NSM 2022
NORDIC STATISTICAL MEETING

Mobil phone position data and official statistics

Pieter Vlag, Ulf Durnell, Jens Malmros, SCB, pieter.vlag@scb.se

Abstract

Anonymised and aggregated mobile phone position data from mobile network operators can be a very suitable data-source for producing official statistics about dynamic populations, commuting, tourism and long-distance travelling. Having this in mind, Statistics Sweden (SCB) and one of the biggest telecom operators in the Nordics (Telia) started a collaboration in 2020, one month before the outbreak of COVID-19. The aim of the project was to analyse the quality of aggregated mobile phone data. Public interest and interest of authorities in these aggregated data considerably grew after the outbreak of the pandemic because these data clearly showed changes in population movements before and after the pandemic. However, it appeared that creating stable time-series out of mobile phone position data is not straightforward in periods of sudden change, hampering an exact quantification of differences in population movements before and after the pandemic. Transparent and well-described processes were needed to tackle this issue. In this context, the collaboration gradually changed into a partnership in which the operator delivers 1) process descriptions and 2) aggregates while SCB develops 1) quality reports and 2) benchmark tables. Benchmark tables compare differences between dynamic population estimates based on mobile phone position data with register-based static populations from SCB and related register-based statistical data from SCB.

The results of this partnership will be discussed in the presentation. The results of a still on-going phase of the partnership will also be presented: development of statistics out of the partnership. Other aspects which will be discussed during the presentation are: involvement of other mobile network operators in this process and standardisation of the use of MNO-data.

Keywords: mobile phone position data, time-series, benchmarking, partnerships

1. Introduction

Four mobile phone network operators (MNOs) are active in Sweden. The biggest one (Telia), which is also active in Denmark, Finland, Norway and the Baltic States, commercially exploits its aggregated and anonymised mobile phone position data (MNO-data) by so-called Telia Crowd Insights platform. Telia Crowd Insights draw general attention from governmental agencies and media in Sweden during the COVID-19 pandemic as these quickly available MNO-data showed plausible changes in population movements after the outbreak of the pandemic.

Statistics Sweden (SCB) and telecom operator Telia collaborate in processing and analysing aggregated mobile phone signalling data since 2020. The first stage of the project was explorative. Its main conclusion was that these aggregated and anonymised MNO-data are a very suitable data-source for producing official statistics about population movements. These statistics may enrich current commuting, tourism and travel statistics and publish smart statistics about dynamic populations: how may people are present at a certain time-point at a certain place. However,

- the exact amplitude of changes in population movements were difficult to quantify
- the transparency of the processes could be improved
- better monitoring of the quality was needed.

The second stage of the collaboration (February 2021 – September 2021) was focused to improve quality and transparency of the processes. It resulted in a detailed quality report for internal use, a more general quality report for external use and improvements for methodology.

The third phase of the project (from September 2021) is part of a government assignment for SCB. The government assignment consists of three tasks which need to be investigated together with data-owners (MNO) and relevant governmental agencies:

- conditions and quality criteria for the use of MNO-data (and new data sources in general) for official statistics
- development of new ‘smart’ statistics
- proposals for reducing statistical surveys with help of MNO-data.

The collaboration with Telia was extended for the government assignment. Another MNO in Sweden (Tre) started to collaborate with SCB on the same condition: data-

deliveries in exchange for knowledge sharing and a limited handling fee which covers the costs for the operator to deliver data. Together these MNO share roughly 50 % of the market, with Telia having a market share of close to 40 % and Tre having a market share slightly over 10 %. These market shares are approximately constant over time at national level since 2019. Regional differences in market shares do exist, partly because not all MNOs in Sweden have a network in the remote parts of Sweden.

2. Legal issues and business model

The ‘open government principle’ and ‘equal treatment principle’ are key values of all Swedish governmental agencies and guaranteed by law. To prevent any discussions about the ‘principle of equal treatment’, SCB has actively contacted all other MNOs and offered them the same opportunity for collaboration as Telia and Tre. Furthermore, other MNOs can make notice of the results of the collaborations between SCB and Telia by reading the public reports. Another consequence of the ‘equal treatment principle’ is that SCB cannot develop a processing system for MNO-data, which is suitable for one operator only.

