Journal of Forensic and Legal Medicine xxx (2016) 1–5

Contents lists available at ScienceDirect

Journal of Forensic and Legal Medicine

journal homepage: www.elsevier.com/locate/jflm

Big data uncertainties

Pierre-André G. Maugis

Department of Statistical Science, University College London, United Kingdom

ARTICLE INFO

Article history:
Received 15 June 2016
Received in revised form
8 September 2016
Accepted 9 September 2016
Available online xxx

Keywords:
Big data
Statistical inference
Selection bias
Sparse dependence

ABSTRACT

Big data—the idea that an always-larger volume of information is being constantly recorded—suggests that new problems can now be subjected to scientific scrutiny. However, can classical statistical methods be used directly on big data? We analyze the problem by looking at two known pitfalls of big datasets. First, that they are biased, in the sense that they do not offer a complete view of the populations under consideration. Second, that they present a weak but pervasive level of dependence between all their components. In both cases we observe that the uncertainty of the conclusion obtained by statistical methods is increased when used on big data, either because of a systematic error (bias), or because of a larger degree of randomness (increased variance). We argue that the key challenge raised by big data is not only how to use big data to tackle new problems, but to develop tools and methods able to rigorously articulate the new risks therein.

1. Introduction

Data offers a quantified outlook on the world we all live in. As scientists, we strive to best articulate and extract important and defensible conclusions from this outlook. Big data—the idea that an always larger volume of data is being constantly recorded—suggests that new problems can now be subjected to scientific scrutiny. However, the variety of ways big data is collected raises issues. These issues are made concrete by the observation that scientists, starting from the exact same very large dataset, and progressing through rigorous statistical arguments, may reach contradictory conclusions. Examples, unfortunately, abound: for instance, the same drugs were found to be successively efficient and inefficient based on the same dataset.[1–4]

It is because data is a quantified outlook on the world that we assume it is an impartial one. Yet, as stressed again recently by the American Statistical Association,[5] "proper inference requires full reporting and transparency". The core of the message is that using statistical methods, however rigorously, does not exempt from providing a full account of how the data was collected, and of the arguments that led from the data to the conclusions. Concretely, it means that statistical methods are not automated oracles, to which one may input data and receive as output a true conclusion. Statistical methods are but tools to be used, as part of a discussion, by experts to justify their conclusion. Using forensic science terminology, "forensic reasoning"[6] cannot systematize "individualization".[7]

When using statistical methods to support and justify a conclusion, the key concept that requires transparency is uncertainty: how confident can one be about the conclusion reached by the statistical analysis? what are the main weaknesses of the arguments going from the data to the conclusion? are the assumptions underpinning the analysis credible? It is tempting to reduce uncertainty to a single number, quantifying how likely is the reached conclusion to be true given the observed data; or conversely quantifying how risky it is to trust the proposed conclusion. However, to produce any such number, assumption will have to be made, so that the uncertainty of these assumptions would have to be quantified too, leading to a circular argument. Therefore, a single number cannot capture the complexity and subtlety of the uncertainty in a conclusion reached by statistical methods. Instead, the uncertainty of the conclusion should rather be discussed as transparently as possible.

Statistical science, as a field, provides many methods and tools to dissect, articulate and quantify uncertainty. Ideally, these methods and tools—be they frequentist or Bayesian—should enable researchers to provide a full report, intelligible by all, discussing the uncertainty of their conclusion. Nonetheless, especially in the case of big data, building such a report is not an easy task: the interplay between the assumptions made, the technicality of the methods used, and the need to produce an accessible yet precise summary, makes it a very hard exercise even for the best of us. For

E-mail address: p.maugis@ucl.ac.uk.

http://dx.doi.org/10.1016/j.jflm.2016.09.005
1752-928X/

intance, a long discussion has recently taken place in the pages of the Proceedings of the National Academy of Sciences[8–14] debating the validity of the statistical arguments used to suport the claim that hurricanes with female names are deadlier.[15] Furthermore, big data being a new type of data, statistical methods especially developed for big data are still in their infancy, and experts agree that much progress still needs to be made in the field.[16,17]

To exhibit how hard it is to fully articulate the uncertainty stemming from using big data, we will focus on two key sources of uncertainty: uncertainty from the data, and uncertainty from the model. The first pertains to how accurate, complete and clean is the data; the second to the fidelity, flexibility and power of the model. Although deeply related, the two types of uncertainties have different natures. The first requires making subtle observations on the structure of the data and how it was collected: it is tied to the more classical statistical concept of bias. The second requires making a rigorous formal treatment of the model, or framework, used, and is tied to the statistical concept of inflated variance.

More precisely we will focus on two concrete issues one has to face when using big data. The first is concerned with how the data was collected, and is technically referred to as selection bias. It asks the question of how statistical methods are affected when the observed data only represents a biased sub-part of the population. The second is concerned with models requiring weak dependence between data points. It asks the question on how uncertainty, or more precisely the variance, is increased—or possibly decreased—by the presence of this dependence. Both problems are pervasive in the era of big data, and we will give examples both of datasets and of statistical methods where this is the case.

