HEINONLINE

Citation: 109 Colum. L. Rev. 2081 2009

Content downloaded/printed from
HeinOnline (http://heinonline.org)
Fri May 25 11:57:45 2012

-- Your use of this HeinOnline PDF indicates your acceptance of HeinOnline's Terms and Conditions of the license agreement available at http://heinonline.org/HOL/License

-- The search text of this PDF is generated from uncorrected OCR text.

-- To obtain permission to use this article beyond the scope of your HeinOnline license, please use:

https://www.copyright.com/ccc/basicSearch.do?
&operation=go&searchType=0
&lastSearch=simple&all=on&titleOrStdNo=0010-1958

VANDERBILT UNIVERSITY

DiscoverArchive

Retrieved from DiscoverArchive,
Vanderbilt University's Institutional Repository

This work was originally published in
109 Colum. L. Rev. 2081 2009.

ESSAY

A PRACTICAL SOLUTION TO THE REFERENCE CLASS PROBLEM

*Edward K. Cheng**

The "reference class problem" is a serious challenge to the use of statistical evidence that arises in a wide variety of cases, including toxic torts, property valuation, and even drug smuggling. At its core, it observes that statistical inferences depend critically on how people, events, or things are classified. As there is (purportedly) no principle for privileging certain categories over others, statistics become manipulable, undermining the very objectivity and certainty that make statistical evidence valuable and attractive to legal actors. In this Essay, I propose a practical solution to the reference class problem by drawing on model selection theory from the statistics literature. The solution has potentially wide-ranging and significant implications for statistics in the law. Not only does it remove another barrier to the use of statistics in legal decisionmaking, but it also suggests a concrete framework by which litigants can present, evaluate, and contest statistical evidence.

Statistics are often at the center of contemporary litigation, whether the case involves toxic torts, predictions of violence, property valuation, or even drug smuggling. Yet, despite the growing prevalence and importance of statistical evidence, a number of commentators have recently raised a serious challenge to its use in the legal system by invoking the so-called "reference class problem."

The problem is perhaps best introduced through the colorful case of *United States v. Shonubi*.[1] In *Shonubi*, Judge Weinstein faced the difficult task of estimating the amount of heroin smuggled by a Nigerian drug "mule." Officials had apprehended Charles Shonubi at New York's John F. Kennedy International Airport (JFK) with 427.4 grams of heroin in his digestive tract, and a jury had subsequently convicted him.[2] The sentenc-

* Professor of Law, Brooklyn Law School; Ph.D. Student, Department of Statistics, Columbia University. Many thanks to Ron Allen, Cliff Carrubba, Jenny Diamond Cheng, Tom Clark, Neil Cohen, Justin Esarey, George Fisher, David Freedman, Susan Herman, David Kaye, Jeffrey Lipshaw, David Madigan, Richard Nagareda, Dale Nance, Mike Pardo, Jim Park, David Rosenberg, Chris Sanchirico, David Schum, Jeff Staton, and Peter Tillers for reading drafts, offering helpful comments, and providing other assistance. This Essay also benefited from workshops given at the University of Utah, the University of Colorado, Emory University, the University of Alabama, Vanderbilt University, the Northeast Law & Society Conference held at Amherst College, and the Brooklyn Law School Brown Bag Series. Thanks to Aran McNerney and Casey Kroma for research assistance. Research support was provided by the Brooklyn Law School Dean's Summer Research Fund and the Project on Scientific Knowledge and Public Policy.

1. 895 F. Supp. 460 (E.D.N.Y. 1995).

2. See id. at 464.

HeinOnline -- 109 Colum. L. Rev. 2081 2009

ing guidelines, however, required the court to determine the *total* amount of heroin imported in the course of conduct,[3] so the court had to estimate the amount carried by Shonubi on seven previously undetected smuggling trips.[4] To do so, Judge Weinstein used statistical data from the United States Customs Service for Nigerian drug mules at JFK during the time period in question[5] to construct a statistical model.[6] The judge then concluded that Shonubi had carried between 1,000 and 3,000 grams of heroin over his eight trips and sentenced him to 151 months in prison.[7]

Judge Weinstein's method may initially appear quite straightforward. In making its statistical estimate, however, the court necessarily had to classify Shonubi into some group or "reference class," and this decision as to which reference class to use is a tricky question. For instance, why was "Nigerian drug mules at JFK during the time period" the correct reference class?[8] The court could have just as easily looked at a bewildering array of alternatives, all of which would have yielded different estimates. The court could have considered the amount carried by all drug smugglers at JFK, all Nigerian smugglers regardless of airport, or smugglers in general. It could have used Shonubi himself as the reference class, meaning that the estimate would be 8 x 427.4 = 3,419.2 grams.[9] It could have even made intuitively odd reference class choices such as all airline passengers or all toll booth collectors at New York's George Washington Bridge (Shonubi's day job),[10] both of which would have yielded very low estimates. As Mark Colyvan and others argue, "Shonubi is a member of

3. Id. at 467.

4. Id. at 466 ("Based on evidence at the trial and at sentencing, the judge found that the defendant had made a total of eight smuggling trips to Nigeria between September 1, 1990 and December 10, 1991."). The specific type of judicial factfinding practiced in *Shonubi* is now surely unconstitutional in the wake of United States v. Booker, 543 U.S. 220 (2005), which requires that such sentencing enhancements be found by the jury. Nevertheless, the fundamental statistical problem remains, irrespective of the context or decisionmaker. Peter Tillers, If Wishes Were Horses: Discursive Comments on Attempts to Prevent Individuals from Being Unfairly Burdened by Their Reference Classes, 4 Law, Probability & Risk 33, 33 n.†, 36 n.12 (2005).

5. *Shonubi*, 895 F. Supp. at 499–504 (describing customs service data).

6. Id. at 521–23 (describing model).

7. Id. at 530.

8. See Mark Colyvan, Helen M. Regan & Scott Ferson, Is It a Crime to Belong to a Reference Class?, 9 J. Pol. Phil. 168, 172–73 (2001) [hereinafter Colyvan et al., Is It a Crime] (discussing possibility of other classifications); Paul Roberts, From Theory into Practice: Introducing the Reference Class Problem, 11 Int'l J. Evidence & Proof 243, 247 (2007) (noting Judge Weinstein's court-appointed expert panel of statistician David Schum and law professor Peter Tillers challenged statistics offered by prosecution because of reference class problem).

9. This basic multiplicative solution is actually what Judge Weinstein applied as an initial matter, United States v. Shonubi, 802 F. Supp. 859, 860–61, 864 (E.D.N.Y. 1992), until he was reversed by the Second Circuit, United States v. Shonubi, 998 F.2d 84 (2d Cir. 1993); see also *Shonubi*, 895 F. Supp. at 466–68 (detailing procedural history).

10. *Shonubi*, 895 F. Supp. at 465; Colyvan et al., Is It a Crime, supra note 8, at 172.

HeinOnline -- 109 Colum. L. Rev. 2082 2009

many (in fact infinitely many) reference classes; some of these classes consist largely of unsavory types while others consist largely of saints."[11]

We thus have a problem—namely, which reference class, and accordingly which estimate, is the most appropriate one? More important, in any given case, how is the judge or jury supposed to decide? The parties can of course argue ad nauseam about which set of characteristics is more important, but ultimately, the choice of category is often seemingly arbitrary, and disturbingly so. Many classifications will appear perfectly reasonable, yet the choice may mean the difference between a long and short prison sentence, or a large or small damage award.

This troubling state of affairs is commonly called the "reference class problem." Statistical inferences depend critically on how people, events, or things are classified. The problem is that there is an infinite number of possible characteristics, and (purportedly) no principle for privileging certain characteristics over others. As a result, statistics arguably become highly manipulable.

