10/2016

Measuring Socio-Spatial Justice: From Statistics to Big Data[1] – Promises and Threats

Elisabeth Tovar

Abstract (180 words)

Whether to be praised or vilified, Big Data is often presented as the coming of a new statistical world, where causality and scientific method will be replaced by powerful statistical algorithms, able to infer probabilistic predictive models from correlations between the mass of our daily numeric tracks. In this paper, as an Economist, we discuss, within the framework of socio-spatial justice measurement, the normative pros and the cons of the 'Old World' of traditional, bottom-up and deductive statistics and those of the 'New World' of Big Data statistics, decentralised and inductive. In the end, Big Data is neither white nor black magic: a social construct, it is only another statistical tool that we can use for a better understanding of the world. As such, it can and should be scrutinised by social scientists. From the consequentialist point of view of Economics, for which measuring well-being is crucial to conceiving social justice, Big Data looks like a genuine improvement on the data scarcity from before… provided that we are able to organise, from a procedural point of view, its regulation.

Keywords: Big Data, Consequentialism, Economics, Procedural Justice, Socio-Spatial Justice

[1] A first version of this discussion was drafted for the workshop on "Liberté, égalité, computer. Gouvernementalité algorithmique and spatial justice" organised by Justice Spatiale/Spatial Justice on 28 November 2014 at Paris Ouest University. The author would like to thank the workshop organisers, as well as all contributors and participants for the contradictory but fruitful debate which led her to write this article. She would also like to thank the two anonymous reporters whose comments greatly contributed to improving this article.

[Introduction] 'Big Data' and Socio-Spatial Justice: Economist Issues

Big Data, which is sometimes introduced as one of the elements of the 4th industrial revolution (Anderson, 2012), is more often defined as the advent of a three-dimensional statistical world characterised by the famous "3Vs", i.e. the growing volume, velocity and variety of exchanged and analysed data (*cf.* the report of the META Group – now known as Gartner Group – drafted by Laney, 2001). Today, we put forward a change in paradigm in the *nature* and usage of data: while traditional data processing (including data mining) relies on deductive reasoning, Big Data supposedly indicates a transition to inductive analysis. In this new statistical world, inferential analysis based on very large quantities of decentralised low density data, makes it possible to infer models with *predictive* capacity (Delort, 2015).

"Big Data is fundamentally different from data mining, a difference that doesn't have anything to do with the volume of data but which is conceptual. A data warehouse, where data mining takes place, relies on a model. Conversely, (…) Big Data consists in *preliminarily and inductively* creating *models with predictive capacity, using masses of low density data.* (…) We are moving from facts to rules and mathematics makes it possible to measure the uncertainty weighing on these rules, depending in particular on the facts on which these rules are based". (Delort, 2015)

The emergence of the new statistical world of Big Data, has given rise to a rich and lively debate on which this special issue of *Justice Spatiale, Spatial Justice* is based.

In *Politique Étrangère*, Mayer-Schönberger positions himself from the point of view of 'digital neo-positivism' (Mosco, 2014, quoted by Ouellet *et al.*, 2014) and welcomes in *La Révolution Big Data* an evolution comparable in magnitude to replacing the Newtonian notion of absolute space and time with Einstein's relativity theory. Increasing the rationality behind decision-taking, Big Data disconnects our perception of the world from our preconceived fragile postulates and, especially, from our unrealistic need for causality. Moreover, where Big Data also gives an economic value to the saga of our lives, juxtaposed, it makes sense and reveals ways of living together we never suspected up to now. As a result, it is essential to measure its economic repercussions and organise its good governance as far as politics is concerned (Mayer-Schönberger, 2014).

By contrast, other authors have been adopting the viewpoint promulgated by surveillance studies (Ouellet *et al.*, 2014): in *Les Cahiers du Numérique*, Carmes and Noyer (2015) highlight in "*L'irrésistible montée de l'algorithmique*" the dangers of outsourcing the processing of data which is "*Too Big to Know*" (Weinberger, 2012). In the tradition of Desrosières' critique of statistical governmentality (2008a and 2008b), these arguments echo the "critique of computational reason" evoked by Bachimont (2008). At the other end of the spectrum, Anderson (2008) was being provocative when he announced the obsolescence of the scientific and deductive method contained within the limits of human thoughts, as a tool for reading and organising the world.

Big Data fundamentally challenges the Social Sciences: Can the advent of this new statistical world truly change completely our way of understanding society and of working on this understanding?

In this line of thought, we would like to contribute to the debate by linking it to the specific field of socio-spatial justice measurement, and by mobilising the tools of economic science (as a specific type of social science). Economic science turns out to be particularly eager for quantifiable data leading to the accurate mapping of the world, with a view to correcting problems related to inequalities or inefficiencies. This positivist approach has been used by many a "standard" empiricist economist, whose scientific production efforts are turned towards creating tools of measurement that are relevant, from the point of view of the social problematics under study, and for whom theory follows and must be amended to become fact-proof, this being sometimes a criticised methodological position compared to other paradigms in economic science (*cf.* in Labrousse, 2010, the critique addressed to Esther Duflot on her positivism).