The ’open governmental principle’ implies that all shared documentation between SCB and MNOs is public. Shared documentation between MNOs and SCB may, however, reveal confidential business processes of an individual MNO. On the other hand, the statistical law prevents and forbids revealing information about individuals and individual enterprises. The MNOs wanted to share documentation about their processes with SCB, but requested that their process descriptions should be treated as confidential. This request appeared to be more complex than expected resulting in discussions with legal experts concerning which conditions the delivered information about MNOs data-collection processes are covered by the statistical law. Consequently, SCB cannot reveal all known information about MNOs processes regarding MNO-data. Another consequence of receiving process descriptions only is that data at low aggregation level are needed for quality assurance and check the actuality of the descriptions.

Data-privacy and GDPR (General Data Protection Regulation) are important issues when dealing MNO-data. All MNOs in Sweden have high standards regarding data-privacy. They interpret GDPR rigidly to prevent image-damage at all costs. MNOs are

therefore only prepared to deliver aggregated data to SCB (together with providing information about their processes). In this context the relationship between SCB and the two participating MNOs is a kind of a partnership in which MNOs deliver aggregated and anonymised data and SCBs role is mainly quality assurance by

- asking information about processes,
- checking plausibility, by receiving MNO-data at low aggregation level and comparing them with other statistical information.
- benchmarking the anonymised and aggregated data-deliveries with other statistical information
- develop ‘smart statistics’ out of the data which serve general interests and are complementary to the commercial exploitation of these data by the MNOs.

3. Data deliveries

SCB receives aggregated data from Telia about activities and trips to carry out the government assignment. The definition of an activity is: A stationary signal measured at least 40 minutes within an hour in the same grid cell. An activity can be related to (being at) home, (being at) work etc. It may last from 40 minutes up to one day. By choosing the 40 minutes threshold the activity is unique for the hour. It can therefore be interpreted as the main geolocation of a device during an hour. For the same reason, it can with help of a weight model related to human populations in districts, municipalities and regions during night-time (in this study defined as between 2 – 5 am) and daytime (in this study being defined between 9 am and 3 pm).

The definition of a trip is a directional movement between two areas. This is the case when the stationary signal is less than 40 minutes within an hour. As a rule of thumb, it can be said that in urban areas trips can be identified if the user moves 300-500 meters, while in rural settings that number is above 1000 meter. Over water or in areas with almost no residents a movement might not be detected in the network for even longer distances, due to the size of the grid cell. Trips can be related to travels.

SCB receives the following date from Telia:

- activities at (geographical) levels: grid, district (DeSO), municipality and region.
- trips between the (geographical) levels: grid, district (DeSO), municipality and region.

Temporal resolution of the data-deliveries: one hour. Data-deliveries are monthly. Time-series from January 2019 up to last month (April 2022 at time of writing).

Data-deliveries from Tre are: activities at municipality level, temporal resolution one hour series from 2021/1.

Data are analysed at municipality and regional level for development of smart statistics. Data-deliveries at lower levels (grid and DesO) are used for quality assurance, detecting breaks and artificial signals in time-series and correcting for them

4. Processing

4.1 Source data

Each day millions of MNOs subscribers generate billions of data points when using their phones. This data is captured, processed and put in a data lake to be utilized in activities for e.g., business purposes such as improving network performance of the MNO. The data generated is coming both from active usage but also from passive use. Examples of active usage is sending a text message or receiving data. Examples on probe data could be if a phone switches connection from one antenna to another or when the cell checks if the phone is still present in the antenna.

In the first stage of the processing at MNOs these raw cellular network data and infrastructure data are combined in so-called antenna signals. In short, antenna signals contain hundreds of millions of rows from all the devices, or SIM-cards, in the network, where one row represents a connection, event or signal from the device. Antenna signals consist of three main components.

- An anonymised key- identifier
- Timestamp (at what point the SIM-card left a trace in the network)
- Cell Id (at which cell tower, or Base Transceiver Station, did the SIM-card leave a trace at the given timestamp)

The anonymisation (=creating the anonymised identifier) is driven by GDPR and carried out at regular time-intervals (nowadays 24 hours at Telia to be 7 days from autumn 2022, 30 days at Tre). Consequently, a device can be followed 24 hours only. Due to GDPR, MNOs in Sweden do not use client data in their process.

3.2 *Processing the source data*

Processing, which is carried out at Telia and Tre, of the source data can be divided into three main steps.

