A proof-of-concept solution for the secure private processing of longitudinal Mobile Network Operator data in support of official statistics

Fabio Ricciato*, Triin Siil**, Riivo Talviste**, Baldur Kubo**, Albrecht Wirthmann*

* Eurostat, Unit A5 — Methodology; Innovation in official statistics, Luxembourg, fabio.ricciato@ec.europa.eu

** Cybernetica, Estonia, baldur.kubo@cyber.ee

Abstract. This paper provides an overview of the joint project conducted by Eurostat and Cybernetica during 2020-2021. The main goal of the project was to demonstrate the feasibility of a Secure Private Computing solution for the privacy-preserving processing of Mobile Network Operator data. The technology of choice for this project was based on Trusted Execution Environment technology with hardware isolation. In this contribution we report on the initial motivations, high level goals, results and lessons learned during the project.

The views expressed in this paper are those of the author and do not necessarily represent the official position of the European Commission.

1 Background

Almost every person nowadays carries one or more mobile device(s) with a subscription to some Mobile Network Operator (MNO). As mobile devices interact with the mobile network infrastructure, they reveal their current approximate position to the MNO. Depending on the configuration of the network infrastructure, such interactions are recorded by the MNO for purposes related to mobile service delivery such as, e.g., billing, optimisation or troubleshooting. Such data are referred here as *MNO data* and include Call Detail Records (CDR) and the more informative *signalling data* when available[1].

[1]Note the difference between MNO data, that are collected *on the network side*, and other sources of location data that are collected *on the device side* by the operating system (OS) and/or by app(s), and then reported remotely to the app manager(s) and/or OS vendor. The latter kind of data are typically based on measurements obtained by on-device sensors (most prominently GPS but also WiFi, Bluetooth, 2G/3G/4G radios, etc.) that can be transformed into location data either directly on the device (e.g. for GPS) or remotely by the app/OS server (e.g. for WiFi

MNO data embed information about the position of mobile devices, and therefore could serve as a base for extrapolating statistics about presence and mobility pattern of human population. Some MNOs are already leveraging such data for the extraction of commercial analytics products related to human mobility and presence patterns, either with in-house resources or by partnering with specialised companies. The European Statistical System (ESS) is considering MNO data as a potential new data source for future official statistics, and several ESS members are already conducting experimental activities and pilot studies in cooperation with MNOs in their respective countries.

Moving from one-off projects or case studies towards regular production of official statistics based on MNO data requires finding solutions to a number of issues. First, there are important methodological challenges stemming from the fact that MNO data are affected by various sources of uncertainty, e.g., limited spatial resolution (device location is not known precisely but only approximately) and incomplete temporal view (device location is not observed continuously but only at a limited set of discrete times), and anyway they are referred to the population of mobile devices that is correlated with, but not perfectly equal to the population of humans (mobile devices do not map one-to-one to humans). Such methodological challenges, though extremely important in the perspective of producing official statistics, fall outside the focus of this paper.

Besides the methodological challenges mentioned above, one must consider the *technical and legal* aspects stemming from the fact that (i) *MNO data are personal data* and (ii) MNO data embed a wealth of *business sensitive information* (e.g., distribution of MNO customer basis, details about network configuration and deployment, etc.). In other words, the confidentiality of MNO data needs to be protected both in terms of individual privacy (of the mobile subscribers) and trade secrets (of the MNO). In this respect, in agreement with the principles of Trusted Smart Statistics [1, 2] Eurostat and the ESS are interested to identify *technical modalities of MNO data access that are able to provide strong data protection guarantees without giving away the possibility to compile accurate and relevant statistics.* Against this background, Eurostat is exploring the feasibility and applicability of so-called Privacy Enhancing Technologies (PET), and specifically Secure Private Computing technologies (SPC) to the processing of MNO data[2].

fingerprinting). The term "Mobile Phone Data" (MPD) has been used by different authors to refer to any kind of location data related to the use of mobile phones, including both network-side and device-side location data. In this sense, MNO data can be considered a subset of the more general class of MPD.

