Computer Simulation and Statistical Modelling: Rivals or Complements?

Thomas K. Burch
Population Research Group
University of Victoria
Victoria, British Columbia, Canada

[Revised, 19 May 2005]

1. Introduction

Social science is rife with polarities, defined by the *Oxford English Dictionary* as 'the possession or exhibition of two opposite or contrasted aspects, principles, or tendencies.' There is the fundamental polarity inherent in all empirical science between data and theory – in John Locke's phrase, 'experience and reflection on experience.' There is a polarity between the micro and macro levels of analysis [Courgeau [2004] speaks of 'une opposition macro-micro'] with at best partial synthesis of the two.

In some of the social sciences, notably sociology and political science, there are polarities between quantitative and qualitative research, between empirical research and critical analysis, and between value-free and explicitly ideological social science.

Less widely discussed are polarities involving different methodological traditions **within** quantitative social science. These affect many scientific disciplines, including demography. They have led to tension and at times hostility, thereby weakening empirical social science in its central tasks, and in its confrontations with post-modernist critics and old-fashioned radical positivists. Not least of these is the polarity between statistical modelling, viewed as the fitting of models to observational data, and computer modelling [simulation], viewed as an attempt to represent some portion of the real world, or some theory about the real world, in a way that goes beyond observational data. The statistician Adrian Raftery, paraphrasing C.P. Snow, refers to 'two main cultures of quantitative research – statistical modeling and deterministic simulation models…' noting that the proponents of the two approaches seldom interact [see Raftery, 2004].

In this paper, I examine the last-mentioned polarity as it manifests itself in demography and related empirical social sciences. My central argument is that statistical modeling and computer simulation are best viewed as complementary not competing modes of analysis. I attribute much of the tension between the two approaches to a continuing misunderstanding of the interrelations among data, models, theory and reality, and to confusion about the epistemological character of different kinds of demographic analysis [Burch, 2003d]. A particular focus is on the failure to recognise that it is not the **form** of an analytic tool, but its **application**, the use to which it is put, that determines the epistemological character of an analysis.

Viewed in this light, much of the tension surrounding simulation and statistical modelling derives from the fact that the two approaches naturally tend toward different uses, simulation toward theoretical analysis, and statistical modelling toward empirical analysis. The polarity is at base the familiar polarity between theory and experiment – or, viewing science as a human institution, between theorists and empiricists -- still very much unresolved in contemporary social science, but seen as a natural division of labour in more mature sciences.

To set the stage for our discussion, Section 2 presents a 2x2 table illustrating the crucial distinction between the form of an analytic tool and its uses or applications. Section 3 considers the dichotomy between models of data and models of the real world [or theories about the real world]. Section 4 softens the earlier dichotomies with examples of mixed forms of analysis. I conclude with a comment on the abstract – and therefore incomplete -- character of all scientific knowledge, whether empirical or theoretical.

2. Analytic tools and their disparate uses

Much of the methodological confusion surrounding the epistemology of theory, models, and data can be clarified by emphasising the distinction between an analytic tool and the purpose for which it is being used. A simple two-by-two table can help. Figure 1 emphasises the fact that the **same** mathematical or statistical tool can be used for **different** kinds of scientific analysis, that **different** tools can be used for the **same** purpose – and so on for the other two cells – **same/same** and **different/different**. The table can be used to classify demographic work, with examples readily available for each of the cells, although the distribution is far from even.

Figure 1: Analytic Tools and Their Uses

a] Same tool, same purpose	b] Different tools, same purpose
Example: standardised use of cohort-component model for population projection	Example: exponential, logistic, cohort component, systems, etc. models for population projection
c] Same tool, different purposes	d] Different tools, different purposes
Example: cohort-component model for forecasting [prediction], simulation, and prospective analysis [Romaniuc, 1990]	Example: Statistical models for data analysis; differential equations for theoretical analysis

Cell a [**same/same**] might be thought of as containing stylised or stereotypical demographic analysis, in which the same analytic apparatus is used over and over for the same purpose. A classic example is the use of cohort-component projection algorithm for the purpose of population forecasting. This use has been canonised by governments and other official organisations, such as the United Nations Population Division and the World Bank. In demographic texts, the technique is routinely presented as the standard forecasting technique, sometimes the only technique.

