Eurostat's Traffic and Mobility Project

Evangelia FORD-ALEXANDRAKI[1], Matyas MESZAROS[1], Miriam BLUMERS[1], Nikolaos ROUBANIS [1]

[1]Eurostat, Luxembourg

Abstract

The availability of new data sources and privately held data has increased exponentially in recent years and accelerated the transition towards using innovative data sources in the ambit of transport statistics. Correspondingly, Eurostat is currently exploring their use and developing methods to produce statistical indicators based on these; as well as implementing automated data collection methods. Concretely, three use cases are being investigated: the (1) density and distribution of recharging stations, the (2) availability and efficiency of public transport and (3) traffic and air quality. The data sources considered are to a large extend publicly available or crowd sourced. In a next step, Eurostat is developing the methodologies for producing indicators for each use case. All the technical implementation is made in the cloud based datalab which provides a flexible environment for prototyping new indicators. These indicators will for example be the maximum and average distance between recharging stations within a defined region; the percentage of a region's area and percentage of population that can be reached within a certain amount of time by public transport; and the average concentration of air pollutants during rush hour.

Keywords: Transport, Experimental, Mobility, Environment, Innovative data sources

1. Introduction

Eurostat is the statistical office of the European Union. Its mission is to provide high quality statistics and data on Europe. One of the core values to achieve this is to drive innovation in official statistics as well as to improve the quality, breadth, and depth of the statistics portfolio.

In recent years, the availability of new data sources and privately held data has increased exponentially and accelerated the transition towards using innovative data sources in the ambit of statistics. Correspondingly, Eurostat is currently exploring their use and developing methods to produce statistical indicators based on these - also in the ambit of transport statistics. In order to reap the benefits of the data revolution and to complement traditional statistical methods and sources with innovative approaches.

In a first step an exploratory study[1] was carried out to identify potentially suitable new/non-traditional data sources that can have application in the transport and mobility statistics domains. During this landscaping activity, data sources not necessarily coming from the

[1] Supported by a contract with GOPA.

transport sector were targeted, but rather sources which could be used to develop further harmonised transport statistics in several domains. First potential uses in official statistics were outlined. One of the areas highlighted was the use of Automatic Identification System (AIS) data for maritime traffic statistics. The paper "Early estimates of maritime traffic using innovative data sources" by Nikolaos Roubanis, Eurostat, and Boryana Milusheva, Eurostat, further describes the Eurostat's work in this ambit. Concretely, Eurostat and the European Maritime Safety Agency (EMSA) established a cooperation agreement to develop methods and produce early estimates of European ports vessel traffic, exploring the use of Automatic Identification System (AIS) and other administrative, and commercial data available in EMSA.

2. Identification of use cases

Building on the first landscaping activity, Eurostat proceeded[2] to the identification of concrete use cases. The aim being to develop methods to produce statistical indicators based on these; as well as implementing automated data collection methods in the cloud based datalab which provides a flexible environment for prototyping new indicators. In a first step the uses cases will be designed and set up i.e. piloted for a small number of MS (one to three) as to realise proof-of-concepts to produce experimental official statistics for mobility and transport. If successful, the uses cases will later on be scaled up to cover a greater number of MS. This is where the data lab environment for refining the methods and data processing pipeline infrastructure will come into play enabling and facilitating any needed adjustments and changes.

To assess potential use cases regarding their feasibility several criterions were used, both from an innovation/business as well as the IT perspective. From the business i.e. transport side the following criteria were used:

- Data availability - Is data directly available, how easily can it be accessed?
  Concretely, are APIs available? Is there a need for a data sharing agreement with the potential data provider? Are alternative sources of data available?
- EU coverage – data for how many Member States (MS) would it be available?
  Is representativeness a challenge?
- Data quality - Is the data reasonably harmonized (structure, frequency)?

[2] Supported by a contract with PwC.

Is the information available for a long enough time? Static or dynamic data? How fragmented is the data (i.e. what scale of aggregation/ cleaning/ harmonisation efforts are needed to make the data usable?)? Could pre-aggregated data be accessed?

- Data to statistics - How to go from data to statistics?

Can the data be used to support existing statistics or to create new experimental statistics?

