DATA WILL DRIVE INNOVATION IN PUBLIC POLICY AND MANAGEMENT RESEARCH IN THE NEXT DECADE

Maureen A. Pirog

While there are many strong forces at play in the field of policy analysis and management, the strongest and most fundamental changes that will shape the field of policy analysis over the next decade are in the types of data that are increasingly available to researchers. Because research designs and statistical/econometric approaches are codependent with data, changes in the types of data available to public policy and management researchers will transform the field. Additionally, continuing research on the ability of econometrics to compensate for selection biases will continue to push researchers interested in the does it work question further down the path toward random assignment studies. Creative work on better quasi-experimental methods could change this movement, but in the absence of better econometrics, researchers will continue the movement into experimental research. Let me begin with the low-lying fruit, changes in the types and accessibility of data.

DATA ARE DRIVING INNOVATION IN POLICY ANALYSIS AND PUBLIC MANAGEMENT RESEARCH

Data are key inputs to the production of public policy and management research. Early in my career there were warnings related to data, typically framed as GIGO or garbage in-garbage out. While this adage still stands the test of time, data quality, and availability is perhaps better today than it has ever been. The data.gov Web site alone lists over 80,000 datasets, searchable by key words, and produced by governments at all levels, universities, nonprofits, commercial, and collaborative efforts. This list is incomplete inasmuch as it does not include the larger number of datasets that are not immediately available because they require the completion of human subjects’ protocols to safeguard the privacy of the data.

Administrative Data Linked Across Agencies and Programs

Administrative data maintained by governments at all levels are not new and in fact, they have been used by policy analysts and management researchers for decades. As such, the changes I describe here are perhaps the least surprising or innovative, but in the short-run, these changes are likely to be adopted by many public policy and management researchers, because the changes are producing data that are quite similar to the types of data in current use. Hence, using these data will not require big changes in research designs or statistical approaches. Researchers will, however, have access to more complete or comprehensive data on citizens, and having a fuller picture of individuals should, all other things being equal, improve public policy and management research.

Efforts to link records across government programs and agencies are occurring at both the state and federal levels. While a seemingly simple concept, most governmental information systems are designed to operate programs, not conduct research. As such, how the data are stored can be inconsistent with data storage for
research purposes. Additionally, many programs do not use social security numbers as unique identifiers and in the absence of a national identification number, how one correctly merges records is not as simple as it may first seem. Nonetheless, efforts to merge data across government agencies and programs are moving forward at both the federal and state levels.

Record linking projects are relatively common in education and workforce-related research due in large part to the Statewide Longitudinal Data Systems (SLDS) Grant Program authorized by the Educational Technical Assistance Act of 2002 (ETA), and subsequently boosted by the American Recovery and Reinvestment Act of 2009 (ARRA). The Institute for Education Sciences (IES) began awarding competitive grants to states in 2005. By 2012 every state had received SLDS funding with the exceptions of Wyoming, New Mexico, and Alabama (National Center for Education Statistics [NCES], 2013a). In 2009, a similar effort to encourage states to collect administrative data across agencies was implemented by the U.S. Department of Labor. The Workforce Data Quality Initiative (WDQI) is focused primarily on longitudinal labor market data systems, although it also supports efforts to match individual records to education data. WDQI is smaller than SLDS: As of 2013, 31 states received $31 million WDQI grant funds, whereas more than $300 million in SLDS grants had been awarded to 47 states (NCES, 2013c). The U.S. Department of Health and Human Services’ Administration for Children and Families (ACF) additionally has a history of providing funding to enhance states’ TANF and related administrative data for a variety of reasons, including research (Wheaton, Durham, & Loprest, 2012).

The State of Washington provides one example of how these federal initiatives are playing out at the state level. $23 million in SLDS funds in 2009 and $1 million in WDQI funding in 2012 allowed the state to create the Education Research and Data Center (ERDC) (NCES, 2013b). The ERDC links individual-level data across 11 state agencies including the Administrative Office of the Courts, Office of the Superintendent of Public Instruction, State Board for Community and Technical Colleges, Washington Student Achievement Council, Employment Security Department, as well as the Departments of Corrections, Early Learning, Health, Licensing, Labor and Industries, and Retirement Systems (ERDC, 2014). Births, deaths, health outcomes, labor market participation, incarceration, education outcomes are just some of the key individual-level linked variables available at through the ERDC.

