Statistical Commission
Forty-seventh session
8 – 11 March 2016
Item 3(c) of the provisional agenda
Big Data for official statistics

Background document
Available in English only

Report of the 2015 Big Data Survey

Prepared by United Nations Statistics Division

Report of the Big Data Survey 2015

June-August 2015

Executive summary

The Statistical Commission approved the proposal that the global working group conduct a global survey on Big Data for official statistics. The objective of the survey was to assess the situation regarding the steps undertaken thus far by statistical agencies in relation to Big Data. It enquired about the strategic vision of national statistical offices and their practical experience with Big Data. The questionnaire contained questions on the management of Big Data, advocacy and communication, linking Big Data with the Sustainable Development Goals, access, privacy and confidentiality, skills and training, and on the most urgent needs of statistical offices regarding the use of Big Data. In addition, the questionnaire contained detailed questions on Big Data projects, intended for those offices that had been engaged in one or more projects.

The survey was conducted from June to August 2015. A questionnaire was sent to each national statistical office with the request that it consult with the relevant stakeholders in its national statistical system. A total of 93 countries completed the questionnaire (32 member countries of the Organization for Economic Cooperation and Development (OECD) and 61 non-OECD countries). In addition, a reply was received from Eurostat. The questionnaire consisted of 18 questions on Big Data management and 24 questions on each reported Big Data project. About half of the countries reported at least one Big Data project, with a few reporting more than four. However, it should be noted that the other half of the countries did not report a project. Nevertheless, a total of 115 Big Data projects were reported: 89 by OECD countries, 22 by non-OECD countries and 4 by Eurostat.

Statistical offices consider "faster, more timely statistics", "reduction of response burden" and "modernization of the statistical production process" to be the main benefits of using Big Data, followed by "new products and services" and "cost reduction". Interestingly, whereas two thirds of the developing countries consider meeting new demands such as the indicators of the Sustainable Development Goals to be a benefit, only one third of the OECD countries share that view. Many offices have already used or considered using Big Data sources, mainly scanner data, web-scraping data, mobile phone data and satellite imagery data. This result corresponds more or less with the actual reported Big Data projects; most used were mobile phone data (42 projects), followed by web-scraping data (31 projects) and scanner data (23 projects). However, scanner and web-scraping data were used much more in OECD countries than in non-OECD countries. These results are also reflected in the statistical domains for which Big Data was most used: price statistics (based on scanner data), followed by tourism (mobile phone data), population (mobile phone data) and transport and labour statistics (web-scraping data).

Given that the use of Big Data in general involves working together with other offices to gain access to data and acquire the necessary technical skills, the survey asked what partnerships respondents had established for Big Data projects. Most often, respondents had partnered with government institutes, followed closely by the research and academic communities. They had established far fewer partnerships with data providers and information technology (IT) companies.

Furthermore, the surveys determined that most reported Big Data projects were in the exploration and research stage. However, if a project was intended to go into production, it complemented an existing data source. In addition, the survey revealed an interesting finding: both developing and developed countries continued to rely on traditional statistical methods in the processing and analysis of Big Data sources. That finding may be interpreted in at least two ways. First, Big Data projects can apparently be processed and analysed using traditional methods, which should significantly reduce the threshold for undertaking a Big Data project. Second, national statistical offices may not yet be very familiar with advanced Big Data tools for processing and analysis, which may hamper the full exploitation of Big Data sources.

In response to the question "Which skills are important for your office to acquire in order to better deal with Big Data?", respondents indicated "methodologist on Big Data issues", "Data scientist" and "mathematical modelling specialist" as the three most-needed skills, with the more IT-oriented skills of "IT architecture specialist", "Data visualization specialist" and "cybersecurity specialist" given less priority. Whereas statistical offices clearly see a need for these new advanced Big Data skills, they have not yet begun to hire candidates who possess them, nor have they sent their staffs to receive training on them.

The survey explicitly asked about the urgent needs for guidance that national statistical offices might have. The respondents indicated "skills and training for Big Data", "quality framework for Big Data", and "access to Big Data" as the three top guidance priorities, followed by "estimation methods", "use of web-scraping data" and "use of mobile phone data". This result is consistent with the needed skills indicated in the areas of Big Data methodologies, estimation methods and data science.

The following conclusions can be drawn from these results: (a) training and capacity-building on Big Data topics (methodologies, estimation and a quality framework) is necessary; and (b) more pilot projects are needed, especially those with participation from developing countries. In addition, easier access to Big Data will lower the threshold for undertaking a Big Data project. Consequently, training, Big Data methodology and quality frameworks, and data access have been chosen as the three top priorities of the programme of work of the global working group and are included in the themes for the next international conference.

Report

Part I and II

Respondents information

The survey was submitted by the National Statistical Office in all countries. In total 94 responses were received, of which 61 from non-OECD countries, 32 from OECD countries and 1 international/regional organization. For a few countries with more decentralized national statistical systems, the survey was coordinated by the NSO and separate responses where submitted by various institutes, departments or ministries involved in statistical production.

In terms of the division/section/unit responsible for overseeing Big Data projects within the National Statistical Office, the most common sections was the sections dealing with methodological issues or the IT department. Only a few countries had established a dedicated Big Data task team.

Big Data opportunities and potential benefits

Overall, the responses to the survey indicated that the statistical community consider that a paradigm shift is in the making and Big Data has been identified as a strategic priority by many offices. At the same time, many statistical offices see a lack of resources to invest in Big Data. The challenge to keep business as usual running with limited resources, while at the same time contemplating the use of Big Data, is a problem for some statistical offices.

Respondents indicated that the main reason to use Big Data for official statistics is to provide faster, more timely statistics. However, many of the benefits from the usage of Big Data are interlinked. Big Data can complement or replace existing statistics in a more timely, detailed and qualitative way, which can contribute to cutting production cost and reducing burden on enterprises and citizens, while at the same time increasing quality and providing new statistical information not previously available.

It was also mentioned that Big Data offers the opportunity to measure the changes in economic development, such as the so-called 'shared economy', and that the creative use of Big Data is strategically important for any innovative knowledge institution.

However, it was mentioned that there is a need to do further cost-benefit analysis of Big Data projects before concluding on its benefits vis-a-vis traditional data sources. The benefits of using Big Data, in particular timeliness, must be compared with statistical risks from using these alternative data sources which can have problems with respect to sampling design, data processing, etc.

Partnerships

OECD counties have to a greater extent started to establish partnerships on Big Data projects with other statistical agencies than have non-OECD countries, in particular, through initiatives such as the Eurostat/ESS Task force on Big Data and the UNECE initiatives related to Big Data. Developing countries have established fewer partnerships, but a few countries cite the

participation in the UN Global Working Group, or partnerships with Paris 21, World Bank and Global Pulse.

The most common partnership on use of Big Data is with other government institutes. In addition, a larger share of countries responded that they are trying to establish a partnership with mobile phone operators, highlighting the fact that not many offices have managed to establish a partnership with mobile phone operators. This could be due to legal obstacles for such a partnership. Further, many countries have established a partnership with retail stores and supermarket chains to obtain scanner data for price statistics.

Big Data sources and statistical domains

The most used sources are scanner data, satellite imagery and web-scraping data. In particular, it is worth noting that more than 80 percent of OECD countries have used or considered using web-scraping data and scanner data. Social media and mobile phone data are for the moment much less used according to the responses to the survey, due to a number of factors, in particular issues related to privacy and confidentiality. Non-OECD countries appear to be focusing on trying to make use of satellite imagery and mobile phone data.

Big Data has been mostly used for price statistics, followed by population/migration and labour statistics. In particular, the survey reveals that 94 percent of OECD countries have used or considered the use of Big Data for the price statistics. While the demand for better data for the monitoring of the Sustainable Development Goals is strong, the statistical community has not started to use or consider using Big Data for indicators related to the SDGs. However, respondents stress that they consider Big Data sources for all types of statistics and that it is quite clear that big data can be beneficial for nearly all of the statistical domains.

