Tracking the digital footprint in Latin America and the Caribbean

Lessons learned from using big data to assess the digital economy

UNITED NATIONS

ECLAC

BigDATA

Digital economy for Latin America and the Caribbean

Thank you for your interest in this ECLAC publication

Please register if you would like to receive information on our editorial products and activities. When you register, you may specify your particular areas of interest and you will gain access to our products in other formats.

Register

UNITED NATIONS

ECLAC

www.cepal.org/en/publications

facebook.com/publicacionesdelacepal

www.cepal.org/apps

Tracking the digital footprint in Latin America and the Caribbean

Lessons learned from using big data to assess the digital economy

UNITED NATIONS

ECLAC

BigDATA

Digital economy for Latin America and the Caribbean

This report was prepared under the overall guidance of Mario Cimoli, Deputy Executive Secretary of the Economic Commission for Latin America and the Caribbean (ECLAC), and coordinated by Sebastian Rovira, Valeria Jordán and Jorge Alejandro Patiño of the Division of Production, Productivity and Management. The big data exercise was carried out by a team of experts, Veronika Vilgis, Yu-Chang Ho, Xin Jin, Kangbo Lu, Karla Rascon-Garcia and Matthew Reese, coordinated by Martin Hilbert. Valuable comments were received from Mario Castillo, Wilson Peres, Fernando Rojas and Nunzia Saporito of ECLAC.
This publication was produced in the framework of the activities conducted under the United Nations Development Account project, entitled "Big data for measuring and fostering the digital economy in Latin America and the Caribbean".

United Nations publication
LC/TS.2020/12/Rev.1
Distribution: L

Printed at United Nations, Santiago
S.20-00380

This publication should be cited as: Economic Commission for Latin America and the Caribbean (ECLAC), "Tracking the digital footprint in Latin America and the Caribbean: lessons learned from using big data to assess the digital economy", (LC/TS.2020/12/Rev.1), Santiago, 2020.

CONTENTS

Tables

Figures

Map

Image

SUMMARY

This report explores the opportunities and challenges of the systematic use of publicly available digital data as a tool for the formulation of public policies for the development of the digital economy in Latin America and the Caribbean (LAC).The objective is to share the lessons learned in order to advance in a research agenda that allows the countries of the region create alternative measuring tools based on the digital footprint.

Its substantial insights stem from the content of its six chapters, which assess the current state-of-art of complementary aspects of the digital economy in LAC, namely the labor market and digital skills; technology prices; micro-, small- and medium-sized enterprises; broadband; cryptocurrency; and social media. We show that the digital footprint left behind by labor market portals, e-commerce platforms, and social media networks allow obtaining unprecedented insights, both in terms of reach and detail. For example, in terms of reach, we are often able to gather naturally harmonized data for more than 15 LAC countries, track 2.5 million small enterprises across the region, or 35 million statements related to the Sustainable Development Goals (SDGs). In terms of detail, we are able to go beyond what official survey statistics contain and track the gender of small business ownership in the Caribbean, or the effective hourly salary for a specific skill like data entry by gender. Beyond what official administrative registries contain, we are able to distinguish active use of 3G and 4G mobile access, or the price for emerging technologies like drones. The report showcases some 30 figures that exemplify the presented opportunities.

The methodological contributions of the report stem from the lessons learned from this exercise of Big Data analytics. In this sense, the report serves a rough guide for practitioners interested in using modern data science for development policies. Naively, when thinking about the 'big data' paradigm, people seem to imagine that, with enough computational skills, one would simply go online to collect data, and the entirety of reality would suddenly stand point-blank in front of any observer in all of its details in real time. In reality, the process is more reminiscent of the proverbial inspection of an elephant by a group of blind people, where data scientists take the role of the blind touching very distinct parts of the whole, never being able to grasp it all at once, but trying to piece together irreconcilable pieces of evidence. Making sense of 'big data' in a meaningful way includes computational challenges, but goes much further, and touches on the definition of data science as the convergence between computer science, statistics, and its substantive application area. Issues of representativeness, generalization, harmonization, data quality, definition of variables and indices quickly become the main concerns of data science in practice. We report some of the lessons learned throughout this exercise, which took place during January and March 2019, and summarize them in the final chapter.

INTRODUCTION

A. Big Data for evidence-based policy-making

Policy proposals not based on empirical evidence are much less likely to be successful than policy designs thoroughly informed by the current state of reality. Empirical evidence is not sufficient, but a necessary requirement in order to assure the success of development interventions. While policy-makers need to know where they would like to go and how to get there, the starting base of all this visionary modelling is knowing where one currently is. Without knowing where one is sailing, it becomes a game of luck to select the favorable wind to get to a chosen destination. The lack of firm knowledge of existing realities contributes to the often-lamented fact that so many public policies are ineffective and miss their aspired impact. Only if policies are firmly grounded in empirical realities can we aspire for them to have lasting impact.

Unfortunately, developing regions like Latin America and the Caribbean (LAC) are notoriously short of empirical data, especially in new and emerging fields like the digital economy. On the one hand, the digital economy has already started to dominate the global economy. At the time of writing of this report (early 2019), seven of the eight most valuable companies in the world are from the digital economy: Amazon, Microsoft, Alphabet (Google), Apple, Facebook, Tencent, and Alibaba. Each of these has higher market capitalization than industrial giants like Johnson & Johnson or banks like JPMorgan Chase (ranks 9 and 10). As such, UN ECLAC has proposed to undertake policies of structural change for equitable and sustainable development that harness the full potential of the digital economy in order to promote a shift in the region's production structure towards more knowledge-intensive industries and higher-productivity sectors (CEPAL, 2018; UN ECLAC, 2018). On the other hand, the quick rise of the digital economy faces a mismatch with the inertia of the existing statistical apparatus. These incipient areas of the development agenda do not yet count with the required attention when we compare them with other more traditional statistical and academic research.

At the same time, it is in these new, emerging, and fast-paced areas where empirical inventories are especially useful. This is because in these innovative areas, policy recommendations themselves are often not yet comprehensively examined. Without a firm standing on empirical realities, policy-makers face the double uncertainty of an unknown starting point and a half-proven policy recipe to act on it. The combination of lacking policy options and evaluations, and the deficiency of the understanding of current realities decisively reduces the probability of success for public policies. If one is uncertain about the most favorable winds, not knowing where one is sailing can quickly become disastrous and push oneself backward.

While the lack of proven policy options is limited by the accumulation of experience (and therefore, ultimately, the passing of time), the main limitation for empirical evidence are resource constraints to obtain data. It therefore seems to be the more feasible challenge to tackle head-on and should not be postponed. Furthermore, resource constrains to obtain data should in theory be solved by the very outcome of the digital revolution: the abundance of data footprints society leaves behind with every digital

step it takes, aka 'big data' in form of digital trace data. We therefore propose to take advantage of this digital footprint produced by digital interactions in order to monitor selected aspects of the digital economy.

This leads to the proposal to continue with research methodologies that incorporate alternative measurement tools based on web data and the digital footprint, which takes advantage of publicly available online data to inform public policy making. Online communication provides a plethora of digital trace data that can be harnessed to obtain actionable insights. For example, job market sites disclose labor market dynamics, e-commerce platforms reveal technology availability and accessibility, online retail-platforms provide insights into the location of small enterprises and market concentrations, and social media depict gender inequalities and network access. We collected and analyzed data that provide substantive insights into six complementary aspects of the digital economy in LAC, namely the labor market; technology prices; micro-, small- and medium-sized enterprises; broadband; cryptocurrency; and social media. While presenting this data, this report also describes some of the lessons learned in the application of big data techniques for public policy purposes. By sharing these, the report pursues two complementary goals. It serves as:

- a current inventory of selected issues regarding the digital economy in LAC; and
- a basis for discussion about the opportunities and challenges of working with digital online trace data for development policies.

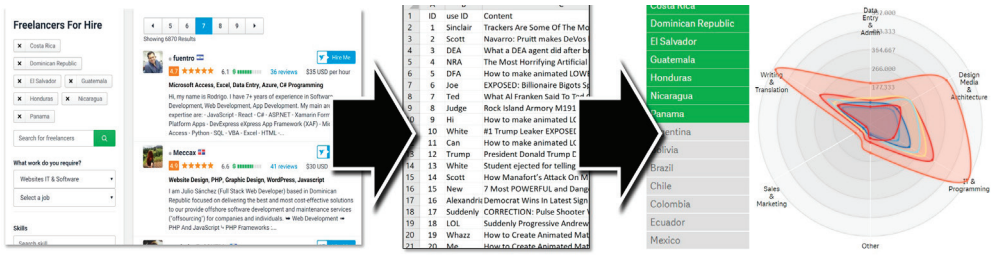
Both goals are complementary and should be discussed together. They represent the evolving character of working with empirical data in a digital age. Traditional approaches to measuring international development have faced data scarcity, but counted with a relatively high degree of quality from well-understood sources. In our undertaking we seek to access an abundance of data, but face severe challenges when it comes to understanding what each kind of data represents, and how it can be generalized, if at all. This leads to methodological challenges.

B. Methodological stepping stones of data science

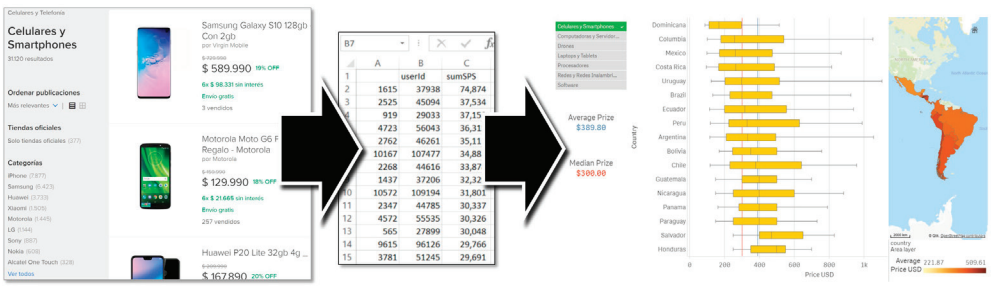
According to the leading complex systems scholar, Alessandro Vespignani, the three "toughest challenge[s]" of understanding today's "techno-social systems" consist in "first… the gathering of large-scale data… second… the formulation of formal models… [and] third… the deployment of monitoring infrastructures" (Vespignani, 2009, p. 428). Implementing and feeding this setup with the help of online trace data and computational analysis always follows a similar workflow, starting with the collection of online data, followed by methodological decisions and data cleaning, to visualization and analysis. Figure 1 presents examples of converting data from a labor market portal into insights on skill levels in Central America (A), and of converting data from an e-commerce site into the accessibility of technology across the region (B).

Figure 1
Schematic presentation of the data science workflow with digital trace data from data collection, Over data processing, to data visualization

A. Converting data from a labor market portal into insights on skill levels in Central America



B. Example of converting data from an e-commerce site into the accessibility of technology across the region



Source: Prepared by the authors.

In terms of the required skills in practical terms, each step can be done in a more or less sophisticated manner and requires different kinds and degrees of knowledge. In short, the first step requires primarily computational skills, the second predominantly statistical skills, and the final third step, requires analytical skills with specific domain knowledge. The result is the definition of data science as the convergence between computer science, statistics, and its substantial application area.

In principle, everyone able to navigate a mouse-curser could execute a simple version of the first step by copy-pasting content from the open internet. There are also several user-friendly out-of-the-box solutions available to automate this step, whose functionalities can be learned within minutes. Several of them offer assistance in online data collection against a service fee. For more sustainable and sophisticated applications, it is useful to have some basic programming experience. A basic knowledge of Python or another multipurpose programming language enables to set up custom-made solutions to obtain online trace data through harvesting web pages or by accessing provided databases. The second step mainly requires statistical skills. Researchers used to work with survey datasets, such as ad-hoc questionnaires or household surveys, usually count with the required skills to tackle this step. Any spreadsheet can be used to store data in a basic form. Going one step further, having knowledge of database management and an ability to work with database languages like SQL, allows practitioners to increase the size of the database, and to create a more sophisticated database design. As for the last step, any data visualization tool will serve to get started, including popular solutions like MS Excel or Google sheets. The main skill set required here is an in-depth understanding of the substantive area of application. This allows researchers to present a chosen set of meaningful variables in a useful way (which, of course, needs to be determined in order to guide the data collection step, feeding back into step 1). More flexible and interactive online dashboards are available on the market. Their use either requires

either programming skills (such as in R) or the acquisition of a commercial license for more user-friendly solutions.

In the following, we discuss these the different options more in depth and explain the implementation options we have chosen for the presented exercise.

1. Data Collection: tools and legal considerations

For this exercise, we gather digital trace data publicly available on the open internet. This digital footprint is often inevitably left behind by every digital step taken in the digital economy (Hilbert, 2016). There are two straightforward ways of accessing digital traces. One is by essentially recording the information displayed on the front-office of a website (either by web-crawling or web-scraping), and the other one is by accessing the database behind it, in the back-office by accessing an API (Application Programming Interface) provided by the data-processing platform.

a. Web scraping and crawling

The act of copy-pasting can be automated with either web-crawling or web-scraping. They essentially consist of machine-enabled copy-pasting of data from webpages. Whatever a user can view in a web-browser and manually copy-paste into some spreadsheet, can be collected by these programs. The difference is that web-crawling is typically used for the indexing of diverse webpage designs, while web-scraping takes advantage of the standardized design of a single platform.

A web-crawler, sometimes called a spider or spiderbot, is a program that systematically browses the World Wide Web, typically for the purpose of web indexing. For example, web search engines use crawlers to update their content by copying the content of webpages, which allows users to search more efficiently. As such, crawlers have been developed to collect content from webpages with the most diverse designs, which is a benefit in terms of reach, but the downside is that the collected content is usually not harmonized, contains many ambiguities and standardization challenges. In this report, we do not work with web-crawlers, but concentrate on web-scraping, which takes advantage of the harmonized design of a specific web-platform. The benefit of this approach is that the collected data is already pre-formatted. Furthermore, this approach well justified in LAC, a region that counts with many transnational portals that serve as natural data sources (figure 2). In order to extract data on technology prices, one can take advantage of this harmonized design. For example, the e-commerce platform mercadolibre.com operates in 18 countries of the region. In all of them, the platform uses a similar web-page design. Web pages are built using text-based mark-up languages (HTML and XHTML), which provide structure to the displayed content in text form. Users take advantage of this harmonized design to quickly assess the content without having to orient themselves anew each time. In the same way, machine programs can take advantage of the harmonized design to collect data. The program then basically accesses page after page, copies the specified content (not necessarily all of it, but the parts specified by the collector), and pastes it into a database (such as a simple spreadsheet, like an Excel or csv file). Naturally, only things that are accessible can be scraped. If the collector does not have access to a specific page (e.g. some hidden page) and cannot copy-paste it manually the scraper cannot access it either.

There are several out-of-the-box web scraping tools readily available that can be learned within minutes. For example, the open source application www.webscraper.io is a popular extension of Google's Chrome browser. The user opens the desired web page in Chrome, and specifies the content to be collected, basically by clicking at the content to be collected. There are several paid services, which facilitate the collection task (an online search for something like 'web scraping service' will provide a selection of those). They are usually offered for free for a limited contingent (such as 500 pages per month, etc), and also allow a for-pay option, which in many cases also provides assistance and customer service (most lay users require assistance when trying to scrape more modern webpage designs, such as pop-up content, or pages with diverse scrolling functionalities). The most sustainable solution consists in programming scrapers in-house, which is usually done with Python, a modern multipurpose programming language. Popular Python packages for web scraping include selenium (www.seleniumhq.org) and BeautifulSoup (www.crummy.com/software/BeautifulSoup). We employed both of these tools in the current exercise. Several online courses are available at no cost to teach web scraping with Python within a few sessions (depending of pre-existing knowledge of Python).

Figure 2
Examples of transnational online platforms in Latin America and the Caribbean

Source: Prepared by the authors.

b. Legal and ethical considerations

It is important to note that web scraping implies that one opens and closes the accessed page, and therefore uses the resources of the server of the company that provide the information. The program literally opens the page (if a page cannot be opened, such as with password protection, it cannot be scraped), then loads its content (if the content is not loaded, it cannot be scraped), and copy-pastes the content. The more people access a portal and load the content, the slower the service (in extreme situations, servers can be overloaded, which prevents other users to access the portal, which is sometimes purposefully done in aimed cyberattacks). Therefore, it is understandable that portal owners do not like data scientists continuously scraping their portal. The signature scraping leaves behind is similar to an automated cyberattack on the server. This is combined with concerns about intellectual property as the provider of the portal often owns the content in a proprietary way.

Therefore, an increasing number of portal providers specify in their terms of service if automated copy and pasting of content is allowed or not. Whether these specifications are merely binding for permission of the use of the service or can have legal consequences depends on the specific legislation and the kind of interaction with the content, as well as on the usage of the content. In principle, everything on the open internet can be seen and therefore is accessible by the public, but the portal and its content are still the private property of the provider. So, under many legislations, the owner has the right to ban users who do not comply with their particular rules of service. Think about it in terms of any brick-and-mortar store that can deny a badly-behaved individual to enter their private building. In this way, it makes sense that online portals can also specify that people who break their terms of service should not be able to interact with their data in their portals.

The question if obtained data infringes on intellectual property rights depends on the ruling legislation, as well on the nature of the collection and its reuse. Again, think about it in terms of an off-line service, such as a paper-based newspaper. If you would sit down on your kitchen table with a piece of paper and write down the price of all products offered in the second-hand section of the newspaper, and then report the average price of those in one of your publications, it is unlikely that the owner of the newspaper would feel you infringed upon any property rights. On the contrary, it might be welcomed that you provided additional visibility and added-value analysis to the service provided by the newspaper. It will draw more attention to the provided service. However, if you would type up all second-hand ads word-by-word and publish them in your own newspaper, this act of plagiarism clearly affects the business model of the newspaper owner, and you might face legal consequences.

In this sense, when copy-pasting any content from the open internet, it is important to inform oneself about the particular terms of service of a portal, and about the ruling legislation of the service provision, just in the same way that it is important to inform oneself about the terms of service of a brick-and-mortar store. If in doubt, we urge any data science practitioner to consult with legal experts of the ruling legislation for their case.

c. APIs: Application Programming Interfaces

Having the foregoing caveats in mind, web scraping or crawling is not the most straightforward, and therefore also not the most sustainable way to collect data. The best possible way to obtain online trace data in a sustainable way is through a so-called API (Application Programming Interface). For our purposes, APIs are essentially official access points to parts of the database of a provider's service or platform.

In data science, obtaining data directly from the back-office through APIs is usually considered a more official, controlled, and sustainable entry through the front door of the data-edifice; while web scraping and crawling from the front-page of services are a more uncontrolled backdoor entry to the displayed data. Online platforms cannot control who scrapes or crawls their content, but they can control who enters their API. Usually so-called tokens are provided, which give temporary access to the data, and, at the same time, track the data retrieval process, including previous registration, etc. This does not mean that one way of accessing online data is less common than the other. For example, Google's omnipresent search services (which have become indispensable to so many other services) are based on Google continuously crawling and indexing the World Wide Web. From the perspective of the data provider, the benefits of providing API access go beyond easing the processing load of the company's servers and providing a certain degree of control over data usage, but include taking advantage of positive network effects that happen when other services connect to the selected services provided by the company.

