DOI: 10.1111/rssa.12762

COMMENTARY

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Series A
Statistics
in Society

Enhancing (publications on) data quality: Deeper data minding and fuller data confession

Xiao-Li Meng

Department of Statistics, Harvard University, Cambridge, Massachusetts, USA

Correspondence
Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, MA 20138, USA
Email: meng@stat.harvard.edu

Abstract

Statistics typically treats data as inputs for analysis, whereas the broader data science enterprise deals with the entire data life cycle, including the phases that output data. This commentary argues that it would benefit statistics and (data) science if we statisticians were also to treat data as products in and of themselves, and accordingly subject them to *data minding*, a stringent quality inspection process that scrutinizes data conceptualization, data pre-processing, data curation and data provenance, in addition to data collection, the traditional objective of our emphasis before data analysis. A concrete step in promoting deeper data minding is to encourage fuller *data confession* in (statistical) publications, that is, to entice—or at least not to disincentivize—the authors into providing more details on the genealogy of a given body of data, including an account of its deliberations, especially with respect to sources of adverse influence on data quality. The collection of articles in this special issue (on data science for societies) provides both the inspiration and aspiration for deeper data minding and fuller data confession.

KEYWORDS

data collection, data conceptualization, data curation, data life cycle, data pre-processing, data provenance, data science ecosystem

1 | WHAT'S THE DIFFERENCE BETWEEN STATISTICS AND DATA SCIENCE?

If one were to pose the question, 'What's the difference between physics and physical science?', it would sound strange even—or especially—for a physics journal. Physics is a pillar of physical science, a collection of disciplines that study systematically the inorganic world at all scales, from quarks to cosmos. But there are other pillars, such as astronomy, chemistry, Earth and planetary science, etc. Whereas there are always cases providing food for thought or fight (e.g. Is material science a branch of physical science?), few informed minds still find themselves contemplating the question of how physics differs from physical science.

The same cannot be said, however, of the question: 'What's the difference between statistics and data science?' That question retains its hold on many minds, whether informed or not. Indeed, some of us still get rather emotional on this matter: 'Data science is just statistics!' or even 'DS is BS!' (And we do not mean Bayesian statistics).

Why do we react this way? Why cannot we answer the question of how statistics differs from data science confidently: 'Statistics is a pillar of data science, a collection of disciplines that study systematically the digital world in every aspect, from what data to collect or conceal to how data inform or mislead'? But statistics is not the only pillar; there are also applied mathematics, computer science, information science, operational research, philosophy, signal processing, etc. Operationally, statistics typically treats data as inputs for the purpose of analysis and decision-making, whereas data science addresses the entire data life cycle, including data considered as output or more broadly as products in and of themselves. Indeed, the increased emphases on scientific reproducibility, replicability and reliability (e.g. see Meng, 2020) and the open science movement compel us to preserve and curate data for others to scrutinize and study, a critical (data) science endeavour that simply is not a piece of statistical cake (or a cup of statisticians' *t*). Granted, the nascent nature of data science generates a longer menu of food for thought or fight, but the general relationship between statistics and data science should be at least as clear as that between physics and physical science, or sociology and social science, or anthropology and the humanities.

This special issue of *JRSSA* on 'Data Science for Society: Challenges, Developments and Applications' provides us with a great opportunity to reflect on the role and place of statistics within data science, and to contemplate how and what we can do to benefit further the broad data science ecosystem and, ultimately, the human ecosystem (Meng, 2019). The call for the special issue identifies several (but not all) major challenges and opportunities for statisticians and data scientists in general, concerning:

(i) researcher access to data while preserving privacy and maintaining confidentiality,
(ii) understanding quality and sources of bias in data derived from multiple sources, and
(iii) handling the complexity and high-dimensionality of large, longitudinal and high velocity data sources.

I was particularly pleased to see in this call an emphasis on understanding the *data* themselves, with a focus on the issues of data privacy and bias, and more broadly, on data quality. The broader data science community is addressing these critical issues with great urgency even relative to the development of methods for processing or analysing data, the traditional focus of statistics, machine learning and the like. When I was given the honour of writing this commentary, I naturally wanted to see how these challenges would be addressed in this special issue. I was also curious whether the term ‘Data Science’ had ended up being employed herein merely as a cosmetic amplifier or to signify a broader investigative framework, with contributions that normally would not be expected in statistical publications.

2 | ENHANCING DATA LIFE CYCLE WITH DATA MINDING

Table 1 documents my attempt to collect some data to address my curiosity. The columns are ordered approximately according to the data life cycle depicted by computer scientist Jeannette Wing (2019), as enhanced by the data conceptualization needs identified by philosopher Sabina Leonelli (2019) as well as data reuse considerations posited by information and social scientist Christine Borgman (2019). For the sake of full disclosure, this trio of perspective articles on data and data science were featured in the inaugural volume of *Harvard Data Science Review* (HDSR), for which I have served as the founding Editor-in-Chief since 2018, an adventure that has heavily influenced my understanding and appreciation of data science. As such, I ask for the readers' indulgence regarding my frequent reference to HDSR, especially my own editorials (and with citations of my own research articles as supporting evidence). As a matter of the fact, this commentary follows the thematic style of my HDSR editorials (see https://hdsr.mitpress.mit.edu/editorscolumn[1]), wherein all articles in a single issue are threaded into a supporting web suggestive of a central theme. Because of this need to (over)fit the theme, I also ask for all authors' forgiveness for my cherry-picking assertions or nitpicking omissions in their articles. I am doing so while being especially aware that criticizing others is much easier than getting things right oneself!

