17

NEW ERA FOR COMMENSURABLE COMPARATIVE URBAN RESEARCH? MACHINE LEARNING AND/OR PROPAGATIONS

Dominique Boullier

Let's jump ahead to a distinction that only clutters the debate or reassures: there are no qualitative or quantitative methods and « mixing » them does not help. There are only various degrees of consideration of the dimensions (variables) of a problem and, therefore, possible comparisons. Indeed, in urban anthropology, it is not impossible to deal with a city, or even a village, or even a district, as I did in my thesis a long time ago. This can produce a case study which is not « qualitative » just because it includes interviews or quantitative just because it includes demographic data, for example. It will, in any case, tend to deal with as many dimensions of the situation or even of a given problem in that geographical area as possible. These dimensions are the facets of a situation, which can be extended to the point of making what Geertz (1973) called "thick descriptions" with the secret hope, never achieved, of recording « the whole » relations between generations in such and such a block of five towers, of following all the movements of drug dealers between such and such a neighbourhood, of recovering all the dimensions of the search for employment in such and such a town, etc. The qualitative approach often has a much greater quantitative claim than it is given credit for, because it involves increasing the dimensions observed, or its variables, even if this terminology is not used.

But some researchers, like myself, also want to make a comparison between two neighbourhoods, if possible fairly close, even if statistically there are differences. This remains labelled as « qualitative », but as soon as we start to compare, we are necessarily obliged to make two neighbourhoods or two populations or two situations or two cities commensurable. In order to do this, one can only enter into a process of *reducing the dimensions* or the variables observed. The question must become more precise, the indicators more reduced, the situations observed more limited in order to remain fairly similar, etc. The only limit to this reduction process will be its feasibility, in terms of time, access to situations, transcription of interviews, etc., along with what data scientists call after Church (1932) and Turing (1937), the « complexity classes » that make a

 DOI: 10.4324/9780429287961-20

problem computable or not (Boullier and El Mhamdi, 2020). When conducting an *ethnography*, one willingly accepts to do work limited to a field since one knows that *ethnologists* will be able to use this data to make it comparable later and even that *anthropologists* will be able to make abstract models that are much more formal and reductive from the observations collected. Mauss (1950) was no different: he never went into the field but was able to read everything published that might have a bearing on the problem he was dealing with, for example, the gift, which led to footnotes that were more important than the main text in his famous essay. What are these footnotes if not databases, not formalized as such nor quantified, but the number of observations on a gift cycle of a particular type necessarily plays a role in validating a certain theoretical statement, whereas a single observed occurrence will lead to consider it as an exception. It is always prudent to avoid pseudo-quantitative expressions in research writings, such as "often", "most", and "in some cases", which by their vagueness really prevent the reader from appreciating the validity of the indicator or the phenomenon described. But to do this, it must be recognized that it is not frequency that is the key to a good case study since its objective is primarily to extend the dimensions observed, not preparing the material for possible comparisons but emphasizing the uniqueness of the case.

ANT, Thick Description: How to Expand the Dimensions of Observations

Actor-network theory has gone even further in this direction, as it does not predetermine any "whole" as a reference (Latour et al., 2012) but lets the actors do so, as by *following* them, we can then verify whether they agree to make groups, classes, definitions hold together and make them natural or whether they remain in controversy, in negotiation over the nature of the situation, the territory, the affiliations, etc. The advantage of this approach lies in the creativity of the observation, which will be attentive to all the mediations of processes that are always "in the making". Thus, an idiosyncratic « individual public transport » project such as Aramis (Latour, 1996, which is still of interest to urban planners and less and less utopian, can be described by the author as a detective search for what caused it to go off the rails, by reviewing all the conditions for a project, just one, this one case, to be sustained or, on the contrary, to be killed off, as Latour proposes, by a lack of love, that is to say, by a lack of interest on the part of the stakeholders. To achieve this type of dynamic description ("following the actors" and not accumulating them as in Geertz's "thick description"), it is necessary to adopt an anti-reductionist point of view, according to the principle of "Irreductions" announced by Latour (1998). we may then say, with regard to all these case studies, these monographs, that this is not "scientific", that this type of work focuses on one specific case and therefore cannot draw laws from it, or at least points of comparison.

In fact, these are different moments of scientific activity that remain equally necessary. The increase in the indicators of observation, in the dimensions of the problem, in the variables of a question is necessary at the beginning because the researcher must doubt the relevant unit of observation. S/he cannot register as such the definition of the actors who would say that to analyze urban traffic, it is necessary and sufficient to use transportation reports and commuter studies, which by a strange chance, are precisely carried out by urban traffic analysis services. In doing so, s/he runs the risk of being biased by the a priori definition of the research object, carried out by others who have formatted the data according to this definition. This time of irreductions, of exploration of all possible dimensions, of repopulation of all mediations that could play a role is therefore necessary and salutary to free oneself from positivism of data or observations. This does not disqualify them, but it is prudent to explore the limits of validity of these categories before taking them at face value.