In general, an API is a software intermediary that allows two applications to talk to each other. Each time one uses any app one is using an API, including any social media platform and messaging service. Many platform providers offer more or less structured and regulated access to their API. This allows users or developers to access the whole or selected parts of the data processed by the platform provider. One common motivation of doing this is that developers from other services and apps can now systematically link their services to it, and therefore create a networked ecosystem that benefits both (including the sharing of customers, etc). For example, think about searching for a flight directly on an airline portal, or using some kind of intermediary travel agency, or a Google search. The travel agency or Google also has access to the flight database, but does not execute searches in the airline's website in order to retrieve the travel itineraries. It accesses some kind of API provided by the airline, which allows to retrieve the data without affecting the service provision on the consumer website of the airline. The API is the interface that can be queried by the online travel service to get information from the airline's database to book seats, baggage options, etc. The API then takes the airline's response to the request and delivers it right back to the online travel service, which then shows the customer the most updated, relevant information. Popular APIs include Google Maps (which is used by many travel and location services, which build their more sophisticated services on top of this data), or Google Translate. In order to take advantage of Google's powerful translation service in an automated manner, there is no need to go through its main customer portal https://translate.google.com. One can as well enter Google's cloud translation API directly (https://cloud.google.com/translate). This allows other service provider to integrate Google's translation service into their website (aka 'powered by Google Translate').

Most APIs are made for developers of other apps and programs, and require some rudimentary knowledge of a common programming language like Python. Since APIs are the more sustainable way to systematically obtain data from a chosen data source, it is useful to have such skills. Besides, API providers also use APIs in order to control who accesses their data, and for what purposes. Many require previous registration, including some personalized data and a statement of purpose of the exercise, and often grant permission only after a human review of the case. In many cases, users are then given so-called tokens (long codes of letters and numbers), which allow to retrieve a certain amount of data during a certain timeframe. Token can expire as quickly as in a few hours, and can be restricted to access for certain variables and/or request frequencies. In this way, the data provider assures that their servers will not get overloaded. Therefore, working with an API can sometimes require a detailed plan with regard to how much of what data can be obtained in what frequency. In this exercise, for some APIs we set up a self-developed protocol of wait-and-retrieval times, which assured us to balance the retrieval bottleneck provided by the API with a continuous influx of data (e.g. make 12 randomly timed data requests within one minute).

Most of the major international portals operating in Latin America and the Caribbean provide APIs, with selected variables, especially those portals that also have global reach. Whenever possible, we preferred API access to obtain our data. Data available through APIs is often restricted (since pre-selected by the provider), and not everything that is visible on the public web page might be available through the corresponding API. At the same time, monitoring digital traces for public policy purposes does not require very detailed data. Working with high level statistics that often summarize a variable on the country level, we are not interested in the personal data of one single individual, including privacy infringing identifiers. We work with highly aggregated data in the form of national averages or distributions. In order for projects like the one presented in this report to be sustainable in the long run, it would be desirable to have regulated and agreed-upon access to the sought-after variables from the APIs of selected service providers in the region.

2. Data Processing

The second step from figure I from above is often referred to as 'data wrangling' in data science. It involves all steps included from converting the collected data into a usable format for analysis. In the daily life of data scientists, this task is often the most time-consuming task and occupies up to 80 % of the data scientist's time in practice.

Like all observational data, online data is often incomplete (not every row and column is complete) or unreliable (not every data point can be taken at face value). However, it is important to remember that the resulting work is not more challenging than the work of traditional researchers working with offline observational data, such as ethnographers, who convert observations into rows and columns in their notebooks, or survey exercises, which must judge the level of truth and reliability of the provided answers. For example, organizations like UN ECLAC have worked for many decades on the intricate statistical challenge of harmonizing the results from household surveys from different countries. Not all variables might be exactly the same in every household survey, and some cases might be missing. The ensuing work of data cleaning and preparation includes decision about how to harmonize discrepancies and if and how to extrapolate over missing values. Statistical knowledge is required in order to be able to judge the representativeness of the obtained data (observational data is, per definition, never a representative random-sample, but always bias), as well as to assess the trustworthiness of the source and its specific content.

One recurrent characteristic is that online data scales very quickly, which often leads to the notorious phenomena of 'big data'. This applies both to many cases and to many interrelated variables. The resulting content lends itself to several intricate database designs. However, for the sake of general usability, in this exercise we stuck to a simple spreadsheet format in form of csv files (comma-separated-values), which are simple text-files in which different columns are separated by commas. Those can be opened and processed with any kind of spreadsheet program, including MS Excel or Google Sheets, but also more advanced statistical packages, like R, Python Pandas, SPSS, or STATA, etc.

3. Data analysis and visualization

The third and final step from our data science workflow consists in converting raw numbers into some format that lends itself to derive actionable insights. In our case, we opt for the deployment of a series of visualizations that allow analysts, policy-makers, scholars and the general public to gain informed insights in a timely manner. This is still a very superficial level of analysis, and most of the dozens of graphs and figures presented in this report could benefit from a much more in-depth analysis in form of a case study (in complement to qualitative insights from specific domains) or more sophisticated analytical methods (including multivariate analysis, like structural equation models, or other econometric examinations).

In the most rudimentary form, the obtained data can be visualized with simple graphs with the help of popular programs like MS Excel or Google sheets. This is precisely what we have done for many of the graphs presented in this report. The fundamental limitation with this approach is that it restricts the richness of presentation by the necessity to choose a predetermined number of variables (usually a selected subgroup of all available variables) and to decide on a certain level of aggregation (the conditioning variables on basis of which averages and distributions are assessed). The presented graph is only one of all possible graphs that could have been presented on the basis of the same data. This is useful to answer a specific research question or hypothesis,

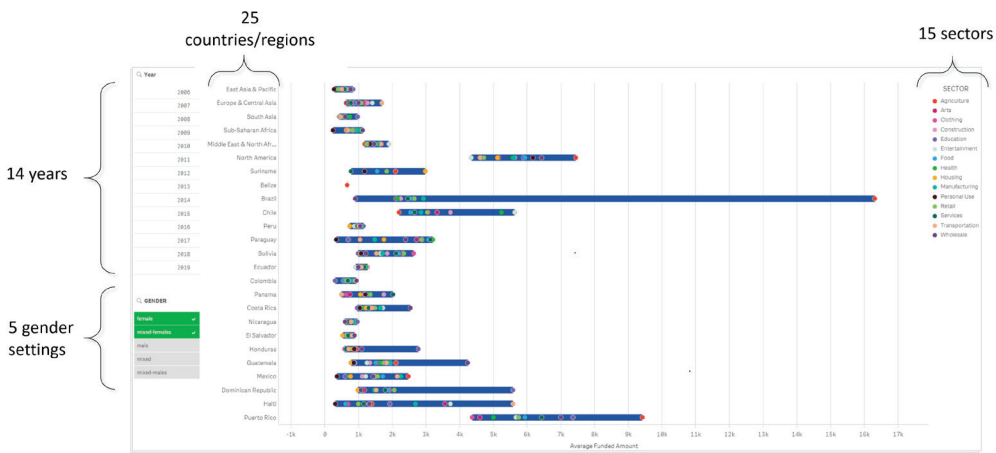
or to formulate a judgment with regard to existing arguments, but goes against the main premises of the 'big data' paradigm: taking advantage of all of the available data.

For example, a typical dataset collected for this exercise might contain 25 countries (or regions), tracking activity of 15 sectors, over 14 years, being able to disaggregate by 5 socio-demographic aspects (see for example the crowd-source finance analysis below, (in chapter III also figure 3). This implies a total of (25+15+14+5) = 59 variables. In the most basic form, we could choose a graph that shows all of them together, including all years, all sectors, disaggregated by all countries and by gender. This would be a very busy graph, most probably intelligible. In contrast, we could only present one single number, e.g. one sector for one year, in one country, only for men. We have exactly 59 of these insightful numbers. This would be hardly a graph at all, and miss important comparative aspects. In between, and in practice, someone with domain knowledge makes a deliberate choice about the selection of variables and their level of aggregation, choosing the most 'interesting', 'insightful', or 'meaningful' years/sectors/aspects, and/or their level of aggregation. This process of selection is inevitably more an art than a science, because the number of choices is almost unfathomable, which makes the selection of any specific graph so restrictive. The number of possible graphs is the sum of all possible values of the well-known 'n choose k' formula, where 'n' denotes the number of available variables, and 'k' the number of the chosen variables presented in a given graph:

$$Number\ of\ possible\ graphs = \sum_{k=1}^{n} \binom{n}{k} = \sum_{k=1}^{n} \frac{n!}{k!\,(n-k)!}$$

Choosing to present the combination of any two of the total 59 variables allows to create 1,711 different graphs. Creating a graph that focuses on three of them, leads to 32,509 different figures. Working with four variables, allows the researcher to choose from half a million different graphs (455,126 to be precise), and selecting any five of the 59 variables makes for more than five million figures (5,006,386). In total, the sum of all possible choices allows to make more than half-quintillion different graphs (exactly 576,460,752,303,423,000), which is more than a billion different graphs for every inhabitant of Latin America and the Caribbean.

Figure 3
Example screenshot of a dashboard that displays the average loan amount from 1,551,384 crowd-financed loans provided through Kiva.org



Source: Prepared by the authors.

Doing data science also implies to make choices among these unfathomable number of options. The first requirement to make this in any meaningful way is domain knowledge of the area under investigation. Without the guidance of informed research questions, it is unlikely that meaningful graphs are being selected. This is why data science goes way beyond computer science and also beyond mere statistics. The second step consists in trying to open the straight jacket provided by data representation tools. Modern data science therefore often does not work with static graphs, but with so-called dashboards. These are interactive tools that allow users to select and deselect variables, and analyze different aspects of the graph. The general public is getting used to such online dashboard assessments, as interactive line-graphs allow them to zoom into a certain time period (e.g. when looking at currency exchange rate histories), or when evaluating the changing number of the total price of their purchase by (de) selecting products in their e-commerce shopping cart.

In our case, we opted to work with such visualization tool. There are open source options available, such as R-shiny (which works on basis of the statistical package R, and has commercial extension: https://shiny.rstudio.com). Popular commercial options require a commercial license and charge a fee per user.

C. Areas of interest & data sources

The selected areas that were chosen for this exercise emerged as the common denominator between the interests express by LAC countries in the Digital Agenda for Latin America and the Caribbean (eLAC2020), and the pragmatic realities of available sources. Because there is an emerging interest both in measuring aspects of the digital economy and gender issues, the study aimed at disaggregating many of the obtained variables by gender. At the same time, we chose topics that are typically difficult or impossible to obtain data for using traditional methods.

For any digital tracking exercise with international development goals, it is beneficial to obtain data from digital platforms that cover several different countries with a single business model. This ensures data comparability and facilitates collection efforts. Fortunately, in Latin America and the Caribbean, several such pan-regional platforms are available, not least because of the common cultural, linguistic and geographical context. Therefore, such platforms quickly became our main focus when looking pragmatically at possible sources. These include Bumeran, Coin.dance, Facebook, Freelancer, Ookla, Kiva, MercadoLibre, Twitter and Workana. However, we also did exploratory exercises with other local sources (such as national portals and platforms), especially to check reliability and the representativeness of sources.

As a result of these different considerations, we selected six main areas of interest that show different aspects of the digital economy in Latin America and the Caribbean (see Table 1) , with a special emphasis on gender issues. All data was collected between January and March 2019 at UN ECLAC's headquarter in Santiago, Chile.

Table I
Topics of digital economy, web sources and number of observations, January – March 2019

Topics	Sources	Observations
A) Labor Market and digital skills		
a) The gig-economy of freelancers	Freelancer.com	N = 74,970 individuals
	Workana.com	N = 19,840 individuals
b) Traditional labor market contracts	Bumeran.com	N = 78,475 individuals
	BNE.cl	N = 1,254 individuals
B) Technology Prices		
a) Price Index	Mercadolibre.com	N = 192,092 products
b) Tech categories	Mercadolibre.com	N = 192,092 products
C) Small and Medium-Sized Enterprises		
a) Online retailer	Mercadolibre.com	N= 2,473,977 sellers
b) Crowd-funding	Kiva.org	N= 1,551,367 loans
	Kickstarter.com.mx	N= 888 loans
D) Broadband	Ookla's netindex	N= 9,360 observations
E) Cryptocurrency	Coin.dance	N=15,840 observations
F) Social Media		
a) Socio-demographics	Facebooks ads manager	N= 110,880 observations
b) Trending Topics: Sustainable Development Goals (SDGs)	Twitter	N = 35,818,261 tweets

Source: Prepared by the authors.

LABOR MARKET AND DIGITAL SKILLS

I

For an assessment of the labor market, we worked with three different sources. Freelancer and Workana are both gig-economy websites, where professionals offer specific skills and clients ask for assistance with specific projects. Freelancer is more global, while Workana is focused on Latin America. Both provide data on the supply and demand side of the labor markets, as the platforms offer professionals the opportunity to post a profile of themselves, and employers the opportunity to post job opportunities. We also worked with Bumeran, a job market site where companies post job opportunities. This only covers the demand-side of job offerings, while the digital footprint does not reveal who gets invited to interviews or who gets hired.

As a first step, we propose a scheme to harmonize categories. We did not work with machine learning classifiers, as done in other exercises (Amato et al., 2015; Kässi & Lehdonvirta, 2018a), since the platforms provided categories directly, which we then matched. Looking for job categories related to the digital economy, we find that there are several thousand categories in Freelancer, but only a couple of dozen in Workana. Bumeran has again a different taxonomy of job categories. We then used the harmonization mechanism in Table 2 to create ten comparable categories for all three sources.

Table 2
Harmonization scheme of the collected job categories related to the digital economy

	Freelancer	Workana	Bumeran (1)
IT & Programming	Websites, IT & Software |Mobile Phones & Computing	IT & Programming	Technology
Writing & Translation	Translation & Languages | Writing & Content	Writing & Translation	Communication
Design Media, and Architecture	Design Media, & Architecture	Design & Multimedia	Design
Data Entry and Admin	Data Entry & Admin	Admin Support	Administration, Secretary
Engineering & Manufacturing	Engineering & Science | Product Sourcing & Manuf	Engineering & Manufacturing	Construction, Engineering, Mining, Production, Health
Sales and Marketing	Sales & Marketing	Sales & Marketing	Marketing, Sales, Call center
Freight, Shipping & Transportation	Freight, Shipping & Transportation		Logistics
Business, Accounting, Human Resources, & Legal	Business, Accounting, Human Resources, & Legal	Legal, Finance & Management	Finance, Management, Legal, Human Resources, Education, Accounting Assistant, Commercial Analyst, Accountant, Admin and HR Analyst
Local Jobs & Services	Local Jobs & Services		Gastronomy, Crafts, Driver, Security, Cashier, Furniture Shipowner, Toilet Assistant, Production Operators, Polisher, Master Kitchen, Welder, Lifeguard, Mechanic
Other	Other		Insurance

Source: Prepared by the authors.

A. The gig economy of freelancers

An important and growing aspect of the digital economy is the so-called gig-economy (Heeks, 2017). Digital networks give workers flexibility and the opportunity to have additional sources of income. For an increasing part of the labor market, what were initially small side-jobs or ad hoc consultancies, have even become a main source of employment. Since many gig-economy opportunities involve services that can be provided from a distance, they also provide access to the global job market, since is an important fact to consider of our case of focusing on LAC. We have looked at two gig-economy platforms, the globally operating crowdsourcing marketplace Freelancer.com (founded in 2009, over 21 million users, headquarters in Sydney and offices in Southern California, Vancouver, London, Buenos Aires, Manila, and Jakarta), and Workana.com, which has its focus on Latin America (main concentration in Argentina, Brazil, Colombia, Mexico, Uruguay, and Venezuela, but being used throughout the region). In contrast to other inventories (Kässi & Lehdonvirta, 2018b), we look at both the supply-side of the labor market, which refers to digital platforms where professionals offer their services, and the demand-side, which consists of employers seek professionals by offering job opportunities.

The global platform Freelancer.com attracts notably more professionals than the Latin American platform Workana.com, even on a per-country basis. Most countries have 4-6 times more professionals on Freelancer than on Workana (with the notable exception of Brazil). While workers can sign up for both sites simultaneously, we speculate that access to global job opportunities through Freelancer.com may be an important incentive for workers in the digital economy.

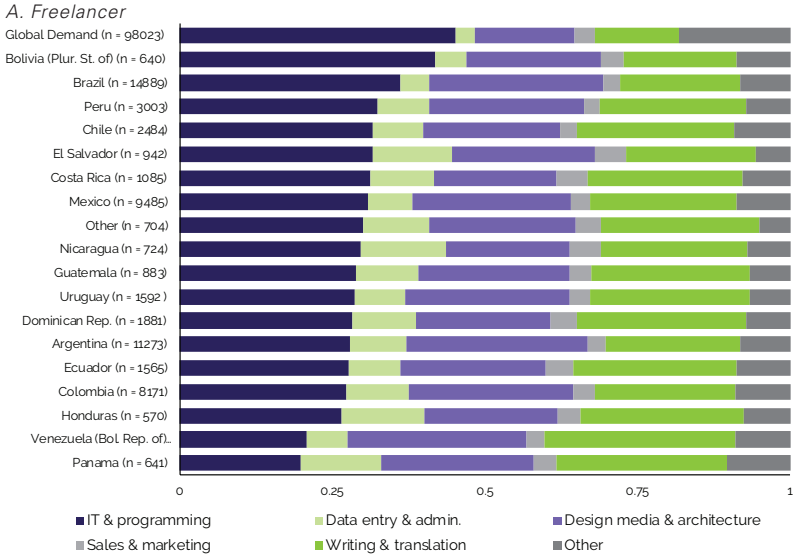
1. Supply of job categories per country

Figure 4 shows that the distribution of the national supply of skills is surprisingly similar among the most diverse countries, big or small, North or South. In most countries, 'IT & Programming' provides the largest pool in the digital economy, followed by both 'Writing & Translation', and 'Design Media & Architecture'. For most countries, and for both sources, the two key-sectors of the digital economy, 'IT & Programming' and 'Data Entry & Administration', represent 40 % of the labor supply. It is interesting to note that some countries surpass this relative share rather unexpectedly, such as Bolivia, El Salvador, or Nicaragua.

It is also interesting to note that 'IT & Programming' consistently occupies a larger share in the regional platform Workana.com, compared to the global platform 'Freelancer.com'. At the same time, 'Freelancer.com' offers more professionals in the related field of 'Data Entry & Administration', which leads to the question whether some of these differences might be due to inconsistencies in the classification system (see Table II). Future studies should check on this possibility, potentially by using a machine-classifier that does not rely on the classification used in the platform, but an in-house classification based on job titles or descriptions (Amato et al., 2015; Kässi & Lehdonvirta, 2018a).

