IDB
Inter-American
Development Bank

Infrastructure, public policy and the challenge of big data

Infrastructure and Energy Sector

Walter Sosa Escudero

TECHNICAL
NOTE Nº
1847

January 2020

Infrastructure, public policy and the challenge of big data

Walter Sosa Escudero

Universidad de San Andrés

January 2020

Inter-American Development Bank

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Sosa Escudero, Walter.
Infrastructure, public policy and the challenge of big data / Walter Sosa Escudero.
p. cm. — (IDB Technical Note ; 1847)
Includes bibliographic references.
1. Big data. 2. Machine learning. 3. Public utilities-Technological innovations. 4. Infrastructure (Economics)-Government policy. I. Inter-American Development Bank. Infrastructure and Energy Sector. II. Title. III. Series.
IDB-TN-1847

Key words: Infrastructure, Public Policy, Big Data, Machine Learning
JEL Classifications: C80, L91, L94, L95

This document is a product of the research program developed for the preparation of the Inter-American Development Bank 2020 flagship report: Infrastructure Services in Latin America. To know all the documents from the research program, see: www.iadb.org/infrastructureservices

http://www.iadb.org

wsosa@udesa.edu.ar

Infrastructure, public policy and the challenge of big data

Walter Sosa Escudero

Abstract: This article discusses the possibilities of using big data and machine learning strategies to improve the provision of public services along many dimensions, including better access or quality and cost reduction. Several examples are presented for the case of water, electricity and transportation. The note also highlights the challenges faced by the infrastructure sector and, in general, by the public sector in relation to the possibilities and limitations of big data.

1. Introduction

In 2016 there were approximately 70.8 million *smart meters* (intelligent devices that measure electricity usage) installed in the US, that produced 6.786 million reads per day every 15 minutes (Yin et al., 2019). Figures like this are clear examples of the *big data* revolution, that has already reached every aspect of society, including, as expected, the provision of public utilities like water, electricity or transport.

The terminology "big data" refers to the most salient feature of this new data driven paradigm: the massive size of information that contrasts markedly with data arising from traditional sources like structured surveys, administrative records or experiments, whose datasets usually contain -at most- a few thousand points.

However, the big data revolution consists of more than just size: it refers to data generated by the spontaneous interaction with interconnected devices. It is this spontaneous nature what is behind the massive size, and what separates big data from other traditional sources of information. Surveys or experiments are *structured* mechanisms of producing data, in the sense of being specific about the relation between

the data and a concrete population to which they refer. That is, *before* data is collected there is either a sampling framework or an experimental design that guides data collection of productions, that is later explicitly used to refer to a certain population. On the contrary, big data involves massive *unstructured* data sets, mediated by the use of a device like a cellular phone, GPS, smart meter or a computer, of a markedly observational nature that, by construction, do not obey to an obvious sampling framework or experimental design. Consequently, big data is not just more of the old data but a completely new phenomenon, to the point that *new data* might provide a more reliable characterization, as suggested in Seth Davidowitz (2018) and Sosa Escudero (2019).

Machine (or statistical) learning, statistics and artificial intelligence are the algorithmic counterpart of the data revolution, understood as the computer or mathematical strategies used to process, store and analyze these massive data sets, and are as an integral part of the big data paradigm as the data themselves. “Data science” is a rather new terminology that, when not used carelessly, refers to a new paradigm that integrates traditional actors of data analysis (like statistics and every field of knowledge or practice that demands empirical studies, from economics to biology or linguistics) with computer science, engineering, and even “soft” but equally relevant disciplines like design or communication, in an iterative and highly interactive fashion.

No part of the society remains absent of the data revolution, and the provision of infrastructure services is not an exception. This note explores the possibilities of using the big data and machine learning strategies to improve the provision of public services along many dimensions, including better access or quality, and cost reduction. The note also discusses the challenges faced by the infrastructure sector and, in general, by the public sector in relation to the possibilities and limitations of big data.

