376, Cambridge, MA 02139. Email: nosgood@mit.edu

heterogeneity in a stylized simplification of a health policy model. The following two sections briefly describe three popular ways in which heterogeneity can be captured in systems modeling frameworks. This is followed by a simple analysis of the computational resources required by each approach. The penultimate section analyzes the scaling of error with the number of distinct types of heterogeneity that need to be captured. The conclusion summarizes these results and discusses their implications for modeling.

2 Motivations for Representing Heterogeneity

2.1 Implications of Non-linearity for Aggregation

One of the most important motivations for the representation of heterogeneity in a model is non-linearity of system evolution with respect to system state. It is well understood in the system dynamics community that non-linearity of this sort is widespread and effects greatly complicate the analytic study of systems by preventing easy decomposition of a system into readily understood and pieces or external inputs into eigenfunctions, and that non-linear systems therefore require the use of numerical integration of the broader system to gain an accurate picture of system evolution. What is less widely appreciated in this community is that non-linearity in the evolution of elements of a population (e.g. people, cars, etc.) often requires that models of such a system represent important components of heterogeneity among the members of that population if they wish to accurately capture qualitative system behavior.

Consider a complex deterministic system that includes a population of *n* components of the same general class (e.g. people, cars, etc.), each of which exhibits some system state comprised of *d* characteristics (for example, people with different ages, weights, genders and smoking status, or cars with different levels of fuel efficiency and age). We term a given member of this population $\vec{x}_i$ (where $\vec{x}_i$ is *d*x1) and take the general case in which the time evolution of that particular member of the population is governed (as above) by a shared but possibly non-linear function *f*: $\dot{\vec{x}}_i = f(\vec{x}_i)$

Now consider the evolution of the entire population $X = \{\vec{x}_i\}$; we use $\dot{X}$ to denote $\{\dot{\vec{x}}_i\}$. From the above, it follows directly that $E(\dot{X}) = E\big(f(X)\big)$. Consider now the situation for a non-linear function *f*. For such a function, in general $f(\vec{a}+\vec{b}) \neq f(\vec{a}) + f(\vec{b})$ and thus

$$E\big(f(X)\big) = E\left(\frac{\sum_i f(\vec{x}_i)}{n}\right) \neq f\left(\frac{\sum_i \vec{x}_i}{n}\right) = f\big(E(X)\big).$$

In spite of this difficulty, it is common to encounter aggregated non-linear models that attempt to characterize system evolution using just the mean states or attributes of the entire population or broad divisions thereof. An example with which the author was associated (and a motivation for the example below) is described in [Tobacco Policy Model] Many of the models suffering from this problem are of high quality and have made their way into prominent model collections ([Kaibab] [Flowers] [Epidemic] [Resources] [Bass Diffusion]). Perhaps the most familiar type of neglect in the treatment of heterogeneity can be found in the widespread tendency to represent of a non-linear stochastic system using the *mean* trajectory of the system, and the calculation of non-linear effects on the basis of this mean trajectory.

As will be shown below, the consequence of ignoring heterogeneity in non-linear systems is inaccuracy and – in some important cases –failure to capture important qualitative components of system behavior.

2.2 A Simple Example

In this section, we examine a very simple but realistic example of a non-linear system and demonstrate how significant systematic inaccuracy can result when simulation is conducted with respect to aggregated attribute values.

As the basis for this example, we take the characterization of smoking behavior – a subject that has served as the basis for several policy-oriented system dynamics models [Tobacco Policy Model] [Roberts, Homer et al]. In a subdivision common to their field, [Tobacco Policy Model] characterize the population into three categories based on current and past smoking status: Never smokers, current smokers, and former smokers. Following this model, we make use of a Markovian model structure where individuals in a given smoking category have a certain probability density of changing their smoking behavior over time and transitioning to a new smoking state. Thus, never smokers are modeled as initiating smoking ("**u**ptaking") with probability density *u*, current smokers have a probability density *c* of quitting ("**c**easing") smoking, and former smokers exhibit probability density *r* of **r**elapsing into smoking.

The resulting system can be characterized as a classical set of 3 first-order linear differential equations:

$$\frac{d\vec{x}}{dt} = \begin{bmatrix} -u & 0 & 0 \\ u & -q & r \\ 0 & q & -r \end{bmatrix} \vec{x}$$

Where $\vec{x}$ is the state vector consisting of the count of never, current and former smokers ($\vec{x} = \begin{bmatrix} N \\ C \\ F \end{bmatrix}$).

For simplicity of this example, we will assume an initial state vector of $\begin{bmatrix} N_0 \\ 0 \\ 0 \end{bmatrix}$ (i.e. that all of the population starts in the never-smoker state).

It is worth remarking that while the differential equation is linear with respect to the smoking status state vectors, it is non-linear when both the state vector and (individually dictated) transition probabilities are taken as variables. More specifically, while the behavior of an individual *i* is described by a $\dot{\vec{x}}_i = A\vec{x}_i$ (where that contents of A are dictated by that individual's values of *u*,*r*,*q*), that of individual *j* may be described by a different equation $\dot{\vec{x}}_j = B\vec{x}_j$ (where A≠B), because the transition probabilities for each individual may differ.[2] Thus $E(\dot{\vec{x}}_i + \dot{\vec{x}}_j) = E(A\vec{x}_i + B\vec{x}_j)$. In general, unless the two individuals have

[2] It is widely recognized in the health community that individuals vary widely in likelihood of changing smoking-related behavior.

identical transition probabilities (and thus A=B), there is no constant matrix C such that $E\left(\left\{\dot{\vec{x}}_i,\dot{\vec{x}}_j\right\}\right)=\frac{\dot{\vec{x}}_i+\dot{\vec{x}}_j}{2}=C\left(\frac{\vec{x}_i+\vec{x}_j}{2}\right)=CE\left(\left\{\vec{x}_i,\vec{x}_j\right\}\right)$ for all individual states $\dot{\vec{x}}_i$ and $\dot{\vec{x}}_j$.

Given this non-linearity, the remainder of the section will examine the implications of heterogeneity for this particular system as it occurs in two parameters and affects two respective types of systems behavior. In particular, we will examine the effects of heterogeneity in initiation rates and its effect on dynamic behavior, and the implications of heterogeneity in relapse rates and its effect on the system steady state. In both cases, we demonstrate the systematic discrepancies that result from simulation or analytic solutions that assume system transition rate averages (i.e. which model the system at an aggregate level, assuming that the aggregate parameters (u, r, q) are associated with the system-wide averages for these parameters).

Consider first the case of initiation behavior. As represented in this stylized model, the stock of never smokers is associated with a single (out) flow. The behavior is thus the familiar one of a first-order delay, consisting of the mean interarrival time of a Poisson process. Considering for the moment this component of the system in isolation. We have

$$\frac{dN}{dt}=-uN$$

The solution is the declining exponential, where c is a constant to be determined by the initial conditions:

$$N=ce^{-ut}$$

To emphasize the dependence of N on parameter u, we will write this as N_u below.