As we can see, although different articles touch upon various phases of the data life cycle, they are all mostly devoted to analysis methods and results, as we normally would expect from a statistical publication. This is a simple reminder that statistics—regardless of how we overvalue or undervalue it—contributes most to the part of the data science that entails data analysis and application. Reading through all the articles, however, has reminded me that both statistics and data science would benefit if we statisticians (or anyone who analyses data) could treat data not just as input for analysis, but also as an output on its own. My main motivation to call for this enlargement of perspective is to push ourselves to be much more mindful about data quality before engaging in data analysis or modelling, that is, to conceptually reduce our reliance upon methodological remedies at the stage of analysis, which typically comes with heavy assumptions (e.g. building an imputation model to handle missing data). If we were to adopt such an enhanced perspective, we might then be able to focus more on scrutinizing data origins, histories, chain of custodies, collection mechanisms and so on, instead of consuming all our mind power almost immediately with processing and analysing data to produce analytical results. I therefore suggest that we add this phase—which perhaps can be termed *data minding*—of holistically scrutinizing data quality right before ‘Analysis & Methods’, with the intention of forming and encouraging a rhythmic and harmonic flow of data minding before—and into—data mining.

Although many of us have emphasized the aim of understanding data quality prior to data analysis, this aim cannot be accomplished fully without considering the ultimate goals of analysis

[1] Yes, MIT, and no, MIT and Harvard have not merged. Harvard University Press only publishes books. Hence I was given the opportunity to launch HDSR with MIT Press through the open-access PubPub platform (https://www.pubpub.org/).

TABLE 1 Percentage of content allocated to each phase, calculated by the generously guesstimated total number of pages on relevant discussion divided by the total number of pages of an article (excluding references and supplementary materials)

Articles and study topics	Conceptualization & collection	Pre-processing	Management & storage	Curation & provenance	Analysis & methods	Results & interpretations	Visualization & communication
Bolin et al.: Pedestrian Flows in Cities	4/23	L	L	L	13/23	2/23	4/23
Geroldinger et al.: Kidney Disease Prevalence	2/28	4/28	1/28	1/28	10/28	10/28	2/28
Iacopini et al.: Public Concerns via Social Media	1/19	L	L	L	9/19	7/19	L
Masselot et al.: Heat-health Warning System	2/32	L	L	L	11/32	8/32	4/32
Tickle et al.: Global Terrorism Incidence	3/22	L	L	1/22	11/22	3/22	3/22
Virtanen & Girolami: London Criminal Activities	3/22	L	L	L	12/22	3/22	3/22
Wright et al.: Danish Aging Health Care Need	3/18	L	L	2/28	4/18	4/18	3/18
You et al.: Global Surveillance of Emerging Diseases	1/14	4/14	L	L	4/14	3/14	2/14

Notes: The letter 'L' indicates very low allocation (e.g. none or brief mentioning). The counting of the relevant pages is neither mutually exclusive (e.g. the same discussion may count towards several phases) nor exhaustive (e.g. literature review may not be included in any of the counting).

itself. Data quality is a purpose-dependent as well as an analysis-method-dependent concept, as made clear by the construction of data defect index (Meng, 2018). However, the broader process of data minding asks for more comprehensive inquiries and deeper reflections, particularly with respect to both study-invariant characteristics (e.g. the origin of a publicly available data set) and study-dependent considerations (e.g. how a pre-processing step may bias a particular study of interest). Of course, without sufficient documentation on data conceptualization, collection, pre-processing, curation or provenance, the process of data minding would be severely limited. Currently statistical journals do not put sufficient emphasis on such documentation—Table 1 provides a glimpse into this reality. To change this situation would take serious collective effort, the most important aspect of which would be to incentivize authors to report in as much detail as possible anything that might have had a negative impact on data quality, from data conceptualization to data provenance. Presently, the authors are mostly disincentivized to do so, presumably on the grounds that such *data confessions*, so to speak, would likely lead to criticism and requests for more work from reviewers. Reporting defects could also weaken or even invalidate the conclusions of the analysis. But that is the very reason that we need these confessions.

The worry here is not 'garbage in, garbage out', to borrow a harsh but common warning. If something is recognized as garbage, most likely it will be treated as such. The real worry is 'garbage in, package out'—when garbage gets packaged nicely with decorative labels, the package can sell, at least to the uninformed. This is a part of the worry underlying all the calls for 'algorithm transparency' (e.g. Rudin et al., 2020), though here 'data transparency' is just as important, if not even more so. We therefore need to be collectively creative in finding ways to incentivize authors to provide more data confessions. One inspiration comes from the medical profession, where there is a well-established culture dictating the disclosure of potential side effects. There is also a general understanding that the more potent a medication, the more serious side effects it may produce. But such trade-offs are necessary when treating severe diseases until a better trade-off is found. Indeed, the need to reduce side effects while not decreasing the treatment potency is a driving force for medical science, just as the quest for better data quality while not reducing problem complexity should be a driving force for statistics and more broadly data science.

