Dear Dr. Kenth Engø-Monsen,

Thank you very much for giving us the opportunity to submit a revised draft of our manuscript titled The Hidden Potential of Call Detail Records in The Gambia to Data & Policy. We appreciate the time and effort that you have dedicated to providing your valuable feedback on our manuscript. We are grateful for the insightful comments. We have been able to incorporate changes to reflect the suggestions provided. Here is a point-by-point response to the comments and concerns.

Comment 1: Section 3 B. contains copies of paragraphs contained in Section 3 A. Remove the redundant text.

Response:
The redundant text was removed.

Comment 2. It is not clear from the descriptions whether only charging data is used (CDRs), or if more detailed location data from network probes is the bases for the pipeline. The significance of this is that CDRs will only be generated whenever a customer initiates a service, whereas data from the network will continuously measure customers' location. The unclarity stems from the following statement in Section 3 C. Technical: "CDR data are massive datasets, which are huge in size and generated with high speed." This is not true for CDR data in general. However, it is certainly true for network data.

Response:
We use only CDR data. The following text was removed to avoid the confusion: "CDR data are massive datasets, which are huge in size and generated with high speed. Depending on the number of subscribers, the size of data can be several hundred gigabytes to terabytes and can reach the limit of ordinary hardware and software systems. To respond to these challenges, the"

3. "Ensuring privacy, the identifiable data field will be anonymized using a hashing algorithm, which is irreversible to original data." Which identifiable data fields are hashed? According to European legislation (GDPR) hashing in itself does not guarantee anonymity, so please clarify the definition of when anonymity has been obtained.

Response:
We revised the original text to "Ensuring privacy, the identifiable data fields such as the IMEI are encrypted and replaced with random numbers before the analysis." Further explanations including when anonymity has been obtained are added to the footnote 3 in Section 5 A where we also explain about the data field used for defining the number of subscribers. We consider

that the anonymity is obtained once the encrypted identifier is replaced with the random number by the regulator (PURA) after combining datasets from two MNOs.

Added footnote 3: IMEIs are used as a data field for defining the number of subscribers for the analysis while the IMEI generally defines the number of devices and the IMSI defines the number of SIM cards. This is because we combine datasets from two MNOs. In developing countries like The Gambia, it is common that one device is used by a person who has these two MNOs' SIM cards. To avoid double counting the same person, who exists across two MNO data, the data was de-identified by respective MNOs and the encrypted identifiers were replaced with random numbers on regulator's premise after the two datasets were combined.

4. "The mobile penetration rate of The Gambia was 94.2% in 2013 and rose to 140% in 2018." This indicates that multi-SIMing is very frequent in The Gambia. In the solution, when counting the number of travelers between locations, how is multi-SIMing accounted for in the counts to make sure the counts are not inflated due to multi-SIMing behavior? It is important to get the counts right, since these are proxies for population travel patterns.

Response:
We think that using the IMEI could mitigate the impact of the multi-SIM holding, In addition to the footnote 3 mentioned above, we added the following description to Section 4 A.
"Like other developing countries, multiple-SIM-card holding is common in The Gambia. We expect a certain overlap between the two-MNO subscribers, which might have resulted in over-representing the multiple-SIM-card holders. In this study the impact of the multi-SIM holding on the analysis result is considered to be limited since the two MNOs primarily market 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."

5. Page 7; line 4-5: "These identifiers are encrypted using a one-way function by the MNOs so the data provided to the regulator do not include any personally identifiable information." Hashing is only de-identifying the data records and not fully anonymizing them. Please note that de-identification through hashing is different from anonymous, and these are two different things. According to GDPR, the de-identified data is still potentially sensitive, and should be considered as personally identifiable information. The reason is that an adversary may possess another dataset that together with the deidentified dataset renders it identifiable.

Response:
Thank you for the note about the difference between the de-identified data and fully anonymized data. We revised the sentence to the following:
- First, identifiers are encrypted using a one-way function by the MNOs on their premises.
- Second, the encrypted identifiers are replaced with the random numbers after the data are combined…..

6. Page 7; line 12: Suggest rewriting "… but the above-mentioned aggregation process lowers the risk of being reverse engineered substantially" to ""… but the abovementioned aggregation process lowers the risk of reverse engineering."

Response:
Thank you for the suggestion. The sentence was revised accordingly.

7. What is the significance of including this sentence, when it is stated that two weeks of data was used? "It could ideally be computed for a period of four weeks before the initial COVID-19 cases were announced, which was 17 March, if the data before March were available."

Response:
The sentence was removed as there is no significance of including it. We wrote in that was as we wanted to use four-week data for computing the baseline to understand normal mobility.

8. Table 1: Indicator 3 Use-case column states "Proxy for population and population movement". This is only a proxy for population count, I believe.

Response:
Yes, it is only about the population count. The sentence was revised accordingly.

9. Page 8; Section C Application in The Gambia – First bullet point: "In our data, we observe no significant fluctuations in total transaction volumes over the data period." What is the significance of this?