Consider a simple case of a population of N_0=1000 in which half the population (500 individuals) have uptake probability density .05%/month and half 1.5%/month. We will contrast the difference between the exact solution to this simple situation and the solution that arises from neglecting the heterogeneity in the population and modeling it as a homogeneous population with uniform uptake transition probability density 1% (a number identical to the *mean* population transition density for the population).

As demonstrated by Figure 1, the numerical behavior between the two cases diverge. In particular, the use of an aggregated transition probability in the model yields a systematically pessimistic estimate of the number of individuals remaining as current smokers.

The systematic nature of this difference arises directly from the system heterogeneity. It reflects the fact that the population of never-smokers is changing over time, and in particularly is being rapidly drained of the "high risk" population with higher uptake rate, leaving an increasingly dense population of individuals at low risk of smoking. Within a homogeneous population (as simulated by the model using aggregated parameters), the aggregate probability of uptake would remain constant over the lifetime of the simulation, and a considerably larger quantity would begin smoking, leaving fewer as remaining non-smokers. In the presence of sufficiently large heterogeneity, the qualitative behavior of these two approaches may differ (with the heterogeneous model continuing for a prolonged period with a sizeable fraction of the "low-risk" population remaining never-smokers, while the never-smoker population in the aggregate model is rapidly and uniformly depleted).

The difference also reflects the underlying mathematics of the scenario: The analytic solution for the count of never smokers N(t) as a function of time depends non-linearly on the value of *u*. Given that the *exp* function is convex, basic calculus and Simpson's rule tell us that $N_u < \frac{N_{\frac{u}{2}} + N_{\frac{3u}{2}}}{2}$. a

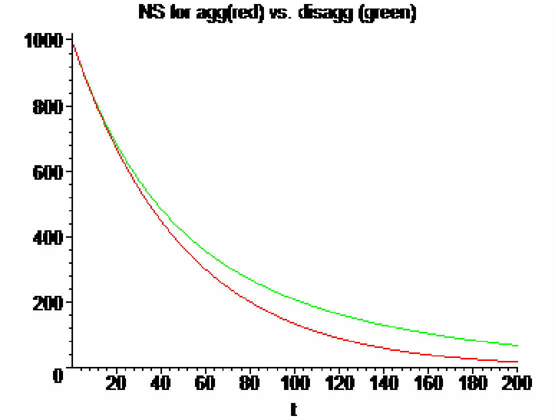

Figure 1: Count of Never Smokers under Assumption of Mean Transition Probabilities (Red) and Heterogeneous Transition Probabilities (Green)

The discussion above has focused on the dynamics associated with a particularly simple subcomponent of the model, exhibiting only exponential decay. Because of the simplified nature of the model – in which the never smoker stock is only drained and never replenished – it is clear that despite the significant differences in uptake rate between an aggregated and aggregated model, the asymptotic behavior of the never smoker stock in both models is identical: The stock depletes to zero over time. We now turn to examine how choice of an aggregated model introduces systematic inaccuracies into our understanding of the steady-state behavior of the system.

Consider again the system of differential equations shown above. For an initial population vector $\begin{bmatrix} N_0 \\ 0 \\ 0 \end{bmatrix}$ exhibiting homogenous transition parameters (*u*,*r*,*q*) the stylized model gives a steady state population distribution of $N_0 \begin{bmatrix} 0 \\ \frac{r}{r+q} \\ \frac{q}{r+q} \end{bmatrix}$ -- in other words, a situation in which a fraction $\frac{r}{r+q}$ of the population consists of current smokers, and $\frac{q}{r+q}$ consists of former smokers. It bears mentioning that these fractions are non-linear in both r and q, and thus we would not expect the steady state of mean transition values to be the same as the mean of the steady state of the full distribution of transition values. To quantify this difference, suppose that we have a heterogeneous population in which one half of the population is at “high risk” of relapse whenever they are former smokers, and the other half is at low risk whenever they have ceased smoking. We will now contrast the implications of modeling such a population using a model assuming a uniform relapse rate across the entire population (a relapse rate equal to the mean of the actual relapse rate over the entire population) , and a disaggregated model which accurately characterizes the relapse behaviors of the high-risk and low-risk groups by following each separately. Noting that only one of the steady-state populations of current or former smokers needs to be specified to determine the other, we will henceforth focus our attention on the current smoker count.

In this stylized model, the exact steady-state behavior of the heterogeneous system can be obtained by simulating the two populations independently and combining the results. Thus, we if have half of the population ($N_0/2$ “high risk” individuals) associated with relapse risk $r+\alpha$ and the other half of the individuals with relapse risk $r-\alpha$, we arrive at an steady state in which the number of current smokers at steady-state is $N_0 \frac{r^2+rq-\alpha^2}{(r+q+\alpha)(r+q-\alpha)}$. By contrast, a model built around the assumption of a homogenous rate of relapse uniform across the population would yield a steady-state estimate for the current smoker count of $N_0 \frac{r}{r+q}$. Thus, despite the fact that both the aggregated and exact estimates made exhibited identical mean relapse rates, a systematic error of value $N_0 \frac{q\alpha^2}{(r+q)(r+q+\alpha)(r+q-\alpha)}$ is introduced between the two estimates (respectively) of the size of the steady-state current smoker population.

Given that *r*,*q*, and α are all positive, this difference will be positive as long as $\alpha<(\underline{r+q})$. Thus, given a real-world Markovian system such as that shown, a model that ignored heterogeneity in relapse rates would consistently over-estimate the number of individuals who remain as current smokers and under-estimate those who successfully remain former smokers.

The value of the difference in magnitude depend heavily on the values the transition probabilities; Figure 2 shows the absolute error expressed as a function of the fractional heterogeneity (α/r). The proportional error (error in the estimate of the current smoker population as a fraction of the current smoker population) will generally be several times higher than that shown in these figures.

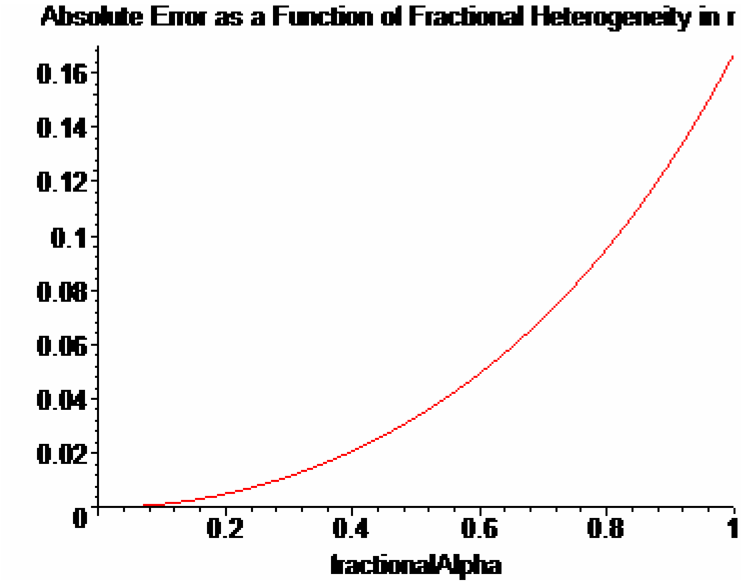

Figure 2: Inaccuracy of Steady-State Estimate of Current Smoker of Population as a Function of the Fractional Population Relapse Rate Heterogeneity (α /r). The Total is Expressed as Fraction of Population for Sample Data with q=r=.1.

