SoftwareX 31 (2025) 102245

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original Software Publication

Population potential on catchment area (PPCA): A Python-based tool for worldwide geospatial population analysis

Joan Perez *, Giovanni Fusco

Université Côte-Azur, CNRS, ESPACE, Nice, France

ARTICLE INFO

Keywords:
Spatial analysis
Accessibility
Population
Catchment areas

ABSTRACT

The Population Potential in Catchment Area (PPCA) protocol is a Python-based methodology designed to evaluate and analyze population distributions within specified pedestrian catchment areas globally. PPCA utilizes OpenStreetMap (OSM) and Global Human Settlement (GHS) data and employs Google Earth Engine for data acquisition and morphometric analysis. Through a series of four automated steps, the protocol cleans, processes, and classifies geospatial data, ultimately yielding refined population estimations within defined catchment regions. This protocol enables researchers and urban planners to assess the population that can be potentially accessed on foot using the street network, within given distances. The protocol allows this assessment globally with minimal input requirements, focusing mostly on bounding box coordinates.

Metadata

Nr	Code metadata description	Metadata
C1	Current code version	v.1.0.5
C2	Permanent link to code/repository used for this code version	https://github.com/perezjoan/Population-Potential-on-Catchment-Area—PPCA-Worldwide
C3	Permanent link to reproducible capsule	https://github.com/perezjoan/Population-Potential-on-Catchment-Area—PPCA-Worldwide/blob/main/Case%20studies%20Coordinate%20Examples.txt
C4	Legal code license	Attribution-ShareAlike 4.0 International
C5	Code versioning system used	git
C6	Software code languages, tools and services used	python
C7	Compilation requirements, operating environments and dependencies	Python 3.9+, Google Earth Engine API authentication, OS: Windows/Linux/Mac
C8	If available, link to developer documentation/manual	https://github.com/perezjoan/Population-Potential-on-Catchment-Area—PPCA-Worldwide/blob/main/README.md
C9	Support email for questions	Joan.PEREZ@univ-cotedazur.fr

1. Motivation and significance

Population accessibility and distribution are central concerns in sustainable urban planning, particularly as cities strive to ensure equitable access to essential services and amenities. When working on international comparative studies, the challenge lies in integrating data from disparate geospatial sources to provide useful insights. On one hand, the Global Human Settlement (GHS) database offers a worldwide estimation of population distribution within 250 × 250 m grid cells, derived from satellite imagery, census data, and machine learning models [1]. On the other hand, OpenStreetMap (OSM) provides a crowdsourced, open-access geospatial database with detailed information on street networks, buildings, and land use, though it may contain inconsistencies in building classification and height attributes. These two databases offer the potential to produce meaningful international comparisons among cities. However, a key problem is how to distribute populations from the GHS grid to possibly incompletely described buildings within each cell and subsequently calculate accessibility to those populations on foot from locations along the street network.

The Population Potential on Catchment Area (PPCA) protocol was developed to address this challenge as it bridges the gap between coarse population estimates and the finer-scale spatial realities of urban environments, enabling a more detailed and accurate assessment of population potential along pedestrian networks. While many existing tools focus separately on population distribution or street networks, PPCA

* Corresponding author.
E-mail addresses: Joan.PEREZ@univ-cotedazur.fr (J. Perez), Giovanni.FUSCO@univ-cotedazur.fr (G. Fusco).

https://doi.org/10.1016/j.softx.2025.102245
Received 6 March 2025; Received in revised form 3 June 2025; Accepted 18 June 2025
Available online 28 June 2025

integrates GHS population grids with OSM building footprints and pedestrian networks, creating a direct link between population data and spatial networks. The protocol automates the scraping and integration of data from both GHS and OSM, a feature not commonly present in traditional urban analysis tools, which often require manual preprocessing and adjustments to fit different urban contexts. This automation allows the protocol to operate globally. Its adaptability to diverse environments, from dense city cores to sprawling suburbs, makes it a versatile asset for urban research and planning. Indeed, population accessibility analysis has particular relevance in the context of the "15-minute city" (15mC) concept, which envisions urban environments where essential services are accessible within a short walking distance [2,3]. Traditionally, the 15-minute city (15mC) has been framed as a policy ensuring maximum population coverage within a 15-minute walk of clustered activities [4]. An alternative perspective emphasizes the need to reach a population threshold within a 15-minute walk from main streets, making commercial activity viable along a network serving the entire urban area [5]. Regardless of the definition, population pedestrian sheds along the street network are essential for implementation of the 15mC concept. The PPCA protocol was developed to enable this calculation in an internationally comparable way.