Following these criteria potential indicators together with potential data sources were analysed and evaluated in terms of business feasibility and technical points of attention. Three use cases were settled on. These are:

-> Use case 1 – the density and distribution of public recharging stations based on crowdsourced and proprietary data in NUTS[3] regions (NUTS 2 and NUTS 3 regions).

-> Use case 2 – the availability and efficiency of public transport using official timetables/schedules and crowdsourced data in NUTS regions.

-> Use case 3 – Air quality traffic pollutants levels: the concentration level of air pollutants at rush hours and the number of km with high traffic and pollution level, using European Environment Agency (EEA) air quality datasets and TomTom speed profiles data.

3. Use case 1 – Recharging stations density and distribution

1.1 Context, indicators and data source

Given the increased importance of electro-mobility and the increasing share of electric vehicles, use case 1 focuses on the distribution of recharging stations, namely on the overall distribution as well as the distribution by category (i.e. maximum power output). The purpose is to examine within NUTS 2 or 3 regions the level of development for the network of recharging stations and identify potential "recharging deserts". Furthermore, the availability of these recharging stations monitored over time could be used to assess the actual usage of this specific type of alternative fuels.

In the context of the 9th Cohesion Report similar analysis have been carried out regarding the distribution of recharging stations such as the availability of nearby electric vehicle recharging points (within 10 km).

The following table shows the concrete indicators to be produced.

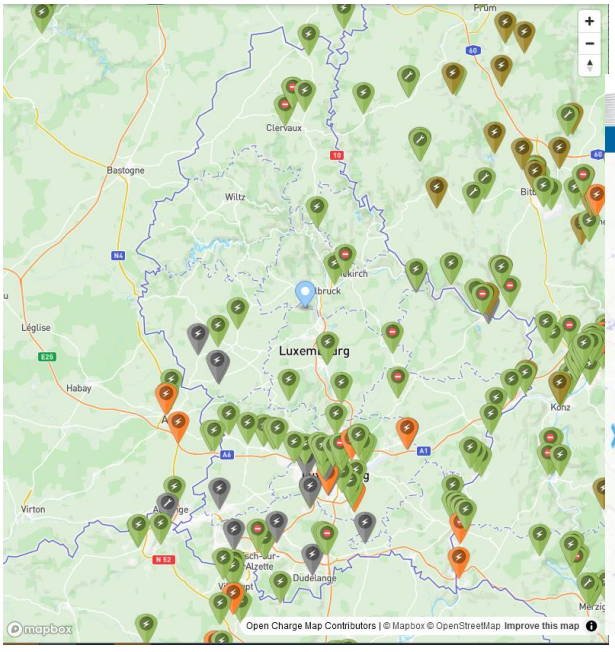
[3] The nomenclature of territorial units for statistics (NUTS classification) is a hierarchical system for diving up the territory of the EU. NUTS 1 regions are major socio-economic regions, NUTS 2 regions are basic regions and NUTS 3 are small regions.

Table 1. Overview of the selected use cases experimental statistical indicators to be produced

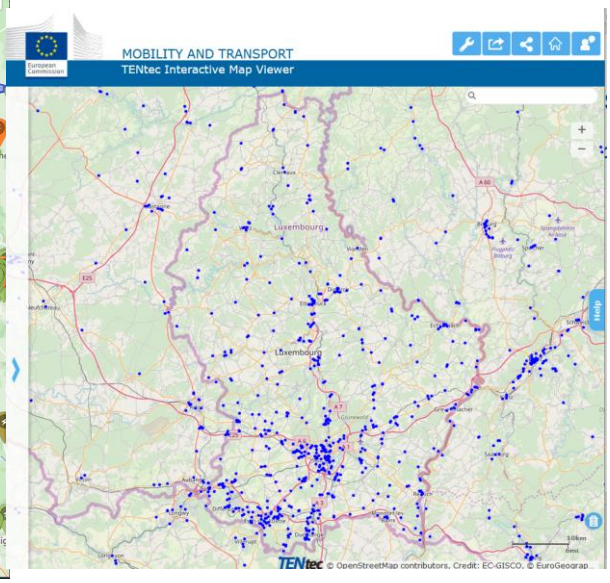
Use case 1 – Recharging stations density and distribution	
Indicator 1.1	Measuring the **recharging infrastructure density** through the number of recharging stations by NUTS region area
Indicator 1.2	Quantifying the **charging infrastructure network capacity** through the number of charging stations per category
Indicator 1.3	Determining the **charging infrastructure distribution** calculating the minimum or maximum, and average road distance between the nearest charging stations within a region