This has resulted in comparability among population forecasts by various governments and other agencies, and insured a certain kind of correctness in procedure. But it also has discouraged – or at least not encouraged -- exploration of other possible approaches to population forecasting, the use of different tools for the same general purpose. The identification of population forecasting with cohort-component forecasting is related to what

Oeppen and Wilson [2003] have termed *reification* in demography – the confusion of a particular abstract measure or model in demography with the underlying real-world process to which it pertains. [1] The continued use of the same technique, and virtually only that technique, for the same purpose is not inherently fallacious, but it tends to lead to reification by sheer repetition and habit. We tend to develop tunnel vision.

For the purpose of population forecasting, cell b [**same purpose/different tools**] would include, as well as cohort-component projection: the Leslie matrix; the exponential and logistic functions, and their extensions [common in population biology and ecology]; systems models for projection [e.g., with feedback from outcomes to inputs]. In the beginnings of modern population forecasting in the early decades of the 20th century, a variety of methods was explored, before the cohort-component approach became dominant almost to the point of monopoly [de Gans, 1999; Burch, 2003a]. Only recently has the field begun once again to explore alternatives to the standard model, such as the addition of feedbacks, socio-economic and environmental variables, and stochastic inputs [with confidence intervals on outputs].

Another classic example of the use of different tools for the same purpose is the use of both the stable population model and cohort-component projection in order to clarify the effects of changing mortality, fertility, and migration on population age structure. Each approach yielded the same general conclusions [e.g., the centrality of fertility decline in past population aging], but each shed distinctive light on various facets of the problem, and some problems were more tractable using one approach rather than the other [e.g., the role of migration and transient dynamics]. [2]

The stereotypical approach to population forecasting and other demographic analyses – the tendency to identify one tool with a particular use and only that use -- has also had the result that cell c [**same tool for different purposes**] is not as full of examples as might be desirable. The regular presentation and use of the cohort-component projection algorithm for projection, for instance, helped obscure its value for other purposes. Romaniuc [1990] has provided the most systematic exposition of this point in his discussion of the standard projection model as 'prediction,' 'simulation,' and 'prospective analysis.' In general, he views the uses of the algorithm as being ranged on a continuum, from the most realistic [prediction] to the least realistic [simulation]. Burch [2003b] has developed the same line of thought with respect to the life table, arguing that at base it is a theoretical model of cohort survival, with the description of period mortality rates only one of its uses. But current demography texts tend to remain in cell a – the projection algorithm is for forecasting, the life table is for measuring current mortality.

---

[1] Abstraction in the standard cohort-component projection model include its deterministic and linear character , the absence of feedbacks and interrelations among input variables, and the absence of any but demographic variables. [environmental, economic, social, etc.].

[2] The habit of thinking in terms of many different tools for the same general purpose is common among demographers when it comes to measurement [cf. the large variety of measures of fertility], but not so common when it comes to other kinds of analysis. The idea of a 'toolbox' of theories and models is central to the model-based view of science among philosophers [see Giere, 1999, 1988]. It is found in the work of some empirically inclined social scientists [Coleman [1964] on 'sometimes-true theory'; Meehan [1968]; Keyfitz [1975]. The more influential doctrine of scientific procedure has been logical empiricism, which aims toward discovery of the one true theory.

There is nothing novel in the idea that a given analytic tool can be used for several different purposes. Coleman begins his classic work on mathematical sociology [1964] by outlining four major – and different – uses of mathematics in science. The idea is implicit in the adoption of mathematics by so many different disciplines and in the use of some mathematical concepts in so many different substantive contexts. Witness the exponential function, used to study radioactive decay, population growth, interest rates and discounting, and fixing the time of death by body temperature, to mention only a few. In demography, the exponential is usually presented as a 'technique,' even though many of its actual uses are speculative and theoretical in character – doubling times, *reductio ad absurdem* projections [e.g., human population for the next 2000 years at 1.5% per annum].

