COLLECTIVE VIOLENCE AND INTERNATIONAL CRIMINAL JUSTICE

An interdisciplinary approach

Edited by

Alette Smeulers

intersentia
Antwerp – Oxford – Portland

CHAPTER 14

LEARNING THE HARD WAY AT THE ICTY: STATISTICAL EVIDENCE OF HUMAN RIGHTS VIOLATIONS IN AN ADVERSARIAL INFORMATION ENVIRONMENT*

Amelia Hoover Green

1. INTRODUCTION

"[T]he Chamber is of the view that such doubt has been cast upon the [prosecution] study's conclusions that reliance upon them would not be appropriate."
– Judgment, *Milutinovic et al.*[1]

On 26 February 2009, judges at the International Criminal Tribunal for the Former Yugoslavia (ICTY) delivered their Judgment in case IT-05–87, Prosecutor v. Milan Milutinovic, Nikola Sainovic, Dagoljub Ojdanic, Nebojsa Pavkovic, Vladimir Lazarevic, and Sreten Lukic (hereafter referred to as *Milutinovic et al.*). The five charges against all five defendants were as follows: deportation, a crime against humanity (1); forcible transfer or "other inhumane acts," a crime against humanity (2); murder, a crime against humanity and a violation of the laws and customs of war (3, 4); and persecution, a crime against humanity (5). Milutinovic

* This research was supported by a 2007–2008 Benetech Human Rights Program Research Fellowship, 2007 and 2008 grants from the Whitney and Betty MacMillan Center for International and Area Studies at Yale, and a 2008–2009 United States Institute of Peace Jennings Randolph Peace Scholar Dissertation Fellowship. Many thanks to the colleagues and advisors who have helped formulate and refine these impressions, particularly Patrick Ball, Elisabeth Wood, Laurel Eckhouse, and Meghan Lynch, as well as the organizers and participants of the 2009 Expert-Meeting "Collective Violence and International Criminal Justice – An Interdisciplinary Approach," VU Amsterdam and the Amsterdam Centre for Interdisciplinary Research on International Crimes and Security.

1 *Prosecutor v. Milan Milutinovic et al.* Case No. IT-05–87, T.Ch. III, 26 February 2009 (hereafter referred to as *Milutinovic et al.).*

himself was found not guilty of all five charges, while two of his co-defendants were found guilty of all charges and two others were found guilty of only two charges.

The four-volume, 1,600-page Judgment (throughout this document, "Judgment" refers to the *Milutinovic, et al.* Judgment) delivered that day included just nine paragraphs related to statistical evidence prepared by the prosecution's expert witness, Patrick Ball, and colleagues at the Human Rights Data Analysis Group (HRDAG).[2] On nearly every substantive point, the Chamber found the statistical evidence unconvincing: the underlying data were incomplete and biased[3]; the correlation between refugee flows and killing had not been proven[4]; the Prosecution had not adequately dealt with alternative hypotheses.[5] For better or worse, these findings were not exceptionally consequential; the Chamber reached guilty verdicts in fourteen of twenty-five charges against the five defendants. However, the court's findings regarding the Prosecution's statistical evidence called into question whether numerical accountings of human rights violations, necessarily plagued by incomplete data, can play a useful role in bringing perpetrators of human rights violations to justice.

The promise of statistical evidence is clear: statistics, as a discipline, provides specific, well-established rules for testing hypotheses about the existence or non-existence of patterns and associations. Large-scale quantitative data can provide a "big picture" of alleged criminal activity that may be unavailable elsewhere. In theory, where patterns and associations cannot be shown using documentary or testimonial evidence, statistical clues could provide important evidence. In reality, however, a number of differences between statistical and scientific reasoning, on the one hand, and legal reasoning, on the other, decrease the effectiveness of human rights statistics in legal settings.

This paper analyzes the HRDAG team's work on Kosovo (including both work on *Milutinovic et al.* and work on Prosecutor vs. Slobodan Milosevic, case IT-02–54, hereafter *Milosevic*)[6] between its initial submission in 2002 and the *Milutinovic et al.* verdict in 2009. It asks whether and how the analysis, or the presentation of the analysis to the Chamber, could have been more successful. More generally, it identifies key similarities and differences between statistical reasoning and legal argumentation, and suggests methods for overcoming these differences, at least insofar as the statistician's methodological and presentational choices are concerned.

---

2 Ball et al. 2002a and 2002b and 2007.
3 *Milutinovic et al.* 27168: para. 26.
4 *Milutinovic et al.* 27168: para. 27.
5 *Milutinovic et al.* 27168: para. 28.
6 *Prosecutor v. Slobodan Milosevic*, case IT-02–54, T. Ch. III (hereafter *Milosevic).*

It is important to note at the outset that this piece is written from the perspective of the statistical analyst, rather than from a legal perspective. While statisticians may find fault with the Chamber's statistical reasoning, the strategies discussed below concern statisticians' choices, rather than jurists' opinions. (While it might be desirable, either intellectually or practically, for judges to alter their mode of assessment, those presenting the evidence bear responsibility for teaching judges to assess the evidence accurately.) The paper proceeds roughly chronologically, outlining Ball et al.'s analytical process and conclusions, the Defense expert's response, and the Chamber's eventual Judgment before moving on to discussions of the basis for decision and its implications for the analytical strategies of human rights statisticians. Throughout, I attempt to minimize technical and mathematical language, referencing relevant resources but focusing the discussion on major analytical and presentational choices rather than technical specifics.

2. THE PROSECUTION'S ANALYSIS

In both *Milosevic* and *Milutinovic et al.*, the empirical issue before the Chamber concerned the defendants' responsibility for large-scale killings and migration of Kosovo Albanians (generally into Albania) during March-June, 1999. Essentially, the question was: did Kosovo Albanians flee as a result of an ethnically targeted policy of killing and forced migration, or for some other reason? If the killings and migration proceeded as a result of policy, at what level was that policy decision taken? Did the defendants order (or know of) the policy?

Given enough data, statistical analysis could determine whether higher levels of violence were associated at a substantively and statistically significant level (i.e., an association of large magnitude, highly unlikely to have occurred by chance) with any of the parties to the conflict. In particular, the Prosecutor hypothesized that troops of the Federal Republic of Yugoslavia (FRY) pursued a coordinated policy of ethnic cleansing, producing deaths and migration independent of combat operations, NATO bombing or KLA actions.

2.1. DESCRIPTIVE ANALYSIS

In order to answer questions of responsibility, Prosecution statisticians first required accurate descriptive statistics regarding exactly what occurred and an estimate of the number of persons killed in, and departing from, each Kosovo municipality during each two-day (or similarly short) period during March-June 1999. On each day (or for each two-day period), how many persons (and in

particular, how many Kosovo Albanians) were killed in each municipality, and how many left as refugees?[7] In addition to information regarding killings and migration, the authors required data on the presence and activity of the several parties to the conflict in each municipality over each two-day period.

Unfortunately, human rights statistics are generally plagued by a number of problems, including non-random under-registration, manipulation for political purposes, and lack of baseline statistics about either violent events or other population changes. (Armed conflicts tend to happen in failed states; failed states tend to lack adequate registries of vital statistics.) The conflict in Kosovo was no exception.

