# WORKING GROUP ON DATA INTEGRATION

*The integration of geospatial data and statistics to compute SDG indicators: requirements and practices*

UN-GGIM: EUROPE

UNITED NATIONS COMMITTEE OF EXPERTS ON GLOBAL GEOSPATIAL INFORMATION MANAGEMENT

# Guidelines for SDG Indicator Calculation

11 SUSTAINABLE CITIES AND COMMUNITIES

## *11.2.1 Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities*

*Compiled by Statistics Sweden*

*with the contributions of:*

Austria (NSI) | Denmark (NMCA) | DGREGIO | Finland (NMCA)

Poland (NSI) | Portugal (NSI) | Slovenia (NSI) | Turkey (NSI)

*December 2021*

# 11.2.1 Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities

## A. INDICATOR CALCULATION AND DISCUSSION

### 1. DEFINITION

This indicator is used to monitor the progress towards SDG 11 on the use and access of public transportation systems and move towards reducing the reliance on the private means of transportation. This includes improving the access to areas with a high proportion of transport-disadvantaged groups such as the elderly citizens, physically challenged individuals, and low-income earners - and reducing the need for mobility by decreasing the number of trips and the distances travelled.

The accessibility-based urban mobility paradigm also critically needs good, high-capacity public transport systems that are well-integrated in a multimodal arrangement with public transport access points located within comfortable walking or cycling distances from homes and jobs for all. Since most public transport users walk from their trip origins to public transport stops and from public transport stops to their trip destination, local spatial availability and accessibility to these stops is often evaluated in terms of pedestrian (walk) access, as opposed to cycle access, park and ride or transfers. Hence, the initial version of the global metadata proposes the access to public transport to be “convenient” when a stop is in walking distance of 0.5 km from a reference point such as a home, school, work place, market, etc[1]. Recently, the updated global metadata proposes a distinction between low-capacity modes (e.g. bus), for which walking distance of 500 m is used, and high-capacity modes (e.g. rail, metro), for which a walking distance of 1 km is used[2].

The global metadata distinguishes the *core indicator* from *complementary or other potential indicators* of “convenient access” that may provide a useful complement to the core indicator. Several complementary indicators are proposed in the metadata. These all represent potential extensions, or more detailed modifications, of the core indicator. However, complementary indicators require more data and, in some cases, may compromise the comparability of the indicator across jurisdictions, since collection techniques may vary. This paper is focussing on the core indicator, particularly in its simplest version, i.e., the initial one, without distinction between low- and high-capacity modes.

### 2. METHODOLOGY

Computation of the core indicator comprises the main methodological components outlined in this section:

Delimitation of the urban area**:** In line with the rationale of Goal 11, the indicator is targeting the urban context. Hence, as a first step, the urban area and its population needs to be defined and delimited. The global metadata does not give a clear recommendation as to how the urban area should be defined. There are several possible ways to define and delimit “urban”:

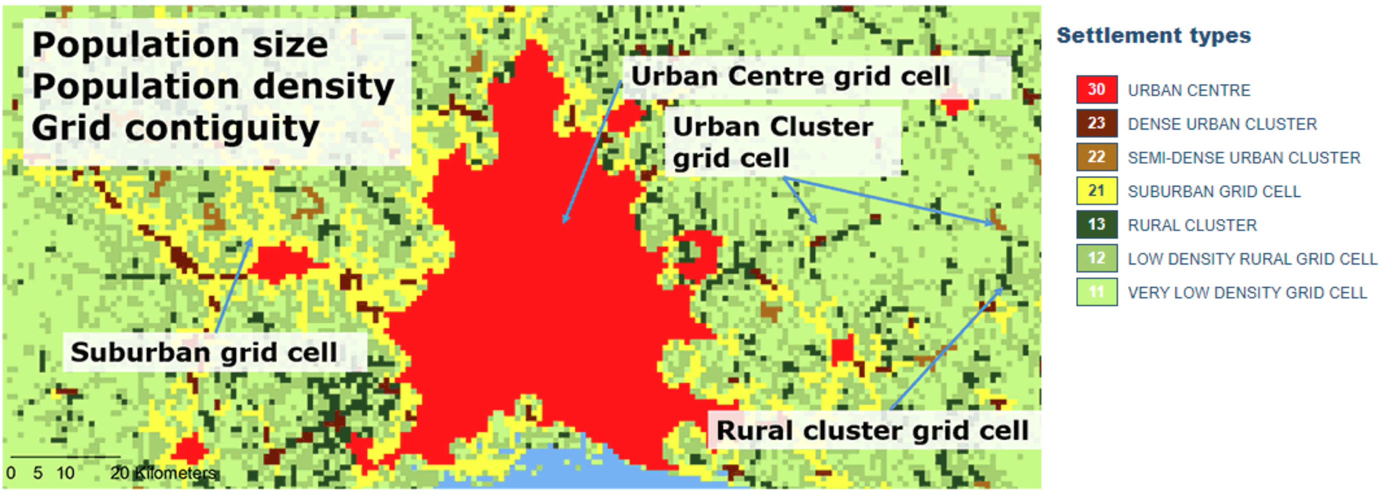
- Existing national urban geographies based on the boundaries of administrative urban units (only viable in countries where cities or towns are explicitly recognised as administrative units)
- Existing national urban geographies based on functional, morphological or land use conditions (such as density of built-up areas, population density etc.)
- Custom created urban boundaries for the purpose of SDG monitoring based on methodology described by UN-Habitat (compare with indicator 11.3.1[3])
- Existing global or European harmonised geographies based on gridded population data (Degree of Urbanisation)