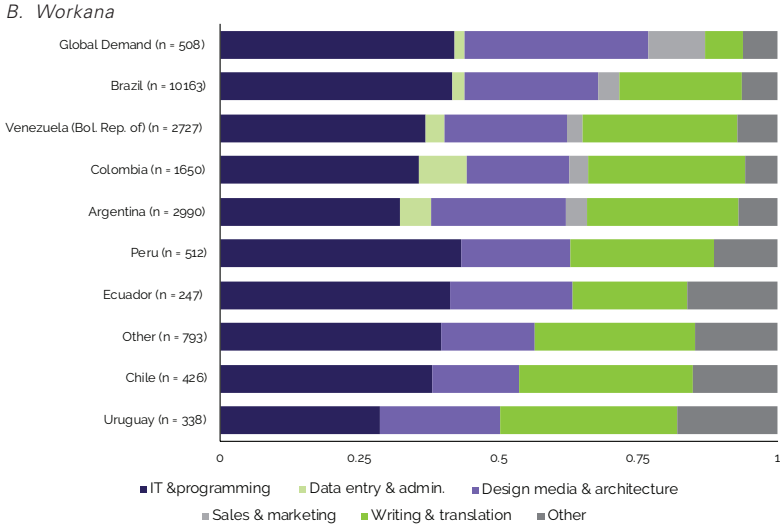
Figure 4
Proportions of professionals' job categories per country, February 2019
(In percentages)

A. Freelancer



Note: Other refers to: a) COUNTRIES - Cuba, French Guiana, Guadeloupe, Haiti, Martinique, Paraguay, Puerto Rico, and Saint Martin; b) CATEGORIES - Local Jobs & Services, Freight Shipping & Transportation, Engineering & Manufacturing, and Business Accounting Human Resources & Legal.

B. Workana



Note: Other refers to: a) COUNTRIES - Belize, Bolivia, Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Puerto Rico, Dominican Republic, Trinidad & Tobago; b) CATEGORIES - Engineering & Manufacturing, and Business Accounting Human Resources & Legal.

Source: Prepared by the authors on the basis of freelancer.com and workana.com.

2. Global demand and national supply

Both crowdsourced marketplaces, Freelancer and Workana, also offer the opportunity for employers to post job opportunities. Since our digital economy jobs mainly consist of services that can be performed remotely, we estimate global demand independently from their country of origin.

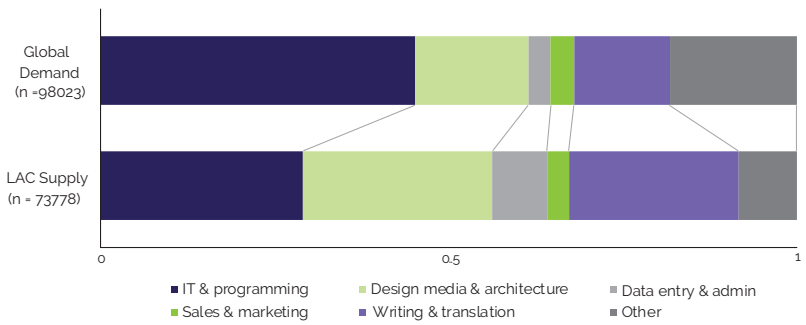
The first row of figures 4.A and 4.B shows the global demand followed by country-level supply in subsequent rows. Maybe the most striking finding is that in the global market of Freelancer.com (figure 4.A), the most demanded job category globally ('IT & Programming') is clearly underrepresented in LAC countries. In both sources, the region is clearly overrepresented with 'Writing & Translation' skills, as well as with 'Data Entry & Administration' (which becomes clear in the case of Freelancer.com (figure 4.A).

Figure 5.A and 5.B show a summary of the regional supply and the global demand for each platform. A correlation analysis reveals that the overall match is similar in both cases (for Freelancer.com $R^2 = 0.590$; for Workana.com $R^2 = 0.607$). However, as shown in the figure, this overall mismatch has different origins. As already mentioned, the 'IT & Programming' skills provided by LAC are not at par with the global demand for this skill, while the 'Writing & Translation' skills are overrepresented. This is a sign that the skill market in the region has not caught up with technological change, in which many writing and translation tasks are infused with artificial intelligence, and programming skills become an important aspect of daily business. The gap in 'IT & Programming' is notably larger for the global market of Freelancer.com. Another difference between both platforms refers to 'Design Media & Architecture'. For the global market of Freelancer.com (V.a), LAC skills are overrepresented, while for the regional market of Workana.com (V.b) the provided skill-level in LAC are not enough. The reason is not clear to us and this would benefit from some further in-depth analysis. There also seems to be relatively more demand for 'Sales & Marketing' in Workana.com, than in Freelancer.com.

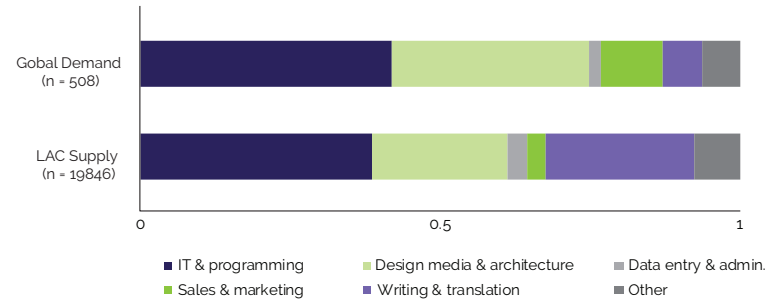
The same data can also be summarized by a heat map, which color-codes the demand-supply differences. In figure 5.C, the darker the cell, the larger the oversupply, the lighter the cell, the larger the undersupply. The underlying measure is proportional, which means that a value of 2.0 (see legend) implies that the national percentage of the supply in this category is twice as large as the percentage of the global demand.

Figure 5
Global job demand and national supply, per job category. a) Freelancer supply average of all countries; b) Workana supply average of all countries; c) Freelancer: heat map of ratios of national supply and global demand. February 2019
(In percentages)

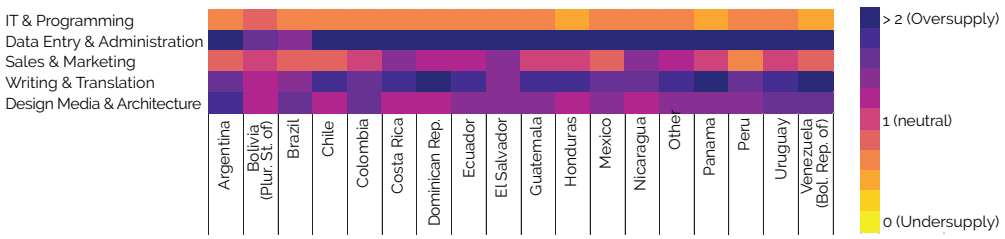
A. Freelancer – Global Demand and LAC Supply



B. Workana – Global Demand and LAC supply



C. Freelancer: Heatmap of Global Demand and National Supply. (Ratio)



Source: Prepared by the authors on the basis of freelancer.com and workana.com.

3. Hourly rate per country and job category

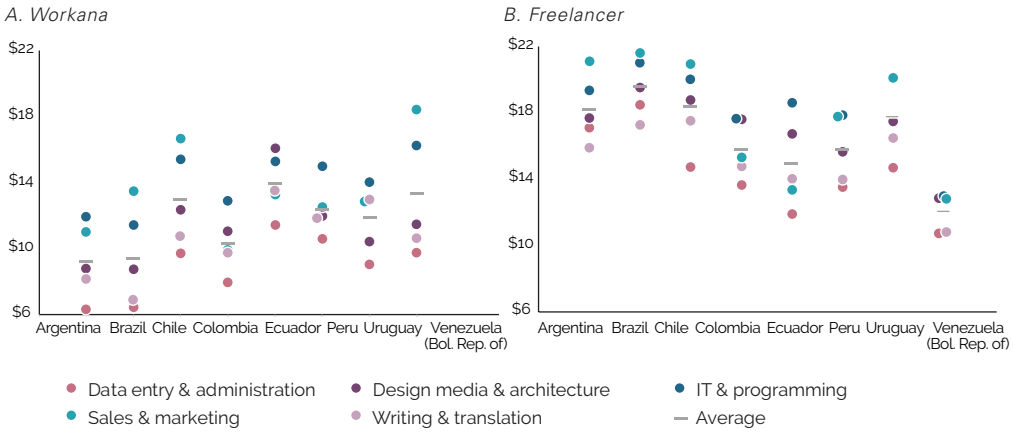
At crowdsourced job marketplaces, professionals post their hourly rate, which is rather to be seen as a wishful proposal and does not have any binding power. This being said, analyzing the resulting rates quickly leads to a problem of data quality. Some users seem to confuse currencies. It is likely that a person asking for "$10,000 per hour" in Chile is actually asking for US$ 15 per hour, or 10,000 Chilean pesos, not for US$ 10,000 per hour. The mistake would result from selecting the wrong currency when the professional inputs the hourly rate. For most currencies, this is easy to detect, especially when the amount asks for more than US$ 1,000, such as for the cases of Chile and Colombia, which had an exchange rate of around 670:1 and 3,200:1 at the time of collection. We calculated the respective US$ amount and we excluded all other cases where the hourly rate remained above US$ 1,000. For Workana.com, this reduced our sample from 19,846 profiles to 19,054 profiles.

As a result, we obtain a regional average hourly rate at US$ 14.8 for Workana.com, and US$ 22.03 for Freelancer.com professionals.[1] One could now rightfully point out that professionals who aim at the global market of Freelancer.com might ask for higher rates than the ones offering their services in the rather regional market of Workana.com. However, looking closer at the data also reveals some issues with data quality that might be linked to data quality differences in user-sourced data collection. It seems like some professionals confuse fixed price per job and hourly rates when filling in the boxes at the platform. In other cases, it seems clear that these rates are exaggerated aspirations from professionals, while those projects that are eventually carried out are quite different. We found cases were professionals from Mexico asked for US$200 per hour, but only ever earned merely US$10 and US$35 on two projects.

1 Bases on hourly rates in Brazil, Chile, Colombia, Ecuador, Peru, Uruguay, Venezuela.

We looked at more profiles individually and found very few profiles of professionals who could credibly ask for more than US$100 per hour. In addition, most projects seemed to have been executed with a lower hourly rate (one cannot derive the hourly rate from total projects, but get a qualitative idea about the involved amount of work). Therefore, figure 6 presents the average hourly rates when deleting those hourly rates that ask for more than US$100. The average hourly rate reduces to US$ 16.5. This is a significant drop, by eliminating less than 2 % of the profiles (N = 55,132) (figure 6.B). It compares with the adjusted hourly estimate for Workana.com with a cut-off of US$100, which is US$ 11.7 (N = 18,994) (figure 6.A).When working with digital trace data there is no true answer to the question of what is the correct hourly rate asked by professionals on freelancer platforms. It turns out that in our exercise, the ordinal rank order of several sectorial hourly rates is in the same order when comparing among countries, with 'Data Entry & Admin' being the least paying sector, followed by 'Writing & Translation', and 'IT & Programming' being the most lucrative sector. This observation would suggest that the differences between both platforms are absolute and stable among skillsets, and that average hourly rate in the regional market of Workana.com is proportionally about US$5 below the average hourly rate on the global market of Freelancer.com.

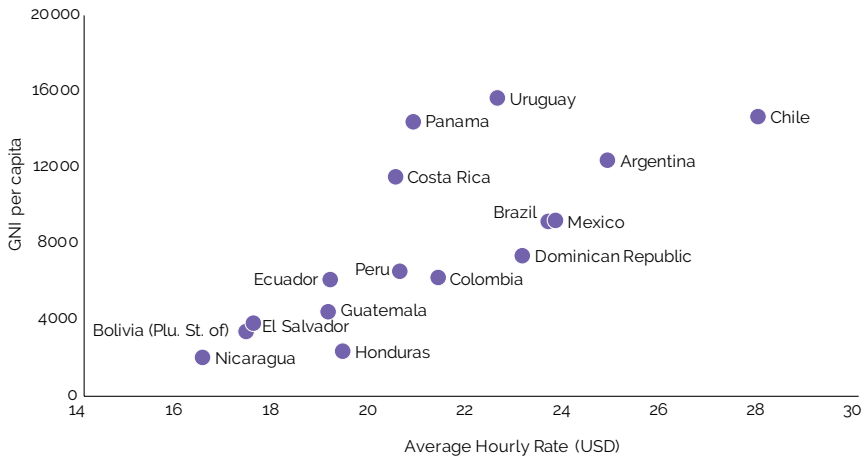
Figure 6
Mean hourly rate (asking price) of freelancers per country and job category, cut-off at US$100. February 2019
(Dollars)



Source: Prepared by the authors on the basis of freelancer.com and workana.com.

Figure 7 puts mean hourly rates from Freelancer.com (US$ 1,000 cut-off) into the context of national income. The hourly rates on the platform correlate strongly with real-world income. Every country more to the right than the average has a relatively high hourly rate, compared to the level of their economy (such as Colombia, Mexico, and Dominican Republic), while countries higher up than the rest have a higher national income, and relatively lower hourly rates (such as Panama, Costa Rica, and Uruguay). In other words, in the context of their national economy (opportunity cost with other income generating sources), it seems that freelancers of the global digital economy in Mexico and Dominican Republic have a comparatively higher salary than the ones in Costa Rica or Uruguay.

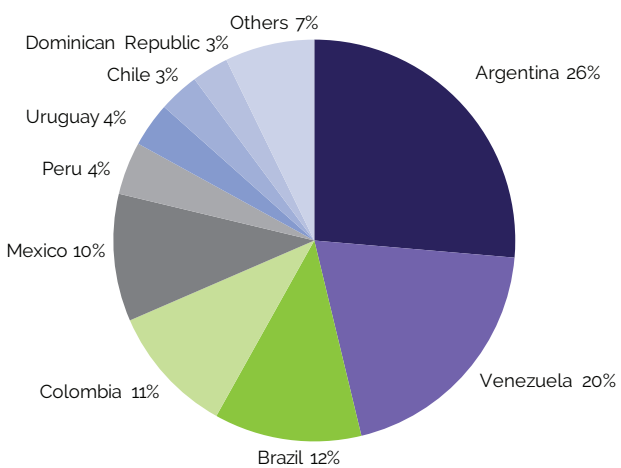
Figure 7
Mean hourly (US$ 1,000 cut-off) rate at Freelancer.com and gross national income (GNI) per capita. February 2019
(Dollars)



Source: Prepared by the authors on the basis of freelancer.com and workana.com.

Figure 8 shows an estimation of the effective size of the gig-economy in Freelancer.com, based on the money earned by freelancers with executed projects. Argentina and Venezuela have taken the greatest advantage of the opportunities to execute distance jobs through the platform, earning almost half of the region's income, US$ 4.2 million and US$ 3.2 million, respectively. While these are important values, it also shows that the gig-economy is still relatively small in LAC. Venezuela's gig-economy workers on the platform report over 250,000 hours, which suggests an effective hourly rate of some US$ 12.80. Putting this in the context of a 40-hour work week, it could be comparable to a well-paid full-time job.

Figure 8
Share of earnings by freelancers by country of origin (N = 6,751). February 2019
(In percentages)



Source: Prepared by the authors on the basis of freelancer.com.

4. Gender issues in the digital economy

a. Identifying gender from digital footprints

Neither of the two crowdsourced labor-marketplace platforms specify the gender of the professionals as a stand-alone category (doing so might actually be worrisome in terms of aspirations for non-discrimination). Therefore, we used a machine learning classifier, NamSor (Name Ethnicity and Gender Classifier)[2] in order to determine the gender of the workers based on their names. Using Namsor's free API, users' first and last names were classified by gender. Three different classifications were possible: male, female, or unknown. For each, a "confidence score" is provided, which was a continuous variable ranging from [-1.0, 1.0], with -1 being certainty for male, and +1 being certainty for female.

We undertook a sensitivity analysis of the accuracy by repeated random sampling (n=20 each time), investigating the types of names classified under each level. As per API default, names with a confidence score between [-0.1, 0.1] were classified as unknown. However, our analysis showed that with the |0.1| threshold, about 2 – 4 % of the names are classified as unknown, while with a more conservative |0.2| threshold, some 10 % are classified as unknown (Table II). We decided to go with the more conservative threshold of (-0.2, 0.2) for unknown gender.

There are several options to fine-tune general classifications. One would be manual inspection of the 10% of the unknown cases. Here it turns out that most of the unknown names would be as difficult to classify by a human coder, such as 'Estauri', 'Zaimari', 'PatriV=c'', 'Luan', 'Maxi', 'Criz', or 'Asistente', while others are due to modifications what might be solved by human inspection, such as 'Romerd', 'Rodrig. , or 'Jakelinne'. This shows that there is a potential for manual fine-tuning for a small percentage of names. We also explored to classify the provided profile pictures with a face recognition software from Google's TensorFlow supervised machine learning package. We trained it with 4,000 hand-coded images from Freelancer.com. The first bad news was that there is a relatively high number of useless profile pictures (anything but a picture of a real face). The second bad news was that the image classifier did not correlate very strongly with the name classifier (overall correlation of $R^2 = 0.40$). Given priority to the name classifier, the additional image classifier was only able to resolve an additional 2% of all cases, leaving us with some 8% of all cases still in doubt. We recommend future exercises to explore these and additional complementary approaches to automated gender classification.

Table 3
Gender prediction-from-names confidence intervals [-1 male, +1 female], difference for confidence intervals shift from |0.1| to |0.2|. February 2019

	Workana (n = 19,846)		Freelancer (n = 73,778)	
Unknown interval:	[-0.1, 0.1]	(-0.2, 0.2)	[-0.1, 0.1]	(-0.2, 0.2)
Male (%)	58.6	55.5	68.1	62.7
Female (%)	37.5	34.6	29.4	26.5
Unknown (%)	3.9	9.9	2.5	10.8
Total (%)	100	100	100	100

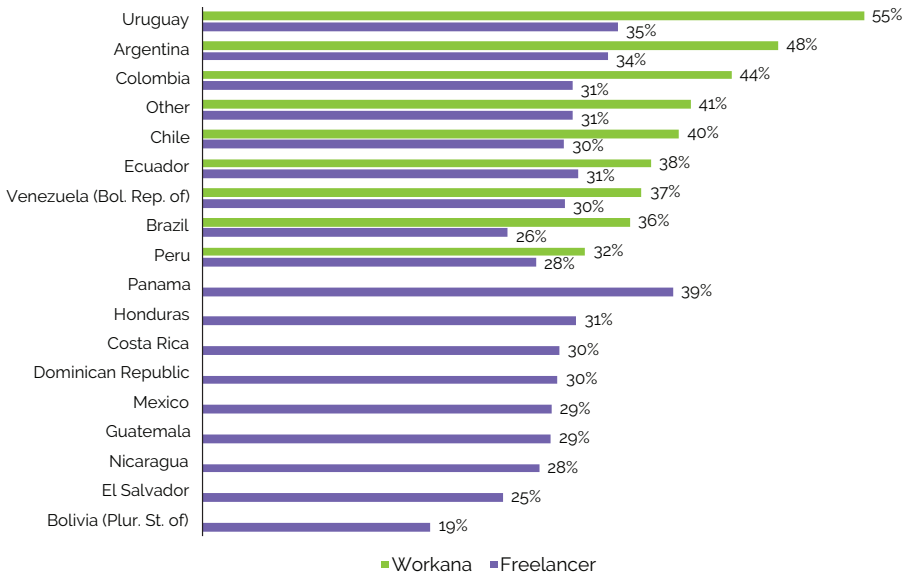
Source: Prepared by the authors.