2. Software description

The PPCA protocol provides a globally applicable methodology to evaluate and refine population distributions within pedestrian catchment areas using data from GHS and OSM. By automating data acquisition, preprocessing, and analysis, the protocol allows users to model population accessibility with minimal input while integrating morphometric and machine learning techniques. Unlike traditional methods requiring predefined datasets, the protocol automatically retrieves data via bounding box coordinates. Users do not need to provide external datasets, as the protocol integrates directly with OSM and GHS.

The PPCA protocol is hosted on a dedicated GitHub repository (see Metadata table). The repository includes information on how to set up a specific environment as it relies on several specialized libraries. Each Python script within the repository corresponds to a distinct step of the protocol. These scripts are modular, allowing users to run individual steps or the entire workflow, depending on their requirements. The protocol also requires a Google Earth Engine authentication with credentials.

2.1. Software architecture

The protocol is divided into 4 steps, each designed to handle a specific task: data acquisition, building classification, floor estimation, and population-over-catchment-area analysis. The four steps of the protocol are as follows: first, it collects and preprocesses geospatial data from OpenStreetMap (OSM) and Global Human Settlement (GHS) using bounding box coordinates, and filtering for key features. Second, it calculates morphometric indicators and uses a Decision Tree Classifier to label buildings as residential or non-residential. Third, it estimates the number of floor of residential buildings using OSM attribute data or predictive modelling when needed. Finally, it maps population estimates onto pedestrian networks to analyze density distribution across catchment areas. At the end of each step, the protocol generates output data in GeoPackage (GPKG) format. Users can also run subsequent steps independently, provided they have data formatted to match the GeoPackage standards produced by earlier steps (Figs. 1–3).

2.2. Software functionalities: data acquisition (step 1)

This step acquires data from OSM and the GHS database, ensuring the information is cleaned and structured for subsequent analysis. The following python libraries are employed to streamline the whole acquisition process: Earth Engine Python API (access to satellite imagery and geographical data [6]), QGIS (open-source Geographic Information System that integrates with Python) and OSMnx (acquisition of OpenStreetMap data [7]).

Users must then define input parameters, including bounding box coordinates, the coordinate reference system (CRS), and the desired year of analysis for the download of the GHS data. OSM data are retrieved as the most recent dataset available at the time of querying, typically reflecting real-time updates from contributors. Once configured, the process begins with the automated acquisition of GHS data. Raster files representing population density are downloaded and subsequently transformed into vector data using a QGIS pixel-to-polygon algorithm.

The protocol proceeds to extract features from OSM using the OSMnx library. This includes data on buildings, streets, and land use within the defined bounding box. Each dataset is cleaned to ensure consistency and usability. For buildings, structures with a footprint smaller than 15 m^2 are excluded, as are those with only underground levels or considered as light constructions. Streets are filtered to retain only those used by pedestrians excluding those with the following attributes: "motorway", "motorway_link", "trunk", "trunk_link", "busway", and "cycleway". In addition, streets are also excluded if they are in tunnels or if their

