Applied Geography 49 (2014) 45–53

Contents lists available at ScienceDirect

Applied Geography

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

Accidental, open and everywhere: Emerging data sources for the understanding of cities☆,☆☆

Daniel Arribas-Bel*

Department of Spatial Economics, VU University, De Boelelaan, 1105, 1081 HV Amsterdam, The Netherlands

Keywords:
Data sources
Open data
Cities

A B S T R A C T

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.

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 surpassed (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 behavior. 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 pressure 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 creativity 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

☆ This article belongs to New Urban Worlds: Application, Policy, & Change.

☆☆ 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.

* Tel.: +31 20 5986090.
E-mail address: darribas@feweb.vu.nl.

[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.

0143-6228/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apgeog.2013.09.012

are appearing. Similar to other fields (e.g. see Edelman, 2012; Einav & 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 Challenges); 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.

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 government 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 & Cheshire, 2012), may prove more fruitful. Section Challenges 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 benefited 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 beneficial, both in private (e.g. geo-targeted 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 (2013). The rest of the text is structured as follows: Sections "Citizens as sensors": collection data from the bottom-up, Businesses moving online (and crating data in the process) and Open Governments, open data 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 initiatives led by different actors that serve as real illustrations; Section Challenges discusses some of the challenges that these new data sources pose when contrasted with the ones traditionally used by the social sciences; and Section Concluding remarks concludes with a few remarks and highlights.

"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 behavior (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,

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

[3] According to a recent study (Meeker, Devitt, & Liang, 2012), the number of mobile users of the Internet is expected to exceed that of desktop users before 2015.