Advocacy and communication needs

The statistical community expressed a need for activities such as identifying key partners and funding sources as well as maintaining a list of Big Data projects. In the present phase of usage of Big Data for official statistics, it was argued that dialogues with private enterprises on the mutual benefits of sharing data are a priority.

The concept of Big Data is still new to many countries, and it needs to be well explained to the public as well as to producers and users of statistics. Also, to generate interest at political level, the statistical community stressed the momentum and demands arising from the Sustainable Development Goals as a great opportunity.

Linking Big Data and SDGs

In terms of using Big Data for measuring progress on the SDGs, very few countries have indicated they are currently doing this. However, when the SDG indicators are approved by the UN Statistical Commission, a number of countries report that they will start considering Big Data, consistently with the Fundamental Principles of Official Statistics on determining the most appropriate source for various indicators. However, related to this issue the use of administrative data for this purpose was stressed, and not to be overlooked due to too much focus on Big Data sources.

Access, privacy and confidentiality

Most countries have a framework for dealing with privacy issues and this framework (with the same strict rules and guidelines for statistical disclosure) also applies to Big Data. Only 28 percent of non-OECD countries and 41 percent of OECD countries have addressed aspects of access to data, privacy and confidentiality specifically for Big Data sources on a case by case basis.

Skills and training needs

Big data pose entirely unprecedented challenges in IT infrastructure, conceptual analysis and practical organisation. In terms of the skills needed to better deal with these issues, Big Data methodologist and data scientist are cited as the the top skills/occupations. Respondents defined that data scientists need a combination of three skills sets, namely (Big Data) methodology, IT and subject-matter knowledge.

Even though data science is considered as a much needed skill, the respondents do not actively hire data scientist since such position has not yet been properly defined. Therefore, as an alternative, statistical offices look for candidates that possess both statistical and IT expertise. Other approaches are to train existing staff, to hire short term consultants, and to establish collaboration and partnership with universities and research institutes. In particular, participation in Big Data conferences and workshops, UNECE Sandbox, inviting Big Data experts for in-house seminar and self-learning are cited as ways to improve the Big Data skills.

Moving forward, the statistical community expressed the need for guidance in the areas of skills and training, quality frameworks for Big Data and access to Big Data. There is also demand for guidance on the usage of specific data sources such as web-scarping and mobile phone data. In particular, lack of access to Big Data sources is cited as the major bottleneck, and many such sources are not restricted to national territories. Access to international Big Data sources has to be coordinated across national and international statistical agencies, hence the high need for international guidance. For less developed statistical systems, a full understanding of the concept and capacity development is the highest priority.

Part III: Projects

Project breakdown

The second part of the questionnaire focused on specific Big Data projects, and a total of 115 projects where submitted from 43 different countries and organizations. 89 of these projects where submitted by OECD countries, 22 by non-OECD countries and 4 by a regional organization. In terms of geographical regions, 47 projects where submitted by European countries and 36 by countries in the Americas.

Sources, statistical domains and partners

In terms of the Big Data sources used in the projects, the most common sources are mobile phone data, web-scraping data and scanner data. Other data sources often refer to administrative data such as tax files, health, social services and education data, but also other data sources such as online search data (Google Trends), high-frequency financial market data as well as simulated data,

The most common partners are other government institutes, academic and research institutes and intermediary Big Data providers. Other partners most often refer to retail chains, but also other partners such as news providers, payment gateways, international/regional organizations and credit card companies.

The projects are applicable to a range of areas of official statistics, but the most common areas are demographic and social statistics, and price statistics. In particular in OECD countries, the focus has been on using Big Data for price statistics, with many projects already in production phase.

Objective and data access

While countries responded that the main reason to use Big Data sources was to produce faster and timelier statistics, this was not the most common project objective. Most of the projects are intended to supplement existing data, or the projects are for exploration and scientific/research purposes.

Given that one of the major challenges in relation to Big Data for official statistics is related to data access, it is interesting to note that most of the projects make use of data that is accessible publicly. Some statistical offices obtain data for free under legislation or they access publicly available data from data providers, but incur costs related to data transfer and data collection. Many of the projects are still in negotiations with data providers and the exact conditions have not yet been decided. Some are currently looking into the changing or reviewing the legislative framework in order to ensure easier access to data from different data sources.

The survey also reveals that while most statistical offices in the OECD countries have ensured broader access rights to data, in most non-OECD countries, the statistical offices only have access to data for the specific project. This can be partly explained by the more common use of publicly available data, in particular using web-scraping, in OECD countries, and the use of administrative data, as well as broad agreements with retailers.

Tools and methods

In terms of methods and tools used in the Big Data project, most respondents indicated that traditional statistics methods and technologies (e.g., relational database, spreadsheet) are currently being used or were used. This may indicated lack of access to and knowledge of Big Data technologies such as Hadoop Clusters or other emerging data mining tools. In addition, data visualization methods and tools are cited a quite a lot which indicates that presentation of Big Data result is quite an important factor.

Notes on the analysis of response

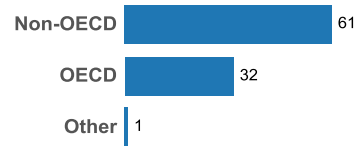
- Whenever a country submitted more than one response to the part I and II of the survey, the responses were consolidated to one response by country, representing the whole national statistical system. This was the case for Japan and Mauritius.
- 62 countries/organizations submitted their responses in PDF format. Their responses were inputted by the analysis team.
- In addition, two countries responded by email that they do not have any Big Data experience and could not complete the questionnaire.
- Not all respondents responded to all questions, and there are a number of 'null' responses. Overall these represent around 10% of the responses, and are included in the analysis for transparency.
- For the analysis of the responses by OECD/non-OECD countries, the responses submitted by the regional organization (Eurostat) was not included.
- The responses to the general part of the questionnaire were presented as a percentage of total responses, while the responses to the projects part of the questionnaire was presented in levels, for analytical purposes and clarity.
- A few questions were asked as open-ended questions. For analytical purposes these questions were classified and categorised. This was the case for the question "Division or unit responsible for overseeing Big Data projects within the organization" and question 33 (the statistical domains of the projects).

