Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities[1]

Daniel Arribas-Bel

darribas@feweb.vu.nl
Department of Spatial Economics
VU University
De Boelelaan, 1105
1081 HV Amsterdam
The Netherlands

Abstract

In this paper, I review the recent emergence of three groups of data sources and assess some of the opportunities and challenges they pose for the understanding of cities, particularly in the context of the Regional Science and urban research agenda. These are data collected from mobile sensors carried by individuals, data derived from businesses moving their activity online and government data released in an open format. Although very different from each other, they are all becoming available as a side-effect since they were created with different purposes but their degree of popularity, pervasiveness and ease of access is turning them into interesting alternatives for researchers. Existing projects and initiatives that conform to each class are featured as illustrative examples of these new potential sources of knowledge.

Keywords: Big data, Open data, Cities

[1]This manuscript was prepared for the special session "Urban Futures 2050", held in August at the 2012 ERSA meeting in Bratislava, Slovakia. The author would like to thank Julia Koschinsky, Ellen Schwaller and Emmanouil Tranos for the comments on a previous version of the paper. All the possible errors remain responsibility of the author.

Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities

Abstract

In this paper, I review the recent emergence of three groups of data sources and assess some of the opportunities and challenges they pose for the understanding of cities, particularly in the context of the Regional Science and urban research agenda. These are data collected from mobile sensors carried by individuals, data derived from businesses moving their activity online and government data released in an open format. Although very different from each other, they are all becoming available as a side-effect since they were created with different purposes but their degree of popularity, pervasiveness and ease of access is turning them into interesting alternatives for researchers. Existing projects and initiatives that conform to each class are featured as illustrative examples of these new potential sources of knowledge.

Keywords: Data sources, Open data, Cities

1. Introduction

These are exciting times to be an urban scientist. Not only is the world as a whole becoming more and more urbanized, once the historical threshold of more people living in cities than in rural areas has been already surpased (UN Department of Economic and Social Affairs, 2008), but the ability we are gaining to look into the inner workings of urban systems grows at even faster rates (Batty, 2012). An increasing amount of aspects of human life can be traced back through diverse digital footprints and, when aggregated, can reveal emerging patterns. Many economic transactions which used to be done *offline* have now been moved into the web, and their archival has created, as a "side-effect", incredible amounts of data that reflect many aspects of human behaviour. Democratic governments have not been completely foreign to technological change either. Many local, regional, national and supra-national public institutions are moving parts of their infrastructure

into the cyberspace and responding to the presure of activists that demand more transparency by releasing some of those data in open formats. All of these recent societal changes did not explicitly intend to redefine the "data landscape" available to urban researchers, but they have, making possible analysis at degrees of detail and scope unthinkable only a few years ago. The traditional creativitiy that applied researchers (geographers, economists, etc.) have developed to measure and quantify urban phenomena in contexts where data were scarce is being given a whole new field of action.

The amount and diversity of new data sources relating to cities that is becoming available grows exponentially[1], to the point it may seem unrealistic to look at all of them as one entity. However, this paper argues that much of them share three key characteristics that make them particularly well suited to current urban research. These include: their accidental nature, their open availability to researchers, and the ubiquity of their presence in everyday urban life. First, unlike a census or an economic survey, specifically created with research and policy analysis in mind, these sources were not originally intended for this end but for other purposes. Its potential usefulness for scientists comes then *accidentally*, as a byproduct. Second, and partly related to the previous one, all of these sources are available to researchers without the need to pay any fee or reach exclusive deals with the company/institution providing them. Finally, given the degrees of pervasiveness that are reaching the technologies and services where they originate, new datasets relating to virtually any quantifiable aspect of human life are appearing. Similar to other fields (e.g. see Edelman, 2012 and Einav and Levin, 2013 for recent reviews in the case of economics), the combination of these three factors creates a significant opportunity for urban and regional scientists to study new phenomena or to examine old questions with a new insight. Very much in line with the views of Overman (2010) in relation to Geographic Information Systems (GIS), these data can in turn help: reduce location measurement error of observations (although they may introduce other biases, see Section 5); avoid the issue of discretizing continuous problems; fill gaps where traditional data are unlikely to exist; and design instrumentation strategies as a source of exogenous variation.

[1] As a notable sign of this increase in the amount of urban data and subsequent research, the long-standing journal *Cities* has created a meta-journal, *Current Research on Cities (CRoC)*, with the aim of summarizing the field and pointing out current concepts in urban research.

The main line of argument is that most of these data sources fall into one of three main groups, based on the basic actor and the nature of the process at which they originate. The first category is comprised by data collected in a *bottom-up* approach from mobile sensors carried by humans. At an *intermediate* level, we can identify databases employed to provide a (usually free) service through the internet by web companies. These are typically aggregated from several primary sources and derive from businesses which either move or base their activity on the internet. The last group is characterized by the *top-down* fashion in which it is collected, and it has to do with data released in an open format by public and government organizations at different geographical levels. This classification is not exclusive and may be combined with other ones as well as inter-mixed (e.g. open govenrment data collected from mobile sensors, as in what is become known as "civic apps"). It is based on the intrinsic nature of the data origins and, although simple, it can be powerful to better interpret their attributes and, particularly, the type of processes or phenomena they may be reflecting. Ultimately, it is the good understanding of what the data can and cannot "tell" that makes it possible to incorporate them into meaningful studies.