This section has demonstrated that for a very simple system that is non-linear with respect to population parameters, making use of a model that fails to capture heterogeneity of the system with respect to model parameters can introduce potentially significant errors into the expectations of system evolution. It is only through the disaggregation of models of such system with respect to key attributes that close fidelity of model simulation to real-life system evolution can be expected. The body of this paper examines the accuracy and computational costs associated with three approaches to representing heterogeneity.

3 Three Techniques for Representing Heterogeneity

In this section, we briefly survey three distinct means for capturing system heterogeneity within a model. Suppose that we have an underlying population of size *n* of some general class of component (people, cars, etc.) in the system of interest. Suppose further that each of these population members is associated with *d* state elements (termed here “characteristics” or “attributes”) the heterogeneity of which we believe likely to have a strong impact on the understanding system behavior and to impact policy selection: For example, a population of people with distinct age, gender, weight, and smoking status (which we take to be categorical for simplicity’s sake). There are three primary ways to characterize heterogeneity in such frameworks:

- **Using Stocks Disaggregated by Attribute Value.** In this case, the user represents components holding different ranges of attributes values using distinct stocks in the model. The framework maintains distinct levels for each stock.[3] Each such level represents the quantity of that particular type of component within the system. Although its exact form differs in different

[3] Note that we confine this discussion to stocks, but the use of arraying extends to other variables. Other types of formulations can either be represented as stocks (e.g. delayed fixed) or need not be represented at all as part of system state vector, and can simply be computed from the state vector.

software packages, *arraying* is frequently used as a convenient means of disaggregating stocks and auxiliary variables according to *attribute value* (for discrete attributes) or range (for continuous attributes). This technique allows for convenient representation of common names and common (vectored) equations shared between instances of an arrayed stock, but is conceptually the same as explicit disaggregation.

- **Agent-Based Disaggregation.** Agent-based modeling (of which we consider micropopulation models [Ackerman] a subset) is a general modeling technique that represents the set of components of a system using a set of *agents*. These agents are associated with sets of user-defined attributes shared between member of a population of a particular class of component. For example, to draw on the examples introduced above, an agent might represent a particular person in the population, and associated with particular values for the attributes of current smoking status, and other characteristics that might influence likelihood of initiation, relapse and cessation (such as age, gender, physiologic responsiveness to nicotine, etc.). An agent to be used in a study of vehicle pollution might represent a particular vehicle, with attributes of age and tailpipe emissions profile. In general, a particular agent can have arbitrarily many such attributes.
- **Using Co-flows.** Co-flows are a common technique for capturing important *statistics* (population means) arising from heterogeneity with respect to population characteristics without the burden of disaggregation.

Each of these techniques will be briefly described below.

3.1 Population Stocks Disaggregated by Attribute Value

The representation of heterogeneity using disaggregated stocks is the most straightforward and likely the most common of the techniques for representing heterogeneity in traditional system dynamics models. It is also the one on which this paper will focus the greatest attention.

We can choose to represent this population within stocks disaggregated according to these attributes. In order to accomplish this process, we create a collection of stocks, each occupying some particular volume in *d* dimensional space. Each dimension of this space is associated with a particular attribute; the attribute (state element) values of each member of the population can be conceptualized as coordinate for this member in this *d* dimensional space.

Within this representation, the level of each such stock will represent the number (or, alternatively, the fraction) of individuals in the population whose attribute values in *d* are such that they fall within the volume of interest. In other words, the stock counts the number of individuals in the population that share particular sets of values for every one of the attributes (age, gender, weight and smoking status).

For continuous attributes and dimensions (e.g. age, weight), a particular stock will count members of the population whose state with respect to that characteristic falls into a particular contiguous range of values for that attribute (e.g. people ages 50 through 59) and who also have the appropriate values specified required for the other attributes to fall in this category.

For discrete attributes (e.g. gender and categorical smoking status), a stock might represent population members with a particular value for that attribute (e.g. females) and appropriate values for all other attributes. Alternatively, a stock might be associated with all population members who belong to a particular countable set of values for that attribute (e.g. current and former smokers) and appropriate values for all other stocks.

The representation discussed here is also very well suited to changes in population size and composition. Appearance and disappearance of population members can be readily represented. For example, births in a living population can be represented as flows into stocks representing the youngest members of the population, with appropriate other attributes; mortality in a living population model simply requires outflows from the appropriate stocks. Population members who change their attribute values (e.g. members of the human population who age, gain or lose weight or change smoking status, or cars that age or degrade) will participate in flows between stocks with appropriate values. (For example, the aging of a single individual *i* aging from the 50-59 year old stock to the 60-69 year old and with other values (g_i, w_i, s_i) for gender, weight and smoking status remaining constant, could be represented as a flow of size one that increments the stock pop_{60-69,g_i,w_i,s_i} and decrements the stock pop_{50-59,g_i,w_i,s_i}).

It can readily be appreciated that while the framework is conceptually simple, very general and easy to implement via techniques like arraying, the number of distinct volumes in *d* dimensional space – and thus the number of distinct stocks – will increase rapidly according to the number of attributes of interest and the number of distinct ranges that must be broken out in this way. This observation will play a key role in the analysis below.

3.2 Co-Flows

Co-flows are system structures form a third important approach for the representation of the impacts of continuous heterogeneity in traditional system dynamics models. Unfortunately, they are also less general than stock disaggregation, which limits their use to certain special cases.

The general structure of coflows is shown in Figure 3. Within this approach, we characterize population means for some attribute or function of interest by keeping track the values of that attribute or function for each of the inflows and outflows into stocks. The next two sections introduce the fundamental coflow mechanisms, discusses tradeoffs on the applicability of the approach.

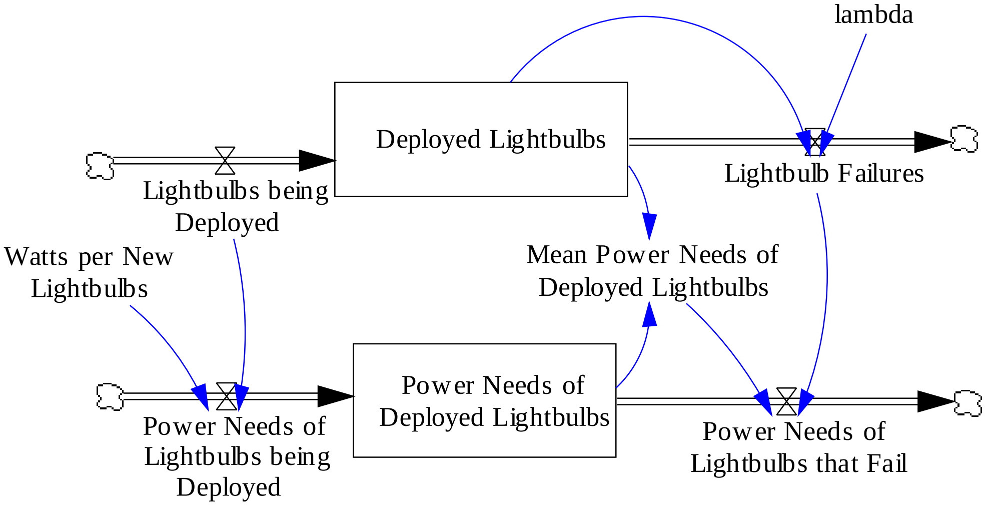
lambda
Deployed Lightbulbs
Lightbulb Failures
Lightbulbs being Deployed
Watts per New Lightbulbs
Mean Power Needs of Deployed Lightbulbs
Power Needs of Deployed Lightbulbs
Power Needs of Lightbulbs being Deployed
Power Needs of Lightbulbs that Fail