[2] https://v2.namsor.com/NamSorAPIv2/index.html.

b. Gender inequalities

Figure 9 shows percentages of females per country for each of the platforms. We can see that there is a larger share of female professionals in the regional market of the digital economy in Workana.com than in the global market of Freelancer.com. The gender distributions by job category confirm the general expectation that female professionals keep the gender biases of labor division in the qualitative sectors of the digital economy, being most represented in 'Writing & Translation', and least present in the most quantitative sector, 'IT & Programming' (figure 10). Both sources agree on this finding. It is interesting to note that 'Data Entry & Administration' shows a relatively high percentage of female professionals in the region. However, this might also originate from biases introduced by our classification system (see table 2), which has a blurry boundary with some general administrative support tasks.

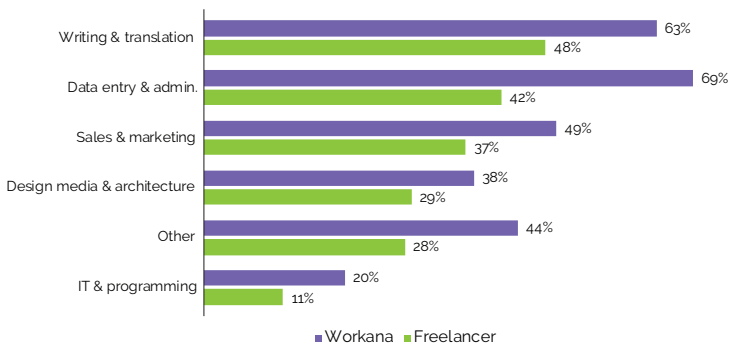
Figure 9
Proportion of female professionals by platform per country. February 2019
(Percentages)



Other: Freelancer - Cuba, French Guiana, Guadeloupe, Haiti, Martinique, Paraguay, Puerto Rico, and Saint Martin; Workana - Belize, Bolivia (Plur. St. Of), Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Puerto Rico, Dominican Rep., Trinidad & Tobago.

Source: Prepared by the authors on the basis of freelancer.com and workana.com.

Figure 10
Proportion of female professionals. Per job category and platform. February 2019
(Percentages)



Other: categories - Engineering & Manufacturing, and Business Accounting Human Resources & Legal

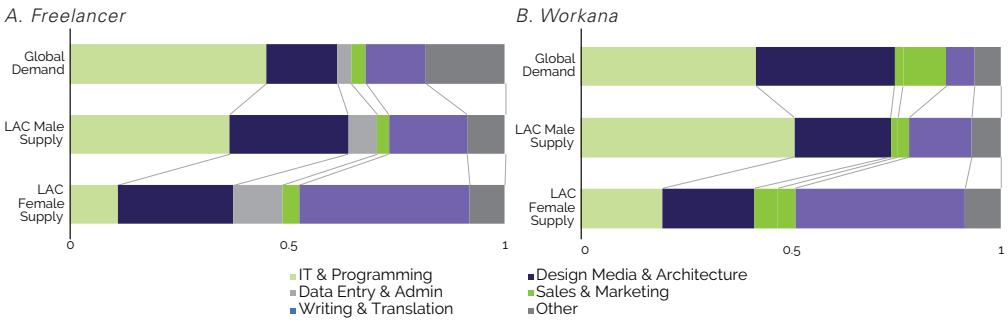
Source: Prepared by the authors on the basis of freelancer.com and workana.com.

c. Gender differences in supply and demand

We now condition the same analysis on gender. Figure 11 makes it strikingly clear that the national supply of digital economy workers is much better aligned with the job-demand for males than for females. This is mainly due to the higher share of men with skills in 'IT & Programming', the main demand of the digital economy. In the regional job market of Workana, the supply of 'IT & Programming' is even proportionally higher than the mainly regional demand. There is clearly more supply for female workers in the digital gig-economy for 'Writing & Translation' than there is demand, on both platforms. Female freelancers also provide more supply than demand for 'Data Entry & Administration' at the beginning of 2019.

Figure 11

Global job demand and regional supply, by gender

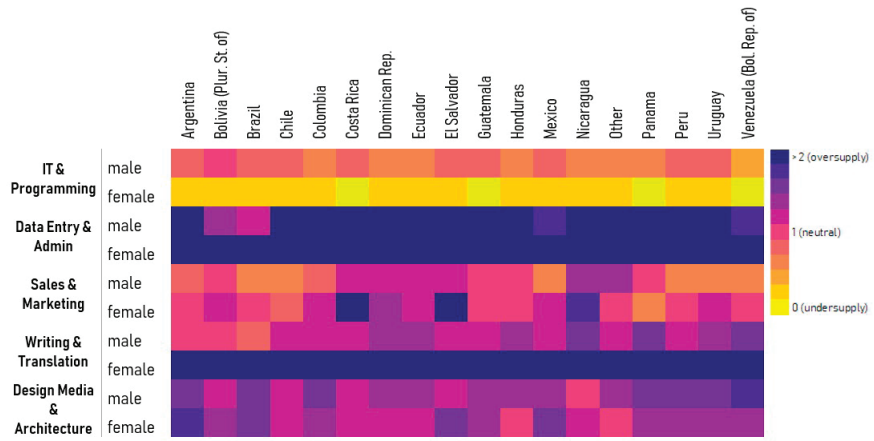


Source: Prepared by the authors on the basis of freelancer.com and workana.com.

The richness of the obtained data lends itself to produce a myriad of different analyses and visualizations. Figure 12 summarizes the supply-demand match with a heat map, which color-codes the differences.

Figure 12

Heat map of global job demand and regional supply, per gender and job category. February 2019 (Ratio)



Source: Prepared by the authors on the basis of freelancer.com.

B. Full-time employment demands

We also collected and analyzed the digital trace data from labor marketplaces where employers that mainly offer full-time jobs post open positions. This is expected to differ from the freelance-based gig-economy as it is more aligned with traditional employment opportunities and jobs are usually located within the country and satisfied by the national workforce.

1. Full-time employment demand

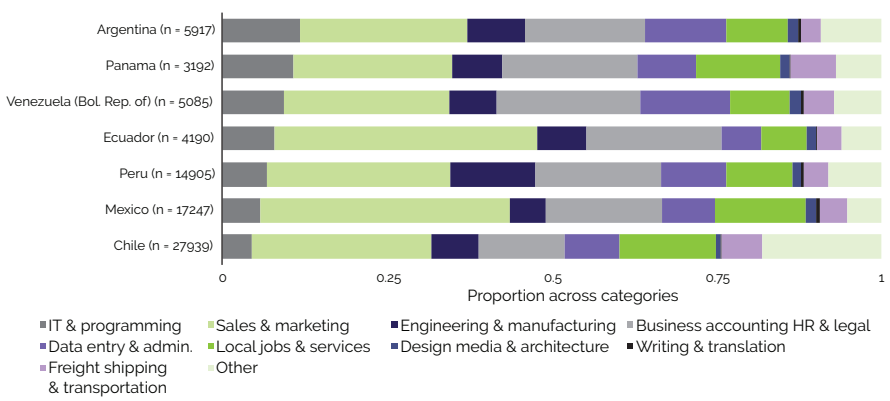
Our main pan-regional source of this kind is Bumeran, which is the leading online recruitment portal for full-time employers in Latin America, operating online in 7 countries, with some 12 million visits per month, more than 60,000 ads per month, more than 13 million candidates, some 260,000 new candidates per month, and some 50,000 client companies per year (it is known as laborum.com in Chile). Our collection confirms that some 90% of Bumeran's jobs are full-time employments, with only about 1% being contract or temporary (which is the market covered by the previously analyzed freelancer portals).

Figure 13 shows employment categories of digital economy positions by country. It is striking to compare these distributions with the equivalent distributions from the gig-economy (e.g. figure 5). Full-time employers in the region look much less for 'IT & Programming' and for 'Writing & Translation', and much more for 'Sales & marketing', and other business administration positions. We also observe differences among countries. For example, in Chile and Peru, there is more demand for 'Engineering & Manufacturing' (which was part of 'others' in the freelancer analysis, given its small sample in that market), than in countries like Ecuador, Panama and Mexico.

Again, it is important to emphasize that the categories of the different sources do not necessarily need to be perfectly aligned (see table I3). But even beyond these choices, it might simply be the case that a skill like 'design' is understood differently by users in differ ent countries and cultural contexts. This is of course similar to traditional survey-based statistics.

Figure 13

Employment per job category of digital economy employment at Bumeran. February 2019 *(Percentages)*



Source: Prepared by the authors on the basis of bumeran.com.

2. Text analysis of job requirements

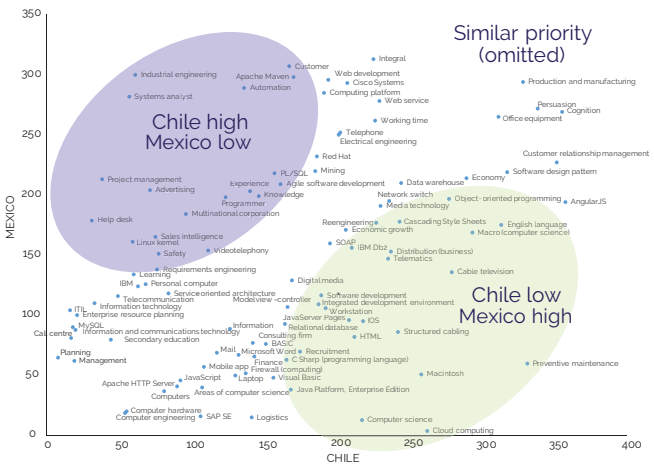
One of the opportunities of digital trace data is the exploitation of non-traditional data sources, such as qualitative text or images. Both are abundantly available online, and can be converted into measurable numbers with the help of a variety of computational tools. We already explored the use of images in our gender classification task, and now turn to a text analysis of job descriptions in Bumeran.

The most straightforward method would be to create so-called bag of words, which essentially counts the appearance of all words, highlighting the most frequent ones. Going beyond simplistic bag of words approaches, we fed the obtained text into an automated semantic machine learning analysis that extracts meaning contained in the job posts. We used a platform called TextRazor, which uses state-of-the-art Natural Language Processing (NLP) techniques to parse, analyze, and extract semantic metadata from the content. In particular, we use TextRazor's topic tagger (which was originally trained with millions of Wikipedia pages) to assign high-level topics to content, using entity and word category relationships. For example, if a job post asks for skills in "Amazon web services, virtualization", the semantic analysis returns "cloud platform" as the best-fit topic tag. If a job post solicits skills in "service desk, support, maintenance, hardware", the Wikipedia trained artificial intelligence returns as the best-fit topic tag "Help desk". TextRazor has an automatic understanding of hundreds of thousands of different topics at different levels of abstraction and works in ten languages, including Spanish and Portuguese. These allows us to convert vast amounts of qualitative unstructured data into quantitatively interpretable insights.

Figure 14 shows the semantically extracted job demands in the Technology area in Chile versus Mexico. In total, the artificial intelligence detected 358 different job categories in Chilean postings and 325 in Mexico, of which 172 were common in both countries. We compared the detected priority for each topic and selected the 100 common categories that were least similar in both countries. The result reconfirms the previous finding and shows that technology job demands in Chile call more frequently for managerial and business skills (like those related to 'Customer', 'Project management', 'Advertising', and 'Sales intelligence'), while job posts in Mexico call more pervasively for technical skills (like 'cloud computing', 'Java Platform', 'HTML', and 'cable television'). The analysis also provides evidence that English language is more commonly required in Mexico, than in Chile. In follow-up studies, it would be interesting to link these insights to complementary traditional statistics.

Figure 14

Semantic analysis of Bumeran job posts in Chile and Mexico (laborum.cl, N = 1,907 postings; bumeran.com.mx, N = 1,640 postings). Axes are ranks (high numbers are low ranks)



Source: Prepared by the authors on the basis of bumeran.com.

AFFORDABILITY OF DIGITAL TECHNOLOGIES

II

A necessary condition for the digital economy is the required technological equipment. It is well-known that in the region the main obstacle to digital access is the cost of technology (CEPAL, 2013; Hilbert, 2010). We consult the pan-regional platform MercadoLibre Inc., an Argentine company that operates online marketplaces dedicated to e-commerce and online auctions in 18 Latin American countries. In 2018, the company employed 7,239 professionals (2,409 on its own information technology and product development staff) and reported a net-income of US$ 1,440 million (MercadoLibre, 2018).

MercadoLibre offers both used and new products. We focus exclusively on new products, since this will give us a common base in terms of product quality. MercadoLibre offers different categories of products, which turn out to be somewhat distinct among their country branches, and ends up leading to a considerable diversity of product categorization among countries. For our purposes of international harmonization, the most important consideration is how different categories are effectively used in different countries by users, which is a data quality problem whose main solution consists in labor intensive human inspection.

A. Technology price index

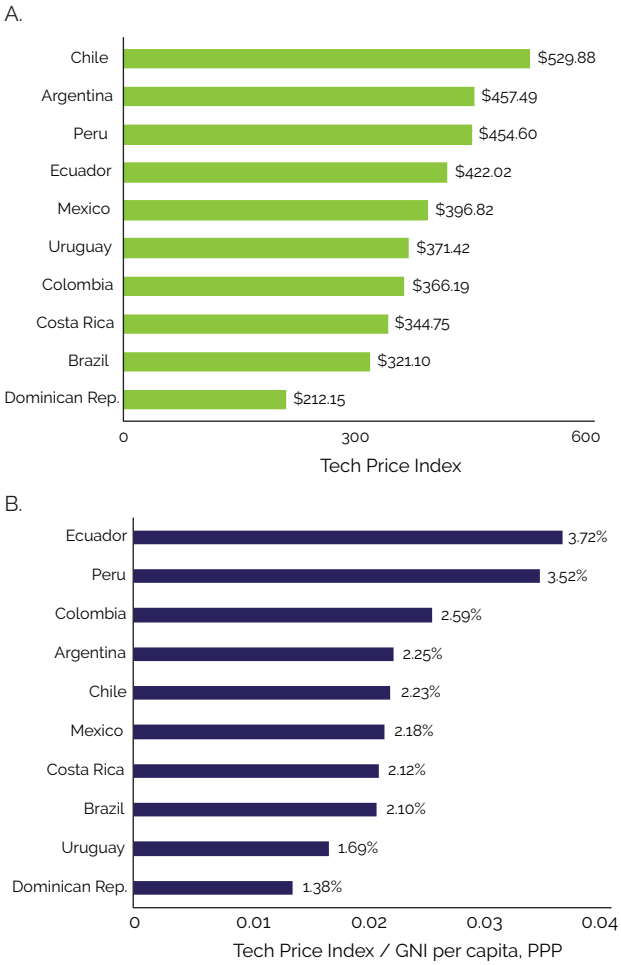
In total we tracked 192,092 products from ten countries and seven sectors, namely mobile phones, laptop & tablets, computers & servers, processors, networks and wireless, software and drones. After much detailed analysis, we created an aggregate index, for which we calculated the simple average of the average prices per category. This provides relatively more weight to smaller categories, such as drones (figure 16). As always, the constitution of an index depends on the goal of the researcher and often pre-determines the result (Minges, 2005). Figure 15.a shows that in nominal values, the average price of technology is clearly higher in some countries than in others. This is related to the characteristics of national markets, such as their degree of development, the degree of competition of the markets, the tax structure, in addition to the level of purchasing power, among others.

Figure 15.B contrasts the tech prize index against national income per capita (with purchasing power parity (PPP), which considers the opportunity costs for other products within the country. We can see three groups. In Ecuador and Peru, tech prices are relatively high in relation to national income. Colombia is coming close to this group. Argentina, Chile, Mexico, Costa Rica and Brazil represent a quite similar mid-field when it comes to tech affordability. In Uruguay and the Dominican Republic, tech prizes are clearly lower than in the rest of the region, considering how much the average person earns. This provides a useful price index for technology affordability, based on a broad selection of options, updatable in real-time.

Figure 15

Technology price index by country. a) nominal values in US$; b) in relative terms as percentage of Gross National Income (GNI, ppp) per capita. February 2019

(Dollars)



Source: Prepared by the authors on the basis of mercadolibre.com.

Figure 16 contrasts the absolute and relative prices per technology sector (i.e. figure 15.A versus 15.B). Again, it shows that for most sectors (including drones; computers and servers; laptops and tablets; mobile and smartphones; processors; software; network products), technology is most expensive in Chile (in nominal values of US$, which results in being more moved to the right on the horizontal x-axis). However, in relative terms considering its purchasing power (up and down the vertical y-axis), the only sector led by Chile is software (which might have to do with intellectual property rights), while Chile is among the more affordable cases for mobile phones, drones, laptops and tablets, and networks. In relative terms, Ecuador and Peru again stand out at two countries where several kinds of technologies are rather expensive in the context of national purchasing power. The medium price of a mobile phone in Ecuador and Peru is the equivalent of almost three percent of the annual national gross national income per capita (PPP).[3] It is interesting to note that in the context of national purchasing power, inhabitants of the Dominican Republic can access to the lowest tech prices of the region for most technologies, including laptops and tablets, drones, computers and servers, mobile and smartphones, and processors.

