Winning the Accuracy Game

Three statistical strategies—replicating, blocking and modeling—can help scientists improve accuracy and accelerate progress

Hugh G. Gauch, Jr.

Thomas Edison famously stated that genius is "one percent inspiration and ninety-nine percent perspiration." In crop science, as in many other fields of research, investigators find much truth in this proverb. The discovery of an improved variety of corn, wheat or soybeans is very much a numbers game. The standard varieties already incorporate many genetic refinements from monumental breeding efforts in the past. Improvements come in small increments from testing large numbers of experimental genotypes. Other sciences have analogous challenges: In pharmaceutical research, for example, many compounds must be screened to find one that might make a successful medicine.

However, measurement errors or chance variations can cause an inferior plant to look better than a superior one. In a large field of contenders, the superior breed can get lost in the crowd. Scientists are aware of this problem, of course, but they routinely underestimate its severity.

Several years ago, I analyzed a trial of seven varieties of soybeans. The variety that appeared to be best was 14 percent better than the average of the other six, and 3 percent better than its

AP Photo/Alan Marler

Figure 1. How can one identify the best cancer drug, safest automotive design or highest-yielding crop variety? Three statistical strategies can increase success in selecting the best treatment or entry: replicating, blocking and the oft-neglected strategy of modeling. The high-quality turfgrass needed for home lawns and elite golf courses such as this Donald Ross-designed course at the Grove Park Inn in Asheville, North Carolina, has emerged from scientific screening of hundreds of experimental and commercial varieties.

replications	confidence of success (percent)
2	60.8
3	66.7
4	70.5
5	73.2
10	80.5
20	86.0
40	90.0
100	93.7
162	95.0
500	97.2

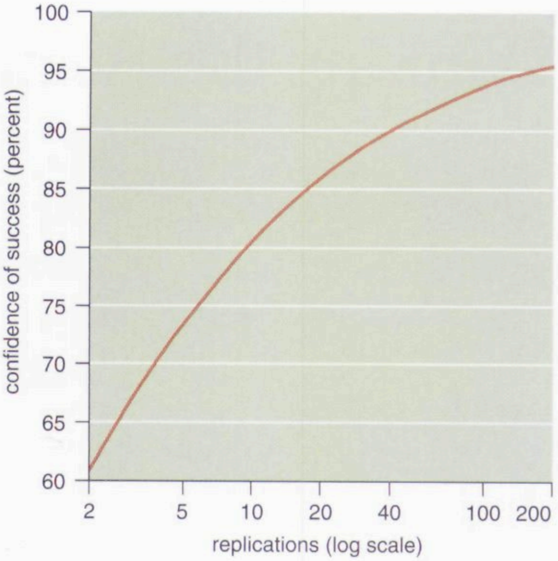

Figure 2. Sir Ronald A. Fisher, shown calculating at left, made Rothamsted Experimental Station in England the proving ground for major statistical techniques used in science today. He established the importance of randomization and of replication—doing an experiment over and over again—to increase accuracy. The table and graph show how the probability of replication successfully helping accuracy, despite noisy data, grows with the number of replications. Although the first few replications pay off handsomely, the rewards fall off rapidly. (Photograph by A. Barrington-Brown, republished by permission of the R. A. Fisher Memorial Trust.)

closest rival. Suppose these numbers accurately reflected its superiority, and the experiment were replicated. Would the same soybean necessarily come out on top? Surprisingly, simulations showed only a 49 percent probability that it would; the fourth-best or worse entry would win 10 percent of the time.

The odds in favor of selecting the best breed (or pharmaceutical compound, or product-safety modification) improve, of course, if the experiment is replicated more times. Replication decreases the effect of chance variation, thereby improving accuracy, efficiency and repeatability. But in the numbers game, this way of achieving accuracy comes at a severe cost. The more times breeders have to run the same experiment, the fewer alternative breeds they can test.

Fortunately, more replication—or more perspiration, to recall Edison's dictum—is not the only way to improve accuracy. A small investment on the "inspiration" side can make a very large difference. Two other strategies, called *blocking* and *modeling*, can provide at least one replication's worth of accuracy (and often more) at almost no cost. Blocking is a method of experimental design that reduces the effects of chance errors. It has become routine in the development of better medicines, safer cars, stronger steels and a host of other applications.

Modeling to gain accuracy is much less familiar to practicing scientists, even though it is frequently applicable and usually improves experimental accuracy more than blocking does. The idea behind it dates back, in some sense, to the medieval master of parsimony, William of Ockham. Scientific data always contain a mixture of *signal* and *noise*; the scientist's job is to discern the signal. It almost always shows up as patterns that are inherently simpler than the noise. Noise is idiosyncratic and complex; the reasons why a particular corn plant produced more grain than the one next to it are often unknowable. But signal is simpler; a single environmental difference may cause dozens of breeds to respond similarly. Modeling is a way of amplifying the signal by placing greater weight on simple patterns in the data.

Different sciences and different experiments vary widely in the amount of accuracy they can attain. In physics, the gyromagnetic ratio of the electron has been measured to 11 significant digits. Many quantities in science and industry are readily measured to three to six significant digits. The data I work with in breeding trials carry only one significant digit. Nevertheless, every science shares the need for greater accuracy. Gains in accuracy translate to safer products, more effective medicines and more food on the world's tables. Accuracy matters.

The Limitations of Replication

In 1919, a 29-year-old statistician named Ronald A. Fisher started a new job at Rothamsted Experimental Station in Harpenden, England. It was the premier agricultural research site in the country and would become, thanks to Fisher's efforts, the proving ground for many of the statistical techniques that scientists take for granted today. Fisher was hired to make some sense out of 76 years of experimental records, which he later called a "muck heap."