Although potentially very advantageous, the use of these data is not free of challenges. Most of them derive from their *accidental* nature, from the fact they were not originally intended for this use. In particular, the major flaw may relate to the quality of the data: depending on what it is that we are trying to measure, the degree of completeness and bias in the population samples can compromise results and lead to misleading conclusions. But those are not the only hurdles to be confronted. Because often times they were not intended to be used in bulk, collection can be tricky and require some programming and database skills to access the sources. Once collected, the characteristics of the data may require methodologies and techniques not very familiar to the field yet. In some cases, as in what is come to be known as "big data", the size and lack of structure of the datasets is such that applying traditional techniques may not be the preferred solution and other methods, such as machine learning (Bishop, 2006) or knowledge discovery from databases (KDD) techniques (Miller, 2010), as well as advanced visualizations (Batty and Cheshire, 2012), may prove more fruitful. Section 5 will discuss these issues more in detail.

When dealing with such a broad topic, it is almost as useful to explicitly state what *is not* included as much as it is to describe what *is* covered. It is important to make clear that the main aim of this paper is neither of

the following. First, it does not intend to be an exhaustive survey of all the literature that has already taken advantage of these new kind of data. Although not vast (yet), the amount of publications using any of these three sources is large and sparse enough that any attempt would be incomplete. Instead, I provide a few illustrative projects as an example of the advantages to be benefitted from and challenges to be assumed. Second, this piece is not about *any* possible new source of data that is becoming available through the web or from public governments. The three categories in which the data sources featured are conceptualized are fairly broad and do include many of the new kinds of data appearing nowadays; however there exist alternative ones that are not best conceptualized into either of the three labels proposed in this work[2]. Third, this will not deal with opportunities arising from the use of these data in contexts other than academic research in the fields of urban and regional science. This is not to say those are nonexistent or irrelevant; on the contrary, applications in other fields can be highly benefitial, both in private (e.g. geo-targetted marketing) and social (e.g. disaster management, social services efficiency) terms. However, the strength of this paper is on bringing into the attention of those two academic communities these new advances in the hope it will ease their adoption for future research and, as such, it will be confined to that specific end.

This paper takes a practical approach by exposing the nature of these data sources in an accessible way. This is done purposely to reach as many potentially concerned regional and urban researchers as possible and stir their interest. For the advanced reader, a more explicit treatment of ontological and epistemological aspects of the use of this kind of data can be found in Warf and Sui (2010), Boyd and Crawford (2012) or Crampton et al. (2013). Equally important aspects such as its political economy or issues underlying their production can be found in Leszczynski (2012) or in a recently compiled edition by Lisa Gitelman (Gitelman, 2013). The rest of the text is structured as follows: Sections 2 to 4 describe the emergence and characteristics of the three different categories mentioned above, suggest how they can be helpful for researchers interested in urban issues and feature projects and iniciatives led by different actors that serve as real illustrations; Section 5 discusses some of the challenges that these new data sources pose when contrasted with the

[2]For instance, although closely related, *volunteered geographic information* (VGI, see Goodchild, 2007 or Sui, 2008) systems are not explicitly covered in this context.

ones traditionally used by the social sciences; and Section 6 concludes with a few remarks and highlights.

2. "Citizens as sensors": collecting data from the bottom-up

The invention of the internet and its ubiquitous presence nowadays, particularly reinforced with the emergence of mobile devices[3] such as smartphones and tablets, has created a platform in which every aspect of life is subject to leave a digital trace. Not only obvious ones like internet behaviour (browsing patterns) or economic activity (in the form of online purchases for instance), but also more traditionally intimate aspects of humans are being stored online: opinions are reported in blog posts, memories in pictures uploaded to social networks and even feelings or moods may be reflected on micro-blogging services such as Twitter, Inc. (2012). When we conceptualize internet-enabled mobile devices as extensions that empower human beings, *citizens* effectively become *sensors* (Goodchild, 2007) that produce streams of data that in turn can help reveal different aspects of their own nature.

This section is dedicated to a subset of these sources particularly promising due to its ease of access: that freely and openly available on the web. Many of these data are broadcast by individuals directly to the internet and may be accessed by other people (in fact that is usually the main aspiration of the "data producers", to be reached). Not only are they readily available but, in many cases, access is even encouraged by the providers. As an example, many social networks, such as Facebook, Inc. (2012) or Twitter, Inc. (2012), offer application programming interfaces (APIs) that allow developers to access (part of) their data in an automated way. Although these APIs were initially designed to build third party applications or services, their existence opens up the door for researchers to access these sources without having to reach any previous agreement. This has a democratizing effect in that the potential set of researchers that may access and work with the data expands beyond those able to reach exclusive deals.

