Model Selection and the Multiplicity of Patterns in Empirical Data

James W. McAllister†‡

Several quantitative techniques for choosing among data models are available. Among these are techniques based on algorithmic information theory, minimum description length theory, and the Akaike information criterion. All these techniques are designed to identify a single model of a data set as being the closest to the truth. I argue, using examples, that many data sets in science show multiple patterns, providing evidence for multiple phenomena. For any such data set, there is more than one data model that must be considered close to the truth. I conclude that, since the established techniques for choosing among data models are unequipped to handle these cases, they cannot be regarded as adequate.

1. Introduction: Techniques for Data Model Selection. A data model is a mathematical representation of an empirical data set (Suppes 1962). In most cases, a data model is deliberately inaccurate to some degree: the function of a data model is not to reproduce the data in all details, but to identify a significant component or pattern in the data set. An example of data model is a line drawn in a graphical plot of a data set. Data models are an important tool in scientific research. Many scientists regard constructing a data model as a means of uncovering the regularity in a data set by empirical analysis, prior to any explanatory hypothesis. A data model, on this view, may be the first step in formulating an empirical law of nature.

Any data set admits various different data models. A central problem in scientific research is therefore how to choose among data models. Re-

†To contact the author, please write to: Faculty of Philosophy, University of Leiden, P.O. Box 9515, 2300 RA Leiden, The Netherlands; e-mail: j.w.mcallister@let.leidenuniv.nl.

‡I presented a previous version of this paper at the 20th Biennial Meeting of the Philosophy of Science Association, Vancouver, November 2006. I am grateful to the audience for constructive discussion. I thank Leiden University students Marjolein Eysink Smeets and Lenneke Schrier for suggesting the cortisol example, and Remko van der Geest for comments on a draft.

Philosophy of Science, 74 (December 2007) pp. 884–894. 0031-8248/2007/7405-0028$10.00
Copyright 2007 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/525630 Published online by Cambridge University Press

searchers in various disciplines, including philosophy of science, have proposed quantitative techniques for determining which data model is the best, where "the best" is usually interpreted as "the closest to the truth," sometimes via the proxies "the most likely to be true" or "the best-supported by the data." All these techniques balance simplicity or parsimony against goodness of fit, and assign a numerical value to each proposed model of a data set: the model that heads the ranking is the best (Myung, Forster, and Browne 2000; Burnham and Anderson 2002; Weakliem 2004; Waldorp and Wagenmakers 2006).

These model selection techniques play an influential role not only in research practice, but also in philosophical thinking about science. They seem to promise a way of interpreting empirical data that does not rely on judgment or subjectivity. In consequence, they seem to offer a foundation for asserting the objectivity of scientific laws and theories. They are also often credited with clarifying the role of simplicity in theory choice, solving the problem of induction, paving the way for automated discovery, and various further achievements.

I will briefly outline three of the several quantitative techniques for choosing between data models on offer. The first technique derives from algorithmic information theory, and exploits the principle of algorithmic compression (Li and Vitányi 1997). According to this technique, the best model of a data set is the shortest algorithm able to reproduce it. Proponents of this technique suggest that the shortest algorithm that reproduces a data set represents the underlying message in the data, and corresponds to a fundamental regularity in the world. Some early contributors to algorithmic information theory, such as Solomonoff (1964), seem to have held that the compression of the data set must be exact, that is, that the algorithm must precisely reproduce every data point. Other proponents of this technique, reasoning that any empirical data set contains a random element, have suggested that the algorithm must compress only some significant component of the data set, discarding the residue. An example is Gell-Mann (1994, 51–61), who advocates choosing the model of a data set that provides the shortest compression of what he calls the "useful information" contained in the data set (McAllister 2003).

A second technique is suggested by minimum description length theory (Li and Vitányi 1997, 351–372). According to this technique, the best model of a data set is the one that minimizes the length of a two-part description of it: the first part states the model, and the second part enumerates the discrepancies of the data from the model. The best model is thus one that strikes the best balance between the simplicity of the model itself (which shortens the first part of the description) and the

https://doi.org/10.1086/525630 Published online by Cambridge University Press

accuracy with which the model generates the data (which shortens the second part).