Why were the data at Rothamsted such a mess? In the 19th century, scientists had little conception of the importance of replication. Frequently the productivity of a given breed in a given year would be represented by a single measurement. With only one

Hugh G. Gauch, Jr., is a senior research specialist in crop and soil sciences at Cornell University. He received a B.S. in botany from the University of Maryland in 1964 and an M.S. in plant genetics from Cornell in 1966. His research specialty has been statistical analysis of ecological and agricultural data. His most recent book is Scientific Method in Practice, *published by Cambridge University Press in 2002. He is a fourth-generation member of Sigma Xi (Cornell University, alpha chapter, 1980), having been preceded by his great-grandfather, Charles Wesley Rolfe (a founding member at the University of Illinois, 1903), maternal grandmother, Susan Farley Rolfe (later Mrs. Horace Graham Butler; University of Illinois, 1909), and father, Hugh Gilbert Gauch (Kansas State University, 1937). Address: Crop and Soil Sciences, Cornell University, 519 Bradfield Hall, Ithaca, NY 14853. Internet: hgg1@cornell.edu*

observation, as scientists now know, it would have been impossible to estimate the amount of error, and therefore impossible to make any meaningful comparisons between measurements. It was hard to know which results to take seriously.

Inaccuracy or error is quantified by two familiar descriptive statistics. The *standard deviation* is the square root of the mean square error of all the individual observations. (*Error* is the observed value minus the true value. Since true values are not known, averages over replications are used in error calculations.) The *standard error* is the root mean square error for an average over *N* replicates, which equals the standard deviation divided by the square root of *N*.

Replication is one of the finest ideas in the history of science, but it faces a severe law of diminishing returns. Halving the standard error requires a fourfold increase from one to four replications. The next several halvings require 16, 64, 256 and 1,024 replications, which rapidly become prohibitively expensive. Scientists are familiar with this square-root dependency and the diminishing returns that follow as a consequence.

A second shortcoming of replication is far less well known. It becomes apparent when one compares the success rate of replicated and unreplicated measurements in estimating true values. Obviously, scientists prefer an average of two replicates to a single unreplicated observation because the former is likely to be more accurate. But that does not mean it will always be more accurate. Just by chance, the first replicate may be quite close to the true value, while the second replicate is far from it. In that case, the average of the two replicates is less accurate than the unreplicated result.

The reader might like to ponder the following three questions before going on. How often is the average of two replicates more accurate than a single measurement? How about the average of five replicates? And how many replicates would be needed to achieve 90 or 95 percent confidence that the average is more accurate than the unreplicated measurement? Curiously and regrettably, few scientists know the answers to these practical questions.

Here are the answers, based on the ordinary assumption that measurement errors are distributed according to a "bell-shaped curve." Two replicates are more accurate than one 60.8 percent of the time. For five replicates, the success rate climbs to 73.2 percent—but this means that a single observation is still more accurate 26.8 percent of the time. To increase replication's success rate to 90 percent, most scientists I have spoken with guess that three to eight replicates would be sufficient. In fact, the actual number is 40. To achieve 95 percent success, a daunting 162 replicates are required, again far beyond what most scientists would expect.

Of course, the lesson here is not that scientists should always replicate their experiments 40 or 162 times. Instead, they should develop realistic expectations about what replication can and cannot accomplish.

A related problem, which I alluded to in the introduction, emerges when scientists face a selection task. In medical research, it is common to perform an experiment in order to select the best treatment. Again, it is obvious that greater accuracy will improve the success rate of the best-treatment selection. But when this simple insight is informed by concrete numbers, most scientists are surprised by the difficulty of selection tasks, and hence the importance of using all the available strategies to gain accuracy.

Consider the simplest selection scenario. An experiment has *T* inferior treatments and one superior treatment. All of the inferior treatments have the same average effect. What is the probability that the experiment will correctly identify the superior treatment, despite noisy data? A field called order statistics answers such questions.

For instance, for T=10, the best of the observed values for the inferior treatments is likely to be 1.54 standard errors above the mean. If the superior treatment is only one standard error better than its rivals, it does not have a good chance of winning, merely 32.4