1. Processing of the data: (anonymisation, imputing 'missing' time slots, filtering for disturbances in the network and data-deliveries etc.)
2. Geolocation: relating signals of the receiving antennas to a geographical grid.
3. Weighting the sample of devices to human populations

Processing the source data at Telia has been described during the second phase of the collaboration between Telia and SCB by structuring existing documentation into a standard reporting standard for official statistics: SIMS (Single Integrated Metadata Structure). This SIMS report contains confidential information about Telias processes and is not public, because it reveals information about Telias business processes. A public report about methodology has, however, been described by Vlag, 2021. The SIMS-report can be considered as a win-win situation for Telia as SCB; Telia has structured information about their processes related to quality. SCB has the required minimum information about the processing of MNO-data. Information about data-processing is needed for transparency when considering publishing statistics out of MNO-data.

Tre, the other operator SCB collaborates with, has a similar process for step 1 and 2. Documentation of Tres processes needs, however, some improvement. This is an ongoing activity.

3.3 Data-processing and geolocation: SCBs contribution in this collaboration

SCBs contribution in this collaboration has been facilitating the documentation and participate in methodological discussion regarding the first two steps: 1) data-processing of the source data and 2) geolocation. The reasoning behind this position is that SCB does not have superior methodological and IT-expertise about these parts to improve the existing processing systems of the MNOs considerably at short notice. However, SCBs contribution in the collaboration has been active regarding the weighting of the MNO-data and benchmarking the results. This is due to

- the complexity of relating mobile devices to human population
- the observation after the first stage of the collaboration that SCB has much more experience and knowledge in relating observations to populations than MNOs

3.4 Challenges in weighting MNO-data

According to data from the Swedish Post and Telecom authority (PTS), there are 1.4 SIMS-cards per capita in Sweden. This finding implies that more SIM-cards than humans do exist in Sweden. Therefore, weighting, relating number of SIM-cards to human populations , is needed even if data from all operators are available. The sample in weighting are devices with a SIM-card registered in Sweden, regardless the client type or type of devices.

The fundamentals of the weight model used by Telia (and what SCB applies now on Tre data) is relatively simple.

- The first activity of a 24-hours period is considered as 'HOME' (some exceptions do exist).
- Σ (HOME) is corrected for stationary devices in a municipality: Σ (HOME*)
- The weight factor *w* per municipality is calculated by:

  w = *N* / *n* with

  N = population according to SCBs population statistics in a municipality

  n = [Σ(HOME*)] or the so-called baseline for a reference period. A reference period is a period outside the holidays seasons for which it can be assumed that people are at their registered address (=home) during the night.

Weight factors are applied to the device for a 24-hour period, regardless of the device moves to other locations during the day. The factor depends on the location (municipality) of the first signal of the device during the day. The same weight factor also applies during movements from one to another location. The weight procedure is repeated every 24-hours.

The main assumption behind this approach is that during weekdays outside holiday periods most persons and their belonging devices spend the night at their registration (home) addresses and the baseline (=corrected sum of devices) can therefore be related to population statistics.

Critical factors in the calculation of the weight factor are:

- stability of the so-called baseline*, Σ (HOME*)*
- the geographical aggregation level for the weight calculations

- updating frequency of updating the weight factors. Ideally, the weight factor
  - should not be recalculated daily, such that real variations in spending nights at a specific location (for example during holidays and weekends) may be captured
  - but should be regularly updated, to keep the weight model up-to-date due to changes in antenna coverages, and changes in client share.

In 2020 it was observed that an unstable baseline of the weighting is the main cause of instability in time-series. This instability was visible by 1) unexplained drops in populations with long recovery times, 2) drift in estimates over time, 3) lacking consistency across products (daily vs. hourly) and geographical levels (regional level vs. municipality level). After investigations, it appeared that the baseline appeared to be unstable due to

- artificial causes such as
  - changes in the network, leading in the geolocation process step to changes in grid assignment and consequently to subtle changes in municipality assignment.
  - temporal drops in supply of data due to network failure
  - slight changes in client share of the MNO at low district or municipality level.
- real changes such as
  - differences in human night populations during weekdays, weekends and due to travelling in holiday periods. In this context an important conclusion was that the MNO-data showed population movements during working periods and holiday periods more variable than initially assumed in the weight model due to flexibility in holiday periods, working at different locations etc.