In this case, neither surveys of displaced populations nor border crossing records could hope to record every refugee leaving Kosovo. However, UN and other estimates of refugees exiting Kosovo closely coincide with Albanian border crossing records for all but extremely high-volume days.[8] On days when more than 20,000 individuals crossed the Albanian border, record-keeping systems could not keep up with the flow and many (in some cases, most) refugees remained undocumented. Consequently, the Prosecution analysts supplemented border-crossing records with data from the United Nations High Commission for Refugees (UNHCR) and the Albanian government's Emergency Management Group (EMG). In all, approximately 400,000 persons were estimated to have fled Kosovo during March through late May, 1999.[9]

Because data on migrants' municipality of origin, and the date on which they left home (as opposed to the date on which they arrived at the border) was incomplete, Ball used data from refugee camp surveys and other sources to estimate migrants' point of origin, travel time and (by implication) the two-day period during which they left their point of origin.[10] The majority of migrants throughout the period reported that they left their municipality on the same day that they exited the country; however, a relatively significant minority reported longer travel times, in a few cases over 70 days. Ball fit reported travel times to exponential distributions for each of three phases of migration,[11] showing that travel time could be imputed with some accuracy. Ball generated timelines showing the number of estimated departures in each two-day period between March and June, for Kosovo as a whole and for separate regions, which showed

7 No records of internally displaced persons were kept; the analysis refers only to those Kosovo Albanians who crossed borders following their displacement and are properly termed "refugees."

8 Ball 2000: 39, graph A1.

9 Ball 2000.

10 Ball 2000: 37.

11 Ball 2000: 42–43.

that departures largely followed patterns similar to those of exits. In Ball's interpretation, these graphs showed three regionally specific "waves" or periods of departure. This observation is consistent with the hypothesis that departures were forced and operated according to a high-level (as opposed to merely local) plan or policy.

As in many conflicts, data on deaths in Kosovo were considerably less complete than data on migration. Consequently, accurate estimates of the pattern and magnitude of killings required very significant analysis of (and inference from) documented casualty lists. Lists of casualties are never entirely complete, and are often drastically incomplete. In addition, each victim's probability of appearing on any casualty list is affected by the crime's accessibility to reporting organizations, as well as the victim's social status and networks (among other factors).[12] Consequently, analysts must never assume that lists compiled from NGO's government sources, displaced persons camp registries, or other sources accurately represent the true pattern of killings or displacement. Not even a "master list" composed of all unique entries on all lists will reliably estimate true patterns of killing or displacement, because "visible" victims may be reported multiple times, while "invisible" victims are never reported at all.

In order to estimate how many people were killed during each two-day period in each municipality, Ball et al. used a standard demographic technique: multiple systems estimation (MSE).[13] In essence, MSE is a statistical representation of the process of selection into three (or, in some cases, more) individual lists[14]; this model is fit using information about the number of cases that were selected into only one list, exactly two lists, or all three lists. Information about these overlaps can be used to estimate the likely number of "invisible victims" – those who were not recorded on any list. However, selection into lists does not occur randomly. For example, "visible victims" – public figures and city-dwellers, for example – are much more likely to be listed in all lists, while some people are very likely to be listed by one organization and very unlikely to be listed by others due to their location or social networks. Hence, the model must account for non-independence of lists. Using MSE analysis of four separate lists of deaths, Ball et al. estimated approximately 9,000–12,000 killings during the March-June

12 Guzmán et al. 2007 and Lynch and Hoover 2008.

13 Chandra Sekar and Deming 1949; Bishop, Fienberg and Holland 1975; Darroch et al. 1993; Fienberg et al. 1999, and, in the human rights context, Ball 2003.

14 The estimated number of total cases, both counted cases and uncounted cases, is the dependent variable in a general linear model with a Poisson (event count) link function (i.e., a Poisson regression). The number of cases not counted in any list, then, is the exponentiated intercept term. For technical discussion of the technique (also termed "multiple recapture"), see Bishop, Fienberg and Holland 1975, Chapter 6; Fienberg et al 1999; Darroch et al. 1993. For other applications in human rights and social science, see Guzmán et al. 2007; Silva and Ball 2006 and Lynch and Hoover 2008.

period,[15] an estimate that was consistent with other, independent estimates of killings during this period.[16]

2.2. CAUSAL ANALYSIS

In contrast to their detailed discussion of estimation procedures for *descriptive* statistics, such as the gross magnitude and individual two-day magnitudes of killings and migration, Ball et al. produced a relatively short analysis of the *causal* question at hand: what group, if any, was most associated with changes in the levels of killing and migration? In other words, which group was most likely to have been responsible for the patterns of violence observed?

A standard statistical tool for testing the strength of associations between (hypothesized) causal variables and outcome variables is linear regression. Under certain assumptions about the quality and distribution of data, linear regression analysis accurately tests the relative strength of hypothesized causal variables. For example, had adequately unbiased data regarding migration, deaths, or troop movements existed, Ball et al. might have chosen an extensive linear regression analysis to test their causal hypotheses. This method has the advantage of familiarity and has a simple, intuitive explanation. It also measures competing hypotheses more or less directly against one another. If, for example, a linear regression model found that KLA presence was a statistically significant variable with a large effect size on violent outcomes, while FRY presence did not prove significant to violent outcomes, then the analyst can (again, assuming the accuracy of the data) rule out the hypothesis that FRY presence was more associated with violent outcomes than KLA presence.

However, Ball et al. performed only an exploratory regression analysis, largely because of missing data issues. The regression analysis tested the relationship between violent outcomes and NATO or KLA activity. However, because it was performed without data on FRY activities, its conclusions are potentially dubious, as Ball acknowledged to the Chamber. The finding, however, supported the Prosecution's contention that KLA and NATO activity were not consistently associated with violent outcomes: estimated coefficients for KLA and NATO activity were neither statistically nor substantively significant. Ball et al. presented this finding graphically, by plotting the observed trend in killings and migration over time alongside the "residuals," the number of killings or migrants left unexplained by the linear regression model.[17] Had KLA or NATO activity

15 Ball et al. 2002.
16 Cf. Spiegel and Salama 2000.
17 Ball et al. 2002, appendix 2.

explained levels of violence with accuracy, the residual line would have hovered around zero over time (i.e., the model would have fit the observed data much more closely). Instead, the residuals closely tracked the observed patterns of violence, indicating that little of the violence was explained by a regression analysis including only KLA and NATO activity.

Ball et al. then turned to an analytical strategy they termed "peaks versus presence." Proceeding from the logical assumption that an armed group's activity must *precede* or *coincide with* events such as killing and migration in order to have caused them, the HRDAG authors tested the temporal relationship between day of peak violence and NATO or KLA presence in each municipality. The authors recorded the "peak" two-day period for killings and migration in each municipality, then analyzed whether KLA or NATO activity preceded or coincided with that peak, followed it, or whether the evidence was inconclusive (i.e., whether KLA or NATO activity had preceded peak killing or migration, but by a significant period of time).

Again, because reliable evidence of FRY troop movements and conflict events was unavailable, Ball et al. chose to make their argument by ruling out competing hypotheses. If (1) KLA and NATO presence did not reliably coincide with peaks, (2) causal hypotheses other than KLA, NATO and FRY presence seemed implausible, and (3) killings and migration proceeded in a patterned way across time and space, this evidence would be consistent with planned FRY responsibility for killings and migration.[18]

Ball et al. found that KLA presence closely preceded or coincided with peaks of killing in 11 of 29 municipalities, followed the peak in 12 others, and was inconclusive in 6.[19] Likewise, KLA attacks preceded or coincided with peak refugee flow in 10 municipalities, followed peak flow in 11, and were inconclusive in 8. NATO bombings preceded or coincided with peak killings in just 3 of 29 municipalities,[20] followed the peak in 20 and were inconclusive in 6. NATO

18 Regarding condition (3) above, the Prosecution hypothesized that killings and migration followed similar patterns over time and space – an observation that, if borne out, would be consistent with the causal hypothesis that killings and migrations were part of a plan or policy. In retrospect, it is unclear whether matching spatiotemporal patterns of killings and migration are a necessary piece of evidence to support the "plan or policy" causal hypothesis; certainly, spatiotemporal coincidence is not *sufficient* evidence to prove a plan or policy. This is so because, while suggestive of a broader plan, killings and migrations could co-occur for a number of other reasons. Ball et al. chose a graphical representation of the similarities between patterns of killings and migration over time. While the relation is visually obvious, the lack of more formal testing in this section of the authors' report would prove consequential at trial.

