Complexity and Scientific Modelling

Bruce Edmonds,
Centre for Policy Modelling,
Manchester Metropolitan University.

1 Overview

There have been many attempts at formulating measures of complexity[1] of physical processes (or more usually of the data models composed of sequences of measurements made on them). Here we reject this direct approach and attribute complexity only to *models* of these processes in a given language, to reflect its "difficulty". This means that it is a highly abstract construct relative to the language of representation and the type of difficulty that concerns one.

A framework for modelling is outlined which includes the language of modelling, the complexity of models in that language, the error in the model's predictions and the specificity of the model (which roughly corresponds to its refutability or to the information it gives about the process).

Many previous formulations of complexity can be seen as either:

- a special case of this framework using particular modelling languages and focusing on particular types of difficulty;
- attempts to "objectify" complexity by considering only minimally complex models or its asymptotic behaviour;
- relativising it to a fixed mathematical structure in the absence of noise;
- misnamed in that they capture the specificity rather than the complexity.

Such a framework makes sense of a number of aspects of scientific modelling.

1. As a result complexity is not situated between order and disorder, as several authors have assumed, but rather such judgements arise given certain natural assumptions about the language of modelling and the desirable trade-offs between the model complexity, its specificity and its error rate.
2. Noise can be seen as that which is unpredictable given the available resources of the modeller. In this way noise is distinguished from randomness. Different ways of practically distinguishing noise can thus be seen as resulting from different trade-offs between complexity, error, specificity and the choice of modelling language.
3. Complexity is distinguished from concerns of specificity such as refutability, entropy and information. Complexity is thus seen to have context-dependent relations with such measures but in general is independent from them.

[1] Throughout I am distinguishing a lack of complexity from that of *simplicity*, which has come to mean something slightly different in philosophy (see section 8).

4. Less complex models are not a priori likely to be more accurate, but rather that given the typical structure of expressive modelling languages and our limitations in searching through such languages, choosing the simpler model can be a useful heuristic.

2 Complexity[2]

There is an understandable wish to measure the complexity of real systems rather than just of models of systems, but if natural systems have inherent levels of complexity they are beyond us. In practice there is no practical upper bound on their complexity as one has only to consider them in more detail or by including more aspects. On the other hand, the effective complexity of systems does depend on our models - the exact motions of the planets may be puzzling when you have to describe them in terms of epi-cycles but much simpler in terms of ellipses.

It may be objected that even if such “real complexity” is so intractable, one can still make comparative judgements; i.e. it is natural to judge some natural systems as more complex than others. An example of this is the claim that a cell is simpler than a whole organism. However in defending such judgements one is always forced into relating that which is compared within a common model. It is only once you have abstracted away what you consider to be irrelevant details, that such judgements of relative complexity become evident. It may be argued that some such models and frameworks are privileged but this pre-judges decisions about relevance and so is not helpful to an analysis of complexity and its place in scientific modelling.

In any case complexity is more critically dependent upon the model rather than what is modelled. So we will approach it from this point of view and leave the reader to judge whether this distinction is fruitful in understanding the processes involved.

For our purposes we will define complexity thus:

> *“The difficulty associated with a model's form when given almost complete information about the data it formulates.”*

This is a special case of the definition given in [5].

The relevant type of “difficulty” depends somewhat upon your goals in modelling, but here it will indicate the difficulty in finding the model in a search starting at the smallest model forms. This could be size, depth or some indication of the computation that is necessary to discover it.

3 A Framework for Analysing Modelling

To frame the discussion of complexity, I will outline a model of modelling.

Firstly it is important to distinguish between the form and predictive meaning of such models. The models themselves are always held in some form. The set of such possible forms can be considered as a language in its broadest sense - frequently it may correspond closely to an actual language, either natural or formal. I will call this the modelling language. Such models are amenable to some form of inference, in that they can be used to predict some property given some other information (even if sometimes some

[2] Some holist like to reserve the word “complexity” for natural systems, if you are one please read “complication” for “complexity” where appropriate.

of the necessary information is only available after the predicted event). At least some of the information comes from what is modelled in the form of measurements. The models correspond, loosely, to what they model via these predictions.

Thus we distinguish two aspects of a model: its form and the correspondence between possible information and the predictions that one could infer from it. This set of information along with the respective predictions can be thought of as defining a subspace of the space of all relevant possibilities. I will call this subspace the model's semantics, because one can draw an analogy between a logic's syntax and its semantics in terms of the set of logical models a statement is true for.