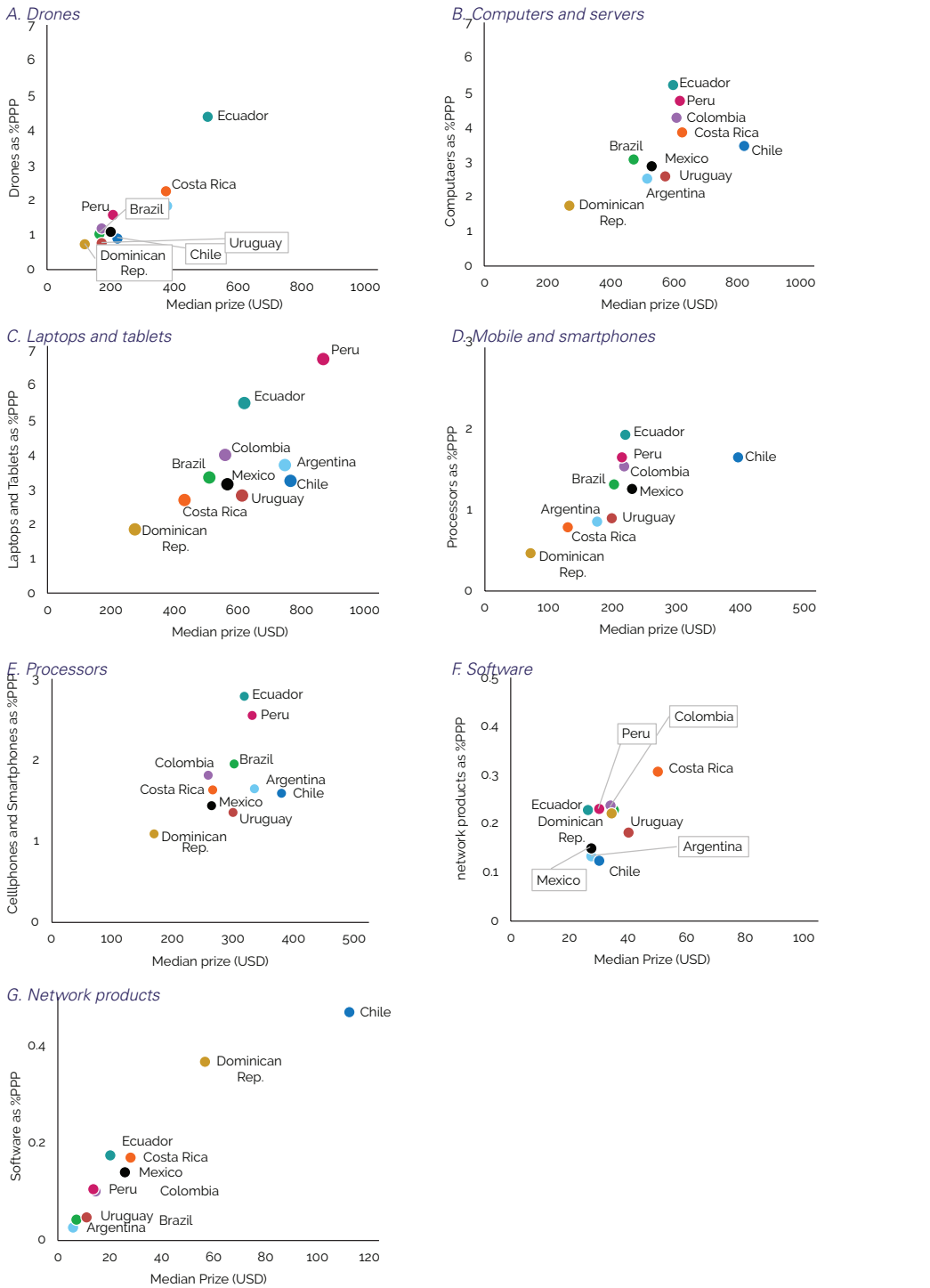
[3] Even 3.5 when not adjusting GNI per capita with purchasing power parity (PPP).

Summing up the different percentages suggests that, on average, some 12% of the regional per capita income would be required if everyone in the region would have access to a medium-sized solution of each one of the seven kinds of technologies tracked here. This shows the potential to build price baskets, and leads us to the next approach.

Figure 16

Median technology prices per sector (horizontal x-axis) versus median prices normalized by GNI per capita PPP. February 2019

(Dollars)



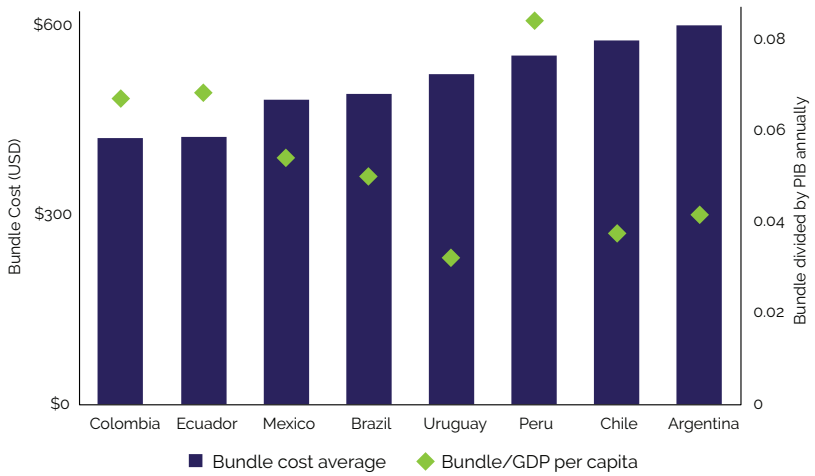
Source: Prepared by the authors on the basis of mercadolibre.com.

B. Bundle Price

We do a follow-up analysis and dig deeper into specific technologies creating a basket comprising a set of access device: laptop (Acer Aspire), tablet (Apple iPad) and mobile phone (Samsung Galaxy J2). It is important to emphasize that this kind of analysis depends on the chosen technology and there might be many confounding factors. Any adequate choice is also quickly changing as technology evolves. However, following the efficient market hypothesis for online markets, even a handful of devices should approximate the average national price.

Figure 17

Bundle price by selected country. February 2019.



Source: Prepared by the authors on the basis of mercadolibre.com.

Taking as reference a basic basket of technological products, it is observed that in the selected countries the cost of this set of devices ranges from 400 to 600 US dollars (see figure 17). By relating these values to the level of income per capita, an indication of the degree of affordability of these technologies by the population is obtained. In this sense, this basket of products is more affordable for the inhabitants of Uruguay and Chile, for whom it represents 3% and 4% of GDP per capita, respectively. They are followed by Argentina and Mexico with 5%, Colombia and Brazil with 6%, Ecuador with 7%, and finally Peru, where the cost of the basket reaches 8% of per capita income.

MICRO, SMALL AND MEDIUM-SIZED ENTERPRISES: ACCESS TO MARKETS AND CROWDFUNDING

III

Small and medium-sized enterprises (SMEs) in Latin America represent a heterogeneous group of economic agents that contribute significantly to the generation of employment in the region, as well as to its gross domestic product (Dini & Stumpo, 2011; Peres & Stumpo, 2002; Vallejos, 2003). In most countries, SMEs contribute between 20 % and 50 % of employment, and between 15 % and 40 % of sales (Ferraro & Stumpo, 2010). Despite their importance in the region, relatively little is known about this important economic actor, especially in comparison to larger enterprises. This lack of information can be reflected in obsolete, late or inadequate policies. This concern only increased in the digital economy with the arrival of new markets for small and medium sized actors. At the same time, this new tendency also opens an opportunity, since actors in the digital economy leave digital traces behind, which can be used to study them.

We focus on two sources. MercadoLibre is the largest online retail platform in the region (MercadoLibre, 2018). We track 2,473,977 retailors from 18 countries. We also look at crowdfunding platforms, especially kiva.org, a non-profit organization that allows people to lend money via the Internet to low-income entrepreneurs in over 80 countries. We analyze 1,684,119 loan projects on Kiva, as well as 2,400 projects from kickstarter.com.mx in Mexico.

A. Online retailers

MercadoLibre.com provides a set of developer APIs, which gives access to merchandise and seller data. However, to get the seller information, a MercadoLibre-defined ID of the seller is required. Note that 'simply' collecting all sellers might not be useful, since it might contain many sellers who never sold anything, or who have been dormant for a long time. Therefore, we need a representative list of seller IDs currently active in the region. This forces us to take some decisive methodological decisions, basically related to the sampling of the sellers.

We decide to sample our seller database by collecting the currently most relevant merchandise items and identify their sellers. Specifically, our procedure of data collection consists of three steps. (1) Obtain the merchandise dataset. We decide to sample a set from each available product category. Within MercadoLibre's 18 LAC countries, there are 239,771 product categories (see table 4). The default sorting algorithm of merchandise per product category in the API is labelled "More Relevant" (with two other options available: "Lowest Price" and "Highest Price"). We collect the 1,000 most "relevant" products per category per country, providing us with 53,425,985 products, of which 43,108,698 were new items, and the rest used re-sell products. Note that many categories contained less than 1,000 or no products in some countries, which is why we retrieved less than the 239,771 * 1,000 theoretically possible products (about 22% of the possible). (2) From the collected merchandise dataset, we retrieved the seller ID, and subsequently removed duplicates within the seller ID set, resulting in 2,473,977 unique retailors from 18 countries. Finally, (3), we used the retrieved seller IDs to collect seller information from the API.

Table 4 shows the sellers of the most relevant items sold in early March 2019, sampled from all available product categories. Other sampling methodologies will lead to a different dataset and might lead to different results. Among the 2,473,977 sellers, we further concentrate our analysis on vendors that are labelled as "normal" and "active" on the platform and exclude international sellers, large brands, car dealers, franchises and real estate agencies.

Table 4
Data collection involved in sampling of sellers on MercadoLibre

	Number of product categories	Collected number of merchandise products	Collected number of sellers
Argentina	44,765	8,147,807	562,869
Bolivia	424	15,123	842
Brazil	65,506	19,953,271	683,344
Chile	19,559	4,265,428	149,614
Colombia	21,360	4,928,705	152,054
Costa Rica	1,506	189,754	11,695
Rep. Dom.	691	146,099	7,539
Ecuador	1,713	431,893	36,189
Guatemala	512	18,951	1,322
Honduras	184	5,694	292
Mexico	39,074	7,895,614	365,733
Nicaragua	176	2,190	220
Panama	458	34,965	2,131
Paraguay	325	11,728	622
Peru	8,439	1,500,016	109,279
El Salvador	264	7,974	628
Uruguay	23,360	3,295,392	122,573
Venezuela	11,455	2,575,381	267,031
SUM	239,771	53,425,985	2,473,977

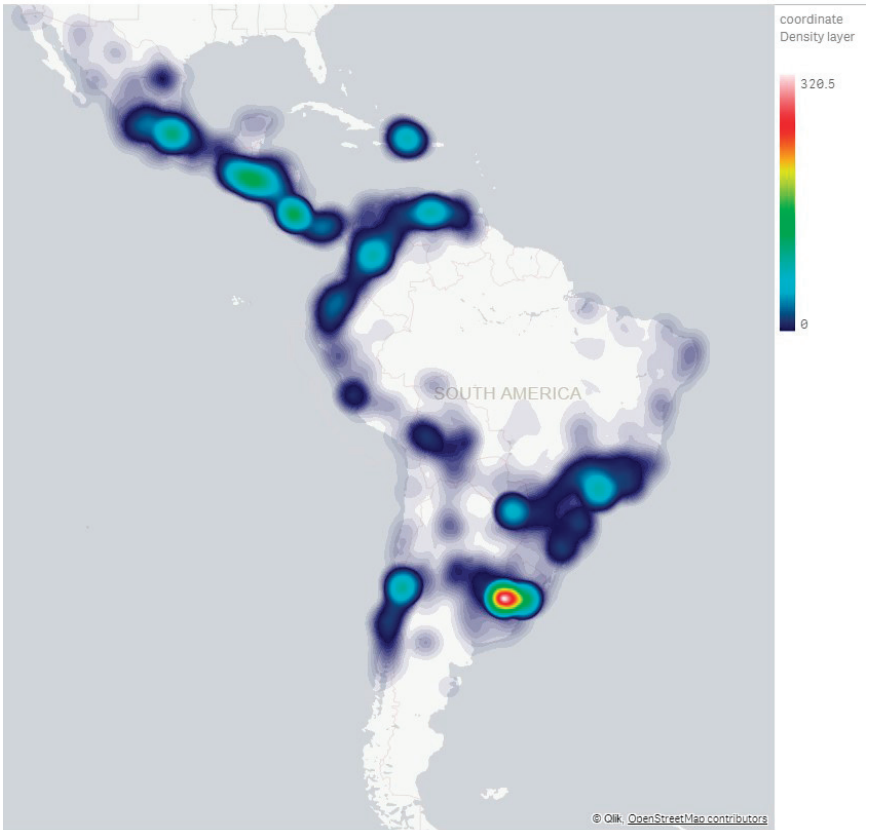
Source: Prepared by the authors on the basis of mercadolibre.com.

From a practical perspective, computing with 53.4 million products migh API is in JavaScript Object Notation (JSON) format. We used a document-oriented NoSQL database named MongoDB to store JSON format data and retrieved the seller IDs by a Python code with the assistant of the MongoDB system. The duplication of seller IDs was computed using the standard Python Pandas library.

1. Geolocations

Geographic distribution of sellers can be obtained by matching the city and country name of the seller, provided by MercadoLibre, with the Google Geolocation API to obtain geo-coordinates. The following maps are density maps, produced in Qlik.com. Map 1 shows the distribution of sellers in Mercado Libre in the countries of the region. The heat map shows that the use of this type of e-commerce platform is widespread, and its use is not limited to the capital cities but also reaches the interior of the countries. This represents greater access to markets, as well as more consumption options. Also, it shows a high concentration of sellers in the region around Sao Paulo and Rio de Janeiro, as expected, which are the region's largest and fifths largest cities, respectively. Somewhat unexpectedly, map 1 also shows a high concentration in Central America, with Costa Rica being its epicenter with 11,665 vendors, mainly in San Jose.

Map 1
Geographical distribution of vendors on MercadoLibre



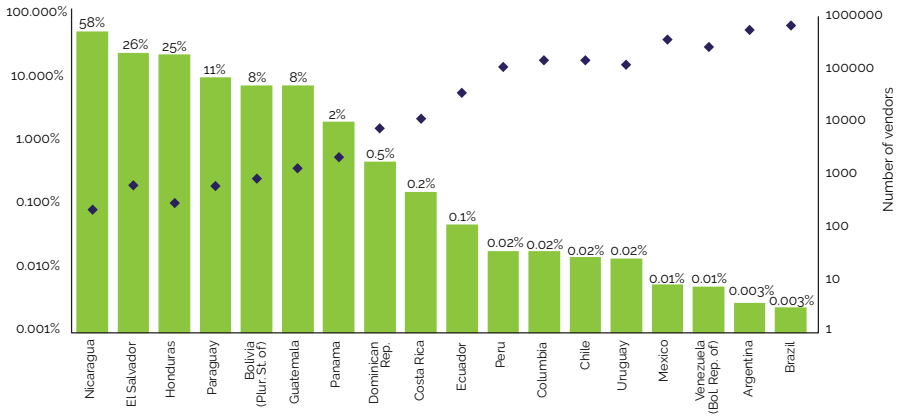
Source: Prepared by the authors on the basis of mercadolibre.com.
Note: The boundaries and names shown on this map do not imply official endorsement or acceptance by the United Nations.

2. Market concentration

Figure 18 shows the concentration among vendors per country. The bars show the market share of the top five-sellers in terms of the numbers of completed transactions on MercadoLibre (all time). The right axis on figure 18 shows the logarithm of the number of all sellers. It is a more complete measure of market concentration in terms of the distribution of transactions (a simplified form of the often-used Herfindahl-Hirschman Index).

Looking at the relation among both indicators, it becomes clear that the more sellers there are in a country (the larger the country), the less concentrated the number of transactions per seller, the less dominant are the top-five vendors. For example, note that in Nicaragua, with 220 vendors, the five sellers with most all-time transactions represent 58 % of the total market, while in Brazil, with 673,528 vendors, the top-five merely capture 0.003 % of all executed transactions. This shows that the provision of a ccess to these new business opportunities is very important, as it seems to translate quite directly into economic opportunities. If there would be more sellers in small countries, the share of the top-five is likely to diminish. With limited access to many micro-, small- and medium-sized retailers, online markets are easily dominated by a few relatively large players.

Figure 18
Market concentration of vendors on MercadoLibre. Logarithmic plot. Market Share of top 5 Vendors in terms of all time sold products (bars); number of vendors per country (right axis). All time
(In percentages (bars) and counts (right axis)

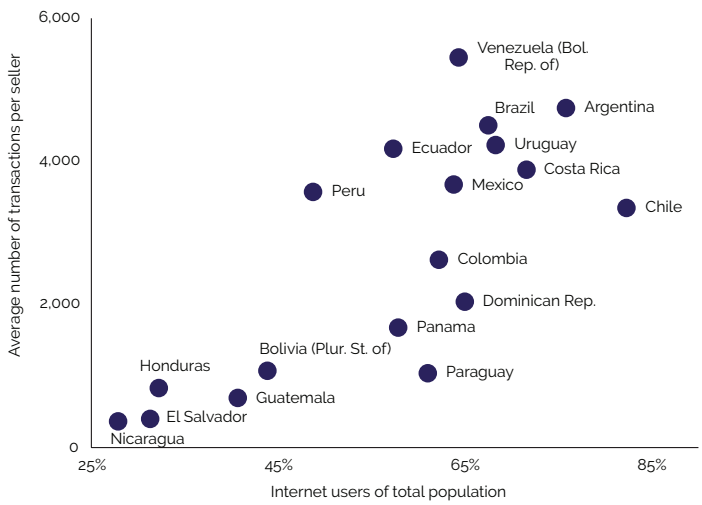


Source: Prepared by the authors on the basis of mercadolibre.com.

3. From access to transactions

This same point is underlined by the results from figure 19, from a complementary perspective. It relates internet users as percentage of total population versus the average number of transactions executed per seller (all time). With higher internet penetration, there are clearly more online transactions, not even in total per country, but also per seller. This suggests that internet penetration promotes both a more vibrant online retail market and also more economic activity per economic agent, probably due to a larger market of potential consumers. This emphasizes the fact that the digital divide in terms of internet access is still a main roadblock for a vibrant digital economy in Latin America.

Figure 19
Internet penetration (ITU, 2017) versus average number of transactions per seller per country. All time

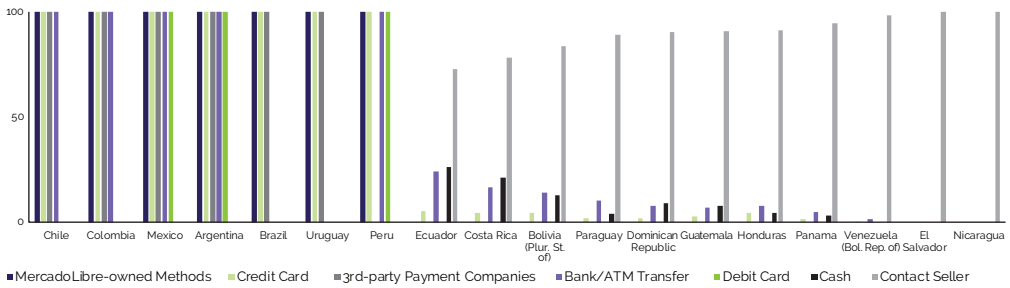


Source: Prepared by the authors on the basis of mercadolibre.com.

4. Enablers of transactions

Figure 20 looks at an important enabler of digital transactions: payment options. The MercadoLibre platforms in some countries, among them Chile, Colombia, Mexico, Argentina, Brazil, Uruguay, and Peru, offer a considerable variety of payment options, while in most smaller countries, buyers have to pay in cash or are asked to contact the seller. This is surely hindering flourishing e-commerce. Comparing this graph with the market share of the top five sellers (figure 18), it shows that limited payment options go hand in hand with a more concentrated market that only benefits a few sellers.

Figure 20
MercadoLibre available payment options by country. January 2019
(In percentages)



Source: Prepared by the authors on the basis of mercadolibre.com.