2. Selection bias

In most contexts “big data” is the data generated by individuals, or sensors, throughout the world; in a sense, it is the byproduct of the presence of technology in our societies. Therefore, despite its volume, it might present some severe “blind-spots”; i.e., the informational content of the data might be incomplete. If the data's informational content is incomplete, the conclusions drawn from it are very uncertain. More precisely, these conclusions are biased in that they focus on only one part of the information.

One interesting case is that of using tweets to obtain live information on large scale events, specifically here the evolution of an hurricane and its consequences using live analysis of tweets.[18,19] Unfortunately, using such approaches may lead to a biased impression of where the damages occurred. As argued by K. Crawford[20] on the analysis of N. Grinberg *et al.*[18] concerning hurricane Sandy, most tweets came from Manhattan and not from “more severely affected locations, such as Breezy Point, Coney Island and Rockaway”.

The biased impression of where the damages occurred is caused not only by the relative density of population, but crucially by the relative density of effective smartphones: power outages meant that smartphones were not recharged, poorer areas had less smartphone owners, severely hit areas had poorer network reception, and so on. Therefore, few tweets came from areas where the hurricane caused power outages or bad network connections, so that severely hit areas are underrepresented in the sample. In the same fashion, poorer areas—areas that might be both more extensive and less robust to hurricanes—are also underrepresented in the sample. This means that using tweets to study the evolution of a hurricane can but produce a biased analysis of the said evolution, biased toward the smartphone rich locations.

In this case, and in many others, the bias comes not from idiosyncrasies in the data, or from issues with the statistical method used, but from the nature of data itself. In the hurricane example, the data itself is complete—the data consists of all tweets in the New-York area—the method used is rigorous, and yet it presents a biased view of the hurricane's evolution since few persons tweeted in the more severely affected areas. It is because the data appears complete that it is so hard to detect that in fact it presents a biased view of events.

Similar examples, where the data appears complete but is in fact biased, abound. This is also true in the field of forensic science, and crucially to evaluate the prevalence of crimes: the chance of a crime being reported will vary depending on many factors, such as the gender, income, religion, cultural origins of the victim and perpetrator, the type of neighborhood the crime took place in, and so on. Furthermore, different crimes will be likelier to be reported than others, even for the same victim. Therefore, the data consisting of all crimes reported to the police most probably presents a biased view of the true rate of crimes, and this bias will change depending on the crime, region or population considered.

We now clarify with an example to what extent such partial, or biased samples, distort estimates and increases the uncertainty of the conclusion of a statistical analysis. Because biased samples falsify the assumption underlying classical statistical methods, the consequence of using biased samples can be surprisingly extreme and counterintuitive. In this sense, it is educating to consider the effect of biased samples on the simple example of correlation.

Correlation has been invented more than a century ago,[21–23] and is today maybe the most used measure of linear dependence in all fields of statistics. Surprisingly, it is very weak to selection bias. To see this, let us consider three real-valued random variables X,Y and Z and call ρ_{XY} the Pearson, or linear correlation, between X and Y; recall that by definition we have:

$$\rho_{XY} := \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\cdot\mathrm{var}(Y)}}.$$

Note that we make no assumption on the dependence between the three variables X,Y and Z, and neither of them need be a mediation factor,[24] instrument or confounder[25] between the others.

Now, assume that we observe a biased sample, such that we only observe X and Y if Z falls within a given subset of its support A. In the hurricane example, X would be the effective damage in an area, Y would be the number of tweets, Z would measure the availability of a smartphone to tweet, and A would take the form $[\alpha,+\infty[$ for some smartphone availability threshold α. In a crime prevalence examples, X would be the number of crimes, Y would be the victim's income level, Z would be the victim's propensity to report a crime, and A would again take the form $[\alpha,+\infty[$ for some reporting probability threshold α. In both examples all variables are strongly interrelated.

In this setting, following the definition presented in the previous equation, the correlation observed in the available sample is

$$\rho_{XY|Z\in A} := \frac{\mathrm{cov}(X,Y \mid Z\in A)}{\sqrt{\mathrm{var}(X \mid Z\in A)\cdot\mathrm{var}(Y \mid Z\in A)}}.$$

However, $\rho_{XY|Z\in A}$ need not be equal to ρ_{XY}. It follows that computing correlation using our partial sample leads us to estimating not the correlation between X and Y (ρ_{XY}), but another parameter altogether: $\rho_{XY|Z\in A}$. Therefore, using our partial sample leads to a systematic error—or bias—of magnitude $|\rho_{XY|Z\in A} - \rho_{XY}|$. Surprisingly, even if (X,Y,Z) is a Gaussian vector, one of the simplest possible distribution and the most amenable to classical statistical methods, then the bias $|\rho_{XY|Z\in A} - \rho_{XY}|$ may be substantial. In Fig. 1 we provide an example where we let A range the deciles of Z. In that case, $\rho_{XY|Z\in A}$ varies widely depending on the decile considered, and

Please cite this article in press as: Maugis P-AG, Big data uncertainties, Journal of Forensic and Legal Medicine (2016), http://dx.doi.org/10.1016/j.jflm.2016.09.005