The observations in 2020 suggested that the quality of the weighted MNO-data was sufficient to detect short-term trends in populations movements due to holidays, and the outbreak of the COVID-19 pandemic. The observations of 2020 also suggested that the quality was insufficient for official statistics, as official statistics are often based on level-estimates and long-term (year-to-year) changes.

3.5 Improvements in weighting and calibration

SCB has actively participated in these discussions to propose improvements in the weighting methodology of MNO-data. These are partly implemented by Telia. At the same time a benchmark report was developed in which the (monthly) estimates derived from MNO-data are compared with related official statistical from SCB such as population statistics, registered based commuting statistics, education and income statistics. This benchmark report can be updated monthly. Its aims, together with the process descriptions, at quality assurance. More specifically, its aim is:

- checking the plausibility of the MNO-estimates and detecting level shifts in time-series at an early stage, which may occur due to subtle changes in client share, population behaviour regarding carrying mobile phones and methodology improvements.
- detecting real changes in population movements and evaluating which 'smart statistics' can be developed out of MNO-data.

Time-series with level shifts may even occur after the improving the weighting technique. These level-shift are often difficult to predict as they may be caused by human behaviour (such as people using phones for separate purposes). An extra quality check has to be added.: calibration. A simple simple 'calibration' model assumes that the average number of night population activities during working in the months October-November corresponds with the SCB-population statistics. If not the time-series have to be calibrated to these values. Calibration methods have to be elaborated further in order to detect level shifts at an early stage and correct for these level shifts. An example is shown in figure 1.

SCB is currently developing a similar weight model itself for the MNO-data of the other operator.

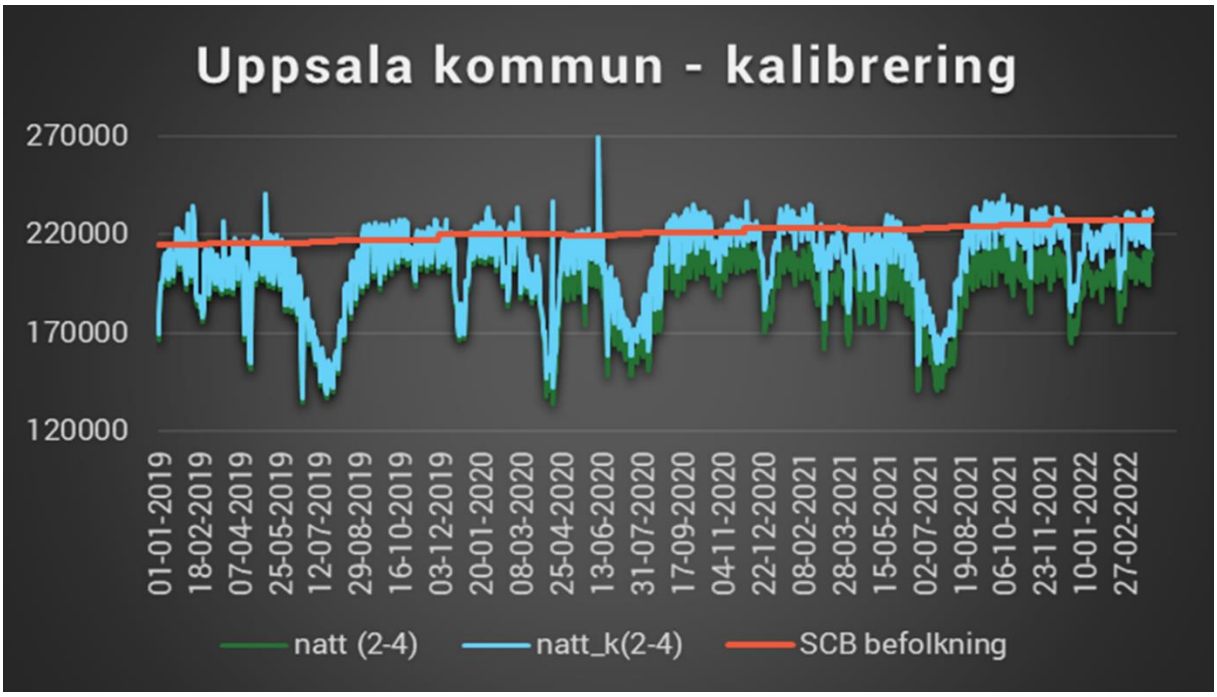

Figure 1 Night population activities of Uppsala municipality, a university city north of Stockholm. The green line denotes the night population activities derived from the standard data-processing (carried out by the operator). The blue line shows the 'night population activities' curve corrected for level shifts. The red line shows the outcome of SCBs population statistics.