2. Big data and machine learning for infrastructure

As stressed previously, big data is not just "more data" but a new paradigm, that has advantages and disadvantages, some known and obvious and some not. Standard statistical methods (including classical econometrics) usually operate with data for which there is an underlying *model,* thought to provide a reasonable abstract representation of the mechanism that generates data, that can be used to describe, forecast or design policy interventions. In this setup, most of the effort of classical empirical analysis was put in recovering unknown aspects of these models ("parameters") through data and statistical techniques. In the infrastructure sector, elasticity demand estimation, efficient frontier analysis or measuring the degree of competitiveness are classical examples of the use of data to inform salient aspects of the industry. Such exercises demand a well-structured economic model that can be taken to the data and, from a statistical perspective, most of these exercises involve *estimation* of unknown aspects of the model.

The availability of massive data sets changes this approach radically. In the previous vision, the model is taken as given, usually provided by some theory of previous experience, and the few data available were exploited to extract the most of a pre-specified model. The *machine learning* revolution together with big data change these roles. Abundant data is now used not just to estimate a model but to *build it* and continuously update it in light of new information, adapting and modifying it towards a certain goal.

By its nature, most machine learning strategies are usually *predictive* or *exploratory*, as opposed to "prospective". That is, machine learning methods are used to provide a representation of the data that can be used to classify, discover patters, form groups or make predictions. On the other hand, prospective strategies are closer to the modern approach of econometrics, whose main goals are related to measure causal effect or to build an empirical "structure" that can be used not only to make predictions but also to understand how an economic phenomenon works and quantify its performance under alternative scenarios.

A relevant case for infrastructure that helps to understand the implementation of big data ideas refers to the increasingly popular *demand response* mechanisms for electricity provision. Such schemes provide incentives to users to modify their electricity consumption according to factors that alter supply and demand, like specific weather conditions or power line damages. Different schemes act through differential rates or "credits" for proper usage. The implementation of such strategy requires considerable information processing that allows consumers to be able to monitor consumption in real time and vary their behavior according to the proposed demand response strategy. The implementation of such process requires complex models to forecast demand in real time, based on factors like weather or other regional information. Yin, Ghatikar and Piette (2019) present a detailed study based on modern predictive analytics techniques based on smart meters data, and estimate the potential benefits from participating in demand response programs for San Francisco and other areas.

This type of analysis is a clear illustration of the potential of big data and analytics in the electricity sector, in the sense that it is based on a massive data set produced by information generated by the *passive* use of the system through a sensor, matched by flexible machine learning methods. By "passive" we mean that there is no active data collection strategy like in the case of a survey or an experiment. It is just the operation of the sensor -that connects consumers and producers in a two/way fashion- what generates the relevant information and allows producers to design and fine tune a pricing/rewards strategy and, simultaneously, consumers to adapt their behavior, resulting in a more efficient match of demand and supply. Throughout the process, producers learn about the incentives mechanisms and consumers to alter they electricity consumption behavior, like, for example, whether to operate dishwashers overnight or, electronically, use an app or interconnected device to react to events that alter electricity consumption (sudden wheather changes, for example) that are properly communicated by the electricity provider. It is the combination of massive, non-structured information and immediate processing what allows these instantaneous and synchronic changes in electricity supply and demand by both consumers and producers.

The previous example is of a markedly predictive/descriptive nature, and is a prototypical case of a modern data science implementation. This contrasts with the imperative use of *causal* studies embraced by the economic profession in the last twenty years. Such "credibility revolution" in economics (Angrist and Pischke, 2014) usually favors "clean" cause/effect mechanisms that help measure the impact of policy interventions, and usually demand the implementation of experiments or quasi experiments, understood as situations where data is generated spontaneously but, through institutional analysis, can be safely thought as if it were generated by an experiment. On the contrary, machine learning strategies are largely observational and hence may fail to distinguish causation from mere correlation.

In an illuminating paper, Kleinberg et al. (2015) point out that policy or project evaluation requires the concurrence of causal impact evaluation studies *and* predictive strategies. In their paper they discuss the case of a policy that facilitates hip replacement surgery, whose evaluation requires to know the potential effect of implementing the surgery in each patient (which requires a causal study) and the life expectancy of the target population (a predictive exercise). In the case of demand response for electricity, a full evaluation of the strategy requires a counterfactual study of the potential changes in demand according to a particular scheme (a causal analysis) together with a predictive strategy for electricity capacity according to weather or system malfunctioning, like the one described previously. Consequently, it is not causal analysis vs machine learning but *both* what is required for a proper evaluation of a policy. Athey (2017) presents a relevant discussion on big data and policy evaluation and implementation.