The speed at which new services and networks appear, gain popularity or dissapear is such that any effort to list or create a full inventory is not only hard but also becomes useless quickly. However, it is possible to identify them by considering the following three characteristics. The defining attribute of

[3]According to a recent study (Meeker et al., 2012), the number of mobile users of the internet is expected to exceed that of desktop users before 2015.

this group, which also represents its main advantage, is the micro-nature of the data: they originate and are contributed at an individual level and, unlike the data sources covered in the next section, when they are accessed, this characteristic is retained. Second, their individual dimension also makes them part of what is come to be known as *Web 2.0*, a concept that captures some of the changes in the internet industry that have turned end-users from mere content consumers into both consumers *as well as* producers. This aspect is important because it is based on the generation of user content that most of the interesting databases are created. Third, they usually have embedded some sort of *social* functionality that connects users and turns the experience from an individual one into a community based one. This feature is a more recent one and offers great opportunities for social network analysis. The main examples are Facebook, Inc. (2012) or Twitter, Inc. (2012)[4] but also more activity-focused ones such as Foursquare (location sharing), Flickr (photography) or GoodReads (books). On the contrary, the condition of free availability effectively rules out other related sources that, due to confidenciality, privacy or security issues, are not publicly available. A particularly close and relevant case is that of mobile phone data. Although it also originates in mobile devices, its access is restricted and, when available, usually requires agreements with service providers. Its potential has already been assessed and reviewed in studies like Steenbruggen et al., 2011, for example.

In the context of this paper, these phenomena become specially relevant when an additional characteristic is taken into account: many of these digital traces incorporate the geographical coordinates of the location where the event occurs. This has been posible due to the popularization of location-aware technologies such as the global positioning system (GPS) and their inclusion in modern mobile devices. Such innovation has clear implications for the nature of the data produced, which immediately gains a spatial dimension. In fact, it is this possibility of connecting events with the location where they occur that appears as the most attractive aspect for urban and regional scientists.

The combination of individual data that reflect different aspects of human

[4]The open nature of this social network in which very few users effectively change their settings to make their content private is probably at the heart of its success and also the reason why it is one of the most evident candidates to use as data for research.

behaviour with the availability of spatial coordinates to geographically locate such activity poses important opportunities to applied urban and regional research. The individual nature has already been mentioned; these data are not only highly detailed in space, but also in time. The large volume generated by these sources and the high frequency with which they are updated means they can be understood as a stream of data in real time rather than as snapshots over periods. This represents a leap forward when compared to the frequency at which other traditional sources are published (e.g. ten years in the case of most censuses). It also has a remarkable potential to inform models with intensive data requirements, such as time geography approaches à-la-Hägerstrand as remarked in Sui (2012), or to bring insight in situations where there is severe lack of traditional data but the degree of pervasiveness for mobile technologies is large, such as developing countries[5]. This degree of detail and scope allows for a fresh approach that is likely to bring new answers to traditional lognstanding questions in the Regional Science and urban literatures, such as commuting or aggomeration economies, for example. In fact, the picture that best describes these sources of data is that of an incredibly detailed lens through which to look at cities. This new capability may represent a shift in urban research in a way that, as Lohr (2012) suggests, is akin to the invention of the microscope four centuries ago.

Although very promising, the scientific use of data of this nature for urban purposes is still at a very early stage. The first explorations into its potentials do not come from traditional urban and regional fields but from computer science researchers. The emerging field of Computational Social Science (Lazer et al., 2009) and, in particular, that of "urban computing" (Cranshaw et al., 2012), of which good examples are Ratti et al. (2010), Cranshaw et al. (2010), Cheng et al. (2011) or Noulas et al. (2011), is at the forefront. Mostly due to differences in traditions, backgrounds and interests, these studies set an emphasis on the *computing* side rather than on the *urban* one. In particular, this is reflected on a combination of expertise from computer science and engineering to study cities. In the next section, I cover in more detail one of its most prominent illustrations as a case of use of this sources of data.

[5] As an illustration of the potential in this realm, a team of researchers at IBM developed the project "AllAboard" (Berlingerio et al., 2013) which implements a system based on mobile phone data that optimizes the public transit network in Abidjan, the capital of Ivory Coast.

Figure `livehoods.png` approximately here (on top of page)

Source: Screenshot captured from `http://livehoods.org/maps/nyc` on September 10th., 2012.

Figure 1: Example of *livehood* in New York City

2.1. The Livehoods *project*

"(T)he 'character' of an urban area is defined not just by the types of places found there, but also by the people that choose to make that area part of their daily life" (Cranshaw et al., 2012). That is the main motivation behind the Livehoods project (`http://livehoods.org`). The project aims at drawing the boundaries of what are called *livehoods*, areas of similar character within a city. Unlike static administrative *neighborhoods*, the *livehoods* redefine cities based on the habits of people who live there.

The delineation relies heavily on data from Foursquare, a location-sharing service in which users can instantly broadcast their location from their smart device in what is termed a *checkin*. Leveraging a database of about 18 million checkin's of many users, the researchers use machine learning techniques to cluster venues based on the users that frequent them. Applying an explicit rule of geographic proximity, the result is the livehoods: subsets of a city whose establishments (restaurants, bars, bookstores, stations, etc.) have a similar clientele or, in other words, where people who go to one of their venues also go very often to the other ones.