Regarding the data sources Open Charge Map (OCM) was identified as suitable given its coverage, the available granularity of data (here for example the type of recharging point) as well as the possibility to access the OCM POI API. Its crowdsource nature is also a plus. The data is to be complemented with Open Street Map (OSM) to enable routing calculations.

In the European Commission's TENtec Interactive Map Viewer Recharging Stations are visualized based on the European Alternative Fuels Observatory (EAFO)/ Eco-Movement data – a commercial data source.



Source: Open Charge Map



Source: EAFO/ Eco-Movement data visualized in TENtec Interactive Map Viewer

Figure 1: Comparison of potential data sources for recharging stations

A mere visual comparison for Luxembourg yields already significant differences – which upon further analysis became even more evident. Given the considerable differences in granularity it was decided to move forward with the data orchestration accounting for both data sources.

Another input data source for the setup is the NUTS raw data in order to attribute recharging stations by their coordinates to corresponding NUTS regions.

1.2 Methodology and data orchestration

A key classification are the categories of recharging points according to the Alternative Fuels Infrastructure (AFIR) Regulation – see table below.

Table 2. Recharging points categories according to the Regulation (EU) 2023/1804 on the deployment of alternative fuels infrastructure

Category	Sub-category	Maximum power output	Definition pursuant to Article 2 of this Regulation
Category 1 (AC)	Slow AC recharging point, single-phase	P < 7.4 kW	Normal power recharging point
	Medium-speed AC recharging point, triple-phase	7.4 kW ≤ P ≤ 22 kW	
	Fast AC recharging point, triple-phase	P > 22 kW	High power recharging point
Category 2 (DC)	Slow DC recharging point	P < 50 kW	
	Fast DC recharging point	50 kW ≤ P < 150 kW	
	Level 1 – Ultra-fast DC recharging point	150 kW ≤ P < 350 kW	
	Level 2 – Ultra-fast DC recharging point	P ≥ 350 kW	

For Indicator 1.3 the starting point is to calculate Origin—Destination pairs between recharging stations. Based on this the minimum, maximum and average distances between recharging stations within a region can be calculated. The travel time is computed using the R5 library.

The next step is to compute isochrone polygons from each recharging station to calculate what percentage of a NUTS region can be covered within a certain time frame (e.g. 10 Minutes) from any recharging station. This is done by combining the individual polygons into a single polygon that is used for area calculation.

By outputting also the greatest travel time to reach a recharging station this provides an answer to the question if there are any “recharging deserts” in a NUTS region.

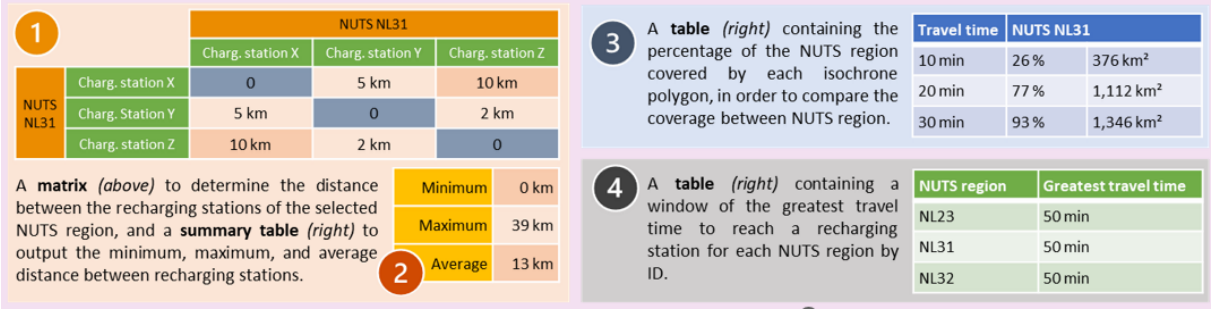


Figure 2: Components of indicator 1.3

1.3 Limitations and areas for further development

A concern in the current implementation is the needed computation time (for BE the required calculation time for indicator 1.3 took 10h for OCM data and 3 days for EAFO data). The needed precision e.g. for NUTS delimitation requires scale and scale -given the number of recharging stations- negatively affects calculation time. Indicator 1.3 furthermore requires a second data ingestion process, to retrieve road distance/ travel time between recharging stations.