Responses

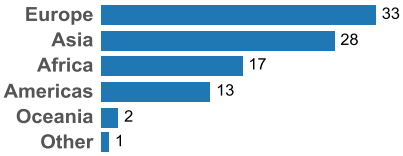
Analysis of respondents

Breakdown of responses by income groups and geographical regions

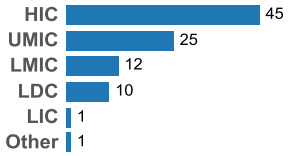
By OECD and non-OECD



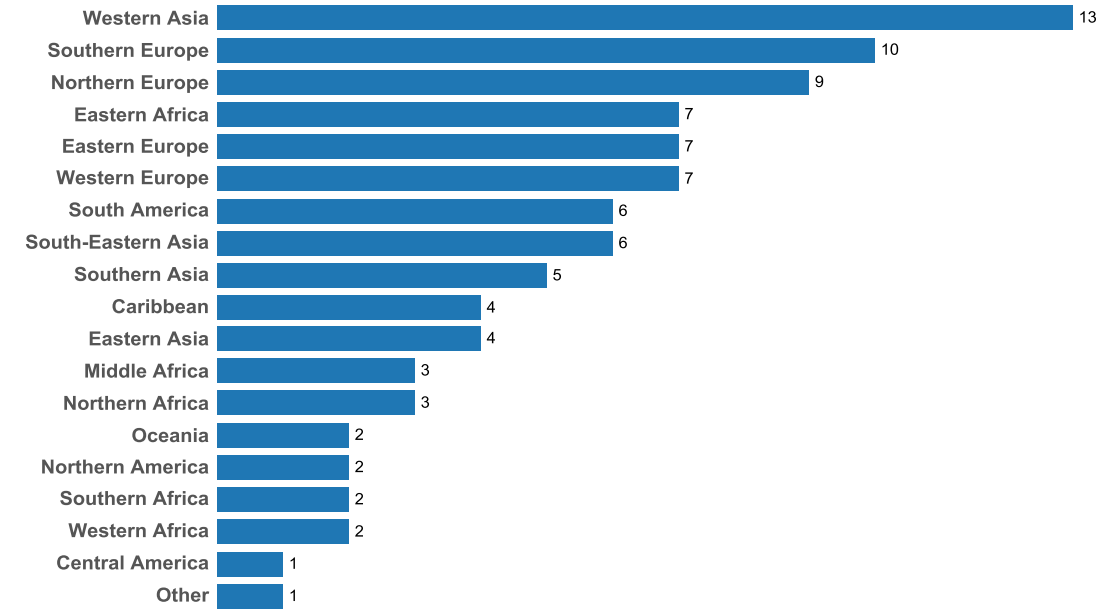
By geographical region



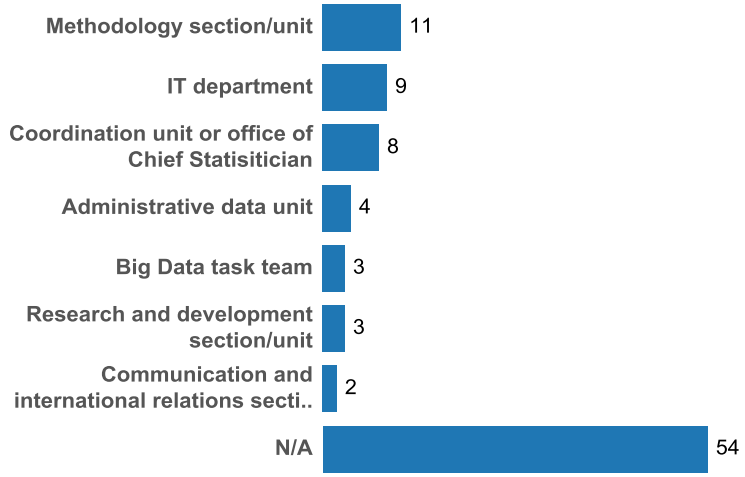
By income group



By sub-region



Division or unit responsible for overseeing Big Data projects within the Office



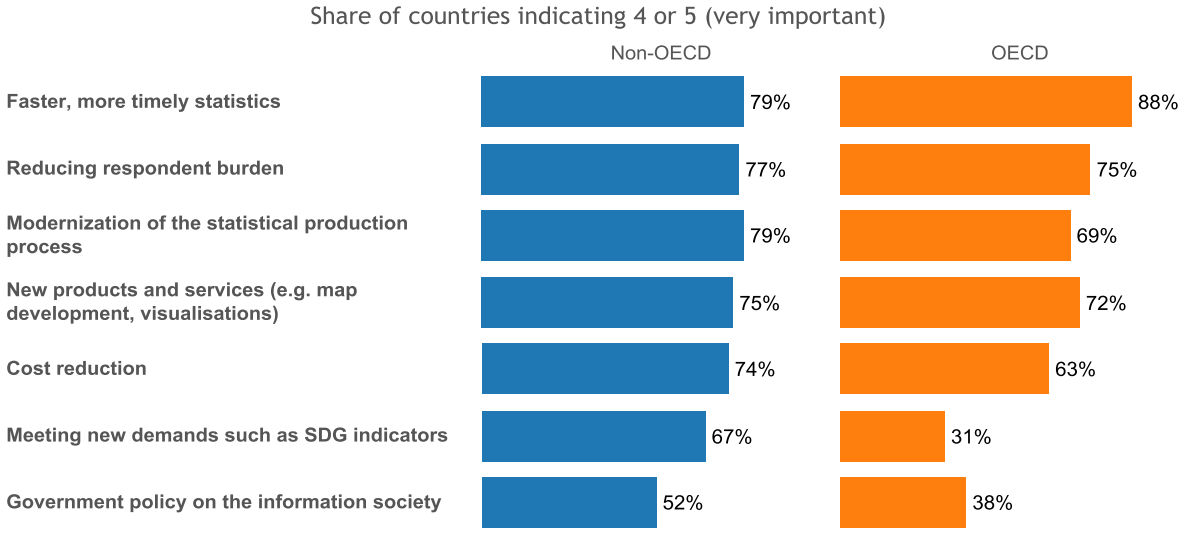
Part 1: Management, advocacy and communications, linking Big Data and SDGs, access, privacy, confidentiality, and skills and training

Office Management of Big Data

Q3 – Main reasons and business benefits of using Big Data

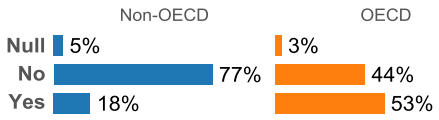
3. What do you see as the main reasons or business benefits for using Big Data in your institute? Indicate on a scale of 1 (not important) to 5 (very important), if these reasons play a role in your office.

	5 (very import..	4	3	2	1 (not import..	Null
Faster, more timely statistics	47%	35%	12%		2%	4%
Reducing respondent burden	46%	31%	13%	4%	2%	4%
Modernization of the statistical production process	44%	32%	18%	1%	2%	3%
New products and services (e.g. map development, visualisations)	43%	32%	16%	4%	2%	3%
Cost reduction	41%	29%	18%	6%	2%	3%
Meeting new demands such as SDG indicators	28%	28%	28%	7%	5%	4%
Government policy on the information society	22%	24%	32%	11%	6%	4%



Q4 - Collaborations

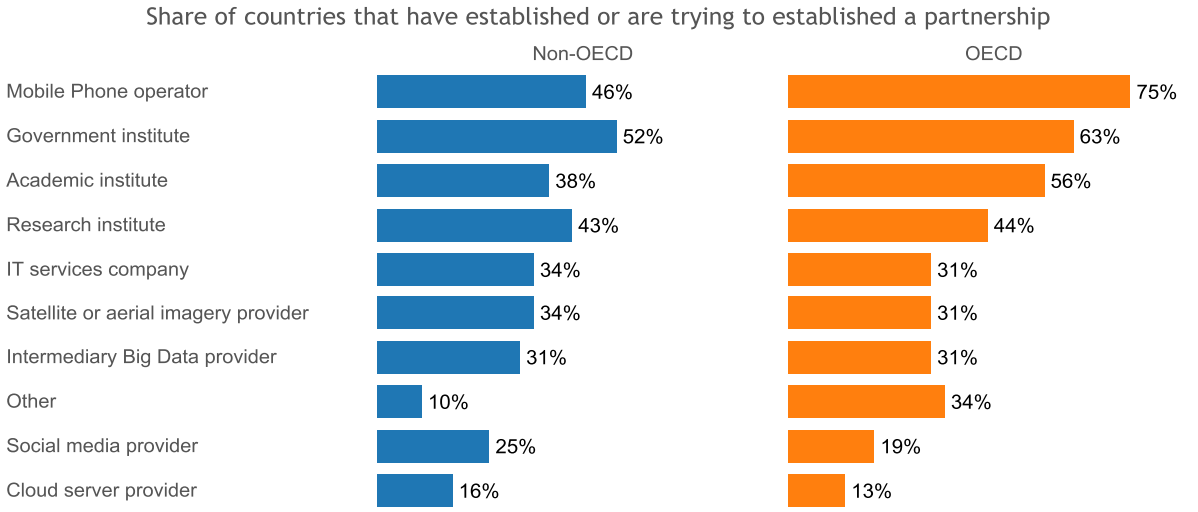
4. Have you established collaborations with other statistical agencies looking to use Big Data for statistical production?