The project is a good example of use of the data sources covered in this section to increase our understanding of cities. Neighborhood delineation has been a challenge for urban social sciences (see, for example, Rey et al., 2011). Taking adavantage of the fine spatial and temporal granularity, as well as its ready availability online (the entire database was scraped from public posts pushed to tweets), the researchers are able to obtain urban pictures that would not be feasible with traditional methods such as surveys, and to gain insight on how the social component can be measured and captured providing, in this case, a different view on the neighborhood construct than it was available before. This is however only one among many other possible uses that urban scientists could find for these kind of data. Alternatively, the project is a good case to hint at how the output of its analysis could be incorporated in other studies. For instance, once constructred, the livehoods could be used as the main units of analysis, in an effort to capture the most

accurate unit of analysis and avoid the so-called modifiable areal unit problem (MAUP, Openshaw and Taylor, 1981).

3. Businesses moving online (and creating data in the process)

Not only individuals' lives are moving online, companies are also hopping on the internet train. In certain sectors, the popularization of the web has created important challenges but also opportunities to the traditional business model. Some firms have embraced them and have significantly increased their productivity and efficiency. Although this technology has been inserted in many diverse ways at different stages of the production chain, its inclusion as an additional factor has always been reflected in an increase of digital data about the economic activity undertaken. In some cases, these data are also exposed to the general public, creating an opportunity for researchers as well. As an example, the real state market has witnessed a transformation in recent years that has greatly improved the availability of information. Consumers nowadays have free access to online databases provided by websites like Zillow, Inc. (2012) or Trulia, Inc. (2012) that aggregate data from local brokers, providing a much larger overview of the market as well as additional information merged from diverse sources. These data are usually freely available via websites or machine readable APIs, which facilitate their extraction. In addition to online companies covering offline businesses, there has also been an outburst of *internet natives* covering new portions of the market. These firms do not have a clear offline counterpart and they are usually data intensive, meaning (digital) data are a key part of their business model. Because many of these sites offer their services free of charge (the revenue is collected through advertising or other means), much of the data are available to researchers as well. This aspect is key because it sets the sources reviewed in this section apart from more traditional data-oriented companies. There is a long-standing tradition of firms whose main business model is to collect and *sell* datasets to researchers or analysts (e.g. ESRI's Business Analyst establishments database or Experian's real state datasets). Although very relevant in some contexts, where their contribution has not only been useful in itself but also in creating synergies with the public sector and in influencing public data collection, their "non-accidental" nature leaves them out of the main focus of this paper.

The spatial as well as temporal availability of these sources of data is much more diverse than for those in Section 2. Fine granularity may be found in

Figure `walkscore.png` approximately here (on top of page)

Source: Screenshot captured from `http://www.walkscore.com/AZ/Phoenix` on September 10th., 2012.

Figure 2: Surface of Walk Score for Phoenix

either time (e.g. Trulia, Inc., 2012) or space (e.g. Walk Score, 2012), but rarely in both. Although, in most cases, the data is not released in "real-time" but aggregated at some sort of scale, the periodicity at which they can be obtained is usually better than that of official sources such as census bureaus, which makes them very attractive for studies in which the interest lays in the temporal evolution of some sort of urban phenomenon. The biggest advantage of this family of sources is the large variety of aspects they can cover: because they originate from the most diverse businesses and range of economic activities, they have the potential to provide measurement on aspects of the economy that used to be unimaginable to capture in data. This better periodicity and sometimes more detail should not be seen as a reason to understand this group of sources as a replacement or substitute of more established ones such as censuses or national surveys. On the contraty, it should be considered as a complement, an alternative that may fill particular needs or that may be merged to more traditional datasets in order to capture the process of interest.

Unlike data of more recent nature, these sources have already been included in some studies within the urban and regional science fields (e.g. Avnimelech and Feldman, 2011) and have proven successful in bringing empirical insight on aspects that traditional data did not allow to capture or identify. Moreover, recent iniciatives such as Feldman et al. (2012), ensure that in the future only more studies will take advantage of their properties. As an example of these sources, the next subsection presents Walk Score, and online company which produces an index of walkability.

3.1. WalkScore.com

Walk Score (2012) was originally a project of Seattle-based company FrontSeat (2012). Its purpose is "to promote walkable neighborhoods" and, ultimately, their aim is for walkability to be included as a typical characteristic of a house (as their website says: "Our vision is for every property listing to read: Beds: 3 Baths: 2 Walk Score: 84"). In essence, the walk score (WS)

is an index that factors in several aspects of walkability such as accessibility and street network characteristics to offer an overall measure of the walkability of a location, as defined by its latitude and longitude coordinates. The data are freely available on their website as well as through an API and may be collected for every point within the set of covered cities[6].

Currently, WS is mostly used by real state brokers and realtors to capitalize on the value of walkability. However, the index is slowly permeating in empirical academic research as well. Some of them (e.g. Duncan et al., 2011 or Car et al., 2010) have first focused on validating their use vis-a-vis more traditional measures, finding very positive results; while others have used the index in applications on fields as diverse as real state (e.g. Rauterkus et al., 2010 and Pivo and Fisher, 2011) or urban design (e.g. Talen and Koschinsky, 2013). Scholars are starting to use it to replace traditional measures, which usually require a larger time and financial investment to collect. This is opening up the door to carry out studies on walkability at levels that were not feasible a few years ago. As an example of large scale analysis, WS has recently been included in a project that aims to evaluate walkability and its relationship with affordable housing at the US national level (Koschinsky and Talen, 2012).