[1] UN-Habitat Metadata on SDG Indicator11.2.1 (March. 2018)
[2] UN-Habitat Updated Metadata on SDG Indicador 11.2.1
[3] UN-Habitat Metadata on SDG Indicator11.3.1

In the context of SDG reporting, there are good reasons to advocate a harmonised concept to define urban in order to enable comparable data between countries. At the same time, local and national administrations may request indicator data based on urban geographies that respond well to their policy needs but are not necessarily well fitted for international comparisons. In order to have meaningful international comparisons of statistical indicators there is an undisputed need for a definition that is nationally relevant and internationally comparable at the same time. One of the few candidates that can balance these somehow contradictory requirements is the Degree of Urbanisation (DEGURBA). The method was proposed by a consortium of international organisations (EU, OECD, World Bank, FAO, UN-Habitat, ILO) led by the EU.

The Degree of Urbanisation classifies the entire territory of a country along the urban-rural continuum. It combines population size and population density thresholds to capture the full settlement hierarchy – 1 km$^2$ grid cells are classified based on population density, contiguity, and population size.

*Figure 1 The Degree of Urbanisation (DEGURBA)*




Source: Joint Research Centre (JRC).

DEGURBA offers a method to classify the entire territory of a country along the urban-rural continuum. It combines population size and population density thresholds to capture the full settlement hierarchy. DEGURBA relies on 1 km$^2$ grid cells that are classified based on population density, contiguity, and population size. The proposed approach for this indicator is to include all the urban settlement types as defined by DEGURBA to define “urban”. In case of limited data availability, particularly of public transport timetables, the indicator could primarily focus on the urban centres (i.e., the high-density clusters).

DEGURBA offers analysis ready data as well as methods and tools to create classifications. Eurostat provides readily classified data on urban clusters based on the 1 km$^2$ GEOSTAT population grid (2011). In addition, Joint Research Centre (JRC) offers classified data based on GHSL (Global Human Settlement Layers) which is a global population grid created through downscaling of census data originally released on coarser areas.

Identification of public transportation stops: The analysis crucially requires data on the location of public transportation stops, but preferably also data on the mode of transportation and the frequency of departures at each stop.

Currently, there are no pan-European data on public transportation stops available[4]. A growing number of transport operators and regional and national data integrators have made significant efforts to provide open and up-to-date data according to a de-facto standard. Despite this, timetables are still unavailable for large parts of the European territory and standardisation of the data models needs to be improved. The data landscape is also quite fragmented, often requiring time-consuming efforts by analysts to find the right data and make them fit for use.

In a recent study for the European Commission, DGREGIO made the following ranking of preferred data sources:

1. Integrated (authoritative) datasets providing stop locations and scheduled timetables for all public transport in an entire country (or region).
2. Datasets providing stop locations and scheduled timetables for major transport operators in a particular region or city.
3. Datasets from authoritative sources, providing stop locations in a particular country or region (without data on scheduled timetables).
4. Volunteered geographical information data on stop locations (e.g., Open Street Map).

An increasing number of timetable datasets is available according to the General Transit Feed Specification (GTFS), including stop location data, timetables, information about the transport mode and dates of the services' activities. The most crucial feature of the GTFS data model is a table of stops including their location (latitude and longitude). Other tables in the GTFS data model need to be related to the stops to retrieve the departure times per stop, to select the relevant days of operation, and to retrieve information on the transport mode available at each stop.

Information on transportation mode is needed to comply with the global metadata suggesting a distinction between low-capacity modes (in particular, bus and tram), for which walking distance of 500 m is used, and high-capacity modes (in particular, rail, metro, and - if relevant - ferry), for which a walking distance of 1 km is used. If a public transportation stop is served with different both low and high-capacity modes, the stop should be classified as high-capacity mode. If no information is available, all stops should be classified as low-capacity mode.
Another reason to relate timetable information to each stop is to be able to assess if the stop should be included in the analysis. Some stops may have a very low traffic frequency, e.g., less than one departure per day and/or only weekend or seasonal services. Such stops hardly contribute to good, reliable, high-capacity public transport systems, hence including them in the analysis will overestimate the accessibility to public transportation.

The global metadata does not require use of thresholds, but UN GGIM: Europe suggests that it is good practice to apply a threshold to separate and exclude stops with irregular or very low service frequencies from the analysis. There is currently no common and agreed standard on such thresholds. Table 1 presents a number of practices used by different countries/organisations to calculate indicator 11.2.1.