This curve shows real changes in night populations such as lower populations during summer, Christmas and Easter. These can be explained by people (partly students) leaving the town. Lower night populations during the weekend in other periods are likely also be related to people leaving town. The same is probably true for the strong decrease night activities at the start of the pandemic (March 2020). Consequently, it can, therefore, be explained that the average yearly night population activities in Uppsala municipality are lower than the register based SCB population estimates (red line).

Artificial features are, however, small level shift in the time-series after the outbreak of the COVID-19 pandemic and each summer. The blue line is corrected for these levels by using a simple 'calibration' model, which assumes the average number of night population activities during working in the months October-November corresponds with the SCB-population statistics.

5. Quality assurance based on data deliveries

Figures 2 and figures 3 show two time-series (2019-2021) of daily night populations estimates derived from the MNO-data from Telia. Figure 2 is for Stockholm, the capital

of Sweden, Figure 3 shows the Jokkmokk municipality, Jokkmokk is a sparsely populated municipality in the northernmost part of Sweden.

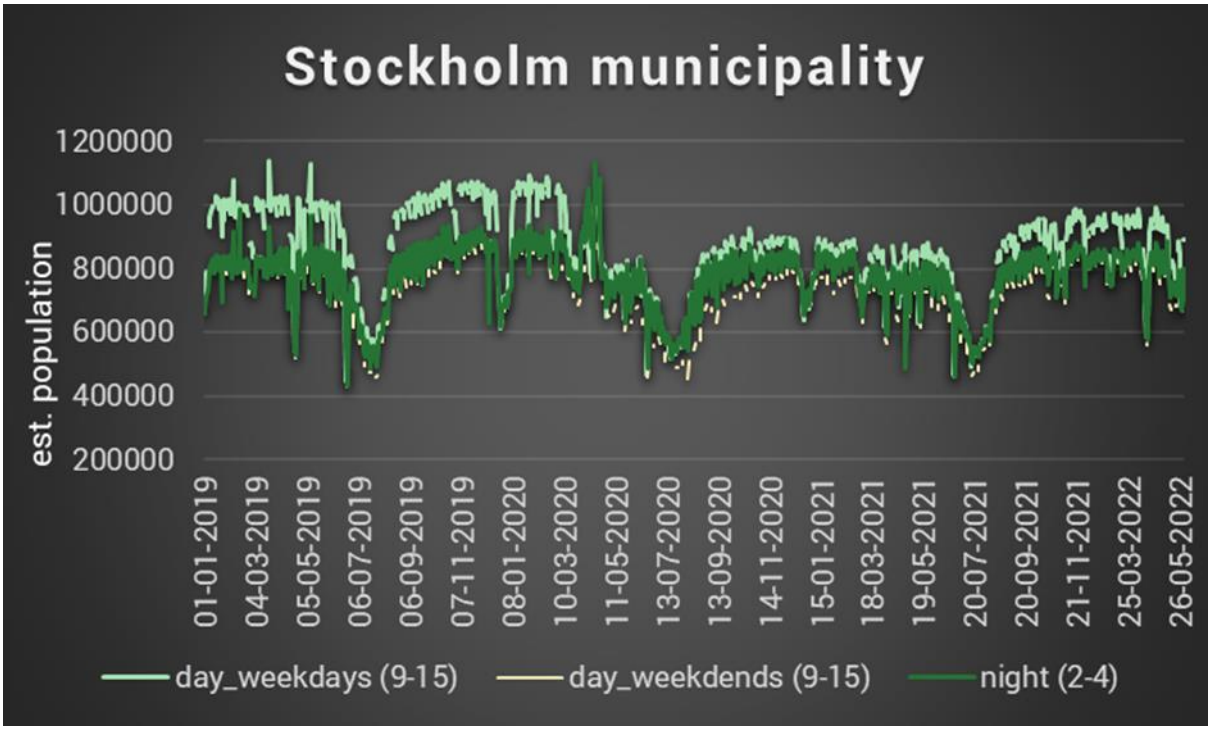

Figure 2 Time-series of day-night populations activities for Stockholm municipality. Note that before the COVID-19 pandemic the day population during weekdays (light green line) are higher in Stockholm municipality, together with the associated trip data suggesting commuting to Stockholm for work. These higher daily population during weekdays disappear during the COVID-19 pandemic and only slowly recover in the most recent periods. During weekends day populations are in general slightly lower than night population (yellow lines).

