Working with a mobile network operator (MNO) to create a privacy-conform method for a better access to MNO data

Lorenz Ade[1], Maurice Brandt[2]

[1,2]Federal Statistical Office of Germany (Destatis), Germany

Abstract

Mobile network operator (MNO) data has huge potential for official statistics. Commercially available data is currently created and processed in a kind of black box, as the aggregation and extrapolation of the data is a business secret. MNOs currently can't share their processing steps, as they employ confidential information like cell tower positions and local market shares for their algorithms. For the usage of MNO data in official statistics, this black box has to be opened if the quality criteria of transparency and comparability are to be achieved.

To solve this problem, DESTATIS has partnered with T-Systems (a subsidiary of the MNO "Deutsche Telekom"). In this collaboration, T-Systems provides DESTATIS, for the first time at all, with access to the secure environments of an MNO to work with anonymized raw signal data. The goal of the collaboration is to create a standardized, transparent and privacy-conform method for an access to MNO data. This work is a part of a large German research cluster on the anonymization of georeferenced data (AnigeD).

The collaboration between DESTATIS as National Statistical Institute (NSI) and T-Systems as MNO guarantees that the necessary expertise in relation to statistics, data protection and mobile network is present in the project. Additionally, the collaboration ensures that the business interests of the MNO are maintained, as they are free to offer their data commercially based on their own algorithms.

Key focus besides achieving the necessary quality criteria for official statistics is the compliance with privacy standards. This is not alone a legal necessity. Ensuring data privacy is fundamental to maintain the trust of the public in official statistics. To achieve this goal when handling the very sensitive MNO data, privacy by design has to be included in the processing of the data from the very beginning. Additionally, data protection officers and institutes are consulted regularly.

Keywords: mobile network data; privately held data; private sector collaborations; data privacy

1 Introduction

Mobile network operator (MNO) data is being analysed worldwide as a potential source for official statistics, as it can be used to analyze a wide range of topics. For example, MNO data could be used to improve statistics on population, mobility or tourism. Additionally, MNO data could lead to the creation of completely new statistics, such as day-time vs. night-time population figures. A major advantage of MNO data is the high level of coverage. For example, 98.1% of households in Germany own a mobile phone. On average, there are 1.87 devices per household (Statistisches Bundesamt, 2022). As a data source it is therefore of great utility for official statistics. But why are there no official statistics based on MNO data yet? MNO data is not primarily produced for official statistics. Instead, it is basically a by-product of the operation of a mobile network. Here the data is mostly used for technical analysis and troubleshooting. Therefore, the data needs to be heavily processed before it can be used in official statistics. These processing steps include the geolocation, deduplication, aggregation and extrapolation of the data. Some MNO's (not only in Germany) offer the usage of anonymized and aggregated data products. In the past, DESTATIS has already produced several promising experimental statistics based on these products, for instance on mobility during the Covid-19 pandemic or population forward projection (see our website at https://www.destatis.de/EN/Service/EXSTAT/Datensaetze/mobility-indicators-mobilephone.html). The methods with which these products are created are however confidential, as they employ information like cell tower positions or local market shares, which the MNOs naturally don't want to share with their competitors. This is however a problem for DESTATIS as National Statistical Institute (NSI), as it has the duty to be transparent about the methods and processing steps it uses in the production of official statistics. To open the black box of the MNO processing, NSI's therefore have to develop their own methods. As the development requires access to the original data, DESTATIS has partnered with T-Systems to gain access to raw MNO data for the first time. Besides providing access to (strongly distorted) data, T-Systems also provides a development environment in their IT-infrastructure and their technical knowledge.

A large part of the data processing to be designed is also ensuring the compliance with data privacy at all steps of the production process. While signalling event data is not as privacy invasive as call details records, the unprotected data would still allow for the tracking of individual movements of cellular device holders. Although the raw data is pseudonymized from the beginning and never to be released, an attacker might be able to identify individuals if insufficient aggregation procedures are employed. To prohibit such reidentification attacks, the

second large focus of this research project is therefore the identification of the anonymization needs and the evaluation of suitable methods to prevent reidentification.

2 Technical aspects

Before the goals and the structure of the project can be detailed, it is important to explain which MNO data exactly is being analysed and which implications this has both for the methodical processing and the anonymization of sensitive information.

MNO data are not the GPS based movement data collected by the device manufacturers like Apple and Google. They are, as already a mentioned, more or less a by-product of the operation of a mobile network. MNO data is constantly generated when cellular devices (anything with a SIM card) interact with the mobile network. There are two different types of data being generated. Signaling event data (SED) and call details records (CDR). SED is generated when a cellular device communicates with the nearest network cells with the purpose of establishing the best connection with the network. This communication is permanently ongoing, independent of the actual usage of the device. CDR are records of the actual usage of the device. In the past, this was mainly information about incoming and outgoing calls. Nowadays, they probably consist mostly of data connections.