3 | MY DATA CONFESSION

To practice what I just preached, here are my data confessions for what is reported in Table 1. Although what I intended to assess is not hard to conceptualize, it is almost impossible to measure precisely. For example, the boundary of data pre-processing and data analysis is not always easy to draw. The boundary is clearer for traditional statistical estimation and hypothesis testing than it is for 'messier' problems such as the global surveillance of emerging disease reported in You et al. (2021). As we all are still in a pandemic, I surmise that few would disagree about the importance of such surveillance systems. However, disagreements likely exist on how to classify the step of text processing for extracting keywords. One may argue that it is a data pre-processing step since it turns the row data (texts) into data for analysis (keywords). But because these keywords effectively serve as 'minimum sufficient statistics' for the subsequent word co-occurrence analysis, one could also argue that it is a data analysis step since keyword extraction is a critical modelling strategy to achieve data reduction. Whereas this distinction may only be of academic interest with respect to the actual application, for my data collection purposes, it creates an ambiguity that requires a judgement call to proceed. I decided to count it towards both, because it reflects the fundamental problem of separating 'pre-processing' and 'analysis' since the decisions made in the data pre-processing step typically have serious analytical and substantive implications, intentional or not (Blocker & Meng, 2013). The need for such a judgement call is also a reminder of the importance of placing a greater emphasis in our research and teaching on data conceptualization because it can directly affect both the meaning of data and their values: the identical numerical value '4/14' in the two entries for You et al.'s article in Table 1 is not a coincidence, but a result of my data collection choice.

However, this conceptual difficulty is not the only data confession that I need to make. There are at least two more. Take '4/14' as an example. To measure the percentage of an article devoted to a particular phase, I need a unit of measurement. Should that be the number of words, lines, pages, sentences or paragraphs? I choose 'page' as the unit for practicality, given the time constraint I am under, and my inability of performing text mining beyond glancing and guesstimating. Readers therefore should take the number '4' with just as many grains of salts, though I have tried to err on the overly salty side to reflect my intention of not taking any details lightly. As for the number '14', since that is the number of pages for the text of the article, that should be much less salty, correct? Unfortunately, life is always more complicated when we get our hands dirty. The articles I accessed were author-generated, with a variety of font sizes, spacing, margins, etc. Therefore, the meaning of page as a metric varies with the articles. Whereas the ratio measure '4/14' makes my data less sensitive to this variation, it does not eliminate the sensitivity, especially for those articles with many tables and figures.

Speaking of figures, I must make yet another confession: I have virtually paired 'data visualizations' with presenting figures and plots, a pairing that is likely as much of an intellectual insult to experts in data visualization (Unwin, 2020) as pairing programming with computer science would be to my CS colleagues. But that is what happens in reality: data are a human product, not a natural product, and hence they come with all the insights and oversights of those who create and process them. Thus, the data quality is governed by the creators' and processors' constraints and limitations, whether worldly or intellectual. However, data, being the victim of all manner of mistreatments, typically are too weak to speak for themselves about the sources of their suffering, even when the bruises and scars are visible. Incentivizing data confession is therefore a small but necessary step in fostering a cultural shift in statistical journals to give data quality attention reasonably comparable to that given to data analysis methods and theory.

'Wait a minute, Xiao-Li, does all this data-quality talk change your assertion that statistics is mostly concentrated on data analysis, and hence cannot be viewed as a data science equal?' Fair enough. The data defects in Table 1 did not change my assertion, but that is precisely because I am aware of what the defects are, and hence can make an informed judgement of their impact. And my data confessions would allow readers to make their informed judgement on how trustworthy my assertions are, and how useful these data would be if they ever wanted to use them for their studies. Without sufficient data confessions, judgements of the impact of data quality (or lack thereof) on our findings may suffer from serious quality issues themselves. Below I will report some difficulties I encountered in this regard while reading articles in this special issue.

But before I proceed, I want first to thank all authors for their tremendous contributions to the task of addressing a wide range of societal problems, and for showcasing the power of harvesting information from data. My discussions aim to further enhance such power and to reduce its misuse by pushing ourselves to set a higher bar in scrutinizing data quality and contemplating its impact on our findings. This higher bar is for all of us, as authors, reviewers, editors and, most importantly, readers, since ultimately that is the largest community through which cultural changes take place.

I of course also invite everyone, from the article authors to their readers, to scrutinize the quality of my discussions, as I am certain that they do not meet the high standards to which I aspire, not because I did not try hard enough, but because any individual's effort and perspective is always limited. Indeed, it is not unlikely that some of my difficulties are due to my misreading or missed-reading of relevant contents in the articles at hand, for which I apologize to the corresponding authors in advance. Nevertheless, I hope the selective data minding exercises reported below might serve the purpose of 'casting a brick to attract jade'[2], but of course without hurting anyone in doing so.