But some tools are closely wedded to particular kinds of analysis, and this leads to cell d [**different tools for different purposes**]. Statistical models, for example, are designed for analyzing data, for summarising and describing data sets, including relationships among variables. But they are not so useful for the statement of theoretical ideas or the representation of real world systems. For this, one needs tools that can deal with unobserved entities or unmeasured variables in a way that goes beyond simply combining them in an error term. Section 3 elaborates on this point.

3. Modelling Data and Modelling Ideas About the Real World

A key distinction is between models of a set of empirical observations [data] and models of a set of ideas about real-world processes. Wickens [1982] provides a nice statement of the point, noting that 'statistical models describe what the data look like, but not how they came about,' and proceeds to suggest ways for 'interpreting the data through a description of the mechanisms that underlie them' [p.9; see also Hedström and Swedberg, 1998]. There is a sense in which data are outcomes of a partly 'black box' whose inner working must be modelled in some other way.

The most common multivariate statistical models such as linear regression and its refinements are often interpreted as though they represent underlying real-world processes. But, as Abbott [1988] has argued convincingly, such an interpretation more often than not involves the fallacy of reification, the invention of a 'general linear reality,' a social world that is linear and largely atemporal. Abbott distinguishes this 'representational' interpretation of linear models from their more appropriate 'entailment' use – if a theory or hypothesis is sound, then I should find a certain statistical structure in a relevant data set. But now statistical modelling is being used not to state some theoretical idea, but to test it. Traditional statistical analysis is pre-eminently a tool of empirical research.[3]

---

[3] A qualification is needed on this point, based on the recognition of two very different kinds of theory in science. Cartwright's [1983] finds a common distinction in physics between 'phenomenological' and 'fundamental' theory. The former deals with empirical regularities and 'laws' [e.g., Newton's law of falling bodies], without delving very deeply into explanatory mechanisms. Coleman [1964, pp.34-52] makes a related distinction between 'synthetic' and 'explanatory' theory. Insofar as statistical analysis yields findings of strong empirical regularities, even universal relationships, it can provide the building blocks of a theoretical system. This has been precisely the logical empiricist program for science. But in the human sciences, strong empirical laws are sufficiently rare that some other approach to theory development is required [see Meehan, Keyfitz, Cartwright, Giere, Teller]. Theory must ultimately be based on empirical research, but, in the face of culture and history and an absence of universal generalisations, it does not

The differential equation, by contrast, is pre-eminently the tool of theory. Lotka puts it as follows:

> In the language of the calculus, the differential equations display a certain simplicity of form, and are therefore, in the handling of the theory at least, taken as the starting point, from which the equations relating to the progressive states themselves, as functions of time, are then derived by integration [1956, p.42].

In a footnote he adds: 'In experimental observation usually [though not always] the reverse attitude is adopted.' The study of data on the progressive states of a system over time or at a point in time [as is often the case in contemporary social science] is a matter of statistical analysis.

As I have argued elsewhere, the infrequent use of differential equations in demographic research may be due in part to demography's relative lack of interest in theory [Burch, 2004].

Cell d has occasioned more than a little confusion regarding the uses of statistics and modelling, including differential/difference equations models. Much of the confusion stems from a mistaken notion that multivariate statistical analysis can yield essentially theoretical propositions as results. As noted earlier, Abbott [1988] has cogently argued against this idea, typified in his quote from Blalock [1960] that 'These regression equations are the "laws" of science" [p.275], a statement firmly rooted in logical empiricism. Abbot's review of the differing fortunes of Blalock's *Social Statistics* [featuring multiple regression] and Coleman's *Introduction to Mathematical Sociology* [featuring differential equations] suggests the popularity of Blalock's view. Coleman's work had gone out of print by the time Abbott wrote; Blalock's text was regularly reprinted and published in new editions. The notion that theory essentially moves beyond the data has been resisted by many quantitative social scientists.