Working with the r5r package also has certain other implications, for example:

- A subsetting of the OSM dataset is required. The chosen method retains complete multipolygons and complete ways, resulting in cases where the road network can extend outside of the NUTS polygon in consideration.
- The polygon outputs from r5r display snapping inconsistencies over large areas. Concretely, they tend to snap to the points in such a way that large areas are covered despite not having legitimate coverage, except for the points themselves. To mitigate this a limit was set at 60 min.
- Further scale and file specifications limitations.

On another note, the current data orchestration, i.e. data ingestion, transformation and indicator calculation could in future also be used for hydrogen infrastructure.

4. Use case 2 – Availability and efficiency of public transport

4.1 Context and indicators

Use Case 2 focuses on the availability and effectiveness of public transport. A starting point is Sustainable Development Goals (SDG) indicator SDG 11.2.1 “Proportion of population that has convenient access to public transport”, as well as the general development and the level of mobility that people can have. The availability and efficiency of public transportation are both valuable indicators for promoting alternative modes of travel. In particular, the number of stops

and lines available, the frequency and the connectivity of those stops are all valuable ways to assess the development and the performance of regional public transport network.

Similar analysis' have previously been done by Eurostat in the context of European Court of Auditor's report "Sustainable Urban Mobility in the EU: No substantial improvement is possible without Member States' commitment" as well as in a working paper by the Directorate-General for Regional and Urban Policy titled "How many people can you reach by public transport, bicycle or on foot in European cities? – Measuring urban accessibility for low-carbon modes".

The following table shows the concrete indicators to be produced.

Table 3. Overview of the selected use cases experimental statistical indicators to be produced

Use case 2 – Availability and efficiency of public transport	
Indicator 2.1	Measuring **availability of public transport** stops per NUTS2 and/or NUTS3 region, (via the average number of stops and lines serving these stops, the frequency, and the percentage of time during the day that public transport is available)
Indicator 2.2	Determining the **efficiency of public transport** via the percentage of region area and the percentage of population that can be reached in certain amount of time

Utilising public transportation networks and timetables information, the availability and efficiency of public transport will be assessed. Regarding the data sources both General Transit Feed Specification (GTFS) data as well as NetEx data is available. Overall, GTFS data is well documented and comes in the form of .txt files. While its EU coverage is not consistent (reported at country, city and/or provider level depending on MS) it is still more consistently available than NetEx. The GTFS data is complemented with OSM data allowing for navigation as well as the NUTS raw data and Eurostat 2021 Census population grid data.

4.2 Methodology and data orchestration

The accessibility of public transport will be examined through the number of stops, lines, and frequency of operation during the day, by NUTS 2 or NUTS 3 region. The efficiency of public transport will be assessed based on distance that can be travelled within a certain time frame from an origin centre in a NUTS 2 or 3 region, and the comparison with travelable distance by car, in the same given timeframe. Moreover, the population distribution is integrated in the efficiency of public transport calculation as percentage of population of a NUTS region reached - complementing the area reached.

In order to do so, first the stops, lines and departures need to be attributed to and aggregated by region. For the calculation of efficiency a uniform starting point needs to be chosen – the population centre of a NUTS region. This is determined by using a clustering algorithm comparing relative densities to determine the population centre. Concretely, the geometric centre of the population centre grid cell (1 km^2) is used as origin point for analysis of reach.