**Biology, Psychology, and Public Policy**

While the sheer number of datasets easily accessible for analyses is exploding, changes in the content of some of these datasets are remarkable and will allow new and important avenues of inquiry. For example, just 10 years ago, the mapping of the human genome was declared complete (Noble, 2003). Fletcher and Conley (2013) note the rapid improvement in these data and indicate that numerous collaborative groups have amassed genetic data on tens of thousands of individuals, making highly powered studies a new and recent reality. In the same article, he also outlines research designs best suited to blending public policy research to capture important genetic and environment interactions. In another article, Fletcher (2012) links genotypes and geocodes with the National Health and Nutrition Examination Survey (NHANES) to use genetic data to explain heterogeneous response to tobacco taxation, and highlighting the limitations of traditional sin taxes in moderating tobacco use.

Such studies are likely to proliferate as policy researchers learn how to use information from the over 300 million biospecimens stored in U.S. biobanks that are publically and privately owned (Maschke, 2008). Most of these data would need to...
be linked with individual-level, socioeconomic data, and policy levers to make them useful for policy researchers as was done in Fletcher’s tobacco tax study. However, other research teams are directly collecting social, demographic, economic, and biological data rather than matching and merging data from different sources. For example, the National Social Life, Health, and Aging Project (n.d.) directly collects nationally representative longitudinal data on over 3,000 individuals and includes a genetic profile as well as blood, saliva, and urine samples in addition to a broad array of traditional social, health, and biological variables. The inclusion of biomeasures with traditional policy variables and controls can, at a bare minimum, illuminate the genetic basis of heterogeneous response to public policies, programs, and managerial approaches. Such work can have profound implications for public policies and is likely to help explain why many blunt policy levers are not as effective as predicted. It is likely that this type of work will profoundly influence the study of risky behaviors such as drug use or unprotected sex, criminal behaviors, and rehabilitation, or failing to save for one’s old age. Undoubtedly, this type of work will push policymakers to find new approaches to dealing with social issues.

Similarly, developmental psychologists and neuroscientists have investigated the biological bases of learning but have yet to build a translational education science to rival medical science in which biological knowledge is translated into medical practice (Roediger, 2013). This gap will undoubtedly be closed somewhat over the next decade or two as education policy researchers seek additional mechanisms beyond smaller class sizes, magnet schools, charter schools, school voucher programs, or performance-based management policies such as No Child Left Behind in an attempt to improve learning outcomes for children. A recent pioneering study (Nelson III & Sheridan, 2011) uses cognitive neuroscience to explore causal links between family and neighborhood characteristics and educational outcomes and draws out the implications for public policies. With advanced brain imaging techniques and subsequent expansions in interdisciplinary research in child development, the potential to translate findings from the lab to large-scale social experiments seems inevitable.

Geospatial Data

Real estate agents long ago knew that the three most important factors in selling homes are location, location, and location. Policy analysts are increasingly aware that spatial relationships between places as well as place-specific context matter and play important role in policy outcomes (Wise & Craglia, 2010). Geospatial data are not new, but again they are increasingly accessible due to the incorporation of geocodes in large social surveys as well as mobile technologies, web applications, and data storage advances that allow policy researchers to access GIS data without certified GIS technicians or expensive software licenses. But the rasterizing of data and the tremendous growth in the availability of such data are already transforming the conduct of policy research. It is hardly surprising that location matters. All other things being equal, growing up in a neighborhood of concentrated poverty produces a different person from one raised in middle or upper income neighborhood.