Q5 - Partnerships

5. Which other specific Big Data partnerships have you established or are you trying to establish? Check all that apply.

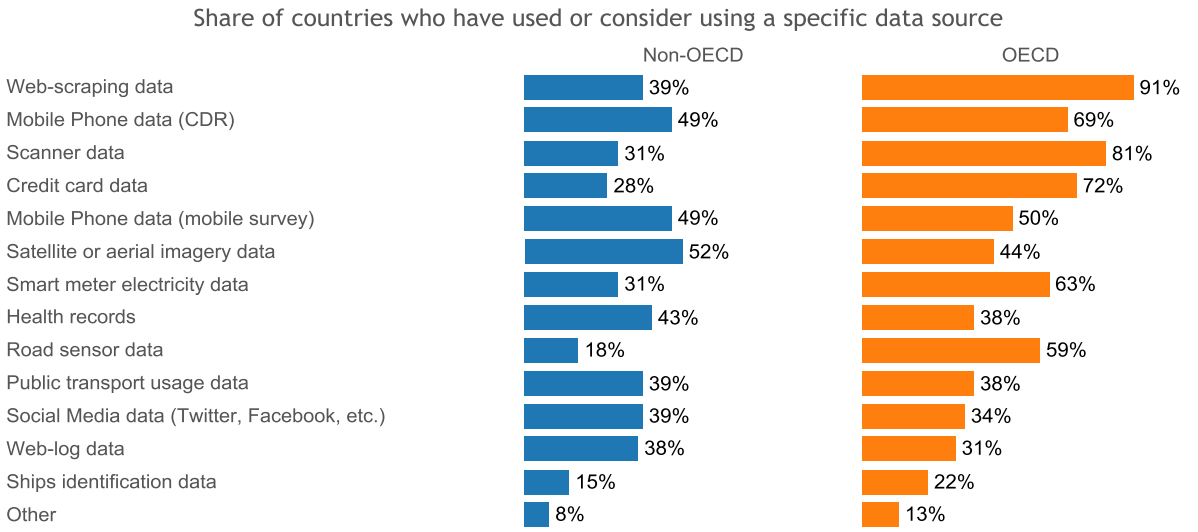
	Established	Trying to establish	Not considered	Null
Government institute	22%	34%	32%	12%
Satellite or aerial imagery provider	17%	16%	52%	15%
Academic institute	12%	32%	40%	16%
Research institute	12%	32%	46%	11%
IT services company	10%	23%	52%	15%
Mobile Phone operator	9%	48%	34%	10%
Intermediary Big Data provider	9%	22%	54%	15%
Other	7%	11%	27%	55%
Cloud server provider	4%	11%	68%	17%
Social media provider	3%	19%	63%	15%



Q6 - Sources

6. Which specific Big Data sources have you used or do you consider using? Check all that apply.

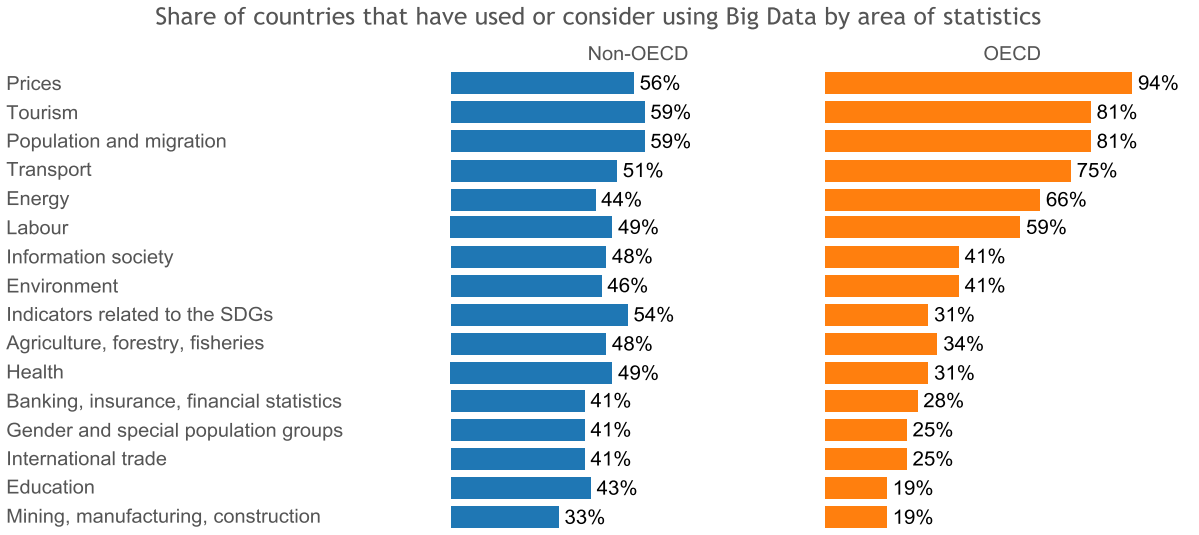
	Used	Consider using	Not considered	Null
Scanner data	22%	27%	37%	14%
Satellite or aerial imagery data	19%	31%	38%	12%
Web-scraping data	17%	40%	32%	11%
Health records	12%	29%	45%	15%
Credit card data	7%	35%	41%	16%
Public transport usage data	6%	32%	46%	16%
Other	6%	4%	28%	62%
Road sensor data	6%	26%	50%	18%
Mobile Phone data (CDR)	5%	51%	33%	11%
Web-log data	5%	30%	51%	14%
Mobile Phone data (mobile survey)	4%	45%	40%	11%
Ships identification data	4%	14%	65%	17%
Smart meter electricity data	4%	37%	46%	13%
Social Media data (Twitter, Facebook, etc.)	3%	35%	50%	12%



Q7 – Statistical domains

7. In which statistical domains did you use or do you consider using Big Data sources for production of official statistics or indicators?

	Used	Consider using	Not considered	Null
Prices	30%	39%	17%	14%
Population and migration	16%	51%	21%	12%
Labour	15%	38%	34%	13%
Health	13%	30%	43%	15%
Transport	13%	47%	27%	14%
Agriculture, forestry, fisheries	12%	31%	40%	17%
International trade	12%	23%	46%	19%
Education	11%	23%	48%	18%
Tourism	9%	59%	26%	7%
Banking, insurance, financial statistics	9%	28%	48%	16%
Energy	9%	43%	34%	15%
Mining, manufacturing, construction	9%	19%	53%	19%
Gender and special population groups	7%	28%	46%	19%
Environment	6%	38%	39%	16%
Information society	6%	39%	39%	15%
Indicators related to the SDGs	3%	44%	39%	14%

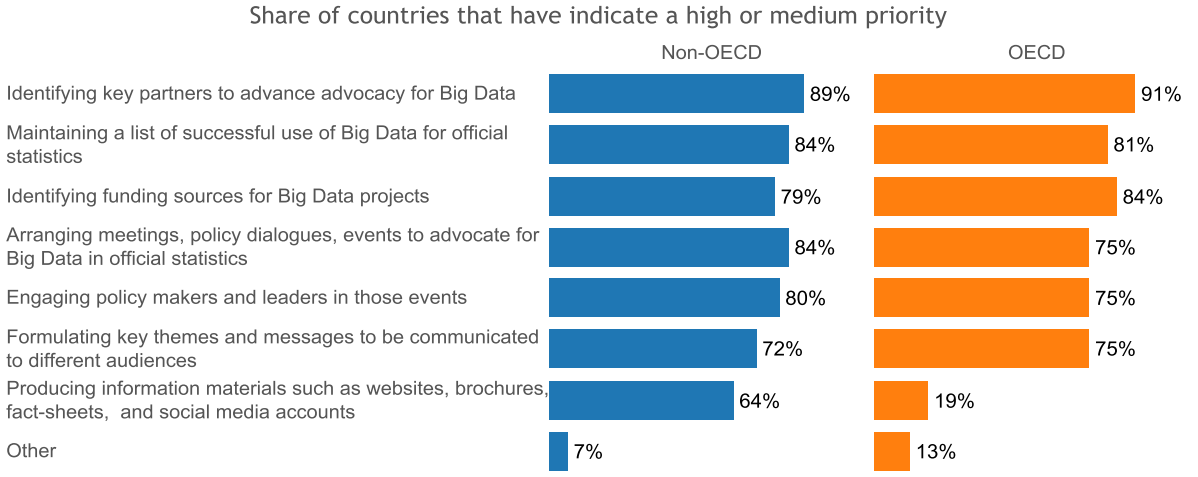


Advocacy and communications

Q8 – Training needs

8. Which activities does your office need to better communicate and advocate for the use of Big Data in your national statistical system? Indicate the level of priority.