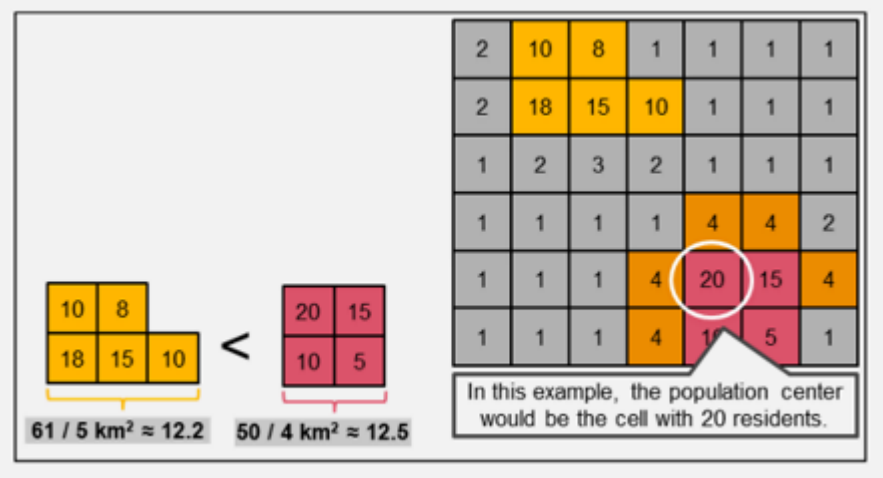


Figure 3: Visualisation of clustering algorithm

From the origin point, isochrone polygons are again calculated using the r5r package based on the OSM network enriched with the GTFS data. By dividing their area by the area of the NUTS polygon, the percentage of a NUTS region area reached within a certain travel time is calculated. And in a second step also the population reached within a set travel time is calculated (by intersecting the polygons with the population 1km^2 grid cells and calculating the sum of the pertinent grid cells and dividing it by the population of the NUTS region).

The same calculations are carried out using the car and both outputs compared.

ID	Isochrone	Public transport			Car		
		Area km²	% area	Pop. Reached	Area km^2	% area	Pop. Reached
SE11	120	1200000	60%	50%	1800000	80%	75%
SE11	90	900000	45%	40%	1400000	60%	55%
SE11	60	600000	30%	20%	800000	40%	35%
SE11	55	550000	27.5%	15%	600000	35%	30%

Figure 4: Example of the isochrone calculation (in min)

4.3 Limitations and areas for further development

The available GTFS data displays differing use of formats making extra data cleaning and treatment necessary before the data can be ingested. Furthermore, there are similar OSM and r5r-derived limitations as for use case 1.

With the addition of GTFS data, the departure time and place are now very important parameters for the calculation. This problem is two-fold: on one side a uniform departure time

needs to be defined to enable comparison (here 10am on a weekday). On the other side, from the chosen origin point (the geometric centre of the identified grid square from the EU Census data), r5r will first either “walk” or “transit” for its routing calculations. This means, that before using any public transportation, there is an inherent walking period that likely must occur and/ or a waiting time for the public transport, and this can be very dependent on the origin point, its nearest public transportation stop and the departure time. In consequence, this might result in XX minutes “spend” on a potentially inconsequential distance.

So, to summarize the results are very sensitive to

- the origin point chosen (geometric center of the population center grid cell)
- the origin time chosen (10am weekday).

Indicator 2.2 could furthermore be further specified by differentiating the mode used and generating isochrones by modes (e.g. train, bus,…).

5. Use case 3 – Air quality and traffic

5.1 Context and indicators

Use Case 3 aims to provide an alternative and automated assessment of the impact of mobility on air pollution, specifically in areas of high traffic density. Locations that have heavier traffic and congestion problems on roadways would likely have higher levels of traffic-related air pollutants.

Measuring these levels during recurrent and peak traffic hours can provide indicators on how traffic may influence pollutant levels. This will be assessed with traffic patterns taken into consideration, specifically peak hours.

The following table shows the concrete indicators to be produced.

Table 4. Overview of the selected use cases experimental statistical indicators to be produced

Use case 3 – Air quality and traffic	
Indicator 3.1	Measuring the **average concentration of air pollutant at rush hours** and the difference from a baseline value
Indicator 3.2	Quantifying the **number of kilometres with traffic and high air pollutant concentrations**

TomTom speed profile data is used to assess traffic behaviour, while EEA’s hourly air quality database is be used for concentration levels of various air pollutants and air quality measuring stations’ locations and characteristics (type and area). For the indicator development only so-called traffic air quality measuring stations – that are located in close proximity to a major road

and, therefore, influenced by the road traffic emissions- are used, to enhance robustness of the values obtained. This is combined again with the NUTS raw data.