Geospatial data are now being used to examine the impacts of a wide variety of policy and management choices. Housing, employment, environmental compliance, and contamination are just a few of the policy areas where geospatial data are used regularly. Over the past several years, in this journal alone, geospatial data (beyond country, state, or county fixed effects) have been used to examine a wide variety of policy and management issues. A few examples provide some insight into the breadth of policy and management questions illuminated by geospatial data: the neighborhood spillover effects of city-supported rehabilitation of housing by for-profit and nonprofit developers (Gould Ellen & Voicu, 2006); the impacts of decentralized
governance on deforestation (Andersson & Gibson, 2007); the neighborhood impacts of a mortgage assistance program (Di, Ma, & Murdoch, 2010); the effects of place-based socioeconomic attributes (redlining), and place-based risk factors on the place-based component of insurance premiums (Ong & Stoll, 2007); the effects of living in ethnic enclaves on access to employment by low-skill workers (Liu, 2009); the effects of living in low-income, predominately minority areas on access to grants to adopt green technologies (Ong, 2012); variation in the employment effects of enterprise zones conditioned on location (Kolko & Neumark, 2010); the effects of teachers’ preferences for proximity in disadvantaging urban schools (Boyd et al., 2005); and the effects of credit scores on residential sorting (Nelson, 2010).

Such studies are the early precursors of what is coming over the next decade. Real-time satellite data, GPS location data, cellular communication data, internet search data, security and traffic cameras, economic transactions and summaries (credit card transactions, credit rating data, and web-vendor data), and photo radar speed enforcement are a few additional examples of geographically situated data that will continue to push geospatial modeling to the forefront. While these types of data have a geographic component, most also fall into the category of big data and are discussed below.

BIG Data

Policy analysts and public management researchers are standing on the precipice of big data, called by Paul Decker (2014) in his Presidential Address, the “data tsunami.” The analysis of big data is labeled in a variety of ways including, among others, data analytics, predictive analytics, or business intelligence. Not surprisingly, one of the key attributes of big data is sheer size. In 2012 about 2.5 exabytes of data were created every day. Put into perspective, more data were generated on the Internet in a single second in 2012 than in an entire year 20 years earlier—and this amount is expected to double every 40 months or thereabouts (McAfee & Brynjolfsson, 2012).

Not only is volume a key characteristic of big data, but also its velocity. Real-time or nearly real-time data requires a fundamental rethinking about how we conduct public policy and management research. Our theoretical approaches to analyzing human and organizational behavior were developed without access to massive longitudinal datasets including geospatially linked, second-by-second interactions for millions of people. Our methodological approaches assume that datasets are mostly stationary with perhaps annual or maybe monthly updates. Big data are far more likely to require an “ongoing conversation” with data which has ramifications for the conduct, replicability, and publishing of policy research.

When we think about data that we have used traditionally, we think about longitudinal, cross-sectional, panel, repeated cross sections, ethnographic, participant observer, and a few other types of data collected by government agencies, researchers, or third-party contractors. The data are largely represented as numbers or text. The sources of big data are, however, far more diverse and many of these sources are relatively new. Among others, big data include sensor readings, satellite images, images posted to social networks, Internet shopping and browsing, and cellular transmissions. Also big data are largely unstructured—that is, they are not available as datasets. The new types and structures of these data will require new approaches to research design and data analysis. It is hardly surprising that these issues are not yet ironed out. Many of these sources of data are relatively new. For example, Facebook was launched in 2004 and Twitter in 2006.

As one small example, we can look at a project run by Nokia from 2009 to 2011 in Lausanne, Switzerland, to collect cell phone transmissions. Nokia collected cell phone records for fewer than 200 young volunteers expressly for the research...
community to study human and social phenomena. The types of data and numbers of observations collected included were calls (in/out/missed), 240,227; SMS (in/out/failed/pending), 175,832; photos, 37,151; videos, 2,940; application events, 8,096,870; calendar entries, 13,792; phone book entries, 45,928; location points, 26,152,673; unique cell towers, 99,166; accelerometer samples, 1,273,333; Bluetooth observations, 38,259,550; unique Bluetooth devices, 498,593; WLAN observations, 31,013,270; unique WLAN access points, 560,441; and audio samples, 595,895 (Laurila et al., 2012). Among many other options, these types of data can be used to ascertain patterns of use of government services (e.g., libraries, health clinics), technical problems encountered in accessing services (e.g., Obamacare, IRS), or ease and popularity of new online service delivery systems (e.g., license bureaus, electronic road tolls).