As a generous producer of statistical indicators, "standard" economic science seems to represent the perfect outlet for Big Data – an additional string to the econometric bow of this field which is already well supported. Conversely, because it questions measurement, positivist economy knows full well that no data is neutral, and that any statistical indicator is only valid from the point of view of its axiomatic properties related to its measurement objective. This leads to a critical look at the – somewhat

fantasised – promise of a world of figures made intelligible, as if by magic, through the dehumanised inductive algorithmic study of Big Data.

More generally, economic science can shed light on the debate through its own analytical methods; in this line of thought, we propose mobilising, to this end, two tools specific to contemporary economic science: marginal reasoning and normative economics. With marginal reasoning, one takes an interest in the *opportunity costs* of the changes under study, and confronts the pros and cons of world countries before and after the Big Data revolution. In addition, the tools of normative economics make it possible to appreciate the opportunity of updated changes: Is the New World situation fairer and therefore more desirable than that of the Old World? What additional evolutions seem necessary to bring it closer to a truly fair situation?

On these methodological bases and as an economist, we try in this article to contribute to the debate, by confronting, within the framework of socio-spatial justice measurement, the normative pros and cons of the 'Old World' of traditional, bottom-up and deductive statistics with those of the 'New World' of decentralised and inductive Big Data statistics.

Bottom-Up and Deductive 'Old world': Quantifying to Define Fairness, Under Procedural Control

Quantifying to Define Fairness

For an economist, it is unthinkable to question the urgent necessity of quantifying reality in order to define what fairness consists in.

Jeremy Bentham, the founder of utilitarianism – a dominant normative paradigm in economics – adopts an explicitly *consequentialist* position where fairness is none other than good, understood in terms of a hedonistic utility principle based on 14 pleasures and 12 pains[2], and which a *felicific calculus* enables each one of us to

[2] BENTHAM's list of pleasure is as follows: "1. The pleasures of sense. 2. The pleasures of wealth. 3. The pleasures of skill. 4. The pleasures of amity. 5. The pleasures of a good name. 6. The pleasures of power. 7. The pleasures of piety. 8. The pleasures of benevolence. 9. The pleasures of malevolence. 10. The pleasures of memory. 11. The pleasures of imagination. 12. The pleasures of expectation. 13. The

compile in the intimacy of our being. On this basis, fairness, as defined on a societal scale, is none other than the greatest utility (i.e. the greatest happiness) for the greatest number. This leads to, in economic models, the famous maximisation of the sum of utilities as the programme of the benevolent planner.

It follows that, in economics, spelling out what fairness consists in necessarily implies measuring well-being; reciprocally, there can be no intelligible economic discourse on fairness without quantification. Quantifying social justice cannot be a wrong, and emerges as a necessary right for reflection on common good. It is therefore essential to devise and use the best normatively founded measuring tools, which explains the 'wars of indicators' that are sometimes denounced by the critics of economic science. Putting forward better measures of well-being and better indicators of its distribution among the members of a society is, in the end, making progress in the appreciation of social justice.

Therefore, reflecting on socio-spatial justice is above all, for an economist, building the most pertinent possible geolocated measures of well-being, as well as spatialised indicators of the distribution of this well-being relying on the most satisfactory possible axiomatic bases.

Data Poverty, Normative Blindness

In this context, the poverty of geolocated data available in the Old World of statistics, represented a problematic obstacle to measuring fairness, and therefore to the possibility of conceiving a public action likely to reduce the gap between the true state of society and ideal and fair society.

A first source of data typical of the Old world comes from the availability of administrative databases. Referring back to their etymological nature, these

pleasures dependent on association. 14. The pleasures of relief." The list of pains is as follows: "1. The pains of privation. 2. The pains of the senses. 3. The pains of awkwardness. 4. The pains of enmity. 5. The pains of an ill name. 6. The pains of piety. 7. The pains of benevolence. 8. The pains of malevolence. 9. The pains of the memory. 10. The pains of the imagination. 11. The pains of expectation 12. The pains dependent on association." (Bentham, 1789, chap. v, § 3).

"*statistics*" [3] (*commune*-based tax income of the Tax Authorities, administrative declarations of corporate social data gathered by social data transfer centres, number and characteristics of job seekers gathered by Pôle Emploi etc.) are built by and for the good public administration. Other data, in lesser quantities, come from national surveys conceived by the statistical system to inform public decision-makers about the reality of the territories they run (general population census, surveys on household mobility, local knowledge of the productive system etc.).

A second source of geolocated data comes from the "bottom-up" work of researchers in the social sciences, which generates the data (surveys, field studies etc.) required for pursuing their research programmes.