The resulting manipulability has the potential to completely undermine the objectivity and certainty that make statistical evidence so promising and attractive for use in judicial determinations.[12] Indeed, although formerly confined to more philosophical discussions, the seriousness of the reference class problem has recently attracted greater attention from courts and evidence scholars.[13] Notably, Ron Allen and Mike Pardo argued in the *Journal of Legal Studies* that the reference class problem significantly limits the usefulness of mathematical or statistical models of evidence.[14] This article in turn spawned a special symposium issue of the *International Journal of Evidence and Proof*.[15] Conference participants universally agreed that the reference class problem is a critical issue con-

11. Colyvan et al., Is It a Crime, supra note 8, at 172.

12. See Ronald J. Allen & Michael S. Pardo, The Problematic Value of Mathematical Models of Evidence, 36 J. Legal Stud. 107, 109 (2007) (questioning usefulness of statistical models because of reference class problem); see also Dale A. Nance, The Reference Class Problem and Mathematical Models of Inference, 11 Int'l J. Evidence & Proof 259, 267 (2007) (noting it would be "catastrophic for any system of trials . . . [to require] that every judgment explicitly or implicitly adopting a reference class . . . be explored and debated").

13. Allen & Pardo, supra note 12, at 109 (submitting that while "application of the probability theory to juridical proof . . . is interesting, instructive, and insightful[,] . . . it also suffers from the deep conceptual problem . . . of reference classes"); Mark Colyvan & Helen M. Regan, Legal Decisions and the Reference Class Problem, 11 Int'l J. Evidence & Proof 274, 276 (2007) [hereinafter Colyvan & Regan, Legal Decisions] ("[G]iven that the different reference classes provide different answers to the probability assignment in question, there is considerable uncertainty about the probability assignment itself."); Tillers, supra note 4, at 48 ("If statistical reasoning based on reference classes is ever to work, such reasoning can work only if resort is made to reference classes that consist at least in part of events that are *not* generated by the choices and behaviour of the individual about whom inferences are under consideration.").

14. Allen & Pardo, supra note 12, at 135.

15. Symposium, Special Issue on the Reference Class Problem, 11 Int'l J. Evidence & Proof 243 (2007); see also Roberts, supra note 8 (introducing symposium).

HeinOnline -- 109 Colum. L. Rev. 2083 2009

fronting statistical evidence in law and that more research needs to be done.[16] Some commentators seemed to hold out hope for a solution,[17] while others were more pessimistic.[18]

In this Essay, I propose a practical solution to the reference class problem in legal contexts. My argument proceeds in two discrete steps. First, I draw a connection between the reference class problem and the problem of model selection in statistics. This link opens a host of new perspectives and tools. For example, the literature frequently treats the choice of reference class as either indeterminate or involving intuitive or subjective judgment.[19] By drawing the analogy to model selection, I show that concrete and reasonably objective criteria exist for judging reference classes.

Second and perhaps more importantly, I argue that model selection methods solve the reference class problem in legal proceedings. Because legal proceedings involve a finite number of possible reference classes proposed by the parties, model selection methods provide us with a tool for determining which proffered reference class is most appropriate. In short, in the legal context, the reference class problem is not as intractable as scholars have assumed.

I should emphasize at the outset that my solution is limited to the legal sphere. The Essay makes no attempt at solving the philosophical reference class problem.[20] For purposes of the legal system, however, my solution is sufficient, and if I am correct, the implications could be substantial. Not only would solving the reference class problem remove another barrier to the use of statistical models in legal decisionmaking, but it would also suggest a concrete framework for litigants and courts to present and evaluate statistical evidence.

---

16. See, e.g., Nance, supra note 12, at 272 (concluding "more work needs to be done on the theory of reference class selection, on how people do and ought to select reference classes for the purposes of assessing probabilities and drawing inferences").

17. Colyvan et al., Is it a Crime, supra note 8, at 172 ("We are not claiming that there is no solution to the reference-class problem, just that there is no straightforward solution . . . ."); Tillers, supra note 4, at 38 n.21 (discussing what "a 'solution' to 'the' reference class problem would do"). Dale Nance seems to be more ambivalent on the possibility of a solution. Compare Nance, supra note 12, at 272 ("[M]ore work needs to be done . . . on how people do and ought to select reference classes."), with id. (remarking that the proposition that "different advocates can always argue for different classes that generate different frequencies . . . is probably true").

18. Allen & Pardo, supra note 12, at 112 ("[N]othing in the natural world privileges or picks out one of the classes as the right one; rather, our interests in the various inferences they generate pick out certain classes as more or less relevant.").

19. Id. at 113 ("There is no a priori correct answer [to the question of which reference class to use]; it depends on the interests at stake."); Colyvan & Regan, Legal Decisions, supra note 13, at 275 ("[T]here is no principled way to establish the relevance of a reference class.").

20. See generally Alan Hájek, The Reference Class Problem Is Your Problem Too, 156 Synthese 563, 564 (2007) (discussing the problem writ large).

HeinOnline -- 109 Colum. L. Rev. 2084 2009

Parts I and II provide introductory material on both the reference class and model selection problems. Part III links the two problems and details the proposed solution to the reference class problem. Part IV addresses criticisms of and limitations on the solution, and Part V offers some applications.

I. THE REFERENCE CLASS PROBLEM

The reference class problem is a fundamental aspect of statistical inference.[21] Inference often involves abstracting a person (or event or thing) to a few salient characteristics, and then comparing that person with others having the same or similar characteristics. However, the problem becomes: How does one choose the comparison group?

To take a simple non-legal example, consider the often discussed statistic that nearly half of all marriages end in divorce.[22] Based on this statistic, what inferences might you draw at a wedding about the bride and groom? Do they actually have a 50/50 chance of remaining married? It depends entirely on how you classify the two newlyweds. As members of the general population, their chance is indeed approximately 50/50.[23] We can seemingly improve our guess, however, by incorporating more individualized characteristics. For example, studies show that a couple's risk of divorce changes based on income, age, education, religious affiliation, and whether their parents have intact marriages.[24]

But which characteristics should we use? The problem is that the divorce rate will change—and sometimes change dramatically—depending on the characteristics or "reference class" chosen. The divorce rate for college graduates will yield one number, while the rate for thirty-year-olds will yield another, and that for couples making between $90,000 and $100,000 who attended small liberal arts colleges in Wisconsin still another. Since every couple falls under an infinite number of classifications, the options become almost paralyzing. One might be tempted to choose the narrowest class since it seems to use all of the information available about the couple. But the narrowest class is in fact the couple itself, which is unique and does not enable us to make any inferences at all.[25]

---

21. According to Alan Hájek, the problem was probably first noted by John Venn, who is most famous for Venn diagrams. The term "reference class problem," however, is attributed to Hans Reichenbach in 1949. Id. at 564.

22. Dan Hurley, Divorce Rate: It's Not as High as You Think, N.Y. Times, Apr. 19, 2005, at F7.

23. Id.

24. David Popenoe, The Future of Marriage in America, *in* Barbara Dafoe Whitehead & David Popenoe, The State of Our Unions: The Social Health of Marriage in America 18 (2007), available at http://marriage.rutgers.edu/Publications/SOOU/SOOU2007.pdf (on file with the *Columbia Law Review*); Hurley, supra note 22, at F7.

25. Branden Fitelson, Comments on James Franklin's 'The Representation of Context: Ideas from Artificial Intelligence' (Or, More Remarks on the Contextuality of Probability), 2 Law, Probability & Risk 201, 203 (2003) (noting that you cannot reduce

HeinOnline -- 109 Colum. L. Rev. 2085 2009

The import of the reference class problem is far from academic and is not merely confined to parlor games at weddings or quirky cases of drug mules arriving in New York. Indeed, the reference class problem arguably arises every day in courtrooms across the country.[26] It has ramifications all over the legal landscape. Consider the following additional examples, which are only the tip of the iceberg.[27]

A. *Property Valuation*

For a variety of reasons, including eminent domain condemnations, tax assessment, tort damages, and insurance coverage, courts often need to assess property values. One prevalent technique is to look at the sale prices from comparable properties.[28] The problem is that each party will offer its own, differing comparison group in an attempt to secure a favorable outcome.[29] At the same time, prevailing doctrine tasks the jury with determining which comparison groups are appropriate, often with little, if any, guidance.[30] As the court in *Loeffel Steel Products, Inc. v. Delta Brands, Inc.* eloquently stated, "care must be taken to be sure that the comparison is one between 'apples and apples' rather than one between 'apples and oranges.'"[31] But as simple as the advice seems, the task is far more complicated.