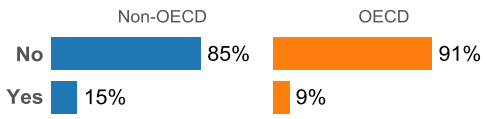
	High	Medium	Low	No need	Null
Identifying key partners to advance advocacy for Big Data	57%	32%		2%	9%
Maintaining a list of successful use of Big Data for official statistics	51%	32%	7%	1%	9%
Identifying funding sources for Big Data projects	59%	22%	9%	3%	7%
Arranging meetings, policy dialogues, events to advocate for Big Data in official statistics	45%	36%	9%	2%	9%
Engaging policy makers and leaders in those events	47%	32%	10%	2%	10%
Formulating key themes and messages to be communicated to different audiences	32%	41%	13%	4%	10%
Producing information materials such as websites, brochures, fact-sheets, and social media accounts	23%	26%	33%	10%	9%
Other	7%	1%	1%	14%	77%



Linking Big Data and Sustainable Development Goals

Q9 - SDGs

9. Has your office actively attempted or discussed linking certain Big Data sources for use of deriving certain indicators to measure progress on the SDGs?



Q11 – Reason to use Big Data for SDGs

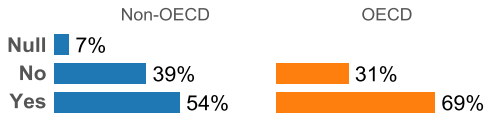
11a. If you use Big Data for compiling SDG indicators, does this improve the frequency?

11b. If you use Big Data for compiling SDG indicators, does this improve the disaggregation?

Access, privacy and confidentiality

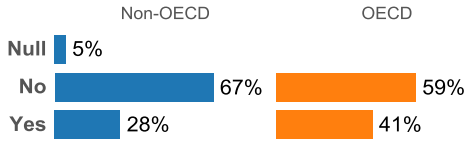
Q12 – Privacy framework

12. Does your office have a framework for dealing with privacy issues which also applies to Big Data sources?



Q13 – Big Data privacy

13. Did your office address aspects of access to data, privacy and confidentiality specifically for Big Data?

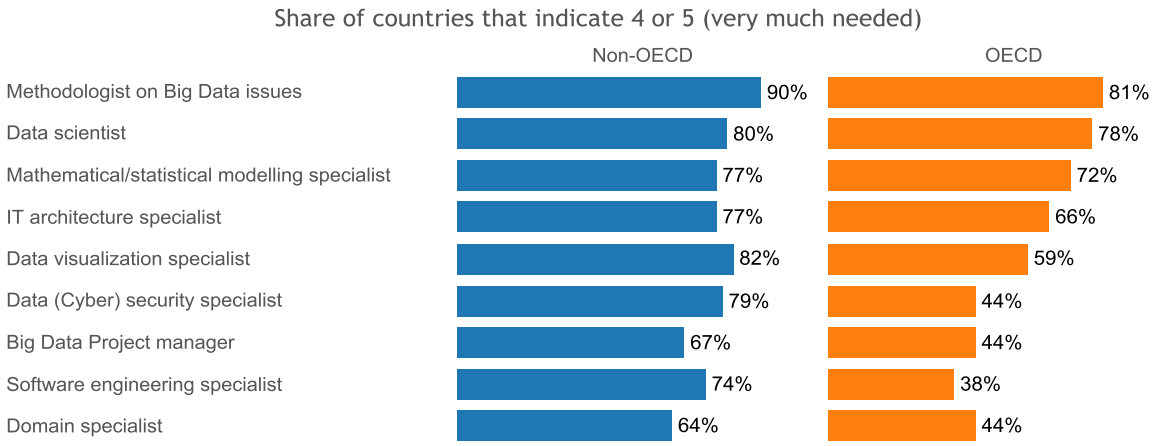


Skills and training for dealing with Big Data issues

Q14 - Skills

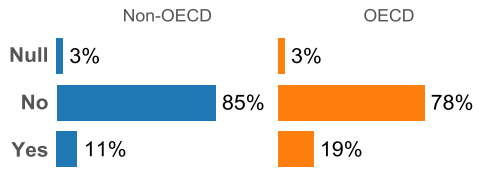
14. Which skills are important for your office to acquire in order to better deal with Big Data? Indicate on a scale of 1 (not needed) to 5 (very much needed), if those skills are needed in your office.

	5 (very much nee..	4	3	2	1 (not needed)	Null
Methodologist on Big Data issues	55%	31%	5%			9%
Data scientist	53%	27%	12%	1%		7%
Mathematical/statistical modelling specialist	47%	29%	14%	3%	2%	5%
IT architecture specialist	35%	37%	11%	4%	3%	10%
Data visualization specialist	33%	41%	14%	3%	2%	6%
Data (Cyber) security specialist	34%	32%	16%	5%	4%	9%
Big Data Project manager	32%	27%	21%	9%	3%	9%
Software engineering specialist	24%	36%	20%	9%	3%	7%
Domain specialist	22%	34%	20%	11%	3%	10%