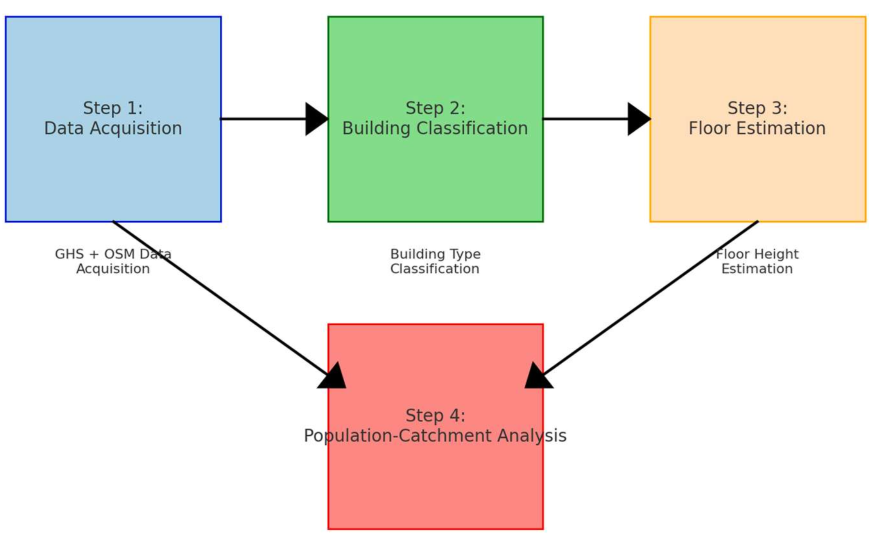

Fig. 1. PPCA Protocol Software Architecture.

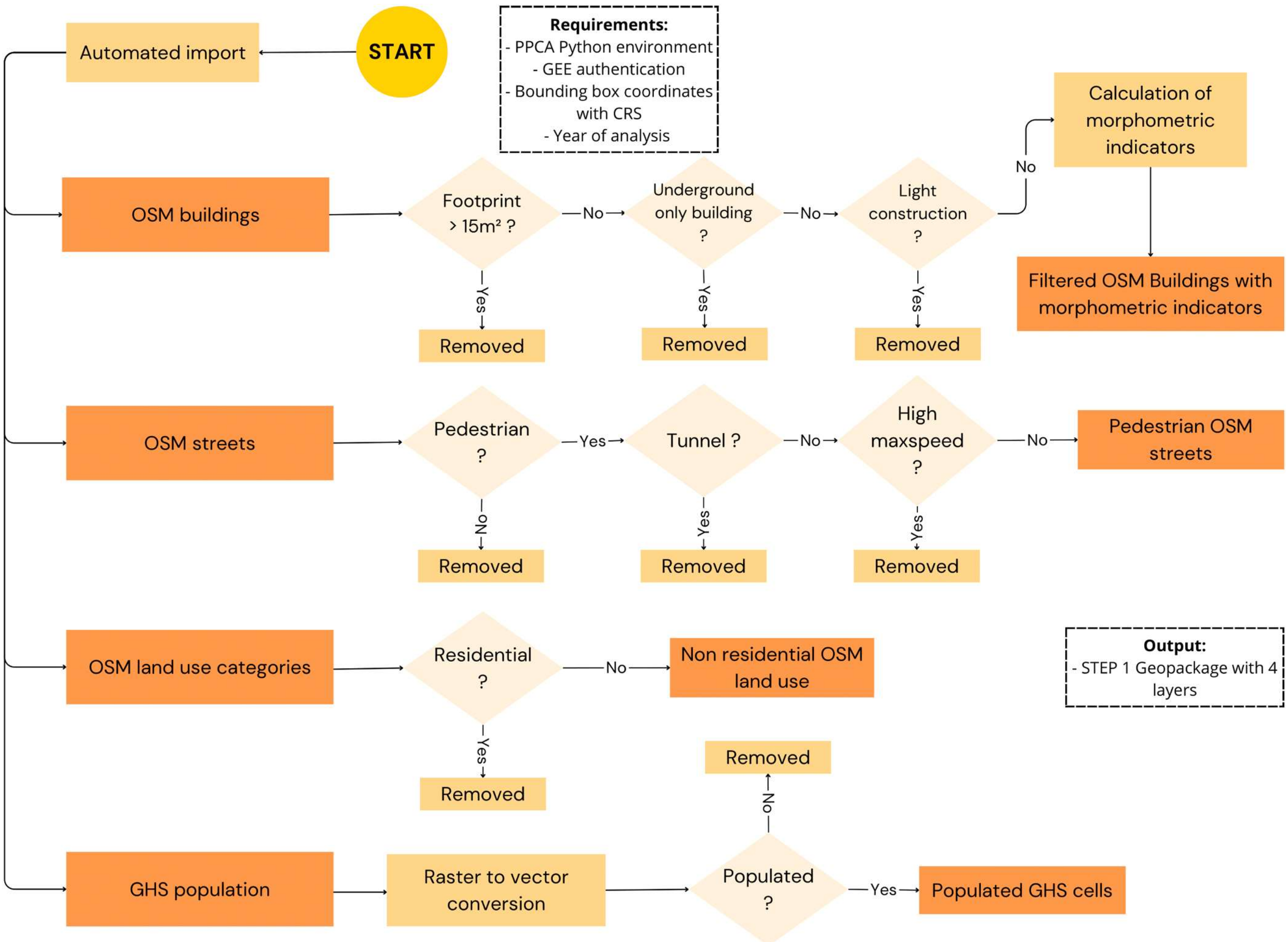

Fig. 2. Flowchart of Step 1.