4 | MINDING DATA HISTORIES

A vivid example of needing deeper data minding is provided by Tickle et al.'s (2021) study on detecting changes in temporal trends for global terrorism events. Time trend analyses are particularly vulnerable to changes in data collection protocols or other confounding changes over time. A well-known example is the change of the definition of AIDS diagnosis in 1987 (CDC, 1987), which created complications for estimating AIDS survival rates (Tu et al., 1993), among other time-dependent studies. I therefore was delighted when I saw Section 2's title 'An Introduction to the Global Terrorism Database'. However, what was revealed in the section made me want more. It stated that the database recorded terrorist events since 1 January 1970, and that during 1970–1997 a number of researchers were employed 'to collate and record events from numerous domestic and foreign reports'. I naturally wondered how 'terrorist events' were defined back then and if the same definition has been adopted consistently over the time, especially considering that the 'terrorist' designation is both country-dependent and time-dependent. What I found is the statement that 'during this period, an event which came to the attention of the compilers was valid for inclusion if it involved "threatened or actual use of illegal force and violence to attain a political, economic, religious or social goal through fear, coercion or intimidation"'.

I fully understand and appreciate the difficulties in collecting such data and the need for hard judgement calls. But this is exactly where the documentation of data collection and data provenance is especially critical. What does 'came to the attention' mean operationally? Were the researchers given unified training or a set of guidelines? If so, did the training protocol or guidelines change significantly over the relevant period of time? Was there a validation or a second-opinion process in place? What were the major credentials and expertise of these researchers? How diverse were their cultural backgrounds or ideologies, and did the range of diversity change over the relevant period of time because of the change of researchers over time? Cultural background and ideological orientations can strongly influence researchers' judgements, especially if there were no strong uniform guidelines in place. What is legal in one country may not be legal in another, and what is considered to be 'fear, coercion or intimidation' by one person may be deemed acceptable by another. As a simple but telling example, there have been several reported cases in the United States where diners at restaurants were confronted by owners or other diners because of their political affiliations or views. Yet depending on which media was reporting such incidences, the confrontations in question were variously characterized either as acts of intimidation or of peaceful protest.

As most of us understand, without knowing how stable a given data collection practice or the inclusion criterion has been, it is difficult to dissect the significance of detected changes. Indeed,

[2] I wish there was an eloquent English counterpart for this awkwardly translated Chinese idiom which means roughly to offer something of lesser value in order to attract something of greater value.

this point is discussed later in Section 6 of Tickle et al. (2021), where it reports that among the very few detected changes, one corresponds to a change of data collection in January of 1998, when it started the practice of only including events retrospectively confirmed by two groups studying terrorism with the proviso that such events must meet at least five of six pre-set criteria. This is a great example of how the known data histories helped to provide a more reasonable interpretation of the analysis results, and I am glad to see that this information was available to the authors and subsequently to all of us. My only suggestion here, mostly to the journal editors, is to encourage the presentation of such data quality information before the presentation of analysis, even though in this case it is nice to see that the authors' method was able to detect this change in the data collection protocol.

A similar data minding exercise is needed for another study involving temporal data, that is, the study by Masselot et al. (2021) on seeking a data-driven threshold to trigger heat wave warnings for senior citizens. I very much enjoyed reading the article and appreciated that it carried out a simulation study to compare the performances of four methods, which is itself a form of 'method minding'. Like the previous article, however, its description on the real data used made me want more. The article states that 'We consider daily data ranging from May to September of years 1990 to 2004 of several administrative health regions around the city of Montreal, Canada'. But there is no description on how the data were collected, especially with respect to the question of whether there was a consistent protocol over the years, for concerns similar to those stated above. But for the purpose of this study, which investigates how heat waves are correlated with the excessive deaths, the information that we need most pertains to the simple remark that the health outcome used in the study is 'the daily count of all cause of deaths in the Montreal region'.

It is unclear from this description alone if 'daily count' means all deaths *occurred* on the day or were *reported* to the city on that day. A heat wave usually lasts no more than a week[3]; yet the reporting delay can easily be on the same order if not more. It should be a simple matter to check (and then document in the article) if Montreal back corrects for the reporting delay in its released daily total, but one should not automatically assume that the corrections are made. For example, for the COVID-19 vaccination counts in the United States, the reporting delay to the Center for Disease Control and Prevention (CDC) is about 5 days. CDC constantly makes the back corrections for released daily counts for the United States as a whole, but not at the state levels (see Bradly et al., 2021). For assessing how a state is doing in providing vaccination, a few days of delays in reporting may not matter that much. But for the correlation that this study is seeking between the heat wave and the excess death counts, a few days of reporting delay might just inject enough noise to significantly weaken the signal that the study was seeking, even with the protection of the predictors being moving averages of daily temperatures. It is not an exceedingly hard problem to address even if the data are uncorrected (e.g. via modelling the reporting delay, as in Bouman et al., 2005), as long as it is a recognized issue. A simple data minding exercise can go a long way to remind ourselves of the importance of such issues.