Whether administrative or scientific, this data production is symptomatic of the deductive nature of 'the Old world', where statistics came from the conscious and well-considered human will for governing on the one hand and understanding on the other.

These data production systems open themselves to much criticism.

Firstly, they are very costly, so much that a National Assembly report recently indicated that "*the cost of the old form of census was one of the main reasons that led to the elaboration of a new method. Indeed, the last general census, which took place in 1999, was initially planned for 1997, but was postponed for budget reasons, where the additional cost linked to its realisation could have led in particular to failing to respect the Maastricht criteria. The estimated cost of the census was in effect concentrated over one year only and therefore required considerable effort. As such, the 1999 census cost around 1,2 billion Francs, i.e. around 180 million Euros*" (Gosselin, 2008).

In the specific case of geolocated data, one also needs to take into consideration the statistical margin of error which limits **their statistical exploitation**, as far as the smaller scales or least populated spatial units are concerned.

[3] The French term 'statistique' was borrowed from the German *Statistik*, which in turn was created by German economist Achenwall (1719-1772) who derived it from the Italian *statista* or "statesman" where, to him, statistics represents all the knowledge a statesman should possess (TLFI, 2015).

Moreover, geolocated data is also **difficult to access,** due in particular to the necessary protection of private life: although the French National Institute of Statistics and Economic Studies (INSEE) is currently working on making summary data available, squared off according to a very thin grid, it is not easy for those who are not affiliated to an institutional research centre and endowed with a solid research project, to obtain personal geolocated data. Similarly, data on the location of individuals (postal address, place of birth, IP address, geolocation etc.) fall under the sensitive data category, for which social science researchers collecting and processing such data need prior authorisation from the French Data Protection Authority (CNIL).

From a normative point of view, these restrictions, which are awkward from a consequentialist viewpoint because in the end they are limiting the capacity of society to know itself, can lead to a social justice optimum if one is to adopt the rival point of view of procedural justice: their existence leads to the rights of individuals within society to be respected.

Furthermore, the main drawback of Old World statistics is that they are most often characterised by the seal of public administration requirements: they only show of society what makes sense for the needs of government intervention. The fact that geolocated data creation is subjected to the practical and political contingencies which underlie the running of the State, justifies all criticisms echoing Desrosières' "*statistical governmentality*" (2008a and 2008b).

For example, data stemming from the general population census makes it possible to evaluate, in great detail, the extent to which disadvantaged households are deprived: it is possible to know which suburbs have the lowest proportion of housing with bathroom installations, or which landlocked towns have the highest rate of non-motorised households. Symmetrically, on the other hand, it is impossible to measure wealth in fortunate households, because the menus of the census questionnaire are systematically truncated upwards: beyond a certain level of comfort, the detailed characteristics of households 'fall off the radar' of public statistics. It is possible to know that a dwelling contains more than one bathroom or more than six rooms, but nothing points to the fact that it can merely be a large apartment, a town house or a

castle with outbuildings. While public statistics make it possible to study places of relegation under every angle, places of abundance are being overlooked by public authorities and researchers alike: it is easy to identify areas of relegation characterised by poverty with a socioeconomic fate diverging from that of the rest of the country (*cf.* for example Préteceille, 2007, 2012 or Tovar, 2014), but it is impossible to properly identify, on the basis of public statistical data, the gated areas where the rich organise their secession from the rest of the country. However, from the point of view of public intervention which governs the construction of such data, there is no need for further knowledge: the collected data is sufficient to guide public policies aiming at opening up and fighting against the geographic concentration of poverty.

Through this example, we can conceive how the origin and nature of "Old World" data can influence how we perceive society. Such data being costly, its impact is conditioned by the aim governing its creation; using this data rigorously requires a deep understanding of its production process – for fear of falling into an undesirable governance of numbers. In the 'deductive' and finalist universe of the Old world, the danger is that statistical blinkers become intellectual blinkers[4].

As a result, there is a relative scarcity of available statistical indicators for measuring socio-spatial justice, while the contemporary Theories of Justice put forward complex and nuanced theoretical definitions of well-being.

Let us take another example, that of 'capabilist' well-being put forward by 1998 Economics Nobel prize-winner Amartya Sen, in contrast to the utility used by standard utilitarian economists (and very often summarised simply as the 'available household income'). The measuring standard of well-being defended by Sen is multidimensional in essence (Sen, 1993, 2010): while comprising the actual

[4] This brings to mind a joke well-known among economists: One night, a policeman sees an economist looking for something on the ground, under a street light. He proceeds to ask him whether he lost something, to which the economist replies: – "I've lost my keys in the dark alley on the other side of the street". The policeman then asks him why he is looking for his keys under the street lamp, and not in the dark alley, to which the economist replies – "Because this is where I can see better to look for my keys".

realisations of individuals measured through various functionings (income, health, education, social recognition, housing etc.), it also includes two dimensions of individual freedom: the matrix of capabilities refers to all the actual potential achievements of a person (i.e. freedom of *opportunity*), while procedural freedom (i.e. freedom of *choice*) reflects the extent to which individuals control their own destiny. Sen's theoretical subtlety makes the capabilist approach attractive in studying people's well-being, although its practical implementation is far from being easy. Where to find geo-located statistical indicators for measuring the wealth and diversity of people's actual achievements?