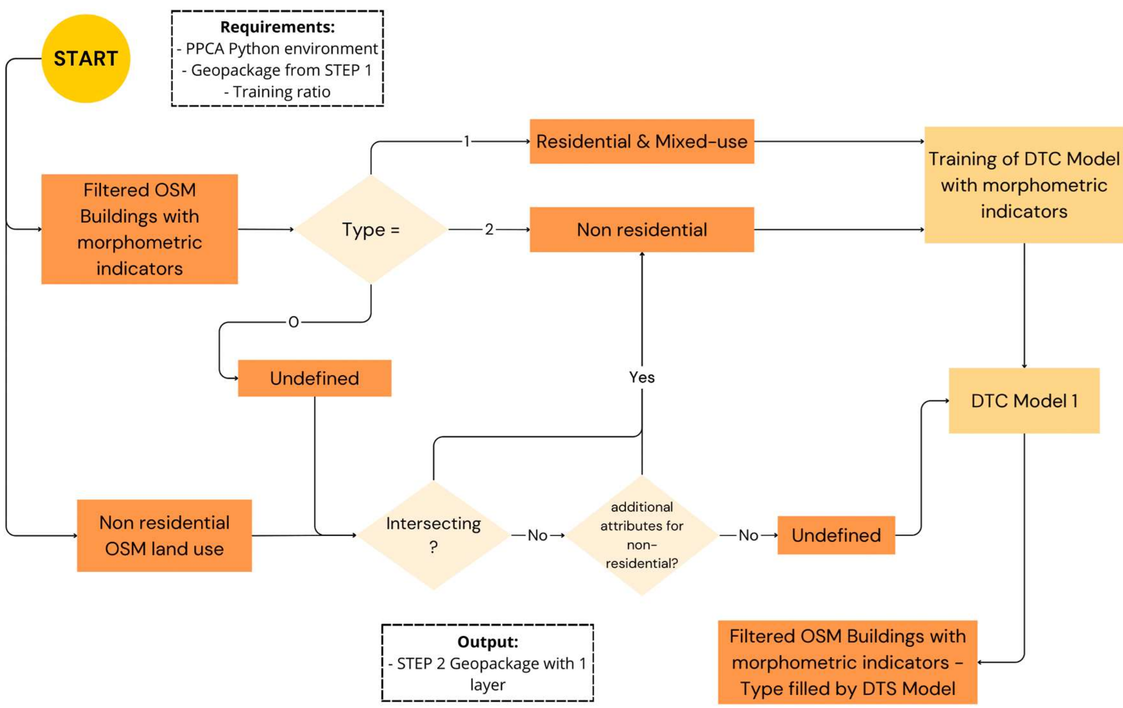

Fig. 3. Flowchart of Step 2.

maxspeed value is greater than or equal to 90 km/h. Land use data is refined to focus on populated categories, excluding industrial and other non-residential classifications based on a list of attributes ("construction", "cemetery", "education", etc.). These filters can be modified within the script by users according to their needs.

For the retained buildings, a series of morphometric indicators are calculated to characterize their geometric and spatial properties. These include FL (number of floors), A (surface area), P (perimeter), E (elongation), C (convexity), FA (floor area), ECA (a composite metric involving elongation, convexity, and area), EA (the elongation-area product), and SW (shared walls ratio). To compute these indicators, MomPy, a Python-based library specialized in urban morphological analysis, is utilized [8]. The output is saved in a standardized GeoPackage (GPKG) format for easy integration into GIS workflows and subsequent steps of the PPCA protocol. The final output consists of filtered and cleaned vector data for GHS population, OSM buildings, streets, and land use areas.

2.3. Software functionalities: building classification (step 2)

The second step involves classifying and predicting building types based on their attributes. This step uses a decision tree classifier to evaluate and fill missing building type information from the OSM building data that have been extracted and filtered in Step 1. Buildings are grouped into three initial categories based on OSM attribute data: residential or mixed-use (type = 1), non-residential (type = 2), and undefined (type = 0).