Figure 3: Example Co-Flow

3.2.1 Co-Flow Basics and Example

Suppose we have a new, higher-efficiency but higher-cost lightbulb design. Suppose we are interested in understanding the impacts of this greater lightbulb efficiency on aggregate lighting efficiency across *n* existing lamps in a warehouse that whose new lightbulb deployments will include some fraction of the new lightbulb (with the fraction depending on the policies in effect).

As depicted in Figure 3, we thus have a stock of deployed lightbulbs (named“Deployed Lightbulbs” in Figure 3) whose inflow corresponds to the introduction of the new, higher-efficiency lightbulbs (“Lightbulbs being Deployed”) and whose outflow is simply burnt-out lightbulbs (“Lightbulb Failures” in Figure 3). Suppose further that we treat the chance of failure in a lightbulbs as a Poisson process (and thus independent of age), where the incidence rate λ (Figure 3’s “lambda”) of failure is identical for the original and new lightbulb types.

While we could certainly represent the population of lightbulbs using stocks disaggregated by age of lightbulb, the structure of the processes are sufficiently simple that characterization of the basic dynamics of aggregate lighting efficiency does not require it. In particular, because the structure of the problem permits us to characterize the energy efficiency of inflows and outflows at any given point in time, we can calculate the energy efficiency of the aggregate stock of lightbulbs at any point in time as well.

The coflow is a mechanism that allows the modeler to take advantage of situations such as this one, where it is possible to maintain aggregate statistics for a heterogeneous population without explicitly disaggregating that population. The coflow does this by maintaining two stocks: One representing the population size (“Deployed Lightbulbs” in Figure 3), and the other the aggregate total for the quantity of interest in the population (“Power Needs of Deployed Lightbulbs” in Figure 3). By keeping track of the quantities of interest (energy use) on the inflows and outflows of the aggregate stock, we can at any time divide one stock by the other and obtain the *population mean* (“Mean Power Needs of Deployed Lightbulbs” in Figure 3) for the quantities.

The co-flow methodology generalizes naturally to the case where there are several attributes of interest on which the statistic whose mean value we wish to calculate depends. In this case, we may be interested in a particular statistic (the total energy use), but we could construct a similar a separate co-flow to calculate the other quantities of interest (such as average age, etc.)

3.2.2 Co-Flow Tradeoffs

By virtue of their elegant and computationally economical representation of complex heterogeneous systems, co-flows form a very valuable component of a system dynamics modeler’s toolkit. Co-flows allow modelers access to important aggregate statistics on systems without the need for expensive disaggregation.

Unfortunately, co-flows also have strict limitations in applicability. Firstly, because they deal with means over attributes in a population co-flows are really only applicable for continuously varying attributes. Heterogeneity with respect to categorical (discrete) attributes is not easily characterized via co-flows.

Secondly, co-flows are limited for use in calculating statistics on attributes with respect to which system evolution is invariant or linear over that attribute. This limitation reflects the need to for the modeler to specify both the internal dynamics of the population stocks and the attribute statistics of population outflows. Coflows are very difficult or impossible to employ to characterize mean population attributes when the population members change attribute values over time in a way that depends on residence time in the stock or the attribute values themselves. For example, while it is possible to model population aging using a co-flow (because the process applies uniformly over all ages of the population), it is not in general possible to maintain a coflow that calculates average weight for a human population over time, for the dynamics of change in weight are not linear in the time spent in the stock or mean weight.

In most co-flows, the mix of heterogeneous elements in the inflows is easily characterizable; as a result, calculation of the mean characteristics of the inflows is straightforward. Calculation of the mean characteristics of outflows is only possible in the presence of special assumptions. The most common case of this (seen in the example above) is where the population is “well mixed”, and the population members exiting the population are associated with attribute values whose mean value is equal to the mean of the attribute values of the population of the whole. It should be clear that this is a highly restrictive assumption; in general, the exit likelihood for a population member will depend on the length of time that member has been in the stock or on the value of the attribute of concern.

For example, in the co-flow example in the previous section, the use of a co-flow was made possible by the fact that the Poisson failure rate λ was identical for both the high-efficiency and low-efficiency lightbulbs. With real lightbulbs, the likelihood of burnout for a given lightbulb may vary significantly with energy efficiency – and perhaps with age. In this general case, it would not be possible to calculate outflow rates or the characteristics of those exiting the stock without knowing the proportion of the stocks occupied by high-efficiency and low-efficiency lightbulbs – and thus the use of disaggregation of the stock of lightbulbs by one or more attributes.

While co-flows are valuable modeling tools, their limitations in applicability greatly restrict their use in representing heterogeneity. For systems in which heterogeneity is important, the rates of stock outflows and rates of change in attribute values by members of the population is frequently directly dependent on attribute values, and not just on the mean of the attribute over some population.

3.3 Stocks Disaggregated by Agent

Regardless of modeling framework, one of the most frequent means of representing heterogeneity over a population consists of maintaining distinct information on the attributes associated with each member of the population. For example, consider a system in which four companies are competing, and where attributes such as employee and customer count, level of efficiency, bank balance, etc. form the dominant causal factors in success. A system dynamics model using agent-based disaggregation might maintain distinct stocks for each attribute of each company (e.g. company A employee count, company A customer count, etc.). Alternatively, most system dynamics frameworks would permit the representation of an arrayed stock for each of the attributes of those corporations. In an explicitly agent-based modeling environment (such as SWARM, Repast, or Netlogo), the agents representing such companies would be represented by objects (associated with classes or structures), each possessing attribute values represented as instance variables. A state equation model created in Matlab might represent agents as a collection of vectors of state variables, each indexed by company.

The agent-based approach is very straightforward; as will be shown below, it also scales effectively to large number of attributes. The approach does, however, require some accommodation for representation of very large populations. In particular, for cases in which the population being modeled by the agents is very large (for example, the population of a country), the computational expense may exceed the statistical benefits of attempting to simulate the full population. For such a simulation, the agent-population can be to be *downsampled* from the real-world, thus forming a “micropopulation”.

Where needed, statistical analysis of data from this approach can be made statistically rigorous and attractive through use of bootstrap [Bootstrap] iterations. For the sake of generality and in line with our desire to examine asymptotic behavior of the techniques, we will assume the use of the downsampling in the analysis in Section 3.3. For example, to characterize smoking behavior in a population (such as that modeled in a simple fashion in Section 2.2), computational and data demands would likely prevent the characterization of the smoking behavior associated with each individual in the entire population.