How Urban Quantitative Studies Face the Dimensions Problem

Let's take a look at how these questions are addressed in more quantitative approaches through some examples in the recent scientific literature. I selected a sample of papers from the journal Urban Studies, the selection being made by the search engine of the journal from the query « comparison » and discarding all research done before the year 2010. Among the first 30 papers, the comparison may rely on countries, cities or blocks, but there is a clear geographical basis for the comparison. Some have a problem with building a comparison at the international level when categories do not fit, like « brownfields » (Adams et al., 2010, p. 98): « Here, we are reminded that the term "brownfield" is generally reserved for potentially contaminated land in North America in contrast to the broader concept of previously developed land employed in England and Scotland. » Despite this obvious limitation, the paper describes three areas of differences, physical, cultural and institutional, the last two ones being standard differences among countries for any kind of topic and aggregating many indicators. All papers reviewed which used statistical sources developed a method based on regression, standard or logistic, and there is no chance for them to make use of other Machine Learning techniques since their datasets are still limited to the traditional registries even though they may assemble in a very clever way. When analyzing the changes of employment subcentres through a comparative study of Houston and Dallas, Liu et al. (2019) select these cities because « the former is the largest US non-zoning city, and the latter, a zoning city of comparable size and regional situation ». But they face a problem of limitation in size: « To avoid the small number of observations when using subcentres as the samples, this study employs traffic analysis zones (TAZs) within subcentres as the basic spatial units to ensure sufficient observations for statistical analysis. (…) It is worth noting that the changing boundary of TAZs may cause data discrepancies over time. » (p. 2546). When constrained to use datasets from public authorities, no scholar can pretend to be in control of its quality and, therefore, the study must be twisted in order to bypass or to comply with these limits. When (Jun et al., 2018) want to « investigate the effects of urban spatial structure on the level of excess commute, (they [decide to]) compare the commuting patterns of two metropolitan areas having very different urban forms, Seoul, Korea and Los Angeles (LA), California, USA. We choose Seoul and LA as case study areas by regarding Seoul as an example of a compact city and LA as an example of a dispersed city » (p. 198). It looks like a very large framework, well designed for large tests of variables at the very general level. But in fact, they adopt a micro level of data collection, which is more adapted to understand the rationales of commuting patterns. However, this choice results in balanced outcomes: « One advantage of using micro data is that there is no intra-zone commuting cost bias and aggregation bias resulting from different zone sizes. (…) One disadvantage of micro survey data is that sample size significantly decreases during the sampling process because we regard each worker's residence or workplace as a zone in itself. » (p. 202). As we can see, when scholars try to build datasets themselves for their own purposes, there is a clear trend to seek to achieve a more detailed and quasi qualitative focus on individual situations and, as a consequence, to augment the number of features or dimensions collected. This balance between the use of detailed records for every case and the reliability of a smaller sample of cases is present at the core of all comparative studies.

Dong and Zhu (2015) manage to merge data from various sources in order to obtain both details on specific variables and a large set of data—we will call them below « fat and tall data »—at the same time): « Data for this study are drawn from several sources. Land use, public transportation, bike route and employment data in the year 2008 were provided by Metro and SCAG, the MPOs for the Portland and Los Angeles regions. Housing and socio-economic data are drawn

from the 2006–2010 American Community Survey (ACS) data set. Local street networks of the two regions are from Census 2008 TIGER/Line shapefiles. » They wanted to compare Los Angeles and Portland regarding smart growth policies in these two contrasting metropolitan areas. The indices they extract from these datasets are aligned with the model of what a smart development should be made of: « Those indices are selected to represent four key physical aspects of smart growth in residential neighbourhoods (residential density, mixed land use, mixed housing, non-autotransportation accessibility) and two socio-economic dimensions (income and racial/ethnic diversities) » (Dong and Zhu, 2015, p. 778). There is a clear normative statement as to what is smart growth embedded within the selection of indicators. This does not allow any room for discovering possible correlations with those features which are not included.

Some models are more open to explore intermediate variables and to combine data analysis with a more qualitative style of study when they introduce the perceptions of the social members within this model in a more Weberian way. This is the case of Liu et al. (2020), who wanted to « examine the relationship between migrants' objective socio-economic status, evaluation of living standards and subjective wellbeing » in several cities in China. But in this case, the city is just a geographical reference for a more representative sample and not the focus of the research as such. What they found is that « the relationship between family income, home ownership in the host city and subjective wellbeing is significantly mediated by perceived living standards relative to the reference groups of schoolmates, neighbours and local urban residents in the destination city » (p. 19).

Big Data: Fat Data and Tall Data

These contrasted examples will help us clarify a misunderstanding between social science and data science since it is assumed that the former does not want to calculate while the latter does not want to question the categories they use. In fact, we are dealing with a division of labour between approaches that is much less watertight than we think. Thus, some of the categories derived from qualitative observations questioning categorization itself end up in the basic repertoire of every quantitative scientist. Experts in quantification, whether they are researchers or professionals in a given field or statisticians in national institutes, cannot afford to change categories for each new problem because the "investment in form" (Thévenot, 1986) to create a stable category is enormous since it is always reductive and therefore debatable.