B. Crowd-funding

Kiva is a popular crowd-funding platform with a special focus on projects in developing countries. Its mission is "to connect people through lending to alleviate poverty" (Waghorn, 2013). It has crowd-funded over 1.7 million loans, with a repayment rate of over 98 percent (Price, 2017), raising about $1 million every three days as early as 2013 (Waghorn, 2013). Kiva is active in 101 countries, 22 of them in Latin America and the Caribbean, with LAC representing 386,498 out of the worldwide 1,551,367 successfully financed projects (about a quarter).[4] We work with 1.5 million successfully funded projects at the time of collection (March 2019), out of the 1.7 listed projects.[5]

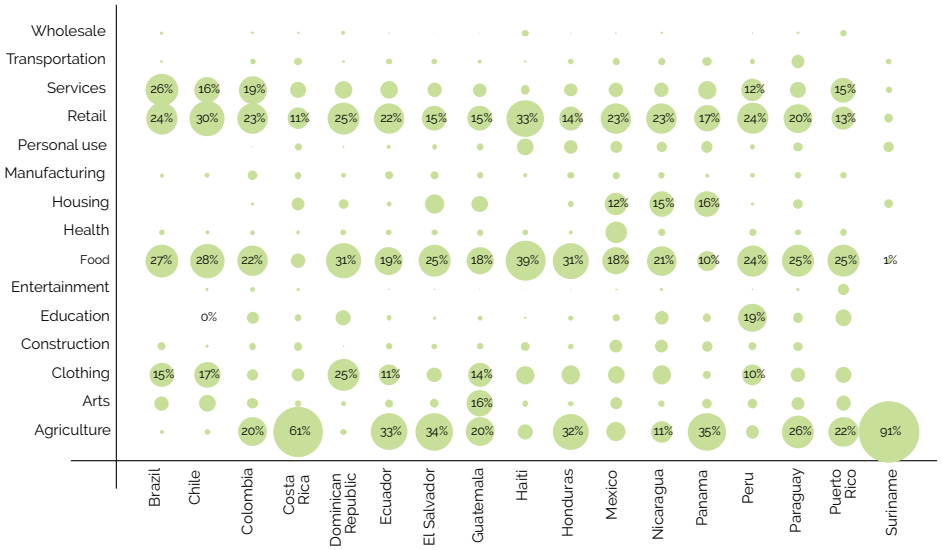
1. Fundraising per sector

Figure 21 shows that most projects in the region are for agriculture, food, and retail. These priorities are similar in Africa and Asia, but different in North America (with less agriculture and more transportation) and Europe (with less entertainment and more housing). Some notable exceptions are a proportionally larger share of clothing projects in Dominican Republic, health projects in Mexico, and education projects in Paraguay.

4 Projects country (all time). Peru: 89,386; El Salvador: 60,423; Nicaragua: 42,100; Ecuador: 38,291; Colombia : 36,732; Paraguay: 28,010; Bolivia: 24,131; Mexico: 18,859; Honduras: 17,066; Guatemala: 13,691; Haiti: 6,160; Costa Rica: 4,856; Dominican Republic: 4,280; Brazil: 875; Chile: 873; Suriname: 265; Panama: 228; Belize: 180; Puerto Rico: 92; East Asia & Pacific: 495,569; Sub-Saharan Africa: 422,556; Europe & Central Asia: 103,904; South Asia: 77,473; Middle East & North Africa: 58,379; North America: 6,988.

5 This excludes projects that are "expired", "fundraising", and "refunded" (money returned).

Figure 21
Proportions of projects per sector in Kiva.org. (2006-2018)
(In percentages)



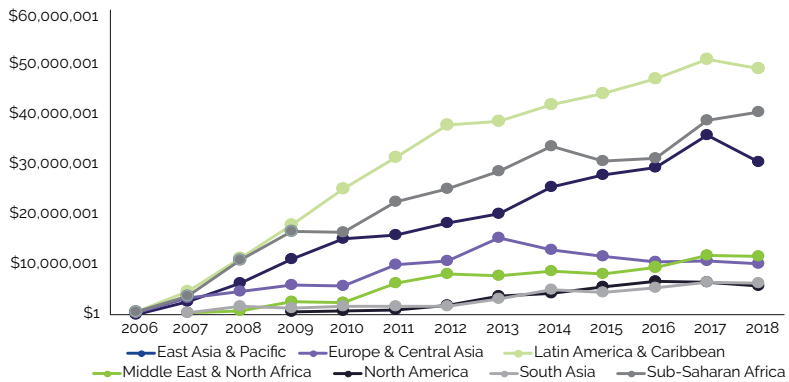
Source: Prepared by the authors on the basis of kiva.org.

2. Fundraising in global comparison

Figure 22 shows that entrepreneurs from Latin America and the Caribbean take very effective advantage of the global networks of micro-enterprise crowd-funding. The region is leading in the sectors of food, retail, agriculture, clothing, services, arts, construction, health and manufacturing. In 9 of the 15 sectors, LAC is leading. No other world region is as successful in the average amount crowdsourced per project across different sectors. Figure 22.A shows that this trend has been sustained over the past decade. This could suggest that in the countries of the region there is an important space for the development of Fintech, in particular credit or loan platforms. Figure 22.C shows that this averages over quite some diversity among countries. It is interesting to note that the inter-sectorial diversity in Honduras and Dominican Republic stems from above-average education projects (in Colombia the smallest), while Chile has a highly crowded-sourced health- and Brazil agriculture sectors, while personal use are often the smallest projects.

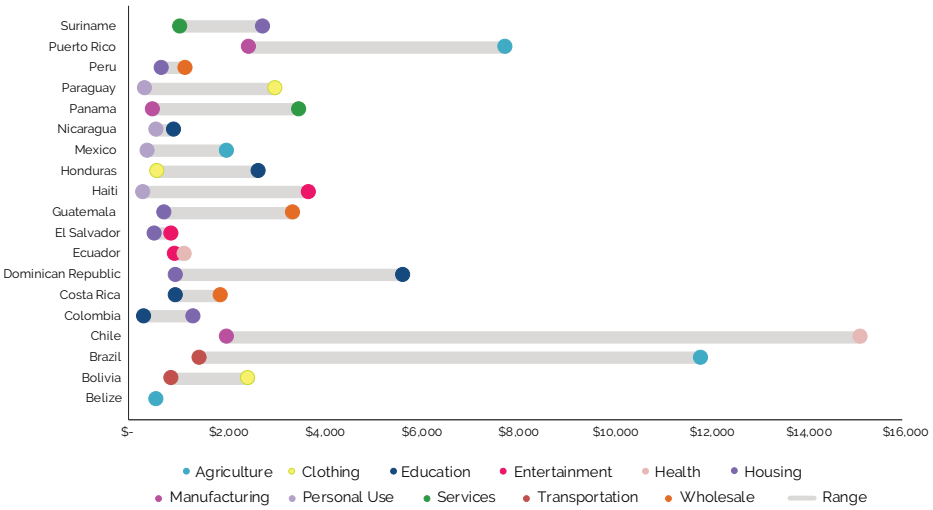
Figure 22
Total Amount of Funding received, 2006 – 2019
(Dollars)

A. By World Region and Year



B. By World Region and Sector

C. Lowest and highest average funded amount by sector and country.



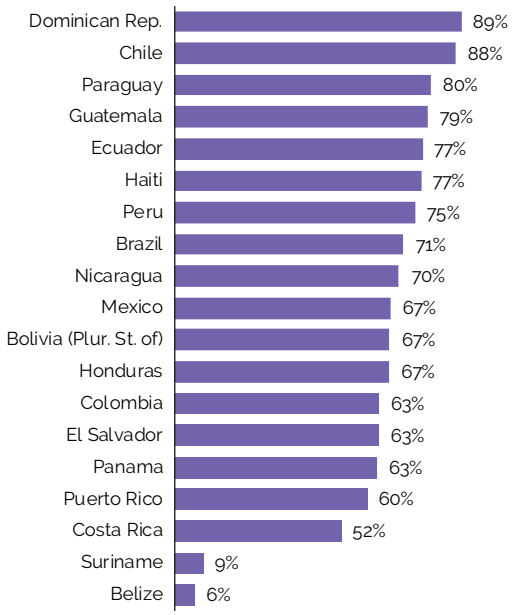
Source: Prepared by the authors on the basis of kiva.org.

3. Gender distribution

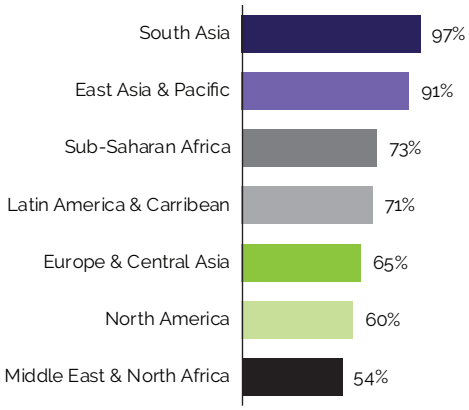
The Kiva database also reports gender of entrepreneurs. Often, projects are run by teams. If a team consists exclusively of women or men, we code that team's gender composition with the dummy variable 0 or 1, respectively. If there is one man and four women, we count this as 0.2, and sum of the resulting values for all teams to get the national average. Figure 23 shows that the majority of crowdsourcing micro-entrepreneurs in the region are women. In Dominican Republic, Chile, and Paraguay, female entrepreneurs represent over 80% (which is similar in Asia). With the exception of the small samples of Belize (n = 180 projects) and Suriname (n = 265 projects), all countries have more active female micro-entrepreneurs, and most also have more active female entrepreneurs on Kiva than in other world regions, including North America and Europe. This shows that an increasing use of these kinds of digital opportunities, bears the chance to fight existing and well-known gender inequalities.

Figure 23
Percentage of female led projects on Kiva
(In percentages)

A. By LAC country. 2006 – 2018



B. By world region



Source: Prepared by the authors on the basis of kiva.org.

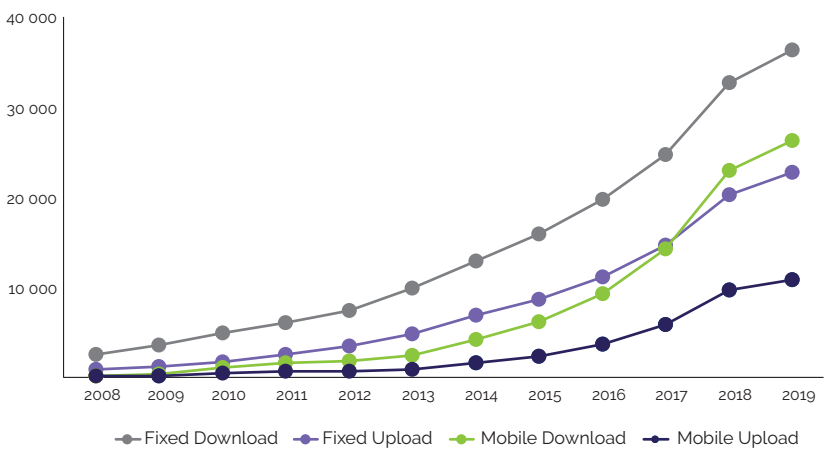
BROADBAND SPEED IN THE REGION AND THE WORLD

IV

The following estimations of broadband speeds follows the methodology developed by (Hilbert, 2014, 2016; Hilbert, López, & Vasquez, 2010), and mainly draws its data from crowd-sourced speedtests collected by Ookla (Ookla, 2019). In other words, they represent the speed obtained during standardized speed tests, executed by users across the world. Historical estimates pre-2018 are taken from (Hilbert, 2019). Much could be said about the following graphs, but this would be a standalone report and beyond the scope of this general exercise.

Figure 24 shows the general global evolution of data upload and download speeds according to access mode, fixed or mobile, over the last decade. It shows that the largest increase in average bandwidth stems from the rise in mobile download speeds. Figure 25 takes a closer look at fixed and mobile download speeds per region, which shows that LAC is part of the slower tier, only providing faster average bandwidth than South Asia and Sub-Saharan Africa. The Middle East and Northern Africa stand out globally in h regard to recent increases in mobile broadband download speeds. Taking a closer look, figure 25 reveals that in LAC, fixed and mobile speeds are comparatively similar, especially compared to more advanced regions, where there is a clear leadership of fixed download speeds. At this point, the reasons are mere speculation, but it might be related to the state of fiber optics development.

Figure 24
Global bandwidth speeds by type, 2008-2019
(in kbps)

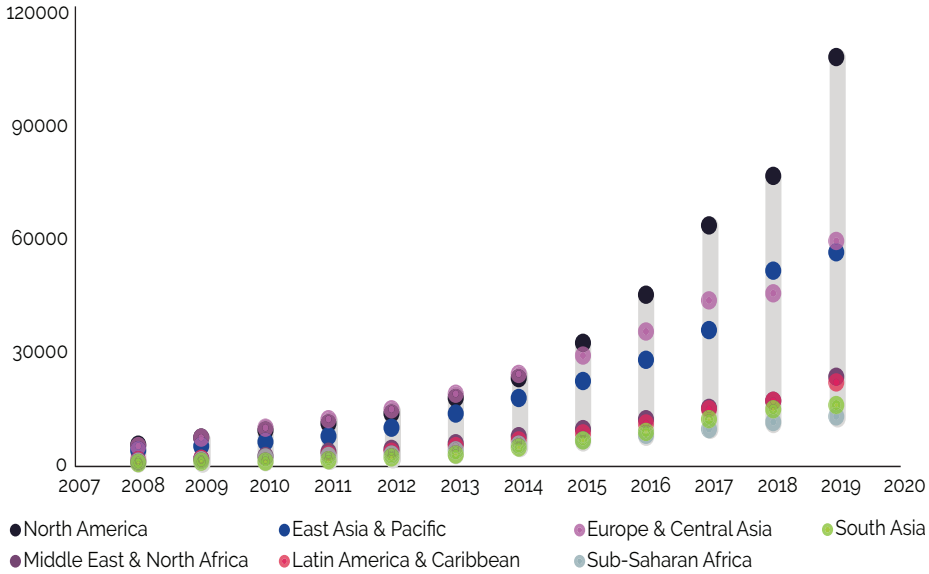


Source: Prepared by the authors on the basis of kiva.org.

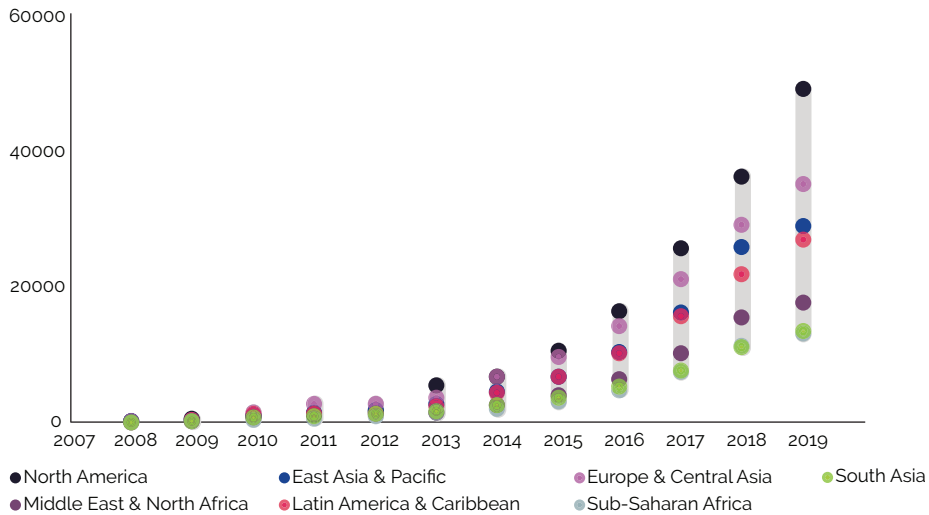
Looking at individual countries (not shown here), Trinidad and Tobago and Chile figure among the world's countries with best broadband capacities. Within the region, we see a leading group of countries with relatively well developed fixed-line download speeds, i.e. Barbados, Trinidad and Tobago, Chile, Uruguay, and Panama. A second tier of countries emerges with better developed mobile download speeds than fixed-line, including Mexico, Belize, Ecuador, Honduras, Peru, Dominican Republic, Colombia, Nicaragua, Guatemala and Bolivia. This strong mid-field is dominated by the fact that mobile download speeds are faster than fixed download speeds. Lastly, we observe a third tier of countries, with comparatively narrower bandwidth, including Paraguay, El Salvador, Costa Rica and Venezuela.

Figure 25
Fixed versus mobile download speed per world region, 2007-2019
(in kbps)

A. Average fixed download rate



B. Average mobile download rate



Source: Prepared by the authors on the basis of Ookla.

There have been important advances in the development of fixed and mobile broadband globally, primarily through the increase in the number of users but also through the rapid technological evolution. This is predominantly apparent by the increase in connection speeds, for example, between 2010 and 2019 fixed broadband increased its download speed by 6.5 times while mobile broadband accelerated 25 times.

However, this technological evolution has not been homogeneous across the globe leading to significant disparities among regions. In February 2019, average fixed broadband download speeds in Latin America were 2.5 times lower than in Europe and Central Asia and almost 5 times slower than in North America. For mobile broadband, average download speeds in Latin America were half those of Europe and Central Asia and 3 times lower than in North America.

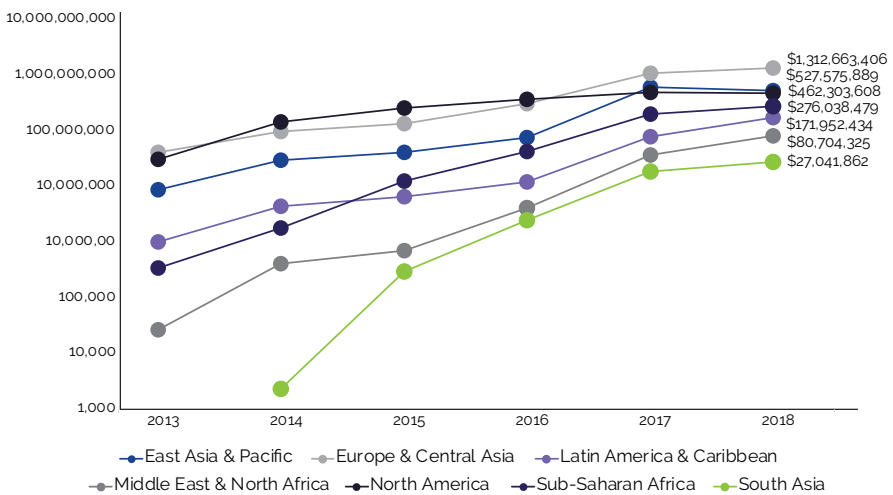
V CRYPTOCURRENCY

Cryptocurrencies are digital assets designed to work as a medium of exchange that uses strong cryptography to secure financial transactions, control the creation of additional units, and verify the transfer of assets through a distributed ledger on a blockchain. A blockchain is essentially a list of records, called blocks, where each block contains a cryptographic hash of the previous block, a timestamp, and transaction data. The redundant copies and decentralized nature of the ledger on a blockchain makes it resistant to modification of the data. As such, cryptocurrencies use decentralized control as opposed to centralized digital currency and central banking systems. Bitcoin is the most popular cryptocurrency, representing 53.2 % of the market cap at the time of collection, March 2019 (Coin.Dance, 2019).

The webpage https://coin.dance/ is a community-driven bitcoin statistics service. It provides several statistics around cryptocurrencies, including the national volume of bitcoins per country (in local currency). The collected data is reported as weekly transaction volume. The source tracks 47 countries, including seven from Latin America, namely Arg entina, Brazil, Chile, Colombia, Dominican Republic, Mexico, and Peru.

