Responses to the reviewers and editor

Editor:

1. Add an 'Author Contributions' section

Done

1. Upload figures as separate files

Not applicable

1. Please upload with your final files a signed version of the Author Publishing Agreement

IRF not yet received

Reviewer 1:

1. Add a scheme for each principle

The principles were derived in relation to the FPOS. This is shown in a standard way for each principle.

1. Explain “for statistical purposes only”

Wording has been modified to better explain “for statistical purposes only” in section 3.3.1

1. Explain GSMA privacy guidelines

Additional explanation is provided at the end of section 3.3.3

1. Please illustrate the projects

Given the length of the paper we did not put figures into the country example. More detailed information and results from the country cases can be retrieved from the websites of the project partners.

Reviewer 2:

1. Some of the principles stem almost entirely from a statistical context, making the principles useful only in use cases that involve statistical agencies.

Section 1.1 has been redrafted to better explain why this paper is written from the perspective of the NSOs

1. **Section 2**. Explain how anonymization and aggregation have a significant effect on the accuracy of data.

In the introductory part of section 3, more information is provided on anonymization and aggregation

1. How do the five principles contribute to the concerns of discrimination of minority groups?

In section 2, page 4, precautions regarding minorities have been explicitly mentioned. The length of the paper did not allow to go into further detail.

1. Explain “burden for respondents”

The last paragraph of section 3.1 gives additional explanation on “burden of respondents”.

1. How should intrusiveness and usefulness be balanced?

Throughout the paper this topic (trade-off between privacy protection and fit-for-purpose) is considered, most explicitly in section 5, last paragraph on page 23.

1. In terms of proportionality, I would also consider the efforts required from mobile operators to provide statistical data

This is addressed in section 3.1

1. Can you reframe the principle of professional independence to also cover other actors in addition to statistical bodies?

Given the focus of the paper and the role of NSOs in society, the responsibility of professional independence within these projects has been put on the NSO (see section 3.2)

1. The three principles of privacy protection (statistical system, data protection authority, private sector) seem artificial and difficult to understand. Does this refer to differences in the responsibilities of the parties in privacy protection?

The introduction of section 3.3 has been redrafted to cover this point

1. In a global context I would also suggest mentioning the international treaties protecting privacy, such as the Universal Declaration of Human Rights.

Reference to the Universal Declaration of Human Rights has been added to section 3.3

1. Explain pseudonymization and anonymization

Explanation regarding pseudonymization and anonymization has been added to section 3.3 (last paragraph) and section 3.3.2

1. Elaborate on the crucial point about need to consider whether the objectives of mobile data analysis could be achieved with aggregate, non-identifiable data

This point was explicitly added towards the end of section 3.3.1 and is mentioned at many places throughout the paper

1. Compliance with applicable laws and accountability. I believe that both of these principles could be addressed in more detail

Accountability is explicitly addressed in section 3.3.3

1. Questions about the selectivity and representativeness of mobile data when collected from individual operators. The argument is well explained but could be deepened by including comparability challenges in combining data from several operators

Challenges in combining data from several operators have been mentioned in the country cases of Estonia and The Gambia. The length of the paper did not allow for a deeper analysis.

1. The importance of verifiability and reproducibility as statistical principles. However, in the context of mobile data analytics there may be practical concerns for operators to share information to fulfill these principles

Open source policy (use of GitHub) would address this, and has been mentioned a few times in the paper; for example, at the end of section 3.4

Country Examples

* 4.1.1 This sentence is difficult to understand and requires restructuring for clarity: “On 17 March 2020, preparations began to use mobile operator data to provide essential mobility information on questions like: how did mobility of people change during the emergency situation and did those people, who returned from foreign countries, not move around, but remain in one place.”

We modified the sentence to the following outcome: “On 17 March 2020, Director General of Statistics Estonia started discussions on how to derive mobility information from the mobile operators’ data to manage better the emergency situation.”

* 4.1.3 The point about the Estonian Electronic Communications Act only allowing anonymous data processing is significant as this is the case in most EU countries greatly affecting feasibility of mobile data analytics in the member states. However, later in the section the authors mention lack of time as an explanation for not being able to conclude an agreement about more specific data. This seems contradictory with the statement that Estonian law prevents the use of non-anonymized data.