B. *Toxic Torts*

The plaintiff's background risk of cancer or some other disease is often central in a toxic tort case, since a doubling of the risk is sometimes

---

reference class to single person because then probability is either 0 or 1, which would correspond to the truth, which is precisely what is unknown).

26. Roberts, supra note 8, at 245 (arguing that because "[e]very factual generalisation implies a reference class, . . . this in turn entails that the reference class problem is an inescapable concomitant of inferential reasoning and fact-finding in legal proceedings").

27. See, e.g., Allen & Pardo, supra note 12, at 113 (discussing reference class problem in determining error rates for eyewitness identifications). Although technically outside the evidentiary context, but no less important, Rob Rhee argues that assessment of case values for settlement purposes falls prey to the reference class problem, because these case valuations "can be framed from the reference point of the judge, the court and forum, the attorneys, the parties (if repeat players), the type of action, the type of injury, the legal framework, and—not the least of which—the evidentiary assessment." Robert J. Rhee, Probability, Policy and the Problem of the Reference Class, 11 Int'l J. Evidence & Proof 286, 289 (2007).

28. E.g., Adcock v. Miss. Transp. Comm'n, 981 So. 2d 942, 947–48 (Miss. 2008) (using comparisons to value property taken by eminent domain).

29. E.g., Engquist v. Wash. County Assessor, No. TC-MD 030303F, 2003 WL 23883581, at *1 (Or. T.C. Magis. Div. July 29, 2003) (illustrating problem in tax assessment context).

30. E.g., United States v. 819.98 Acres of Land, 78 F.3d 1468, 1472 (10th Cir. 1996) ("A dissimilarity between sales of property proffered as comparable sales and the property involved in the condemnation action goes to the weight, rather than to the admissibility of the evidence of comparable sales.").

31. 387 F. Supp. 2d 794, 812 (N.D. Ill. 2005) (invoking time-honored admonition in case involving comparison groups for calculating lost profits).

HeinOnline -- 109 Colum. L. Rev. 2086 2009

used to prove specific causation.[32] But in calculating this background risk of cancer, what characteristics of the plaintiff should the court use?[33] Is the relevant risk that among white males, forty-five year olds, smokers, residents of Idaho, or some combination of these and many other traits?

C. *Class Actions*[34]

Under Rule 23 of the Federal Rules of Civil Procedure, courts may certify classes only where there is sufficient commonality between class members.[35] The problem is determining which shared characteristics are important for assessing commonality.[36] For example, in *Dukes v. Wal-Mart, Inc.*, an employment discrimination case, the court needed to find that all members of the plaintiff class were similarly injured.[37] Among other things, the plaintiffs offered a statistical analysis showing gender disparities in compensation and promotion practices at the "regional level," an aggregation of eighty to eighty-five stores.[38] Wal-Mart, in contrast, argued that a store-by-store analysis was more appropriate.[39] Who was right?

D. *DNA*

One of the most fundamental aspects of DNA evidence is the random match probability (RMP), the probability that a person chosen randomly from the population will have the same profile as the one found at

---

32. E.g., In re Silicone Gel Breast Implants Prods. Liab. Litig., 318 F. Supp. 2d 879, 893–94 (C.D. Cal. 2004) (discussing how doubling of background risk provides evidence that substance caused specific plaintiff's disease).

33. See, e.g., Colyvan & Regan, Legal Decisions, supra note 13, at 275 (using probability of contracting lung cancer as example of reference class). In an entertaining article, evolutionary theorist Stephen Jay Gould wrote about his diagnosis of abdominal mesothelioma in 1982. See Stephen Jay Gould, The Median Isn't the Message, Discover, June 1985, at 40–42. His initial shock at the short median lifespan, eight months, quickly faded away as he sharpened his reference class, learning that he "possessed every one of the characteristics conferring a probability of longer life": youth, early diagnosis, good medical care, and a healthy outlook on life. Id. Gould lived another twenty years before his death in 2002. Carol Kaesuk Yoon, Stephen Jay Gould, 60, Is Dead; Enlivened Evolutionary Theory, N.Y. Times, May 21, 2002, at A1.

34. Many thanks to Richard Nagareda for suggesting this example.

35. Fed. R. Civ. P. 23 (permitting court to certify class "only if . . . there are questions of law or fact common to the class" and common questions "predominate over any questions affecting only individual members"); see Richard A. Nagareda, Class Certification in the Age of Aggregate Proof, 84 N.Y.U. L. Rev. 97, 102–03 (2009) (discussing difficulties in obtaining class certification).

36. Class certification questions are arguably more complex than other reference class questions. For one thing, they are not strictly factual, but are instead highly interwoven with issues of administrative efficiency and substantive policy. As such, reference class issues in this context are largely beyond the scope of the Essay.

37. 509 F.3d 1168, 1195 (9th Cir. 2007).

38. Id. at 1180 & n.5.

39. Id. at 1181 (noting Wal-Mart's expert apparently looked at the "*sub-store* level by comparing departments to analyze the pay differential").

HeinOnline -- 109 Colum. L. Rev. 2087 2009

the crime scene. Yet, which population is appropriate for calculating the RMP? The entire human population? The defendant's racial subgroup? The city in which the crime occurred? In *Darling v. State*, the defendant, a Bahamian native, was on trial for the rape and murder of a woman in Orlando, Florida.[40] The defendant argued that instead of using the FBI's African American population database, the DNA expert should have used a Bahamian database.[41] Should the expert have done so?[42]

Furthermore, what about the reporting of lab error rates in DNA cases? Should those be national, regional, or specific to the individual lab or technician? Should those statistics be further narrowed if studies show that technicians work more reliably midweek than on Friday before quitting time?

E. *Modus Operandi Evidence*

In *United States v. Trenkler*, at issue was the identity of the maker of a bomb that had killed a police officer.[43] At trial, the prosecution introduced an expert who had used a government database system to compare the bomb's attributes with a bomb that the defendant had previously detonated.[44] For example, both bombs had been placed underneath vehicles, used magnets, and involved remote controls.[45] Among the over 40,000 incidents in the database, only seven had all of these characteristics, and only the defendant's prior bomb had also used "duct tape, soldering, AA batteries, toggle switches, and 'round' magnets."[46] Although the First Circuit ultimately found this database evidence to violate the hearsay rule, Chief Judge Torruella in dissent raised what was essentially a reference class issue. He observed that the database "list[ed] approximately twenty-two characteristics . . . but [that the expert], inexplicably, chose only to query ten of those characteristics."[47] He further noted that selection of other characteristics might have suggested that the two bombs were not a match at all.[48] In *Trenkler*, the prosecution's expert clearly chose those characteristics that would most inculpate the defendant; the defendant appeared to offer no counter class in return.

---

40. 808 So. 2d 145 (Fla. 2002).

41. Id. at 159.

42. The Florida Supreme Court largely evaded the problem by arguing that any error associated with use of the African American database was negligible. See id. David Kaye has suggested that the proper reference class was men living in Orlando, while acknowledging that if there had been a Bahamian community in Orlando, the class may have required further narrowing. D.H. Kaye, Logical Relevance: Problems with the Reference Population and DNA Mixtures in *People v. Pizarro*, 3 Law, Probability & Risk 211, 212 & n.15 (2004).

43. 61 F.3d 45 (1st Cir. 1995).