4. Open Governments, open data

Opposite to data in Section 2, the last family of sources is the reflection of a "top-down" process, in which public organizations release some of their internal data in open format. In effect, governmental organizations, from the national level down to local authorities, are making available increasing parts of the data they collect while developing their activities. This process if fueled mainly by four main strategic drivers (Shadbolt, 2010): transparency and accountability, economic and social value, public service improvement and creation of new industries and jobs. The first one is a response to citizen demands and may be seen as a tool to build trust (Cabinet Office, 2012). By allowing external parties to access, review and study internal data, it becomes easier to identify and attribute responsibilities in cases of, for instance, corruption. Closely related to this goal is the emerging field of "data

[6]See Front Seat (2011) for a detailed description of the methodology and `http://www.walkscore.com/rankings/cities/` for a list of US cities for which WSs are provided (last accessed: September 5th., 2012).

journalism" (Rogers, 2011, Gray et al., 2012), which has played a role in the development of open government data and is based on the idea that data, just like text or photographs, can be a powerful tool to inform and hold governments accountable in cases where needed, hence collaborating to the function of press in a democratic society. On the other hand, there is a more pragmatic reason as well that oversees the last three main drivers. The opening of data can be a succesful strategy not only to serve the democratic goals of a government but to turn it into a more efficient and impacting organization. This has been pointed out by the so called "Government 2.0" movement, which aims at improving the effectiveness of governments by the introduction of technology and practices borrowed from the computer world in their processes. As O'Reilly (2010) puts it: "Government 2.0, then, is the use of technology -especially the collaborative technologies at the heart of Web 2.0- to better solve collective problems at a city, state, national, and international level".

At the moment of writing, `http://data.gov.uk` (UK) exposes 8,680 datasets and `http://data.gov` (US), 378,529 raw and geospatial datasets, and those are only two of the main government portals. This vast amount of data is as large in quantity as it is in diversity. A quick browse through the index reveals items as disparate as "US DOE/NNSA Response to 2011 Fukushima Incident: Radiological Air Samples"[7] and "Central Contractor Registration (CCR) FOIA"[8]. Clearly, not all of this is potentially relevant for urban and regional research. However there are still reasons to consider these sources. Many of the data have location information and can hence be geographically pinpointed, which provides them with a regional dimension. In addition, the incredible diversity and abundance of data in these portals, makes them a good archive when looking for proxy variables of phenomena for which data are not at hand or sources of exogenous variation within an identification strategy, for example.

The wave of opening data does not stop at the national level. Many regional agencies and local administrations are joining this trend as well. Cities like New York (`https://data.cityofnewyork.us`), Chicago (`https://data.cityofchicago.org`) or Paris (`http://opendata.paris.fr`) have

[7] `https://explore.data.gov/Geography-and-Environment/US-DOE-NNSA-Response-to-2011-Fukushima-Incident-Ra/u9mw-zn8r`

[8] `https://explore.data.gov/Information-and-Communications/Central-Contractor-Registration-CCR-FOIA-Extract/3hqn-qzh6`

Figure `nyc.png` approximately here (on top of page)

Source: Screenshot captured from `http://data.fabernovel.com/nyc-subway/` on September 17th., 2012.

Figure 3: NYC Transit System

started open data portals in which they upload datasets about many diverse aspects of the city. As an example of clear interest for urban researchers, next is considered the iniciative of the Metropolitan Transportation Authority of New York City.

4.1. MTA transit data

In 2010, the Metropolitan Transport Authority announced the release of part of the data about the public transit system in New York City in an open format (Metropolitan Transport Authority, NYC, 2010). Basically, this includes service and schedule data as well as other characteristics of the transit system such as hourly volume of *swap-in*'s and *swap-out*'s in subway stations. In this case, the original purpose is to put the data out so some services can be outsourced at low cost, increasing thus efficiency in the provision ("make data available to software developers who are interested in creating smartphone or web applications – or "apps" – that help our customers"). However, free access to this kind of data represents a tremendous opportunity that used to be restricted to the few researchers that were able to reach agreements with the agency. More than one in two New Yorkers (55%) commute by public transit (U.S. Census Bureau, 2010). Having direct access to data that describes and characterizes this phenomenon at such refined scale certainly makes possible interesting analysis for transport engineers, urban economists and planners.

The MTA iniciative is only one example of a more general trend that is perhaps best exemplified in the public transit case[9] but that spans a much larger case scenario. From local finance data to urban tree canopy metrics, the release of open public data has the posibility to positively impact much of the applied research conducted about cities and regions by allowing access to previously restricted data to a much wider scientific community.

[9]For a detailed list of municipalities offering transit data in open format, see `http://code.google.com/p/googletransitdatafeed/wiki/PublicFeeds`.