5 | MINDING MISSING DATA

A third study (Wright et al. 2021) involving temporal data is on the prediction of health care needs for seniors in Denmark, which has many well-established nation-wide registers. The study

[3] For example, a heat wave with mean temperature over 38°C and lasting more than a week occurs about once per century in Chicago; see https://www.nature.com/articles/s41598-019-50643-w/.

used eight such registers, from education registration to death registration. The article provides a brief summary of each register, which is all currently expected by most statistical journals. To encourage statistical communities to pay more attention to all aspects of data, I would again suggest that statistical journals ask the authors to also provide a summary of how they actually processed such multi-source data. Data linkage usually is not an easy task (e.g. Christen, 2019), though I would surmise that the unique identifier system for Danish citizens has made the issue much easier to handle than, say, in the United States. Nevertheless, life is never simple with multiple databases, especially when they were established at different times, which is the case for the eight Danish registers, established over a period of four decades. At minimum, this creates the problem of different individuals with different missing data patterns in terms of their registered histories.

Varied missing patterns are no longer a grand challenge for statistical analysis, as long as we understand the mechanisms responsible for such patterns, since we have developed an array of methods for dealing with general missing data patterns (e.g. Little & Rubin, 2019). However, they do pose conceptual or computational challenges for those who are unaccustomed to them or only have access to methods for regularly patterned data. A popular workaround is to exclude data that do not fit the regular patterns, the so-called 'complete case analysis', which is known to sacrifice statistical efficiency or validity, and typically often end up doing both (e.g. Meng, 2012). Indeed, this article deals with a left censoring problem by 'limiting the length of the exposure time window w to be short enough such that the registered history of all contributing persons is equally long', where w indexes the duration before the time horizon (for prediction) during which the registered history is collected for modelling the prediction. Such forced uniformity selects a particular sub-population of individuals. I am glad to see that the article provides a clear example to illustrate this selection: immigrants are excluded if they immigrated to Denmark during the time window w before the time horizon, so too is anyone who emigrated out of Denmark during the window even if they immigrated back before the time horizon, because either case does not permit the calculation of the full w-histories. Ideally, I would wish, however, to be informed about how many cases were excluded, and to see some discussion of potential selection bias and its impact on the findings. One may wonder, for example, if this exclusion can result in underestimating the health care need, because those who immigrated back may have done so partly because of the Danish health care system. It is entirely possible that in this case the selection bias may not be much of a concern, because, for example, the number affected is rather small. If so, such a reason still should be stated, to discourage less experienced readers from blindly adopting complete case analysis only because of its apparent endorsement by a top statistical journal without any warning.

Among the methods used for handling general missing data patterns, a popular one is multiple imputation (MI, Rubin, 2004), which was adopted by Geroldinger et al. (2021) in their study on kidney disease prevalence among diabetic patients in Australia. The study's goal to combine hospital data with laboratory data is clearly a sensible one, but it is also a very ambitious one because the lab data were available for no more than 1.2% of the patients. Worse, those who have the lab data are shown to be not representative of either the prevalence population or the hospitalized population, and we all should thank the authors for providing such an explicit warning. All of this means that the validity of imputing over 98% missing lab data depends heavily on the model assumptions made. Whereas much has been written on checking imputation models, there is a deeper data minding exercise about which much more needs to be written. That is, there is an issue of 'uncongeniality' (Meng, 1994) when imputation is treated as a data processing step, and analysis (e.g. prevalence estimations) is thus carried out separately after the imputation is done, which is the case for MI. It is known that even if both the imputation model and analysis approach are valid on their own, if they are not derived/derivable from a coherent joint model (i.e. they are uncongenial), then the resulting MI variance estimation via Rubin's rule is not guaranteed to be valid (e.g. Xie & Meng, 2017). As far as I can tell, the authors' setup is an uncongenial case (and I should add that, indeed, 'congeniality' is a rather stringent requirement, and hence difficult to achieve in practice). A practical way of examining the consequences of this uncongeniality is to run simulations to empirically observe how far off is the coverage probability of the MI intervals. Whereas such a data minding exercise is certainly a more advanced one, there is a simple and practical remedy to ensure at least the nominal coverage, that is, to double Rubin's variance estimator (Xie & Meng, 2017).

Another more advanced data minding exercise pertains to Virtanen and Girolami's (2021) spatial–temporal joint modelling of criminal activities in London. I cannot agree more with their emphasis on modelling crime in multiple categories, which is statistically more principled and more efficient than analysing each category separately. I also appreciate that their 'Data' section is longer than the 'Model' section, a practice that I would love to see more in statistical publications. However, I wonder if there was a missed opportunity here with respect to their preference for collecting data from 'a single police force to avoid data duplication and management issues'. I fully understand that logistically it is much easier to focus on a single data source—I would have done the same or even less, being as I am a statistician incapable of managing almost any database. But this is where the broader data science community can help in a substantive way, since the duplications are likely sources of information useful for addressing the underreporting in crime activities, as inspired by the dual-system estimation for census undercount (e.g. Zaslavsky, 1993). It may even be the case that the missing at random assumption has a better chance to hold with the joint modelling of multi-category crimes, since the under-reporting mechanism for one category might be partially predictable from observed crime activities in a different category.