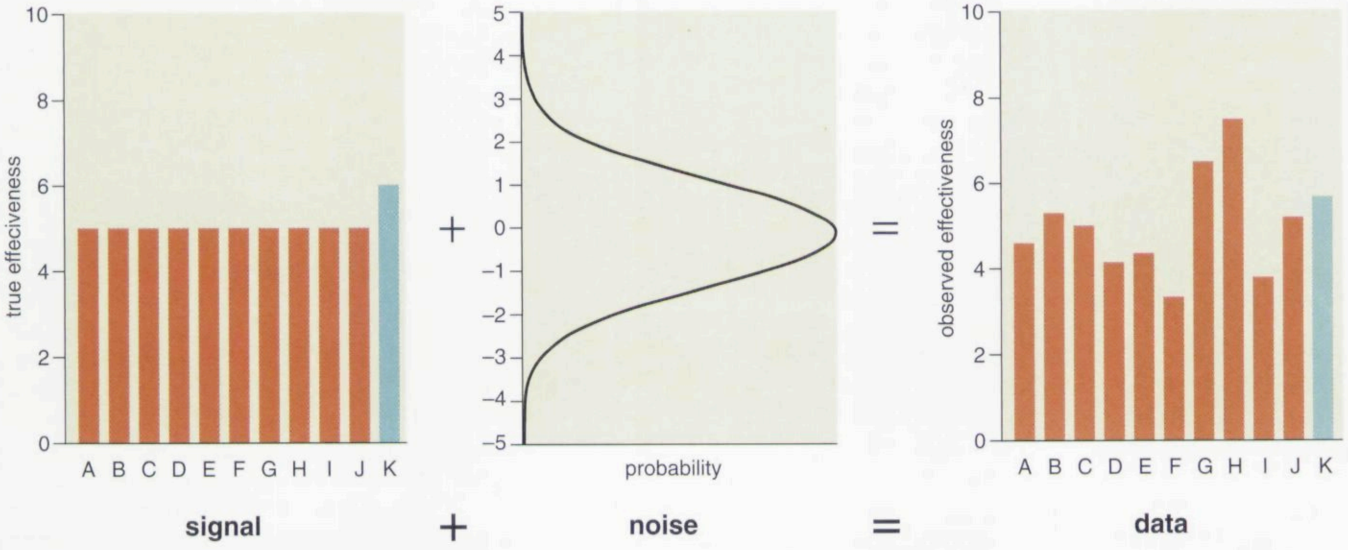

Figure 3. Experimental data have both signal and noise components, the former attributable to imposed treatments and the latter arising from uncontrolled factors. In a simulated experiment, ten treatments, A through J, have equal true effectiveness. Entry K meanwhile is superior by one unit (one standard error). Random noise is added to each treatment following a standard normal distribution with a standard deviation of one unit. Entry K appears to rank third, when it is actually best—a typical outcome, since the simulation shows that on average 2.4 inferior treatments surpass the best treatment. Noise hinders the search for the best treatment.

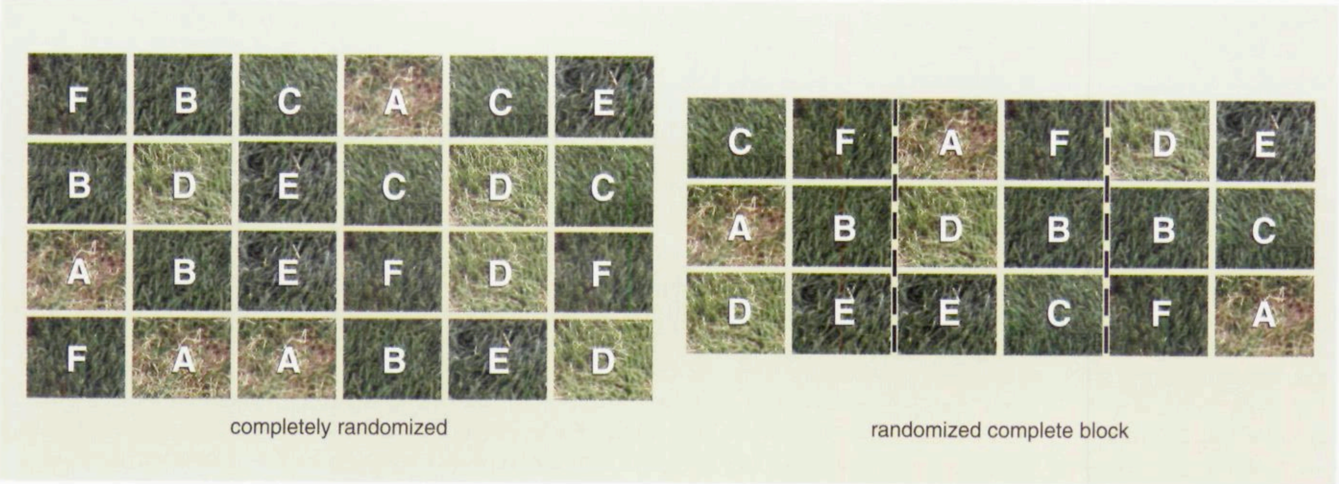

Figure 4. Blocking builds accuracy into an experiment. The simplest method, the completely randomized (CR) design, applies treatments at random to experimental units. At top left, 24 hypothetical plots of turfgrass are randomly assigned six treatments, A through F, for a total of four replications. Accuracy can be increased by the randomized complete block (RCB) design, subdividing the experimental units into as many blocks as there are replications. Statistical efficiency increases because the smaller blocks are more uniform than the whole field. The RCB trial shown at top right, with only three replications arranged from left to right, typically provides about the same level of accuracy as the four replications of the CR trial shown at top left. Shown at left is a trial for perennial ryegrass *(background)* and Kentucky bluegrass *(foreground)* from the National Turfgrass Evaluation Program, which uses RCB designs with three blocks. (Photograph courtesy of Scott Ebdon, University of Massachusetts, Amherst.)

percent. Yet in many actual experiments the odds are even worse. Realistic tasks, including many in desperately important medical research, often involve hundreds of competitors and small margins of superiority.

Suppose now that the number of replications were increased by a factor of four. This would halve the standard error of the estimated effects. The difference between the inferior and superior treatments would increase to two standard errors. (The difference itself has not changed, but the units it is measured in have gotten smaller.) Now the probability that the experiment will identify the superior treatment more than doubles, to 65.8 percent.

Given the above considerations, scientists clearly need other strategies for battling with measurement error, besides the expensive brute-force method of increasing the number of replications. However, the concept of replication does provide a useful common currency for quantifying and comparing the benefits of other error-control strategies. The error reductions or accuracy gains achieved by any method can be expressed in terms of the number of additional replications that would be required to achieve the same improvement.

Designing for Accuracy

The second strategy of blocking was also developed from crop research at Rothamsted Experimental Station, by Fisher and his protégé Frank Yates in the 1930s. It requires scientists to carefully think out the design of the experiment before beginning.

A typical experiment has one or more deliberately controlled factors of interest, such as diets for chickens or chemicals for reactions, which can be given the generic label of "treatments." In agricultural yield trials, each combination of genetic and environmental variables is considered to be a treatment. Thus, an experiment may test 30 *genotypes* (varieties differing in one or more genes) in 20 different environments, for a total of 600 treatments.

The experimental design specifies in advance how the treatments will be allocated to experimental units (plants, chickens, people). It usually incorporates both randomization and replication to minimize bias and increase accuracy. Although intended to be the same, units given the same treatment always vary because of uncontrolled factors. In the realm of the life sciences the units are never completely identical, and the application of treatments is never completely uniform. The purpose of an experimental design is to minimize the consequences of uncontrolled variation.