[2]Following [3] we divide the family of PET solutions into the distinct sets of *Input Privacy* and *Output Privacy* solutions. The terms "Secure Private Computing" (SPC) and "Privacy-Preserving Computation" (PPC) refer to technological solutions for the Input Privacy problem — how to compute the exact desired output without having access to the input data and without leaking any other information other than the final desired output (e.g., intermediate data).

2 Project motivations and goals

The focus of the project was on Input Privacy solutions and specifically on Trusted Execution Environment (TEE) with hardware isolation. According to [4] TEE is among the most mature technologies in the field of privacy-preserving computation.

The high-level goal of this project was to verify the feasibility of a TEE for the processing of longitudinal data from a single MNO[3]. More specifically, the project aimed at contributing answering the following questions:

1. Are TEE solutions sufficiently scalable to deal with the data volumes and data rates that are encountered in real-world MNO settings?

2. Are TEE solutions sufficiently mature and flexible for adoption in a statistical production setting?

3. What does the adoption of TEE solutions entail on the legal side?

To address these general questions with respect to a concrete application, Eurostat formulated an *hypothetical reference scenario* whereby one NSI cooperates with one MNO of the same European country. The reference scenario, detailed later in Section 3, involves a number of assumptions about technical, business and legal aspects that are meant to capture the essential elements of potential would-be real-world scenarios. It includes, among other elements, the formulation of a "toy example" of statistical methodology for MNO data processing[4]. A detailed description of the "toy" methodology is available in [6].

The reference scenario and the statistical methodology were provided by Eurostat as input to the project. They represent the main specifications and requirements for Cybernetica to design and implement the prototype of a technical solution based on a commercially available TEE product, namely Sharemind HI[5], that in turn is based on Intel SGX technology[6]. The solution was installed in a laboratory environment with commercial-off-the-shelf hardware. It was then tested with a stream of synthetically generated data emulating the format, volume and rates that may be expected to be encountered in real-world application scenarios. The detailed characteristics of

[3]This use-case is different from, and complementary to, the use-case considered in [5] in the context of combining inbound roaming identifiers across multiple MNOs.

[4]The "toy methodology" was defined exclusively for the purpose of testing and demonstrating the feasibility of the technological solution developed in the project. While it attempts to anticipate possible elements of future "serious" methodologies and statistical definitions tailored for MNO data, it should not be considered representative of official methodologies and definitions. The development of the latter falls outside the scope of this project and remains responsibility of ESS bodies (e.g., Working Groups, Task Forces) that were not involved in this project.

[5]`https://sharemind.cyber.ee/sharemind-hi`

[6]`https://www.intel.com/content/www/us/en/architecture-and-technology/software-guard-extensions.html`

the test data were defined with the main goal of stressing the solution and therefore tend to be worse-than-real.

In addition to the technical feasibility, the project included a work package focusing on the analysis of the legal aspects connected with the prospective adopting this solution in a real-world setting. This strand of work resulted in a legal study, conducted by the legal experts within the Cybernetica staff, and in the model of a close-to-real Data Protection Impact Assessment (DPIA) document for the solution at hand.

3 Overview of reference scenario and requirements

In this section we provide a high-level view of the reference scenario that was defined by Eurostat as input to the project.

It is assumed that, before the deployment of the new solution, a legacy workflow is in place at the premises of the MNO for processing location data for business purposes. The pre-existing data flow is sketched in Fig. 1. The location data are generated by the MNO Network Department (MNO-ND) that is in charge of operating the mobile network infrastructure. The MNO-ND staff has full access to the raw data for (primary) purposes related to network operation, troubleshooting, billing, etc. We assume that a distinct department within the MNO organisation, namely the Value Added Services (MNO-VAD), reuses location data for the (secondary) purpose of producing commercial statistics and business analytic products. It is assumed that, in order to comply with data protection regulations, the location data are pseudonymised before they are passed to MNO-VAD, and that the pseudonymisation keys are periodically changed every period T. Periodic change of pseudonymisation keys reduces both the *risk* and potential *impact* of individual re-identification, and for this reason such a technique is used by several MNOs. In real-world settings the value of the pseudonymisation period T ranges from one day to a few months, depending on the MNO. In our scenario we assume a very conservative value, namely $T = 24$ hours, in order to stress the solution with high computation load.