A similar confusion is found in some writings of the systems dynamics school of modelling, based on the pioneering work of Jay Forrester at MIT. In several of their works, representatives of this school extol the superiority of systems dynamics modelling over the use of statistical analysis, as if the two served the same scientific function. In one recent manual, multiple regression is held up to ridicule [for an elaboration, see Burch, 2002]. But there is little recognition that most of their systems dynamics models were in fact theoretical, in the sense of speculative and untested. Although they contain some data and often try to reproduce some data series as output, they are not fundamentally empirical. They were often criticised by empiricists, who found them fanciful, and by theorists, who found them lacking in firm theoretical grounding. As Hummon [1990] remarks of computer simulation, 'Applications in the past have tended to focus on large, complex systems. Simulation models of complex organizations, cities, the world environmental system, were the standard fare…. Of course, coherent theories for these phenomena did not exist.' He goes on to note that more recent applications of computer modelling to theory construction are more focused, 'often tied to existing theoretical models.'

---

simply flow from it through the application of inductive logic. See the contemporary theoretical physicist Roger Newton [1997] on theory as an act of the creative imagination.

4. Hybrids and Mixed Forms: Revisiting the Dichotomies

The distinction developed in the previous section is important but it is not absolute. One should think of it in terms of the emphasis or even spirit of a particular piece of analysis. Is it mostly about the data, or is it mostly about the theoretical ideas? Some analyses lie at one or the other pole, for example, regression analysis of data with few if any guiding theoretical ideas, or at the opposite extreme, pure theoretical speculation oriented only casually toward empirical observation.

The advance of any science involves an ever closer intermingling of data and theory. Theory must ultimately be evaluated in terms of its ability to explain or predict empirical observations. And data, at least in a scientific context, is meaningful only if it is collected with an eye to theoretical development. Recent developments in statistics, theory and computer modelling have tended to blur the distinction and promote healthy intermingling.

Some statistical models, for example, have moved a little closer to theoretical models. Structural equations models such as path analysis are a case in point. When properly used, they are fit only after one has formulated at least a primitive model of the process at hand. Assumptions must be made about temporal ordering, causal linkage, and direct and indirect effects. A path model begins to unravel the complex process that links outputs to inputs. They are rudimentary mechanistic models. Such models remain largely statistical and empirical, however, insofar as they include only variables that have been measured [at least indirectly, as in factor analysis, latent variables, and similar techniques].

By the same token, computer modelling as theoretical elaboration has begun to incorporate elements of statistical analysis. The development of stochastic population projections is a case in point [see, for eample, Raftery, Givens, and Zeh, 1995; Lee, 1999; Keilman, Pham, and Hetland 2000].

These blendings can require a re-examination of categories like *stochastic* and *deterministic*. Consider a largely deterministic systems model that has been slightly modified by the addition of random terms to one of the key relationships [$y = (f)x + \alpha$, where α is some sort of random term], and by the inclusion of one or more conditional statements involving such a 'randomised' variable. This could represent some sort of threshold, such that very different outcomes result from the variable being above or below the threshold. An example might be differential equations models of species extinction once the population declines below a certain size. Now it is a case of a random event triggering a qualitatively different response in an otherwise deterministic model. Do we call such a model deterministic or stochastic?

Another example of a movement of computer modelling toward statistics can be found in more recent incarnations of dynamics systems software, characterised earlier as somewhat hostile toward statistical methods. Goldsim, a relatively new product, emphasises the addition of random terms to model variables [as described just above], and the running of multiple simulations to yield both average values and variance of results. Recent versions of older software such as Modelmaker or Vensim emphasise statistical tests for goodness of fit to the data. More importantly, they include optimisation routines to find some parameter values that yield the best fit [with already known parameter values fixed, at least within given ranges]. Statistical methods are being used to specify some key elements of a complex model. But many

elements – perhaps the majority – remain beyond the ken of empirical observation. One has a theoretical model that is partly estimated statistically.

The emerging field of computational modelling [agent-based or rule-based models] provides many more examples of an intermingling of determinism and chance [in demography, see Billari and Prskawetz, 2003]. In earlier micro-simulation such as Monte-Carlo models, events occur mainly by chance, according to various probability distributions [but see some early models of family and household formation, in which kinship rules play a central part; see, for example, Wachter, 1987]. In agent-based models, chance still operates at many points. But central to most models are strong behavioural rules which are determinative, that is, not subject to chance.