In contrast to co-flow or attribute-based approaches, the representation of heterogeneity in agent-based models requires no special techniques: It emerges from the description of agents as entities with attributes and the use of agents to characterize specific population members. In contrast to these other approaches for representing heterogeneity, there is no need to quantize the attribute values for agents into ranges for placement into discrete stocks: Each agent represents the data needed for characterizing its behavior to whatever level of precision desired.

What is clear is that an agent-based approach to representing heterogeneity requires a very substantial investment of computational resources for even the most conceptually simple models characterizing behavior in a large population. The sections below quantify this investment and contrast it to what is required in models which represent heterogeneity using attribute-based disaggregation.

3.4 Representational Choice in Two Popular Frameworks

While establishing criteria for the choice of frameworks lies outside the scope of this paper, it is worth noting that modeling frameworks vary broadly in the degree of choice and convenience they offer to modelers in capturing heterogeneity. A few words are in order about systems likely to be of greatest familiarity to the audience: Agent-Based modeling and System Dynamics.

As would be expected, agent-based frameworks permit great expressiveness in characterizing agent-based disaggregation, but offer only the most incidental support (if any) for other approaches. While recourse to general-purpose programming mechanisms in principle allow the use of all techniques described for representing heterogeneity, use of alternatives to agent-based disaggregation is not facilitated by the frameworks and in many ways runs counter to natural programming practice within agent-based frameworks.

System dynamics models offer modelers a richer set of choices, by permitting highly accessible means of representing simple forms of all three representations discussed. Unfortunately, within current system dynamics frameworks the agent-based disaggregation approach is somewhat rigid, and only grudgingly accommodates fluctuating populations: Regardless of whether an explicit or implicit arraying strategy is used, the size and structure of the model is “hard-wired” to the size of the population being modeled. This rigidness somewhat limits the appeal of agent-based disaggregation in system dynamics models, making it challenging to represent populations with birth and death processes (and especially those in which agents dynamically merge or split). System dynamics packages could add significant modeling expressiveness by permitting dynamically adjustable array dimensions.

4 Computational Resource Demands

The sections above have discussed the mechanics of representing heterogeneity in two modeling paradigms. In the ideal world, the techniques discussed above would be used to disaggregate the representation of a system to whatever degree desired for system simulation. Unfortunately, computational resources in the form of performance and space are limited, and simulation models exhibit great appetites for both. As a result, the modeler is forced to compromise precision to throttle computational resource demands. This section demonstrates the means by which these paradigms compromise accuracy in order to accommodate limits in computational resource availability. The results of this analysis will play a role in comparing the relative advantages of these two modeling approaches when provided with a particular level of computational resources.

In the case of attribute-based disaggregation, the compromises occur in the form of using less finely disaggregated stocks. It is important to emphasize that this "lumping" is made *statically* – and the scheme for associating stocks with volumes in *d*-space is frequently made independent of the knowledge of the frequency distribution of the population at hand over the space of attribute values.

In the case of agent-based disaggregation, the means by which accuracy is compromised to ensure computational feasibility is rather different. Rather than ignoring differences among certain groups of the population, agent-based models will *downsample* a population to bring the simulated population of agents to manageable size. This approach (which the analysis below will term as *sampling*) loses accuracy by the fact that particular population members may be omitted from the downsampled population. It is important to emphasize that this downsampling is typically performed *dynamic* (during simulation) and typically draws uniformly from all elements of the population. As a result, the downsampled population typically represents a uniformly random sample of the overall population.

Given that there is a practical tradeoff between computational resources and accuracy, it is worth considering the magnitude of the computation resource demands imposed by each approach. To this end, this section of the paper derives simple formulas for the relationships between computational processing time and storage space required for each paradigm. These relationships are not precise, but express how the resource demands vary for different sizes of populations and attribute values, as well as different levels of lumping and downsampling.

The next section turns to examine how the *error* associated with each paradigm varies with the size of the problem. The results of both of these sections help to suggest regimes in which each paradigm has strategic advantages. This may help guide the choice of one technique over the other by shedding light on loss of accuracy required under each technique in the presence of a fixed amount of computational resources.

Because of the contrasting representations involved in capturing heterogeneity, the computational resources vary with respect to different model parameters. Because both simulation approaches allocate data in memory and then update each of these data items in each timestep, performance is directly proportional to the amount of space required – and both scale according to the same functional form over model parameters.

- **Agent-based Disaggregation**: Computational resource demands in agent-based disaggregation vary with the number of attribute fields that must be updated across the population of agents within each timestep. If there are D attributes and N members of the population, performance and space usage scale with $\Theta(ND)$. If for the sake of speed or space we downsample the population by a factor of s (essentially sampling the population such that one out of every s members of the population are included), then the number of operations and amount of memory required scales with $\Theta\left(\frac{ND}{s}\right)$.
- **Coflows**: In those specialized cases where a co-flow based approach can be applied, the amount of space and number of expression evaluations required to represent D attributes in a population

of size N is much smaller. Within a co-flow framework, one stock is required for the population count itself, and one stock for each linearly independent statistic required for a given attribute. If f_{LI} is the count of independent statistics required, this gives a total of $\Theta\left(1+f_{LI}\right)$. In realistic situations, a modeler will typically seek a report on at least one statistic per attribute, (so as to report on the population statistics) and thus $f_{LI} > D$. For most practical models, the number of statistics reported per attribute will be a small constant and will not vary with D; in other words $f_{LI} < \alpha D$ for some $\alpha > 1$. Thus co-flows can be realized with time and space $\Theta\left(D\right)$.

- **Attribute-based Stock Disaggregation**. As noted in Section 3.1, the use of attribute-based stock disaggregation is widespread in the systems modeling community. Understanding the scaling of computational resource demands with respect to this approach is thus particularly important. In this technique, a stock is arrayed – either literally or figuratively – with a separate array index for each attribute dimension. If there are D dimensions and dimension i has d_i subdivisions, the total count of stocks to be calculated is $\Theta\left(\prod_{i=1}^{D} d_i\right)$. Consider that any non-trivial dimension i will have $d_i \geq 2$. Thus, $\prod_{i=1}^{D} d_i \geq \prod_{i=1}^{D} 2 = 2^D$. As a result, we have a space and computational time demand of $\Theta\left(\prod_{i=1}^{D} d_i\right) = \Omega\left(2^D\right)$ This suggests that the computational time and space rises exponentially in D.

5 Error Scaling Analysis

For modeling of populations that are large and exhibit a large degree of heterogeneity, it is frequently infeasible to run a simulation as complete as one would otherwise wish. This constraint is particularly binding for stochastic model that require the Monte Carlo simulations of large sets of realizations. In such cases, acceptable performance is traditionally achieved by lowering the precision with which the heterogeneity in population is represented. We have seen that representation of heterogeneity is important for understanding system behavior, and lowered precision in capturing such heterogeneity within a model typically leads to the divergence of model behavior from the behavior of the same model if system heterogeneity was fully captured.