19 Ball et al. 2002: 11, fig. 8.

20 Ball et al. 2002: 12, fig. 9.

activity preceded or coincided with peak refugee flow in 9 municipalities, followed peak refugee flow in 13 others, and was inconclusive in 7. Neither KLA nor NATO actions, then, were consistently associated with peaks in either killings or refugee flow.

Ball et al., and the Prosecutor argued that this analysis effectively ruled out KLA and NATO actions as causes of the patterns of killings and migrations observed. In particular, while the "peaks versus presence" approach suffers from flaws such as selection bias (it does not consider low-violence outcomes) and an over-reliance on peak periods of violence to the exclusion of other high-violence periods, these flaws tend to overestimate, not underestimate, the effect of KLA/NATO presence on violent outcomes. In testimony on 20–21 February 2007, Ball attempted to explain this phenomenon:[21]

> JUDGE: *"When I read this report, I – where does it say that this is unsatisfactory? I've read this thinking this is a genuine update of your conclusions, but I'm now getting from you that this is an exercise that you're not happy with. Now, where do I see in this report that you're unhappy about this?"*
>
> WITNESS: *"I think that if we read the section on page 5, which I'm looking for now. I will point it to you directly. When we stress over and over again the excessively conservative nature of this report, that is what we are saying. We are saying that we believe that this overstates – that this analysis overstates any potential relationship between NATO and KLA activity and potential killing and migration."*
>
> JUDGE: *"Sorry. And I really hadn't got from this the message that this was unsatisfactory, but thankfully it's been clarified in the oral evidence."*

Throughout Ball's testimony, questions from the Chamber frequently indicated confusion or skepticism. In the exchange above, for example, Ball's "dissatisfaction" stems from the fact that the test he has chosen privileges the Defense, in that it overstates the significance of KLA/NATO activity to violent outcomes. The judge, however, interprets this statement as an admission that the analysis, as a whole, is "unsatisfactory".[22]

---

21 *Milutinovic et al.* 10253, paras 2–15.

22 Regarding the Chamber's reaction to statistical evidence, other analysts have questioned why other significant statistical evidence, including two datasets upon which the Ball et al. reports were based, were ruled out as "hearsay," given that Ball's work was admitted as evidence.

3. DEFENSE CHALLENGES

In cross-examination, attorneys for the defendants focused on Ball's association with human rights organizations and his statements to various organizations regarding the situation in Kosovo.[23] However, the Defense's primary critique of prosecution statistical evidence came from its expert witness, Eric Fruits.[24] Fruits, by training a business economist, had served as an expert witness in a number of trials, but had never taken or taught a course on mathematical demography.[25] Importantly, Fruits did not contribute any independent analysis; as the Defense expert, his role did not require that he prove the inaccuracy of the Ball et al. findings, and Fruits produced no statistical evidence of his own, instead relying on criticisms of the Prosecution experts' methods.

This section critically summarizes the Defense expert's critique of the Ball et al. reports[26], and in doing so describes the information environment at the ICTY. By "information environment," I mean the amount and quality of information presented to the Chamber, as well as its emotional or rhetorical content. The term "information environment" is one most commonly used in research on marketing and consumer decision-making,[27] but it applies equally well to other decision-making contexts. Literature from social, cognitive and organizational psychology has closely investigated several aspects of the information environment, including the effects of information overload,[28] the effects of high-stakes versus low-stakes decision-making,[29] the effects of familiar versus unfamiliar facts and opinions on decision-making style and capacity,[30] and the effects of polarized information on decision-making.[31] In assessing the Defense criticisms, I attend particularly to aspects of the criticisms that affect the Chamber's mode of decision-making.[32]

Fruits' Report claims, first and most generally, that the statistical work undertaken by Ball et al. was not consistent with standard practice in the discipline of statistics. In subsequent sections of his Report, Fruits criticizes the

---

23 *Milutinovic et al.* 10259–10278.
24 Fruits 2007.
25 Fruits had previously *employed* demographic statistics (e.g., statistics regarding population parameters such as those produced by the United States Bureau of the Census) in an econometric analysis. He had no specialized training in the estimation procedures that *produce* vital statistics, which is the province of mathematical demography.
26 Ball et al. 2002 and 2007.
27 Jacoby 1984 and Lurie 2004.
28 Lau and Redlawsk 2001.
29 Allison 1969 and Staw 1981.
30 Zaller & Feldman 1992 and Gross 1991.
31 Simon 2004.
32 See Hoover and Ball 2007 for a more substantive assessment of Fruits' critique.

Prosecution experts' descriptive findings on migration and mortality, the underlying data on KLA and NATO activity, the Prosecution experts' finding of similar patterns of migration and mortality, the "peaks versus presence" causal analysis, Ball's use of linear regression techniques,[33] and the authors' analysis of linear regression residuals. Importantly – but not unusually for legal argumentation – the tone and framing of the Defense expert's Report suggests that each criticism of the Prosecution's statistical analysis is equally, and extremely, important.

3.1. DEFINING "STATISTICS" AND "THE SCIENTIFIC METHOD"

The most important claims underlying the Fruits Report[34] concern the definition of "statistics" and "the scientific method." Fruits outlines an econometrics-based summary of "the scientific method".[35] and/while/on the other hand Hoover and Ball argue that Fruits assumes that the scientific method is based entirely on statistical testing, that statistical testing is limited to the practice of econometrics (the branch of statistics typically used by economists), and that econometrics is based fundamentally on the use of linear regression techniques.[36] Following this definition, Fruits states that the hypothesis-testing strategies used by Ball et al. are "unscientific" and "meaningless".[37]

Fruits' framing of "the scientific method" depends crucially on certain assumptions about available data on conflict processes, most particularly (1) the accuracy of existing quantitative data and (2) the supposition that conflict processes and the data they generate are susceptible to testing by linear regression models (cf. "general linear reality[38]). For an econometrician, these are usually sensible assumptions; economic data are significantly more complete and accurate than data on violent conflict. Furthermore, economic data measure phenomena that are neither exceptionally rare nor (in most cases) extremely unevenly distributed. However, violent events during armed conflict are rare, unevenly distributed, and often unknown to organizations collecting data, meaning that data about conflict violence are likely to be biased to a large, and unknown, degree. Given the limitations of violence data, standard econometric testing can (and does) produce unreliable results. Uncritical acceptance of Fruits'

33 Ball et al. 2002.
34 Fruits 2007.
35 Fruits 2007, sec. 4 and sec. 12.
36 Hoover and Ball 2007.
37 Fruits 2007.
38 Abbott 1988.

definition of the scientific method would rule out nearly any analytical method appropriate to these data.

In addition to privileging linear regression as a (the) method of testing, Fruits emphasizes tests of statistical significance, rather than tests of effect sizes or existence claims. This analytical choice reflects Fruits' expertise: significance testing, unlike testing for the existence of a phenomenon, is central to econometric practice. Meaningful tests of significance are indeed vital to the econometrician's definition of "real" evidence. However, this rather narrow view of the scientific method works, unsurprisingly, to the defense's advantage.