To address undefined building types, a two-phase classification approach is employed. The first phase applies rules based on OSM attribute data and spatial intersections with non-residential land-use data. Buildings classified as undefined but intersecting non-residential zones are assigned a non-residential type. The same is done for buildings inside residential areas but identified by OSM as non-residential through additional attributes (tourism, parking, shop, office).

Next, a decision tree classifier (DTC) is trained to model relationships between building characteristics and their types. The dataset is split into training and testing subsets. Residential and non-residential buildings serve as labeled data for training. Morphometric indicators are used as predictive features in training, while the target variable is the building type. The classifier is trained and evaluated using metrics of accuracy (i.e. how well the model predicts building types). Once trained, the model is applied to buildings with undefined types. Predictions are made based on the same attributes used during training. The output dataset contains the original characteristics of the buildings, the original building-types (including missing values), and the predicted types.

2.4. Software functionalities: floor estimation (step 3)

This step focuses on estimating the missing number of floors for buildings to recalculate the floor-area (already calculated in Step 1) without missing values. This is once again achieved by training a decision tree classifier on observed data to predict missing values for the number of floors. The process begins by ensuring that both "height" and "building:levels" are numeric and consistent. Missing values in "height" are filled by multiplying "building:levels" by 3 (assuming an average floor height of 3 m), and missing values in "building:levels" are filled by dividing "height" by 3. Records where the number of floors remains missing are isolated into another dataset, as highlighted in Fig. 4. A decision tree classifier is then trained using buildings with observed values for the number of floors. Key morphometric indicators A, P, E, C, ECA, SW and EA are once again used as features for training. The classifier is evaluated on a test subset of the training data to determine its accuracy. Once this second decision tree model is validated, it is applied to predict missing floor values for the isolated part of the dataset. The floor-area is then recalculated for all buildings.

2.5. Software functionalities: population-over-catchment-area (step 4)

The final step integrates population estimates into a pedestrian street network analysis, mapping population potential across catchment areas of different radii. The process begins by distributing population estimates from the GHS dataset to individual residential and mixed-use buildings. Building centroids are generated and spatially joined with the GHS population data to associate each building with the total population value of the cell. Population estimates are then distributed among buildings based on the ratio of each building's floor area to the total floor area within the same GHS grid cell, yielding an estimated population value for each structure.

These estimates are then integrated into a pedestrian network analysis, as detailed in Fig. 5. The methodology involves generating points at regular intervals along the retained streets, with an optional offset applied to the first point to prevent unnecessary superpositions at intersections. The interval length is customizable by the user, recommended values for pedestrian accessibility analysis range between 5 and

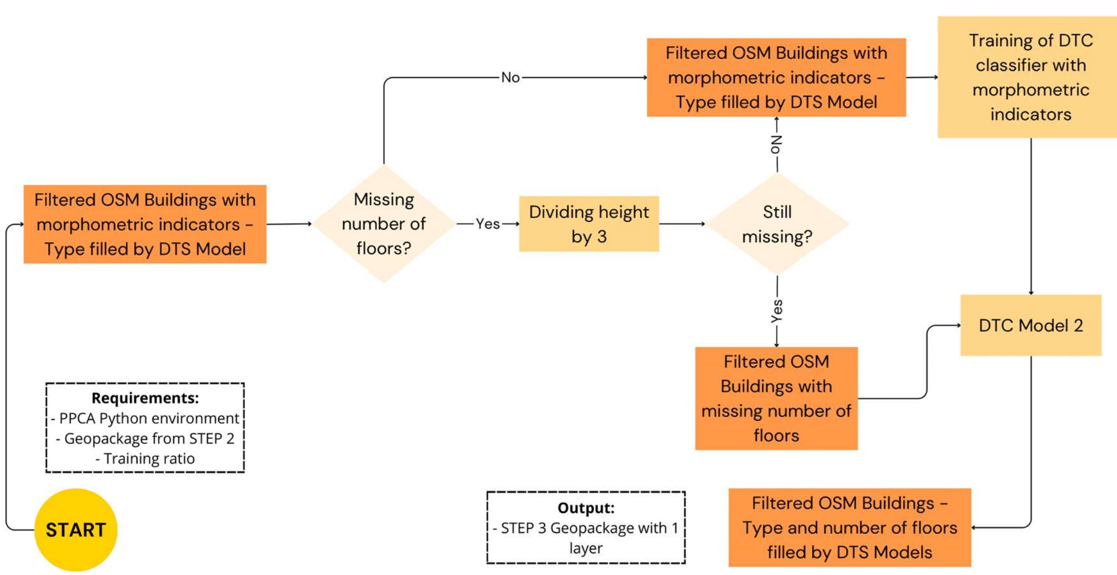

Fig. 4. Flowchart of Step 3.