All things considered, if we adopt a consequentialist normative point of view – that of the economic analysis where the fair or unfair nature of a social organisation depends on its effects on the human beings making up that organisation – then, in the Old world, the production of geolocated data needed to evaluate socio-spatial justice can in no way be considered as satisfactory.

Procedural Guarantees

However, if fairness is evaluated from a procedural point of view (where what matters is the fairness of the rules governing the way the world is organised), the judgment is far more nuanced. In this regard, it is interesting to relate the argument proposed by Sen concerning the difficulties encountered with the practical implementation of his approach. Sen defends, for fear of *paternalism*, the theoretical indecision of statistical indicators required for measuring capabilist well-being. He then points out that, because statistical data cost so much to collect, the evaluator is compelled in the end to make do with information on the *available* elements of well-being, i.e. information which is necessarily that which was given the highest value by society.

In a way, because of its deliberate and hierarchical nature, the production process of Old World statistics guarantees the 'traceability' of the data produced, *a fortiori* in a democratic political system, where public intervention submits to public deliberation, under the forever critical eye of researchers.

Despite this procedural absolution, which makes it possible to put up with the current situation, we can lament the fact that, as researchers, we only have access to

data that has been reasonably developed by society, via the filter of scientific argumentation, political intervention and democratic discussion. By contrast, we could emphasise the significance of dealing with less "orthodox" data in order to explore innovative or marginal dimensions, as far as well-being is concerned, detached from any a priori end.

This is precisely what is being promised by the era of the 'New World of Big Data statistics'.

A Decentralised and Inductive 'New World': Consequentialist Promises, Procedural Questioning

Statistical Abundance at the Service of Knowledge

The emergence of Big Data opens up breathtaking perspectives for the geolocated measurement of well-being. For an economist, it represents invaluable progress for the measurement of socio-spatial inequalities and, from a consequentialist viewpoint, for better public policies.

First of all, the quantity and especially the *nature* of geolocated data, have undergone a deep change: the ocean of data making up Big Data stems from a novel decentralised production process characterised by "low-intensity" intentionality. In contrast to the well-thought-out collection of public data, the countless statistical traces left forever by our digital existence are just a click away for the Social scientist: contents of our research carried out on browsers, emails and profiles, or activities on our online family, professional, social or sentimental networks, our purchases, online bank accounts, travelling or biodata gathered via connected equipment... Because today our life is partially taking place online, the infinite memory of the Web keeps an accurate record of our tastes and political opinions, of the many entanglements of our sociability, of our productive activity, of our role of *homo economicus* trading on markets, but also of the footprint of our physical existence and, tomorrow perhaps, of our physiological intimacy. Most of the time, this accumulation of data takes place independently of our will (even if the tools enabling us to conceal our digital traces exist), even if it can also be the result of the conscious construction of our digital

identity: social and professional network profiles, blogs and online data storage service subscriptions among others.

From now on it is possible to perceive, measure and quantify like never before the (geographic) reality of our existence: because digital communications increasingly gives media coverage to the reality of our lives, they produce anarchic accumulations of uncontrolled, decentralised and 'spontaneous' data of incomparable wealth and pertinence with that of the Old World.

This "economist's dream", this "brave new world" (Shearmur, 2015) in the field of residential segregation measurement, can be illustrated as follows: two Estonian researchers recently compared the measurement of the social interactions of residents in Tallinn, stemming from census data (night segregation), with that of the residents geolocation gathered through their cell phones (day segregation) (Silm and Ahas, 2014). They were able to show that, although places of residence are clearly segregated, different ethnic groups were sharing the city during the day, with a high probability of inter-ethnic contacts.

This study, grafting knowledge stemming from 'new world' statistics to knowledge from the "Old world", opens up new promising questions for understanding segregation in more detail. Many authors (such as sociologist Préteceille, 2007 and 2014) explain that residential segregation is mainly the consequence of the economically dominant categories seeking *entre-soi*, while the working-class categories live in poor suburbs neglected by the rich. If, as shown by this study using 'new world' statistics, day segregation is lower than night segregation, does that mean that the dominant groups are less performant in their avoidance strategies when they use the urban space during the day? Does that mean that they perceive differently their day projection through the urban territory and their night static withdrawal in 'their' residential suburb? Does that mean that the dominated groups are integrated into the economic functioning of the city but are relegated as soon as sociability comes up? Admittedly, with the help of well-tried tools such as monographs and interviews, social science could shed light on carefully selected territories, groups and individuals as far as these questions are concerned. Beyond

this, in the Big Data era, we can hope for more global answers by bringing to light statistical regularities evaluated on the scale of society as a whole.