The primary way in which these models can be judged is by the degree of correspondence between what is modelled and the predictions of the models - its error. This however does not rule out the default model, that "anything can happen". Such a model is always trivially correct (and thus is typically chosen as a starting point). Thus we also need an additional goal of preferring the more specific (or refutable) model. I will call this the model's specificity. A modeller with infinite resources and time need only use these two measures as guides in its choice of model. In some cases, of course these dual aims might be in conflict. In a given modelling language one might be forced to choose between a vague but accurate model and a specific but more erroneous one. In some accounts the specificity of models are sometimes left out of analyses of modelling because the types of modelling languages considered are inherently precise.

For us more limited beings, with very distinct practical considerations, the complexity of our models become important. As we shall see below we have to balance the complexity, the error and the specificity of out models. Note that here, complexity is a property of the form of the models while the error and specificity are properties of its corresponding model semantics.

4 Other Formulations of Complexity

There is not room to do anything but mention but a few of the other formulations of complexity here (see [4] for a reasonably complete listing). I will only look at three here (I am distinguishing here between complexity and simplicity as it is traditionally used in philosophy, for this see section 8, below).

The Algorithmic Information of a pattern can be considered as the difficulty of storage of a program to generate the pattern (or alternatively the difficulty in finding such a program when working though possible programs in order of length).

Grassberger [6] defines the Effective Measure Complexity (EMC) of a pattern as the asymptotic behaviour of the amount of information required to predict the next symbol to the level of granularity. This captures an aspect of the scaling behaviour of the information required for successful prediction by a markov process model. This thus captures the asymptotic behaviour of a special case of my definition. A similar approach is taken by Badii and Politti [1].

The topological complexity described by Crutchfield [3], is a measure of the size of the minimal computational model (typically a finite automaton of some variety) in the minimal formal language in which it has a finite model. Thus the complexity of the model is both 'objectivised' by considering only minimal models but also related to the fixed hierarchy of formal languages. This has a number of disadvantages. *Firstly* this does not give a unique complexity for any pattern, as there is not necessarily such a “minimal” formal language, *secondly* in some formal languages the minimal model is uncomputable and *thirdly* in stochastic languages the minimal model will frequently be a completely random one, so one is forced to trade specificity with complexity to get a reasonable result. He also defines a measure of specificity similar to EMC above, as complementary to the topological complexity.

In each case the desire to attribute complexity purely objectively to a physical process seems to force a relativisation to either some framework for which privilege is claimed (e.g. a Turing Machine), to some aspect of the problem (e.g. granularity of representation) or by considering only the minimal size. This, of course, does not completely eliminate the inherent subjective effects in the process of modelling (principally the language of modelling), and obscures the interplay of complexity, specificity and the error involved.

5 Order and Disorder

It has been frequently asserted that complexity lies somewhere between order and disorder [6] (what is sometimes called “the edge of chaos” [7]). Thus in figure 1 below many people judge that the middle pattern is more complex than the other two.

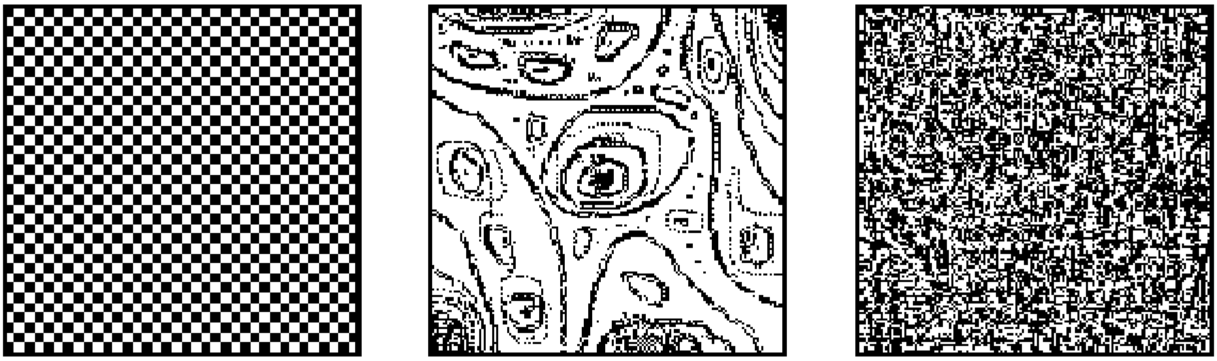

Figure 1: Three patterns (after Grassberger [6])

As a result of this idea graphs like figure 2 have dominated the literature about complexity in physics. However, as I shall argue, for such a situation to be true you at least need some other assumptions.