But once an agreement has been reached to decide to differentiate « medium-sized cities » from « small towns » and « metropolises », or zoning cities and non-zoning cities according to a few robust and discrete (not fuzzy) criteria, it becomes easier to compare, since entities have been made comparable, even though, from the point of view of lived experience, they will nevertheless remain definitively unique. But this categorization debate presupposes a qualitative analysis because, at this point, all the attributes of the entities in question could be unfolded again to check that we are not aggregating incoherent entities or else combining observations in a supercategory (TAZ, Traffic Analysis Zone, for instance) which itself is only valid in relation to the other classes. Better still, when different variables are taken into account, it is possible to extend or reduce them, and this is most often a function of the time available, their accessibility, and above all, the calculation time for the computer (which itself, therefore, enters into the extension or non-extension of the dimensions).

Under these conditions, we can say that, in all approaches, we are tending towards Big Data in the sense that we look for volume and variety. But for what we call qualitative studies, we are talking about Fat Data (many variables, dimensions about a single case) and in quantitative

studies, we will speak of Tall Data (many cases compared on few variables, few dimensions). Fat Data visually corresponds to the increase in the number of columns in the spreadsheet and Tall Data to the number of cases that make up as many rows in the table (Griepentrog et al., 2016). The dream of totality (a maximum of dimensions for a maximum of cases) does not make much sense since the question of point of view remains essential, as we will see; however, very often, scholars will try to assemble both as we saw previously. The robustness of these analyses is not judged solely on the volume of variables taken into account (when we repopulate an endless description) or on the number of cases accumulated (when we collect all the occurrences of a phenomenon or all the population of a given city). But now, the recently available computing capacities make it possible to believe in this possibility of a « complete » dataset, which may even lead one to think that one can do without theory (Anderson, 2008) or standpoint. This is the great promise and the great illusion of the new methods using Big Data and Machine Learning (the technical basis for Artificial Intelligence).

It is even possible to revise the qual-quant division by claiming to exploit textual, verbal or even visual data, in a direct way, in their so-called "natural" environment that are digital networks of all kinds. Machine Learning and in particular neural networks, or connectionist approaches, have demonstrated their computational power in image recognition operations that can also be used for research or in natural language processing (NLP), which goes so far as to claim to process emotions in expressions (sentiment analysis). All these claims are often pretentious, as we have shown (Boullier and Lohard, 2012), but the advances are continuous and the demonstrative effects remain powerful and very seductive. The works of Manovich based on an exhaustive set of Instagram images of Hurricane Sandy in New York City or of comparisons among standard popular views of cities all around the world (Hochman and Manovich, 2013) will deliver a very disturbing and powerful vision of the event over time, but it clearly addresses questions about images and semiotic features more than about urban issues (such as the urban experience of disasters, their management and so on). Daniel Gatica-Perez uses a lot of large Instagram datasets that are sorted out by machine Learning algorithms in order to make visible the cultural differences in food and drink habits when observed in cities through the lenses of local inhabitants or those of tourists, for instance (Phan and Gatica-Perez, 2017), but one may consider that this is limited to confirming well documented cultural features.

These new computational possibilities can, therefore, undoubtedly inform comparative method provided that researchers accept several reorientations of their approaches: 1/The end of a search for exhaustiveness 2/An emphasis on an inductive approach.

The Price of Machine Learning: The Loss of Exhaustiveness

Machine Learning can only work with a lot of data, i.e. a lot of volumes, so that the algorithms can be trained on the part of the dataset in order to establish what is known as a « ground truth » (Jaton, 2017), which will be validated by experts in the field (most algorithms work by supervised learning). The algorithms can then be made to work on the dataset to be explored, which is still unknown. But to reach the required volume, one needs to accumulate a lot of data, often much more than a city department or a national statistical service has on a particular issue. The volume is achieved by accepting to collect very heterogeneous sources, as is done by digital platforms, such as Google or its subsidiary Waze for transport. It is the signals from mobile phones and their geolocation that provide this volume, although it is not possible to guarantee any kind of exhaustiveness in relation to vehicles on the move at a given time in

a given city, or even representativeness, since the references of an application's accounts are either inaccessible for regulatory reasons (RGPD) or unavailable because they have not been filled in (Waze does not need socio-demographic categories). As we can see, obtaining volume will only come at the cost of abandoning exhaustiveness and representativeness, which is not a problem for data sciences but can be for social sciences that want to exploit traditional methods for testing causality.

Indeed, Machine Learning will not test hypotheses or exploit a priori defined categories or variables but only correlations between all possible dimensions from any available feature. This approach, therefore, requires the adoption of an inductive method in which the hypotheses arising from the comparison will emerge from the exploration of the data. This culture of induction will pose many problems for social scientists trained in hypothetico-deductive approaches. One could even say that it brings ML closer to exploratory qualitative methods since it is also about finding new dimensions and correlations that would open up new and unexpected avenues. This is, in fact, what ML does best: either it confirms relationships that are already well known and does not add much, or even trivialities, which makes it very disappointing, or if used with finesse, it allows the validity of correlations to be tested that it brings to light and that could never before be raised.