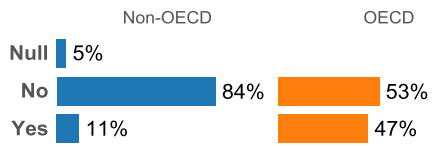
Q15 – Data scientists

15. Are you actively hiring data scientists?



Q16 – Big Data training

16. Are you actively training existing staff on Big Data topics?

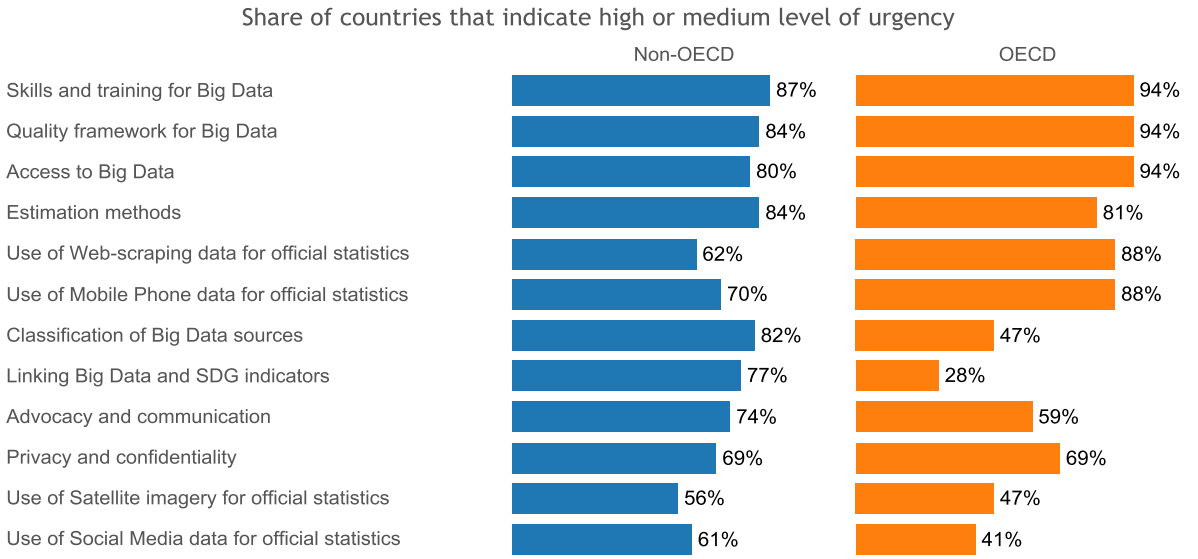


Part II: Needs for guidance on Big Data for Official Statistics

Q18 – Guidance needs

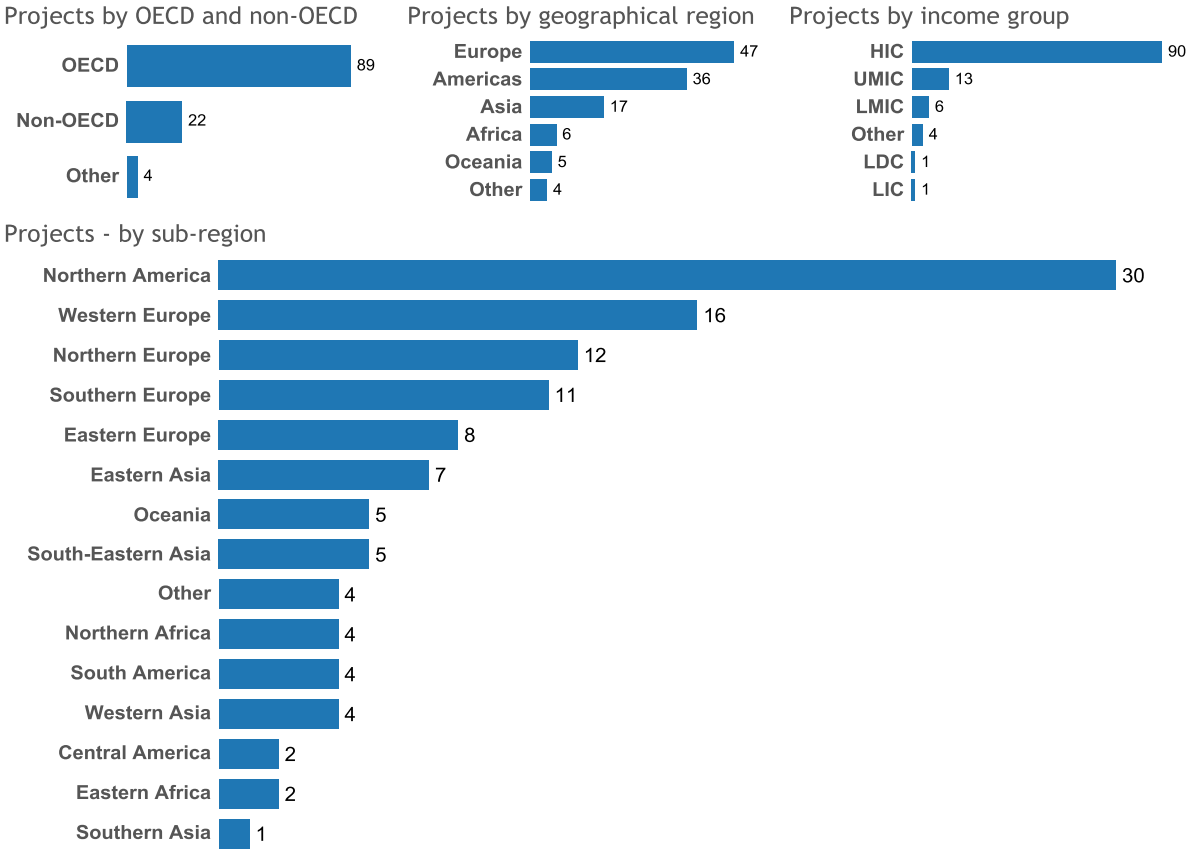
18. On which topics do you see an urgent need for guidance for your office or national statistical system? Indicate the level of urgency.

	High	Medium	Low	No need	Null
Skills and training for Big Data	69%	20%	2%	2%	6%
Quality framework for Big Data	62%	26%	5%	2%	5%
Access to Big Data	60%	26%	4%	2%	9%
Estimation methods	54%	29%	9%	2%	6%
Use of Web-scraping data for official statistics	43%	29%	15%	3%	11%
Use of Mobile Phone data for official statistics	41%	35%	13%	1%	10%
Classification of Big Data sources	36%	33%	18%	6%	6%
Linking Big Data and SDG indicators	31%	29%	29%	4%	7%
Advocacy and communication	30%	39%	15%	7%	9%
Privacy and confidentiality	29%	40%	14%	6%	11%
Use of Satellite imagery for official statistics	28%	26%	32%	6%	9%
Use of Social Media data for official statistics	27%	28%	30%	6%	10%