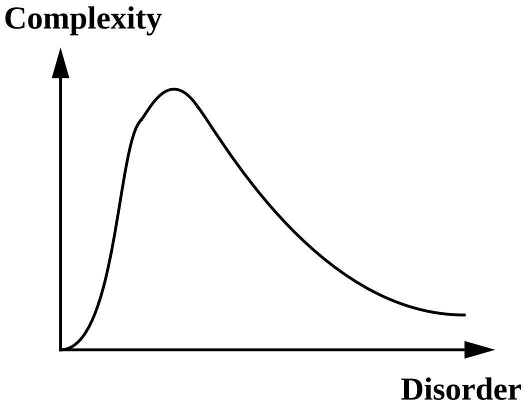

Figure 2: The presumed relationship between complexity and order

As can be easily seen, in the modelling framework above there is absolutely no need for this to be true. Although the highly ordered data might well correspond to the simplest models, it will also often be the case that the most disordered data corresponds to the most complex model forms.

To see this possibility consider the following situation. A modeller has an infinite and precise symbolic language with a limited number of symbols and some fixed grammar such that it includes some small expressions, but expressions of increasing size can be constructed. Suppose this language describes members of a class of data strings of any length of any sequence of symbols taken from a fixed alphabet.

A simple counting argument shows that most such patterns are disordered (as defined either by something like Shannon information or algorithmic information measures), but a similar counting argument shows that only a few of these patterns can correspond to models with relatively small minimal representations. That is, most of the disordered patterns will correspond with the models with the relatively large minimal representations. Whatever the ordering in terms of ease of search, in general the bigger forms will be more difficult to find, i.e. more complex.

Thus, in this case, far from complexity and disorder being antithetical, one would be hard pushed to arrange things so that any of the most complex models would correspond to even slightly ordered patterns.

So if complexity does not necessarily lie between order and disorder, where has our intuition gone wrong? *Without any prior knowledge about the process that produces the data we have no reliable way of distinguishing what is merely very complex behaviour and what is irrelevant noise*. The diagrams above mislead us because our experience about the patterns we typically encounter, has led us to recognise the noise, and separate it out from the relevant pattern. That this is not necessarily so, see figure 3, where we show each pattern as a magnification of a section of the one to its right.

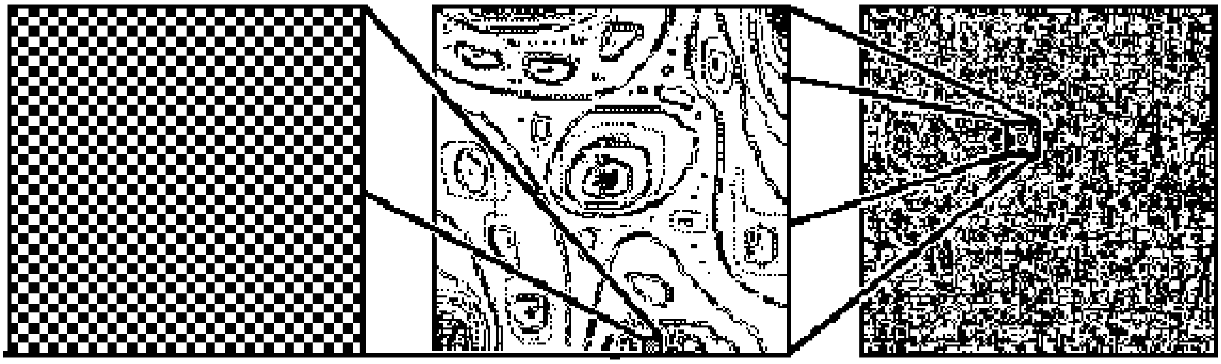

Figure 3: Three patterns with the inclusions shown

Faced with this new information one might change one's mind and say the rightmost pattern is the most complex. The initial judgement of the middle pattern comes not because such disordered patterns do not correspond to complex models in some precise languages, but because we are beings with limited resources used to receiving noisy data. We know it is not usually sensible to try to describe such patterns in such precise languages. This is for two reasons: *firstly* we do not have the time and *secondly* broader experience has taught us such models do not predict well on new data. In other words, we know that such “overfitting” of data is not likely to be a profitable activity. The association of what is apparently disordered with simplicity is thus the result of applying a natural heuristic and thus does not represent a necessary relationship.