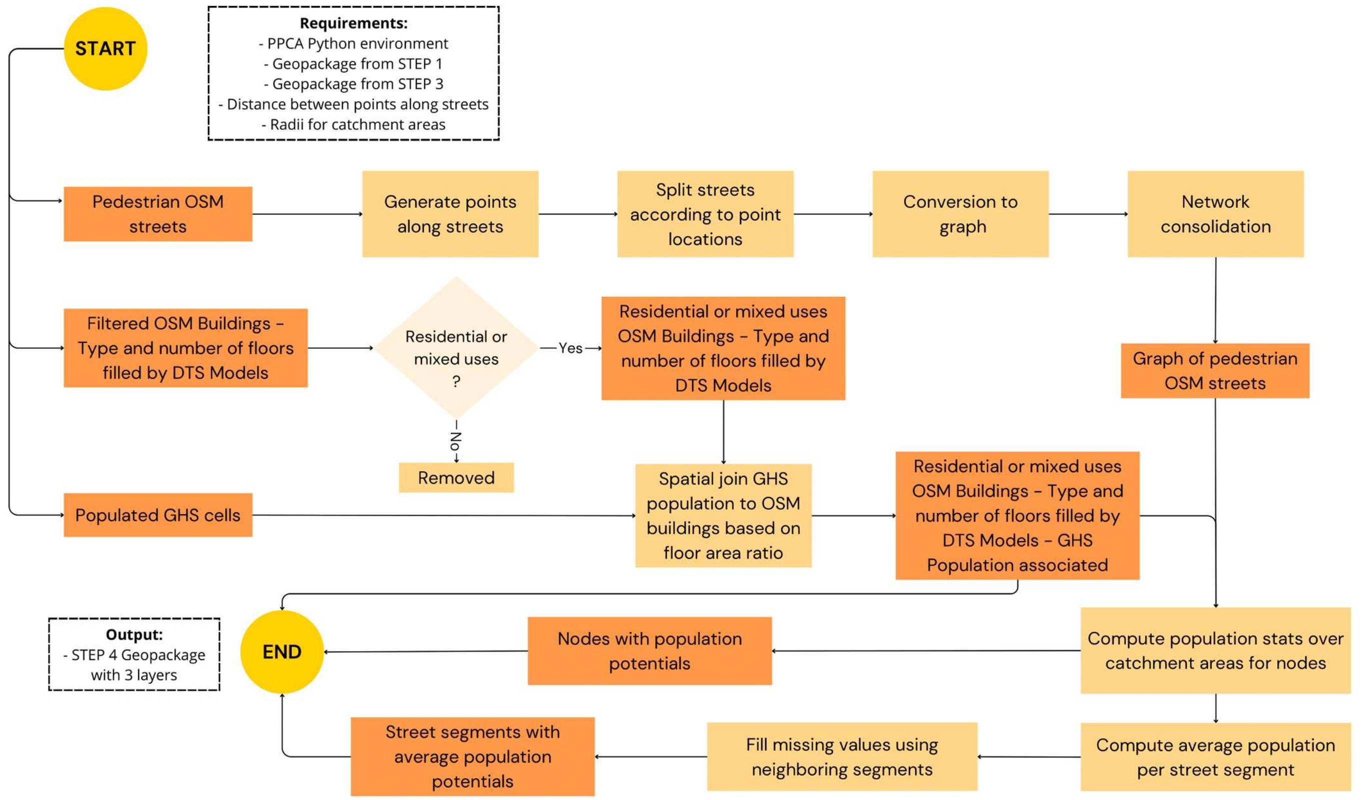

Fig. 5. Flowchart of Step 4.

50 m, with a trade-off between accuracy and computing time. A graph representation of the pedestrian network using the generated points as nodes is then constructed. NetworkX is leveraged through CitySeer, a spatial analysis library that enables network-based accessibility computations [9], to further refine the graph by consolidating nodes and edges to remove unnecessary complexities and redundancies. Buildings are assigned by CitySeer to the closest pedestrian node.