Another aspect of digital evolutions needs to be highlighted: the simplification of access to all the databases of the Old World: public statistics, databases created by government-owned and private enterprises or by isolated researchers. Recently organised by the actual Government (*cf.* www.data.gouv.fr), this data is more visible, more easily exploitable, especially by non-specialists.

Decentralisation and Democratisation of the Discourse on Statistics

We are currently progressing towards greater democratisation as far as societal knowledge is concerned. The digital era involves the diffusion of data processing technologies, especially geolocated data, as is the case with free software R (www.cran.r-project.org) and its constantly more and more sophisticated modules of cartography and statistical analysis of geolocated data. Thanks to this production of collaborative tools, the practice of statistical analysis is facilitated by the emergence of communities of users who popularise and diffuse econometric tools, methods and good practices. The production of statistical discourses is no longer the prerogative of those who know, whether public administration experts or accredited academic researchers.

On the one hand, with 'Big data' and the digital revolution, we can hope for the emancipation of data production (geolocated data in particular) vis-à-vis State supervision, and for the advent of an era of decentralised and democratic *statistical abundance*, with consequentialists welcoming the elimination of technical and political obstacles hampering knowledge on socio-spatial inequalities. By following this reasoning through, Big Data could constitute a tool for the defence of citizens were they to use such data to denounce the potential abuses of the State machinery.

Procedural Threats: Between Fantasy and Reality

On the other hand, this raises very serious procedural issues such as consenting to the diffusion of personal data, commodifying this new economic resource,

controlling data access and utilisation, protecting individual freedoms and controlling econometric analytical systems peculiar to Big Data.

We can start by highlighting the digital invisibility of those whose lives have been out of sight of the virtual 'all-social'. This is the danger of a new form of relegation, this time between the connected and those who are excluded (by choice or by lack of communication or consumption) from shared virtuality. How can those who do not use a smartphone be included in the measurement of segregation as proposed by the Estonian researchers? In this case, spatial justice issues arise between space and cyberspace. More generally, Shearmur (2015) explains that Big Data can only process codifiable and quantifiable information, but that it cannot lead to understanding Humanity without human mediatisation.

Symmetrically, there is the issue of cyberspace opacity and temporal-spatial depth. In the physical space, it is possible to break away (even temporarily) from the social world and to protect one's intimacy in private places, away from public view[5]. The unequal distribution of the possibility of "living hidden away to live happily" is a key for reading spatial justice because the privilege of invisibility is a prerogative of the dominant categories. In cyberspace, the barriers guaranteeing our intimacy appear rather thin: computers have little security; our emails and browsing habits on the Web are systematically being traced and checked; and surveillance equipment is being connected, among others. The consequence are that we live our lives under an 'eye of Sauron' which records everything that ought to remain untold: actions and opinions, dislikes and weaknesses... As a result, we could reinterpret the spatial justice imperative as necessary procedural imperviousness between intimate space and a public space fossilised by the virtual world.

A third procedural preoccupation concerns controlling this new data. The informational abundance of Big Data offers the means to Orwellian knowledge of all the spaces of our existence: that of the real world, that of the virtual world and, soon, that of the intimate world of our bodies and our minds. From a procedural viewpoint,

5 So much so that, one of the reasons given by the German statistical office, when abandoning an exhaustive census, was the multiplication of barriers (doors, caretakers, digicodes, etc.) which prevented contact between citizens and census enumerators, the statistical eyes of the State.

the question is open to say whether such a level of transparency is a bad thing in itself; from a consequentialist viewpoint, everything depends on what is done with it.

One of the elements of the debate concerns the commodification of what Mayer-Schönberger (2014) qualifies as "new economic resource": Big Data information « *can be used indefinitely to multiple and new ends, an additional value being produced with each operation. (…) [Their reutilisation] opens the way to creating new products and services, and therefore to new flows of income for companies – which could lead to their business model evolving.*" As such, the economic model of major digital actors offering "free" online services (search engines, commercial Websites, social networks, software programmes etc.) relies increasingly on gathering and selling the precious data.

The commodification of our virtual footprints is not well known by citizens and, de facto, escapes the authorities endowed with the procedural legitimacy guaranteed by universal suffrage. In this context, we could propose a new procedural ideal of spatial justice, with the possibility, for administrations representing democratic legitimacy over a territory, of imposing procedures of fair and transparent control upon entities that appear, wrongly, as being able to exist "outside", disconnected from the real world.

However, this is not a fundamental issue: although implementation is slow, in the end we can hope for the imminent democratic regulation of this new digital world. On the other hand, the nature of the statistical analysis of Big Data raises another tricky issue.

With Big Data, as seen previously, it would no longer be necessary to seek in order to find, since computer algorithms make it possible, better than human intentionality, to bring out statistical links without having recourse to any prior causal model, with the mass of available data leading to the use of probabilistic models able to *predict* events.