5. Concluding Comments

Our conventional views of empirical social science need revision. In particular, demography and mainstream empirical sociology need to develop a more sophisticated approach to theoretical models. We need to reconsider the superordinate position we have granted highly flawed and limited statistical analyses, which have regularly be used to 'disprove' perfectly sound and useful theory. Statistical analysis may be closer to the data, but that does not necessarily mean it is closer to reality. A carefully crafted theory or model, which can include unmeasured – even unmeasurable variables – may be a better representation of reality for many purposes.

Demographers and other highly quantitative social scientists often think of statistical analysis of data as solid and hard-headed, firmly grounded in reality – in sharp contrast with the verbal speculation of theorists or the 'made-up' numbers of simulators. But the epistemological differences between theory, modelling and statistical analysis are not as great as our conventional thinking would have it. Statistical analysis is not the bedrock it often is taken to be.

Empirical data sets which we subject to statistical analysis are abstract representations of concrete reality; they are partial, selective, over-simplified depictions of some complex concrete real-world system. The data set does not constitute theory in any meaningful sense, but it often is shaped by the influence of implicit theoretical assumptions as to what data are important. There is no such thing as pure empirical description. There always is selection. Each datum is empirical – and real to the extent of its precision – but the assumption that the overall data set represents an object or system is just that, an assumption.

The standard multivariate statistical models also are highly abstract creations of the human mind. They assume a particular mathematical structure among a limited set of variables, whether that structure exists in the real world or not. If they are thought to represent the world, they are almost certainly grossly over-simplified representations, which is not to say that they may not be useful for some purposes. But an abstraction – say, a linear or log-linear model – added to an abstraction – a data set as described above – does not yield absolute truth about the concrete world. Like theory, they are selective and partial representations. They may be useful for some purposes, but that does not make them true. If such multivariate models are viewed,on the other hand, as 'entailments' of some well-developed theory, then they become analogous to experimental results. They do not of themselves explain anything; they only indicate that an explanatory theory has some plausibility.

7

These limitations of statistical analysis of data are the basis for the notion of the underdetermination of theory by empirical research [see Turner, 1987]. A striking statement of the problem is provided by Bracher, Santow, Morgan and Trussell [1993]. After a state-of-the-art analysis of unusually rich survey data, they comment: 'However detailed and comprehensive the "explanatory" factors that we have had at our disposal, they are, after all, only dim reflections of the possibly unmreasurable factors that keep marriages together or drive them apart…' [p.423]. It is precisely the role of theory to go beyond such dim reflections.

But, as suggested above, the two kinds of work are not so much diametrically opposed as lying toward the opposite ends of a continuum of human attempts to describe and understand the real world of human behaviour. All human knowledge, including empirical statistical analyses, is a human invention, a construct. Like theory or modelling, it is selective, abstract, limited, incomplete, provisional – in short relative, not absolute. There are important differences among different kinds of social scientific knowledge and the processes that generate them. But their fundamental epistemological character is the same. Many of our beliefs often are held as absolutes, as fundamentally true. Scientific knowledge, by contrast, can aspire at best to 'realism without truth' [Giere, 1999].

References

Abbott, Andrew. 1988. Transcending general linear reality. *Sociological Theory* 6:169-186.

Billari, Francesco C.; and Prskawetz, Alexia. 2003. *Agent-Based Computational Demography Using Simulation to Improve Our Understanding of Demographic Behaviour.* Heidelberg: Physica-Verlag.

Bracher, Michael; Santow, Gigi; Morgan, S. Phillip; and Trussell, James. 1993. Marriage dissolution in Australia: models and explanations. *Population Studies* 47:403-426.

Burch, Thomas K. 2002. Computer modelling of theory: explanation for the 21st century. In Robert Franck [ed.], *The Explanatory Power of Models: Bridging the Gap Between Empirical and Theoretical Research in the Social Sciences.* Boston: Kluwer Academic Publishers. Pp.245-265.

Burch, Thomas K. 2003a. The cohort-component population projection: a strange attractor for demographers. In Fleischhacker, de Gans and Burch [eds.] *Populations, Projections and Politics.* P. 39-58.