The simplest design, called the *completely randomized* (CR) design, simply assigns treatments to experimental units at random. The *randomized complete block* (RCB) design subdivides the experimental units into as many blocks as there are replications. Within each block, each treatment is allocated to one unit at random. The blocks are chosen to minimize uncontrolled within-block variations. For instance, in an agricultural trial, plots that are near to one another are likely to be more similar than plots separated by a greater distance. Thus the blocks are simply compact parcels that are smaller and more uniform than the whole field.

Extensive experience with RCB designs shows that one can typically achieve a "statistical efficiency" of 1.3, which means that the accuracy of the experiment is comparable to a CR design with 1.3 times as many replications. Thus a RCB with three actual replications is about as accurate as a CR with four, so the researcher has gained a full replication "for free."

And yet there are even better approaches than RCB. So-called "incomplete block designs" (in which each block receives only a subset of the treatments) also reduce the residual error, but can accomplish two other things that RCBs cannot. First, they allow the investigator to adjust estimates of treatment effects closer to their true values. Second, these adjustments can improve rankings, increasing the probability of the truly best entry winning the trial. Although blocking is pervasively popular in scientific research, scientists are rarely aware of these additional benefits of incomplete blocking.

From Frogs to Shrinkage Estimators

Though the idea of parsimony is old, its modern expression in statistics had to wait for two breakthroughs around 1955: the theory of shrinkage estimators, initiated by Charles Stein of Stanford University, and the advent of modern digital computers. Modeling is not as easy to grasp as replication and blocking, so I will begin with a "toy example" before proceeding to some real case studies.

Suppose that you are collecting frogs from a pond for a jumping-frog contest. You collect several frogs and race them over a 10-yard course. They complete the course in a range of times from 30 to 90 seconds, with an average of a minute. What is your best estimate of the time that each frog would take if it hopped the course again?

Naively, you might expect a frog that took 40 seconds the first time to finish in 40 seconds the second time. (For simplicity, let us ignore learning effects.) Indeed, that would be your best guess if you had not collected any other frogs. Also, the actual measurement is the best estimator of the true value if there are two frogs. However, Stein and other statisticians proved that it is *not* the best estimate if you observed three or more frogs. Instead, you should "shrink" each frog's deviation from the mean by a certain amount, determined from the number of frogs and the variability of the times. Suppose the formula told you to shrink your frog's deviation by 30 percent. Then you would estimate that it would finish 14 seconds faster than the mean the second time—that is, in 46 seconds instead of 40. Such a procedure is called, naturally enough, a "shrinkage estimator."

Shrinkage estimators have been proven both empirically and theoretically. Why do they work, and what do they have to do with parsimony? Without a shrinkage estimator, you would in effect be assuming that every frog is a unique animal, unrelated to every other frog, so that the other frogs' times are irrelevant to determining your frog's ability. But it is much more reasonable—and parsimonious—to assume that all of the frogs come from one genetic talent pool. Their abilities are distributed along a bell-shaped curve, with most frogs close to the mean ability and only a small percentage being especially fast or slow hoppers. Thus your frog's true ability probably lies closer to the mean than its 40-second time in the first race would suggest. Note that the shrinkage estimator uses all of the data when calculating each frog's estimate. This more vigorous use of the data increases accuracy.

Incidentally, although merely shrinking estimates toward the average value does not change rankings, more complicated experiments or models can change rankings and hence can change winners. Also, numerous statistical analyses, including multiple-regression models, can produce adjusted estimates different from the actual data, even though these may not be called "shrinkage estimators." Fitting models to data often requires millions or even billions of arithmetic steps, making computers essential.

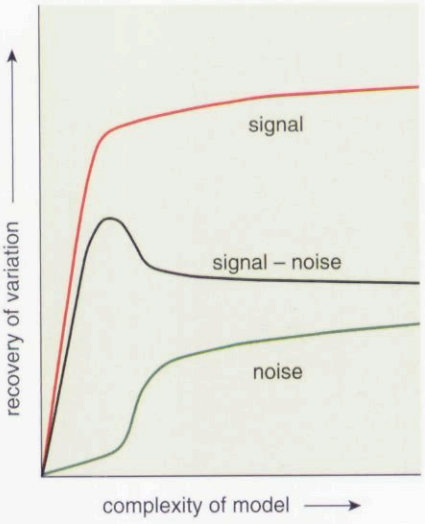

Figure 5. Modeling recovers variation from signal and noise differently depending on the complexity of the model. Signal is relatively simple, so it is captured rapidly by early model components. But noise is very complex. Noise is captured slowly at first, suppressed by considerable signal recovery, then more quickly by exploiting chance correlations in the noise, and slowly thereafter. Predictive accuracy is improved by capturing signal but degraded by capturing noise. The implied response for signal minus noise is a unimodal response called Ockham's hill, shown in Figures 7 and 9.

Moving from this toy example to real experiments, scientists in many fields routinely use statistical models for various purposes other than gaining accuracy, such as testing for significant effects or summarizing and visualizing complex information. For example, a common data format is a two-way layout or data matrix, such as the yields of 30 genotypes in 20 environments or sales of 19 products in 57 stores. Principal-components analysis, factor analysis, correspondence analysis (also called reciprocal averaging), nonmetric multidimensional scaling and other popular multivariate analyses can reduce the high-dimensional data to a two-dimensional graph that often successfully captures most of the structure in the data.