A frequently quoted example concerns the important socio-spatial justice issue of geographic targeting and ethnic profiling, as carried out by the police force. Used increasingly by police forces in the United States, Predictive Policing software programmes reveal, within one or two blocks of houses, the geographic location of

crimes even before they are recorded (Pearsall 2010). Resorting to these Big Data techniques is obviously carried out in the name of the consequentialist principle of better policing, although in this case procedural considerations also come into play. It is indeed also important to guarantee the respect of the 4th amendment of the American Constitution, which protects citizens from searches not based on 'probable cause'[6]. Predictive Policing objectively guides police interventions, which supposedly is an improvement on police intuition which is based on subjectivity, partiality, arbitrariness or simply on racism (Guthrie Ferguson 2012; Koss 2015).

This immediately raises many problems as far as protecting civil rights is concerned (Guthrie Ferguson 2012; Koss 2015; Sprague 2015), since used in this way Big Data gives police forces statistical legitimacy to harass the (essentially Black and Latino) populations living in the most disadvantaged – and therefore crime-encouraging – suburbs (Crawford and Schultz 2014; Sprague 2015; Barocas and Selbst 2016).

However, from a strict statistical point of view, Big Data infers no causality and offers only probabilistic inferences. The software does not in any way say that a person in possession of whatever characteristics, moving on foot in whatever American suburb, is necessarily a criminal; it only indicates that, at a given time and place, the probability of such a person committing such a type of crime is high. As such, the "dictatorship of data" does not follow from Big Data itself, but from the imperfections of the humans manipulating it: incapable of thinking probabilistically, irremediably contaminated by a visceral need for causality, the human users of Big Data use it wrongly to confirm their prejudices (Mayer-Schönberger, 2014).

"There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyse the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot. (…) The opportunity is great: (…) Correlation supersedes causation, and science can advance even without coherent models,

[6] This amendment establishes the right of citizens "to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized".

unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?" (Anderson, 2008)

This extract from Anderson's (2008) famous forum in *Wired*, illustrates pro-Big Data dogma and, at the same time, the fantasies of its critics. It offers a glimpse of a world where one would need to put an end to Science, to that human obsession for modelling and searching for causality, and let the machines do the thinking for us.

This breathtaking image of a society regulated by infallible algorithms presents a problem. Indeed, the dogma and fear of Big Data rely on two fragile beliefs: the 'naturalism' of data collected by Big Data, and the superiority of the powerful inductive correlations of Big Data on causality, the Grail of scientific method.

Yet, as recalled by Crawford, Miltner and Gray (2014), social science researchers and regulars of Old World statistics have, for a long time, established the eminently social nature of statistical data production and use: "*Raw Data is an Oxymoron*", as summarised in the title of the collective work edited by Lisa Gitelman (2013), and data collated as Big Data is no exception.

Furthermore, while Big Data leads to revealing new correlations concealed in the mass of digital data, there is nothing truly revolutionary from the viewpoint of what Statistics is, according to Cournot's 1843 definition: "*A set of techniques for mathematical interpretation applied to phenomena for which an exhaustive study of all factors is impossible*". Although Big Data can offer a probabilistic vision of the world, it seems rather hasty to deduct any kind of superiority from correlations as tools for understanding the world.

Moreover, as testified by the improbable and humorous correlations published on the Spurious-Correlations website (Vigen, 2015), pretending to be able to manage without human rationality in grasping the consequences of results offered by Big Data, as both the zealots and the pessimistic opponents of Big Data would have us believe, hardly seems serious.

"The big data team simply uncovered better, more meaningful correlations. (…)" Big data analysis can be about correlations OR causation—it all depends, as it has always been, on what question we are asking, what problem we are solving, and what goal we are trying to achieve. I don't think big data will do anything to—and has little to do with—our obsession with causation. But as Big Data successfully demonstrates, this is one technology-driven

phenomenon that can improve our lives and require all of us to pay attention and start engaging in a meaningful conversation of what to do about its potential risks." (Press, 2013)

Algorithms are not some new digital divinity; they are social objects, built and used by humans for specific reasons. Far from subscribing to the dystopic spectre of Philip K. Dick's sci-fi novel *Minority Report* (1956), we can highlight the intentional neutrality of what, in the end, is a mere tool of statistical analysis. From a consequentialist point of view, Big Data, as the potential weapon of oppressors, can just as well be a tool of liberation in the hands of the oppressed. Confronted with the abuses of 'Predictive Policing', the American civil society has developed police monitoring software programmes such as Cop Watch, and uses Big Data methods to predict the irrational usage of force by the various departments of the municipal police force.