We have seen above that different modeling approaches capture information on system heterogeneity using different mechanisms. It is therefore to be expected that the effects of lowered precision may lead to significantly different levels of error in different modeling frameworks. In this section, we consider how the errors associated in estimating a population statistic within the different modeling frameworks scale with the computational resources required by the simulation.

In particular, we consider the error associated with totaling a statistic over the attributes of the population represented in the model. For simplicity, we assume that the computation of the statistic for each takes time $\Theta\left(n\right)$. For example, we may wish to calculate the sum or mean or variance of the age of the population being modeled.

In order to focus the discussion, we assume that the number of attributes to be represented is a constraint in the analysis below. That is, we assume that we represent a certain number (D) of characteristics within the model to *some* degree, and simply vary the degree to which a given attribute is represented.

Suppose we now have a fixed level of computational resources at our disposal. We now consider how accuracy scales for each of the techniques for capturing heterogeneity.

5.1 Co-Flows

For the special case in which co-flows can be used to capture the statistic of interest (e.g. where the population is well-mixed with respect to the statistic of interest and the dynamics and attributes of entry and exit can be adequately captured), there is no need to disaggregate the system to represent heterogeneity. Computational resource demand is linear in D, and the statistic of interest can be calculated exactly over the population as one of the co-flow calculations.

5.2 Attribute-Based Stock Disaggregation

For attribute-based stock disaggregation, the computational resource demands depend only on the number of dimensions D and the level of detail with which the modeler represents each dimension (the number of width Δx of divisions into which each dimension is divided). In order to analyze the scaling behavior of the binning approach, we make the simplifying assumption that all dimensions (attributes) are continuous and are subdivided into segments of uniform size (Δx), thus dividing the attribute space as a whole into equal-sized hypercubes ("bins") of dimensionality D equal to the number of attributes being modeled. This restriction is less onerous than one might think: Given that the patterns of heterogeneity in the underlying population (and thus the density of the population across the attribute space) will in general be varying dynamically over the course of the simulation, the most advantageous decomposition of attribute space into stocks is frequently not clear a priori. Moreover, any attribute dimension can be linearly scaled to accord with the uniform size restriction.

Given these assumptions, the amount of error associated with calculating the population statistic varies as $\Theta\left(c^{-\frac{2}{D}}\right)$. The derivation of this mirrors the calculation of quadrature error bounds in basic calculus.

It is notable here that despite the sharp scaling with D, there is no scaling with N: We maintain the same number of stocks regardless of the size of the population.

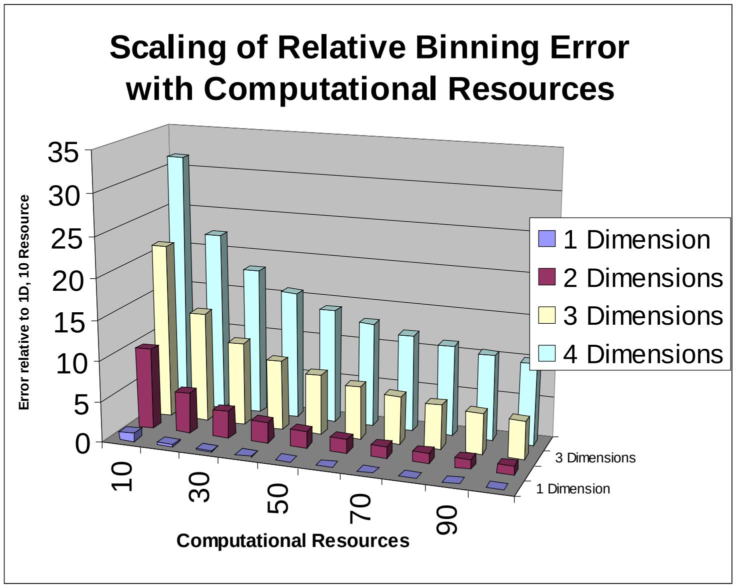

Figure 4: Scaling of Error in Computation of Population Statistics using Attribute-Based Disaggregation. Scaling is depicted over Computational Resources and Count of Dimensions of Heterogeneity

Intuitively, the formula for error means that as the number of dimensions to be represented doubles, the new error is the square root of the old error (for example, while the error in computing the statistics might be 10^{-4} for 2 dimensions, it will be 10^{-2} for 4 dimension, and 10^{-1} for 8 dimensions). Looked at in another way, the relationship between dimensions and error means that for any dimension D greater than two, a given increase in computational resources will not be met by a proportional decrease in error. Consider an increase in computational resources by a factor of α (say, 1%). Now consider the fractional decrease R(c, D) in error that results

$$R(c,D)=1-\frac{E(c,D)}{E\left((1+\alpha)c,D\right)}=\Theta\left(\left((1+\alpha)^{2}\right)^{-\frac{1}{D}}\right)$$

Figure 5 illustrates the scaling of R(.01,D) with D. For just one dimension (D=1), a 1% increase in computational power permits a 2% decrease in error. For two dimensions, the fractional decrease in error is very close to proportional to the increase in computational resources. For dimensions four and above, an increase in computational power yields a considerably smaller decrease in error.

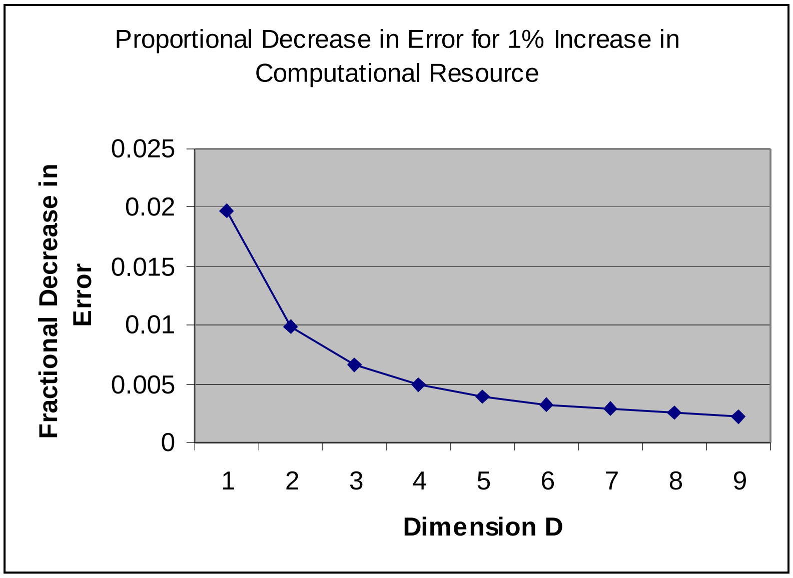

Figure 5: Scaling of Decrease in Error Associated from a 1% Increase in Computational Power with Number of Attribute Dimensions (Attribute-Based Disaggregation)

As shown in Figure 6, the scaling of error over dimension also means that the increase in computational power required to halve the observed error rises exponentially with the count of heterogeneity dimensions D. $R(c,D)=.5$ yields $\alpha = 2^{\frac{D}{2}} - 1$.