44. Id. at 49–50.

45. Id. at 50 & n.6.

46. Id. at 50 & nn.6–7.

47. Id. at 64–65 (Torruella, C.J., dissenting).

48. Id. at 66–67.

HeinOnline -- 109 Colum. L. Rev. 2088 2009

* * *

Stepping back, the reference class problem seems almost paradoxical.[49] Theoretically, choosing a reference class is supposedly intractable. Yet, as Dale Nance observes, ordinary people make statistical inferences every day without becoming paralyzed by the reference class problem, and they presumably make reasonably good classification choices.[50] For example, some of the possible reference classes in *Shonubi* just seem obviously wrong. Using statistics on the average amount of drugs smuggled by all airline passengers (i.e., approximately zero) seems naïvely broad. Using statistics on George Washington Bridge tollbooth collectors seems strikingly irrelevant.

The problem with a purely intuitive approach, however, is that it only provides a rough guide. In many cases, multiple reference classes will appear plausible, and if each of these classes leads to different conclusions, the problem remains.[51] For example, whether Judge Weinstein should have used statistics on Nigerian drug mules at JFK or drug mules with Shonubi's height and weight characteristics is a much more difficult question. Similarly, whether a home should be classified as a three–bedroom with one bath or a three-bedroom set on a hill for valuation purposes has no obvious answer.

Furthermore, if the only method for selecting appropriate reference classes was intuition, that result would foreclose meaningful reasoned decisionmaking and run the danger of becoming highly subjective. This subjectivity is what motivated Allen and Pardo to call the usefulness of statistical models of evidence into question. If statistics depend on the choice of reference class, and selecting a reference class depends on "argument and, ultimately, judgment,"[52] then mathematical models have not advanced the ball by much.

Nevertheless, the existence of an intuitive sense about proper reference classes should be an encouraging sign. It suggests that we may be able to distill more objective criteria for selecting a reference class. As it turns out, those criteria can be found in the statistical concepts surrounding model selection.

---

49. Colyvan et al., Is It a Crime, supra note 8, at 172–74 (disclaiming that reference class problem is unsolvable, but showing that obvious solutions fail).

50. Nance, supra note 12, at 263 n.9 ("[P]eople drawing inferences routinely order reference classes as better or worse relative to their inferential task. Presumably, they do so with some success.").

51. See Colyvan & Regan, Legal Decisions, supra note 13, at 275 (showing intractability of reference class problem where multiple classes appear plausible); Hájek, supra note 20, at 565 (same).

52. Allen & Pardo, supra note 12, at 115.

II. Model Selection

A. *The Model Selection Problem*

As its name implies, model selection is a problem about how we can best statistically model a given phenomenon.[53] For a basic example, consider the problem of fitting a line to a set of data points. Assume that we would like to predict a student's GPA based on the number of study hours he puts in. Through observations, we have the data shown in Figure 1a, which suggest some relationship between the number of study hours per week and GPA. An upward trend is clearly present, but what exactly is the relationship?

The simplest and most obvious choice might be a line, and often a linear relationship is assumed, as in Figure 1b.[54] The slight curve among the data points, however, might suggest that a quadratic relationship may be more appropriate, as in Figure 1c. Indeed, nothing in theory prevents us from fitting ever more complex curves, including the fourth-degree polynomial in Figure 1d, or an nth-degree polynomial that exactly passes through every point in the dataset as seen later in Figure 2. We thus have multiple candidates for models.

53. See generally Walter Zucchini, An Introduction to Model Selection, 44 J. Mathematical Psychol. 41 (2000) (offering short and less technical introduction to concepts in model selection).

54. Determining exactly which line best "fits" the data points presents another issue of inference, but this problem is reasonably well understood. A common method is least-squares estimation, in which the sum of the squared distances between the line and each of the points is minimized. See Malcolm R. Forster, Key Concepts in Model Selection: Performance and Generalizability, 44 J. Mathematical Psychol. 205, 210 (2000) [hereinafter Forster, Key Concepts] (discussing link between least-squares method and maximum-likelihood method often favored by statisticians).

HeinOnline -- 109 Colum. L. Rev. 2090 2009

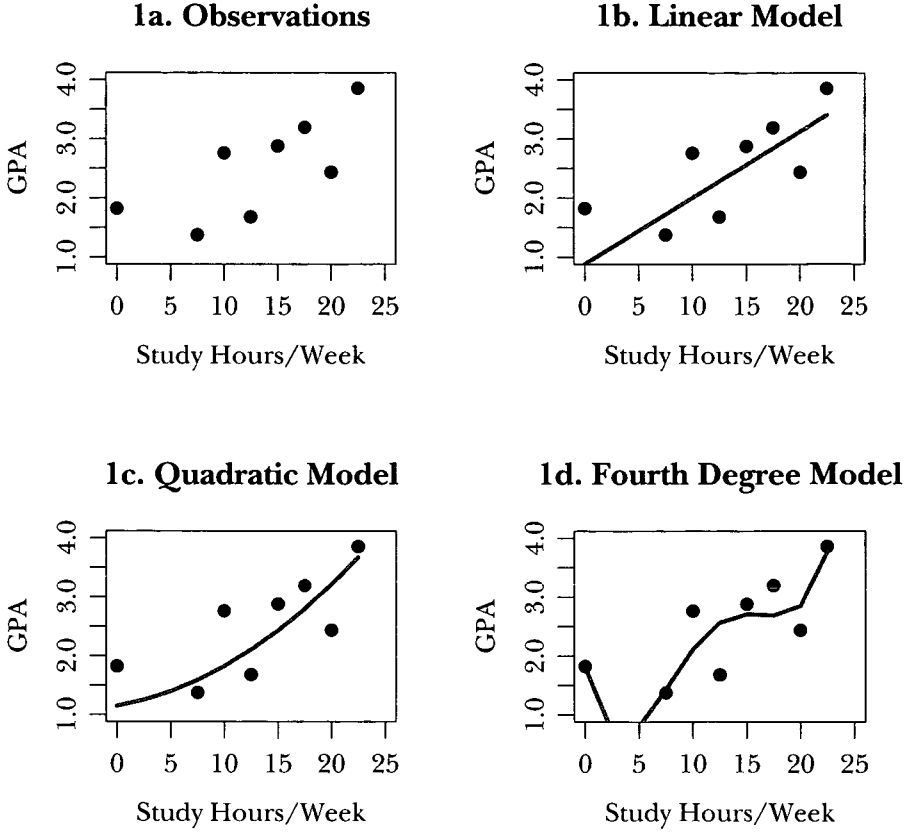

Figure 1: Example Fits to Observed Data Points

At least initially, there appear to be infinitely many models and no obvious principle for choosing one over another.[55] Our intuitions suggest, however, that some of the curves are more plausible than others. For example, the fitted curve in Figure 1d seems excessively complex: Study hours and GPA are unlikely to be related in this way. This intuition may be the basis for the time-honored principle of Occam's Razor, which favors simpler explanations.[56] But how and why is the curve *excessively* or *unnecessarily* complex? Mere intuition falls short. Although intuition may

55. Mathematically speaking, for n data points, an nth order polynomial is all that is required for the curve to pass through all the points. However, one can certainly fit higher order polynomials, which will simply allow for more—for lack of a better term—squiggles between the data points.

56. See, e.g., Lewis S. Feuer, The Principle of Simplicity, 24 Phil. Sci. 109, 109 (1957) (stating Occam's Razor: "Entities are not to be multiplied unnecessarily" (emphasis omitted)). Technically, Occam's Razor excludes needlessly complex models, which means that it may only exclude models that have variables beyond those necessary to pass the fit line through all of the data points. See supra note 55 (discussing fitting a line to pass through all points). Occam's Razor does not necessarily select simpler models on the assumption that some of the variation is due to random error. Nevertheless, the spirit of

suggest which models are plausible or desirable, it is neither precise nor objective. The higher-order curve in Figure 1d may be easily excluded, but intuitively choosing between the linear (Figure 1b) and quadratic (Figure 1c) curves is far more difficult.