Introducing specificity into the account makes sense of this. We are naturally good at distinguishing noise from relevant information (we have been practising since birth), so that we do not realise when we are doing it. The moral is that it is extremely useful to use less precise languages to describe such patterns, even at the cost of some accuracy with respect to the data. If the language allows expressions of varying degree of precision, then an overfitting model may well be more complex than a more suitable lower accuracy one. The most appropriate model for a pattern with high entropy might be a very simple and very vague model. If the complexity of the patterns is judged by complexity of their most suitable model one comes to the intuitive judgement that complexity is “between” order and disorder. It is only when one takes the unnatural step of leaving the specificity out of the account (e.g. by restricting the models to a uniformly precise language) that one is faced with the unnaturalness of the complexity of the best model.

6 Noise

The above account illustrates the importance of making a judgement as to what in the data may be considered as noise. If we take noise as what is effectively unpredictable in a pattern, then we will see that different approaches and conceptions of noise naturally emerge given different responses to excess error.

Let us imagine we are faced with an unacceptable level of error in the predictions of our best current model. What could be done?

Firstly, we could search for more accurate models by widening the search to look at other equally precise models. It is sensible to try the easier models first, but if we exhaust all the models at our current level of complexity we will be forced to try more complex ones. In this case we are effectively discounting the case that the unexplained elements of the data are unpredictable and treating noise as what is merely *currently* unexplained due to its complexity. This is a view taken in the light of many chaotic processes which can produce data indistinguishable from purely random data.

Secondly, we could decide to look for models that were less specific. This might allow us to find a model that was not significantly more complex but had a lower level of predictive error. Here we are essentially filtering out some of the data, attributing it to some irrelevant source. This might correspond to a situation where you know that there is an essentially random source of noise that has been imposed upon the data. This is the traditional approach, used in a wide range of fields from electronics to economics.

Thirdly, and most radically, we could seek to change our language of modelling to one that we felt was more appropriate to the data. Here we have noise as the literally indescribable. for example, sometimes a neural network (NN) is set up so that extreme fluctuations in the data are not exactly capturable by the range of functions the NN can output. In this way the NN is forced to approximate the training data and overfitting is avoided.

Thus randomness may be a sufficient characterisation of noise but it is not a necessary one.

7 Complexity vs. Information

The above framework distinguishes between the complexity of the model form and its specificity. The specificity of a model has been characterised in many ways, including: the information a model provides, the system’s entropy, and the model’s refutability.

Such measures of specificity have often been linked to a model’s *simplicity*, where by *simplicity* we mean *that property of a model which makes it more likely to be true than another, given that they have equal evidential support*. This property is called “simplicity”, because it is traced back to the principle of parsimony attributed to William of Occam. Thus Popper characterises simplicity as a model’s refutability [11], while Sober has associated it with the minimum extra information to answer a given question [14]. This tradition has been continued by several authors who have used various measures of information to capture it including Shannon information and algorithmic information[3]. It is clear that such *simplicity* is not necessarily the opposite of complexity, as described above (see section 8).

That complexity is not rigidly linked to the specificity of a model can be shown by considering any modelling language which has terms explicitly denoting nonspecificity (frequently called “error terms”). Clearly, the introduction of such terms can make an expression simultaneously more complex and less specific.

This is not to say that there might not be good reasons to prefer a model which is more specific, just that it is neither directly linked to either a model’s complexity or its error rate. Rissanen makes a case for a particular trade-off between the specificity of a model and its complexity - namely that one should seek the size of the minimal description which includes the model and the deviations from the model.

8 Complexity and Induction

From this framework it is clear that a lack of complexity is, in general, not a reliable guide to a model's error rate[4], unless the problem and modelling language happen to be constructed that way[5]. On the other hand experience has frequently shown us that the less complex theory often turns out to be more useful. The answer to this riddle becomes clear when one takes into account the process whereby models are typically developed.

An ideal modeller without significant resource restrictions might well be able to attempt a fairly global search through possible model forms. Some automatic machine-based systems approximate this situation and there it does indeed seem to be the case then that a lack of complexity is no guide to truth (e.g. [13]) - in fact, if anything the reverse seems to be true since there are typically many more complex expressions than simpler ones.