The Type of Problems

But this will hardly allow the social sciences to claim to describe the general state of a problem within a given society and thus to give classical statistical robustness to the comparison. To do this, it is necessary to understand that the types of problems posed by the data sciences up to now are primarily intended to solve problems in the most operational sense. When questions of urban morphology are raised, data scientists may use historical comparisons about the growth of cities and mobilize fluid physics models for that purpose (Hernando and Plastino, 2014). But, from the point of view of data scientists, they will not do so in order to test a sociological hypothesis but to seek to optimize a model of anomalous propagations and help make decisions about transportation system design. The dimensions selected and the robustness of the results are therefore not of the same type since the aim is either to improve a model, to test sociological hypotheses or to help make a decision, which will be taken anyway, but by reducing uncertainty, using probabilities which are not sometimes even explained in the presentation of the results. Scholars who focus on urban form as a theoretical issue, in the wake of B. Hillier's works about space syntax, do not deal with the data in the same way. They are really concerned about reaching an understandable pattern of the connection between the « city » and the « urban », or the « space » and the « social ». For sure, they will process data with reductions based on their mathematical model of segregation/integration, but they are still able to downscale it at such a level (an estate, a block) that interpretations from social sciences become feasible and may help correct the model.

Approximations are also found in the quality control of the data itself. Data cleaning is a key step that is not always necessary when one thinks of relying on official databases such as those of statistical services or urban services. It is even suggested that it is possible to retrieve all this data in raw format and that once it is deposited in open data, all interested parties could make productive use of it. However, this is not true: there is no such thing as "raw data" (Denis and Goëta, 2017) (Goëta and Davies, 2016) but always data embedded in a network of meanings and purposes that have led to the selection of certain variables and thresholds for the needs of a specific service (Courmont, 2021). It must therefore be remembered that any reliance on data

provided for other purposes (which the social sciences have been doing constantly since the censuses themselves, Desrosières, 1993) requires very strict control and a validation period to avoid falling foul of the well-known adage "garbage in, garbage out". But this problem is even more acute for data/traces collected by platforms that use Machine Learning for their calculations. The cleaning time is, therefore, considerable, and the reduction in the volume of the dataset after cleaning is also sometimes enormous, which makes it even more important to focus on a certain number of variables and not to think that everything can be correlated. But any reduction in the volume of the dataset risks leading to a reduction in its variety, which is the second criterion of data quality for Big Data, a degraded version of representativeness since there is no longer a real reference population in this case.

Velocity of Big Data and Flow Analysis

However, a third property of Big Data concerns its 3V's: beyond volume and variety, is velocity. This notion does not make much sense in traditional statistical quality criteria and yet it will have to become so if the social sciences are to exploit the new data sources provided by digital networks. One of the first dimensions of velocity might relate to the flow character of the domains of observation themselves, even if this is not what is meant in Big Data by velocity. However, the computing capacities and traceability devices installed on all urban systems and increasingly with connected objects (Internet of Things) make it possible to monitor flows in their own temporal characteristics (minutes or seconds if needed), whereas until now, we have tended to use volume data over fairly long periods (years or months). The measurement of exchanges between cities, particularly in terms of trade, but also in terms of flows of people, transport flows, financial flows or even energy flows (water, electricity), allows for interesting comparisons, but until now, has been limited to aggregate calculations: so many entries or exits per year, per month or even per week, rarely decomposable beyond the classes provided by the administrations. The very fact of being interested in these flows already constitutes a conceptual turn for which John Urry (2007) has worked by founding an entire sociology on mobility. Very often, the sources that collect these flows are difficult to access, and, above all, the social sciences lack the concepts to think about them and, even more so, to think about the comparison between flows. The dissemination of certain innovations between countries, regions and cities can sometimes address the question of space and make it possible to compare, for example, household media equipment from one city to another, but this is rarely the entry point selected since it is preferred to refer to an explanation based on classic socio-demographic categories, among which the urban dimensions are rarely a priority. However, we have seen during the pandemic that changes in the organization of work, which were reflected in very different flows of commuters (Deng et al., 2020), had to take into account the location of jobs and homes, but also the very different qualities of housing, which may or may not allow teleworking.

The shift towards a sociology of flows is therefore overdue, whereas economists are more used to dealing with these issues from this angle. However, as they ensure that all their measures are made equivalent by recourse to monetary values, the singular properties of the goods and persons that circulate end up disappearing rapidly. Here too, as we can see, in order to make their material calculable, economists accept a reduction, which is certainly arbitrary, but which makes entities of totally different status comparable. The other social sciences rarely accept this reductionism because they want to keep the more or less fine grain of the different outcomes of the experience and give themselves a chance to account for other acting structures or other influential actors.