5. Challenges

So far, this paper has stressed only the benefits of these new sources of data. The reader up to this point would be tempted to wonder why, other than their novelty, they have not been used more intensively in urban research. One of the most obvious answers is that they, as many other types of data typicallly used, also have some drawbacks and imperfections that may prevent their use in some contexts. In order to obtain a full picture of the characteristics and nature of the data reviewed above, this section presents three main obstacles that pose challenges for their direct application in scientific research about cities[10]. The first one relates to the quality of the data, the second to the set of skills required to take advantage of these databases and the third one reflects on the suitability of traditional methods that were meant for traditional data. In addition to these, issues about governance and ethical questions are becoming of greater concern, particularly as it is becoming clear that it is very difficult to maintain privacy even in anonimized datasets (e.g. see Montjoye et al., 2013 for the case of cell-phone data). Although very relevant, these aspects are not directly related to the adoption in research of the data sources reviewed and, consequently, are left out of the paper. The interested reader can find a more extensive treatment in the literature on smart cities, of which a good recent overview is Batty et al. (2012) or Kitchin (2013).

One of the most prominent preocupations when trying to understand phenomena through data is to what extent the sample is representative of the population of interest. This is of particular concern when it comes to data that, for the most part, requires very particular characteristics in the user for it to be generated (e.g. to own a smartphone), to the point of posing deeper questions on the conceptualization of the world online (Graham and Zook, 2013). The extent to which this is a problem depends of course on both the exact type and source of data as well as the particular question being analyzed. While some of the sources reviewed in this context could raise issues of representability (e.g. Foursquare data as a representation of the preferences of a whole population), others do not suffer from that problem (e.g. MTA data on subway usage as a representation of subway usage in NYC). In addition, there are two more reasons to be positive about the use

[10]For a more complete review of challenges to the new age of large datasets in social sciences, see King (2011).

of these data. First, the increasing degree of penetration that the technologies powering these data are reaching can only improve the current situation. Second, in some cases the bias introduced by the data runs in a positive direction. As an example, Ahlfeldt (2013) discusses how the overrepresentation of young highly skilled professionals in the users of photo-sharing services such as Flickr or Picasa can in fact favour the analysis of the influence of urbanity in housing prices by better capturing the preferences of the segment of population that is a potential buyer of houses located in attractive areas. In any case, although present and potentially important, the existence of quality and representability issues in these new sources of data should only be one more characteristic to take into account and properly deal with but not a total deterrent that precludes its use in contexts where it is sensible.

A more subtle underlying cause of the lack of research using open web data is the different barriers to access them in raw format. Many of these sources are exposed to the world through APIs or in a form that require some pre-processing before becoming tabular data (e.g. raw text or html). The interested researcher then needs to have some basic general programming skills that allow him or her to write simple scripts to query the database, usually over the internet. Although this is not a particularly difficult task, it is certainly more intricate than a simple bulk download from a data portal, as most researchers are used to when it comes to obtaining more traditional data. The relevance of this aspect however, is also bound to diminish over time. As computational methods and larger datasets increase in importance and amount in social science research, the returns to learning basic programming capabilities and aquiring expertise on databases other than the traditional data sheets will increase. In fact, it is possible that eventually, they become part of the standard set of tools an applied quantitative social researcher is required to master in order to qualify as such, similar to the way typesetting systems (e.g. LaTeXor Microsoft Word) and statistical packages (e.g. Stata, Matlab or R) are nowadays.

Finally, there is also a case to be made about the suitability of the current methods to analyze and obtain insight out of the databases arising from some of these new sources. Many of the statistical techniques in use in regional science and urban analysis nowadays were created in a context in which data were characterized for its limited availability rather than for its over-abundance. This paradigm might be shifting and, if that is the case, new methods to complement the existing ones will be required. As Skupin and Agarwal (2007) mention in the context of large georreferenced datasets,

"traditional inference methods are either failing or have become obstacles in the search for geographic structures, relationships, and meaning" (Ch. 1). Such new generation of modelling techniques will have to expect continuous rather than discrete, large rather than small and, in some cases, real-time rather than delayed data. This will translate into a family of analytics in which the assumptions about the structure of the data will be traded for the ability to be applied fast and at a large scale. In some cases, as it happens with monitoring systems as those reviewed in Kitchin (2013), the approach will have to be revisited not only in terms of the algorithms used but also in relation to the data infrastructure required to support real-time analysis based on a continuous stream of data. The response from the industry to this phenomenon has been the new and emerging field of "data science", a blend of statistics, engineering and computer science that aims at creating value from the streams of data generated by the online economy[11]. If more of these types of data are to be included in regional and urban studies, researchers will also have to embrace these techniques in order to exploit all the meaninful information, and "borrowing" from fields like machine learning or information visualization, for example, will have to become a more common practice than it is nowadays.

6. Concluding remarks

This paper has reviewed the emergence of three new sources of data that may be useful for the regional and urban scientific communities. These are data coming from individuals carrying location-aware devices, from businesses moving (some of) their activity online and from governments releasing an increasing share of their data in open formats. For each source, a detailed characterization has been given as well as a real world case that serves as an example. A particular focus has been set on the subset of these sources that may be openly and freely accessed by researchers. The overview is complemented with a set of challenges posed by their nature and characteristics that are precluding its ready use in applied urban research.