In the considered reference scenario, the National Statistical Institute (NSI) and the MNO have agreed to cooperate in order to let the NSI produce official statistics based on the location data held by the MNO. To this aim, it is necessary to augment the legacy workflow. The new data flow must fulfill the following main requirements:

1. The legacy workflow for the production of commercial analytics should remain untouched. In other words, the MNO-VAD must be able to use the same algorithms and produce the same analytics as in the legacy workflow.

2. The new workflow must support a statistical methodology based on the processing of longitudinal data traces for individual mobile users over periods

longer than the pseudonymisation cycle T. Processing of longitudinal data must be enabled exclusively for the purpose of producing official statistics and is motivated by the need to determine the individual "usual environment" and the "usual place of residence" based on a long-term view of the visited locations during several months.

3. The new workflow must support the possibility to combine aggregate data derived from MNO location data with confidential NSI data (e.g., high-resolution grid census data). This possibility must be enabled exclusively for the purpose of producing official statistics and is motivated by the need to calibrate MNO data aggregates with official population data at a small spatial resolution.

4. The statistical methodology adopted for production of official statistics is designed to deliver non-personal aggregate data in the output. In our implementation this is achieved by including a Statistical Disclosure Control (SDC) module based on k-anonymity that suppresses elements with value below a threshold k in the final data. In this way the final statistics can be considered (also from a legal point of view) fully anonymised.

5. Confidential NSI data should not be visible to the MNO.

6. Neither the input location data nor any intermediate data other than the very final output statistics shall be visible to the NSI.

The second requirement is particularly critical, since long-term analysis is in direct conflict with the short-term peudonymisation approach that is in place in the legacy workflow. The third requirement is also critical, as it entails the integration of confidential data held by two organisations (namely, NSI and MNO) that are not allowed to access each other data in a joint computation task. The motivation for resorting to Secure Private Computing technologies, and specifically TEE with hardware isolation, stems from the need to address mainly these two requirements.

4 Proposed solution

4.1 Processes

The new workflow is sketched in Fig. 2. The green lines denote cryptographically protected communication channels. At the heart of the TEE technology is the so-called "secure enclave", i.e., a special portion of the processor where only cryptographically signed code can be executed. The process memory as well as any intermediate computation results produced during code execution are cryptographically protected and cannot be accessed by other hardware or software components, unless explicit instructions for their release are included in the signed code itself.

In a nutshell, the secure enclave performs two distinct processes, called "PT" (for pseudonymisation task) and "AT" (for analysis task) as described below. In the first PT process the secure enclave generates the short-term pseudonymisation keys and sends them to the MNO-ND system in charge of performing the periodic pseudonymisation of the location data. The short-term keys sent to MNO-ND are based on a *long-term master key that is generated internally to the enclave and never exported outside*. That means that only the secure enclave is able to reverse the short-term pseudonyms. From the perspective of the external world (outside the secure enclave) the short-term pseudonymisation keys appear to be produced randomly.

In the second AT process, the secure enclave reads the short-term pseudonymised data from the MNO-VAD database, reverses the short-term pseudonymisation (it can do so because it knows the long-term key), ties together consecutive short-term data chunks for each individual mobile subscriber, performs the analysis at individual level (i.e., determines the set of "usual places" for each mobile user), aggregates the results across all mobile users, combines them with the confidential NSI data (which were imported from the NSI in encrypted form) and finally applies SDC checks before exporting the final (non-personal) results to the NSI.

4.2 Methodology

The high-level sketch of the toy statistical methodology defined (by Eurostat) for the projet is shown in Fig. 3. The first block in the processing pipeline, i.e., Module A, runs outside the secure enclave and operates on short-term pseudonymised data. It takes in input the individual MNO data records and produces a summary of the main locations visited by each mobile user within a single pseudonymisation period. Each location is associated to a vector of scores indicating the frequency and total duration of the visits within each period. If the goal is to build statistics based on the "usual environment" of individuals, transit locations (e.g., intermediate location within a trip) can be discarded already at this stage. The following modules B, C and D run within the secure enclave. Module B integrates together the short-term data chunks for every individual user over a long-term observation interval. To do so, Module B must reverse the short-term pseudonyms of each data chunk. For each mobile subscriber, Module C takes in input the long-term summary of visited locations produced by Module B and selects the top visited locations according to some criteria. In this way, the chain of Modules B and C identify the long-term ´´usual environment" of each individual mobile user. Finally, Module D aggregates the individual summaries into aggregate statistics. In so doing, Module D can optionally calibrate the aggregate values based on confidential data provided by the NSI (e.g., detailed census grids at low resolution). Module D includes also SDC functions in the form of k-anonymity filters.