The above example only involves one predictor variable, study hours. The selection problem, however, readily generalizes to the multivariate context, in which we have many potential predictors—e.g., hours of sleep, availability of tutoring, socioeconomic background, etc. The question then becomes not only how complex the model should be in terms of polynomial degrees, but also which variables should be included in the model. Conceptualizing the problem, however, only requires the single variable case, so I will consider a single variable in the discussion that follows.

B. *The Problem of Overfitting*

The statistics literature offers some perspective on the model selection problem beyond sheer intuition. Complex models are problematic not just because they violate some vague preference for simplicity, but because they are an example of what statisticians term "overfitting." Given a dataset, one can actually always improve the fit of a model until the fitted curve passes through every point in the dataset, as seen in Figure 2. Fitting a model to the so-called "training data" is therefore trivial.

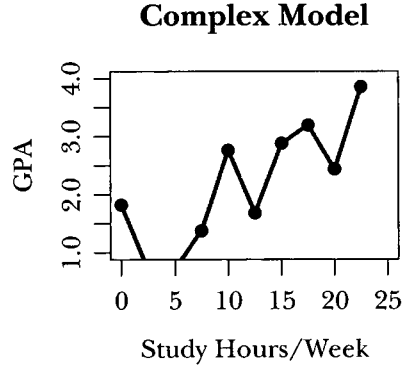

Figure 2: Overfitting Example

The problem is that overfitted models capture not only the relationship of interest, but also the random errors or fluctuations that inevitably accompany real world data. So for example, in Figure 2, rather than modeling the simple linear relationship that gave rise to the data

---

Occam's Razor is toward simpler models, and researchers often invoke it in this broader vein.

points,[57] the complex model erroneously incorporates all of the errors and fluctuations as well. The penalty for this overfitting is lower predictive accuracy. Presented with a new set of students, the excessively complex model will make more errors in predicting GPA than a simpler model that ignores the noise. And arguably, predictive accuracy is the key measure of a model's worth, because if a model is actually a good representation of reality, it should predict well for all datasets, not just the one used to train it.[58] So a tradeoff exists. Too simple a model will fail to identify the underlying relationship and have low predictive accuracy. Too complex a model will incorporate too much random noise into its inferences about future observations and will also be inaccurate. What we need is the optimal balance between fit and complexity.[59]

C. *Model Selection Criteria*

Fortunately, statisticians have been thinking about the model selection problem for quite some time, and they have developed various criteria for comparing and selecting statistical models. Model selection criteria thus provide a principled way of dealing with the overfitting/underfitting problem beyond intuition. They operate as rating systems that score potential models. Conceptually, model selection criteria have two main parts: One part measures how well the model fits the observed data, while the other measures its complexity, reflecting the fit-complexity tradeoff. As previously discussed, as we increase model complexity, we will always improve the model's fit to the observed data. The key question asked by model selection criteria, however, is whether the resulting increase in model complexity is worth the improvement in fit.

For example, one commonly used criterion is Akaike's Information Criterion (AIC).[60] The derivation of AIC is rather technical, and more

---

57. The generating function for the datapoints in Figures 1 and 2 is GPA = 1 + 0.1*HOURS + ε, where $\varepsilon \sim N(0, 0.5)$.

58. In constructing a predictive model, the given dataset is only a sample of the population. Thus, constructing a model that tracks the current data too closely, we are likely to make inferences that are too strong, hampering the model's ability to accommodate future data.

59. This result can also be considered from a slightly different perspective. Given the current data, a model can "attribute" the variation in the response variable to either a (deterministic) predictor variable or an (stochastic) error term. Since the given dataset is only a sample of the population, constructing a model that is too deterministic—i.e., one that tracks the current data too closely—will cause it to lack the flexibility needed to handle future observations. At the same time, constructing a model that is too stochastic—i.e., one that just blames chance for everything—will fail to use all of the available information. The key question is whether the structural information in the data justifies use of a predictor variable over the error term. See Kenneth P. Burnham & David R. Anderson, Model Selection and Multimodel Inference 31–33 (2d ed. 2002) (discussing balance between overfitting and underfitting).

60. The formula for AIC is AIC = $-2l_s + 2p$, where l_s represents the maximum log-likelihood for the model, and p is the number of predictors or parameters in the model. E.g., W.N. Venables & B.D. Ripley, Modern Applied Statistics with S 173–74 (4th ed. 2002).

HeinOnline -- 109 Colum. L. Rev. 2093 2009

comprehensive mathematical and philosophical treatments of AIC are available elsewhere,[61] but two conceptual points will suffice for our purposes here. First, although the fit-complexity tradeoff may initially seem like a rather crude (and subjective) cost-benefit analysis, AIC performs the tradeoff with some mathematical foundation. Specifically, the criterion derives from information theory concepts about how the observed data and the fitted model "match" each other. Second, AIC helps score different models. In the single variable context in Figure 1, AIC can help select the appropriate polynomial degree (i.e., whether the model should be linear, quadratic, or more complex). In multivariate problems, it can help select which predictors to include in the model.

It is important to understand that AIC rests on certain (albeit reasonable) assumptions.[62] Under different assumptions, researchers have developed a variety of other selection criteria, including the Bayesian Information Criterion (BIC)[63] and the Deviance Information Criterion (DIC).[64] Additionally, these criteria are *heuristics* for accuracy, which is the real goal.[65]

The maximum log-likelihood (l_s) term measures how well the model fits the observed data, while the number of predictors (p) measures its complexity.

61. For technical discussions of AIC, see Burnham & Anderson, supra note 59, at 353–71, as well as the original paper, Hirotugu Akaike, A New Look at Statistical Model Identification, 19 IEEE Transactions on Automatic Control 716 (1974). For a terrific discussion of its philosophical implications, see generally Malcolm R. Forster & Elliott Sober, How to Tell When Simpler, More Unified or Less Ad Hoc Theories Will Provide More Accurate Predictions, 45 Brit. J. Phil. Sci. 1 (1994).

62. As Elliott Sober notes, AIC makes three major assumptions. First, it takes Kullback-Leibler distances (or relative entropy) as the measure between two probability distributions. Second, it makes the "Humean 'uniformity of nature' assumption" that the data are drawn from a relatively stable world and that the mechanism that links the predictor and outcome variables stays constant. Third, AIC makes a normality assumption, which is that "repeated estimates of each parameter are normally distributed." Elliott Sober, Instrumentalism, Parsimony, and the Akaike Framework, 69 Phil. Sci. S112, S116 (2002) [hereinafter Sober, Instrumentalism].

63. BIC rests on entirely different theoretical foundations from AIC, but arrives at a strikingly similar tradeoff. BIC is defined as: $BIC = -2l_s + p * \log n$, where l_s is the maximum log-likelihood of the model, p is the number of parameters, and n is the number of observations in the datasets. Venables & Ripley, supra note 60, at 276; see also Forster, Key Concepts, supra note 54, at 220–24 (discussing when different model selection methods are better than others).

64. See David J. Spiegelhalter et al., Bayesian Measures of Model Complexity and Fit, 64 J. Royal Stat. Soc'y: Series B (Stat. Methodology) 583, 602–05 (2002) (proposing the Deviance Information Criterion and a method for assessing models).

65. In theory, one could dispense with the various model selection criteria and estimate the accuracy of each proposed model directly through a technique called cross-validation. Cross-validation roughly proceeds along the following lines: Randomly divide the available data into two (not necessarily equal) parts. Use the first part, known as the "training set," to fit the model. Then use the fitted model to make predictions on the second part (the "testing set") and to determine the resulting error. Perform this procedure repeatedly to obtain an average "cross-validation" error for the model. By comparing the cross-validation errors of one proposed model to another, one can estimate which one is likely to be more accurate. Cross-validation for model selection is a well-

III. A Practical Solution

With the introduction to the reference class and model selection problems out of the way, we can now move to this Essay's two principal claims. First, the reference class problem is just a subspecies of the model selection problem. Second, model selection criteria such as AIC effectively eliminate the reference class problem as it arises in legal contexts.