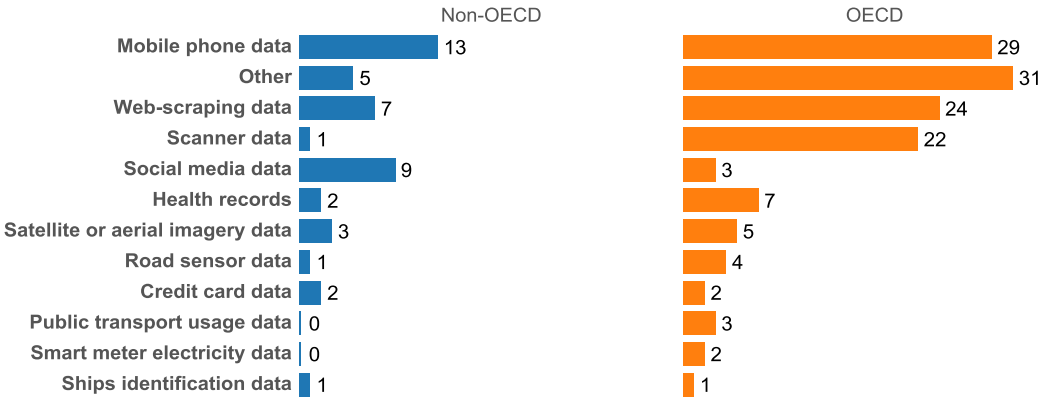
Part 3: Projects

Breakdown of projects by income groups and geographical regions



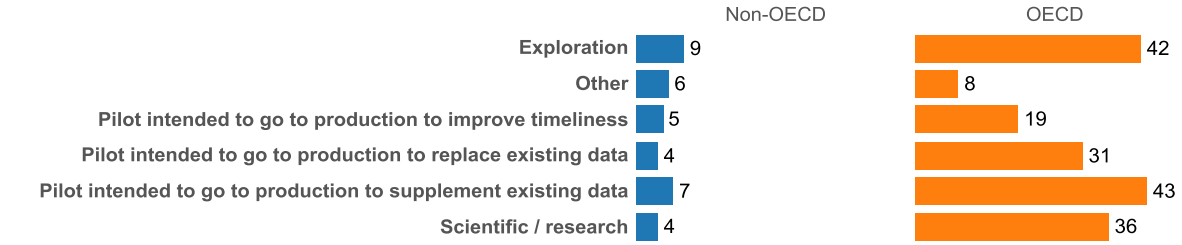
Q22 – Project sources

22. Type of data source(s):



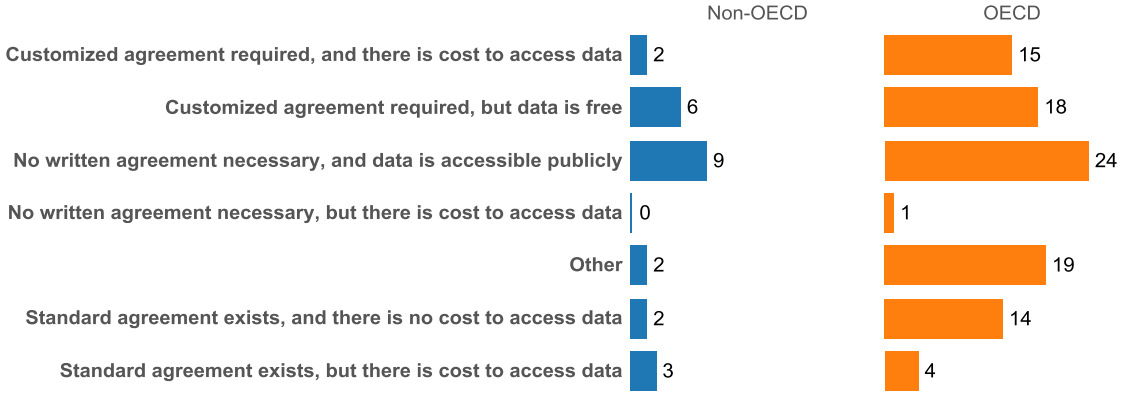
Q23 – Project objectives

23. Objective(s)



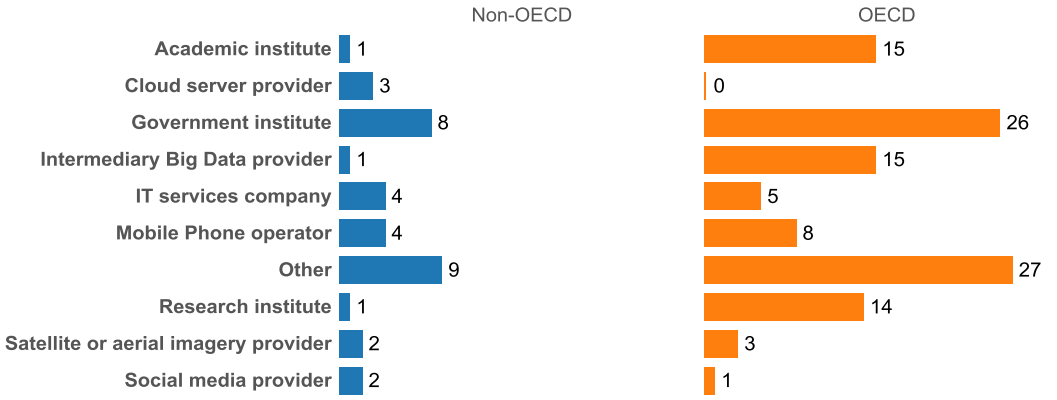
Q25 – Data access

25. Data Access: How did you obtain access to the data?



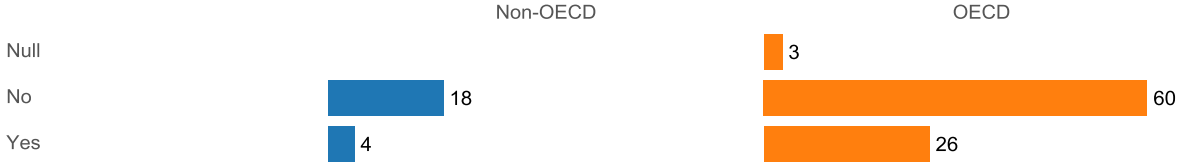
Q26 – Project partners

26. Partner(s):



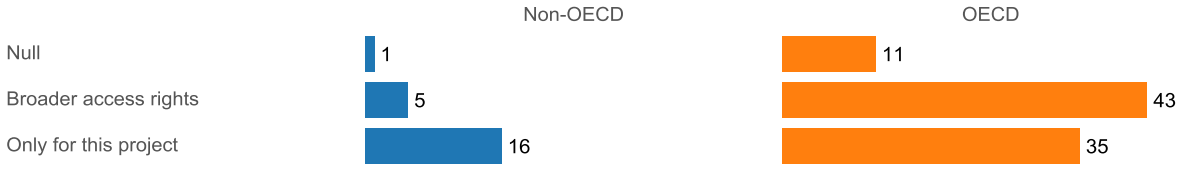
Q27 – Project intermediary

27. Do you use an intermediary (company or research institute) to obtain and prepare the data for your office?



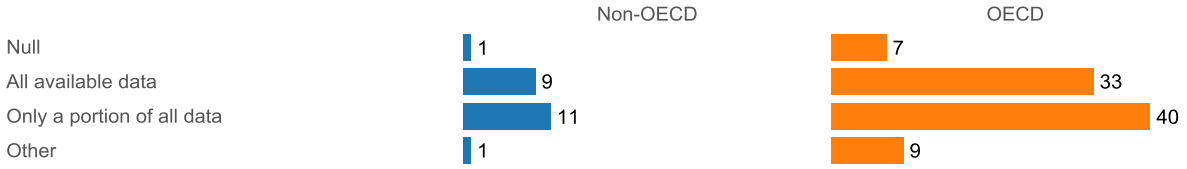
Q28 – Project data access

28. Do you have access to the data only for this project and its purpose?



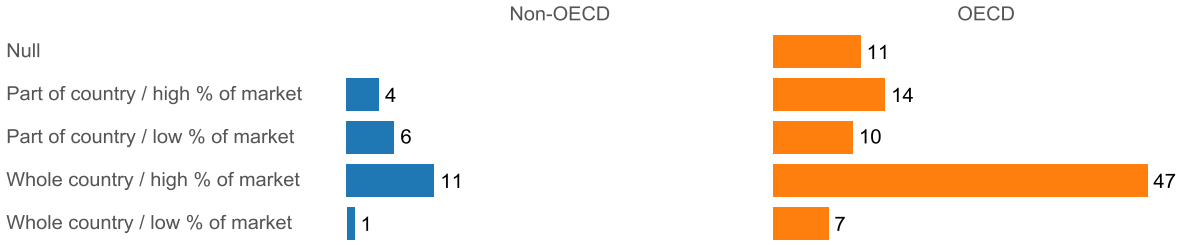
Q30 – Project data coverage and frequency