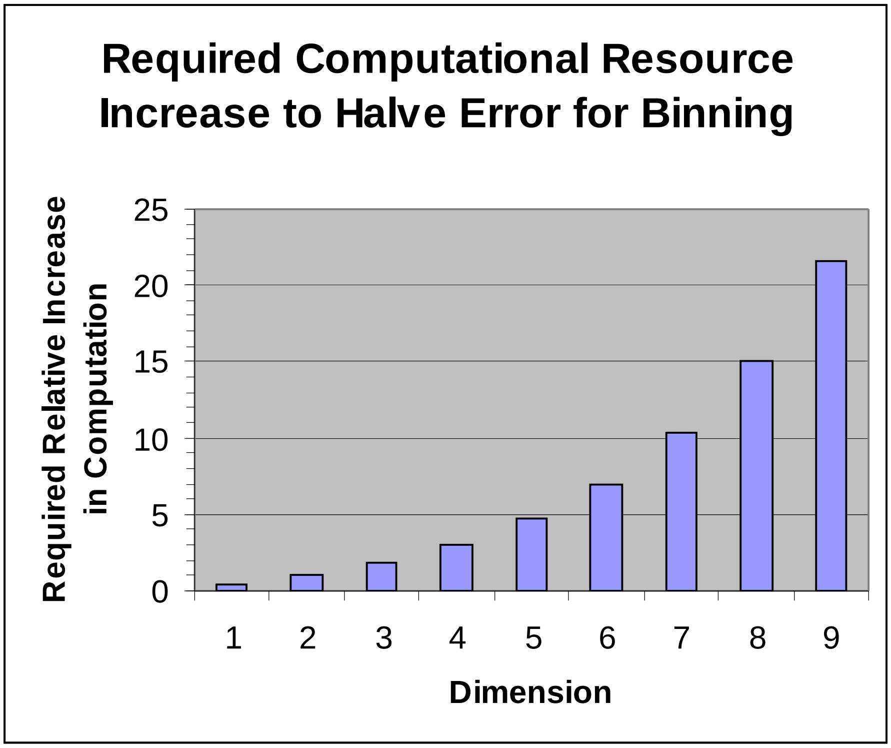

Figure 6: The Difficulty of Decreasing Error in Attribute-Based Disaggregation Approach to Representing Heterogeneity: For Dimensions Higher than 3, Halving Error Requires Increasing Considerably More than Doubling Computational Power.

5.3 Agent-Based Stock Disaggregation

For those cases in which agent-based disaggregation is being used, the primary means of lowering computational resource requirements is through *downsampling*. Using this method, we can form a bootstrap population by sampling with replacement from the full population of individuals, and then run the simulation on this bootstrap population.[4]

Now consider the estimate of the total of some population statistic over the true (real-world) population. Within the model, we estimate this total by computing a statistic over the population of agents in the model.

In this case, the standard error of the sample mean of this statistic over the downsampled population is $\Theta\left(\frac{1}{\sqrt{\frac{N}{s}}}\right)=\Theta\left(\sqrt{\frac{s}{N}}\right)$, yielding an estimate for error for the total over the population as $\Theta\left(N\sqrt{\frac{s}{N}}\right)=\Theta\left(\sqrt{Ns}\right)$. It is important to stress that in contrast to the case of attribute-based stock disaggregation, the error scales with the size of the population being simulated and *not* with the count of attributes being considered. Intuitively, the formula above means that if we halve the number of samples (agents) that we consider for a given population (and thus double the downsampling factor *s*), the error goes up by a factor of $\sqrt{2}$, or approximately 1.4 Conversely, in order to reduce the size of the error by a factor of 2 for a given sized population, we need to simulate 4 times as many samples (in other words, reduce *s* by a factor of 4).

Now consider a case where we have a fixed amount *c* of computational resources available. We further assume that demand for computational resources rises linearly with the number of agents whose rules are computed and linearly with the count of attributes that require updating for each such agent, and that it is too expensive to compute the evolution of the entire population. As a result, to remain using the same level of computational resources, the level of downsampling *s* required varies as $\Theta\left(\frac{ND}{c}\right)$. Given these assumptions, the error for the total over the population rises as $\Theta\left(\sqrt{Ns}\right)=\Theta\left(\sqrt{N\frac{ND}{c}}\right)=\Theta\left(N\sqrt{\frac{D}{c}}\right)$. This implies that for the regime in which our computational resources are insufficient to handle a full sample of the population, error in computing a statistic over the population rises linearly with population size N, with the square root of the dimension count and only inverse to the square root of the computational resources that are available.

In contrast to the case for attribute-based disaggregation, the scaling of error with computational resources is invariant in the number of dimensions involved. Increasing computational resources by a factor of α reduces error by a factor of $\sqrt{\alpha}$. Conversely, halving the error in such a situation requires

[4] We consider here only the error for the trivial case where we compute statistics using only a single bootstrap realization, leaving to a forthcoming paper the more general case of repeated bootstrap samples drawn from the population.

increasing computational resources by a factor of four –independent of the number of dimensions involved.

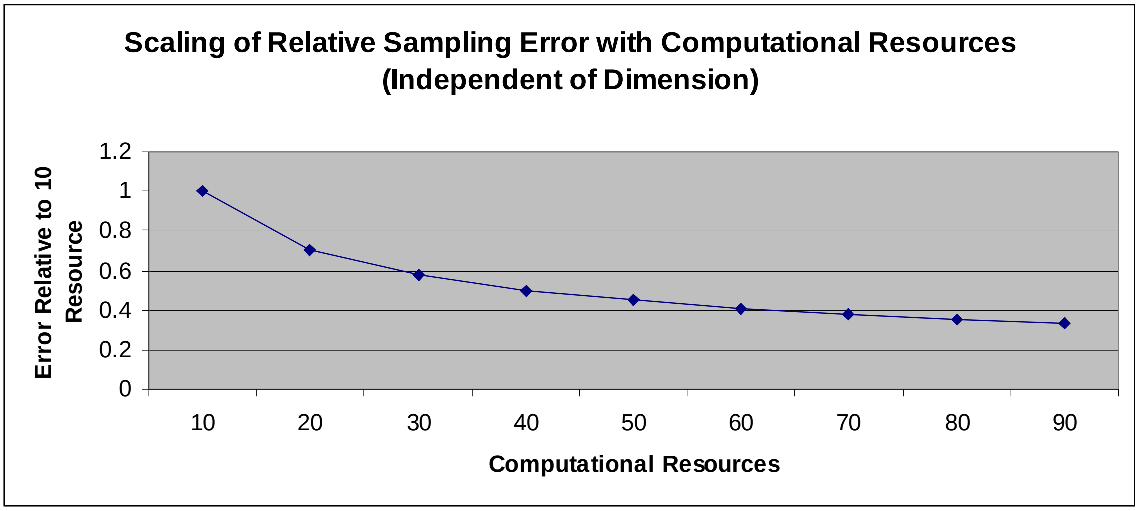

Figure 7: Scaling of Relative Error with Respect to Increase in Computational Resources (Agent-Based Disaggregation). The Results Shown are Independent of Count of Attribute Dimensions.