A third technique comes from Malcolm Forster and Elliott Sober's appeal to the Akaike information criterion in solving the curve-fitting problem. This technique makes use of the concept of a family of curves. Curves belong to the same family if their equations contain the same freely adjustable parameters: examples of families of curves are straight lines and parabolic curves. We first estimate the "expected predictive accuracy" of each family of curves. This is equal to the degree of fit of the curve in each family that best fits the data set, minus the number of adjustable parameters of curves in that family. The best model of the data set is the model corresponding to the best-fitting curve that is a member of the family with the highest expected predictive accuracy (Forster and Sober 1994). For example, the Akaike technique suggests that the Copernican model is a superior model of data on planetary motions than the Ptolemaic model: although the two models fit the data approximately equally well, the Copernican model has a smaller number of adjustable parameters, and thus scores more highly on the Akaike criterion.

All these techniques have some features in common. First, each technique regards all possible models of a data set as being in competition with one another for the status of "model closest to the truth." In other words, none of these techniques has the conceptual resources to identify models of a data set that are not in mutual competition, but that coexist alongside one another. Second, save in the rare case of a numerical tie in the appraisal of two or more models, these techniques are designed to identify a single model of a data set as being the closest to the truth. All other possible models of the same data set are regarded as having less claim to be close to the truth. These features have been noted by Burnham and Anderson: "Very little of the extensive model selection literature goes beyond the concept of a single best model, often because it is assumed that the model set contains the true model. This is true even for major or recent publications" (2004, 262).

How can we account for these features? The established techniques are based on the assumption that any empirical data set provides evidence for only one phenomenon, or objective feature of the world, and that the task of model selection is to identify the model of the data that corresponds best to that phenomenon. (I use the term "phenomenon" in the sense of Bogen and Woodward 1988). Only on this assumption does it make sense to search for the single model that is "closest to the truth."

In this paper, I argue that the assumption that an empirical data set provides evidence for just one phenomenon is mistaken. It frequently occurs that data sets provide evidence for multiple phenomena, in the form of multiple patterns that are exhibited in the data with differing noise levels.

https://doi.org/10.1086/525630 Published online by Cambridge University Press

This means that, in these cases, several different models of a data set must be regarded as equally close to the truth. In the light of this fact, none of the standard techniques for selecting among models of data sets can be considered adequate, since none allows for the possibility that a data set may admit multiple models. I will conclude that a minimal requirement for the adequacy of such techniques is that they should be supplemented by a specification of the noise level that the researcher wishes to tolerate.

2. Data Sets Exhibiting Multiple Patterns. In this section, I present three examples of empirical data sets that have been found to exhibit more than one pattern. Each of the patterns corresponds to a well-confirmed effect that is of substantial importance in the relevant field of science. This means that the models that pick out each of these patterns must be considered close to the truth.

The first example consists of a data set comprising all atmospheric temperature measurements made on earth. Climatologists, meteorologists, and researchers in other scientific disciplines have long analyzed this data set and portions of it. They have concluded that it exhibits many different patterns; furthermore, they have been able to identify with high confidence the causal processes that give rise to each of these patterns. The patterns can be divided roughly into three categories: highly regular cyclical, less regular cyclical, and noncyclical patterns. The highly regular cyclical patterns include a pattern with a period of a day, associated with the earth's rotation about its axis; a pattern with a period of a year, associated with the earth's orbit around the sun; patterns with periods of 11, 90, and 180 years, attributed to the sunspot cycle; a pattern with a period of approximately 21,000 years, attributed to the precession of the earth's orbit; and various patterns with periods of between 40,000 and 100,000 years, attributed to fluctuations in the inclination of the earth's axis of rotation and the eccentricity of the earth's orbit (Bryant 1997; Burroughs 2003). The less regular cyclical patterns include patterns associated with individual weather systems, which have a life span of a few days or weeks; patterns associated with El Niño events, occurring at intervals of between four and ten years; and patterns associated with explosive volcanic eruptions, such as those of Mt. Tambora in 1815, Krakatau in 1883, and Mt. Pinatubo in 1991. The noncyclical patterns include various patterns associated with changes in the major geography of the earth, the earth's rate of rotation, and the characteristics of the sun, which manifest themselves over timescales of between 10^7 and 10^9 years; and the "hockey stick" pattern associated with anthropogenic changes in the composition of the atmosphere in the twentieth century (Mann, Bradley, and Hughes 1998).