This framing of statistical testing underlies Fruits' claim that "Dr. Ball does not demonstrate statistically the existence of 'patterns' of deaths or 'patterns' of migration…what Dr. Ball describes as a pattern has no meaning…someone can 'eyeball' the data and find patterns contrary to Dr. Ball's conclusions"[39] and that that "Dr. Ball's conclusion about the patterns of deaths and migration are supported only by his interpretation of a visual inspection of the data series".[40] Yet a significance test is an inappropriate standard if the question concerns the existence of a pattern. In their rebuttal, Hoover and Ball state that:

> *"Fruits' suggestion that a traditional test of significance would be more dispositive than inspection of the data trends across municipalities, time, and violations is simply incorrect. Indeed, asking whether the observed data are significantly different from a random pattern (this is a typical null hypothesis of the significance type) sets an egregiously low bar for the prosecution, in that a random pattern of deaths and migration would appear as oscillation about a single mean, showing little variation across time or space and no association between violations or across municipalities. The differences between the observed data and randomly generated data could be calculated but would provide no useful evidence."*

Fruits' assessment of "standard practice" in the discipline of statistics is, similarly, tied to his experience with economic statistics. Fruits writes that "Dr. Ball unnecessarily inflates the estimated number of deaths" by using multiple systems estimation to reach a corrected mortality count.[41] Multiple systems estimation is presented, incorrectly and without reference to its extensive use in demographic statistics, mathematical demographics, and population biology, as an ad hoc technique. This assessment of multiple systems estimation ("inflation"), like Fruits' assessment of the HRDAG authors' hypothesis testing strategies, depends on an exceedingly specific definition of "the discipline of statistics," grounded in the practice of econometrics. This presents a serious problem for the Chamber's

39 Fruits 2007, sec. 9.2.
40 Fruits 2007, sec.10.
41 Fruits 2007, sec. 6.1.

decision process. Judges should not and cannot be expected to enter the courtroom with an accurate sense of the breadth and depth of various statistical sub-disciplines. Yet their ability to correctly assess what constitutes "standard practice," not to mention their ability to correctly assess the value of novel testing methods, depends crucially on their perceptions of disciplinary boundaries – boundaries presented by factions with no incentive to fairly inform the Chamber.

3.2. "FUNDAMENTAL FLAWS"

Fruits' Report catalogues a number of shortcomings in the Ball et al. reports,[42] each of which is held to be indicative of "fundamental flaws" in the Prosecution's evidence or argumentation. In an interview, Fruits noted the differences between the goals of academic argumentation (truth-seeking) and the goals of adversarial legal argumentation (invalidating the opponent's argument).[43] If the goal of the statistical expert is to discredit the opponent's argument, then an expert has no incentive to distinguish between truly fundamental flaws and quibbles. The Chamber is then faced with a huge volume of criticism, without a reliable guide as to how (or whether) each criticism should be prioritized.

For example, Fruits introduces a list of municipalities in which the Ball et al. estimates of killings or migration differed from recorded totals as proof of "fundamental flaws" in statistical reasoning, despite the fact that the Ball et al. descriptive analysis is specifically (and avowedly) designed to estimate *unrecorded* killings and migration. Similarly, Fruits claims that Ball et al.[44] produced estimates of deaths that were "inconsistent" with other scientific estimates.[45] On the basis of his reading of a graph,[46] Fruits estimates that the Spiegel and Salama mortality estimates fall outside the Ball et al. 95% confidence interval (approx. 9,000–12,000) for the March-June 1999 period. Fruits does not, however, note that Ball et al.'s estimates fall within the wide 95% confidence interval of Spiegel and Salama, or that Spiegel and Salama wrote extensively about potential downward biases in their survey results. None of this is to say that no serious questions about the Ball et al. analysis exist. Fruits correctly points out the poor quality of the underlying data, as well as potential errors introduced by imputation of place-of-origin statistics. From a statistically inexpert perspective, however, it is impossible to distinguish these criticisms as any more or any less important than others.

42 Ball et al. 2002 and Ball et al. 2007.
43 Personal Communication 2009.
44 Ball et al. 2002 and Ball et al. 2007.
45 Fruits 2007, sec. 6.2.
46 Spiegel and Salama 2000.

Prosecution experts responded with similarly harsh language. Hoover and Ball claim that Fruits has essentially no relevant expertise[47] and each and every criticism in the Fruits Report is the result of Fruits' self-serving definition of "statistics." Neither the Defense nor the Prosecution experts' hyperbolic language represents an accurate view of the competing analysis. Given the complexity, uncertainty, and nuance of the substantive questions, these polarized views produce little useful information for the Chamber to consider. Non-expert Judges are left to decide which side is correct on the basis of their (again, non-expert) perceptions of experts' trustworthiness, professional reputation or likeability.

Had a truly impartial expert, rather than experts employed by the Prosecution or the Defense, attempted a retrospective judgment of the data and methods in either Ball et al.[48] or Fruits[49], perhaps the Chamber could more accurately have prioritized its assessment of the statistical evidence. In retrospect, it seems likely that the descriptive estimation procedures described by Ball et al. would largely withstand criticism, while at least some parts of Fruits' assessment of the Prosecution analysts' causal hypotheses would stand. In particular, Fruits correctly notes that at various points Ball et al. should have employed a more traditional test (such as a standard test of significant differences, or a simple correlation coefficient) rather than relying on graphical representations of their findings. However, rather than raising criticisms that lead to increased clarity of analysis, the ICTY's adversarial procedures produce potentially spurious criticisms that are overconfidently stated but never themselves tested.

4. THE JUDGMENT

As noted above, the Judgment in *Milutinovic, et al.* contained only a brief section relevant to the statistical findings of Ball, et al. and Fruits.[50] The Chamber rendered its judgment of several person-years of work, and several full days of testimony, in nine paragraphs. Two of these paragraphs simply lay out the prosecution and defense hypotheses regarding the statistical evidence. Paragraph 23 lays out what the Chamber viewed as "five key issues" concerning this evidence:[51]

47 Hoover and Ball 2007: 1–2.
48 Ball et al. 2002 and Ball et al. 2007.
49 Fruits 2007.
50 *Milutinovic et al.*, 27166–27170.
51 *Milutinovic et al.*, 27172.

> "*Ball's potential bias; Fruits' alleged lack of qualification; the integrity and completeness of the underlying data; the soundness of the applied methodology; and, most importantly, the persuasiveness of the conclusion reached.*"

On all but the first of these issues, the Chamber sided with the Defense expert, Fruits, in his assessment of the evidence.

Before discussing the outcome of *Milutinovic, et al.*, it is important to acknowledge the possibility that the Kosovo cases are outliers, i.e., that "lessons learned" from these are inapplicable to other uses of statistics in human rights law. I wish to suggest that this is not the case – that adversarial legal institutions (including, for example, both American courts and the ICTY, but *not* the International Criminal Court) use fundamentally similar styles of reasoning that present particular challenges for statistical evidence. To be sure, the specific facts of the Kosovo cases underpinned both the statistical analysis and its reception in Chambers. But in the statistical complexities presented, these cases are hardly exceptional. In the context of human rights statistics, serious difficulty achieving, and convincingly conveying, accurate descriptive and causal inferences is the norm. However, because of the small number of human rights cases in which statistics have played a role, the generalizability of the lessons I draw below remains open to debate.

Reading the expert reports and trial transcripts from *Milosevic* and *Milutinovic et al.* gives the impression not of two scientists engaged in debate, but of witnesses residing in separate and mutually exclusive universes. In one universe, the Prosecution's prejudiced expert witness invents numbers to serve his political purposes and fails basic statistical tests of due diligence, so that none of his conclusions stands. In the other universe, the patterns and magnitudes reported by the Prosecution are correct and all plausible alternative hypotheses have been convincingly ruled out. Despite this, the expert witness for the Defense condemns the Prosecution's analysis, relying on distraction and even outright fabrication.

Which universe is the real one? In an extremely polarized and complex information environment, a statistically inexpert judge must determine for her (him)self where the truth lies. There exist cases in which common sense, logic and legal wherewithal are sufficient to make this type of determination. In these cases, the question at hand concerns a specific narrative about a specific individual, with little or no role for randomness (e.g., "On date X, did he or did he not steal the car?"). Psychological research shows that most juror decisions are framed around a particular narrative or story,[52] which allows for easier testing of facts ("Does this fact fit the story I'm considering?"). However, in the case of abstract or indirect "actions" (such as instituting, or ignoring, a policy

[52] Pennington and Hastie 1991 and 1992.

involving illegal noncombatant loss of life), concrete stories are more difficult to discern. In these cases, answering the "which universe" question requires understanding abstract, multi-part arguments for which there may exist no reliable documentary or testimonial evidence. Moreover, statistical evidence may rely on techniques that are highly specialized, extremely technically demanding, complex, or non-intuitive.