A. *Reference Class As Model Selection*

By now, the similarities between the reference class problem and the model selection problem should be somewhat apparent. The goal in both contexts is predictive accuracy, and that task requires optimizing the model or reference class so that it neither underfits nor overfits.

Revisiting the *Shonubi* case illustrates this point. The court in *Shonubi* needs to predict the previously smuggled amounts as accurately as possible. It thus must choose a reference class that optimizes the trade-off between fit and complexity. If the court uses characteristics to describe Shonubi that are too broad, such as "all airline passengers," then it will fail to use all of the discriminating information available, resulting in poor accuracy. However, if the court uses characteristics that are too narrow, then it will run the risk of incorporating noise and random coincidences. For example, if Shonubi likes playing basketball and eating spaghetti, using the two other basketball-playing, spaghetti-loving heroin smugglers is suspect. Given three people and an infinite number of personal characteristics, one can always find some characteristics in common. The real question is whether those characteristics have any predictive accuracy going forward.

As it turns out, however, the reference class problem and the model selection problem are not just similar or analogous; they are actually one and the same. Reference-class-style reasoning is equivalent to using a highly simplified form of regression modeling. When one chooses a reference class, one effectively selects a specific regression model, namely a model that has a binary (yes or no) variable indicating membership in the reference class.

To see this more clearly, consider the following example. Assume that we classify Shonubi as a "heroin smuggler in 1998." To estimate the amount of drugs carried by Shonubi on previous trips, we would then simply average the amounts seized from heroin smugglers in 1998. This procedure, however, is precisely equivalent to setting up a simple regression model, $DRUGS = \beta * HEROINSMUGGLER98$, where *DRUGS* is the

established area of study. For more information, see the citations given in Burnham & Anderson, supra note 59, at 36.

The problem with cross-validation is that depending on the size of the dataset and the complexity of the models, it can be quite computationally involved. Id. Thus, heuristics like AIC are often preferred. Fortunately, for large datasets, researchers have shown that the results of AIC closely approximate those of cross-validation. Id. at 62.

amount of heroin seized, *HEROINSMUGGLER98* is a binary variable for whether the person was a heroin smuggler caught in 1998, and β is the effect that being a heroin smuggler caught in 1998 has on the amount of heroin seized.

All of the candidate reference classes in *Shonubi* are thus equivalent to simple regression models. The class "Nigerian drug smugglers" is the same as the model *DRUGS=β*NIGERIAN*; the class "toll booth collectors" is the same as the model *DRUGS=β*TOLLCOLLECTOR*. We have thus transformed the reference class problem in *Shonubi* into a model selection one. So trying to select a reference class is no different that trying to select a regression model.

B. *Model Selection Methods As the Solution*

If the reference class problem is merely an instance of the model selection problem, then model selection methods handle the reference class problem in the legal system, for all practical purposes. Contrary to the existing legal commentary on the issue, principled methods for preferencing some reference classes over others do exist. These methods are none other than the model selection criteria. Thus, choosing a reference class need not be a matter of subjective or intuitive judgment, but can be an objective and quantifiable endeavor.

Let me reemphasize, however—my claim is limited only to the *legal* context and does not extend to the broader philosophical reference class problem per se. No one has yet determined how to do optimal model selection generally, as in finding the single best model for a given phenomenon. That global optimization problem is exceptionally difficult, if not intractable, because the number of potential predictors for any phenomenon is limitless.[66] Indeed, even if we limit ourselves to a finite population of predictors, determining the optimal model can be challenging. Given n potential predictors, the number of candidate models is 2^n, which means that the number of models we need to test grows exponentially. For example, if there were twenty potential predictors included in a dataset, an exhaustive search would involve considering over a million models.[67]

Fortunately, we do not need to find the global optimum to solve the reference class problem in the legal context. Owing to adversarial system values, courts do not determine the truth writ large, but rather only mediate disputes between two parties (or in complex litigation, a large but

66. See David Draper, Assessment and Propagation of Model Uncertainty, 57 J. Royal Stat. Soc'y: Series B (Stat. Methodology) 45, 51 (1995) (discussing how the range of possible models grows at "a rate much faster than that at which information about the relative plausibility of alternative structural choices accumulates").

67. This analysis does not even account for transformations, which again result in an infinite set of possible models. For example, while the dataset may provide height as a potential predictor, we could also potentially use height squared, the square root of height, etc.

HeinOnline -- 109 Colum. L. Rev. 2096 2009

finite number of parties).[68] Courts therefore never need to determine the optimal reference class. They just need to decide which reference class among those presented by the parties is *better*. This assessment is importantly a comparative one, and conveniently, model selection criteria exist for comparing models, and now by extension, reference classes.

IV. Discussion and Limitations

The two major claims presented above have potentially wide-ranging implications for the legal system and beyond. Linking the reference class problem with the model selection problem injects a new body of research into the discussion. For example, the concept of overfitting explains what people intuitively do when they find some reference classes plausible and others not. At the same time, the wide variety of tools used for performing model selection (along with the mathematical theories underlying them) can now be brought to bear on the reference class problem.

In addition, solving the reference class problem in the legal context potentially opens the door to greater acceptance of statistical models. Allen and Pardo's critique of mathematical models of evidence based on the reference class problem was devastating because it suggested that statistical models were of questionable value.[69] Finding a solution eliminates this cloud and arguably reinstates statistical methods as an important alternative to traditional methods of proof.

That said, the solution is not without its limitations, and in this Part I address the likely criticisms and acknowledge some of the solution's limitations.

A. *Available Data*

The most significant limitation to this practical solution to the reference class problem is that it depends heavily on available data.[70] Model selection criteria can mediate among various reference classes, but only if sufficient data exists to make the assessment. For instance, suppose that

68. E.g., Kilcoyne v. Plain Dealer Pub. Co., 678 N.E.2d 581, 586 (Ohio Ct. App. 1996) ("A judicial proceeding resolves a dispute among the parties, but does not establish absolutely the 'truth' for all time and all purposes."); Morrison v. State, 845 S.W.2d 882, 902 n.2 (Tex. Crim. App. 1992) (Benavides, J., dissenting) ("It is widely accepted that the primary goal of adversary process is a fair resolution of disputes between litigants, not the discovery of objective historical fact."); Judicial Panel Discussion on Science and the Law, 25 Conn. L. Rev. 1127, 1132 (1993) (statement of Connecticut Superior Court Judge Martin L. Nigro) ("Don't misconceive the purpose of a trial. . . . The trial is really a dispute settlement. It's got to come to an end.").

69. See supra note 14 and accompanying text (questioning the usefulness of mathematical models due to the reference class problem).

70. Allen & Pardo, supra note 12, at 119 ("[T]here may be no data for other plausible reference classes, which means that the mathematics can be done only by picking these or some variant."); Sober, Instrumentalism, supra note 62, at S118 (noting that the answer of optimality in AIC depends on the data available).

HeinOnline -- 109 Colum. L. Rev. 2097 2009

the defendant in *Shonubi* argued for using tollbooth collectors as the reference class. Unfortunately, no database of day jobs of arrested heroin smugglers exists. As a result, the court has no way of considering the defendant's proposed reference class, even though the proposed class theoretically (however unlikely) might have been a better class than the one ultimately used.

In many ways, the "practical" qualification to the title of this Essay derives from this limitation. Given the available data, model selection criteria readily mediate debates over which reference class is appropriate to use. However, they cannot determine what the optimal reference class would be if we had perfect information and an infinite dataset. Arguably, however, a solution to the reference class problem, particularly in the legal context, should not have to meet such transcendent goals. Any legal setting involves limited statistical information, and the legal system always trudges on with the information that it has.

An opposing party can dream up a new reference class for which no data is available and thus challenge the statistical evidence presented. The legal system, however, may choose to deal with this problem through evidentiary presumptions that place the burden on the challenging party to produce the contrary evidence (or in this case, more data).[71] Since model selection methods have chosen the best reference class given the available data, it is not clear that a speculative suggestion with no supporting data should be heeded.[72]

B. *Selective Data Collection*

Another possible concern raised by the solution's dependence on available data is selective data collection. Due to informational asymmetries, one party may have a far greater ability to collect or access relevant datasets than the other. For example, only the government is positioned to collect and produce data on heroin smuggling. Consequently, a critic might question whether the government could skew results in its favor through its choice of data collection methods.