The EEA air quality data -specifically, the nitrogen dioxide (NO_2) concentrations- are currently used in Eurostat's Recovery Dashboard to measure the changes in mobility pattern during and after the pandemic and their environmental effects. NO_2 specifically is highly linked to traffic - according to an EEA briefing "Sixty-four per cent of all NO_2 exceedances [of air pollution limits] reported were linked to emissions from road traffic".

5.2 Methodology and data orchestration

For the first indicator, a monthly baseline for average air pollutant concentrations per station needs to be calculated first. Then, in order to calculate the average air pollutant concentration during traffic and compute the difference to baseline, the roads around an air quality station need to be identified. For this a radius of a 100m is used to assure the representativeness of a traffic station. Then, the traffic on those roads needs to be inferred (traffic is assumed when there is a deviation in the TomTom speed profiles of <= 70% from free flow). For the detected rush hours the average air pollutant concentration is calculated and compared against the monthly baseline.

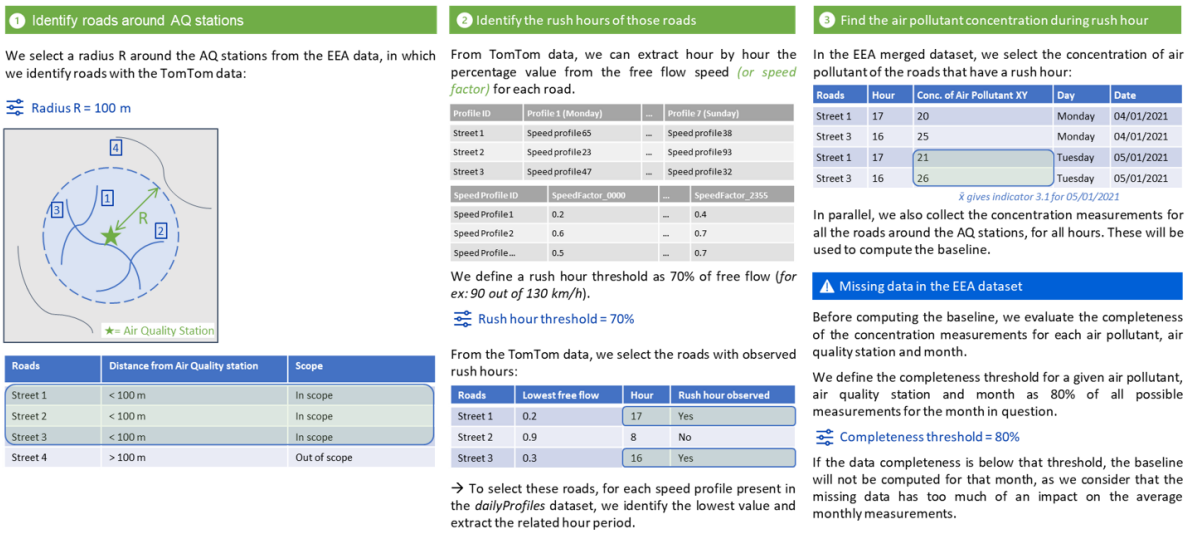


Figure 5: Detailed methodology to calculate average air pollutant concentration during traffic

For the second indicator the roads with observed rush hours are selected and narrowed down to those displaying at the same time an air pollutant concentration above the monthly baseline. Based on this, the sum of kilometres of roads and minutes spend in traffic around the air quality

stations are computed. This indicator can only be calculated if the completeness threshold for the baseline calculation is met. Given considerable data gaps, this is not assured.