A New Statistical Quality: Traceability

But we must then accept the need to work at a finer granularity than the mere reference to flows. The speed of Big Data should both oblige the social sciences to do this, which multiplies the domains (the flows of travellers and those of conversations about a city are very different issues) and, at the same time, make phenomena comparable according to the attributes of any propagation (Boullier, 2023). For what Big Data introduces into the social sciences is a strong and constant reference to time and rhythm. In this respect, all comparisons are possible without having to make the reduction to value that economists do. In this case, curves become the key representation, organized along a time axis and allowing calculations on specific forms of propagation, the most famous of which are the S-curves, phase transition curves, found in all fields. Considerable mathematical work is now being carried out around these formalisms representing rhythms. The flows can then be granularized to the extreme and it is for this reason that Big Data speaks of velocity because calculations are carried out on the fly, no longer on stabilized databases but from a constant increment of data from sensors installed in all places and for all purposes. The methods and calculation requirements are very different, but so is the conceptual approach since we are no longer interested in comparing aggregated data over a very coarse and ultimately little studied time interval, but in the permanent variation and the patterns of these variations over very fine time intervals. For example, the comparison of waste flows per month between different cities makes it possible to correlate this with indicators structuring the city, its inhabitants and its waste collection and treatment services. But when we enter into on-the-fly, permanent traceability, on fine degrees that exploit the data from all the sensors (right down to the individual dustbins, as we are already experimenting with, or even on each piece of packaging that we have previously tagged individually), we can hope to test totally different correlations, which explore more individual behaviours or very specific moments or phases (holidays, habits, changes in packaging policies, etc.) from the point of view of the waste that circulates. As we can see, we must therefore accept that the traditional explanatory factors that have been mobilized up to now (socio-demographic, spatial or governmental properties of a city and its inhabitants) and that remain valid are no longer the only correlations that deserve to be tested to account for all the dimensions of a phenomenon.

Changes in Sensors, Changes in Calculation and Changes in Point of View: Propagations

In this case, we must agree to change our point of view and enter into a new generation of social sciences, those that will be based on the capacity to act, the agency, of entities that circulate, and whose paths and rhythms can be traced. This is what we developed extensively in our proposal for a new paradigm in social sciences based on propagations (Boullier, 2023).

Making entities as vast as cities, economic sectors and populations comparable implies reducing their internal diversity a priori. But it is no longer necessary to make these reductions a priori. Traces are stored not only by social network platforms but also by all urban information systems based on the collections made by sensors now placed in all city equipment but also in connected objects. It is even desirable to keep the granular state of all the events recovered by the sensors as long as possible. Indeed, it is thanks to this multiplication of elementary traces that new correlations can be tested. This is the imperative of urban services and user service platforms, but it can also become the imperative of researchers, who must now change their standpoint to take a close interest in all the entities that affect behaviour and events. These social sciences are emerging

thanks to the capacity to collect elementary traces *en masse*, thanks to the speed that makes traceability possible on the fly, and thanks to the computing capacity that makes it possible to process and model without any a priori hypothesis. These social sciences track and seek to detect patterns in the propagations of all types of entities. If they are interested in this, it is because their role in the composition of social situations and worlds is no longer disqualified.

The social sciences of previous « generations » (Boullier, 2017) were built with other imperatives and other computing resources. The censuses equipped with Hollerith's machines from 1890 onwards made it possible to make the « whole » of the social accessible, to statistically validate "laws" made robust by the exhaustiveness of the data collection. Durkheim's holistic approach made it possible to assert a point of view centred on the role of "society" and, more precisely, of social structures, an agent that acts upon individual behaviours through inheritance of position, representations and socialization. In this approach, comparisons are possible between large social entities delimited by administrative boundaries (nations, regions, cities) and sectors (work, industry, health, etc.) that impose their own categories and make them more commensurable (except for national differences, most of the time). Then, in the 1930s, statisticians finally agreed to use only samples, taking care to validate their representativeness and producing confidence intervals. These techniques were used to great effect by Gallup and Lazarsfeld in particular, who thus made opinion polls possible for the purpose of surveying media readership (and radio audience metrics in particular) and electorates. Individual preferences can therefore be collected and aggregated to form what will be called "public opinion" (Osborne and Rose, 1999). New techniques and new entities (the opinions of individuals) give rise to a new collective entity, public opinion. This does not replace the structural approaches, but this new standpoint, that of individual preferences, constitutes a form of integration of a more Weberian point of view in the social sciences. It will be mobilized above all by political science, by marketing and by economists who thus manage to make the strategies of the calculating agent computable on a large scale. The comparisons can then move to new aggregates, such as lifestyles associated with affiliations of various kinds (political ones especially).

Today, the availability of online traces of behaviours of all kinds makes it possible to detect patterns, for instance, from the activity of accounts on social media, based on likes, links, republication, etc., and no longer on public opinion. It is thus the messages that are propagated and attract reactions from users (who can sometimes be robots). More broadly, all traces collected by connected objects are no longer necessarily connected to precisely identifiable human behaviour but to flows, propagations of entities, such as vehicles, waste, noise, natural processes, brands, leaders, etc. Their only propagation produces effects on databases, on automated decision processes or on individual perceptions and behaviours (think of the vanity metrics that display the popularity of a post). Social scientists are expert in explaining all these processes by using the conceptual framework of social structures or the one of individual preferences of agents. We advocate here the opportunity to take a standpoint from the propagations themselves so that we get access to these horizontal phenomena generated by proximity at a specific place, moment or media. Here rumours, ideas or collective impulses (crowd movements) can become viral with a considerable reach, even though sometimes limited to a short time (such as buzz evident on the internet through the power of replications favoured by the platforms). These aspects of social life cannot be reduced to social structure effects nor to individual choices, they display specific patterns of propagation that are so frequently experienced in urban settings. Traditionally, one could say that the urban experience is increasing the number of opportunities for « superficial » encounters and influences as Simmel put it. This exposure to so many flows of propagation must become part of urban studies, all

the more when data are collected *en masse* through digital information systems. Comparisons will thus address the patterns of the propagation curves (spread, speed, tipping points, etc.), to describe cities as a flow and a rhythm, not far from Urry's vision. Comparisons of propagations and rhythms on any specific issue or at the level of the global city, deliver insights about the signature of the city, closer from the experience of inhabitants and at the same time easily connected to the governance methods, always sensitive to schedule, programmation of projects, events and so on.