Findings from social and cognitive psychology suggest that high-volume; highly polarized information environments are inhospitable to complex understandings, and to facts derived from non-intuitive processes. Allison, for example, noted that high-stakes decision-making led many members of the Kennedy administration to become extraordinarily inflexible during negotiations over the Cuban missile crisis, adhering to "implicit conceptual models" even after those models were proven ineffective.[53] Janis found that decision-makers facing stress or threat, including expert decision-makers, become less flexible in their thinking.[54] Staw outlines a number of seminal works in social and cognitive psychology in which individual and group problem-solving capability are undermined by threats, including "threats" such as increasing the pace at which problems are solved.[55]

In a key critique of rationalist models of decision-making, Tversky and Kahneman discuss the prevalence of several modes of heuristic thinking, showing "that people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probability and predicting values to simpler judgmental operations.[56] In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors".[57] Tversky and Kahneman outline several common heuristics, the most relevant of which in the court setting is cognitive accessibility: items or instances that are easy to recall (for any of several reasons) are typically judged as more common or likely. Familiarity with concrete items or instances leads to increased recall of those items and, crucially in the present context, "imaginability" leads to increased judgment of likelihood for abstract concepts.[58] This well-replicated finding suggests that counter-intuitive information will be more difficult to store, more difficult to retrieve (less accessible) from memory, and consequently more difficult to employ in making judgments.

More recently, Lau and Redlawsk showed that heuristic use increases with information complexity and volume.[59] Lau and Redlawsk's subjects were political

53 Allison 1969: 689.
54 Janis 1972.
55 Staw et al. 1981.
56 Tversky and Kahneman 1974.
57 Tversky and Kahneman 1974: 1124.
58 Tversky and Kahneman 1974: 1127.
59 Lau and Redlawsk 2001.

decision-makers. Each was presented with an overwhelming amount of information about a policy option on which they had previously specified a preferred goal. The subjects were also presented with several potential sources of heuristic information, including subject-matter experts, as a means toward judging which option would satisfy their goal. Heuristic use *increased* the accuracy of well-informed subjects' judgments, but *decreased* the accuracy of novices' judgments, actually magnifying their lack of expertise. The implication for statistical experts presenting complex information to inexpert judges is clear.

In their Judgment, the *Milutinovic et al.* Chamber relied implicitly on several factors that were cognitively accessible but potentially irrelevant or misleading. The Chamber set a low bar for expert credentialing. The Judges appear to have used both explicit and implicit heuristics, including expert trustworthiness and concept accessibility, in considering their decision. Several pieces of evidence cited as particularly compelling in the Judgment are actually demonstrably irrelevant to the statistical findings in the case.

The Chamber adopted a relatively lenient standard in assessing expert witnesses' credentials. In particular, the court took no notice of differences in training and expertise between Ball, the Prosecution's key expert witness, and Fruits, for the defense. While both hold Ph.D.'s in the social sciences, Ball had served as a statistical expert for numerous large-scale documentations of human rights abuses. Fruits, on the other hand, is a business economist who had never taken or taught a course in statistical demography or dealt with human rights matters prior to this case. Over the Prosecution's objection, the court held that Fruits was a suitable expert because "he [ ] taught a higher education course concerning the problems associated with linear regression analysis, [ ] consulted on projects involving statistical analyses of demographic data, and [was] admitted as a statistical expert in courts in the U.S."[60]

The Chamber relied on the American standard for expert witnesses, laid out in *Daubert v. Merrill Dow Pharmaceutical*. In *Daubert* the Justices overturned the previous standard ("generally accepted practice," *U.S. v. Frye*, 1923) for expert witnesses, charging trial judges with conducting "far-ranging" inquiries into the potential expert's qualifications.[61] However, rather than applying the *Daubert* standard itself, the Chamber based its admission of Defense expert Fruits on the fact that he had previously been admitted as an expert in American courts. (Note, however, that Fruits had been admitted as an expert in civil insurance cases pertaining to insurance claims – not criminal trials, and certainly not war crimes trials.) As has been extensively discussed in both scientific and legal

60 *Milutinovic et al.*, 27168, para. 25.

61 *Daubert v. Merill Dow Pharmaceutical,* 509 U.S. 579 (1993).

forums,[62] *Daubert* asks that Judges assess potential experts' credibility and qualifications, despite their own lack of expertise.[63]

The defense attacked Ball's impartiality with some success, pointing out instances in the past during which Ball had made remarks (formally or informally) about his negative opinions of the defendants. The Judgment held that Ball's evidence would be considered on a substantive basis, because his testimony in court "displayed no bias",[64] but detailed at length Ball's confused response regarding a negative comment about Slobodan Milosevic from 2001, and found that "the evasive nature of the witness's responses casts doubt upon his objectivity".[65]

The complexity of the debate over statistics in the Kosovo cases, as well as the sheer volume of information considered at trial, negatively affected the Chamber's capacity to assess the statistical evidence. In particular, the Chamber failed to determine which Defense criticisms of the Ball et al. analyses were substantively important (or potentially substantively important) and which were irrelevant – to say nothing of which were accurate and which were erroneous. The Judgment suggested that the court had closely considered several criticisms of the statistical evidence that are demonstrably irrelevant to the major substantive findings of Ball et al.[66]

For example, the Judgment concurred with Fruits' assessment that lack of data on FRY activities constituted "omitted variable bias" in Ball et al.'s regression analysis.[67] However, Ball et al. did not dispute that omitted variable bias constituted a problem with any potential regression analysis; the Prosecution had never attempted to make its case using linear regression analysis.

The Chamber also found compelling Fruits' demonstration that at "at least three" times during the March-June 1999 period the time series graphs of killings and migration showed different patterns. While matching patterns of killing and migration are consistent with the hypothesis that the two patterns were part of the same policy, evidence (including traditional statistical tests) of coinciding patterns is neither necessary nor sufficient to support that hypothesis. Importantly, the Fruits criticism did not allege the existence of statistical evidence that would merit rejection of the hypothesis ("very similar patterns of killing and migration"), but rather criticized the Prosecution analysts' choice not to conduct formal tests

62 Jasanoff 2005; Kaufman 2001 and Tochterman 1996.

63 For a more general (pre-*Daubert*) account of problems with adversarial expert evidence, see Gross (1991), who argues for the appointment of impartial experts and notes that complicated evidence is unlikely to be correctly assessed in an adversarial setting.

64 *Milutinovic et al.,* 27167, para. 24.

65 *Milutinovic et al.,* 27167, para. 24.

66 Ball et al. 2002 and 2007.

67 See *Milutinovic et al.* 27168, para. 26.

(accusing Ball et al. of merely "eyeballing" their analysis). However, the Chamber found this criticism sufficiently damning to reject Ball et al.'s graphical evidence that patterns of killing and migration were closely related.

These criticisms were damaging to the Prosecution largely because the Prosecution bears the burden of proof. A Defense expert need offer no exonerating evidence; he need only show that the Prosecution's conclusions are overhasty or potentially vulnerable to criticism. Ball et al. did not prepare an extensive arsenal of statistical proofs, instead relying (sometimes necessarily, sometimes unnecessarily) on novel hypothesis testing methods of which the court was suspicious. The court attentively considered statistical criticisms that were demonstrably irrelevant largely because results actually demonstrating their irrelevance were never produced.