[Conclusion] What to Do? Working Programme for Economists

Big Data is neither black nor white magic: it is merely a new statistical tool and, as such, a social construct with limited impact which can and should be scrutinised by social scientists:

"This points to the next frontier: how to address these weaknesses in big data science. In the near term, data scientists should take a page from social scientists who have a long history of asking where the data they're working with comes from, what methods were used to gather and analyse it, and what cognitive biases they might bring to its interpretation (…). Longer term, we must ask how we can bring together big data approaches with small data studies - computational social science with traditional qualitative methods. (…)This goes beyond merely conducting focus groups to confirm what you already want to see in a big data set. (…) Social science methodologies may make the challenge of understanding big data more complex, but they also bring context-awareness to our research to address serious signal problems. Then we can move from the focus on merely "big" data towards something more three-dimensional: data with depth". (Crawford, 2013)

How can an economist contribute to the collective effort of intelligibility called for by Crawford (2013)?

From the consequentialist viewpoint characterising economic science, the participation of the greatest number in the great trace left by the entanglement of

our virtual lives is necessary. This would indeed make knowledge about the reality of our societies more accurate, and make it possible for all to contribute to measuring social justice, "*each counting for one, and none for more than one*" according to the utilitarian maxim. This requires to support the fight against the digital gap, by making relegates a part of the all-connected society, and to organise the equal recording of everyone's existence.

At the same time, respecting public freedoms and the procedural nature of social justice must not be forgotten, with the enactment of fair rules to supervise participation in this new public space (normative economics highlights the importance of the reversibility, publicity and participation criteria[7]).

Moreover, following on from the axiomatic work conducted on 'Old world' statistical indicators, economists could contribute to formulating the normative properties which Big Data algorithms should respect, deconstructing their image of divine black box impervious to human thought.

Finally, if we compare the compilation of our virtual existence to a common good to be protected and shared, we must prevent the market from deciding alone on its collection, storage, exchange and development. The data of the digital era is non-rival: it is not consumed or destroyed through its utilisation. On the other hand, data is to a lesser extent excludable in that it is difficult to prevent its utilisation. As such, data is at least a "club good" and at most a "public good". For an economist, the Government must necessarily be in charge of data regulation, in order to guarantee its efficient production and fair distribution.

To conclude this discussion, Big Data and the new world of statistics symbolised by it come up as an improvement on the data scarcity of Old world statistics. However – and in the end this is fairly satisfying for an economist – it concerns more a marginal improvement of which we should not overestimate the scope but control the development.

7 Among many others such as conformity, possibility of recourse, contradiction, motivation, proof, independence, impartiality, expertise and legality.

About the author: **Elisabeth Tovar**, Senior Lecturer; Université Paris Ouest and EconomiX (UMR 7235)

To quote this article: "Measuring Socio-Spatial Justice: From Statistics to Big Data[8] – Promises and Threats", *justice spatiale | spatial justice*, n°10, July 2016, http://www.jssj.org

Bibliography

ANDERSON Chris, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », Wired, 2008, [URL: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory].

ANDERSON Chris, Makers: The new industrial revolution, New York, Crown Business, 2012.

BACHIMONT Bruno, « Formal Signs and Numerical Computation: Between Intuitionism and Formalism. Critique of Computational Reason ». In : H. Schramm, L. Schwartz et J. Lazardzig (éds.), Theatrum Scientiarum: Instruments in Art and Science, on the Architectonics of Cultural Boundaries in the 17th Century, 362-382. Berlin: Walter de Gruyter Verlag, 2008.

BAROCAS Solon et SELBST Andrew D., « *Big Data's Disparate Impact* », California Law Review, n°104, 2016 [URL : http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899].

BENTHAM Jeremy (1789) Introduction to the Principles Wof Morals and Legislation, 1789, [URL : http://oll.libertyfund.org/titles/278].

CARMES Maryse, NOYER Jean-Max, « L'irrésistible montée de l'algorithmique. Méthodes et concepts en SHS», Les Cahiers du numérique, Vol. 10, n°4, 63-102, 2014.

COURNOT Antoine-Augustin, *Exposition de la théorie des chances et des probabilités*, Paris, Hachette 1843. [URL : http://gallica.bnf.fr/ark:/12148/bpt6k285042]

CRAWFORD Kate, MILTNER Kate et GRAY Mary L., « Critiquing Big Data: Politics, Ethics, Epistemology », Introduction au numéro spécial, International Journal of Communication 8, 1663-1672, 2014.

CRAWFORD Kate, SCHULTZ Jason, « Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms », Boston College Law Review, vol 93, 93-128, 2014.

[8] A first version of this discussion was drafted for the workshop on "*Liberté, égalité, computer. Gouvernementalité algorithmique and spatial justice*" organised by *Justice Spatiale/Spatial Justice* on 28 November 2014 at Paris Ouest University. The author would like to thank the workshop organisers, as well as all contributors and participants for the contradictory but fruitful debate which led her to write this article. She would also like to thank the two anonymous reporters whose comments greatly contributed to improving this article.

CRAWFORD Kate, « The hidden biases in big data », Harvard Business Review, 1, 2013 [URL : http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data].

DELORT Pierre (2015) *Le Big Data*. Paris, PUF, Collection Que sais-Je ? n°4021, 128 p.