Information asymmetry, however, should have little if any effect on the reference class selected. Ex ante, there is no way for the party with access to the information to skew the dataset in its favor. For example, the Customs Service in *Shonubi* chose to record the airport of entry in its

---

71. Cf., e.g., Floorgraphics, Inc. v. News Am. Mktg. In-Store Servs., Inc., 546 F. Supp. 2d 155, 172 (D.N.J. 2008) ("When challenging the admissibility of . . . expert testimony, a party must move beyond empty criticisms and demonstrate that a proposed alternative approach would yield different results.").

72. See Nance, supra note 12, at 268 ("[I]t is plausible that, in the absence of suggestions by the accused, jurors ought to accept the figures provided by the prosecution's witness."); see also Jonathan J. Koehler, Why DNA Likelihood Ratios Should Account for Error (Even When a National Research Council Report Says They Should Not), 37 Jurimetrics J. 425, 431–33 (1997) (arguing broader DNA population statistics should be used if local ones are unavailable).

HeinOnline -- 109 Colum. L. Rev. 2098 2009

dataset. In some cases, that attribute will result in a reference class with higher predicted quantities of heroin smuggled, whereas in others it will result in lower. Ex ante, however, the government has no idea whether the charged smuggler will come from the higher or lower quantity airports, so the decision to collect data on airports of entry cannot be driven by outcomes. Ex post, there are few opportunities for government manipulation. Absent outright suppression or spoliation, the dataset will be available through discovery, and the model selection methods will have been well established and fixed. Thus, the government will be unable to pick and choose when to use the airport of entry predictor. If the predictor benefits the government's case, then the government will advocate for its use; if it benefits the defense, then the defendant will do the same. And irrespective of advocate, model selection methods will determine the predictor's appropriateness.

Sometimes, of course, new data can be collected ex post, which obviously advantages those actors with greater resources. This imbalance, however, is arguably no different that the usual resource disparity issues that plague all litigation, and in any event, the data collector still must show that the new model (that incorporates the ex post data) has a superior model selection score.

C. *Alternative Model Selection Criteria*

Another possible objection goes back to the fact that model selection criteria are non-unique and heuristic. As such, they are imperfect—while model selection criteria help determine which model or reference class is *likely* to be optimal, they are not always right, and they do not always agree.

The lack of uniqueness among the model selection criteria should not be cause for concern. First, the criteria will often either concur or select negligibly different reference classes. In these cases, the presence of multiple criteria is irrelevant. Second, to handle conflict cases, the legal system could arbitrarily preselect a reasonable model selection criterion as the governing method for reference class selection. Establishing such a rule is perfectly within reason, particularly because none of the criteria is consistently better than the others, and uniformity and predictability militate in favor of a universal method.[73]

One potential additional complaint is that the use of model selection criteria merely shifts the problem down one level.[74] Rather than dealing with the problem of conflicting reference classes, the proposed solution merely transforms the problem into a conflict between model selection

73. Alternatively, methods exist for combining multiple models. See Zucchini, supra note 53, at 60 (suggesting combination of multiple models); see also Nance, supra note 12, at 263 (discussing use of multiple reference classes); Robert Tibshirani, Regression Shrinkage and Selection via the Lasso, 58 J. Royal Stat. Soc'y: Series B (Stat. Methodology) 267, 267–68 (1996) (showing method for combining multiple models).

74. Thanks to Jeff Lipshaw for prompting this discussion.

HeinOnline -- 109 Colum. L. Rev. 2099 2009

criteria, and then arbitrarily chooses one. Arguably, however, the elegance of the solution lies in this very shift. Directly mediating between reference classes is nearly impossible. Given the diversity of factual issues handled by the legal system, we could never hope to have the foresight necessary to specify reference classes ahead of time. Even if we could, because reference classes directly affect legal outcomes, the ex ante choice of reference classes would be substantively charged and value laden. After all, the decision never to account for a home's number of bedrooms in valuating property would surely provoke the ire of those with six-bedroom homes. By appealing toward a meta-rule—in this case, a model selection criterion—we can preselect a reasonable and neutral rule for mediating these disputes. Ex ante, no one knows whose ox will be gored by the preselection of AIC over BIC, since they differ only through rather abstract mathematical assumptions, and once settled, the rule operates mechanically without any taint of unfairness or favoritism.

D. *Extrapolation and Stability*

Another limitation is that the proposed solution only solves the *reference class* problem. That is, when an individual is a member of multiple groups, model selection methods can help determine which reference class is more useful for making statistical inferences. The solution does not, however, address problems involving extrapolation, in which an individual does not belong to the group, but we *reason* that statistics about the group might be helpful in assessing the individual anyway. For example, assume that in the *Shonubi* case, no data was available for heroin traffickers at JFK. Instead, the only available data was for Los Angeles International (LAX) and Miami International (MIA) airports. The proposed solution does not help determine which dataset (or both in combination) is more appropriate. Relatedly, the proposed solution does not help assess whether we should extrapolate at all. If in addition to the LAX and MIA data, we also had national data, the proposed solution does not help choose between whether we should use a narrower set of data that requires extrapolation (for example, LAX only) or whether we should use a broader dataset that encompasses the individual (the national data).

Model selection implicitly assumes that the phenomenon in question is relatively stable over time.[75] The reason that selection criteria and cross-validation provide useful estimates of a model's predictive accuracy is that we assume relationships to be relatively stable. For example, we assume that if house prices have historically been positively correlated to square footage, then that relationship continues today. Strictly speaking, that assumption is not always correct. A wave of environmentalism could suddenly sweep a region and make large houses highly unfashionable, much like what happened to large SUVs in the face of high oil prices. In

75. Elliott Sober calls this the "Humean 'uniformity of nature' assumption." See Sober, Instrumentalism, supra note 62.

HeinOnline -- 109 Colum. L. Rev. 2100 2009

doing model selection, we assume this will not occur. Arguably, however, the assumption is a reasonable one. Most reference class problems in the legal system are repeated inquiries into past, stable phenomena. Housing prices, drug smuggling quantities, and cancer risks all fall under this category.

E. *Accuracy as the Goal*

Finally, it is worth noting that the model selection solution assumes that accuracy is the overwhelming goal. After all, the criteria are heuristics that choose models for their predictive accuracy. Nevertheless, one can imagine other objectives that would argue in favor of other reference classes. For example, one of the primary aims of class action litigation is administrative efficiency,[76] so in that context, a broader reference class may be chosen to reduce the number of trials and thus overall litigation costs, even though such efficiencies may come at the expense of accuracy in individual cases.[77] Even in these cases though, the accuracy-oriented model selection methods continue to serve a useful purpose, as they provide a baseline from which other value judgments can depart.

V. EXAMPLES

The discussion thus far has largely been conceptual. In this section, I offer two examples beyond *Shonubi* that hopefully demonstrate how model selection techniques solve the reference class problem in practice.

A. *Fake Barn Town*

In their *Journal of Legal Studies* article, Ron Allen and Mike Pardo use Alvin Goodman's Fake Barn Town to illustrate the reference class problem.[78] Suppose that an agent drives around town to identify barns. In this bizarre hypothetical, however, the town contains both real barns and "barn facades, which, although they look like barns from the front, are just fake barn fronts and not real barns."[79] The agent is unable to distinguish real from fake barns, so his accuracy rate depends on the background ratio of real to fake barns.

Assume arguendo that the agent has identified a given barn as real. Allen and Pardo suggest that the agent's likelihood of being correct is then an instance of the reference class problem. If the agent is in Fake Barn Town, where most barns are fake, the agent will be unreliable. But

76. E.g., Fed. R. Civ. P. 23(b)(3) (allowing class action when it "is superior to other available methods for fairly and efficiently adjudicating the controversy").