In addition, the Chamber failed to draw connections between non-statistical findings of fact and the statistical evidence. The Judgment stated that "the exclusion of the first two hypotheses [that KLA or NATO activity caused killings and refugee flow] by Ball – even if it is based upon correct data and methodology – is of little value because it still leaves a number of *potentially plausible* options unexplored" (emphasis mine).[68] Few other plausible hypotheses for the pattern of killings and migrations were offered during the course of the trial. Fear of crossfire could potentially explain population movements that occurred in the absence of killings, but in general the two patterns coincided closely over both time and space.[69] A second plausible alternative explanation might be *interaction* between two or more parties to the conflict, suggesting that non-combatant deaths were due to crossfire rather than ethnic targeting. But the Chamber's decision noted that (essentially) only Kosovo Albanians were dying or migrating; if all civilians were in danger, we would not expect to observe ethnic selection.

Given these facts, the Chamber's assessment of Ball et al.'s mode of hypothesis testing suggests a misreading of the principles of statistical testing. The implicit assumption in this section of the Judgment is that statistical tests must rule out all *conceivable* explanations in order for hypothesis testing to proceed by elimination, rather than only those that are *relevant* competitors.[70] Yet in this case, non-statistical (and uncontested) facts ruled out several hypotheses that might have been plausible, and the Chamber apparently did not connect the two sources of evidence as it considered its Judgment.

Comparing the information environment at the ICTY to the style of argumentation in statistics or academic social science is instructive, and helps to

68 *Milutinovic et al.* 27169, para. 28.
69 Ball 2000 and Ball et al. 2002.
70 *Milutinovic et al.* 27168, para. 26.

situate some of the strategic concepts introduced below in Section 5. Arguments between experts in statistics and social sciences, though not necessarily civil, typically focus on one or a few key disputes. At best, interlocutors carefully explain the extent to which the criticisms presented damage the key conclusions of the research, and separate minor empirical quibbles from major theoretical disagreements or methodological critiques. The "judges" in the academic setting are fellow experts, whose background knowledge (again, in a best-case scenario) permits judgment based on the merits of the work, rather than on heuristics such as the perceived trustworthiness of the source.[71] The importance of background knowledge leads to disciplinary divisions and has some negative repercussions for "outside the box" thinking (e.g., Fruits' non-recognition of a standard demographic technique, despite significant statistical experience, because his training is specific to economics), but it permits more accurate judgments of the relevance of evidence than would an inquiry conducted without this baseline expertise.

In addition, while not every academic interlocutor is well-intentioned, all have reputational incentives to engage other experts with some semblance of collegiality, and with a clear recognition of which criticisms are both important and appropriate. Arguments are situated in (or in relation to) long-standing lines of research and theory, which help other experts organize and prioritize their disagreements. Not so in the courtroom, where judges are situated in long-standing lines of *legal* research and theory, but must rapidly assimilate the facts (including technical facts) of each individual case anew.

5. LESSONS LEARNED

Given the difficulties faced by statistical expert witnesses in human rights trials, one might be tempted to argue that statistical evidence should play little or no role in international legal settings. Certainly, non-expert judges in adversarial settings will find it exceptionally difficult to reach a nuanced or abstract judgment based on incomplete information. Still, there are lessons to be learned from the experience of statistical experts in *Milosevic* and *Milutinovic et al.* As a statistician and social scientist, I might hope for changes in legal institutions. For example, both elimination of the rather minimalist *Daubert* expert credentialing currently in use at the ICTY[72] and the creation of independent experts, as opposed experts retained by the Prosecution and the Defense, would go some distance toward ordering judges' information environments. More pragmatically, however, it seems clear that the burden of change falls in the same location as the burden of proof – namely, with the Prosecution, and particularly with statistical experts for the Prosecution.

---

71 But see also Michele Lamont (2009), on the influence of emotion and social networks in academic assessments.

72 *Milutinovic et al.* 27167.

There is no reliable way to evaluate the hypothetical outcome in which Ball et al. prepared their evidence, or their presentations of the evidence, differently. But drawing on the documents presented as evidence, trial transcripts, the Judgment, and findings from social and cognitive psychology, it is possible to envision some improvement. The statistician's general goal should be extraordinary sensitivity to the overwhelming information environment that jurists face. Below I outline a number of factors related to this general rule, which proved consequential in this case and are likely in my estimation to be consequential in others as well: (1) presentation of background information vital to judges' understanding; (2) redundant, exhaustive hypothesis testing; (3) focus on causation, rather than description, where possible; (4) privileging of simple, familiar, or traditional modes of statistical testing, where possible; (5) total transparency with respect to data analysis processes; and (6) training for statisticians presenting their work in court settings.[73]

Effective presentation of background knowledge may be the most challenging piece of the statistical expert's extremely difficult puzzle. This is doubly so in the case of human rights statistics, which are seldom well-suited to traditional modes of analysis and therefore often rely on methods that are unfamiliar, counterintuitive, or both. It is important, then, to make the case for those methods simply and convincingly, in advance. Imperfect data are endemic to human rights statistics, and entirely unavoidable. Identifying different categories of quantitative data and clarifying their benefits and drawbacks is a key first step to justifying complex imputation strategies.

Lists of known victims (perhaps the most familiar category of quantitative data for judges) are seldom representative of the full (unknown) population of victims.[74] Although they have the benefit of verifiability, and it is simple to understand how they are generated, these lists seldom portray either magnitudes or patterns of violence accurately. Another category of "data", "back of the envelope" extrapolations not based directly on lists (e.g., claims of "400,000 dead" in Darfur) is neither verifiable nor representative. It is impossible to generate meaningful confidence intervals (or "margins of error," the hallmark of statistical reasoning) for either of these data types. Census correction techniques such as MSE, on the other hand, are based directly on verifiable lists, carry confidence intervals, and can be tested through representation or sequential model changes.

Like statistical techniques, important underlying concepts should be explained simply but thoroughly *before* results are presented. For example, Ball attempted

73 I thank both Patrick Ball and Eric Fruits for their observations on the importance of training to statisticians testifying at trial.

74 Lynch and Hoover 2008.

to explain the concept of "selection on the dependent variable" in the midst of testimony;[75] his analogy was effective, but lost in a broader discussion of the substantive impact of his findings.

Statistical experts must present not only a concise overview of types of data, but also specific illustrations of the ways in which convenience data are biased and the effectiveness of census correction and imputation techniques for correcting these biases. There exists a voluminous statistical and demographic literature on multiple systems estimation for elusive populations, including both empirical demonstrations (e.g., experiments with population counts in gated animal enclosures)[76] and computerized simulations.[77] Given the cognitive constraints of the environment, the point is less to convince judges of the merits of the techniques at issue, and more to give a clear statement of the intellectual pedigree of each method. Recall that judges, consciously or unconsciously, will be using credentialing (of both experts and information) and concept familiarity as heuristics in their judgments.

Necessary background information also includes coherent narratives connecting non-statistical evidence to statistical questions or results. For example, statisticians must clearly present non-statistical (e.g., documentary or testimonial) evidence that rules out plausible alternative hypotheses.

The second key strategy is redundant, exhaustive hypothesis testing. Whereas judgments of hypothesis tests in the academic environment are made at least ostensibly on the basis of shared expert knowledge (i.e., many hypotheses are implicitly ruled out), this is not the case in court. In the Kosovo cases, rather than relying on graphical evidence of coinciding patterns of migration and mortality – a choice that proved extremely damaging in the Chamber's Judgment – analysts could have produced a number of statistics and tests.

As noted above, Fruits suggested that significance testing of some sort should have been performed;[78] although the Prosecution's expert correctly noted that the null hypothesis would be discarded *too easily* (see discussion above at Sec. 3) using a test of significance, the Prosecution experts could and should have performed such a test. They could also, however, have chosen multiple graphical representations of killing and migration statistics, including scatter plots or histograms, or other statistics such as correlation coefficients between killing and migration – or indeed linear regression coefficients using killing, or lagged

75 *Prosecutor v. Milutinovic et al.*, Public Transcript of Hearing, Case No. IT-05–87-T, T.Ch. III, 20 February 2009.
Transcript: 10247.