The population potential over catchment areas, parameterized by customizable radii (e.g., 400 m, 800 m, 1600 m), are then computed through the graph for each pedestrian node. An additional step refines the dataset by filling missing values in small or disconnected street segments using an averaging approach. The final output consists of three layers, including building population estimates, statistics at points along the pedestrian street network, and averaged population potential per pedestrian street segment.

3. Illustrative examples

The computational efficiency of the PPCA protocol was evaluated across different urban environments to assess its scalability and performance. Table 1 presents the running times for each processing step, measured in hours, across five European case studies: Nice, Vienna, Gothenburg, Versilia, and Lille with different numbers of pedestrian streets and residential and mixed-use buildings. The settings of the algorithms were the same for each case study: training ratio of 0.7, points generated along pedestrian streets every 20 m and catchment areas of 160, 400, 800 and 1200 m, corresponding to a 2, 5, 10 and 15 min walk, respectively (using a conventional walking speed of 4.8 km/h). The total processing time varies significantly depending on the number of buildings and pedestrian streets included in each case study.

The execution time for Step 1 is relatively low, ranging from 0.08 to 0.33 h. Step 2 follows a similar pattern, requiring only 0.02 to 0.09 h while step 3 exhibits slightly higher variability, with runtimes between 0.04 and 0.13 h. Step 4 represents the most computationally intensive phase, taking between 1.5 and 8.76 h, as it involves performing graph-based spatial computations. The optional Step 4 NA Value Filling, which averages missing values for small street segments, introduces an additional computational cost ranging from 0.6 to 16.27 h, depending on network density and missing data prevalence.

Table 1
Processing time per step for population potential analysis in various urban areas.

	Nice	Vienna	Gothenburg	Versilia	Lille
Filtered buildings	342,183	495,946	268,160	251,141	674,625
Missing type (%)	90.65	67.08	62.22	32.62	77.19
pedestrian street segments	184,319	386,963	373,066	97,547	210,136
Surface (sq km)	3504	3953	7996	1087	1795
STEP 1	0.157 hrs	0.334 hrs	0.212 hrs	0.082 hrs	0.291 hrs
STEP 2	0.029 hrs	0.068 hrs	0.021 hrs	0.021 hrs	0.095 hrs
STEP 3	0.041 hrs	0.083 hrs	0.043 hrs	0.044 hrs	0.131 hrs
STEP 4	3.836 hrs	8.761 hrs	4.670 hrs	1.504 hrs	3.248 hrs
STEP 4 Optional fill NA values	2.142 hrs	16.268 hrs	10.834 hrs	0.607 hrs	4.317 hrs
TOTAL	6.207 hrs	25.516 hrs	15.780 hrs	2.261 hrs	8.084 hrs

Overall, the total execution time varies from 2.26 h (Versilia) to 25.51 h (Vienna), demonstrating the scalability of the PPCA protocol. The results indicate that areas with a larger number of buildings and a denser pedestrian network require increased processing time, particularly in Step 4. The tests were performed on an Intel Core i9–10,900 K CPU @ 3.70 GHz with 128 GB RAM, ensuring consistent hardware conditions across all study areas.

Figures presented in this subsection are automated outputs provided by the algorithms. Fig. 6 (top) presents an example of Step 1 data