DESROSIERES Alain, Gouverner par les nombres. L'Argument statistique II, Paris, Presses de l'École des Mines de Paris, 2008b.

DESROSIERES Alain, Pour une sociologie historique de la quantification. L'Argument statistique I, Paris, Presses de l'École des Mines de Paris, 2008a.

DICK Philip K., « Minority Report », Fantastic Universe, 1956.

DIMINESCU Dana et WIEVIORKA Michel, « Le défi numérique pour les sciences sociales », *Socio. La nouvelle revue des sciences sociales*, n°4, 2015. [URL : https://socio.revues.org/1254]

GITELMAN Lisa (éd.) *Raw data is an oxymoron*. Cambridge, MIT Press, 2013.

GOSSELIN Philippe, « Rapport d'information sur la nouvelle méthode de recensement de la population », Rapport d'information de l'Assemblée Nationale, n°1246, Assemblée Nationale, 2008.

GUTHRIE FERGUSON Andrew, « Predictive policing and reasonable suspicion », Emory Law Journal, 1, 2012.

HU Han, WEN Yonggang, CHUA Tat-Seng, LI Xuelong « Towards scalable systems for big data analytics: a technology tutorial » IEEE Access Vol **2,** 652–687, 2014.

KOSS Kelly K., « Leveraging Predictive Policing Algorithms To Restore Fourth Amendment Protections In High-Crime Areas In A Post-Wardlow World », Chicago-Kent Law Review, vol 90, n°1, 301-334, 2015.

LABROUSSE A., « Nouvelle économie du développement et essais cliniques randomisés : une mise en perspective d'un outil de preuve et de gouvernement », *Revue de la régulation*, n°7, 2010,[URL : http://regulation.revues.org/index7818.html].

LANEY Douglas, 3D Data Management: Controlling Data Volume, Velocity, and Variety, META Group, 2001 [URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf].

MAYER-SCHÖNBERGER Viktor, « La révolution Big Data », Politique étrangère, n°4, 69-81, 2014

MAYER-SCHÖNBERGER Viktor, CUKIER Kenneth, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Eamon Dolan/Houghton Mifflin Harcourt, 2013.

MOSCO Vincent, To the Cloud: Big Data in a Turbulent World, Paradigm Publishers, 2014.

OUELLET Maxime, MONDOUX André, MÉNARD Marc, BONENFANT Maude, RICHERT Fabien, « *Big Data*, gouvernance et surveillance », Cahiers du CRICIS, n°2014/1, 2014, [URL : http://www.archipel.uqam.ca/6469/1/CRICIS_CAHIERS_2014-1.pdf].

PEARSALL Beth, 'Predictive Policing: The Future of Law Enforcement?', National Institute of Justice Journal, n°266, 2010 [URL : http://www.nij.gov/journals/266/Pages/predictive.aspx]

PRÉTECEILLE Edmond, "Segregation, social mix and public policies in Paris", In T. Maloutas et K. Fujita (éds.), Residential Segregation Around the World. Making sense of contextual diversity, Ashgate, 153-176, 2012.

PRESS Gil, « Big Data News Roundup: Correlation vs. Causation », Forbes Tech, 19 avril, 2013 [URL : http://www.forbes.com/sites/gilpress/2013/04/19/big-data-news-roundup-correlation-vs-causation/]

PRÉTECEILLE Edmond, « Is gentrification a useful paradigm to analyse social changes in the Paris metropolis? » Environment and Planning A, Vol 39, n°1, 10-31, 2007.

SEN Amartya, Éthique et économie, Paris, Presses universitaires de France, 1993.

SEN Amartya, L'idée de justice, Paris, Le Seuil, 2010.

SHEARMUR Richard, « Dazzled by data: Big Data, the census and urban geography », Urban Geography, vol. 36, No. 7, 965–968, 2015 [URL http://dx.doi.org/10.1080/02723638.2015.1050922]

SILM Sirii, AHAS Rein, « The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset », Social Science Research, Vol. 47, 30–43, 2014.

SPRAGUE Robert, « Welcome to the Machine: Privacy and Workplace Implications of Predictive Analytics », Richmont Journal of Law and Technology, 21, 2015 [URL : http://jolt.richmond.edu/v21i4/article13.pdf]

TOVAR Élisabeth, « Mesurer la pauvreté : l'apport de l'approche par les capabilités. L'exemple de l'aire urbaine parisienne en 2010 », *Informations Sociales*, n°82 (Mars-Avril), 2014.

UNIVERSITÉ DE LORRAINE et UMR ATILF (Analyse et Traitement Informatique de la Langue Française), « Statistique », Dictionnaire Trésor de la Langue Française Informatisé (TLFI), [URL : http://www.cnrtl.fr/etymologie/statistique]

VIGEN Tyler, Spurious Correlations, Hachette Books, 2015.

WEINBERGER David, Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room, New York, Basis Books, 2012.