77. See generally David Rosenberg, A New Sampling Method for Reducing the Cost of Resolving Differing Claims Against a Defendant 2–3 (2008) (unpublished manuscript, on file with the *Columbia Law Review*) (proposing use of sampling methods with "no structuring or pre-screening of the group of claims" to promote greater administrative efficiency).

78. Allen & Pardo, supra note 12, at 111–13.

79. Id. at 111.

HeinOnline -- 109 Colum. L. Rev. 2101 2009

if the agent is "on Real Barn Street (in Fake Barn Town), in which all the barns are real . . . [then h]e is a reliable reporter on that street."[80] Simultaneously, the agent might also be in Real Barn County and Fake Barn State, which make him reliable and unreliable respectively. As Allen and Pardo ask:

> [S]uppose an empirical test were being run as to the ability of our agent (a witness at trial, for example) to identify barns accurately. What is the "proper" baseline (base rate) for running such a test? Is it the proportion of true barns on Real Barn Street, Fake Barn Town, Barn County, or Fake Barn State (or maybe the United Barn States of America)? There is no a priori correct answer; it depends on the interests at stake.[81]

In a sense, Allen and Pardo are correct, but I would argue that their position is only a function of their question's ambiguity. As a context-independent question—what is the agent's accuracy rate generally—perhaps no answer exists. But legal proceedings are emphatically not context-free. At trial, we care about the reliability of the agent to identify barns *in this case*, and the case tells us the street, town, county, and state in which the particular barn is located.

Once we have a specific context, then choosing the reference class proceeds easily using model selection criteria. The issue becomes what level of detail (street, town, county, state) is appropriate, and the selection criteria negotiate that question readily. In an extreme case, like that proposed by Allen and Pardo where the accuracy changes with each subsequent reference class, we will choose the narrowest class possible (without reducing the class to a single person), because each additional individualizing piece of information is highly probative and useful in determining the accuracy rate of the agent.

B. *Estimating the Value of a Destroyed House*

To further illustrate the use of model selection techniques in property valuation, imagine the following (simplified) litigation hypothetical. A jury has found the defendant negligent in setting fire to the plaintiff's house in Newton, Massachusetts in 1998. The question now is damages—what was the value of the plaintiff's house when it was destroyed? The town has 114 home sales on record from 1983 onward.[82] The dataset contains the price, year of sale, year of construction, total rooms, bedrooms, bathrooms, and total square footage.

The plaintiff's house was built in 1925, had three bedrooms and one bath, and had six total rooms totaling 1,200 square feet. Immediately, we

---

80. Id. at 112.

81. Id. at 113.

82. The dataset used for this hypothetical was from Newton, Massachusetts, and was generously made available on the MIT OpenCourseWare website. MIT OpenCourseWare, Hedonic.dta, at http://ocw.mit.edu/OcwWeb/Economics/14-33Fall-2004/Labs/ (last visited Sept. 13, 2009) (on file with the *Columbia Law Review*).

HeinOnline -- 109 Colum. L. Rev. 2102 2009

encounter reference class problems. Certainly, taking the mean home price ($341,139) in the dataset seems inappropriate, since it mixes houses of varying size and age, and it fails to account for changes in real estate prices over time. At the same time, individualizing the inquiry to the plaintiff's unique house is no help either, since its price is precisely what we seek. In between these extremes, the relevant characteristics are unclear. Should the house be compared only to houses with three bedrooms and one bath? What about the square footage and the age?

Again, because the context is litigation, we need not determine the absolute best reference class. The parties will raise differing reference classes, and our job is to determine which one is *better*. The defendant argues that the key attribute is that the house has three bedrooms and one bath. This reference class will lead to an estimate of $294,392. In contrast, the plaintiff argues that the older houses in the area are better built and have more charm, and that a second bath is not as important in a three-bedroom house. The plaintiff asserts that the proper reference class is houses with three bedrooms built before 1960. This reference class generates an estimate of $304,642.

So whose reference class is preferable? We can of course argue forever about whether the number of baths is important to the price and whether older houses sell better. Model selection criteria, however, offer a way out. Assuming that we choose AIC as our criterion for model selection, the answer is straightforward. The plaintiff's reference class yields an AIC score of 3093.5, whereas the defendant's reference class yields a score of 3094.5.[83] Lower AIC scores are more desirable, so the plaintiff's reference class is superior, and the court should side with the plaintiff's estimate of $304,642.

* * *

As a final observation, it is worthwhile to reiterate the earlier observation that reference-class-style inference is a highly cramped form of regression modeling. The use of reference classes, while unquestionably simple and accessible, is actually quite crude. Rather than being limited to a single binary variable (i.e., whether you are a member of the given reference class or not), regression models ordinarily are far more flexible and can account for more subtle effects. For example, rather than looking only at houses with three bedrooms, the model could incorporate houses of all sizes by raising the estimated price by $25,000 for each additional bedroom. This kind of flexibility enables regression models to be considerably more accurate.

For instance, in the destroyed house example, we can develop a regression model that uses square footage, bedrooms, bathrooms, and year

83. Incidentally, using the mean for the entire dataset, which would have yielded an estimate of $341,139, generates an AIC score of 3096.37, which is less optimal than either model proposed by the parties.

of sale with a dramatically improved AIC of 2675.61. (As it turns out, adding other variables, such as year of construction, turn out to be unnecessary and would contribute to overfitting.) Plugging in the attributes of our house—1,200 square feet, three bedrooms, one bath, destroyed in 1998—yields an estimated price of $238,017. Ideally, a court should use this estimate instead.

Nevertheless, this quibble is quite beside the principal point, which is that model selection criteria solve the reference class problem. Whether we like it or not, reference-class-style reasoning is both a common and straightforward method for making inferences, and thus it is unlikely to disappear any time soon. Perhaps courts should really prefer regression analyses to reference classes, but that is another fight altogether.[84] And even if courts were to adopt regression analyses as their method of choice, the selection criteria would still remain relevant as a method of choosing among models.

VI. CONCLUSION

Peter Tillers once suggested two possible criteria for a solution to the reference class problem: A strong solution would "generate an autonomous procedure for wrestling probabilities about individual events out of the appropriate reference class or combination of reference classes," whereas a weaker one would "merely tell us how human beings (often) make probabilistic sense out of experience—without providing a replica of or substitute for human judgment."[85] The proposed solution in this Essay appears to satisfy both. By drawing a tight analogy to model selection, we see that model selection criteria solve the reference class problem for all practical purposes in the legal system. Although they cannot guarantee (nor do they help us find) the absolute "best" reference class, these ranking systems autonomously and powerfully mediate among the classes raised by the parties, which is more than sufficient for legal purposes. The proposed solution also represents a starting point for how human beings judge the appropriateness of various reference classes. The concern about overfitting and the fit-complexity tradeoff are certainly plausible explanations, although that hypothesis obviously requires empirical study.

Although the reference class problem and its solution may appear academic at first blush, the wide-ranging practical ramifications of the solution should not be overlooked. With the growth of computing power and database storage, the availability and use of statistical data and evidence will only grow in the legal system. Concomitantly, disputes about which statistics to use will also increase. Ultimately, many of those disputes will boil down to debates about reference class, including the often

84. Whether the estimation method involves reference classes or regression analyses, model selection methods can be helpful for improving predictive accuracy.

85. Tillers, supra note 4, at 38 n.21.

HeinOnline -- 109 Colum. L. Rev. 2104 2009

heard claim that a litigant's case is "unique." With model selection methods in hand, courts now have a powerful method for assessing and deciding those disputes.

HeinOnline -- 109 Colum. L. Rev. 2105 2009

HeinOnline -- 109 Colum. L. Rev. 2106 2009

COLUMBIA LAW REVIEW

VOL. CIX—2009

NEW YORK

2009

HeinOnline -- 109 Colum. L. Rev. i 2009