76 Burnham and Overton 1978.

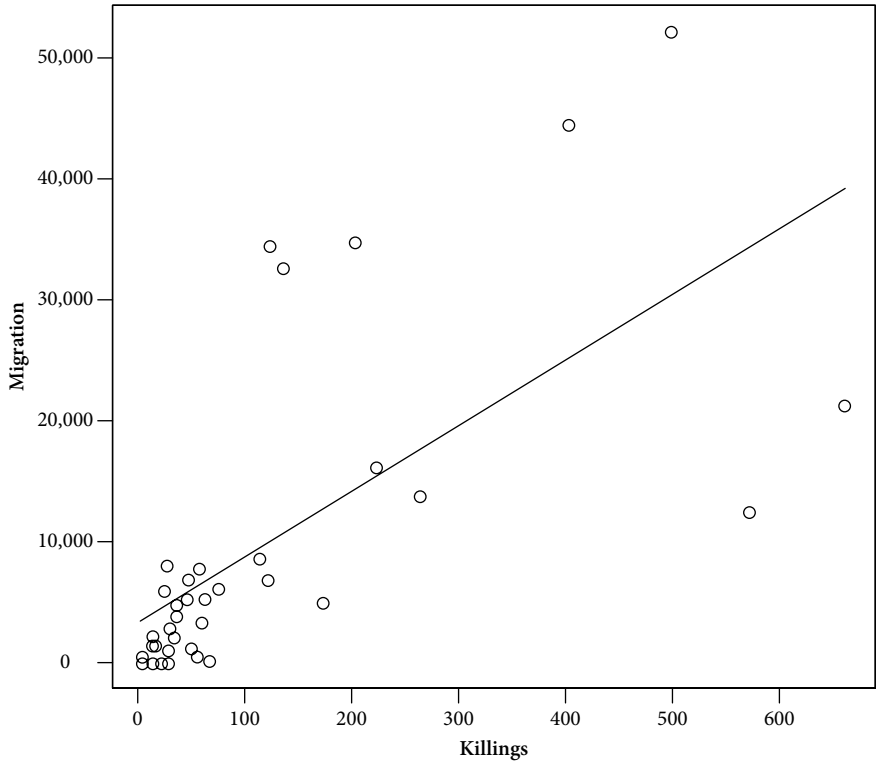
77 Chao et al. 1992 and Fienberg et al. 1999.

78 Fruits 2007.

killing, to explain migration. Any of these choices would have supported the prosecution's hypothesis and pre-empted the defense's claim, that an overarching pattern was not shown.

For example, the correlation coefficient between total killings and total migrations over time (for each two-day period between March 20 and May 31, 1999) is approximately 0.70.[79] In a simple bivariate regression model, regressing the number of persons leaving on the number of persons killed, the coefficient of the number killed is estimated at approximately 54, meaning that for every person killed, about 54 left home during the same two-day period. While killing cannot explain all the variation in migration, it explains a great deal. Furthermore, the probability that the true coefficient is zero (i.e., that there is no relationship between killings and migration, the implicit hypothesis of the Defense expert's Report) is less than 0.0001. Below is a simple scatter plot showing killings on the x axis, migration on the y axis, and the regression line.

Relationship Between Killings and Migration



79 Data and code for all illustrations in this section are available from the author: amelia.hoover@yale.edu.

The "peaks versus presence" analysis that the Chamber found unconvincing might also have been supplemented by any number of graphical illustrations and numerical tests of the same hypothesis. Prosecution statisticians might have analyzed the distribution of violence by testing whether average violence given KLA or NATO presence was significantly different from average violence given no KLA or NATO presence, or average violence more generally. In fact, it was not. The table below compares municipality-days with NATO presence or bombing (group 1) and municipality-days without NATO presence or bombing (group 2), and finds that in no case can we reject the null hypothesis of non-association between NATO presence and killings (N=1392 municipality-days).

Null Hypothesis	**Mean With NATO presence**	**Mean Without NATO presence**	**P-value**	**Reject the null?**
True difference in mean number killed (between groups 1 and 2) is equal to zero	3.9	3.3	0.47	NO
True difference in proportion of days on which anyone is killed (between groups 1 and 2) is equal to zero.	.37	.33	0.29	NO
True difference in proportion of days on which more than five persons are killed (between groups 1 and 2) is equal to zero.	.16	.13	0.27	NO

These data can be presented in another way as well: the total proportion of days with killings is 0.34; the total proportion of days with NATO activity is approximately 0.20. If NATO activity and killings were unrelated (statistically speaking, independent), we would expect the proportion of days with both killings and NATO activity to be approximately 0.34 * 0.20, or 0.068. There are 1392 municipality-days in total, so we would expect about 0.068 * 1392, or 94.8, municipality-days with both killings and NATO activity if the relationship between NATO activity and killing were entirely random. In fact, the observed count of days with both killings and NATO activity is 94, supporting the hypothesis that NATO activity is unrelated to killings. These simple, compelling analyses can be repeated with KLA activity, although data on KLA activity is significantly sparser than data on NATO activity.

Another aspect of redundant testing is testing with both uncorrected and corrected data. While corrected data is more scientifically reliable, given the cognitive accessibility of uncorrected lists or "minima," it may be advisable to complete each step of the analysis using both types of data. If the results are similar (as they were in the Kosovo analyses of Ball et al.) then this strengthens

the analyst's conclusions; if not, unless the census correction results are dubious for other reasons, there is a strong case to be made that corrected results provide better information for model fitting.

Finally, "redundant, exhaustive" testing involves testing even minimally plausible hypotheses if at all possible. In this case, the prosecution's analysis was based on a somewhat academic understanding of competing hypotheses: those that were widely considered plausible were tested if at all possible (although the key causal hypothesis, that FRY activity caused the observed killings and migration, could not be tested directly), whereas those that had been ruled out, whether by other factual information or because they were not deemed sufficiently plausible, were not dealt with.

As we have seen, this proved disastrous for the prosecution statisticians' strategy of ruling out competing hypotheses rather than confirming the key hypothesis. During the course of the trial numerous other hypotheses were brought forward, many of which could have been discarded without comment if they had been tested. Instead, the failure to test these hypotheses introduced doubt as to whether enough hypotheses had been eliminated; the Chamber cited the existence of other (undefined but not discarded) hypotheses as a key reason for its view that Fruits' criticism had discredited the prosecution's statistical evidence and conclusions. Had prosecution statisticians or prosecutors either tested more hypotheses or more carefully drawn the connections between non-statistical findings of fact and the statistical evidence, this outcome might have been averted.

A third key strategy involves focus. While statisticians (and especially demographic statisticians) often focus on the intricacies of descriptive inference, the Chamber in the Kosovo cases was less focused on *what* had happened than on *why* and *how* it had happened. The Prosecution statisticians expended very significant energy on descriptive estimations, but offered only two tests (linear regression analysis and "peaks versus presence") of the relevant causal hypotheses. In addition to adjusting to the information environment of a court case, statisticians must carefully attend to judges' preferred focuses.

Having performed exhaustive, redundant testing focused on causal hypotheses rather than descriptive statistics, the analyst must also carefully choose what information to emphasize, and what to hold in reserve. Given the success of defense criticisms of unfamiliar or unorthodox statistical techniques, it appears likely that prosecution statisticians would have benefited from presenting reliable results of more traditional statistics first, prior to introducing graphical illustrations, "peaks versus presence"-style analysis of sequencing, or other less familiar strategies. To the extent that they effectively (neither over-confidently

nor under-confidently) represent the statistician's key conclusions, simple statistics such as correlation coefficients, linear regression coefficients and tests of significant differences should be presented as key statistical evidence *before* resorting to unfamiliar or complex estimations.