6 | MINDING DATA CONCEPTUALIZATION AND COLLECTION

Like the issue of dealing with missing data, data conceptualization and data collection are areas in which we statisticians have made significant contributions, from questionnaire designs to design of observational studies (e.g. Rosenbaum, 2010). However, the advances of digital technology and digital media are pushing us to do more, or at least to apply what we already have developed to newer areas and help others to do the same. The study by Bolin et al. (2021) on pedestrian counts on streets in Amsterdam, London and Stockholm was made possible by tracing Wi-Fi signals of unique mobile phones at both street crossings (of a selected location segment). Whereas this is clearly a viable approach, I cannot help wondering about the relationship between the pedestrian counts, which are the stated estimand, and the traced Wi-Fi signal counts, which are the estimators. The article briefly mentioned the issue of removing 'noise' (e.g. signals from Wi-Fi printers) and 'scaling' based on manual counts, which I assume is an attempt to reduce bias, such as the one due to pedestrians carrying multiple phones. But since the relationship between the Wi-Fi signal counts and pedestrian counts is critical for all the subsequent analysis, I would certainly like to have information on how the scaling was done.

For example, if the scaling was simply to bring the counts to the manual counts (presumably regarding the actual pedestrians), then it is unclear why the effort of tracing Wi-Fi signals is necessary. Furthermore, if scaling factors are somehow estimated by aggregating over different streets, then one needs to be rather careful not to mix apples with oranges. For example, I carry two cell phones, one personal and one for business. I assume for streets in business districts, there would be more pedestrians like me, and hence that the number of Wi-Fi signals would tend to be larger than the number of pedestrians. However, when I travel internationally with my family as tourists, we almost always leave at least one cell phone in the hotel for multiple reasons (e.g. avoiding costly spam calls; taking turns to charge battery because we only have one plug adaptor), and I assume that we cannot patent such behaviours. The scaling factors in tourist areas therefore are more nuanced, since for some groups of pedestrians, it could be less than one. Since a major purpose of this study is to enhance urban planning, it seems important to understand how pedestrians' phone carrying behaviours vary in different areas. Of course, all these considerations may be already made, but just not reported. But one consideration still warrants a question—namely, if the data contain both Wi-Fi signal counts and manual counts, why not model them jointly, instead of relying on scaling?

The digital age has brought us not only new technologies to collect data, but also an astronomically large amount of data, especially via social media. A serious challenge for the data science community is how to deal with the double whammy of a greater deterioration in data quality (because social media data are biased by the self-opt in design), coupled with a sharply increased sense of false confidence by virtue of their sheer volume. A most recent example is reported in Bradley et al. (2021), where it is demonstrated that, compared to the benchmark provided by the CDC on the rate of vaccination in the United States, the worst estimates came from the largest survey, consisting of about a quarter of a million Facebook users. This survey overestimated the vaccination by 17 percentage points in May 2021, whereas the spot-on estimates were from a knowledge panel of about 1000 people, which was carefully managed by Axios-Ipsos. This is a case of *Big Data Paradox*, meaning that the bigger the data, the more surely we fool ourselves (Meng, 2018). In this case, the enormous apparent survey sizes led to vanishingly narrow confidence intervals (as traditionally calculated), ensuring that we never got close to the actual vaccination rates. Indeed, the Facebook survey of 250,000 responses has the same mean-squared error (for estimating the national rate) as would a simple random sample of no more than 200 responses or even ten responses.

With these issues in mind, I naturally wondered about the quality of social media data when I started to read the article by Iacopini et al. (2021), which uses Twitter data to study public concerns about country-risk and their effect on financial markets. In almost all aspects, this is a tremendous article, ranging from modelling to computation and, beyond that, to data collection, description, processing, management and analysis. Indeed, the supplementary material was almost twice as long as the main article, providing much documentation of and for a meaningful data minding process. For example, I particularly appreciated the italicized remark on the adoption of the term 'risk' as the search phrase instead of alternatives, though I wish that the journal had given the authors more space to highlight such an important discussion in the main text, instead of the supplementary materials, which many readers may overlook.

My strongest wish, however, is for more space in which to conduct a careful discussion on the relationship between the 'Twitter public' and the 'general public', for the purpose of studying the financial markets and in general. A simple data minding exercise would be to investigate how these two publics differ. The mentioned Facebook study found that the percentage of people with no more than high school degrees in the sample was only about 50% of the rate in the US population according to the census, suggesting that the 'Facebook public' is considerably more educated than the general public. It would certainly be useful to see if similar differences exist for the 'Twitter public', and if the difference varies from country to country. A deeper data minding exercise, however, would be to investigate as to whether the 'Twitter public' better represents the 'investing public' than does the 'general public', given that, ultimately, what affects the financial markets are those who actively invest, whether on behalf of institutions, clients or themselves. One may argue, for example, that people with lower educational levels may have less impact on financial markets because they are less likely to participate actively in investment decisions. Of course, such stereotyping might be wrong, especially with multiple countries involved. But this is exactly how deeper data minding might help to generate new research questions, and perhaps along the way enable us to better answer the original ones, if only because the answers to the newer questions may provide deeper insight into what our data can and cannot tell us.