Let us now conceive of formulating a model of this data set. Clearly, multiple models may be formulated, each corresponding to one of the

https://doi.org/10.1086/525630 Published online by Cambridge University Press

patterns exhibited by the data and listed above. Each of these models must be regarded, in the light of our current knowledge, as very close to the truth: there are strong grounds for considering each pattern to be a genuine component of the data, and for regarding the hypothesized cause of the pattern to be a real physical phenomenon. It is indisputable, for example, that atmospheric temperature data show a pattern with a period of one year, and the earth's orbit around the sun is well attested. The "hockey stick" pattern has attracted more controversy, but there seem good grounds for believing that it is real and that it is caused by anthropogenic changes in atmospheric composition. The same holds for the other patterns listed above.

From this it follows that, depending on the focus of his or her research, a scientist is fully justified in choosing any of these models as the one closest to the truth. A meteorologist studying Atlantic weather systems will select a model picking out a cycle with a period of a few days, whereas a geologist studying northern hemisphere glaciations will choose a model picking out a cycle with a much longer period.

For a second example, consider a data set consisting of microwave radiation intensity readings of the sky. Radio astronomers have identified several different patterns in this data set, and have attributed these patterns to specific mechanisms. One pattern consists of a number of discrete peaks with an intensity of a few kelvin, spread over the sky, caused by astronomical radio sources. A second pattern, named "isotropic background," consists of a constant intensity distribution, independent of direction, with the spectrum of a black body at approximately 2.7K. Cosmologists regard this pattern as evidence of the decoupling of matter and radiation in the early universe. A third pattern, named "dipole anisotropy," consists of one warm and one cool pole in opposite directions in the sky, and is modeled by a first-order spherical harmonic function with an amplitude of approximately 3.4×10^{-3}K. Cosmologists regard this pattern as evidence of the motion of the earth relative to the matter that last scattered the microwave radiation. A fourth pattern, dubbed "quadrupole," consists of two warm and two cool poles, and is modeled by a second-order spherical harmonic function with an amplitude of around 1.6×10^{-5}K. Cosmologists attribute this pattern to the radio emissions of the Milky Way. A fifth pattern, called "ripples," consists of irregular fluctuations of radiation temperature with an amplitude of approximately 3.1×10^{-5}K on angular scales of 10° and above. Cosmologists regard this pattern as evidence of inhomogeneities in the matter distribution in the early universe, from which galaxies and clusters of galaxies condensed. The sixth pattern identified up to now, dubbed "acoustic oscillations," consists of irregular fluctuations of radiation temperature with an amplitude of approximately 6.9×10^{-5}K on an angular scale of around 1°.

https://doi.org/10.1086/525630 Published online by Cambridge University Press

Cosmologists regard this pattern as evidence that the universe is flat, that is, that the total mass and energy density of the universe is equal to the so-called critical density (Partridge 1995).

As in the case of the atmospheric temperatures data set, it is possible to formulate a model of the microwave radiation data set corresponding to each of these patterns. Again, each of these models must be regarded, in the light of our background knowledge, as very close to the truth: the effect predicted by each model is well confirmed and—although our grasp of causal processes in this area may be somewhat less secure than in climate history—the cause attributed to the effect is probably real.

Thus, a scientist is warranted in choosing any of these models as the one closest to the truth in the light of his or her research interests. For example, it is justified for one cosmologist to describe the cosmic microwave background as isotropic, and to cite this isotropy as evidence for the big bang, and it is equally justified for another cosmologist to describe the background as anisotropic, and to cite this anisotropy as evidence for the uneven matter distribution in the early universe.