Response:
We removed that sentence as it does not make sense. Instead, we explained how we used this indicator for adjusting other indicators as follows; "We compute this indicator at the national level for examining how the number of active subscribers as a whole country changes over time and for adjusting the result of other indicators."

10. Page 8; Section C Application in The Gambia – Fourth bullet point: "we do not use this indicator for generating Origin-Destination (OD) matrices as we were not able to examine how the OD matrix is impacted by missing links between the origin and final destination regions". OD matrices are the most important empirical tool for mapping and understanding the travel patterns in a country, and very important in epidemiological modelling to forecast disease spread. Hence, the reviewer is very puzzled by this statement, and believe that an elaboration is needed to give more details into why the OD matrices have not been used.

Response:
We revised the fourth bullet points and added explanations about the limitation of this indicator as the following: “This indicator can be used for constructing Origin-Destination (OD) matrices but has limitation in capturing long-distance trips. This is because a trip for constructing a OD matrix is defined by each consecutive pair of records, meaning that a long-distance trip is transformed into a set of several short trips, and thereby a link between the origin and destination of the long trip is missed.”

11. Page 8; Section C Application in The Gambia – Last paragraph: Suggestion for general improvement is to highlight and extend findings and the indicators’ specific relevance to COVID-19.

Response:
We revised the descriptions of each indicator and explained the relevance to COVID-19 (for each bullet point). In addition, we added an explanation on the importance of Indicator 3 for adjusting the value of other indicators.

12. Page 8; Section 5 A – First sentence: I believe “population movement” should be “population distribution”.

Response:
Yes. Revised accordingly.

13. Page 8; Section 5 A: Is it IMEIs that is being used? This will count the number of unique handsets, whereas IMSI will count the number of subscribers. Clarification needed.

Response:
Our response overlaps with what we wrote as the responses to Comments 3 and 4. We added the following descriptions as footnote 3: IMEIs are used as a data field for defining the number of subscribers for the analysis while the IMEI generally defines the number of devices and the IMSI defines the number of SIM cards. This is because we combine datasets from two MNOs. In developing countries like The Gambia, it is common that one device is used by a person who has these two MNOs’ SIM cards. To avoid double counting the same person, who exists across two MNO data, the data was de-identified by respective MNOs and the encrypted identifiers were replaced with random numbers on regulator’s premise after the two datasets were combined.

14. Figure 4: There is made reference to a baseline, without defining what this baseline is. Please explain.

Response:

We added the description on the baseline for each indicator as below:
Indicator 3: We use the average of the number of active subscribers for the first two weeks of March as the baseline.
Indicator 6: The average number of residents of the first two weeks of March is used as the baseline.
Indicator 7: We use the average distance traveled for the first two weeks of March as the baseline.
Indicator 10: We use the average of population inflows for the first two weeks of March as the baseline.

15. Section 6 B Technical constraints: The sentence "In addition, we had to employ complex techniques and multiple steps to complete a simple task, which means a single step easily run by an available code was divided into multiple steps with intermediate results. This is because such a simple task requires a lot of time for computation once it starts running, which could be easily interrupted due to an unstable network environment." is hard to comprehend. Please consider rewriting.

Response:
We revised the sentence to the following by elaborating the context.
"Finally, working in a limited capacity context required flexibility, and from times to times involved compromises. In terms of hardware, an initial server provided to PURA for piloting purposes became the go-to for data storage when COVID-19 impeded the acquisition of additional server capacity. In terms of analytics, although there was a tool for producing standardized mobility statistics available on the GitHub repository, we chose to write our own script. This was to accommodate the system parameters on PURA's premises. Since much of the analytical work was done through remote access, it was restricted by network capacity and could be interrupted by electrical outages. This required breaking computationally intensive tasks into multiple steps with intermediate results, so that any interruption would only disrupt the current computation and not the whole script."

16. Section 6 C Policy dialog – towards the end of the section:
oPopulation movement patterns … could inform targeted testing initiatives, …
o … this can also inform where constraints on mobility should be enforced ….
My question is: How was any of this information used by the Government or health authorities in The Gambia? Having this information is not the same as acting upon it. Did the right stakeholder within the Government have access to the information? Were the findings and insights shared with the decision makers putting in place testing policies? How was mobility information used when deciding on the socialdistancing policy implemented on March 18th, 2020 in The Gambia?

Response:
As noted in the revised text: "When the COVID-19 crisis struck, the groundwork was already in place. As results were made available, findings were shared and discussed among members of the government task force for COVID19. They provided an important input into the policy

discussion and prompted requests for a scaled-up version providing real-time data. Unfortunately, efforts to quickly turn this case-study into a fully function data pipeline were delayed due to constraints in implementation capacity until late 2020."
Because of these delays, inputs from the CDR data did not inform the social distancing policies implemented on March 18th 2020 and initial rollout of testing. Rather they contributed to the policy dialogue on the trade-offs in relaxing social distancing restrictions given their severe constraining effects on economic activity.

Best regards,
Ayumi Arai (on behalf of authors)