In practice, big data are being used to optimize emergency evacuations, criminal apprehensions (e.g., using crowd-sourced photos at the Boston Marathon bombing), and predict election outcomes (Lane & Stodden, 2013). Although subject to considerable controversy, the U.S. National Security Administration and the UK Government Communications Headquarters regularly collect cellular and internet transmissions (Google, Facebook, Microsoft, Apple, Yahoo, and Skype) for national security purposes. Initially, Google Flu trends predicted the spread of influenza globally using search engine queries faster and more accurately than the Center for Disease Control (Ginsberg et al., 2009).

In contrast, policy research using big data are just beginning to appear. For example, using online seating maps, Garrow (2013) shows that airline practices of blocking seats for premier customers forces other flyers to purchase premier seats even when the blocked seats are empty, a topic addressed in a U.S. Senate panel in 2012 by Transportation Secretary LaHood. Undoubtedly, keystroke analysis will provide insights into the types and concentrations of persons adversely impacted by the management problems associated with the rollout of Obamacare. Access to government services as well as their use and misuse and how changes in public policies change these patterns will open up enormous opportunities for policy and management researchers over the next decade.

Aside from reconceptualizing data and devising new approaches to data analyses, there are other barriers to the use of big data. Much of the data are proprietary and how to secure access is unclear. Additionally, most of these data were produced by individuals with some expectation of anonymity. Protocols to protect human subjects using big data are likely to differ from regular university protocols. Informed consent is at the heart of the issue as is the ability of private vendors to anonymize the data. These and other related issues are currently under discussion by the American Statistical Association’s Committee on Privacy and Confidentiality.

Methodological Trends

The second major change that is already occurring in policy research is methodological. Most of the econometric corrections for self-section bias do not reliably reproduce experimental findings. As editor-in-chief of JPAM, one of the most common, if not the most common, reason for rejecting papers that are sent out for review is that the referees do not have confidence in the identification strategies used by the authors. In other words, they do not have confidence in the instrumental variable(s), ability of propensity score matching to capture important unobserved variables, or the exogeneity of policy intervention. The increasing awareness of the limitations of econometrics is pushing government agencies and large swaths of researchers to conduct field experiments, alternately referred to as randomized clinical trials (RCTs).
In the 1990s the federal Department of Health and Human Services required states to conduct experimental evaluations of state welfare waivers as a condition of receiving a state waiver. The What Works Clearinghouse of the U.S. Department of Education’s Institute for Education Sciences provides an unambiguous statement that random assignment is the method of choice in the conduct of evaluations. There has been an independent groundswell of support for experimental approaches to evaluation in the research community. Perhaps the most visible manifestation of this is the growth of the Abdul Latif Jameel Poverty Action Lab (J-PAL) initiated at the Massachusetts Institute of Technology and a global network of researchers committed to rigorous impact evaluations through the conduct of RCTs. From a modest venture with four affiliated professors in 2003, J-PAL’s 91 affiliated professors and staff have 441 completed or ongoing RCTs in 52 countries (See http://www.povertyactionlab.org/History).

In summary, new types and better data are here and access to BIG data will transform the conduct of policy and management research. The real-time aspect of new data sources will move us closer to an ongoing conversation, ballad or dance with our data. In the absence of real methodological breakthroughs with quasi-experimental research designs, the field will continue its current movement toward field experiments.

MAUREEN A. PIROG is Rudy Professor at Indiana University, 1315 East 10th Street, Bloomington, IN 47405; Affiliate Professor at the University of Washington, Seattle, WA; and Research Advisor, Sanlam Centre for Public Management and Governance, University of Johannesburg, Johannesburg, South Africa (E-mail: pirog@indiana.edu.)

REFERENCES