For our third and final example, we turn to a set of measurements of a person's blood cortisol level over time. Cortisol is a steroid hormone that, in humans, is primarily responsible for regulating metabolism and the body's response to inflammation and stress. Endocrinologists identify multiple patterns in such a data set. One pattern consists of a 20–50 percent increase of mean 24-hour cortisol levels through the human lifespan. A second pattern is a cycle with a period of 24 hours: cortisol levels are lowest during the night, climb rapidly to reach a maximum between 8 a.m. and 10 a.m., and decline through the rest of the day. A third pattern consists of 15–20 "spikes" of varying magnitudes in each 24-hour period, corresponding to discrete cortisol secretion events, believed to be controlled by the hypothalamus (Weitzman et al. 1971; Van Cauter, Leproult, and Kupfer 1996).

As in the previous cases, it is possible to formulate a model of the plasma cortisol data set corresponding to each of these patterns. Again, each of these models must be regarded, in the light of our medical knowledge, as very close to the truth: both the effect predicted by each model and the causal processes attributed to them by endocrinologists are probably real.

Thus, physicians are warranted in choosing any of these models as the one closest to the truth in the light of their research or clinical interests. A geriatrician who wishes to interpret the high cortisol level in an elderly patient is justified in selecting the model that picks out the lifetime increase; a physiologist studying the human circadian system is justified in selecting the model that picks out the 24-hour cycle; and a physician who wishes to study the body's response to psychosocial stress, or who wishes to

https://doi.org/10.1086/525630 Published online by Cambridge University Press

determine whether a patient has suffered hypothalamic damage after brain surgery, is justified in selecting the model that picks out the spikes.

3. Inadequacy of Current Data Model Selection Techniques. The cases described in Section 2, in which scientists have interpreted data sets as exhibiting several different patterns, are clearly paradigms of good science. The patterns are robust and stable, and scientists have been able to identify causal factors for them all. Some patterns, moreover, have provided the basis for practical applications.

To what extent are these cases typical of empirical data sets gathered in the sciences? I believe that we should regard data sets that exhibit multiple different patterns as the norm, rather than an exception. Most data sets are the causal outcome of a multiplicity of phenomena, acting in combination. Each phenomenon leaves a trace in the data set in the form of a pattern. Thus, most data sets should be expected to exhibit more than one pattern. Each of these patterns is a legitimate object of investigation by scientists.

Awareness that many empirical data sets show multiple patterns seems to be especially strong among social scientists. Research in the social sciences often pertains to structures that form nested hierarchies. Studies of educational attainment, for example, must disentangle the contributions of individual pupils, teachers, classes, schools, and education authorities. Structures on each of these levels are sources of variation in the empirical data. Social scientists often use multilevel statistical tools to deal with data sets showing patterns at different scales (Goldstein 2003).

In the light of these considerations, we should place a new demand on quantitative techniques for choosing among data models. Any such technique should have as a possible outcome the verdict that multiple different models of a given data set are very close to the truth.

Unfortunately, standard quantitative techniques for choosing among data models, such as the three reviewed in Section 1, do not pass this test. They lack the conceptual resources to allow for the possibility that a data set can be correctly analyzed in several different ways.

For example, imagine applying the technique based on minimum description length theory to the microwave radiation intensity data set. This technique would yield a unique model of the data set—the one for which the sum of the length of a specification of the model and the length of a description of the discrepancies of the data from the model is lowest. If we are lucky, this model will correspond to one of the patterns that cosmologists identify in the microwave data set. In yielding this pattern, however, the minimum description length technique disregards all the other patterns that cosmologists identify in the data set, which have equal claim to correspond to structures in the world.

https://doi.org/10.1086/525630 Published online by Cambridge University Press

Similarly, imagine applying the Akaike technique to the atmospheric temperature data set. This technique will also yield a unique model of the data set—the one for which the sum of the degree of fit between the model and the data and a negative quantity proportional to the number of adjustable parameters of the model is highest. Again, this model may correspond to one of the patterns identified by climatologists in the temperature data; but it will not correspond to the other patterns, which are equally real.

In passing, we note that this example shows how a possible counterargument on behalf of the Akaike technique is blocked. The first step of the Akaike technique involves selecting the family of curves from which the best model is to be found. It might be argued that this step gives the Akaike technique the necessary flexibility to recognize multiple patterns in a data set as equally real, if each pattern belongs to a different family of curves. However, the atmospheric temperature case shows that this argument does not run through in general. All the regular cyclical patterns identified by climatologists in the atmospheric temperature data set correspond to harmonic curves with different periods and amplitudes, so they would all compete with one another in the Akaike technique for the status of the model closest to the truth.