It is relevant to remark that the spontaneous nature of big data leaves ample room for causal analysis, either in the form of explicitly designed experiments implemented by the data collecting system, or by exploiting massive data sets to select proper subsets of information that obey a (quasi) experimental design. In the demand response example, causal analysis might be conducted by designing experimental exogenous variations in prices or rewards using randomized control strategies, or by exploiting panel (temporal and cross sectional) variability, that is by collecting data on electricity usage as a response to controlled variation in prices or rewards not caused by supply effects (like

weather forecasts, for example). Einav (2014) is a relevant example on how big data sources can be used to exploited to extract relevant data for causal analysis, in a quasi experimental fashion.

Public transportation is a fertile area for the interaction of big data and public policy. Accessibility to educational, health and labor market institutions is a crucial aspect of policies aimed at increasing social inclusion, a particularly challenging issue in most large urban areas of South America. Cordoba is one of Argentina's largest cities, actually the largest in terms of administrative surface (242 square kilometers) that relies heavily on the efficiency of its bus system. Deviations from prescribed bus routes are common and unexpected, due to road obstructions, traffic accidents, political demonstrations and pickets. Such problems affect relative more the poorest neighborhoods of the city, where no alternatives of public transport are available (trains, taxis, etc.). These deviations affect not only the efficiency of access but the safety of the population: on January 2019 a woman was murdered in the outskirts of Cordoba, when a bus left her in a dangerous area after deviating from its established route due to heavy rain.

A multidisciplinary team of data scientists was formed to help the Secretary of Transport of the city of Cordoba, with the coordination of CAF Banco de Desarrollo. The team exploited a database of real time bus tracking information obtained through the mandatory GPS devices that monitor the position of 600 buses every 20 seconds (and, at the time of the project, a largely unexploited source of information). The city was partitioned in a grid, so prescribed routes constitute a contiguous sequence of cells in such grid. An algorithm was programmed to detect departures in real time, and soon a convenient app for cell phones was made publicly available to alert travelers of route changes. Interestingly, the output of this project is currently being exploited to help re-design routes, to predict possible deviations (in addition to just monitor them) and to study the social impact of transportation by matching the GPS database with socioeconomic information of neighborhoods and households in the city of Cordoba. Quite remarkably, the project was implemented at a very low cost and in a period of only two months, and will soon be open to enable its implementation in other cities. This case, toghether with other implementations of data science projects in public

policy, is part of the Hands on Data project implemented in Argentina, and described in Berniell et al. (2019)

As a final example of the uses of big data technology in infrastructure, the Barcelona Supercomputing Center collaborates with the city's Aigues de Barcelona in several data science initiatives to improve water provision. One relevant project exploits sensor data from the Sant Joan Despí plant to predict dirt accumulation in membranes used to filter water, a factor that increases the power needed to pump water, at a higher cost. Using big data information it is possible to predict the state of membranes and plan the cleaning process in advance, which optimizes energy and has environmental benefits from lowering the carbon footprint (Deign, 2019)

All the examples discussed point towards the enormous potential of big data, machine learning and artificial intelligence in the infrastructure sector, in terms of improving the quality of services, help predict and manage failures, optimize pricing, tailor services to customers and help with environmental concerns. The sectors of the society that are most optimistic about big data usually overemphasize similar successful examples, mostly coming from the private sector, where efficiency and profits play a dominant role in strategic design. Nevertheless, the highly regulated and politically complex environment in which the infrastructure sector operates imposes particular restrictions on the adoption of data based technologies. A discussion of these challenges to adopt big data technologies in infrastructure and the public sector is the subject of the next section.

3. Big data challenges for infrastructure public policy

As stressed in the Introduction, the adoption of big data, machine learning and artificial intelligence technologies reaches all sectors of the society, but, as expected, the public sector lags behind compared to its private counterpart. Efficiency and cost reduction benefits are obvious goals in both sectors, but public policy usually involves multiple and sometimes contradictory goals where, for example, efficiency benefits are evaluated along transparency, privacy, political and communicational concerns.