Most projects of MNO data usage for official statistics concentrate on using SED, as it is independent of device usage and less privacy sensitive. SED is basically a list of connections with the encountered network cells and a timestamp. To create useful statistics, the data has to be geolocated, which requires information about the coverage of the connected cells. The coverage of the cells can be determined by different methods, most of them depending on confidential information about antenna locations of the MNO's. Alternatively, crowdsourced datasets on cell location also exist. While they might not be as accurate and complete as data provided by an MNO, they are well suited for the purpose of this project. They are easy to use and don't require a complicate approval procedure. As location estimation is not the research focus of this project, the saved time can be invested elsewhere.

3 Project Goals

Essentially, this project aims to transfer the usage of MNO data from experimental statistics to official statistics by opening the black box of mobile network signalling processing. This requires designing and implementing a processing and anonymization procedure for the usage of anonymized georeferenced MNO data. To design this process, the conception and set-up of a development environment at the data provider is necessary. Lastly, the project also serves

as an attempt to set up a model process for future cooperation's between private data providers, especially MNO's.

Such a cooperation, as it exists with T-Systems as MNO in this project, is vital to reach the planned project goals. Not only is the project dependent on the provided IT-Infrastructure and data but it also benefits heavily of the combination of the technical expertise of T-Systems with the methodical expertise of DESTATIS as NSI. A cooperation with a commercial data provider is not common for DESTATIS and comes with its own challenges. At the heart of the problems is the duty of DESTATIS as NSI to employ and document the used methods transparently, whereas all MNO's want keep their algorithms and additional information used, e.g. local market shares, confidential.

To achieve the overall goals of the project, four work packages have been designed. The first work package is designed to develop the necessary IT-Infrastructure for the project. This step is necessary due to the natural differences in IT-infrastructure between DESTATIS and Telekom as MNO. To test the standardized processes developed in this project therefore requires a development environment that is comparable with the T-Systems infrastructure. Additionally, this also makes gaining approval from data privacy regulators easier, as the raw data doesn't leave the T-Systems infrastructure.

The aim of work package two is to define and implement the necessary interfaces for the transfer of the computed aggregates to DESTATIS. In addition to the technical and content-related results, it is also necessary to comply with IT and data protection regulations

Work package number three is leaning heavily on the work conducted in the ESSnet Big Data 2 and the work currently conducted in the ESSnets Multi-MNO and MNO-MINDS. Its goal is to methodically develop standardized algorithms for processing and preparation that are as flexible as possible for the daily processing of mobile network signal data, particularly for the purposes of official statistics. The methodological solutions that need to be implemented include, for example, the ability to correct distortions, the calculation of location probabilities or the recognition of movement modes. Additionally, data users in the future should be able to control and adjust their data requirements by using standardized parameters, without a need to access the raw data.

The fourth and final work package aims to develop an improved, more flexible anonymization procedure for mobile network signal data in the context of the applicable data protection regulations. This requires a sensible trade-off between flexibility and precision of the data and the protection-compliant processing of the data. To identify the needs in regard to anonymization, dependent on the required scope of the data, the project is part of a German

research cluster on the anonymization of georeferenced data called AnigeD, which itself is part of a large German research network on the anonymization of data.

4 Project Status and Next Steps

The first big step of the project has already been taken. Both the corporate security of our partner T-Systems and the BfDI (Federal Commissioner for Data protection and Freedom of Information) have accepted our data protection plan. This is already a huge hurdle taken, as a failure to coordinate a sufficient data protection plan could have delayed the project timeline severely.

The data protection plan includes additional measures to prohibit a reidentification of individuals besides the original pseudonymization of the data. These measures include both a distortion in time and space as well as a minimum event size. The distortion in space consists of moving the signal events to a neighboring network cell in a circle with a radius of 1 kilometer. If less than 10 network cells are in this area, the radius of the area is increased until 10 network cells are included. The distortion in time is realized by moving the individual events up to 15 minutes in the past or the future while keeping the order of signal events intact. Additionally, the year of data creation is unknown. To protect information in sparsely populated areas, network cells with less than five devices per hour are not part of the data.

With these measures coordinated, the next step is to set up the development environment and the secure access for DESTATIS in the T-Systems infrastructure.

5 Conclusion

With this project, DESTATIS moves closer to transferring MNO data from the realm of experimental statistics to an incredibly valuable data source for official statistics. To achieve this goal, a standardized processing and anonymization procedure for MNO data is developed. The project benefits immensely from both its cooperation with the MNO T-Systems and the German Research Cluster AnigeD. The cooperation with T-Systems as an MNO allows DESTATIS not only to access (distorted) raw MNO data for the first time, but also to benefit from their technical expertise. AnigeD and the affiliated research network give DESTATIS access to state-of-the-art knowledge on the anonymization of geodata.

6 References

Statistisches Bundesamt. (2022). *Laufende Wirtschaftsrechnungen (LWR), Ausstattung privater Haushalte mit Informationstechnik*. Von https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Ausstattung-Gebrauchsgueter/Tabellen/liste-infotechnik-d.html#115470 abgerufen