Ultimately, these new data have the potential to bring new answers to old longstanding questions in the different fields of urban analysis, and that has been the main premise behind the motivation of this paper. The ability to

[11]Manyika et al., 2011 review its emergence and particularly note the shortage of properly trained labor force in comparison to the existing and future demand

look at urban phenomena through the potentially much more detailed and granular lens these data allow for should be put at the service of existing theoretical premises. However, given the particular characteristics outlined in the previous sections, these sources of data can also be a positive force towards higher integration between disciplines. In fact, they could be seen as a sort of interdisciplinary "glue" that favours cross-pollinization between fields where such interaction has been coming for a long time (e.g. economics and geography, Rodríguez-Pose, 2011) or that induces new creative collaborations within the Humanities as suggested in DeLyser and Sui (2012). Equally, the popularization of such data is likely to also strengthen the linkages between GIS and spatial analysis noted in Goodchild and Haining (2004). Related to interdisciplinarity is the question of data availability and transparency in science. Although not exclusive to the sorts of data reviewed in this text, the tremendous increase in the volume and variety of origins and quality these sources are bringing with them calls for a policy of transparency and, when possible (i.e. when not limited by terms of use or licensing issues on the data providers' end), of reproducibility (Peng, 2011)[12].

Rather than claiming discovery or exhaustiveness, the main purpose of the paper is to bring the attention of researchers who are actively conducting regional and urban analysis to the existence, availability and usefulness of these sources as a complementary alternative to those already in wide use (such as population censuses or surveys). In that sense, it should not be viewed as a plea to completely replace the existing data used in the field but, rather, to incorporate these new ones and to develop strategies to combine the best of both worlds in search of new insights. In an increasingly complex world, we need every possible tool at hand to understand it and be able to deal with the problems of the new Century. There is a new microscope available, it is now up to the researcher to use it.

References

Ahlfeldt, G. (2013). Urbanity. LSE-SERC working paper.

Avnimelech, G. and Feldman, M. (2011). The impact of institution quality,

[12]In this regard, it is remarkable the new open access journal *Sientific Data* (`http://www.nature.com/scientificdata/`), whose mission is to publish descriptions of scientifically valuable datasets.

cluster strength and TLO licensing capacity on the rate of academic staff spin-offs. In *Science and Innovation Policy, 2011 Atlanta Conference on*, pages 1–1. IEEE.

Batty, M. (2012). Smart cities, big data. *Environment and Planning B: Planning and Design*, 39(2):191–193.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., ... and Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, 214(1), 481-518.

Batty, M. and Cheshire, J. (2012). Visualisation tools for understanding big data. *Environment and Planning B: Planning and Design*, 39(3):413–415.

Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F. and Sbodio, M. L. (2013, accessed Jun. 10 - 2013). AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. `http://researcher.watson.ibm.com/researcher/view_project_subpage.php?id=4746`.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Boyd, D. and Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5):662-679.

Cabinet Office (2012). Open Data White Paper. Unleashing the Potential. Technical report, HM Government.

Carr, L. J., Dunsiger, S. I. and Marcus, B. H. (2010). Walk Score as a global estimate of neighborhood walkability. *American Journal of Prenentive Medicine*, 39(5):460–463.

Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proceeding of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Barcelona.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., and Zook, M. (2013). Beyond the geotag: situating big

data and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2):130-139.

Cranshaw, J., Schwartz, R., Hong, J., and Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, ICWSM*, volume 12.

Cranshaw, J., Toch, E., Hong, J., Kittur, A., and Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 119–128, New York, NY, USA. ACM.

DeLyser, D. and Sui, D. (2012). Crossing the qualitative- quantitative divide II: Inventive approaches to big data, mobile methods, and rhythmanalysis. *Progress in Human Geography*, 37(2):293-305.

Duncan, D. T., Aldstadt, J., Whalen, J., Melly, S. J., and Gortmaker, S. L. (2011). Validation of Walk Score® for Estimating Neighborhood Walkability: An Analysis of Four US Metropolitan Areas. *International Journal of Environmental Research and Public Health*, 8(11):4160–4179.

Edelman, B. (2012). Using Internet Data for Economic Research. *Journal of Economic Perspectives*, 26(2):189–206.

Einav, L. and Levin, J. D. (2013). The Data Revolution and Economic Analysis. *National Bureau of Economic Research, Working Paper Series*, No. 19,035.

Facebook, Inc. (2012, accessed Sept. 5 - 2012). `http://Facebook.com`.

Feldman, M., Graddy-Reed, A., McLauring, G., Nelson, K., and Reamer, A. (2012). Innovative data sources for regional economic analysis. Conference Guide. `http://maryannfeldman.web.unc.edu/files/2012/05/Participant-Contact-List2.pdf`.

Front Seat (2011). Walk Score Methodology. White Paper (accessed Sept. 5 - 2012). `http://www2.walkscore.com/pdf/WalkScoreMethodology.pdf`.

FrontSeat (2012, accessed Sept. 5 - 2012). `http://frontseat.org`.