This is important remark, as indeed, the discussions around how to interpret Electronic Communication Act clauses to allow processing of MNO data only when it is anonymous. Some legal teams say that as any operation with the data is processing then even anonymization is not allowed. Some legal team say that if anonymous aggregated data is end result of processing then that would comply with the Act. In reality, everyone is waiting for the new EU ePrivacy Regulation that would be more specific and when ePrivacy Regulation would come into force then the old Act would lose its power. But as we have been waiting for ePrivacy Regulation for years now, then Statistics Estonia has started to change Electronic Communication Act so that there would be written more clear exception on using MNO data without user consent for producing official statistics. We modified this section so it would be better understood:

“While GDPR allows analysis of pseudonymous data for statistics, the Electronic Communications Act (ECA) only allows the processing of anonymous location data. Pseudonymous data means that direct personal identifiers (e.g name, ID-code, address or mobile number) have been replaced with codes that could not be linked back to identities without additional information and substantial effort. Anonymization is processing data using disclosure control techniques (e.g micro aggregation, local suppression, etc.) such that it is not possible to identify persons from the data directly or indirectly. There is ongoing legal debate if using these techniques on MNO data is in accordance with ECA. Since there was not enough time to work out an agreement with the MNOs on using non-aggregated data, it was decided to retain only the mobility monitoring mission requesting the MNOs just to carry out a simplified calculation to be aggregated and forwarded to Statistics Estonia. Statistics Estonia’s role as an intermediary was crucial as the MNOs did not want to show each other the regional distribution of their customer base.”

* 4.1.3 This section presents a good argument about the importance of Statistics Estonia acting as an intermediary due to the operators’ unwillingness to share data with each other. I assume this is a very common practical concern that could be raised also in the context of the general principle about professional independence.

We added under adherence paragraph “In this project Statistics Estonia was important intermediary to aggregate MNO data and not to disclose their business secrets.”

* 4.1.6 The summary section contains good reflection on the project and suggestions for improvement in future projects. What I found particularly interesting was the analysis on differences in data between operators due to each of them anonymizing data on their own terms. The suggested solutions of extensive documentation of methodology, testing, and quality analysis seem effective but may not be fully feasible due to legal restrictions.

We believe that there are no legal restrictions on doing extensive data quality analysis. These would add costs to all parties involved, but we think that principles of commitment to quality is very important factor if we want to produce statistics from mobile phone data for long term and win public trust over it.

* Data availability statement: The linked website contains Covid-related statistics from Estonia but does not provide anything on mobile data or the project described (site visited on 18 December 2020).

The mobile operator data is not available and even the aggregates from MNOs has been deleted by Statistics Estonia. We modified the statement “The aggregated mobility data and general methodology are described on Statistics Estonia’s website. Any other data that was used within the project has been deleted by involved parties as it was agreed in the beginning of the project”

4.2 Ghana

* 4.2.1 It is interesting to compare the longer-term projects in Ghana and The Gambia to the expedited implementation in Estonia

This point is address under “Principles of necessity and proportionality” on page 22.

* 4.2.1 The text mentions that the project was conducted in a legally compliant and privacy-preserving manner. It would be interesting to have a short description about the main concerns and solutions in this regard. I also noticed that compliance with the GDPR was mentioned instead of local laws. Was there a particular reason for that?

Reference has now been added to Li et al. (2021) which provides more details.

* 4.2.4 The section states that it was not possible to make statistical inferences about the whole population using only the available dataset (it was previously mentioned that the data came from a single operator). This is an important issue to acknowledge. However, sometimes single-operator data can be statistically extrapolated to the whole population. Were there any particular reasons why this was not possible in this case?

We are currently working on statistically extrapolating to the whole population using survey data. This update has been included in the write-up. At the time we performed the initial analyses to support COVID-19 intervention, this milestone of scaling to the whole population was a phase still in planning.

4.3 The Gambia

* 4.3.2 The list of mobility indicators (population distribution, home location, mobility) is interesting but could be clarified by explaining how this information was derived from mobile data.

The following descriptions on the indicators were added as a footnote:

“Proxy for population count uses the number of unique IDs active (call, SMS, and data communication) within a day and region. Proxy for home location is defined as the modal location of the last observation on each day of the week, where the user is most frequently in the evenings or at night. For more detailed explanation please see the World Bank COVID19 Mobility Task Force repository on GitHub (with hyperlink to https://github.com/worldbank/covid-mobile-data).”

* 4.3.3 It is stated that the data was produced from the CDRs by PURA (regulatory authority). This sounds to deviate from the other two projects where I understood that the operators provided the data and no raw CDRs were processed by external parties. If this is correct, how was the Gambian setup decided and did the different implementations affect the final analyses?

Access to raw CDRs: Raw CDRs were processed by respective operators to ensure that data provided to PURA were de-identified and no personally identifiable information was included. In this regard, PURA did not have access to raw CDRs.