7 | LET US PROUDLY CALL OURSELVES A DATA SCIENTIST, THAT IS, WHEN WE EARN IT…

As members of a leading discipline in data science, we statisticians surely can all call ourselves data scientists, just as physicists can call themselves scientists. However, at least in the coming decade or two, this term will carry an expectation that we statisticians still need to work hard to attain a more diverse and dynamic professional persona than has traditionally been connoted by the terms 'statistics' or 'statistician'. Some of us may consider such an expectation or rather perception unfair, considering the impact that statistics has had across the spectrum of human inquires and on our global societies before the term 'Data Science' was even coined. Having been a professional statistician for 30 years, I certainly share this frustration. However, the last 3 years of my experience as Editor-in-Chief of HDSR has revealed to me that that such frustrations are shared by almost every (major) contributing discipline in the ecosystem of data science, from signal processing to operations research (OR) and to information science. For example, few informed minds would not acknowledge the tremendous contributions made by the OR community to optimization methods, which are the workhorse of much of the machine learning algorithms. Yet how many people have (ever?) seen a data science Venn diagram that mention OR in some way?

The lesson I have learned along the way is perhaps best captured by a remark that one commonly sees in fine print after a financial firm has demonstrated how wonderfully its products have performed: past performance is no guarantee of future results. We statisticians have done a lot, but so have scholars from other disciplines, who are just as brilliant and perceptive as we are—not just about subjects of their expertise, but also about *data*. Just to demonstrate my point, here are the section titles of an article in HDSR published in Fall of 2019:

- *Data Don't Just Emerge*: *They Must Be Created*
- *Data Are Physical Objects*, *Subject to Friction and Corruption*
- *The Stability and Reliability of Data Is an Accomplishment*, *Not a Natural State*
- *Data Aren't Insulated from Questions of Judgment and Expertise*, *but Intertwined With Them*
- *New Data Analytics Might Offer Qualitatively Different Discoveries*, *but Are Often Built on the Same Structure as Existing Technologies*
- *Even as the Human Sciences Are Becoming Data Sciences*, *the Data Sciences Remain Inescapably Human Sciences*

Can you guess the disciplinary identity of the author, without looking it up? Regardless, I surely invite readers to enjoy this article, as it contains many points that (implicitly) echo the call for deep data minding, as one can tell from these section titles.

My main point simply is that when we statisticians (or indeed any scholarly group) actively expand our investigative angles while building upon our disciplinary strength, we can achieve a broader impact and higher disciplinary standing, with less time for (self-consuming) frustration. The process of data minding is designed with that in mind, since it pushes us to engage in broader inquiries beyond the constraints of tradition, while at the same time taking advantage of our traditional emphasis and practice of thinking critically about data collection mechanisms at all stages.

Undoubtedly data minding is a demanding process to do well but it is essential for ensuring and enhancing scientific reliability, especially for complex problems with many confounding factors and many traps for overfitting or self-fulfilling prophecies. It would not be an easy one for those of us---statisticians or not—who have gotten too comfortable with what we are accustomed to do. But there is a great hope in the rising generations, both because of the complexity of the problems they face and their desire to surpass preceding generations. At the risk of relying on n=1, I invite readers to dive into Chan (2021): 'Combining Statistical, Physical, and Historical Evidence to Improve Historical Sea-Surface Temperature Records', a data science journey of a Ph.D. student in Earth and planetary science. I would not want to ruin the readers' fun in learning about this fascinating journey, but it suffices to indicate the intense work that was involved in carrying out the data minding process to note that the author had to learn Japanese.

In a nutshell, I am not suggesting that statisticians—or anyone for that matter—need become an expert in every domain of data science. That is neither a practical nor a healthy aspiration, as it amounts to seeking a 'data unicorn' (e.g. Davenport, 2020). But I do suggest that we use our expertise in everything data science, starting with, well, *data*. The expanded data pursuit is intellectually much more demanding than its narrower counterpart, but it is also profoundly more satisfying, as it essentially entails being the Sherlock Holmes of the digital world. We have missed the opportunity to lead in data mining (and some may argue that we never wanted to do so), but we surely can lead in data minding. I would even venture to suggest that this is one of the areas where the future of statistical leadership in data science lies.

ACKNOWLEDGEMENTS

I am deeply grateful to the editors of the special issue for giving me this opportunity to venture outside of HDSR to write an HDSR editorial style commentary, and for their infinite patience and flexibility. It is both a rewarding and challenging experience, since unlike the articles in HDSR, I do not have 'data provenance' on them (e.g. I do not know if any important omission—in my judgement—was due to some editorial requests, e.g. to save space). I thank Christine Borgman, Sabina Leonellie, and Jeannette Wing for educating me about the broader data science perspectives. I am also grateful to Joe Blitzstein, Radu Craiu, Aleksandrina Goeva, Robin Gong and Suzanne Smith for very helpful comments and editing, especially under 24 h. Of course, all remaining Chinglish and errors are mine (even though I really do not want them).