Gitelman, L. (Ed.). (2013). "Raw Data" is an oxymoron. *MIT Press*.

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Goodchild, M. F. and Haining, R. P. (2004). GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science*, 83:363–385.

Graham, M. and Zook, M. (2013). Augmented realities and uneven geographies: exploring the geolinguistic contours of the web. *Environment and Planning A*, 45:77–99.

Gray, J., Chambers, L., and Bounegru, L. (2012). *The Data Journalism Handbook*. O'Reilly Media.

King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, 331(11):719–721.

Kitchin, R. (2013). The Real-Time City? Big Data and Smart Urbanism. SSRN Working Paper Series. Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2289141`

Koschinsky, J. and Talen, E. (2012). Affordable Housing and Walkable Neighborhoods. A National Urban Analysis . `https://geodacenter.asu.edu/projects/hud`.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstyne, M. V. (2009). Computational Social Science. *Science*, 323(5915):721–723.

Leszczynski, A. (2012). Situating the geoweb in political economy. *Progress in Human Geography*, 36(1):72-89.

Lohr, S. (2012, accessed Sept. 5 - 2012). The age of big data. *The New York Times*, pages `http://www.nytimes.com/2012/02/12/sunday--review/big--datas--impact--in--the--world.html`.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). *Big data: the next frontier for innovation, competition and productivity*. McKinsey Global Institute.

Meeker, M., Devitt, S., and Liang, W. (2012). Internet Trends. Technical report, Morgan Stanley, http://www.morganstanley.com/institutional/techresearch/pdfs/Internet_Trends_041210.pdf.

Metropolitan Transport Authority, NYC (2010). MTA Launches New Website. Press release; http://www.mta.info/mta/news/releases/?en=100113-HQ2.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50(1):181–201.

Montjoye, Y. A. de, Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports*, 3.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Openshaw, S. and Taylor, P. J. (1981). The modiable areal unit problem. *Quantitative geography: A British view*, 9:60–69.

O'Reilly, T. (2010). Government as a Platform. In Lathrop, D. and Ruma, L., editors, *Open Government. Collaboration, Transparency, and Participation in Practice*. O'Reilly Media.

Overman, H. G. (2010). "GIS a job": what use geographical information systems in spatial economics? *Journal of Regional Science*, 50(1):165–180.

Pivo, G. and Fisher, J. D. (2011). The Walkability Premium in Commercial Real Estate Investments *Real Estate Economics*, 39(2):185–219.

Peng, R. D. (2011). Reproducible Research in Computational Science *Science*, 334(6060): 12261227.

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., and Strogatz, S. H. (2010). Redrawing the Map of Great Britain from a Network of Human Interactions. *PLoS ONE*, 5(12):e14248.

Rauterkus, S. Y., Thrall, G. I., and Hangen, E. (2010). Location Efficiency and Mortgage Default. *Journal of Sustainable Real Estate*, 2(1).

Rey, S. J., Anselin, L., Folch, D. C., Arribas-Bel, D., Gutierrez, M. L., and Interlante, L. (2011). Measuring Spatial Dynamics in Metropolitan Areas. *Economic Development Quarterly*, 25(1):5464.

Rodríguez-Pose, A. (2011). Economists as geographers and geographers as something else: on the changing conception of distance in geography and economics. *Journal of Economic Geography*, 11(2):347–356.

Rogers, S. (2011). *Facts are Sacred: The power of data*. Guardian Books, first edition.

Shadbolt, N. (2010). The Linked Data Revolution. In *Innovating Through Information Lecture Series, London School of Economics*.

Skupin, A. and Agarwal, P. (2007). Introduction: What is a Self-Organizing Map? In P. Agarwal and A. Skupin, editor, *Self-organizing Maps: Applications in Geographic Information Science*. John Wiley, Chichester, Sussex.

Steenbruggen, J., Borzacchiello, M., Nijkamp, P., and Scholten, H. (2011). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, 78(2):223-243.

Sui, D. Z. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, 32:1–5. Editorial.

Sui, D. Z. (2012). Looking through Hägerstrand's dual vistas: towards a unifying framework for time geography. *Journal of Transport Geography*, 23:5–16.

Talen, E. and Koschinsky, J. (2013). The Neighborhood Quality of Subsidized Housing. *Arizona State University: GeoDa Center Working Paper*.

Trulia, Inc. (2012, accessed Sept. 5 - 2012). `http://trulia.com`.

Twitter, Inc. (2012, accessed Sept. 5 - 2012). `http://twitter.com`.

UN Department of Economic and Social Affairs (2008). World Urbanization Prospects. The 2007 Revision. Highlights. Technical report, United Nations, New York.

U.S. Census Bureau (2010). American Community Survey, 5-year estimates. Accessed using American FactFinder; http://factfinder.census.gov/home.

Walk Score (2012, accessed Sept. 10 - 2012). http://walkscore.com.

Warf, B. and Sui, D. Z. (2010) From GIS to neogeography: ontological implications and theories of truth. *Annals of GIS*, 16(4):197-209.

Zillow, Inc. (2012, accessed Sept. 5 - 2012). http://zillow.com.