Usually, however, and certainly in the case of human modellers they do not have this luxury. They can check only a very limited number of the possible model forms. Fortunately, it is frequently the case that the meaning of the models allows us to intelligently develop and combine the models we have, to produce new models that are much more likely to produce something useful than a typical automatic procedure. Thus it is frequently the case that it is sensible to try elaborations of known models first before launching off into unknown territory where success is, at best, extremely uncertain.

On its own elaboration is, of course, an inadequate strategy. One can get into a position of diminishing returns where each elaboration brings decreasing improvements in the error rate, but at increasing cost.

[3] Variously attributed to combinations of Solomonoff [15], Kolmogorov [8] and Chaitin [1].

[4] In this I agree with Quine [12] and Judea Pearl [10].

[5] For example, if we knew that nature had developed a certain class of systems starting simply and elaborating from this and our language of modelling reflected this structure we would expect there to be a threshold of complexity such that models below this were more likely. Such a principle would then be a contingently true.

At some stage preferring simpler and more radically different models will be more effective. Thus sometimes choosing the simpler model, even if less precise and accurate is a sensible heuristic, but this is only so given our knowledge of the process of theory elaboration that frequently occurs.

9 Conclusion

Complexity is usefully distinguished from both the probability of correctness (the error) and the specificity of the model. It is relative to both the type of difficulty one is concerned with and the language of modelling. Complexity does not necessarily correspond to a lack of *"simplicity"* or lie between order and disorder.

When modelling is done by agents with severe resource limitations, the acceptable trade-offs between complexity, error and specificity can determine the effective relations between these. The characterisation of noise will emerge from this. Simpler theories are not a priori more likely to be correct but sometimes if one knows that the theories are made by an agent, for whom it is easier to elaborate than engage in a wider search, preferring the simpler theory at the expense of accuracy can be a useful heuristic.

References

[1] Badii, R. and Politi, A. 1997. Complexity: hierarchical structures and scaling in physics. Cambridge University Press, Cambridge.

[2] Chaitin, G.J. 1966. On the Length of Programs for Computing Finite Binary Sequences, *Journal of the Association of Computing Machinery,* 13, 547-569.

[3] Crutchfield, J.P. 1994. The Calculi of Emergence: Computation, Dynamics and Induction. *Physica D,* 75, 11-54.

[4] Edmonds, B. 1995. A Hypertext Bibliography of Measures of Complexity. Accessible at URL: http://www.fmb.mmu.ac.uk/~bruce/combib

[5] Edmonds, B. (forthcoming). What is Complexity?: the philosophy of Complexity per se with application to some examples in evolution. In F. Heylighen & D. Aerts (eds.): The Evolution of Complexity, Kluwer, Dordrecht. Also available at URL: http://www.fmb.mmu.ac.uk/~bruce/evolcomp

[6] Grassberger, P. 1986. Towards a Quantitative Theory of Self-Generated Complexity. *International Journal of Theoretical Physics,* 25(9), 907-938.

[7] Kauffman, S.A. 1993. *The Origins of Order.* Oxford University Press, New York.

[8] Kolmogorov, A.N. 1965. Three Approaches to the Quantitative Definition of Information, *Problems of Information Transmission,* 1, 1-17.

[9] Murphy, P.M. and Pazzani, M.J. 1993. Exploring the Decision Forest. Computational Learning and Natural Language Workshop, Provincetown, MA, September 1993.

[10] Pearl, J.P. 1978. On the Connection Between the Complexity and Credibility of Inferred Models, *International Journal of General Systems*, 4, 255-264.

[11] Popper, K.R. 1968. *Logic of Scientific Discovery,* Hutchinson, London.

[12] Quine, W.V.O. 1960. Simple Theories of a Complex World, in *The Ways of Paradox*, Eds., Random House, New York, pages 242-246.

[13] Rissanen, J. 1990. Complexity of Models. In Zurek,WH; (ed.). *Complexity, Entropy and the Physics of Information*. Addison-Wesley, Redwood City, California, 117-125.

[14] Sober, E. 1975. *Simplicity.* Clarendon Press, Oxford.

[15] Solomonoff, R.J. 1964. A Formal theory of Inductive Inference. *Information and Control*, 7, 1-22, 224-254.