REFERENCES

Blocker, A.W. & Meng, X.L. (2013) The potential and perils of preprocessing: building new foundations. *Bernoulli*, 19(4), 1176–1211.

Bolin, D., Verendel, V., Berghauser Pont, M., Stavroulaki, I., Ivarsson, O. & Håkansson, E. (2021) Functional ANOVA modelling of pedestrian counts on streets in three European cities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–23. https://doi.org/10.1111/rssa.12646

Borgman, C.L. (2019) The lives and after lives of data. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.9a36bdb6

Bouman, P., Dukic, V. & Meng, X.L. (2005) A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statistica Sinica*, 15(2), 325–357.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. & Flaxman, S. (2021) Unrepresentative Big Surveys Significantly Overestimates COVID-19 Vaccination in the US. *Nature*, to appear.

Centers for Disease Control. (1987) Revision of the CDC surveillance case definition for acquired immunodeficiency syndrome.

Chan, D. (2021) Combining statistical, physical, and historical evidence to improve historical sea-surface temperature records. *Harvard Data Science Review*, 3(1). https://doi.org/10.1162/99608f92.edcee38f

Christen, P. (2019) Data linkage: the big picture. *Harvard Data Science Review*, 1(2). https://doi.org/10.1162/99608f92.84deb5c4.

Davenport, T. (2020) Beyond unicorns: Educating, classifying, and certifying business data scientists. *Harvard Data Science Review*, 2(2). https://doi.org/10.1162/99608f92.55546b4a

Geroldinger, A., Hronsky, M., Endel, F., Endel, G., Oberbauer, R. & Heinze, G. (2021) Estimation of the prevalence of chronic kidney disease in people with diabetes by combining information from multiple routine data collections. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–23. https://doi.org/10.1111/rssa.12682

Iacopini & M., Santagiustina, C.R. (2021) Filtering the intensity of public concern from social media count data with jumps. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–20.https://doi.org/10.1111/rssa.12704

Leonelli, S. (2019) Data governance is key to interpretation: reconceptualizing data in data science. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.17405bb6

Little, R.J. & Rubin, D.B. (2019) *Statistical analysis with missing data*. (Vol. 793), Hoboken: John Wiley & Sons.

Masselot, P., Chebana, F., Campagna, C., Lavigne, É., Ouarda, T.B.M.J. & Gosselin, P. (2021) Machine learning approaches to identify thresholds in a heat-health warning system context. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–21. https://doi.org/10.1111/rssa.12745

Meng, X.L. (1994) Multiple-imputation inferences with uncongenial sources of input (with Discussions). *Statistical Science*, 538–558.

Meng, X.L. (2012) You want me to analyze data I don't have? Are you insane? *Shanghai Archives of Psychiatry*, 24(5), 297.

Meng, X.L. (2018) Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2), 685–726.

Meng, X.-L. (2019) Data science: an artificial ecosystem. *Harvard Data Science Review*, 1(1), https://doi.org/10.1162/99608f92.ba20f892

Meng, X.-L. (2020) Reproducibility, replicability, and reliability. *Harvard Data Science Review*, 2(4). https://doi.org/10.1162/99608f92.dbfce7f9

Rosenbaum, P.R. (2010) *Design of observational studies*. 10, New York: Springer.

Rubin, D.B. (2004) *Multiple imputation for nonresponse in surveys*. (Vol. 81), Hoboken: John Wiley & Sons.

Rudin, C., Wang, C. & Coker, B. (2020) The age of secrecy and unfairness in recidivism prediction (with Discussions). *Harvard Data Science Review*, 2(1). https://doi.org/10.1162/99608f92.6ed64b30

Tickle, S.O., Eckley, I.A. & Fearnhead, P. (2021) A computationally efficient, high-dimensional multiple change-point procedure with application to global terrorism incidence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–23. https://doi.org/10.1111/rssa.12695

Tu, X.M., Meng, X.L. & Pagano, M. (1993) The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association*, 88(421), 26–36.

Unwin, A. (2020) Why is data visualization important? What is important in data visualization? *Harvard Data Science Review*, 2(1). https://doi.org/10.1162/99608f92.8ae4d525

Virtanen, S. & Girolami, M. (2021) Spatio-temporal mixed membership models for criminal activity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–25. https://doi.org/10.1111/rssa.12642

Wing, J.M. (2019) The data life cycle. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.e26845b4

Wright, M.N., Kusumastuti, S., Mortensen, L.H., Westendorp, R.G. & Gerds, T.A. (2021) Personalised need of care in an ageing society: The making of a prediction tool based on register data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–21. https://doi.org/10.1111/rssa.12644

Xie, X. & Meng, X.L. (2017) Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? (with Discussions), *Statistica Sinica*, 1485–1545.

You, J., Expert, P. & Costelloe, C. (2021) Using text mining to track outbreak trends in global surveillance of emerging diseases: ProMED-mail. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–15. https://doi.org/10.1111/rssa.12721

Zaslavsky, A.M. (1993) Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88(423), 1092–1105.

How to cite this article: Meng, X.-L. (2021) Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 00, 1–15. https://doi.org/10.1111/rssa.12762