Unfortunately, precious few scientists—apart from scientists in a few specialties such as signal processing—know and exploit the fact that these familiar analyses can also serve the purpose of gaining accuracy. The models become more complex and less parsimonious as further components are added. Because the signal is

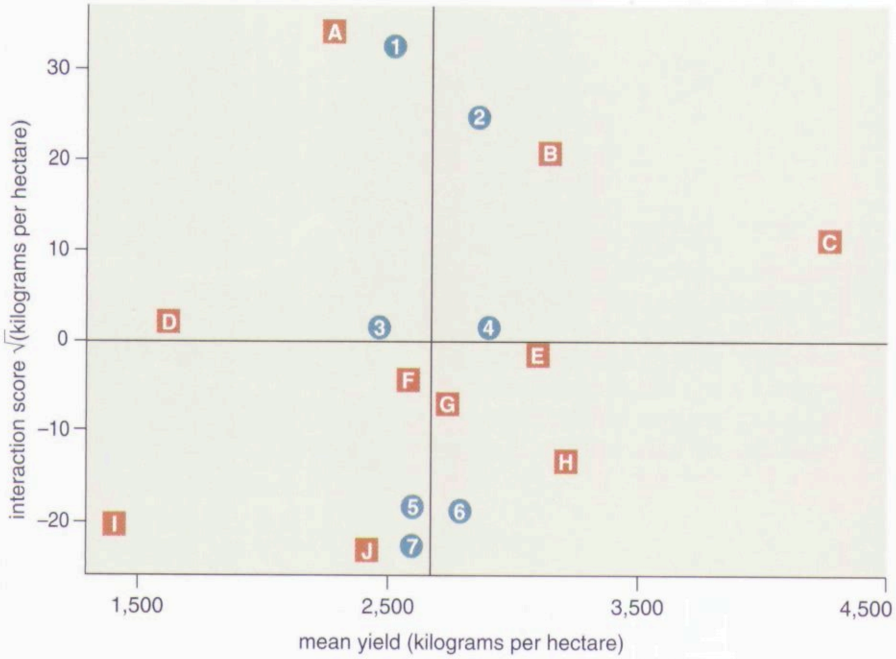

Figure 6. Statistical modeling can help in visualizing complex data as well as increasing accuracy. The AMMI (*A*dditive *M*ain Effects and *M*ultiplicative *I*nteraction) model is useful for two-way data tables whenever main (average) and interaction (differential) effects present researchers with different problems and opportunities, as in agricultural yield-trial research. AMMI first accounts for main effects and then applies principal-component analysis to the interactions. On the horizontal axis, this AMMI1 biplot shows main effects for seven soybean genotypes *(circles)* in 10 New York State environments *(squares)*. The vertical line marks the grand mean (2,678 kilograms per hectare). Those varieties and environments to the left performed poorly overall, whereas those to the right performed well. On its vertical axis, this biplot shows interaction scores. The interaction for a given genotype and environment is estimated by multiplying their scores; thus genotypes and environments with scores of the same sign have positive interactions, but opposite signs indicate negative interactions. The interaction scores reveal a trend from early-maturing varieties and their preferred short-season environments toward the top to late-maturing varieties and their preferred long-season environments toward the bottom. This biplot successfully captures 96.9 percent of the total variation in this complex data set.

relatively simple and the noise is very complex, the early model components selectively recover signal whereas the late model components selectively recover noise. By using only the early components—a form of shrinkage—one can obtain a parsimonious model that makes more accurate predictions than the data. Let us see how this plays out in three real examples.

Agricultural Trials

The turf industry in the United States is huge; about $40 billion is spent annually for lawn care. There are about 50 million acres of maintained turf in the United States, which would blanket a square 280 miles on a side. The National Turfgrass Evaluation Program (NTEP) conducts trials with several hundred entries at about 25 locations for its ongoing research to improve turfgrass quality.

Scott Ebdon, of the University of Massachusetts, and I have analyzed the NTEP data with a variant of principal-components analysis called AMMI. The salient point is that AMMI produces a model family—AMMI0, AMMI1, AMMI2, AMMI3 and so on—with more and more parameters until reaching the full model, which is identical to the data matrix. The number of parameters in the most accurate models is usually a small fraction of the number of treatments. Choices with fewer parameters underfit the real signal, whereas models with more parameters (including the full model that equals the data!) overfit spurious noise. This response, fittingly called Ockham's hill, has been explored in detail in my previous *American Scientist* article ("Prediction, Parsimony and Noise," September–October 1993) and my recent book on scientific method.

In an actual turfgrass quality trial, where we used the various members of the AMMI family to predict cross-validation data, AMMI2 was most accurate, with a statistical efficiency of 5.6. This means that it produced the same accuracy gain as collecting 5.6 times as much data. To collect that much extra data would have cost NTEP over $1 million. Hence, modeling proved to be an extremely cost-effective stratagem for accelerating improvements in turfgrass quality. Inspiration *does* save perspiration!

Parsimonious models typically achieve statistical efficiencies of 2 to 4; blocking designs achieve an efficiency of 1.3 or so. But this comparison actually understates the superiority of parsimonious modeling. We should really be comparing the *gain* in accuracy using a baseline that comes from the experiment itself. The statistical benefit from modeling (equivalent to adding 1 to 3 times the replications in the experiment itself) greatly exceeds that from blocking designs (0.3).

Despite consistent and impressive results, the adoption of aggressive statistical analysis in crop science has been slow. As Donald Nielsen noted in his presidential address to the

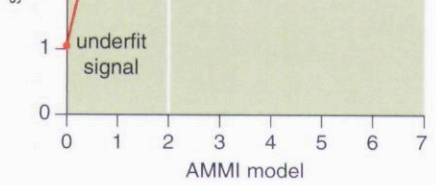

Figure 7. The AMMI model family exhibits Ockham's hill, with a relatively parsimonious model most accurate and efficient. A quality trial of perennial ryegrass, a common species for golf courses and home lawns, provides an example. In this case, AMMI2 with two principal components is most predictively accurate *(white line)*, achieving an impressive statistical efficiency of 5.6. This means that applying the model to the data produced the same accuracy as gathering 5.6 times as much data and not using the model. Models with fewer components are less accurate because they underfit real signal, whereas models with more components are less accurate because they overfit spurious noise.