Machine Learning techniques are much easier to adopt when processing this kind of material. These techniques will test correlations and propose some patterns based on variables within the features of the circulating entity (the semiotic features of a message about an urban project, for instance). However, we should not forget that the two other standpoints can also pretend to account for parts of the patterns: the density of the network (which is a structural feature) or the role played by specific influencers (which is close to individual preferences). Therefore, the comparisons will focus on very different entities and distribute their agency in different ways, depending on the standpoint on the social (structures, individual preferences or propagations). Of course, this will trigger the hope of combining all these standpoints in a general point of view, always looking for the « whole » (Latour et al. 2012). However, the entities, the methods, the tools, the indicators remain different for each of these approaches (structures, individual preferences or propagations). Mourning the vision of a « whole » or of the total interpretation and balance between all agencies is a requirement of scientificity that must be analytical in order to become computable, being more aware of the framing produced by the calculation devices. The emergence of new computation opportunities and methods will force social sciences to think back more about the limitations of any quantification operation, and this will make a difference with plain data science.

Can Machine Learning Be Used for Comparisons in Social Sciences?

Comparisons are potentially innumerable when one goes down to this finely grained level of traceability. The challenge of artificial intelligence, and more precisely of Machine Learning, is to make any correlation between the patterns of these circulating entities calculable, comparable and testable. Machine Learning is massively a comparison undertaking, particularly through what is known as similarity detection, which can be done using several methods that allow the number of classes that we want to obtain to be chosen (clustering by SVM in particular). The classes in question will only be labelled, i.e. interpreted, at a later stage when the algorithm has tested several groupings according to their statistical performance. This approach is therefore inductive, even though the learning process, once carried out, makes it possible to give priority to some of these classes. The culture of machine learners (Mc Kenzie) is not to get back to experts of the fields or to social sciences to validate these choices. Usually, some purely formal optimization procedure seems enough to make these decisions. It may generate some selection of features, some aggregation of variables just for feasibility requirements or along with data scientists' preferences. This is what looks like an unsupervised process that creates the illusion of a fully inductive process. However, it is not just a convenient way of avoiding conceptual or empirical discussions about the relevance of these classifications as processed by the algorithms. Indeed, this is an opportunity for experts or social scientists to get back in the game by contributing to the validation of these classes from their field or theoretical expertise. Induction, in this case, may find its limits. The classification proposals made by the ML algorithms might be considered and fruitful but should be controlled and re-embedded in the knowledge of experts or social scientists. This has

to rely on an organizational design of the data investigation process that is not yet established as state-of-the-art.

What ML techniques and Big Data availability offer is an opening of the variables that are used for the comparisons because many more correlations can be tested. The interest of this flexibility can be understood when exploring non-conventional or transversal domains. For example, the « watershed » entity hardly existed statistically before it was considered, first from an ecological point of view and then from a governmental point of view, that this dimension of the territory deserved a separate status. It was then necessary to collect and aggregate the data according to this new class, which is no longer that of a municipality or a classic administrative territory but which makes it clear that a "city" as an aggregated and built area, from the point of view of water, is not the relevant territory since it is organically linked to its countryside and its rivers. The sensors installed in this way can therefore be closer to the natural phenomenon, which forces us to redeploy dimensions of social life that have hitherto been described solely through the classic statistical instruments used by administrations. The exploration of correlations can also become stimulating for any field that is not yet subject to traceability, and discovering the role of certain mediations can then reveal a new way of dealing with problems. For instance, comparing urban segregation patterns should require ethnical classification that is not legal in France. Many ways of bypassing this limitation are tested, but an ethnography of the choices made by real estate agents demonstrates that they are among the ones who carry these segregating categories as a natural social knowledge. Accessing the datasets of the activity of these agents can therefore become a more reliable way of exploring the construction process of the segregation without necessarily knowing exactly where and when it takes place in their decision processes. But ML techniques can help test and select the valid hypothesis and make the comparison among more or less segregated cities, provided that the datasets have the volume required for learning and testing the validity of the correlations. This is at least the gamble that Big Data is announcing, even if some are content to make it a much finer and more permanent instrument of generalized surveillance. However, the contribution of domain experts is essential, as we know the risk of spurious correlations, which lead either to totally unrealistic causal links or to the omission of a correlation common to a hidden variable. In this respect, data scientists cannot do without the social sciences if they are willing to take the risk of entering this more uncertain but, at the same time very stimulating universe in order to monitor new entities. We could thus have the advantages of the qualitative approach and have rich descriptions from trace sensors available in networks and connected objects. Then we could reduce the exploration of data sets since we have to reduce the dimensions anyway to test correlations but now without doing so a priori, while benefiting from an additional robustness thanks to the volume of data and the testing of all these calculations. In this regard, comparisons are much more open and available because sources are more extended, and they can achieve more robust outcomes since the computation process using ML techniques may help validate classifications or unexpected correlations from these comparisons.