A second possible response to my criticism of the available data model selection techniques is to claim that there is indeed a unique best model of any such data set—the one corresponding to the sum of several or all the patterns that can be identified in the data—and that the quantitative techniques can be expected to pick out this pattern as the closest to the truth. This response does not succeed either, however, for two reasons. First, the sum of all patterns that can be identified in the data would probably coincide with the complete data set itself, since any discrepancy between a data set and a pattern identified in it can be endlessly analyzed as a sum of further patterns. Second, scientists adduce individual patterns in data as evidence for claims about the contributions of individual causal factors. The evidence for a claim about the existence and effect of a causal factor consists of the component pattern that is determined by that causal factor alone: it does not consist of the resultant pattern determined by the combination of several or all causal factors operating in a physical system. For example, the evidence that the universe is flat consists of the component pattern labeled "acoustic oscillations" in the microwave radiation data set, and not of the resultant pattern determined by the combination of several or all astronomical and cosmological processes. The same becomes evident in cases where we wish to manipulate causal factors to alter a situation. The geriatrician concerned at the cortisol level in a patient rightly focuses on the lifetime increase pattern, and ignores the 24-hour cycle. For these reasons, the notion of a sum of several or all

https://doi.org/10.1086/525630 Published online by Cambridge University Press

patterns does not nullify the reality or the significance of each component pattern.

4. Coexistence of Patterns at Different Noise Levels. In Section 3, I argued that three quantitative techniques for choosing among data models—based on algorithmic information theory, minimum description length theory, and the Akaike information criterion—are inadequate for their purpose, because they are incapable of doing justice to data sets exhibiting multiple patterns. In conclusion, I wish to sketch what enhancement would be minimally required for similar techniques to pass this test. The enhancement required is a facility to take "noise level" into account, in a sense that I will specify.

The multiple patterns exhibited in the data sets discussed in Section 2 differ, and can therefore be distinguished, in one simple respect: each is exhibited in its data set with a different noise level. This noise level is the mathematical difference between the pattern and the data set. In the following illustration, $F(x)$ is an empirical data set, the harmonic function $a \sin \omega x + b \cos \omega x$ is a pattern that a scientist might discern in the data set, and $R(x)$ is the noise term:

$$F(x) = a \sin \omega x + b \cos \omega x + R(x).$$

Note that the term "noise" here denotes not a margin of error or factual inaccuracy in the data, but purely a mathematical difference between two terms. The statement that an empirical data set can be described as the sum of a pattern and a corresponding noise term thus holds true regardless of the incidence of observational, experimental, and other errors in the data set.

Different patterns are exhibited in a given data set with different noise levels. For example, the patterns associated with individual weather systems are exhibited in the atmospheric temperature data set with a low noise level: each pattern tracks the measured temperature variations closely. Any individual such pattern has admittedly a short temporal extension, but, while it holds, it has high accuracy. By contrast, the pattern with a period of 21,000 years, related to the precession of the earth's orbit, is exhibited in the data with a higher noise level: that is, this pattern accounts less precisely for the atmospheric temperature at any one time. Similarly, the pattern called "ripples" is exhibited in the microwave radiation data set with a lower noise level than the pattern corresponding to the 2.7K isotropic background. Of course, patterns exhibited with a higher noise level are no less real for this reason.

As these examples show, patterns that are exhibited by data sets with different noise levels are not in competition with one another for the status of "best model of the data": they coexist with one another. It is the case

https://doi.org/10.1086/525630 Published online by Cambridge University Press

both that plasma cortisol levels in humans display a 24-hour cycle, and that they rise with age. There is no contradiction between these statements, as there is also no incompatibility between the models of the empirical data of which these statements are brief descriptions. Models of data that pick out patterns exhibited with different noise levels should thus not be placed in a single ranking with the aim of determining which one of them is the best: instead, each should be evaluated individually for robustness and stability. If each of several models passes this test, then they should all be accepted.