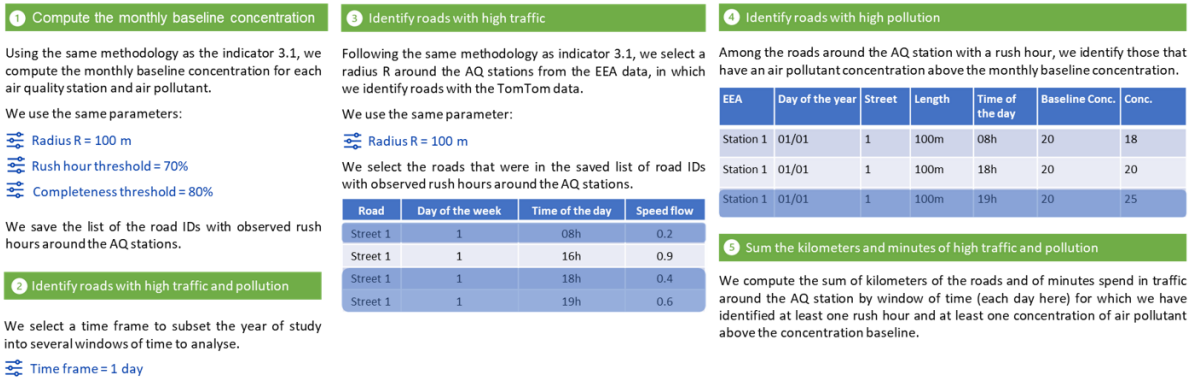


Figure 6: Detailed methodology to calculate kms and minutes of high traffic and pollution

5.3 Limitations and areas for further development

Air pollutant concentrations are sensitive to geographical and local conditions, including for example potential local restrictions for specific cars. The available data furthermore refers to pandemic times with unusual mobility patterns, rendering the interpretation more complicated.

The hourly EEA air quality data are validated annually. This annual revision of the source data brings risks of data incoherence throughout the year and could entail additional revisions of the indicators calculated in the previous periods.

A strong limitation are the considerable size of missing data in EEA data set, which is why a completeness threshold was implemented. Besides, the number of traffic air stations rather limited (11 for Belgium), rendering an aggregation of indicator 3.1 to e.g. city level currently not expedient and the calculation of indicator 3.2 often impossible.

6. Results and conclusion

Use case 1 has evidently shown that commercial data sets tend to have better and richer content than free data sets, but they also entail financial costs. Use case 2 has demonstrated the clear dependency of comparable results on harmonized input data as well as the implications of parameters assumed for comparability purposes. Overall, the use cases showcase how benchmarking of results and transparency of methods is key.

The use cases outlined above are currently being piloted for one to three MS and the data orchestration is being set up in the data lab. First preliminary outputs will need to be thoroughly assessed. Promising indicators could in a next step be presented to Eurostat's Methodology Advisory Committee for approval to be published as experimental statistics. In parallel or

thereafter, said promising indicators could be scaled up to test the implementation for more MS and -if needed- the corresponding methodology further refined and adjusted.

References

European Commission (2024), Ninth Report on Economic, Social and Territorial Cohesion, Publications Office, https://ec.europa.eu/regional_policy/sources/reports/cohesion9/9CR_Report_FINAL.pdf

European Commission, Eurostat (2024) Recovery Dashboard (now: European Statistical Monitor), https://ec.europa.eu/eurostat/cache/dashboard/european-statistical-monitor/

European Commission, Directorate-General for Regional and Urban Policy, Poelman, H., Dijkstra, L., Ackermans, L. (2020), How many people can you reach by public transport, bicycle or on foot in european cities? – Measuring urban accessibility for low-carbon modes, Publications Office, https://data.europa.eu/doi/10.2776/021137

European Court of Auditors (2020), Sustainable Urban Mobility in the EU: No substantial improvement is possible without Member States' commitment, Publications Office, https://op.europa.eu/webpub/eca/special-reports/urban-mobility-6-2020/en/

European Environment Agency, Air quality time series (E1a & E2a data sets), https://www.eea.europa.eu/data-and-maps/data/aqereporting-9/estatistics-e1a-e2a

European Environment Agency (2023), Briefing on managing air quality in Europe, https://www.eea.europa.eu/publications/managing-air-quality-in-europe/managing-air-quality-in-europe

Regulation (EU) 2023/1804 of the European Parliament and of the Council of 13 September 2023 on the deployment of alternative fuels infrastructure, and repealing Directive 2014/94/EU, http://data.europa.eu/eli/reg/2023/1804/oj

Roubanis, N., Milusheva, B., (2024) Early estimates of maritime traffic using innovative data sources.