Figure 26 shows the increase of global bitcoin volumes on a log-scale. It shows that pre-2017, North America was leading, while East Asia and Pacific has overtaken it since then, together with Europe, which is globally leading in terms of bitcoin volumes. LAC possessed over 10 % of global bitcoins back in 2014 coming fourth among the seven global regions, but has since been overtaken by Sub-Saharan Africa which has seen more bitcoin purchases than LAC.

Figure 26
Annual bitcoin volume purchases by world region. 2013-2018
(In dollars and year)



Source: Prepared by the authors on the basis of coindance.com.

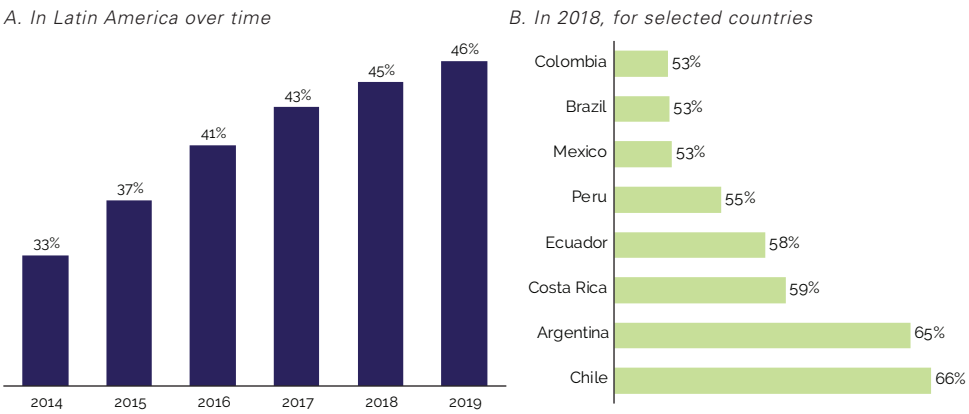
The socio-economic situation of the countries of Latin America and the Caribbean largely explains the interest in cryptocurrencies such as Bitcoin and the acceleration of the volume of transactions in recent years. These assets are perceived as solutions to address the problems of economic instability, political crises, informality of the economy and emigration, which make this region the most unequal in the world. With such obstacles to progress, which result in a crisis of confidence, and a young population with access to technologies, cryptocurrencies emerge as investment instruments and intermediaries for currency exchange.

We consulted the APIs of Facebook-ads to derive relevant socio-demographic and technology characteristics of the region, and Twitter in order to track trending topics related to the Sustainable Development Goals (SDGs).

A. Socio-demographics

At the time of our inventory, Facebook is the fifth most valuable company on Earth and had 2.4 billion monthly users (1 out of 3.5 people worldwide). In Latin America as a whole, Facebook reaches about half of the population in 2019 (see figure 27.A), while in some countries, it is used by two-thirds (figure 27.B). Facebook's business model consists in providing tailor-made advertisement. The Facebook Ads Manager provides potential clients with estimate of "potential reach" of a commercial message within a user segment with certain socio-demographic, economic and cultural characteristics. We used these "potential reach" estimations to reverse engineer Facebook's approximations of an array of social stratifications.

Figure 27
Facebook penetration



Source: Prepared by the authors on the basis of facebook.com.

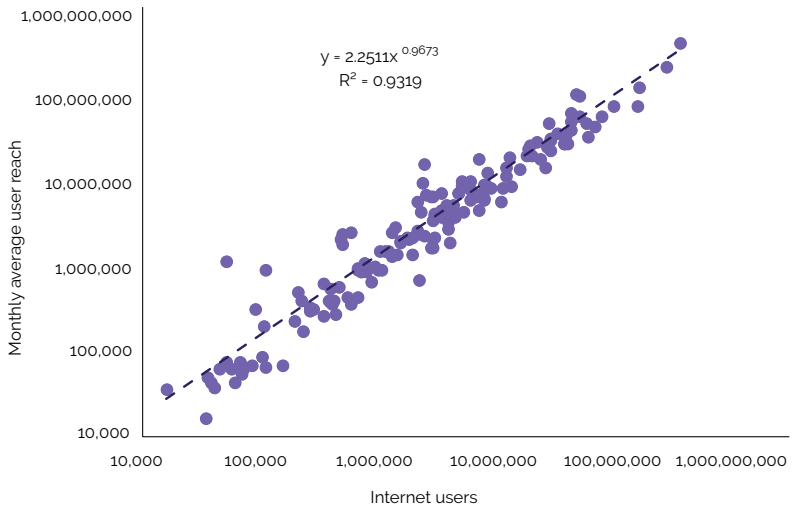
We systematically accessed Facebook-ads manager API in order to obtain estimates about people's interests, technologies and other socio-economic characteristics. We tested for the relationship between Facebook's "potential reach" and the actual population. According to Facebook, "potential reach is an estimation of how many people are in an ad set's target audience. This estimation is a unique calculation by Facebook".[6]

We tested for the relationship between Facebook's "potential reach", actual population and number of internet users (figure 28). Figure shows that the correlation between potential reach and internet users is quite strong ($R^2 = 0.93$). The correlation between potential reach and inhabitants is also strong, but naturally less ($R^2 = 0.78$).

6 https://www.facebook.com/business/help/1665333080167380.

The correlation between the number of Facebook users and potential reach is even stronger (R^2 = 0.9975). All of this suggests that potential reach estimates can be taken as being quite reliable, at least in relative terms. In absolute terms, it best represents the Facebook penetration within a country.

Figure 28
Facebook's potential reach estimates for 164 countries versus internet users (ITU, 2018)



Source: Prepared by the authors on the basis of facebook.com.

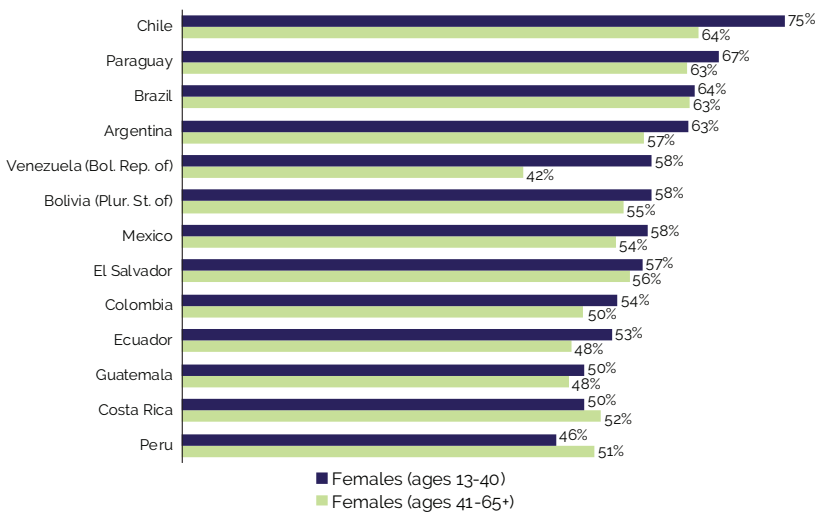
In total, we obtained 70 variables for all 33 countries of LAC,[7] and always segment the potential reach by gender and by six age groups (13-20, 21-30, 31-40, 41-50, 51-60, 61+). We collect daily average reach and monthly average, and see more variation and less reliable results in the daily number, so we work with the potential reach of monthly average users (MAU). If the monthly average user reach is below 1,000, the reported default value by Facebook for MAU is 1,000 (often even if the daily average user is 0, 1 or 2 people per age group and gender, per country). Therefore, we collected the data twice, within one week, to get more reliable results, and report either the average of both collections for MAU, or the number that is different from the 1,000 default. If countries have more than 50 % of the reported fields filled with the unreliable 1,000 default, we exclude these countries from the analysis, due to a small sample size (this is the often case for some Caribbean countries).

[7] Facebook_access_(network_type):_2G (behaviors), Facebook_access_(network_type):_3G (behaviors), Facebook_access_(network_type):_4G (behaviors), Facebook_access_(network_type):_WiFi_ (behaviors), Facebook_access:_older_devices_and_OS (behaviors), business owners (interests), Administrative_Services (industries), Architecture_and_Engineering (industries), Arts_Entertainment_Sports_and_Media (industries), Business_and_Finance (industries), Community_and_Social_Services (industries), Computation_and_Mathematics (industries), Construction_and_Extraction (industries), Education_and_Libraries (industries), Farming,_Fishing_and_Forestry (industries), Food_and_Restaurants (industries), Government_Employees_(Global) (industries), Healthcare_and_Medical_Services (industries), IT_and_Technical_Services (industries), Installation_and_Repair_Services (industries), Legal_Services (industries), Life_Physical_and_Social_Sciences (industries), Management (industries), Military_(Global) (industries), Production (industries), Protective_Services (industries), Sales (industries), Transportation_and_Moving (industries), Civil_Union (relationship_statuses), Complicated (relationship_statuses), Divorced (relationship_statuses), Domestic_Partnership (relationship_statuses), Engaged (relationship_statuses), In_a_relationship (relationship_statuses), Married (relationship_statuses), Open_Relationship (relationship_statuses), Separated (relationship_statuses), Single (relationship_statuses), Unspecified (relationship_statuses), Widowed (relationship_statuses), Technology_early_adopters (behaviors), Food_&_Restaurant_page_admins (behaviors), Bodybuilding (interests), Meditation (interests), Physical_exercise (interests), Physical_fitness (interests), Running (interests), Weight_training (interests), Yoga (interests), Computer_memory (interests), Computer_monitors (interests), Computer_processors (interests), Computer_servers (interests), Desktop_computers (interests), Free_software (interests), Hard_drives (interests), Network_storage (interests), Software (interests), Tablet_computers (interests), Audio_equipment (interests), Camcorders (interests), Cameras (interests), E-book_readers (interests), GPS_devices (interests), Game_consoles (interests), Mobile_phones (interests), Portable_media_players (interests), Projectors (interests), Smartphones (interests), Televisions (interests).

1. Business ownership

Facebook could be a source of statistics about actual business ownership. Figure 29 shows the disaggregation by gender and age. In general, we find that women compare favorably in terms of gender balance in business ownership (figure 29), at least when it comes to the digital economy. In all countries of the region, at least one of the two age groups obtain more than 50%. Older entrepreneurs, more than 41 years of age, outpace their younger counterparts in Costa Rica and Peru (and seemingly in many Caribbean countries, but we decided that the sample size in these countries is not reliable enough). It is important to note that this favorable gender balance might be influenced by the fact that these entrepreneurs are active on Facebook. In the traditional offline world, there might be more male entrepreneurs. If this would turn out to be true, then even more so do digital tools provide potential to become big equalizers in terms of gender issues (Hilbert, 2011).

Figure 29
Percentage of female business ownership by age group. March 2019
(In percentages)

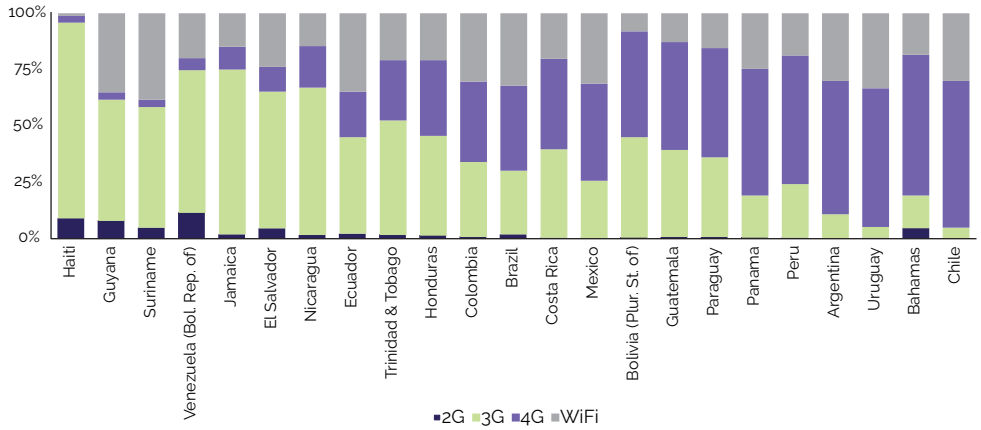


Source: Prepared by the authors on the basis of facebook.com.

2. Network access

One reliable digital trace data from Facebook-ads manager is directly linked to Facebook usage, and is related to the question through which kind of network users access Facebook (figure 30). With a penetration of around half of the population in LAC, this digital footprint provides an unprecedented picture of the actual usage of access technologies. This data provides more room for in-depth analysis with regard to the actual use of regional telecommunication infrastructure.

Figure 30
Access to Facebook according to the access network per country. March 2019
(In percentages)



Source: Prepared by the authors on the basis of facebook.com.

As shown in figure 30, 3G and 4G mobile technologies are the most used means of access. It is noteworthy that in several countries there is still 2G access, although in most cases this corresponds to less than 5%. It is possible to differentiate two groups of countries. The first corresponds to those countries in which access is mainly through 3G networks, most countries in the region fall into this group. The second group, comprising Argentina, Bahamas, Chile, Panama, Peru and Uruguay is marked by a predominance of 4G connections. For most of the countries, connection via WiFi represents about one third of the total access. Taking these data as proxies of connectivity indicators, the development of public access WiFi points could be a solution to increase internet access at the population level.

B. Trending topics

We used the Twitter API to obtain insights about the social media visibility of the 17 Sustainable Development Goals (SDGs) in the region.

1. Sustainable Development Goals (SDGs)

We used the real-time filter through the steaming option of the Twitter API.[8] We collected as many tweets as Twitter would give us related to a set of keywords. These keywords included all 17 SDGs and related words in English, Spanish, and Portuguese (mainly the

[8] https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html.

words of the main SDG title).[9] We collected them together with location information. For this, we specified country names in the Twitter API, as well as for the names of the largest cities.[10] Of the more than 35 million collected tweets, we were able to assign 4,612,966 tweets to one of 29 countries from Latin America and the Caribbean.

Image 1
Sustainable Development Goals (SDGs)

Source: United Nations, 2016.

2. Number of tweets

There are clear differences between the tweeted concerns going on in the digital public spheres of different countries, but also clear similarities. Topics related to SDG16 (peace, justice and strong institutions) is clearly the most commonly voiced concern in the region. This agrees with the rest of the world, while some countries in the region clearly have more concerns about this issue than others. The second and third largest concerns that can be distilled when listening to the global twitter-sphere are good health and well-being (SDG3) and climate action (SDG13). The latter is particularly prominent in some Caribbean countries, such as Barbados, Cayman Islands, Dominica, Trinidad &

9 Collected keywords and keyphrases: 'sdg', 'ods', 'sdg 1', 'ods 1', 'poverty', 'pobreza', 'sdg 2', 'ods 2', 'hunger', 'hambre', 'fome', 'sdg 3', 'ods 3', 'health', 'salud', 'well-being', 'saúde', 'saude', 'bienestar', 'sdg 4', 'ods 4', 'education', 'educación', 'educacion', 'educação', 'educacao', 'sdg 5', 'ods 5', 'derechos de la mujer', 'abuso sexual', 'feminicidio', 'femicidio', 'sdg 6', 'ods 6', 'clean water', 'agua limpia', 'sanitation', 'saneamiento', 'água potável', 'agua potavel', 'saneamento', 'sdg 7', 'ods 7', 'clean energy', 'energía asequible', 'affordable energy', 'energia asequible', 'energía no contaminante', 'energia no contaminante', 'energias renováveis', 'energias renovaveis', 'energias acessíveis', 'energias acessiveis', 'sdg 8', 'ods 8', 'decent work', 'trabajo decente', 'economic growth', 'crecimiento económico', 'crecimiento economico', 'trabalho digno', 'crescimento económico', 'crescimento economico', 'sdg 9', 'ods 9', 'industry', 'industria', 'innovation', 'innovación', 'infrastructure', 'innovacion', 'infraestructura', 'indústria', 'inovação', 'inovacao', 'infraestruturas', 'sdg 10', 'ods 10', 'inequality', 'desigualdad', 'inequalities', 'desigualdades', 'sdg 11', 'ods 11', 'sustainable', 'sostenible', 'sustentáveis', 'sustentaveis', 'sdg 12', 'ods 12', 'responsible consumption', 'produccion responsable', 'responsible production', 'producción responsable', 'consumo responsable', 'produção sustentáveis', 'consumo sustentáveis', 'producao sustentaveis', 'consumo sustentaveis', 'sdg 13', 'ods 13', 'climate', 'clima', 'climática', 'sdg 14', 'ods 14', 'ocean life', 'submarin', 'marine life', 'aquatic', 'aquátic', 'marinha', 'sdg 15', 'ods 15', 'ecosystem', 'ecosistema', 'habitat', 'terrestre', 'forest', 'biodiversidad', 'biodiversity', 'bosque', 'environment', 'medio ambiente', 'ecossistema', 'biodiversidade', 'floresta', 'meio ambiente', 'sdg 16', 'ods 16', 'peace', 'paz', 'justice', 'justicia', 'strong institutions', 'instituciones sólidas', 'instituciones solidas', 'justiça', 'justica', 'instituições eficazes', 'instituicoes eficazes', 'sdg 17', 'ods 17'.

10 For small countries (where there were none or only one city with a population over 300 thousand), we worked with a list of large cities. For medium-sized countries, we queried for all cities over 300 thousand inhabitants in a country. For large countries (total population over 30 million), we queried for the names of the capitals of all states or provinces, along with all major cities over 1 million inhabitants. Then, we fine-tuned the result through manually inspecting the results for each city/state, checking for confusion between similar places (adding conditions like "AND" or "AND NOT LIKE" between city and country names, etc.

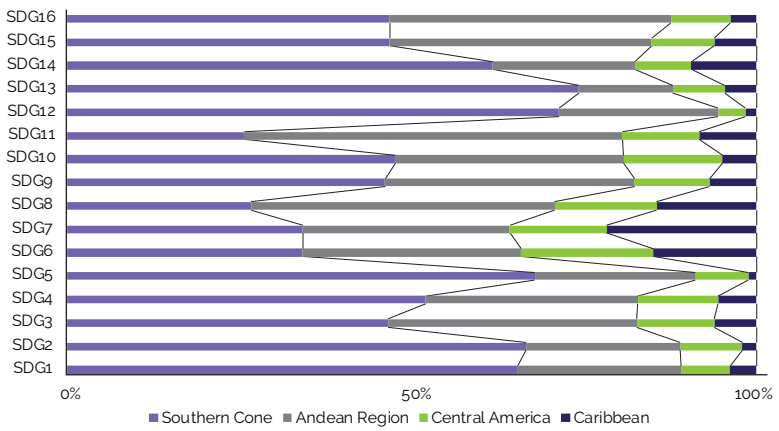
Tobago and Virgin Islands, but also in Paraguay. These three priorities are followed by quality education (SDG4) and industry, innovation and infrastructure (SDG9) worldwide. Compared with the proportions of tweets from the rest of the world, gender equality (SDG5) plays a more prominent role in LAC specifically in Chile and Costa Rica during the period of observation. Industry, innovation and infrastructure (SDG9) is a less relevant concern in social media discussions in the region.