American Society of Agronomy several years ago, agricultural scientists still rely mainly on statistical techniques developed before 1940—that is, in the era of Ronald Fisher. At present, worldwide breeding efforts account for an average yield increase of about 1 percent per year in the major crops. I conservatively estimate that aggressive statistical analysis of the same data would make it possible to increase this average to 1.4 percent per year, at virtually no extra cost. Over a decade, that incremental gain would translate into enough food for millions of persons.

QTL Searches

In genetics, the inheritance of a simple trait may involve a single gene, as in Gregor Mendel's classic experiments with peas. But a quantitative trait (such as yield) involves multiple genes, which are called *quantitative trait loci* (QTLs). Numerous genes with relatively small effects are considerably more difficult to locate on chromosomes than a single decisive gene. Searching for these genes is an extremely important problem, with numerous applications to crop improvements and human diseases.

QTL searches proceed as follows. All individuals in the experiment are screened for numerous genetic markers that have already been located in a chromosome map. Each individual is also measured for the quantitative trait of interest (color, disease resistance or whatever), called its *phenotype*. If groups of individuals with different versions of a given marker gene also have different phenotypes, it is reasonable to infer that a QTL for the trait exists near that marker.

Numerous statistical methods have been proposed for QTL searches during the past 15 years. Recently, Min Zhang at Cornell University and her collaborators have developed a new method with several notable advantages. The most intriguing feature, from the point of view of this article, is that its superior performance results from parsimonious modeling.

From experience, geneticists know that only a small proportion of the marker genes are actually near QTLs affecting the trait being studied. Zhang used a kind of statistical method (a Bayesian method) that can readily incorporate this crucial biological information. Her approach explicitly favors a parsimonious model with few QTLs. Extensive testing shows that this analysis efficiently detects QTLs while largely avoiding false detections. By contrast, previous methods lacking this emphasis on parsimony performed worse, and were more vulnerable to problems of missing data and small samples.

All three strategies—replicating, blocking and modeling—are relevant for accurate QTL searches. Replicating and blocking increase the accuracy of the phenotypic data, which in turn improves QTL searches. And modeling also helps, on two counts. First, modeling the phenotypic data increases accuracy before the search, as explained in the previous section on agricultural trials. Second, modeling QTLs parsimoniously improves the robustness and accuracy of the search itself, as explored in this section.

Molecular Shapes

For much biological and medical research, including drug design, scientists need to determine the three-dimensional shape of a protein or other large molecule with great accuracy. The basic shape of a molecule is constructed from information on electron densities. This initial picture is then refined using two kinds of data: noisy, empirical data on x-ray diffraction intensities for the large molecule of interest, and a data base of typical distances and angles between atoms in small molecules. Note the analogy with the toy example, which also balanced two kinds of data: the data on a given frog and the data on other frogs.

To increase accuracy, crystallographers must choose an appropriate trade-off between these two kinds of data. In a molecule with 1,000 atoms, there are 3,000 parameters to be determined—namely, the three coordinates in space of each atom. But the database can reduce the number of independent parameters by constraining or shrinking all estimates of the chemical bonds of a given kind toward the same length. Crystallographers estimate that for proteins the number of parameters

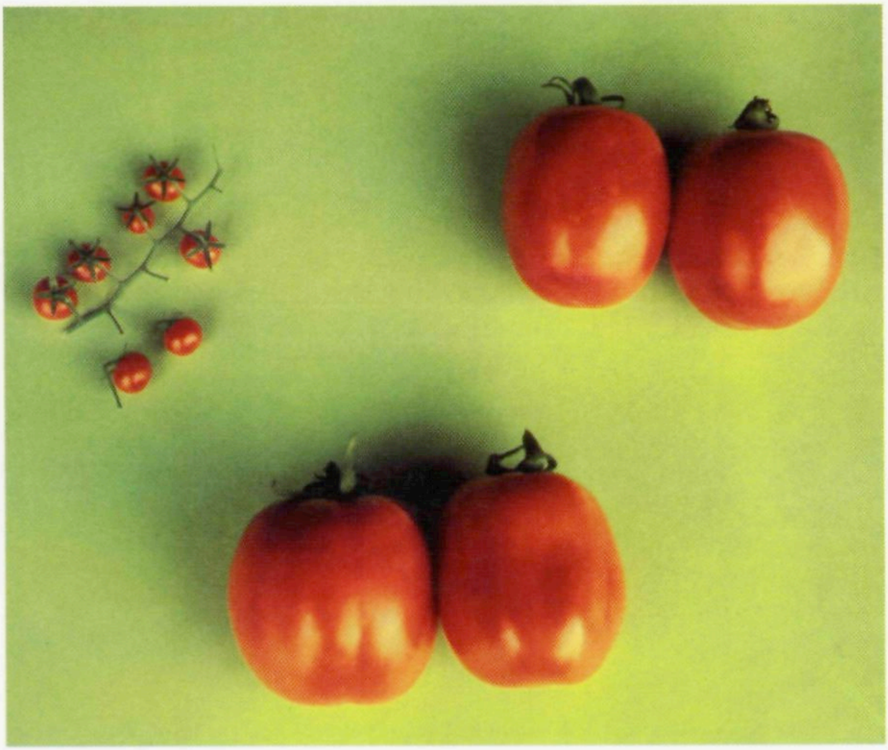

Figure 8. Quantitative trait loci (QTL), which control multiple-gene traits, are of great importance in crop improvement and human disease. By mapping QTL for yield, Steven Tanksley of Cornell University and his collaborators discovered beneficial genes in a small wild tomato from Peru *(top left)*. Transferring the genes into the commercial variety at upper right produced fruit about 10 percent larger, shown at the bottom. Parsimonious modeling, incorporating the background information that rather few marker genes are near QTLs for a given trait, has been used to greatly improve the effectiveness of QTL searches. Such techniques could be applied to searches for medically important QTLs in the human genome, such as the genes underlying high blood pressure, improving the prospects of new treatments. (Photograph courtesy of Steven Tanksley, Cornell University.)

can usually be reduced by a factor of approximately 7.5 (in this case, from 3,000 to 400). Reduction of parameters, as in the agricultural example, is a hallmark of a parsimonious model.