*Table 1 Examples on threshold rules used by different countries/organisations to calculate indicator 11.2.1*

| Country/organisation | Threshold rule |
|---|---|
| Austria | No threshold applied |
| Denmark | Minimum one departure between 06:00-20:00 on a typical working day |
| DG REGIO | No threshold applied but disaggregation of data by frequency classes |
| Estonia | Minimum one departure between 06:00-20:00 on a typical working day |
| Finland | Minimum one departure per hour on average between 06:00-21:00 on a typical working day |
| Israel | No threshold applied |
| Norway | Minimum one departure per hour on average between 06:00-20:00 on a typical working day |
| Portugal | No threshold applied |
| Slovenia | Minimum 8 departures per day |
| Sweden | Minimum one departure per hour on average between 06:00-20:00 on a typical working day |
| Switzerland | Minimum one departure between 06:00-20:00 on a typical working day |
| Turkey | No threshold applied |

Source: Written contributions submitted to UN GGIM: Europe Working Group on Data Integration and the GEOSTAT 3 project.

In the case of Sweden, around 30% of the total number of public transportation stops existing in the country (~50 000) are discarded due to low service frequency. However, it should be noted that the majority of these public transportation stops are located in rural areas. The "loss" of public transportation stops in urban areas, due to a low service frequency, is only around 5 %.

There are several possible ways to define a "typical weekday". One simple way is to select a reference day, e.g., Wednesday in the second week of October etc. When using this approach, it is important to choose a reference day that is not subject to public holidays, school holidays or other events that may cause deviations from the normal service frequency. For consistency over time, the same reference day should be used whenever the indicator is updated.

The location of stops can be portrayed in various ways depending on how the data model is used. For instance, if a bus stop is located on both sides of a street (i.e., one stop for each direction), some service providers will consider this to be one single stop, while others will provide separate data for the actual location of each stop. Something similar happens when representing bus stations or railway or metro station platforms. To create more homogeneity in the data and enhance the comparability of the results, it is recommended to aggregate all stop points located very close to each other to a single "cluster" of stops (for instance stops within 50 meters from each other). This will significantly reduce the number of calculations later without compromising the spatial quality of the data.

Python scripts for processing of GTFS data, provided by DG REGIO, are available by clicking **here** or at direct link to python script package available **here.**

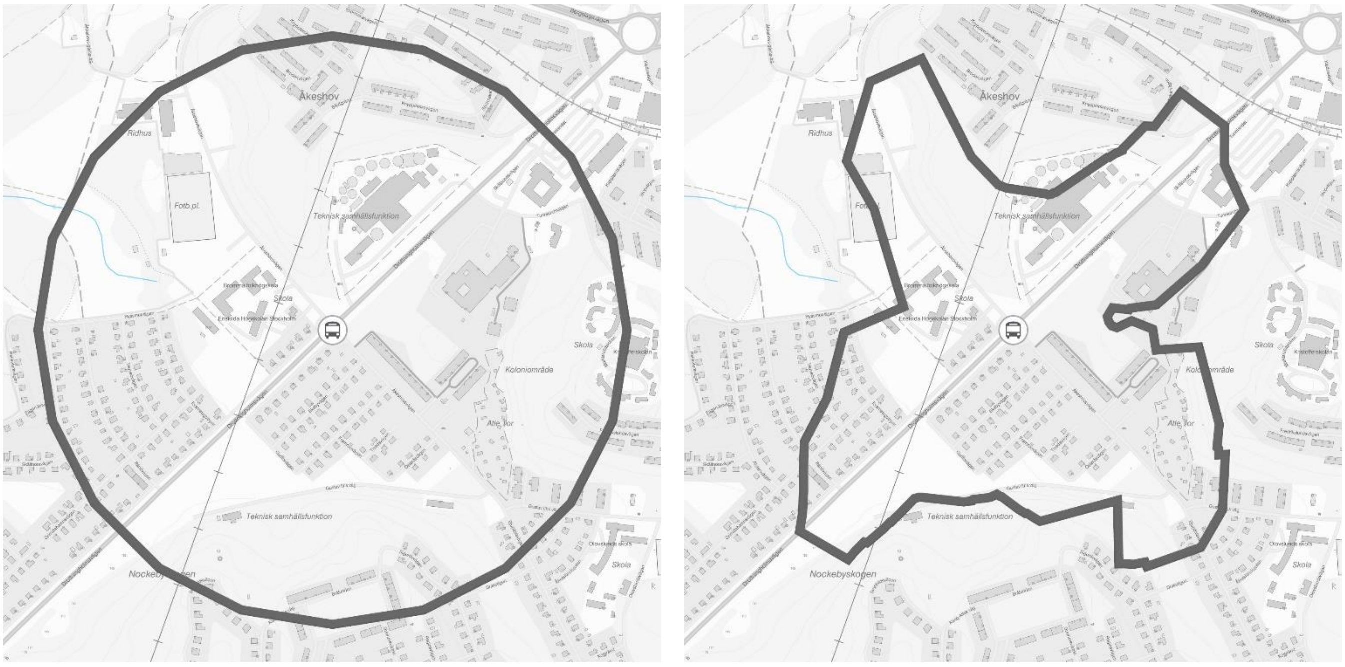
Creation of service areas: Service areas are typically identified using a buffering operation (in GIS), which creates a zone of equal proximity around each public transport stop (left map in Figure 2). However, the global metadata proposes to identify the size of the service area by the network distance – instead of using a buffer distance (equal proximity) around the transport stop (right map in Figure 2).