Two final strategic suggestions – transparency of the analysis process and increased training for statisticians in legal settings – require little explanation. Transparency in analysis is a fundamental requirement of quantitative academic analyses, and a key factor in assessing the merit of statistical work. In that light, one advantage to using simpler or more simplistic analytical techniques is the ease with which one's interlocutors can reproduce them. Lastly, both expert witnesses in this case noted the importance of training for experts who need to present complex, detailed technical information in an adversarial environment.

6. CONCLUSIONS

As with most work undertaken in the aftermath of conflict or human rights abuses, statistical analyses of violence carry extremely high stakes. Incorrect analyses can and do lead to counterproductive policies and misallocations of resources, which cost lives. In court cases involving human rights abuses, incorrect analysis may allow extremely bad actors to avoid responsibility – or it may lead to wrongful convictions. Unfortunately, the formal and informal institutions of legal reasoning may make it exceptionally difficult for statisticians to present their work accurately in court. Statistical work involves significant technical detail in an already overwhelming information environment; in adversarial legal settings, statisticians are "employed" by only one party to the case and have incentives (not to mention responsibilities) to present their work in a light that, to the extent possible, advantages that party.

This analysis raises three major questions for future research. The first concerns credentialing: can courts (including special courts such as the ICTY) reasonably rely on the *Daubert* standard when there may be large gaps in experience and expertise between "experts" who meet the *Daubert* standard? Judges cannot and should not be expected to assess the difference between a mathematical demographer and a business economist (for example), but the difference can make a serious impact in court, as relatively inexpert "experts" pronounce confidently – and incorrectly – about important matters. A second question is whether non-adversarial environments (e.g., European court systems or the International Criminal Court) will experience more success than adversarial legal systems as they assess expert evidence. A significant amount of theoretical background suggests they will. However, especially at

the international level, there exists no systematic study of judicial decision-making about statistics. Given the large and increasing role of statistical debates in the international human rights community, this seems an important avenue for future research.

A final, more normative, question raised by this analysis concerns the role and responsibility of special tribunals such as the ICTY. The question whether the fate of statistics in the international legal community *matters* turns on the importance of statistics for international courts' work. In the Kosovo cases, statistics occupied little of the Judgment and appeared to make no impact whatever on the eventual verdicts. However, if the responsibility of special courts is not simply to render verdicts, but to create a definitive history, the treatment of the statistical record matters very much indeed.[80] In Kosovo and elsewhere, if the court intends to do justice by finding truth (rather than simply assigning responsibility, or not), then its ability to consider of evidence about patterns, magnitudes and causes of violence must certainly improve.

REFERENCES

Abbott, A. (1988). Transcending general linear reality, *Sociological Theory 6*(2), 169–186.

Allison, G. T. (1969). Conceptual models and the Cuban missile crisis, *American Political Science Review 63*(3), 689–718.

Ball, P. (2000). *Policy or panic? The flight of ethnic Albanians from Kosovo, March-June 1999*, Washington, DC: American Association for the Advancement of Science.

Ball, P., W. Betts, F. Scheuren, J. Dudukovich & J. Asher (2002a). *Killings and refugee flow in Kosovo, March-June 1999*, Report to the International Criminal Tribunal for the Former Yugoslavia, The Hague: International Criminal Tribunal for the Former Yugoslavia.

Ball, P., W. Betts, F. Scheuren, J. Dudukovich & J. Asher (2002b). *Killings and refugee flow in Kosovo, March-June 1999*, CORRIGENDUM. Report to the International Criminal Tribunal for the Former Yugoslavia, 1–15.

Ball, P., J. Asher, D. Sulmont & D. Manrique (2003). *How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000* (AAAS Science and Human Rights Program Report), Washington, DC: American Association for the Advancement of Science.

Ball, P., M. Lynch & A. Hoover (2007). *Revisiting killings and refugee flow in Kosovo: responses to additional data and analysis*, Report to the International Criminal Tribunal for the Former Yugoslavia, The Hague: International Criminal Tribunal for the Former Yugoslavia.

Bishop, Y., S.E. Fienberg, & P.H. Holland (1975). *Discrete multivariate analysis: theory and practice*, Cambridge, MA: MIT Press.

[80] I thank Michael Scharf for this observation.

Burnham, K. & W. Overton (1978). Estimation of the size of a closed population when capture probabilities vary among animals, *Biometrika 65*(3), 625–633.

Chandra Sekar, C. & W.E. Deming (1949). On a method of estimating birth and death rates and the extent of registration, *Journal of the American Statistical Association 44*(245), 101–115.

Chao, A., S. Lee & S. Jeng (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal, *Biometrics 48*, 201–216.

Darroch, J., S. Fienberg, G. Glonek & B. Junker (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability, *Journal of the American Statistical Association 88*(423), 1137–1148.

Fienberg, S., M. Johnson & B. Junker (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists, *Journal of the Royal Statistical Society 162*(3), 383–405.

Fruits, E. (2007). *Expert report of Dr. Eric Fruits*, Report to the International Criminal Tribunal for the Former Yugoslavia, The Hague: International Criminal Tribunal for the Former Yugoslavia.

Gross, S. (1991). Expert evidence, *Wisconsin Law Review*, 1113–1232.

Guzmán, D., T. Guberek, A. Hoover & P. Ball (2007). Missing people in Casanare, *HRDAG White Paper*, 1–21.

Hoover, A. & P. Ball (2007). *Lines of questioning for Dr. Eric Fruits*, Report to the International Criminal Tribunal for the Former Yugoslavia, The Hague: International Criminal Tribunal for the Former Yugoslavia.

Jacoby, J. (1984). Perspectives on information overload, *Journal of Consumer Research 10*(4), 432–435.

Janis, I. L. (1972). *Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes,* Houghton Mifflin Boston.

Jasanoff, S. (2005). Law's knowledge: science for justice in legal settings, *American Journal of Public Health 95*(S1), S49-S58. (doi: 10.2105/AJPH.2004.045732).

Kaufman, H. (2001). The expert witness: neither frye nor Daubert solved the problem. What can be done? *Science and Justice 41*, 7–20. (doi: 10.1016/S1355–0306(01)71844–8).

Lamont, M. (2009). *How professors think*, Cambridge, MA: Harvard University Press.

Lau, R. & D. Redlawsk (2001). Advantages and disadvantages of cognitive heuristics in political decision making, *American Journal of Political Science 45*(4), 951–971.

Lurie, N. (2004). Decision making in information-rich environments: the role of information structure, *Journal of Consumer Research 30*, 473–486.

Lynch, M. & A. Hoover (2008). Counting the uncounted: multiple data systems and the analysis of count data, presented at the Annual Meetings of the American Political Science Association, Boston.

Pennington, N. & R. Hastie (1991). A cognitive theory of juror decision making: the story model, *Cardozo Law Review 13(2–3)*, 519.

Pennington, N. & R. Hastie (1992). Explaining the evidence: tests of the story model for juror decision making, *Journal of Personality and Social Psychology 62*(2), 189–206.

Silva, R. & P. Ball (2006). *The profile of human rights violations in Timor-Leste, 1974–1999.* Report to the CAVR, 1–202.

Simon, D. (2004). A third view of the black box: cognitive coherence in legal decision making, *University of Chicago Law Review 71*(2), 511–586.

Spiegel, P. & P. Salama (2000). War and mortality in Kosovo, 1998–99: an epidemiological testimony, *Lancet 355*(9222), 2204–2209.

Staw, B., L. Sandelands & J. Dutton (1981). Threat rigidity effects in organizational behavior: a multilevel analysis, *Administrative Science Quarterly 26*(4), 501–524.

Tochterman, R. (1996). Daubert: a (California) trial judge dissents, *UC Davis Law Review 30*, 1013.

Tversky, A. & D. Kahneman (1974). Judgment under uncertainty: heuristics and biases, *science, New Series 185*(4157), 1124–1131.

Zaller, J. & S. Feldman (1992). A simple theory of the survey response: answering questions versus revealing preferences, *American Journal of Political Science 36*(3), 579–616.