4.3 Actors and roles

The presentation above has introduced the three main actors involved in the data flow, namely the MNO-ND, MNO-VAD and NSI. Each of these actors provides some input data and/or receives output data, therefore they appear explicitly in the data flow sketched earlier in Fig. 2. Besides these three actors, other two actors participate to the process without providing or receiving any data, as explained below.

The TEE technology adopted for this project implements the set of pre-defined "roles" listed in the rows of the Table shown in Fig. 4. Each role defines a set of access rights and powers. Each actor (stakeholder) can be assigned one or multiple roles, and the same role can be assigned to multiple actors. The table in Fig. 4 summarizes the assignment of roles to the different actors (or stakeholders) in the form of a matrix.

The roles of Input Provider and Output Consumer are taken by the three main actors seen above, depending by the specific process task (PT or AT). The "Enforcer" and "Auditor" roles subsume, respectively, the ex-ante (before process execution) and ex-post (after process execution) controls over the whole operation. Ex-ante controls include e.g. the responsibility to approve and sign the code to be run by the secure enclave and to authorize its actual execution. Ex-post controls include the possibility to access and audit the system logs in order to verify correct operation.

The Enforcer role is particularly important: *only code that is approved and signed by all actors endowed with Enforcer role can be executed in the secure enclave*. By way of assigning the Enforcer role to multiple actors the solution implements the principle of *control sharing* (as opposite to control delegation) that is a pillar of the Trusted Smart Statistics concept [1, 2]. In this way, each actor endowed with Enforcer role maintains direct control over the code that is executed by the secure enclave and can verify *before execution* that it complies with the requirements.

In our case the Enforcer role is assigned to all three actors introduced above, namely MNO-ND, MNO-VAD and NSI, plus a fourth independent entity called "External Auditor" (EA). Furthermore, MNO-VAD, NSI and EA are also assigned the Auditor role[7]. In practical settings, the EA may represent a separate controlling department within the MNO organisation or a completely independent organisation.

Having assigned the Enforcer role, the NSI can verify directly that the code implements precisely the intended statistical methodology, while the MNO can check directly that the processing logic reveals no business sensitive information. The EA can verify that the reported output qualifies as non-personal data, or in other words that the SDC filters and the corresponding parameter setting (threshold k in the $k-$anonymity filters) are sufficient to prevent the reporting of small value data

[7]The term "Auditor" refers here to a logical role. It should be distinguished from, and not confused with, the term "External Auditor" that represents an actor (not a role). In our solution the External Auditor is assigned both Enforcer and Auditor roles.

elements. Each enforcer can verify that the code includes explicit instructions for deleting intermediate data and other sensitive data (e.g. keys) and does not contain hidden reporting primitives.

Having assigned the Auditor role, the NSI, MNO and EA can independently access and inspect the execution logs. It is important to foresee, already in the solution design stage, the possibility to produce "verbose" log files rich of details and meta-data about the processing operation (e.g., details about the size of input and output data). Richer and more informative logs increase the probability to detect and trace back incidental processing errors, deliberate mis-use attempts and attacks, and in this way translates into higher deterrence, robustness and security levels.

Configuring which roles are taken by each actor, or equivalently designing the role-to-actor mapping, is a key aspect of the solution design. We demand from the technology that the role-to-actor mapping is properly *implemented* by the software and hardware levels, but its *design* must be accomplished at the "humanware" level by taking into considerations the mandate and interests of each involved organisation (actor, stakeholder) and their mutual business and legal relationships.