30. Coverage and frequency of the data



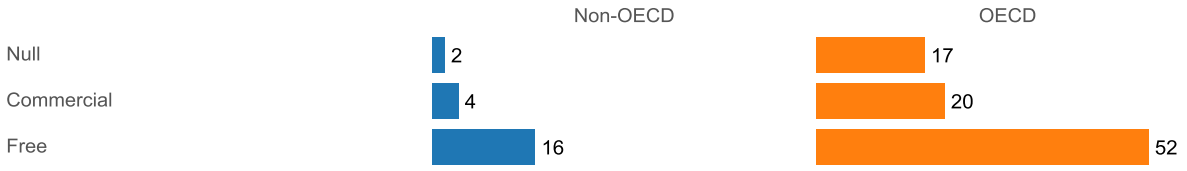
Q31 - Project geographic coverage

31. Coverage (geographic and population)



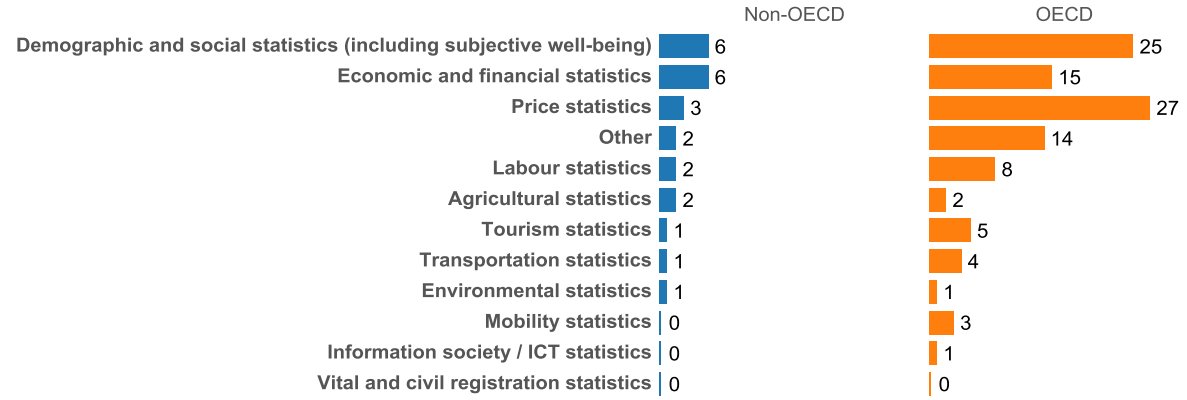
Q32 – Project data costs

32. Cost implication:



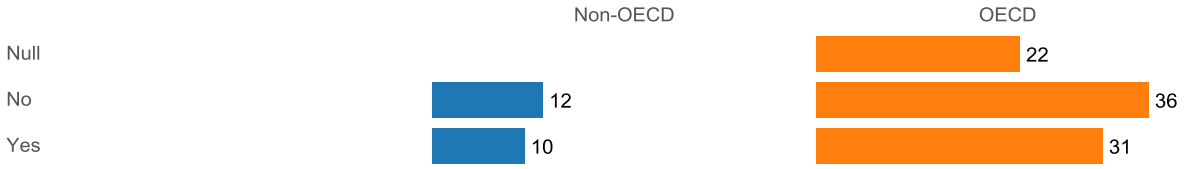
Q33 – Project statistical domain

33. Applicability to which area of official statistics



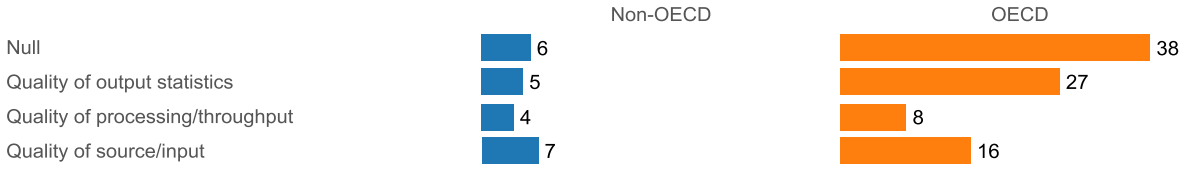
Q34 – Project validation

34. Are you using "training" or "ground truth" data to validate the results obtained?



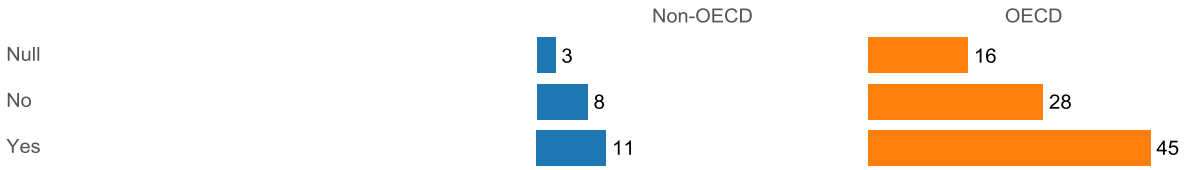
Q35 – Project quality frameworks

35. Have any existing quality frameworks been applied to this source, in terms of:



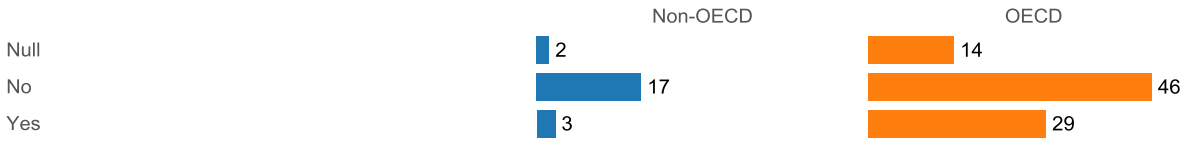
Q36 – Project data concerns

36. Do you have concerns about the quality of the data you have obtained?



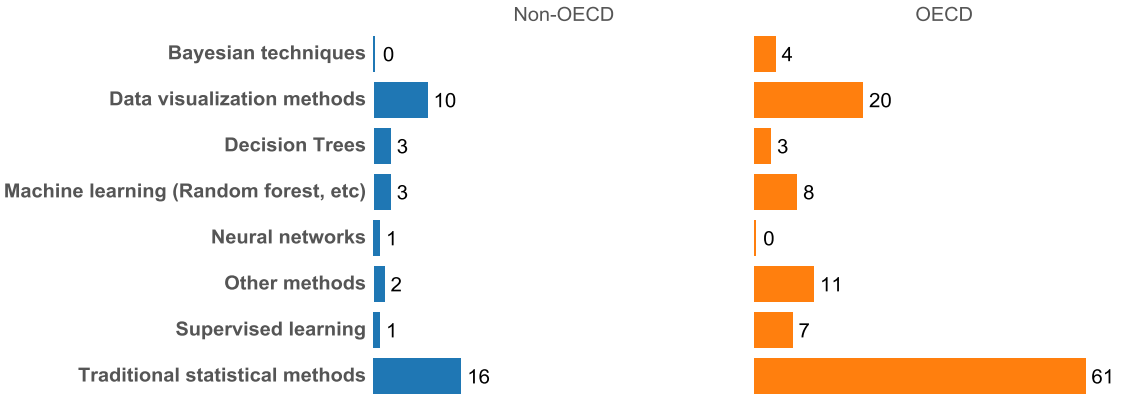
Q37 – Project estimation methods

37. Have you developed new estimation methods or a methodological framework specifically related to the data source(s) used in this project?



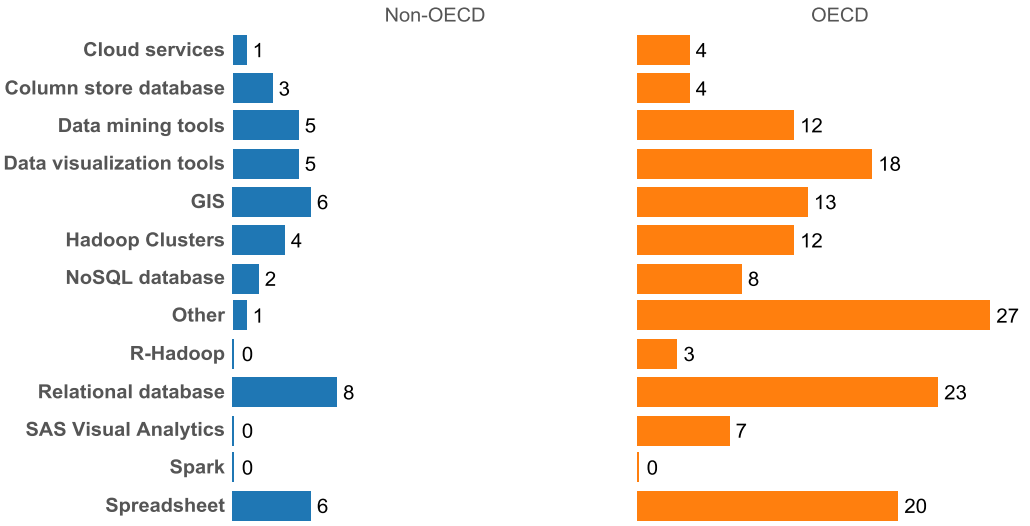
Q38 – Project methods

38. In this project, what methods were used or are being used during the Big Data processing life cycle?



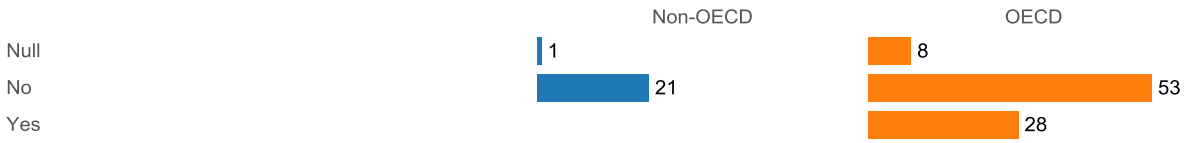
Q39 – Project technologies and tools

39. In this project, what technologies and tools were used or are being used during the Big Data processing life cycle?



Q40 – Project publications

40. Did you publish any research articles arising from your project and (separately) publications that have been very influential in your project design.



Q41 – Project evaluation

41. In the overall evaluation of your Big Data project do you evaluate the following quality aspects? Please check all that apply