But if we remain confined to the same sources, i.e. the institutional or operational records of urban services and national statistics, we will not be able to claim to exploit the computing power of machine learning, and it will be easy to show that simple regression analysis can work just as well (Boelaert and Ollion, 2018). To benefit from the contributions of the new probability and prediction techniques of machine learning, we need to 1/change granularity (high frequency and finely grained traces of behaviour), 2/change our point of view (adopt propagations as a standpoint besides social structure and individual preferences) and 3/accept to follow entities whose role we had ignored until now (understand social life as relying on non-human entities such as objects or « natural life » components).

Of course, we must not forget to take into account the dependence on the sources that produce these traces, private firms, platforms or urban services that, of course, tend to format this data for their own purposes. But this can be said of all the research that exploits data from states, for example, what Sassen and Beck call « methodological nationalism », or from cities or private operators of transport, water distribution, or real estate agencies, etc. Vigilance is required in all cases, but it cannot be achieved by basing research only on data natively produced for research, which hardly exists.

Conclusion

There are limits to the validity of the uses of these new traces because they are not produced in the traditional environment of registries and statistical institutes nor of poll organizations that have built strong rules to control the quality of their surveys. But the main concern should be the style of comparisons urban studies would like to produce through the use of these data.

If their concern is to address the traditional type of issues about the style of governance, the respective efficiency of some policies for specific population, the economic attractivity of cities, the compared level of social segregation and so on, urban studies will have to rely on data sets delivered by public institutes and local governments that most often cannot be considered as big data and cannot use any Machine Learning techniques. The topics, the sources, the techniques and the standpoint (some structural features of the cities are the main variables) are always aligned. Other trends in comparison will get more into network analysis and try to compare the patterns of the networks in some decision-making processes or innovation adoption and so on among various kinds of cities. These patterns will be detected from datasets that are usually easily available, but they will rely on a specific standpoint: some individuals, each node in the network, play a specific role through their individual preferences. Topics, sources, techniques and standpoints are coherent again but very different from the previous ones. These methods are almost never interested in using ML techniques because the datasets do not reach the required volume.

This has a clear signification regarding the relevance of Machine Learning techniques for urban comparative studies: it will make sense and be productive only when different kind of issues and a different standpoint are adopted. When comparing the activity of police patrols in different cities, the official datasets about crime, patrol activities may be available, but if used only with a traditional approach, they do not need to make use of ML techniques if their concern is just to test the relationships between performance of patrols and neighbourhood social features or the effects of different management styles. It is only when data are used at a very finely grained level and from a standpoint on correlations regarding propagations that ML techniques may become profitable. For instance, the second-by-second traceability of activity—using voices recordings between policemen and the plaintiffs or the witnesses, finely grained localization of police vehicles, detailed history of specific types of crime in very specific areas, diffusion of new techniques of burglary, indicators of activities for bars, clubs, parking lots and so on—can be assembled in large datasets, heterogeneous, not validated by officials, with a huge amount of situations and a large number of variables. This will give the opportunity, for instance, to test some variables of the broken window theory. Not from a structural point of view, at a large scale, but as a process of propagation of alerts, of signals that may trigger some change in the minds of all stakeholders, some change in the perception of the environment. Not just to validate or discuss the theory at a general level but to get into the process of signalling, of labelling, of stigmatization and so on, through the channels of propagation of simple indicators along time. This kind of issue cannot be addressed without machine learning techniques, provided that all the datasets that we mentioned

are assembled, checked and cleaned up. The comparison will focus on a dynamic process in different settings, the process of labelling an area, a street, a specific location as « risky, bad, rotten, disadvantaged, dirty », etc. These comparisons do not help to account for the structure of the cities, they account for the processes of propagation among these structures with the resources of a thick description in the form of a timeline of the process that helps obtain specific patterns depending not on the type of cities but maybe on other features or causes. This is where ML techniques become powerful at testing all possible correlations within a dataset and delivering insights about causes that were never considered as potentially relevant. For instance, when exploring these patterns of propagation of labelling, the first moment someone attributes the label, the channel in which it is communicated may play a role and help us understand and anticipate the consequences, as Becker (1963) largely documented. A specific suburb or place may have a career that can be tracked in the conversations, the reports, and the signals that all stakeholders select in a period of time.

This is what Machine Learning announces: a new era of quantification that cannot go without a new standpoint on the social. A new world of data is emerging and gives the opportunity to adopt a new standpoint on the propagations that produce urban life as much as social structures and individual preferences: propagations that sometimes deeply affect urban life and behaviours, as the pandemic has made apparent.

References

Adams D, De Sousa C, Tiesdell S. Brownfield development: A comparison of North American and British approaches. Urban Studies. 2010;47(1):75–104.

Anderson Chris. The end of theory: The data deluge makes the scientific method obsolete. Wired, 23 June 2008.