The question then arises of what relative weights one should place on the x-ray data and the data base in order to optimize accuracy. For decades, a measure called the R factor had been used, which was based on the amount of agreement between the original data and the restrained model. Unfortunately, in flagrant disregard of parsimony, the R factor can be made arbitrarily good by adding more parameters (overrestraining the model). In 1992, Axel Brünger, then at Yale University but now at Stanford, introduced the "free R statistic," which avoids overfitting the noise by

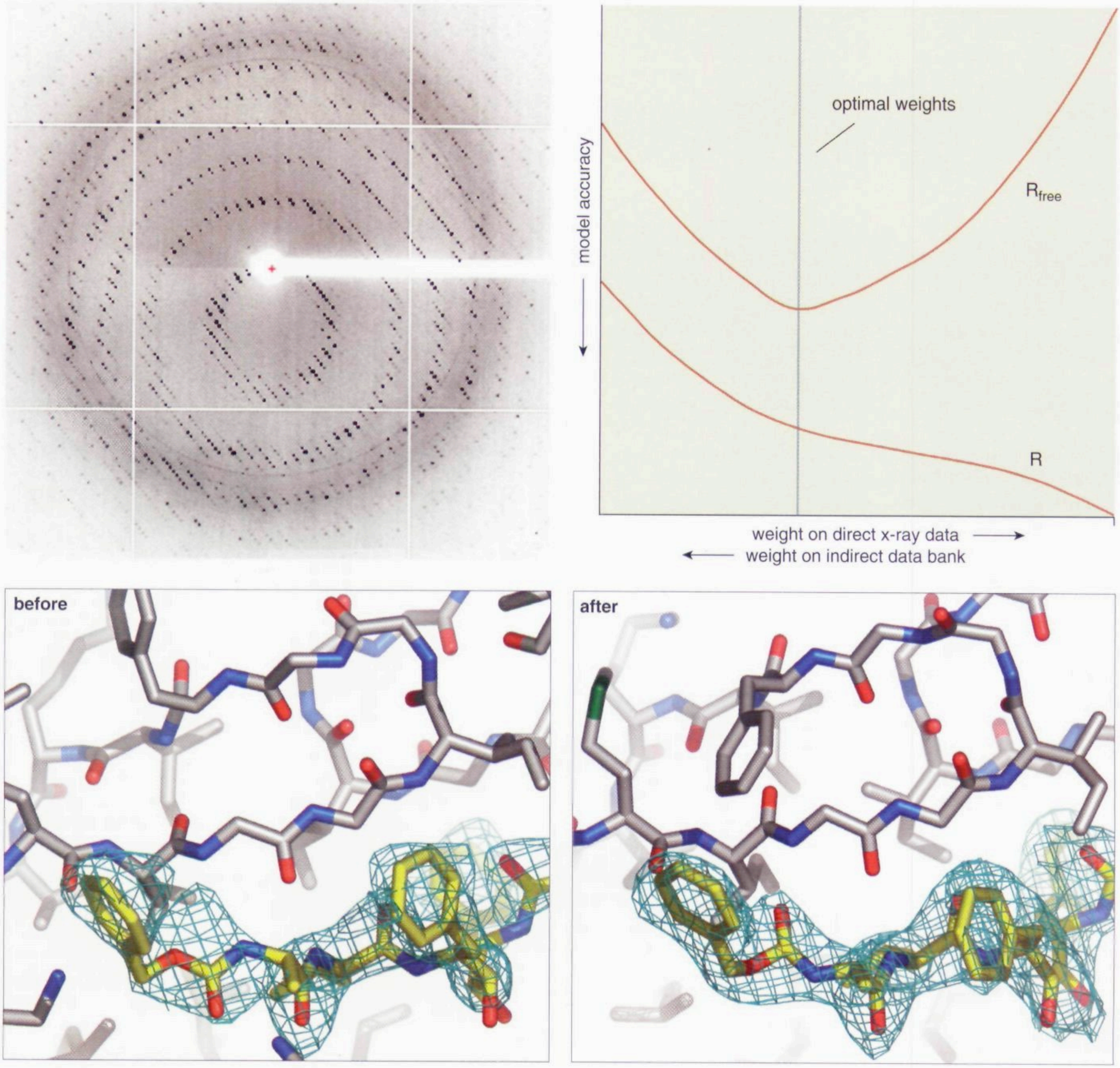

Figure 9. Accuracy is essential in determining the three-dimensional shapes of molecules for drug development. For example, protease inhibitors are an important class of drugs for combating HIV infection, and the inhibitor molecule's interaction with an HIV protease depends on its shape. Biologists determine a protein's shape by combining information from x-ray diffraction, which gives a noisy reading of the crystal structure *(top left)*, with information from a data bank containing known bond lengths and angles. Shown are two cutaway views of a model of the TL-3 inhibitor at the site where it interacts with the protease, a known structure shown extending into the background. In the "before" image, an electron-density map *(blue mesh)* calculated from the diffraction pattern is overlaid on a structural model based on information from the data bank. When proper weights are applied to these two kinds of data, an accurate structure can be fitted *("after")*. A statistic called the R factor measures the agreement between the raw diffraction data and calculated electron densities. A measure called R_{free}, which uses cross-validation to assess predictive accuracy, is now widely used to improve the accuracy of modeled protein structures. Before the adoption of the more parsimonious R_{free}, modelers typically continued to add parameters to reduce R. As the R factor decreased beyond the optimal weights now found by R_{free}, model accuracy actually declined because the models overfitted the noise *(graph)*. (Model images courtesy of Holly Heaslet and Justin Chartron, Scripps Research Institute; diffraction image courtesy of C. David Stout of Scripps.)