6 Conclusion

The results presented above are scaling relationships and are not meant to numerically estimate the exact amount of work required for any method; there are likely to be significant constant factors hidden within the scaling notation that have significant performance implications for a given size of problem. Nonetheless, the results above can provide substantial insights into tradeoffs between the methods.

	Generality	**Computational Resource Demand** **For dimension count (D), Population Size(N)**	**Error** **For computational resources(c), dimension count (D), Population Size(N)**
Agent-Based disaggregation	High	$\Theta\left(\frac{ND}{s}\right)$	$\Theta\left(N\sqrt{\frac{D}{c}}\right)$
Attribute-based disaggregation	High	$\Omega\left(2^{D}\right)$	$\Theta\left(c^{-\frac{2}{D}}\right)$
Co-flows	Low	$\Theta(D)$	$\Theta(1)$

What is clear is that as the number of dimensions rises, agent-based disaggregation methods rapidly dominate both the accuracy and performance of attribute-based disaggregation approaches. In particular, we have seen that for the attribute-based stock disaggregation, the size of estimation error rises rapidly as dimensions increase. Decreasing errors by a given factor requires an increase in computational power that is exponential in the number of dimensions D. This reflects the familiar "curse of dimensionality".

By contrast, the error associated with agent-based disaggregation is independent of the number of dimensions of heterogeneity being considered. For populations large enough to prevent direct simulation, decreasing errors by a certain factor α requires increasing computational resources by α^2 – independent of the dimensionality of the model.

While attribute-based disaggregation fares poorly for large number of dimensions of heterogeneity, attribute-based methods have an advantage for models of low dimension and large population. In particular, the costs associated with agent-based representations rise linearly with population size while the resource demands of attribute-based methods are invariant with respect to population size, as the size of the population to be simulated (N) rises, attribute-based methods will slowly gain an edge in accuracy, performance and storage requirements over agent-based methods.

Broadly speaking, we can say that attribute-based stock disaggregation models are best suited to problems with low number of attribute dimensions (say, less than 3 or 4) or very large population size, while agent-based models using simple downsampling to throttle computation are well adapted to models that exhibit important heterogeneity with respect to medium or high numbers of attribute dimensions and population sizes within a few orders of magnitude of that which can be simulated directly. Co-flow based techniques offer a highly precise and computationally frugal approach for capturing heterogeneity in those specialized instances in which few statistics on attribute values are required, in which the population is well-mixed with respect to the statistic of interest and where the dynamics and attributes of entry and exit can be adequately captured.

Before concluding, it is worth noting that this paper has explored just two of many tradeoffs that require consideration in the selection of a modeling technique to capture heterogeneity. Discussion of these other considerations lies outside the scope of paper, but it is worth mentioning a few of them. The first of these is data availability. Partly to achieve reliable sample sizes in descriptive statistics, secondary data is sometimes available only following attribute-based disaggregation (see, for example, [Census Projections]); in many other cases (such as in many modern health data sets, [Current Population Survey] [NHIS] and particularly in longitudinal studies such as [NHANES]), individual sample data is available to researchers. The form in which data is received can strongly influence the convenience – and sometimes the feasibility – of different schemes for representing heterogeneity. Although the computational error resulting from representing heterogeneity in the scheme most closely matched to the data may be higher than for other techniques, selection of this technique can lower the risk of error in data manipulation and provide the modeler with additional time to refine the model and increase overall model accuracy. A second additional consideration involves the selection of a modeling framework. As touched on in Section 3.4, modeling frameworks often differ in their expressiveness with respect to different schemes for representation of heterogeneity. If considerations such as other characteristics of a system, software accessibility, tools or client requirements play to the favor of one type of modeling framework, the modeler must consider the impact of framework choice on the representation of heterogeneity, and must balance the tradeoffs described here with these other considerations. Finally, the desired scheme for model calibration must be consistent with the approach chosen for representing heterogeneity. For example, agent-based models simulating large populations require considerably different model calibration techniques than do highly aggregated models of the same populations. Given that model calibration can require considerable computational resources and significant impact model error, when planning a model, the modeler would do well to at the outset to jointly select a calibration method and a technique for representing heterogeneity.

7 Bibliography

[Ackerman] Ackerman , E. Simulation of micropopulations in Epidemiology: A series of tutorials illustrated by coronary heart disease models. Tutorial I: Simulation: An introduction, *Int J Biomed Comput*, 36 (1994) 229-238.

[Bass Diffusion] Sterman, John. Bass Diffusion Model. Section 9.3.3. of *Business Dynamics*. Boston: McGraw-Hill Higher Education. 2000.

[Bootstrap] B. Efron and R. J. Tibshirani. Introduction to the Bootstrap. Chapman and Hall, New York, 1993.

[Census Projections] U.S. Bureau of the Census, Population Division, 1996. Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050 -- Middle Series Vital Rate Inputs. Available at:http://www.census.gov/population/www/projections/natvital.html.

[Current Population Survey] U.S. Bureau of the Census. Current Population Survey [Computer file]. ICPSR version. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 1994. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1995. Available at: http://www.icpsr.umich.edu.

[Epidemic] Sterman, John. Modeling Acute Infection. Section 9.2.2 of *Business Dynamics*. Boston: McGraw-Hill Higher Education. 2000.

[Flowers] Ford, A. S-Shaped Growth. Chapter 6 of *Modeling the Environment*. Washington DC: Island Press, 1999.

[Kaibab] Ford, A. The Kaibab Deer Herd. Chapter 16 of *Modeling the Environment*. Washington DC: Island Press, 1999.

[NHANES] National Health and Nutrition Examination Survey I Epidemiologic Followup Study. Atlanta, GA: Centers for Disease Control and Prevention; 1992. Available at:ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHEFS/NHEFS92.

[NHIS] National Health Interview Survey. Health Promotion and Disease Prevention Supplement [Computer File]. ICPSR version. Washington DC: U.S. Dept. of Health and Human Services, National Center for Health Statistics; 1991

[Resources] Ruth, M. and Hannon, B. Managing Open Access Resources. Chapter 25 in *Modeling Dynamic Economic Systems*. New York: Springer-Verlag, 1997.

[Tobacco Policy Model] Tengs, T., Osgood, N. and Lin, T. Public health impact of changes in smoking behavior: results from the Tobacco Policy Model. *Medical Care*. 2001 Oct;39(10):1131-41.

[Roberts, Homer et al] Roberts, E., Homer, J., Kasabian, A., and Varrell M. A Systems View of the Smoking Problem: Perspective and Limitations of the Role of Science in Decision-Making. *Int. J. Bio-Medical Computing* (13) (1982) 69-86.

[Shepard Zeckhauser 1982] Shepard, D.S. and Zeckhauser, R.J. The Choice of Health Policies with Heterogeneous Popu1ations. In Economic Aspects of Health. Fuchs V.R., ed. Chicago: University of Chicago Press, 1982, 255-297.

[Shepard Zeckhauser 1980] Shepard, D.S. and Zeckhauser, R.J. Long-Term Effects of Interventions to Improve Survival in Mixed Populations. *Journal of Chronic Diseases* 33, 1980, 413-33

Back to the Top