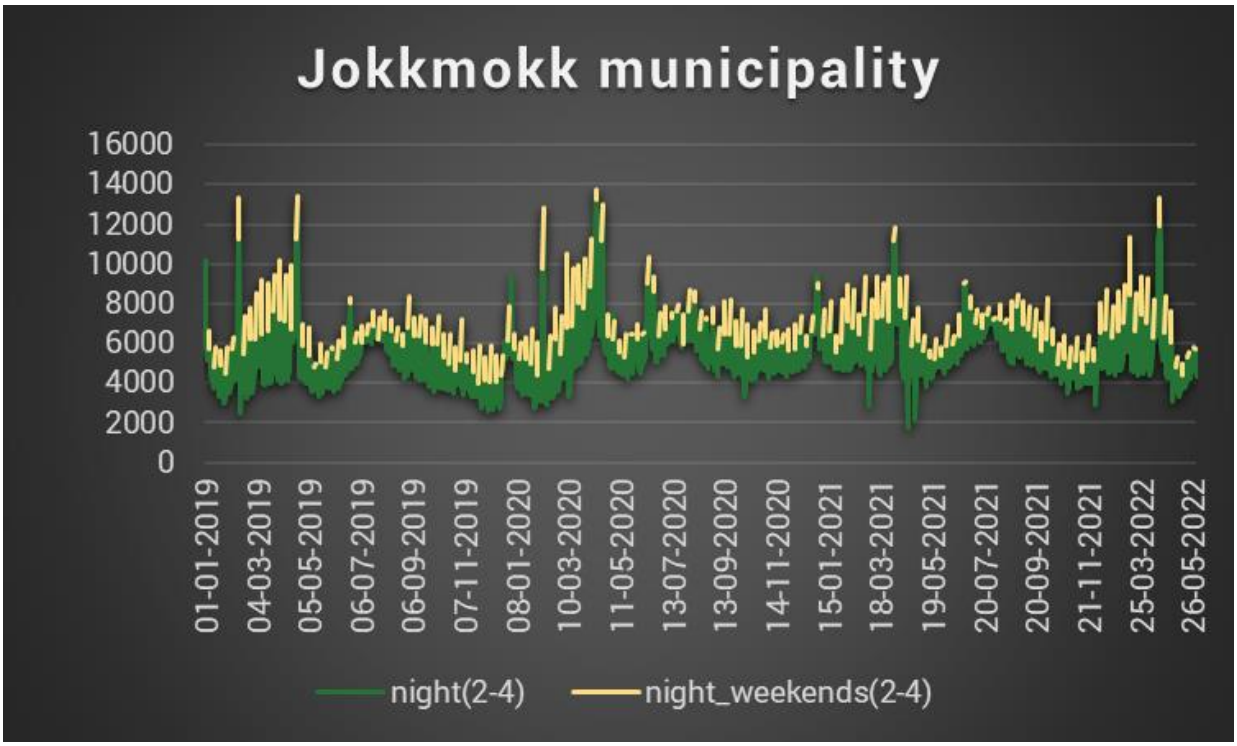

Figure 2 Night activities of Jokkmokk municipality, a sparsely populated municipality in the northernmost part of Sweden. Seasonal patterns in this time-series do exists with higher populations during late-winter/early spring than during the summer. However, the most striking feature of this series are the higher night populations during the weekends. This can be explained by people spending the weekends at their house in Jokkmokk but during the week working in the towns along the Bothnian Gulf (or in the mines in the nearby Gällivare and Kiruna municipalities). Consequently, it can be explained that for this municipality the yearly average of night population activities are higher than the register-based SCB population estimates (red line). A small level shift, though, is detected before and after the outbreak of the COVID-19 pandemic, hampering exact quantification.

As explained in the figure captions, these examples do – in contrast to the register-base population statistics - reveal seasonal variations in populations and explain that for some municipalities monthly, quarterly and even yearly average night populations are higher than the register-based population estimates from SCB, while for other municipalities the average night population activities are lower than the register-based population estimates.