Burch, Thomas K. 2003b. The life table as a theoretical model. Paper presented at annual meetings of the Population Association of America, 1-3 May 2003, Minneapolis, Minnesota.

Burch, Thomas K. 2003c. Demography in a new key: a theory of population theory. *Demographic Research* 9[11]:263-284.

Burch, Thomas K. 2003d. Data, models, theory and reality: the structure of demographic knowledge. In Francesco C. Billari and Alexia Prskawetz [eds.] *Agent-Based Computational Demography.* Heidelberg, New York: Physica-Verlag, 2003. Pp.19-40.

Burch, Thomas K. 2004. Does demography need differential equations? Seminar presentation at the Center for Studies in Demography and Ecology, University of Washington, 12 March 2004. 12pp.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Clarendon Press.

Coleman, James S. 1964. *Introduction to Mathematical Sociology.* New York: The Free Press.

Courgeau, Daniel. 2004. *Du Group À L'Individu: Synthèse Multiniveau.* Paris: Èditions de L'Institut National D'Ètudes Démographiques.

De Gans, Henk A. 1999. *Population Forecasting 1895-1945: The Transition to Modernity.* Dordrecht: Kluwer Academic Publishers.

Fleischhacker, Jochen; de Gans, Henk A.; and Burch, Thomas K. [eds.]. 2003. *Populations, Projections and Politics: Critical and Historical Essays on Twentieth Century Population Forecasting.* Amsterdam: Rozenberg Publishers.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach.* Chicago: University of Chicago Press.

Giere, Ronald N. 1999. *Science Without Laws.* Chicago: University of Chicago Press.

Hedström, P.; and Swedberg, R. 1998. *Social Mechanisms: An Analytic Approach to Social Theory.* Cambridge: Cambridge University Press.

Hummon, Norman P. 1990. Computer simulation in sociology. *Journal of Mathematical Sociology.* 15:65-66.

Keilman, Nico; Pham, Dinh Quang; and Hetland, Arve. 2000. Why population forecasts should be probabilistic: illustrated by the case of Norway. *Demographic Research* 6(15):409-453.

Keyfitz, Nathan. 1975. How do we know the facts of demography? *Population and Development Review* 1:267-288.

Lee, Ronald. 1999. Probabilistic approaches to population forecasting. In Wolfgang Lutz, James Vaupel and D. Ahlburg [eds.] *Frontiers of Population Forecasting.* Supplement to volume 24, *Population and Development Review.* Pp. 156-190

Lotka, 1956. *Elements of Mathematical Biology.* New York: Dover Publications. [First published in 1924 under the title *Elements of Physical Biology.*]

Meehan, Eugene J. 1968. *Explanation in Social Science: A System Paradigm.* Homewood, Ill.: The Dorsey Press.

Newton, Roger. 1997. *The Truth of Science: Physical Theories and Reality.* Cambridge, MA: Harvard University Press.

Oeppen, Jim; and Wilson, Chris. 2003. On reification in demography. In Fleischhacker, de Gans, and Burch [eds.] *Population, Projections and Politics*. Pp. 113-129.

Raftery, Adrian E. 2004. See relevant section of personal web page at http://www./stat.washington.edu/raftery/research/dsm.html

Raftery, Adrian E.; Givens, Geof. H.; and Zeh, Judith E. 1995. Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association* 90:402-416.

Romaniuc, Anatole. 1990. Population projection as prediction, simulation and prospective analysis. *Population Bulletin of the United Nations.* 29:16-31. See also revised version in *Canadian Studies in Population* 30(2003) 35-50.

Teller, Paul. 2001. The twilight of the perfect-model model. *Erkenntnis* 55:393-415.

Turner, S.P. 1987. Underdetermination and the promise of statistical sociology. *Sociological Theory* 5:172-184.

Wachter, Kenneth W. 1987. Microsimulation and household cycles. In John Bongaarts, Thomas K. Burch [eds.] *Family Demography: Models and Their Applications.* Oxford: Clarendon Press. Pp. 215-227.

Wickens, Thomas D. 1982. *Models for Behavior: Stochastic Processes in Psychology.* San Francisco: W.H. Freeman.

*** *** ***