Hence, for the core indicator, public transport is considered "convenient" for those living within a 0.5 km (or 500 m) walkable distance of the nearest "low-capacity stop" and within 1 km of a "high-capacity stop". Using network distance (that is, the walking distance computed using the street network to reach a public transport feature) will help to reflect the configuration of the street network in a realistic way and to recognize the presence of any barriers preventing direct access to public transport features.

The following figure shows a 500 m service area calculated as a buffer around the public transportation stop (Euclidian distance) and a 500 m service area defined through network distance.

[4] A UN-Habitat training module on how to capture Public Transport System is available.

*Figure 2 Service areas (500 m) calculated as buffer around the public transportation stop using Euclidian distance (map on the left) and through network distance (map on the right)*




Source: Lantmäteriet, base map © Lantmäteriet.

For the core indicator, the global metadata recommends use of network distance to the public transport stop to define the service area. UN GGIM: Europe endorses the use of service areas based on network distance as the first-hand choice. However, it is important to notice that using network distance to measure access to public transport stops requires a reliable and high-quality network dataset, including even small streets and attributes enabling a selection of streets accessible to pedestrians, with complete coverage and equal quality across the country.

Low quality, or incomplete, network data will induce poor quality to the calculations. Likewise, a network dataset that is subject to substantial quality improvement over time will risk inducing “false” changes in time series. Under such circumstances, a simpler use of Euclidian distance buffering provides a viable option.

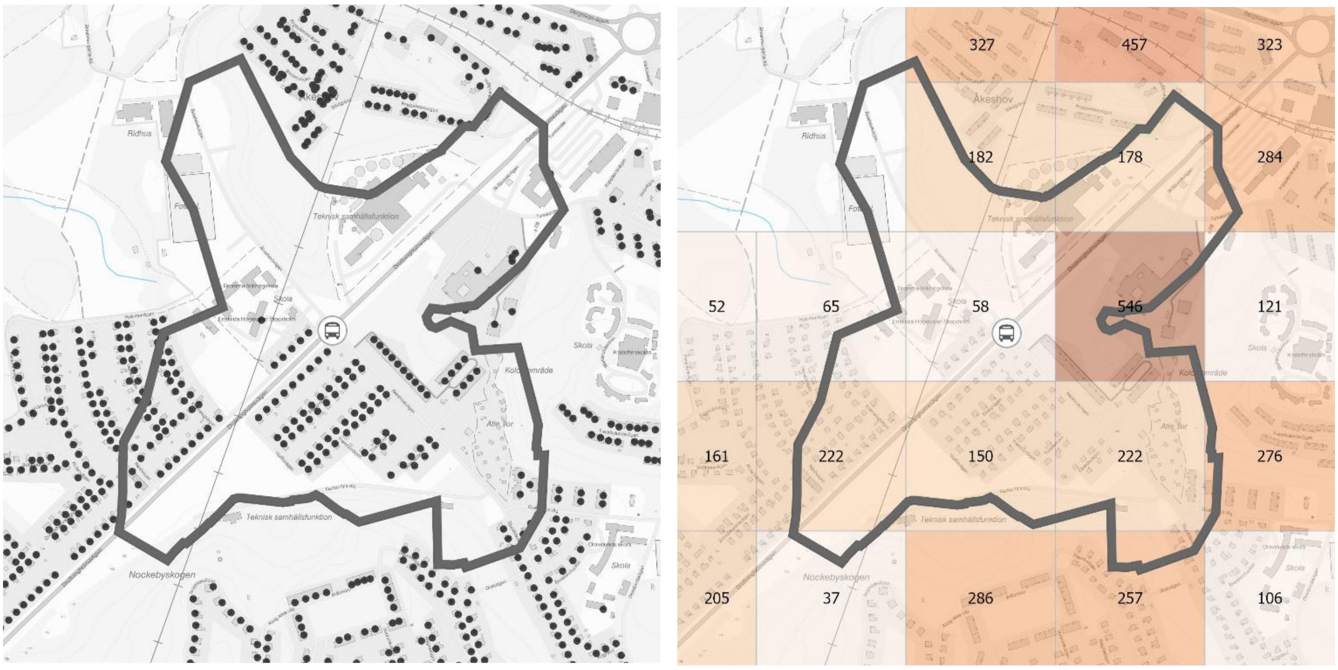
If information on mode of transportation is available, separate the public transport stops in two groups: a) stops of low-capacity modes (in particular, bus and tram) and b) stops of high-capacity modes (in particular, train and metro). From these two groups create service areas using 500 m distance for group a) and 1 km for group b).

The identification of the population served: Once a service area is created, the next step is to overlay the service area onto a spatial data layer describing the distribution of the urban population. The population that falls inside the service area is considered to have convenient access to public transport.

Data that describes distribution of the urban population can vary in terms of resolution and accuracy, from the most fine-grained level, where each person is located with single x and y coordinates (point-based), to a coarser level where the population has been aggregated to enumeration areas, districts or grid cells (area-based). Both types of population data can be used for the calculation, however point-based data enables a more straightforward computation and undisputedly provides a more accurate result. Therefore, it is recommended to use point-based population data as the first-hand choice. However, if point-based data are not readily available, optimised disaggregation processes may be a suitable alternative.