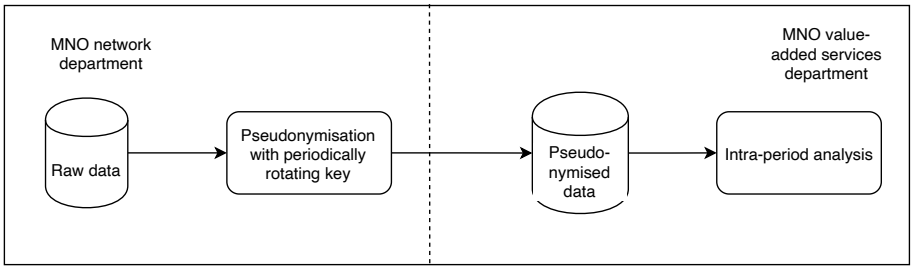

Figure 1: Legacy workflow, before deployment of the proposed solution.

5 Results and key learning points

In this section we summarize the main project results and the key learning points. Further details can be found in the full technical reports[8].

5.1 Scalability

The data processing method corresponding to modules B-C-D in Fig. 3 was first translated into python code and then implemented in a software code suited to run in the secure enclave. It was then fed with a stream of synthetic data chunks emulating the output of module A (operating on short-term pseudonymised data) for a population of 100 million mobile subscribers over a period of 3 months. Such an

[8]The technical reports are publicly available from `https://ec.europa.eu/eurostat/cros/content/eurostat-cybernetica-project_en`

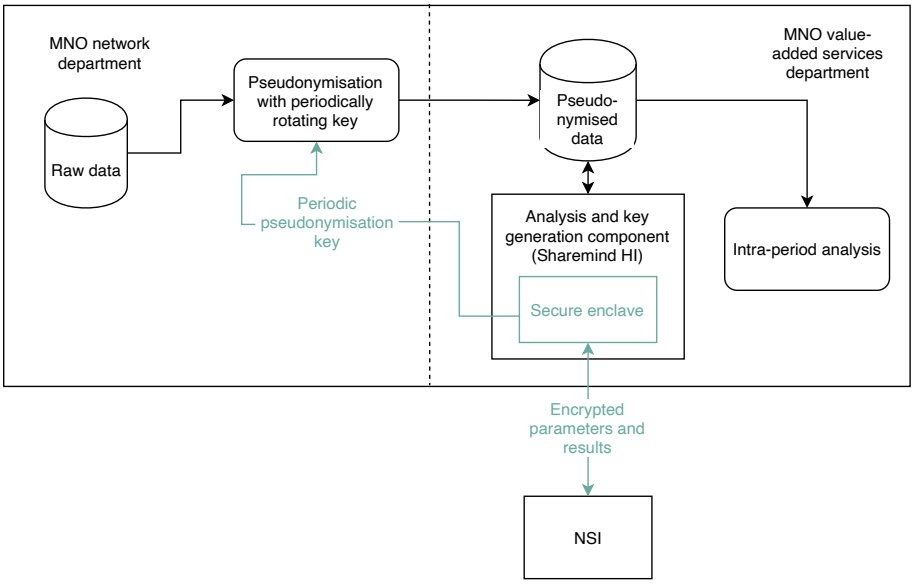

Figure 2: New workflow with the proposed solution.

amount of data was successfully processed by the solution running on a commercial-off-the-shelf machine in few hours, indicating that quasi real-time analysis is feasible in the considered scenario. In principle, the computation time could be further reduced with the implementation of certain software optimisations that were identified but not implemented during the course of the project.

This result did not appear to be obvious at the project inception, since the very limited amount of memory on board the hardware enclave could have represented a bottleneck for the processing task. In fact, the processing logic involves one pseudonym decryption operation for every input data chunk representing the summary of visited locations from a single mobile subscriber during the period T. When the internal memory does not suffice (e.g., to store large intermediate data) the enclave can make use of external memory to store data in encrypted format. In other words, the secure I/O channel between the enclave and the external memory embeds hardware accelerated encryption/decryption functions, and its capacity was more than sufficient to cope with the processing load in our test scenario.