Figure 31.A shows that stakeholders from the Southern Cone are very vocal about topics related to SDG12 (responsible consumption and production) and SDG13 (climate action). It is interesting to detect a very strong concern about SDG11 (sustainable cities and communities) from the Andean Region, and a proportionally large and vocal concern of Caribbean social media users with regard to SDG6 (clean water and sanitation), SDG7 (affordable and clean energy) and SDG8 (decent work and economic growth).

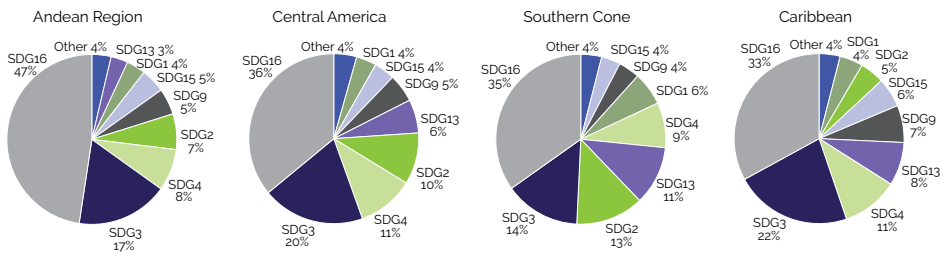
Looking at the SDG-priorities aggregation on the sub-regional level (figure 31.B), it shows that the concerns are quite similar, and pretty much in line with global priorities. Gender equality (SDG5) is larger across LAC sub-regions than globally, but, notably, concerns about climate action (SDG13) are less across all sub-regions, compared to the global level of concern.

Figure 31
Shares of tweets
(In percentages)

A. Per SDG



B. Per sub-region. February - March 2019



Andean Region (Bolivia (Plur. St. of), Colombia, Ecuador, Peru); Caribbean (Antigua and Barbuda, Bahamas, Barbados, Cayman Islands, Cuba, Dominica, Dominican Rep., Haiti, Jamaica, Trinidad and Tobago, Virgin Islands); Central America (Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, Panama, Puerto Rico); Southern Cone (Argentina, Chile Paraguay, Uruguay)

Source: Prepared by the authors on the basis of twitter.com.

VII LESSONS LEARNED

In this report we have discussed and showcased examples of how publicly available online data sources can help inform policy-making in the region. We discovered potentially large benefits from working with digital trace data, but we also have seen that they do not come easily, and that further discussion and sustained effort are required to reap such benefits. After all, we feel that–especially when compared to traditional data sources–the benefits outweigh by far the required efforts and resource demands.

A. Summary of benefits and challenges

1. New insights

We have shown that digital trace data provides unprecedented insights in terms of:

- **thematic coverage**: including areas that were previously difficult or impossible to measure, like service delivery or gender imbalances in small enterprises, or supply and demand for the labor market;
- **geographical coverage**: our pan-regional sources provided sizable and comparable data for 14-18 countries, and national portals provided insights into national sub-regions;
- **level of detail:** different shades of public opinion, relationship status by age, and gender imbalances by job category;
- **alternative sources:** text (and also images) have traditionally been treated as qualitative information, and useless for statistical purposes, while modern machine learning techniques allow us to convert those sources into measurable statistics; and
- **timeliness**: we produced graphs the same week we collected the data, and are able to publish them quickly through interactive online interfaces.

Additionally to the qualitative advantages of working with digital trace data, the required resources are comparably smaller, especially when considering the number of different policy-making discussions that can be provoked with the plethora of obtainable graphs. In this exercise, we mainly focused on the digital economy and gender issues, but have already seen that much broader development goals could easily benefit from publicly available footprint data, e.g. related to the Sustainable Development Goals.

2. Old challenges

Naively, when thinking about the big data paradigm, many tend to attribute the challenges of working with digital trace data to simply computational skills, and possibly required technological infrastructure. Once these are solved, the expectation seems to be that the digital footprint acts like a real-time crystal ball. Our experience shows that the real challenges lie elsewhere. We did not collect or analyze data with any sophisticated infrastructure (we mainly used mid-performance laptops), and computational skills were mainly focused on data collection through Python (be it through web-scraping

or API access) and data cleaning through R. Python has become the standard general-purpose programming language and comes with a plethora of free-online courses for interested parties. Instead, we found that the most severe challenges are more in line with the usual suspects from traditional statistics. Priorities change a bit, but the basic discussions remain the same.

- **Representativeness.** While traditional development statistics is mainly concerned with the representativeness of random survey samples, digital trace data is never a random sample. It is observational data, and the ambition is usually to obtain the entire population completely. We have tracked some 80 thousand job offers on Bumeran and all 200 thousand technology products offered on MercadoLibre. Sample representativeness and associated significance tests lose their leading role in a big data paradigm. Technical skills shift from mastering how to collect a representative sample and how to perform a significance test, to collecting digital traces and how to effectively retrieve, clean and store them.
- **Generalizability.** While observational data including the entirety of a source always represents this source very well, it only represents what it represents, and nothing more. Traditional statistics is used to generalizing from a sample to a larger population. Here we track the entirety of the population, and while it is tempting to generalize to broader settings, this is often very deceptive. We have seen that the global labor market from Freelancer.com is different from the regional market from Workana.com, which is again different from the national labor markets of Bumeran.
- **Harmonization.** International organizations are used to challenges related to international harmonization of indicators, and digital trace data is no exception here. We took advantage of several pan-regional and international platforms, but even here, there are national differences. Incompatible indicators were often an impediment for broader and more detailed coverage.
- **Variable definitions.** The definition of the chosen indicator and the weighting of any aggregate index can pre-determine much of the output. Working with digital trace data, challenges are similar to traditional ones (how to weigh an index?) and add new dimensions (what should machine learning detect when converting qualitative into quantitative data?).

B. Pay special attention to...

In the following, we offer several more detailed reflections, which are mainly related to the aforementioned long-standing statistical challenges, but we also highlight some of the new measurement opportunities. We hope that this more in-depth discussion can serve as guidance for practitioners interested in making digital trace data work for international development.

In this first exercise, we shed light on a diverse range of issues that had not previously been measured. Substantively, each one of the graphs deserves a much deeper analysis. Many of the graphs provide significantly more questions than answers, and will require a more in-depth analysis from respective domain experts (be it for gender issues, labor markets, technology prices, or small and medium sized enterprises, with domain knowledge from different countries, etc.). Each one of these areas could fill a separate publication. We also found several aspects that warrant follow-up analysis to clarify certain points. This was to be expected, since the more we see, the more we discover about what we do not yet know, and where we need to look even closer. In

this sense, while the goal of this report was not to provide an in-depth discussion of these different areas, we were able to successfully showcase that digital trace data is able to further crucial discussions about new topics in the region, which can inform policy-making.

That being said, we can already see from the graphs presented here that the focus of substantive analysis changes in its nature. Traditionally, the main challenge for domain experts of development in the region was data scarcity. A new graph was something valuable and rare to come by, and difficult and costly to produce. Here we have shown more graphs than we were able to do justice in analyzing with the care they would deserve. For example, we had different sources for the labor market in different countries, and each country could have filled several pages with an in-depth domain specific analysis, contrasting the different numbers with national insights. In a big data age, data is more abundantly available. At the same time, one needs to be careful with what each data source represents.

1. Sources matter

Digital trace data is observational data. This is very different than randomly sampled survey data, especially when it comes to generalizations to other or larger populations. Digital trace data is always biased toward its specific source. For example, hourly rates differ in Freelancer.com and Workana.com indicating that they measure different things (global demand, more regional demand). We also found clear and meaningful differences between the labor markets represented by the gig-economy and full-time employment marketplaces. Which one of them is the 'right' representation of the 'labor market'? All of them, for the different aspects they represent. Such differences are not noisy annoyances. When working with observational data, sources allow us to observe what happens in this source, and in nothing else.

Additionally, every source can only provide what it provides. For example, for the gig-economy, we can nicely compare supply and demand, since both are offered on the same platform. However, for the full-time employments offered at Bumeran, we can only track the demand, and do not have access to the national supply.

It is also very important to note that the source-dependency largely influences the interpretation and conclusions for policymaking. For example, the fact that countries like Bolivia, El Salvador and Nicaragua show a higher percentage of 'IT & Programming' and 'Data Entry & Administration' specializations than Brazil, this does not necessarily mean that this is representative of the ground-level truth of these professionals. It might just be that a higher share of these professionals offers their services on the given online platform. In Brazil, they might work full-time employment jobs, or use a different national site where they offer their services. In this sense, it is telling that for inhabitants of countries with specific national conditions (economic situation, productive structure) gig-economy platforms offers options that transcend the limiting conditions nationwide. Of course, this is an interesting finding in its own right. The most important fact to consider is that, at the end, the obtained sources only represent what they represent, and generalizations must be taken with care.

2. Indicators and indexes matter

Definition of indicators matters and depending on weighting of categories in indices, higher/lower in different countries. The eventual choice of the most useful indicator is of course in the eye of the beholder. The best we can do is to eliminate the most

unreasonable ones (but even this runs the risk of overlooking some special need) and to publish all the data that might be useful, in dynamic dashboards that allow users to look for answers to yet unasked questions. While doing this, it is important to be very transparent to explain the source and methodology behind the presented numbers, but the interpretation should and can only be done by the researcher with specific domain knowledge.

3. Harmonization matters

In theory, one of the benefits of digital trace data is the unprecedented level of detail it can attain, which allows for fine-tuned analysis. However, in practice, we learned that details often have to be sacrificed for the sake of international harmonization. Even when a plethora of tech-categories is offered by small countries, at the end, only the largest countries have sufficient overlap in definitions in order to make them comparable among each other. In job categories, we were forced to work on quite an aggregated level of categorization, in order to harmonize among different sources.

While the solution to variable harmonization often consists in aggregation, other times normalization can be useful. This can even help to harmonize among different sources. For gender issues in the gig-economy, we found differences in absolute numbers among two different sources. However, in relative numbers, the ordinal ranking was quite consistent among both sources. But this trick would not work in other cases, as in the case of hourly pay of freelancers. Here we have seen that it seems we detected two different markets, which is also an interesting finding.

4. Domain knowledge matters

Getting the data requires some technical skills, but is doable. Visualizing it in meaningful ways requires some fundamental domain knowledge, which exists. However, in-between these two ends are where most of the challenging tasks happen. This includes the identification of mistakes and inconsistencies that occurred during data collection, the time investment to really understand the data in order to be able to make decisions on how to clean it and on what makes meaningful sense to look at given the data. There were many occasions where we would have liked to ask more questions that the digital trace data alone was not able to answer. This suggests that digital trace data will not strictly replace existing statistics and on the ground research efforts. It will rather complement it. It will also certainly not replace, but rather require the need for more in-depth knowledge. Domain experts are required to make sense of all the gathered data. Data science will become a natural part of this important task, and most of the data business will not be confronted with scarcity, but with abundance. Therefore, we expect that the scarcity of domain knowledge, with in-depth experience of regional and local conditions, will become the main bottleneck of this kind of analytical work in the future.

BIBLIOGRAPHY

Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., … Picariello, A. (2015). Challenge: Processing web texts for classifying job offers. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 460–463. https://doi.org/10.1109/ICOSC.2015.7050852.

CEPAL, N. (2018). *The Inefficiency of Inequality*. Retrieved from https://repositorio.cepal.org/handle/11362/43443.

Coin.Dance. (2019). Coin Dance Statistics. Retrieved from https://coin.dance/stats.

Dini, M., & Stumpo, G. (2011). *Políticas para la innovación en las pequeñas y medianas empresas en América Latina*. Retrieved from https://repositorio.cepal.org/handle/11362/3868.

Ferraro, C. A., & Stumpo, G. (2010). *Políticas de apoyo a las PYME en América Latina entre avances innovadores y desafíos institucionales*. Retrieved from https://repositorio.cepal.org/handle/11362/2552.

Hilbert, M. (2011). Digital gender divide or technologically empowered women in developing countries? A typical case of lies, damned lies, and statistics. *Women's Studies International Forum*, *34*(6), 479–489. https://doi.org/16/j.wsif.2011.07.001.

___(2014). Technological information inequality as an incessantly moving target: The redistribution of information and communication capacities between 1986 and 2010. *Journal of the Association for Information Science and Technology*, *65*(4), 821–835. https://doi.org/10.1002/asi.23020.

___(2016). The bad news is that the digital access divide is here to stay: Domestically installed bandwidths among 172 countries for 1986–2014. *Telecommunications Policy*, *40*(6), 567–581. https://doi.org/10.1016/j.telpol.2016.01.006.

___(2019). Digital Data Divide Database (SSRN Scholarly Paper No. ID 3345756). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=334575.

___Hilbert, M., López, P., & Vasquez, C. (2010). Information Societies or "ICT equipment societies"? Measuring the digital information processing capacity of a society in bits and bytes. *The Information Society*, *26*(3). Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01972241003712199.

Kässi, O., & Lehdonvirta, V. (2018a). Online labour index: Measuring the online gig economy for policy and research. *Technological Forecasting and Social Change*, *137*, 241–248. https://doi.org/10.1016/j.techfore.2018.07.056.

___Kässi, O., & Lehdonvirta, V. (2018b). Online labour index: Measuring the online gig economy for policy and research. *Technological Forecasting and Social Change*, *137*, 241–248. https://doi.org/10.1016/j.techfore.2018.07.056.

MercadoLibre. (2018). *Annual Report, submitted to United States Securities and Exchange Commission, Form 10-k* (p. 141). Retrieved from http://investor.mercadolibre.com/financial-information/annual-reports.

Ookla. (2019). *NetIndex source data*. Retrieved from http://www.netindex.com/source-data/.

Peres, W., & Stumpo, G. (2002). *Las pequeñas y medianas empresas industriales en América Latina y el Caribe*. Siglo XXI.

Price, S. (2017). Lending Pioneer Kiva Hits The One Billion Mark And Launches A Fund For Refugees. Retrieved March 20, 2019, from Forbes website: https://www.forbes.com/sites/susanprice/2017/07/06/lending-pioneer-kiva-hits-the-one-billion-mark-and-launches-a-fund-for-refugees/.

UN ECLAC, (United Nations Economic Commission for Latin America and the Caribbean). (2018). *Data, algorithms and policies: Redefining the digital world*. Retrieved from https://conferenciaelac.cepal.org/6/en/documents/data-algorithms-and-policies-redefining-digital-world.

United Nations (2015), Transforming our World: The 2030 Agenda for Sustainable Development. Retrieved from https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf.

Vallejos, Z. (2003). *Micro, pequeñas y mediana empresas en América Latina*. Retrieved from https://repositorio.cepal.org/handle/11362/10874.

Waghorn, T. (2013). Premal Shah: Loans That Change Lives. Retrieved March 20, 2019, from Forbes website: https://www.forbes.com/sites/terrywaghorn/2013/11/04/premal-shah-loans-that-change-lives/.

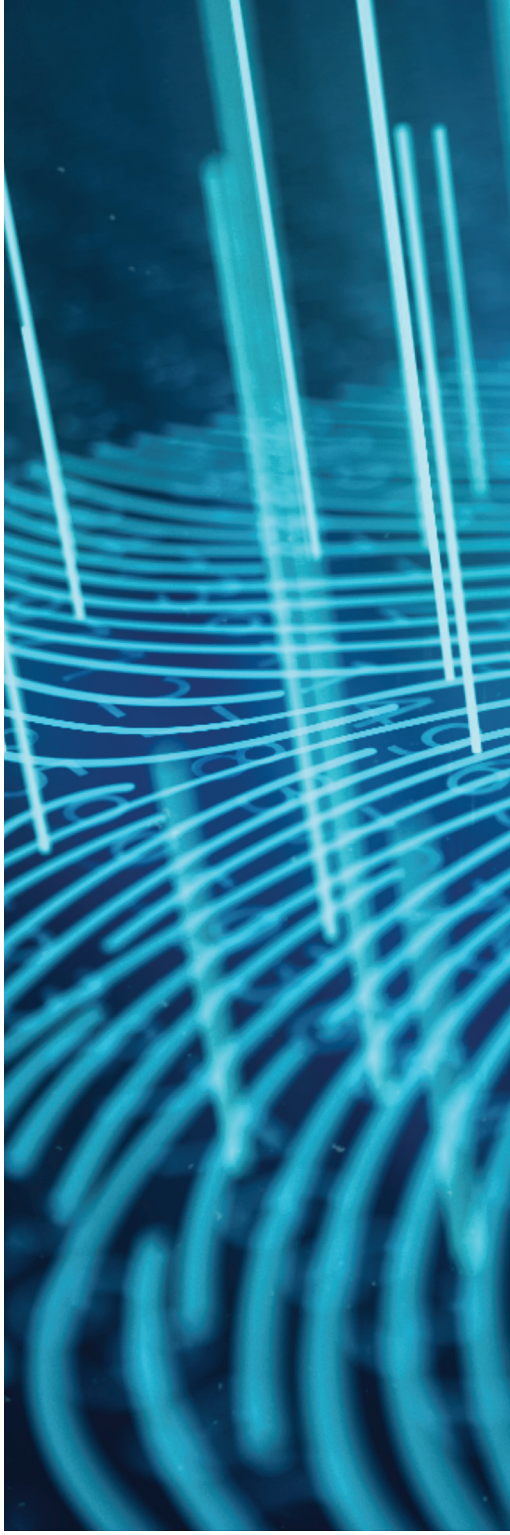
The global economy is becoming increasingly dominated by digital technologies, which are now the main platforms for the majority of socioeconomic activities, with positive effects on economic growth and well-being. In order to maximize these positive effects, empirical evidence is critical for development policies as it affects decision-making in areas ranging from resource allocation to impact evaluation.

This report explores the opportunities for and challenges of the systematic use of publicly available digital data as a tool for formulating public policies for the development of the digital economy in Latin America and the Caribbean. The objective is to share lessons learned in order to advance a research agenda that allows the countries of the region to create alternative measuring tools based on the digital footprint. Using big data techniques, the digital footprint left behind by labour market portals, e-commerce platforms and social media networks offer unprecedented information, both in terms of scope and detail.

The report pursues two complementary goals. It serves as (i) a state of the art on selected topics regarding the digital economy in Latin America and the Caribbean; and (ii) a basis for discussion about the opportunities for and challenges of working with online digital trace data for development policies.

The substantial contributions of the report are in line with the proposal made by the Economic Commission for Latin America and the Caribbean (ECLAC) to undertake policies that promote structural change for equitable and sustainable development by harnessing the full potential of the digital economy in order to shift the region's production structure towards more knowledge-intensive industries and higher-productivity sectors.

Making sense of big data in a meaningful way includes not only computational challenges, but also touches on the definition of data science as the convergence between computer science, statistics and their extensive application. In practice, it has quickly become clear that, from the perspective of data science, issues of representativeness, generalization, harmonization, data quality, and the definition of variables and indices are key concerns. The methodological contributions of the report stem from the insights gained from a big data analytics exercise. The report therefore serves as a rough guide for practitioners interested in using modern data science for development policies.



ECLAC

Economic Commission for Latin America and the Caribbean (ECLAC)
Comisión Económica para América Latina y el Caribe (CEPAL)
www.eclac.org

LC/TS.2020/12/REV.1