using cross-validation. Most of the x-ray data are used to construct the model of the molecule, but about 10 percent of the data, selected at random, are withheld in order to check the predictive accuracy of the model. Just as one would expect, as the emphasis given to the direct x-ray data increases (and the emphasis on the database decreases), the free R statistic displays Ockham's hill. The peak on Ockham's hill (here inverted) indicates the optimal choice of weights for the two kinds of data.

At the time I wrote my previous article for this magazine, Brünger's method had been introduced just the preceding year. That year, only 1 percent of the crystal structures deposited in the international Protein Data Bank had used the free R statistic. Just three years later, adoption of R_{free} had reached 33 percent, and after five years, 71 percent. By 2000, adoption had reached 92 percent, and at present R_{free} is nearly always reported. Thus crystallography constitutes an encouraging case study in how rapidly parsimonious modeling can become accepted (or even required) when data are expensive and limited, but computation and modeling are cheap.

Science Education

Two facts are evident. On the one hand, all statisticians know that parsimonious modeling can increase accuracy and efficiency. On the other hand, few scientists know about this great opportunity. What explains this mismatch, and what is the remedy?

In part, I would suggest that modeling is neglected because of scientists' complacency. Many scientists think that once they have done replication and blocking, they are finished with their statistical homework, thank you very much. The goal of this article is to disturb this unfounded complacency. Replication accomplishes less than many scientists expect it to, blocking is often done by suboptimal designs, and modeling to gain accuracy is routinely neglected.

Why do scientists sometimes fail at the accuracy game? I would pin the blame on a common deficiency in their training. Scientific research requires mastery of both the general principles of science and the specialized techniques of a particular discipline, but the emphasis can fall too heavily on the latter. The community of investigators in a given specialty can remain quite unaware of fine examples of the uses of statistical methods that can be found in other disciplines or literatures, and thus miss opportunities to gain accuracy through modeling in their own work. Astronomers or geologists who see an example of successful modeling in agriculture or chemistry ought to be able to distinguish transferable general principles that they can import into their own specialties. Parsimony and its relation to accuracy (as described by Ockham's hill) is one such principle, with pervasive relevance in science and technology.

Innovation comes not only from inventors of new ideas, but also from importers of relevant ideas. This fact has been recognized by recent position papers on U.S. science education that emphasize versatility in the scientific workforce and the transfer of knowledge between disciplines.

Manifestly, it is precisely the general principles of scientific method that are the most cosmopolitan facet of science. These constitute the most frequent and promising candidates for enhancing knowledge transfer and greater adaptability.

When a scientist plans an experiment with a statistical consultant, or when a student plans research with a thesis advisor, the conversation is likely to include recommendations for replicating and blocking to gain accuracy. After settling that, the scientist or student should draw but one breath before using it to ask, "And what about also gaining accuracy from parsimonious modeling of the treatments?" Scientific research is expensive, so it is imperative to use resources efficiently. Anything less is irresponsible.

Surely at least 25 percent of scientific projects have one or more key steps where parsimonious modeling is applicable, but not yet implemented. It may well be that the agricultural example, in which the current rate of progress could be improved by 40 percent simply from better use of data already in hand, is a typical case. If so, simple multiplication suggests that we are missing out on 10 percent of the potential return from our investment in scientific experiments. A modest investment in training, with attention paid to such general principles of scientific method as parsimony, could reap tremendous benefits for scientific knowledge and for the development of new commercial and medical products. Extra accuracy at trivial cost is a great bargain.

Bibliography

Berger, J. O., and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76:159–165.

Dean, A., and D. Voss. 1999. *Design and Analysis of Experiments*. New York: Springer-Verlag.

Ebdon, J. S., and H. G. Gauch. 2002. Additive main effect and multiplicative interaction analysis of national turfgrass performance trials: II. Cultivar recommendation. *Crop Science* 42:497–506.

Gauch, H. G. 1992. *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. New York: Elsevier (Chinese edition 2001, Hangzhou: China National Rice Research Institute).

Gauch, H. G. 1993. Prediction, parsimony and noise. *American Scientist* 81:468–478.

Gauch, H. G., and R. W. Zobel. 1996. Optimal replication in selection experiments. *Crop Science* 36:838–843.

Gauch, H. G. 2002. *Scientific Method in Practice*. Cambridge, U.K.: Cambridge University Press (Chinese edition 2004, Beijing: Tsinghua University Press).

Glazier, A. M., J. H. Nadeau and T. J. Aitman. 2002. Finding genes that underlie complex traits. *Science* 298:2345–2349.

Jefferys, W. H., and J. O. Berger. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80:64–72.

Kleywegt, G. J., and T. A. Jones. 2002. Homo Crystallographicus—Quo Vadis? *Structure* 10:465–472.

National Academy of Sciences. 1995. *Reshaping the Graduate Education of Scientists and Engineers*. Washington, D.C.: National Academy Press.

National Science Foundation. 1996. *Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology*. Arlington, VA: National Science Foundation.

Tanksley, S. D., and S. R. McCouch. 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* 277:1063–1066.

Zhang, M., K. L. Montooth, M. T. Wells, A. G. Clark and D. Zhang. 2005. Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* 169:2305–2318.

For relevant Web links, consult this issue of *American Scientist Online*:

http://www.americanscientist.org/IssueTOC/issue/821