The following figure shows a service area on top of a dataset with point-based population data (e.g., population geocoded to address locations). The population inside the service area can be calculated easily using point-in-polygon techniques. It also shows the same service area imposed on a population grid (area-based). To calculate the population inside the service area, the data first needs to be downscaled using ancillary data.

*Figure 3 Service area on top of a dataset with point-based population data (map on the left) and service area imposed on a population grid (map on the right)*




Source: Lantmäteriet and Statistics Sweden, base map © Lantmäteriet.

Use of population data aggregated to areas typically require some sort of procedure to compute the proportion of the population within an area that partially intersects a service area. Such procedures can be based on simple rules, for example that the population in the area that falls inside the service area is proportional to the share of the area that falls inside the service area. However, there is a risk that such simple assumption will produce less accurate results, especially when the population within an area is unevenly distributed. To reduce this risk, it is important to use population data at the highest spatial resolution possible.

Consequently, to avoid a less accurate result, area-based population data may first need to be downscaled to smaller spatial units using ancillary data. Such processes should make the best use of detailed land use/land cover information. In particular, a good thematic distinction of built-up classes such as industrial and commercial areas and areas of mixed use (residential versus non-residential) should improve the results. Data on the height and/or on the function of buildings is also very valuable in improving the disaggregation process.

When downscaling population to smaller spatial units, using input data as the highest possible spatial resolution is of crucial importance. Input data at the level of small census tracts, enumeration areas, or high-resolution grids (i.e., cells of less than 1 km$^2$) are particularly appropriate. The target areas of the disaggregation process should be fit for the purpose of combining them with the road network and public transport data. Hence, using building blocks (i.e., areas surrounded by streets) or polygonal parts thereof as target areas is an interesting choice, particularly when ancillary data about land use / land cover are available at building block level. These land use/cover characteristics can then be combined with data on the density of buildings, and if possible, of their height. Data on building density (and height) can be acquired from remote sensing or from georeferenced cadastre data[5].

A comparative study conducted by Statistics Austria and DG REGIO[6] concludes that the difference in output using point-based population data versus gridded population data downscaled to Urban Atlas polygons is rather small. Use of downscaled population data somewhat underestimates the indicator value compared to point-based data. For the seven cities studied in Austria, the underestimation is approximately 1%. To a certain extent, the limited differences in outcome can be explained by the fact that the urban centres are – by definition – relatively homogeneous areas, at least in terms of minimum population density. In such areas, downscaling tends to perform better than in more heterogeneous areas (in particular, where a large variety of land use/cover classes is present).

## 3. DATA SOURCES

The following tables summarise available geospatial and statistical Pan-European data sources (Table 2 and Table 3) and the required national geospatial and statistical data sources (Table 4 and Table 5) for indicator computation.

*Table 2 Pan-European geospatial data sources*

| Name | Source | Periods of reference available | Frequency | Spatial Resolution | MMU | INSPIRE Data Theme | Core data theme |
|---|---|---|---|---|---|---|---|
| Urban Atlas | Copernicus | 2006, 2012, 2018 | 6 years | - | 0.25 ha | Land cover | Land cover |
| GHSL/GHS-POP | Joint Research Centre | 1975, 1990, 2000, 2015 | 15 years | 250 m, 1 km, 9 arcsec, 30 arcsec | - | Population Distribution | - |
| GHSL/GHS-BUILT | Joint Research Center | 1975, 1990, 2000, 2015 | 15 years | 30 m, 250 m, 1 km | - | Land cover | Land cover |
| GHSL/GHS-BUILT-S2 | Joint Research Centre | 2017-2018 | No time series yet | 10 m | - | Land cover | Land cover |
| GHSL/GHS-MOD | Joint Research Centre | 1975, 1990, 2000, 2015 | 15 years | 1 km | - | - | - |
| ESM | Joint Research Centre | 2015 | No time series yet | 2 m, 10 m | - | Land cover | Land cover |

*Table 3 Pan-European statistical data sources*

| Name | Source | Periods of reference available | Frequency | Max territorial granularity | Other relevant disaggregation | ESS regulation reference |
|---|---|---|---|---|---|---|
| Gridded Population data | Eurostat | Decennial | Decennial | Gridded population data (1 $km^2$) | - | Regulation (EC) No 763/2008, Regulation (EU) 2018/1799 |

*Table 4 National/regional geospatial data sources*

| Name | Source | Periods of reference available | Frequency | Spatial Resolution | MMU | INSPIRE Data Theme | Core data theme |
|---|---|---|---|---|---|---|---|
| Public transportation stops | Transport ministries/agencies, public transport service providers | Vary | Vary | - | - | Transport networks | Transport networks |
| Road network | Transport ministries/agencies, commercial data or Open Street Map | Vary | Vary | Vary | - | Transport networks | Transport networks |
| National Land Cover data | Geospatial agencies, environmental agencies | Vary | Vary | Vary | Vary | Land cover | Land cover |

[5] A disaggregation process using Copernicus Urban Atlas data in combination with building densities from remote sensing has been applied by EC DG JRC: Batista e Silva, F. & Poelman, H., Mapping population density in Functional Urban Areas. A method to downscale population statistics to Urban Atlas polygons. JRC Technical Report, EC JRC Ispra, 2016. For an illustration of the use of data on functions and height of buildings for the purpose of population downscaling, see the paper by Hugo Poelman on Mapping high-resolution population and employment data in urban areas, using Copernicus Urban Atlas as a framework: Some tests taking into account the third dimension.