Data processor: In this setting all parties agreed that PURA was a relevant data processor who aggregated the de-identified CDRs of two operators as the operators were competitors in the mobile market. In The Gambia there were existing data collection protocols between the MNO and the PURA where PURA already had put in place a centralized repository of statistics derived from CDRs, which was plugged into the respective MNOs systems as part of its mandate to monitor the quality of their services.

Impact on the final analysis by this setup: The Gambian setup is different from other countries but the data flow, in which external parties had access to only aggregated data, is similar to other country cases. The implementations do not affect the final analysis.

To clarify above mentioned, the following sentence was added to the first paragraph of this section: “CDR data are de-identified by respective MNOs to ensure that no individually identifiable information is included in the data used for producing statistics by PURA.”

* 4.3.3 It is mentioned that the implementation setup eliminates privacy issues. Could you elaborate the privacy concerns and their mitigation especially in the context of exporting CDRs from operators to PURA?

Measure taken to mitigate privacy risks are mainly summarized in 4.3.5 (Adherence to the principles - Principle of privacy protection) but the original text was not clear about what process was made by which organization. So, the name of data processors was added:

“The procedure of data pre-processing and aggregation enabled to maintains privacy. Any individually identifiable information was de-identified by respective MNOs, and de-identified data were aggregated at the regional level by PURA.”

* 4.3.4 The names of the two participating operators are not mentioned unlike in the other two projects. If there is a particular reason for this, such as confidentiality agreements, it could be mentioned for clarity.

We do not have any confidential agreements on the disclosure of operators’ names. To reflect this comment, we added the following sentences to elaborate the nature of populations covered by the CDRs of the two operators in addition to the market coverage.

“The population covered by the two MNOs are primarily different socio-economic groups. One of them is a leading MNO in The Gambia and is popular in urban areas with high-speed internet services. The other MNO provides only voice and short-messaging services with inexpensive plans, which are much popular in rural areas.”

We consider that the above information has similar significance as the names of operators.

* 4.3.5 Principle of professional independence: I welcome the publicity of the project and find it an important feature in building trust to this kind of projects. However, the comment does not seem comparable to how the other two projects assessed this principle.

The description of “Principle of professional independence” was elaborated as the following so as to be comparable to other country cases as below.

“The methodologies employed through this initiative were examined by GBoS and PURA with their strictly professional considerations. GBoS ensured the scientific principles and professional ethics while PURA oversaw the data access and computation process to protect privacy. As part of the process to develop the data pipeline, PURA and GBoS established a platform to strengthen policy relevance of the use of CDR data. Building consensus among all stakeholders and using strategic alliances embedded in the country dialog helped foster ownership and sustainability of this initiative. These procedures including methodologies employed for computing mobility statistics are disclosed to the public with documentations and accessible at any time.”

Chapter 5 – Conclusions

* Principles of necessity and proportionality: The differences in implementation between the three projects, most importantly the advantage of existing data pipelines in Ghana and The Gambia, provide a productive starting point for comparisons. I would like to hear more about the concrete features and results the longer implementation timeframe allowed in Ghana and The Gambia as compared to Estonia.

Wording under this point has been elaborated

* Principle of privacy protection: I find it surprising that the protection of privacy was similar in all three cases. This sounds contradictory with the project descriptions earlier in this article, from which I understood, for example, that Estonian data was anonymized already by the operators, whereas in The Gambia the data was processed by the authorities. Also, I assume considerable regulatory differences may exist between these three countries. I suggest that the argumentation leading to this conclusion of similar privacy is documented in more detail.

This point has been now been better explained and is more consistently applied throughout the paper.

* Principle of commitment to quality: The finding that the input data was different between the Estonian and the two African projects is significant, and probably merits emphasizing in context of other principles too. See my previous comment about privacy.

The trade-off between privacy protection, fit-for-purpose and commitment to quality has now been further explained in the last paragraph on page 23.

* Principle of international comparability: The difficulties in international comparison due to different contexts and populations is a relevant finding. I believe this observation could offer more value if it was expanded beyond these three projects. Do you find incomparability a general challenge in this kind of project and how does it affect the relevance of this principle altogether? Was there something special in these three projects that made them incomparable?

This would be a good topic for a separate paper.

* The suggestion that full open access to the data for statistical bodies and their partners could help develop relevant indicators is probably true. However, there could be practical concerns for the operators to open their data, such as regulatory considerations or business sensitivity. Furthermore, broader access to the data could also have a detrimental effect on privacy.

The sentence that “Full open access to the data would give the NSO and its service provider all possibilities to develop detailed, high-quality and relevant indicators” was made as a theoretical statement not as a proposition.