For any quantitative technique for choosing among models of a data set to be adequate to this scenario, it must incorporate conceptual resources to take account of the noise levels with which different patterns are exhibited in a data set. It is clear that a unique model can be identified as "the best" only if a certain desired noise level is specified—the noise level with which the pattern picked out by the model is exhibited by the data set. In the absence of this specification, talk of "the best" model of a data set is ambiguous. It follows that quantitative techniques for choosing among models of a data set must contain provision for the specification of the desired noise level at the outset of the process. Since neither the algorithmic compression technique, nor the minimum description length technique, nor the Akaike technique have this provision, we must rule them all out as inadequate.

As usual, of course, actual modes of reasoning and judgment used by scientists are more sophisticated than their attempted formalizations by philosophers. I suspect that most experienced scientists would immediately recognize that most data sets exhibit multiple patterns, corresponding with and constituting evidence for multiple phenomena. Scientists deal with this complexity, in most cases, by explicitly or implicitly stipulating what noise level they are willing to tolerate in models of the data. This stipulation enables them to evaluate different models against one another at that noise level, while not ruling out that quite different models may hold at other noise levels.

REFERENCES

Bogen, James, and James Woodward (1988), "Saving the Phenomena", *Philosophical Review* 97: 303–352.

Bryant, Edward (1997), *Climate Process and Change.* Cambridge: Cambridge University Press.

Burnham, Kenneth P., and David R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

——— (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection", *Sociological Methods and Research* 33: 261–304.

Burroughs, William J. (2003), *Weather Cycles: Real or Imaginary?* 2nd ed. Cambridge: Cambridge University Press.

Forster, Malcolm R., and Elliott Sober (1994), "How to Tell When Simpler, More Unified,

https://doi.org/10.1086/525630 Published online by Cambridge University Press

or Less *Ad Hoc* Theories Will Provide More Accurate Predictions", *British Journal for the Philosophy of Science* 45: 1–35.
Gell-Mann, Murray (1994), *The Quark and the Jaguar: Adventures in the Simple and the Complex.* New York: Freeman.
Goldstein, Harvey (2003), *Multilevel Statistical Models*. 3rd ed. London: Arnold.
Li, Ming, and Paul M. B. Vitányi (1997), *An Introduction to Kolmogorov Complexity and Its Applications*. 2nd ed. Berlin: Springer.
Mann, Michael E., Ray S. Bradley, and Malcolm K. Hughes (1998), "Global-Scale Temperature Patterns and Climate Forcing over the Past Six Centuries", *Nature* 392: 779–787.
McAllister, James W. (2003), "Effective Complexity as a Measure of Information Content", *Philosophy of Science* 70: 302–307.
Myung, In Jae, Malcolm R. Forster, and Michael W. Browne, eds. (2000), "Special Issue on Model Selection", *Journal of Mathematical Psychology* 44, no. 1.
Partridge, R. Bruce (1995), *3 K: The Cosmic Microwave Background Radiation.* Cambridge: Cambridge University Press.
Solomonoff, Ray J. (1964), "A Formal Theory of Inductive Inference", *Information and Control* 7: 1–22, 224–254.
Suppes, Patrick (1962), "Models of Data", in Ernest Nagel, Patrick Suppes, and Alfred Tarski (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress.* Stanford, CA: Stanford University Press, 252–261.
Van Cauter, Eve, Rachel Leproult, and David J. Kupfer (1996), "Effects of Gender and Age on the Levels and Circadian Rhythmicity of Plasma Cortisol", *Journal of Clinical Endocrinology and Metabolism* 81: 2468–2473.
Waldorp, Lourens, and Eric-Jan Wagenmakers, eds. (2006), "Special Issue on Model Selection: Theoretical Developments and Applications", *Journal of Mathematical Psychology* 50, no. 2.
Weakliem, David (ed.) (2004), "Special Issue on Model Selection", *Sociological Methods and Research* 33, no. 2.
Weitzman, Elliot D., David Fukushima, Christopher Nogeire, Howard Roffwarg, T. F. Gallagher, and Leon Hellman (1971), "Twenty-Four Hour Pattern of the Episodic Secretion of Cortisol in Normal Subjects", *Journal of Clinical Endocrinology and Metabolism* 33: 14–22.

https://doi.org/10.1086/525630 Published online by Cambridge University Press