5.2 Security

All hardware and software security technologies are exposed to the discovery of new vulnerabilities during their lifetime, and the TEE technology used in the project was no exception. Face to the inflow of newly discovered vulnerabilities, the level of security depends also from the ability of the technology provider to recognise and readily provide fixes, patches and information to the technology adopters. During the course of the project there were 2 patches released by Intel SGX to address recently

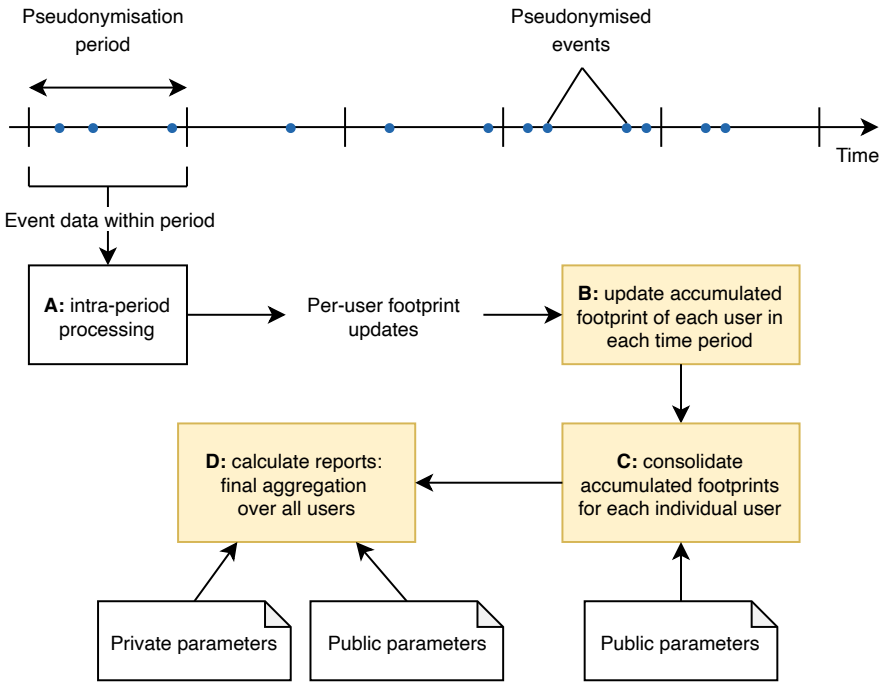

Figure 3: Sketch of the "toy methodology" used for the project. The modules B, C and D (marked in yellow) are executed within the secure enclave. Module A is executed externally to the enclave as it operates on the same short-term pseudonymised data from the legacy workflow.

discovered vulnerabilities, and in both cases the patches were readily communicated by the provider. Like with all other security technologies, a continuous support by the technology provider is an essential element to achieve a high level of security and maturity of the technology.

5.3 Configurability

The biggest part of the work in the project was aimed at defining the roles of the various entities and procedures to enforce these roles. This involves the specification of "who sees what", "who controls what", what security risks are in place and how to address them from the technical and organisational level.

The choice to rely on a relatively mature technology that is already available for production use, for which general-purpose functions and primitives were already implemented, allowed the project team to stay focused on the design of the project-specific elements of the solution, such as the in-depth analysis of the business processes in the specific application context. In this way, the limited project resources would be focused and used more effectively.

It is important to ensure that the adopted technological solution is configurable

		Stakeholders				
		MNO-ND	NSI	MNO-VAD	External Auditor	Intel via Cybernetica proxy
Roles	Sharemind HI server Host			+		
	Coordinator		+			
	Enforcer	+	+	+	+	
	Input Provider	PT[21]	AT	AT[22]		
	Output Consumer	PT	AT	PT[21] AT[23]		
	Runner	PT		AT		
	Developer		+			
	Auditor		+	+	+	
	Attestation Service Provider					+

Figure 4: Assignment of roles to actors (stakeholders.) The entries labelled as PT and AT identify role assignments that are specific for each process, while the marker "+" indicates assignments that apply to both processes.

and flexible. First, it should provide the possibility to define roles, introduce actors (stakeholders), assign roles to actors and change them when needed without excessive additional costs. In this way the solution can be effectively tailored to the needs of specific use-cases and adapted to changes of scenarios. Second, it should allow to make changes to the statistical methodologies at an acceptable marginal cost. In a real-world production environment, the statistical methodology may need to be adapted, for example in response to changes in the input data flow or to reflect methodological improvements. The cost of implementing methodological improvements must be considered in the total cost equation of the solution.