[6] Un-published paper written by Hugo Poelman and Ingrid Kaminger for the UN GGIM: Europe Working Group on Data Integration.

*Table 5 National statistical data sources*

| Name | Source | Periods of reference available | Frequency | Max territorial granularity | Other relevant disaggregation | ESS regulation reference |
|---|---|---|---|---|---|---|
| Population data | NSOs/Census data | Annual to decennial | Annual to decennial | Non-aggregated Geocoded to point-location | | |
| Population data | NSOs/Census data | Annual to decennial | Annual to decennial | Aggregated to small grids or other small areas | | |

## 4. COMPUTATION

The step-by-step description provided below can be considered as a short version of the methodological description in section 2. For methodological considerations and references, please see Section 2.

1. Chose an appropriate method to define urban areas and make an inventory of readily available data to support the selected method. If analysis ready data does not exist for the desired area or time-period, use spatial analysis to define urban areas following methodological guidance and tools for DEGURBA.
2. Make an inventory of the best possible data on public transport stops (including timetable information and information on mode of transportation if available) and prepare data for analysis.
   Apply a threshold if needed to discard public transportation stops with a very low or irregular service frequency.
   Create service areas from the selected public transportation stops using a) *network distance* calculation or b) *Euclidian distance* calculation.
   If information on mode of transportation is available, separate the public transport stops in two groups: a) stops of low-capacity modes (in particular, bus and tram) and b) stops of high-capacity modes (in particular, train and metro).
   From these two groups create service areas using 500 m distance for group a) and 1 km for group b). If no information on mode of transportation is available, use 500 m distance for all stops.
3. Make an inventory of the best possible data on population. If available, use point-based population data for efficiency and accuracy. If point-based population data is not available/accessible, use area-based population data combined with ancillary data (e.g., land cover data) to downscale population to smaller spatial units.
4. Use spatial analysis to overlay a) *urban areas* + *population* to calculate the total urban population and b) *urban areas* + *service areas* + *population* to calculate urban population with convenient access to public transport.
5. Compute the final indicator value for all urban areas (national level) and if desired, disaggregated by region or for individual urban agglomerations (cities). Use the following formula:

$$100 * \frac{Population\ with\ convenient\ access\ to\ public\ transport}{Total\ population}$$

## 5. RESULTS

Though all calculations need to be undertaken on a precise local scale, the result is typically presented on an aggregated level (e.g., national level). Data may also be presented on city level.

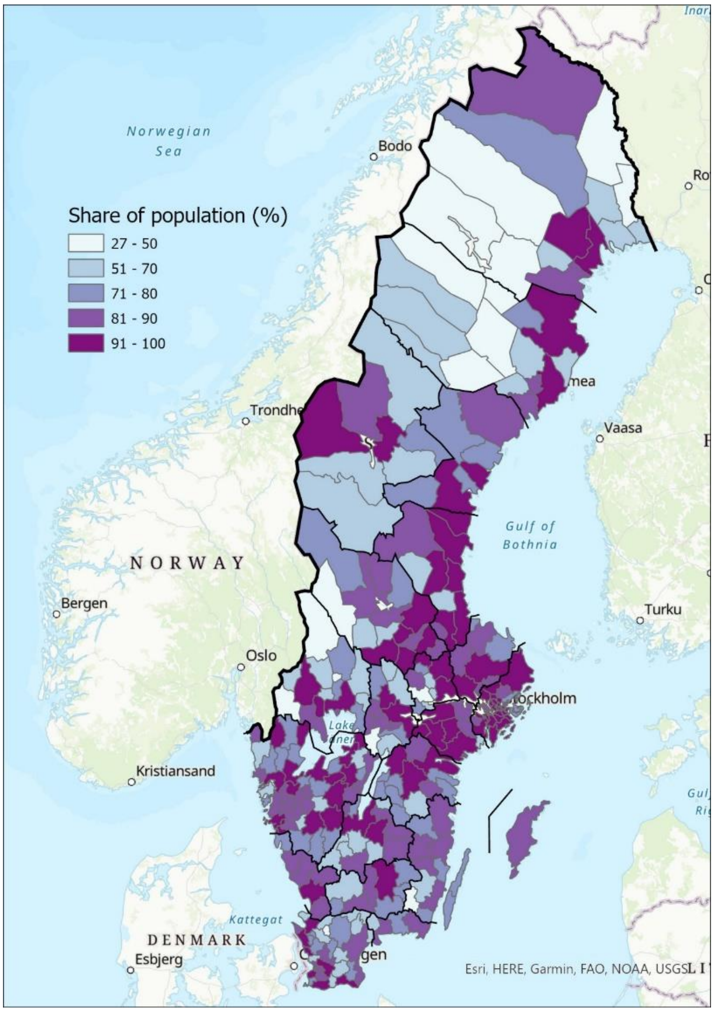
In the case of Sweden, indicator 11.2.1 is calculated annually and presented for the purpose of national monitoring of Goal 11. The indicator is presented as a national aggregate, disaggregated by sex and by a couple of coarse age groups. However, the production pipeline allows disaggregation of any regional or statistical units. When the indicator is calculated, the result is stored as open data in a statistical database allowing users to access data for counties and municipalities.