For example, in terms of efficiency, machine-learning methods exploit data and modeling in a way that, in pursue of efficiency or accuracy, might converge rapidly to "black box" mechanisms that, though reasonable for experts, may be suspicious or difficult to comprehend and accept by consumers, politicians or professional communicators. This concern is particularly relevant in less developed countries with weak institutions, where the "cocktail" of complex and obscure algorithms fed with suspicious data may open the door to difficult political negotiations for the adoption of new technologies, even when otherwise (in terms of efficiency) beneficial.

All these concerns affect the infrastructure sector, which operates either under the structure of the public sector itself or in a heavily regulated framework. In a recent study on the electricity sector, Shuelke-Leech at al. (2015) remark that "utilities are conservative organizations that are generally not digitally savy", due to the fact that new technologies are usually imposed on rigid and many times traditional organizational structures. The authors also suggest that due to their financial structure and regulatory mechanisms, the infrastructure sector operates with low profit margins, hence "electric utililties have relatively low internal incentives to adopt new technologies, particularly when the longer-term benefits accrue to other stakeholders" (Shuelke-Leech et al. 2015, pp. 942).

As expected, privacy concerns are also important for public policy. Sensor based data contains microscopic data that, as discussed previously, is key to implement efficient policies like demand response incentive schemes. Such data sets matched with socioeconomic information (from both, standard and big data sources) might be extremely helpful in the design and implementation of key policies like social tariff schemes. Alejo, Marchionni and Sosa Escudero (2008) find that, for the case of Argentina, based on existing data sources (household surveys) it is extremely difficult to target the poor based on their gas or electricity consumption, due to the low correlation with income, which complicates the design and monitoring of a social tariff scheme. Richer data sets like those available from smart sensor matched with standard survey data might help overcome this difficulty, through more accurate and flexible machine learning methods. But the detailed level of information required for this type of strategy

has to deal with dangerous threats like identity theft, cyber-stalking or other privacy concerns that may increase the political costs of negotiating and adopting social tariffs policies based on big data and artificial intelligence strategies.

The non-structured, sometimes anarchic nature of big data (in contrast, again, to survey or experimental data) may lead to problematic *biases*. In 2012 the city of Boston adopted the Street Bump system (Simon, 2014). Based on an app for cellular phones, using geolocalization, the system detects the existence of bumps in streets automatically, through information sent online by users of the app. The idea was to replace "physical" street inspections by the spontaneous and automatic reports sent by users of the provided app. A successful implementation would just be a matter of inducing drivers to use the app, with an appropriate rewards system, much like that used by Waze. The results of implementing the system were surprising: more bumps were detected and street repair workers sent to richer neighbors, not because streets were more deteriorated there but because cell phone usage was more intense than in poorer regions, biasing the provision of infrastructure services towards the rich. As expected, such biases are not difficult to solve, but the case points to the importance of not using big data information naively, as if it were survey or experimental data. To the point, a relevant challenge of machine learning is to provide a reasonable structure to otherwise unstructured information. In this case, the interaction of big data information (like that providing from an app) and a relatively small structured (ideally random) survey could have help to correct biases and fine-tune the algorithm.

A final challenge for the implementation of modern data science in infrastructure affects the public sector in general. Both operate in bureaucratic and rather conservative environments, with well-established hierarchies, usually a byproduct of the complexity of its operation and the multiplilcity of its goals. Successful modern data science implementations require a truly multidisciplinary approach that might be difficult to replicate rapidly in the complex structure of traditional organizations. The main goal of the Hands on Data initiative (reported in Berniell et al. 2019) was to facilitate the interaction between public officers and data scientists. A salient aspect of this project is the wide differences in language, background, motivation and culture of these two

worlds that require a considerable effort to negotiate rapidly. It is crucial to facilitate the interplay between these two groups of actors to foster cooperation instead of competition. The successful case of buses the city of Cordoba, discussed in the previous sections, is one of the products of the Hands on Data initiative and a promising example of what can be obtained when collaboration between the worlds of the public sector and that of modern data science is cleverly facilitated.