Becker Howard S. Outsiders: Studies in the sociology of deviance. The Free Press of Glencoe, New York. 1963.

Boelaert Julien, Ollion Étienne. The great regression. Machine learning, econometrics, and the future of quantitative social sciences. Revue française de sociologie. 2018;59(3):475–506.

Boullier D. Propagations. Un nouveau paradigm pour les sciences sociales. Armand Colin, Paris. 2023.

Boullier D. Big data challenges for social sciences: From society and opinion to replications. ISA esymposium. 2017;7(2). https://esymposium.isaportal.org/resources/resource/big-data-challenge-for-social-science/

Boullier D, El Mhamdi EM. *Le machine learning et les sciences sociales à l'épreuve des échelles de complexité algorithmique*. Revue d'anthropologie des connaissances [En ligne], 14-1, 2020, mis en ligne le 01 mars 2020.

Boullier D, Lohard A. Opinion Mining et Sentiment Analysis: méthodes et outils. Open Editions Press, Paris. 2012.

Church, A. A set of postulates for the foundation of logic. *Annals of Mathematics*. 1932;33, 346–366.

Courmont Antoine. Quand la donnée arrive en ville. Open data et gouvernance urbaine. PUG, Grenoble. 2021.

Deng Z, Messacar D, Morissette R. "Running the economy remotely: Potential for working from home during and after COVID-19." STATCAN COVID-19: Data to insights for a better Canada, no. 26. Statistics Canada Catalogue no. 45280001. Statistics Canada, Ottawa. 2020.

Denis J, Goëta S. Rawification and the careful generation of open government data. Social Studies of Science. 2017;47(5):604–629.

Desrosières A. La politique des grands nombres. Histoire de la raison statistique. La Découverte, Paris. 1993.

Dong H, Zhu P. Smart growth in two contrastive metropolitan areas: A comparison between Portland and Los Angeles. Urban Studies. 2015;52(4):775–792.

Geertz Clifford. Thick description: Toward an interpretive theory of culture. The interpretation of cultures: Selected essays. New York, Basic Books. 1973, pp. 3–30.

Goëta S, Davies T. The daily shaping of state transparency: Standards, machine-readability and the configuration of open government data policies. Science & Technology Studies. 2016;29(4):10–30.

Griepentrog Brian, Marsh Sean, Turner Sidney Carl, Evans Sarah. Using social marketing and data science to make government smarter, in: Alvarez RM. (ed), Computational social sciences. Discovery and prediction, Cambridge University Press, New York. 2016, pp. 246–265.

Hernando A, Hernando R, Plastino A. Space–time correlations in urban sprawl. Journal of the Royal Society Interface. 2014;11:20130930.

Hernando A, Plastino A. Thermodynamics of urban population flows. Physical Review E. 2012;86:066105.

Hochman N, Manovich L. Zooming into an Instagram City: Reading the local through social media by. First Monday, Volume 18, Number 7—1 July 2013 https://firstmonday.org/ojs/index.php/fm/article/download/4711/3698

Jaton Florian. We get the algorithms of our ground truths: Designing referential databases in digital image processing. Social Studies of Science. 2017;47(6):811–840.

Jun M-J, Choi S, Wen F, Kwon K-H. Effects of urban spatial structure on level of excess commutes: A comparison between Seoul and Los Angeles. Urban Studies. 2018;55(1):195–211.

Latour B. Aramis or the love of technology. Harvard University Press, Cambridge, MA. La Découverte. 1996. (in French: Aramis ou l'amour des techniques. Paris, 1992).

Latour B. The pasteurization of France. Harvard University Press, Cambridge, MA. A.M. Métailié-Pandore. 1998. (in French: Les Microbes: Guerre et Paix Suivi de Irréductions, Paris. 1984).

Latour B, Pablo Jensen, Venturini Tommaso, Grauwin Sébastian, Boullier Dominique. "The whole is always smaller than its Parts"—A digital test of Gabriel Tarde's monads. British Journal of Sociology. December 2012;63(4):590–615.

Liu Y, Liu Y, Lin Y. Upward or downward comparison? Migrants' socioeconomic status and subjective wellbeing in Chinese cities. Urban Studies. 2020;58(12):2490–2513.

Liu X, Pan Q, King L, Jin Z. Analysing the changes of employment subcentres: A comparison study of Houston and Dallas. Urban Studies. 2019;56(12):2532–2548.

Mauss M. Essai sur le don, in: Mauss M (ed), Sociologie et Anthropologie. PUF, Paris, 1950.

Osborne T, Rose N. Do the social sciences create phenomena? The example of public opinion research. British Journal of Sociology. 1999;50(3): 367–396.

Phan H-T, Gatica-Perez D. #Healthy #fondue #dinner: Analysis and inference of food and drink consumption patterns on Instagram. In Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia (MUM '17). Association for Computing Machinery, New York, NY. 2017, pp. 327–338.

Thévenot L. Les investissements de forme, in: Thévenot L. (ed) Conventions économiques. CEE-PUF, Paris. 1986, pp. 21–71.

Turing, A. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 1937;2(1), 230–265.

Urry John. Mobilities. Polity, Cambridge. 2007.

Further Reading

Crawford K. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven. 2021.