5.4 Legal aspects

The legal study conducted within the project [7] showed that the legal aspects related to the (re)use of MNO data by NSI are rather involved due to the interplay between European and national norms from three different legal domains — namely statistical legislation, telecom legislation and data protection legislation. In the articulation between norms produced at different times there are gaps and points exposed to different interpretations. Furthermore, there are national differences — mainly in the statistical legislation, but to a lesser extent also in the national transpositions

of e-Privacy Directive — that concur to make the situation heterogeneous across different EU countries. For this reason, it was not possible to provide a conclusive analysis and a complete DPIA that is valid for all EU countries. However, the (partial) DPIA reference model developed within the project [8] covers the main technology-specific aspects that are independent from legislation and provides a solid starting point for the development on complete DPIA in future follow-up activities. For further details the interested reader is referred to the full project documents [7] and [8].

The study has highlighted the need to strengthen at the European level the legal basis by which statistical offices can reuse MNO data. GDPR allows the use of personal data for "statistical purposes" subject to implementing "adequate technical and organisation measures" to safeguard the data. The solution developed in this project can serve as a safeguard measure, and in this way help to comply with GDPR. However there is still a need to define officially the statistical results that require the reuse of MNO data by NSI. This aspect could be addressed in the context of future revision of legislation on European statistics. Furthermore, there is a need to clarify the legal basis that enables MNOs to let NSIs reuse mobile location data for official statistics.

6 Outlook and follow-up

Solutions based on Secure Private Computing technologies, like the one developed as proof-of-concept in this project, could help to secure the processing operation and reassure the various stakeholders involved in the reuse of MNO data for official statistics, including but not limited to MNO and NSI. Coupled with procedures aimed at ensuring openness and transparency of the methods and purposes of the processing, these technologies would be key to ensure public acceptance and trust [1, 2].

The project results show that solutions based on TEE with hardware isolation are sufficiently scalable and capable of crunching large rates of input data, representing a large population of mobile subscribers with commercial-off-the-shelf hardware. The proof-of-concept solution developed in this project will be made available by Eurostat to NSI/MNO partnerships interested to perform in-field testing and pilot study projects based on real-world data. To this aim, the reference DPIA model drafted within the project will facilitate and speed-up the preparation of the complete DPIA that will be likely needed for conducting in-field testing with real data. Interested NSI/MNO consortia are invited to contact Eurostat for follow-up.

All project deliverables and technical reports will be publicly available from the CROS portal (the exact URL will be given in the final version of this paper).

References

[1] F. Ricciato, A. Wirthmann, and M. Hahn. Trusted smart statistics: how new data will change official statistics. *Data and Policy*, 2, 2020. https://www.cambridge.org/core/journals/data-and-policy/volume/3555B26C32D5A8774BD517B4052CAAD8.

[2] F. Ricciato, A. Wirthmann, K. Giannakouris, F. Reis, and M. Skaliotis. Trusted smart statistics: Motivations and principles. *Statistical Journal of the IAOS*, 35(4), 2019.

[3] F. Ricciato, A. Bujnowska, A. Wirthmann, and M. Hahn. A reflection on privacy and data confidentiality in official statistics. 62nd ISI World Statistics Conference, Kuala Lumpur, https://ec.europa.eu/eurostat/cros/system/files/isi_paper_ricciato_bujnowska_final.pdf, August 2019.

[4] Big Data UN Global Working Group. Un handbook on privacy-preserving computation techniques. https://tinyurl.com/y3rg5azm, 2019.

[5] Mobile phone data meets sharemind hi: tourism statistics innovation in indonesia. https://sharemind.cyber.ee/mobile-phone-data-meets-sharemind-hi/, 2019.

[6] Specification of use-cases. Technical Note for project ESTAT 2019.0232. Available from https://europa.eu/!RDkywK, 2021.

[7] Data Protection Impact Assessment – Scoping Report. Deliverable of project ESTAT 2019.0232. Available from https://europa.eu/!RDkywK, 2021.

[8] DPIA Evaluation Report. Deliverable of project ESTAT 2019.0232. Available from https://ec.europa.eu/eurostat/cros/content/eurostat-cybernetica-project_en, 2021.