All these concerns may help explain why the adoption of big data technologies is still slow in the infrastructure sector. In the study cited previously, Schuelke-Leech et al. (2015) conduct a detailed empirical study on the prevalence of big data in the communications between policymakers, regulators and companies in the electricity sector of the US. More concretely, the authors use modern computational linguistic tools to analyze the *corpus* of transcripts of the Congressional Record and Committee Hearings, that record hearings and reports at the federal House of Representatives and Senate, and testimony and questions in committee hearings.

Quite surprisingly, the authors find that the first appearance of the term “big data” is in 2008, well after such technology reached the interest of the public sector. Overall, the study finds that “data issues are really not an integral part the dialog around electric utilities with policymakers”.

4. Final remarks

There are enormous opportunities for the use of data intensive technologies in the infrastructure sector, as the examples discussed in this note point out. Sensor technology is found to help to better forecast customer behavior and external factors (like weather in the case of electricity or political demonstrations in the case of transportation), to target and tailor customers, to improve the efficiency and quality of the services provided, and to deal with environmental concerns. The potential is considerable and comparable to the benefits shown by the implementation of artificial intelligence and machine learning technologies in the private sector, usually the most optimistic part of the society in terms of adoption of data science.

Still, the highly regulated environment in which the infrastructure sector operates puts it in a comparable situation to that of the public sector, which operates in a politically complex, bureaucratic and sometimes conservative framework that requires a delicate balance regarding the adoption of innovations. All these concerns are particularly relevant for the case of Latin America, due to its relatively weak institutions, which further delay the adoption of new technologies that require considerable consensus to function properly. In particular, big data and artificial technologies call for a truly interdisciplinary approach where the complexity of the public sector plays a key role, comparable to that of technology itself. Clever public interventions that facilitate the interaction between public officers, regulators, customers, sector representatives and data experts are of utmost importance.

References

Angrist, J. and Pischke, J., 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." Journal of Economic Perspectives, 24 (2): 3-30.

Athey, S., 2017, Beyond prediction: using big data for policy problems. Science, 355(6324), 483-485.

Berniell, L. Acion, L., Lombardi, L., Altszyler, E., Sarraute, C., Vazquez, A., Gravano, A. and Sosa-Escudero, W., 2019, Hands-on-Data: Artificial intelligence for the design of public policy in Latin America, CAF Working paper.

Deign, J., 2019, How big data and AI are improving the city's water supply, The Network, Cisco.https://newsroom.cisco.com/feature-content?type=webcontent&articleId=2001827

Einav, L., Dan Knoepfle, Levin, J., and Sundaresan, N., 2014. "Sales Taxes and Internet Commerce." American Economic Review, 104 (1): 1-26.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." American Economic Review, 105 (5): 491-95.

Marchionni, M., Alejo, J. and Sosa Escudero, 2008, La incidencia distributiva del acceso, gasto y consumo de los Servicios Públicos, in Navajas F. (editor), La Tarifa Social en los Sectores de Infraestructura en la Argentina, Temas Grupo Editorial, Buenos Aires. 2008. ISBN: 978-950-9445-58-1.

Origlia, G., 2019, El desvío de un colectivo resultó mortal para una joven cordobesa, La Nacion, 1-8-2019.

Schuelke-Leech, Beth-Anne & Barry, Betsy & Muratori, Matteo & Yurkovich, B.J.., 2015, Big Data Issues and Opportunities for Electric Utilities. Renewable and Sustainable Energy Reviews. 52. 937-947. 10.1016/j.rser.2015.07.128.

Simon, P., 2014, Potholes and Big Data: Crowdsourcing Our Way to Better Government, Wired magainze. Available online at:

https://www.wired.com/insights/2014/03/potholes-big-data-crowdsourcing-way-better-government/

Sosa-Escudero, W., 2019, Big data, Siglo XXI, Buenos Aires.

Yin, R., Ghatikar, G. and Piette, M., 2019, Big dat analytics for electric grid and demand-side management, Berkeley Lab China Energy Group.