As a consequence, it can be concluded that mobile phone position data are a valuable data source for smart statistics about dynamic populations and population movements

6. Possibilities for new statistics

Investigating the development of smart statistics out MNO-data is part of the government assignment. Potential smart statistics which can be developed out of these data are monthly or weekly statistics about 1) night populations activities, 2) the ratio day/night populations and related travel statistics. The hight population activities shows among others clearly seasonal differences in population activities per municipality, region and district (fig. 3). They show for Sweden a population movement from cities and suburbs to the countryside during the summer. Statistics on this topic serve the public interest (for example Road Agencies and Regional Health Care Organisations). The ratio day/night populations activities can, be related to daily commuting and shows effect of the pandemic in our daily commuting patterns (fig. 2). This interpretation is supported by associated trip data.

It should, however, be noted that these smart statistics can be published only if the benchmark methods to detect and correct level are tested. Another challenge is the marketing of these potential 'smart' statistics. They should be positioned as a high-quality complement to the use of MNO-data by MNOs themselves for commercial activities. Future partnerships with MNOs should be based on this relationship

.

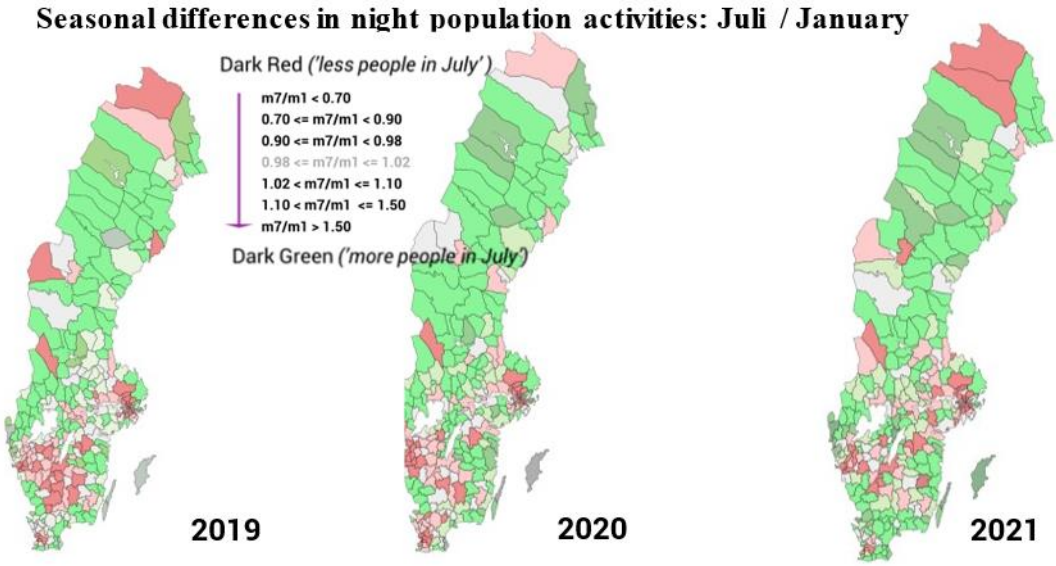

Figure 3 Example of seasonal variations in night population activities. Comparison between July and January in the same year. Green areas are municipalities with higher population activities in July, These are in most cases municipalities on the coast, in touristic areas of densely populated municipalities with many ‘leisure houses’. Red areas are municipalities with lower population activities in July. These are in most cases cities and suburbs or in the northernmost part municipalities with mines.

7. Conclusions

Statistics Sweden works with MNO-data since 2020. It is a fact that MNOs in Sweden want to deliver aggregated and anonymised data to SCB only. They are, however, willing to share documentation if the legal aspects are covered. Role of SCB in these partnerships is facilitating in describing the processes but active in relating the weighted MNO-data to official statistics and improving benchmark techniques to monitor the quality of the results. Having access to aggregates at low aggregation level is also needed to carry out this task. Final aim is for SCB to develop smart statistics about seasonal dynamic populations out of these data and statistics about commuting and trips in urban areas. For this purpose, partnerships with MNOs are needed and a business model in which statistics are positioned as a high-quality complement to commercial activities on MNO-data by the MNOs themselves.

8. Appendix: List of references

Vlag P., 2021, QUALITY IMPROVEMENT IN MOBILE PHONE POSITION DATA: A STEP FORWARD TOWARDS USE FOR OFFICIAL STATISTICS, SCB report