*Table 6 Proportion of urban population that has convenient access to public transport, by sex and age*

| Year | Female | Male | Total |
|---|---|---|---|
| 2014 | 88.4 | 87.7 | 88.0 |
| 2015 | 89.0 | 88.2 | 88.6 |
| 2016 | 88.9 | 88.3 | 88.6 |
| 2017 | 89.0 | 88.4 | 88.7 |
| 2018 | 90.0 | 89.4 | 89.7 |
| 2019 | 90.4 | 89.8 | 90.1 |

| Age group | Share (%) |
|---|---|
| 0-17 years | 87.4 |
| 18-64 years | 89.4 |
| 65- years | 88.2 |
| Total | 88.7 |

*Figure 4 Proportion of urban population that has convenient access to public transport, by municipality*




Source: Statistics Sweden.

# B. NORMATIVE GUIDELINES

## 1. ALGORITHM WORKFLOW

The following charts systematize the process and the options to define the data sources and methods to apply to address the different three components needed for indicator calculation, namely to identify public transportation stops (Figure 5), to capture population distribution (Figure 6) and to create service areas (Figure 7).

*Figure 5 Chart to define data and method to identify public transportation stops*

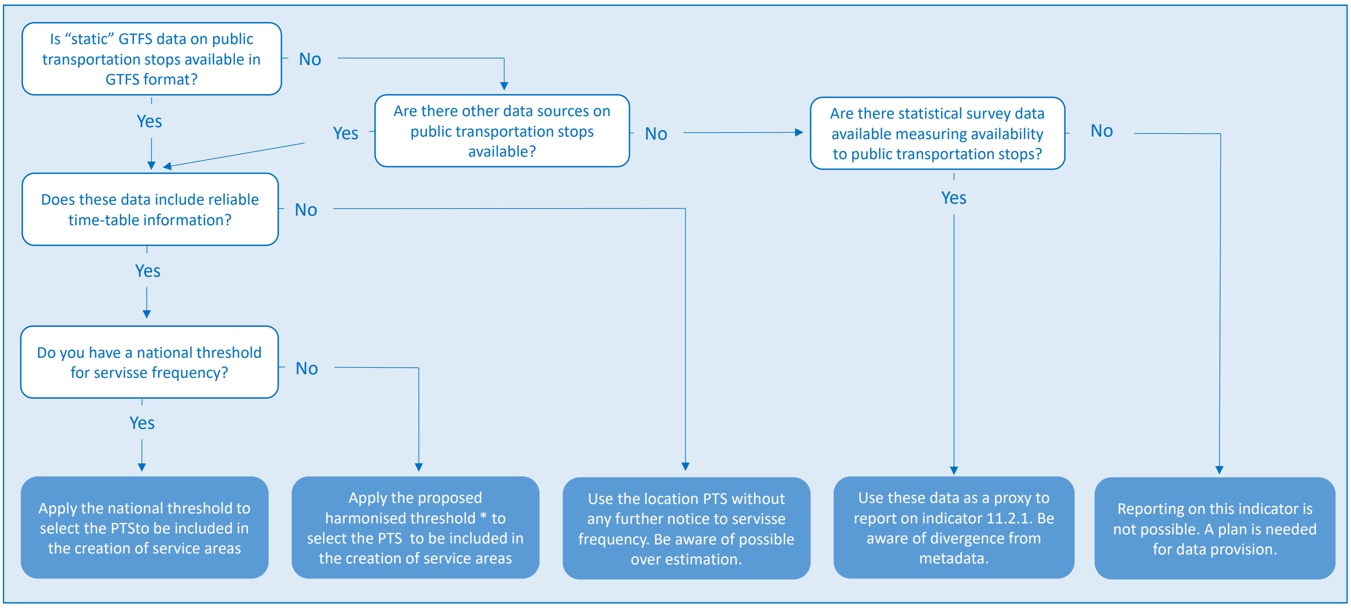



*Figure 6 Chart to define data and method for population distribution*

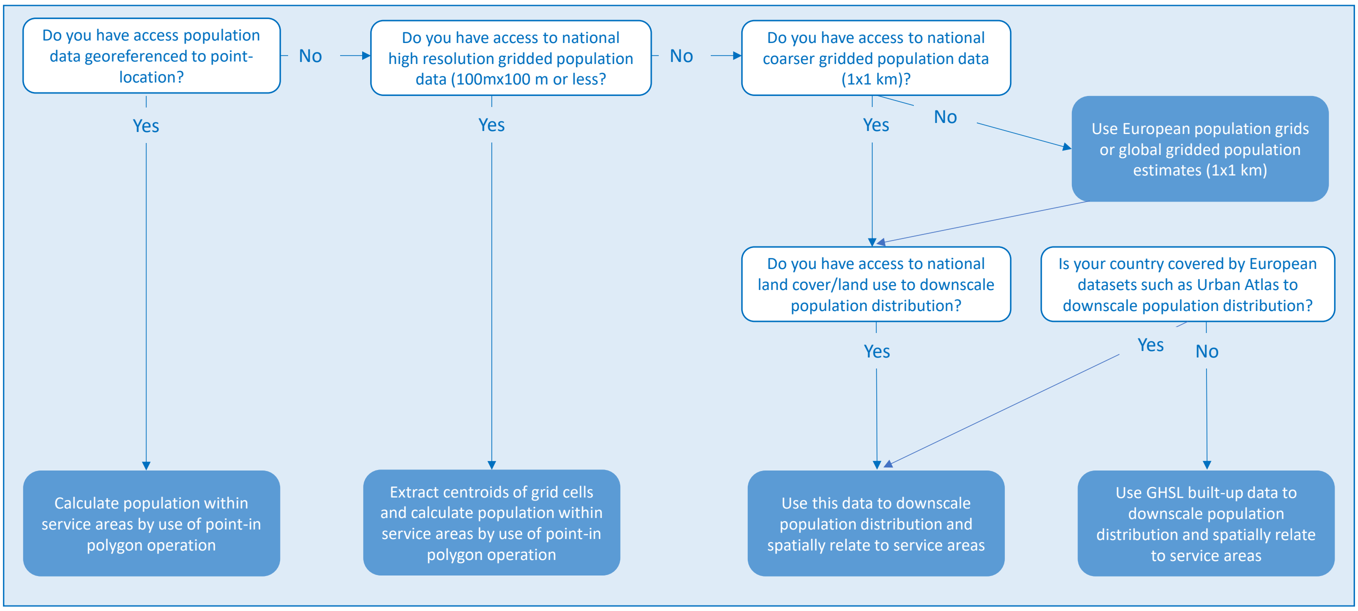
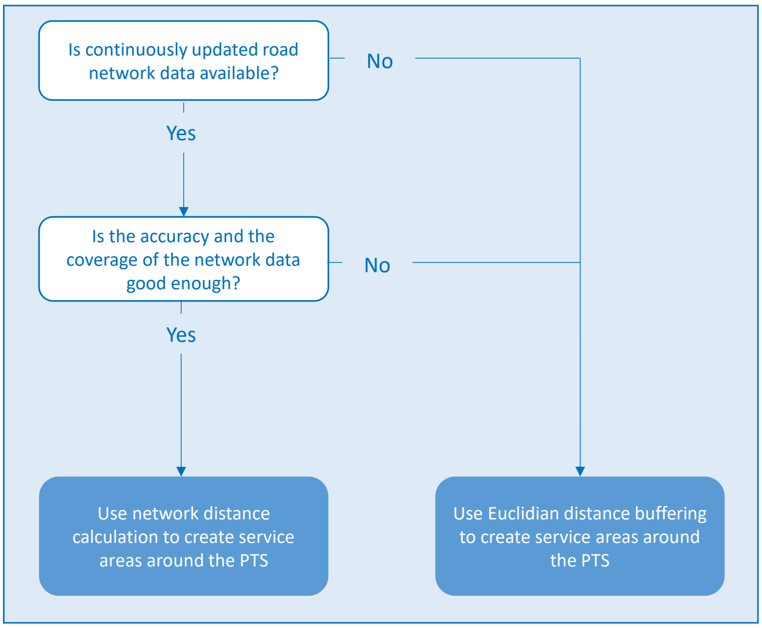


*Figure 7 Chart to define method for creation of service areas*




## 2. RECOMMENDATIONS

- To achieve consistent international comparability in SDG monitoring, the first-hand choice of urban geographies should be the Degree of Urbanization (DEGURBA).
- If available, use integrated (authoritative) datasets providing public transportation stop locations and scheduled timetables for all public transport in an entire country (or region).
- If available, it is good practice to use timetable information to identify and discard public transportation stops with very low and/or irregular (e.g., seasonal) service frequency.
- If available, it is good practice to use information on mode of transportation to make a distinction between PTS with low- (in particular, bus and tram) and high-capacity (in particular, train and metro) modes. For low-capacity PTS the service area should be calculated using a 0.5 km distance and for high-capacity PTS the service areas should be calculated using 1 km distance. If information on mode of transportation is not available, it is recommended to apply a distance of 0.5 km for all stops.
- To create more homogeneity in the data and enhance the comparability of the results, it is recommended to aggregate all public transportation stops located very close to each other to a single "cluster" of stops (for instance stops within 50 m from each other). This will significantly reduce the number of calculations later without compromising the spatial quality of the data.
- Creation of service areas should preferably be based on network distance calculation. However, in absence of proper and high-quality network data, service areas should be created using Euclidian distance buffering.
- It is recommended to use point-based population data as the first-hand choice. However, if point-based data are not readily available, optimised disaggregation processes may be a suitable alternative.