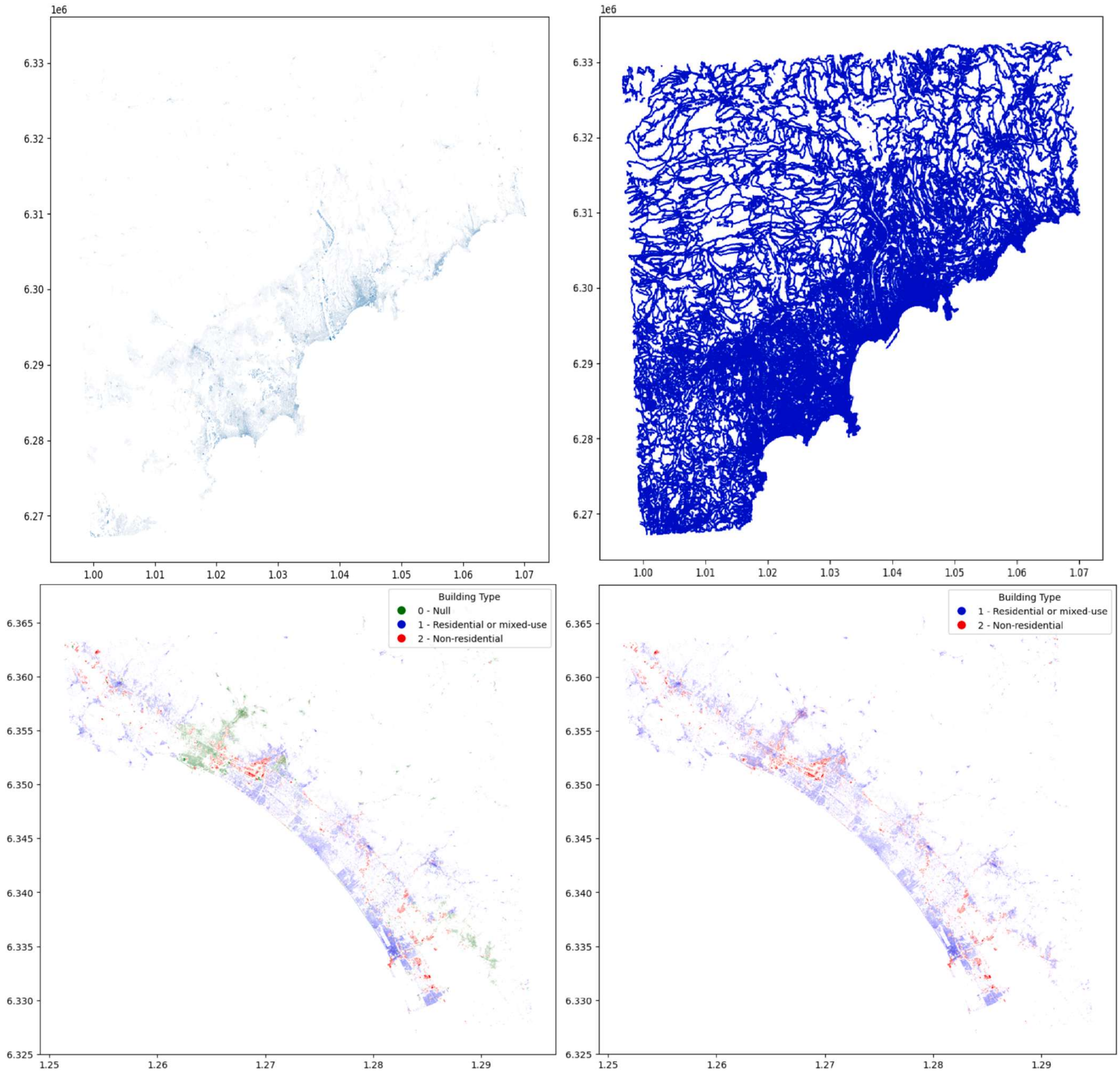

Fig. 6. Retained OSM Building Data and Extracted Pedestrian Street Network. Nice, Franc (top) & Building Classification in Versilia, Italy (bottom).

processing in Nice (France), comparing retained OSM building data (left) with the extracted pedestrian street network (right). The OSM dataset contains 342,183 building-related observations, after applying the filters of Step 1, while the pedestrian street network, filtered based on predefined attributes, includes 184,319 segments.

Fig. 6 (bottom) presents a comparison of building classification results before and after applying the first decision tree classifier to predict missing building types. The left panel illustrates the initial classification based on OSM attributes and land-use data, while the right panel displays the refined classification after implementing the DTC model. In the initial classification (left panel), 32.62 % buildings are still unclassified. Residential and mixed-use buildings (blue) represent 58.97 %, while non-residential buildings (red) make up 8.42 %. The high proportion of unclassified buildings highlights gaps in OSM attribute completeness. After applying the DTC, the model successfully assigns building types to previously unclassified structures, reducing the proportion of undefined buildings to 0 %. The updated classification indicates that 86.06 % are residential or mixed-use, while 13.94 % are non-residential. The overall accuracy reaches 83.6 % with a training ratio at 0.7, while the precision for residential and mixed-uses buildings reaches 90.6 %.

Fig. 7 visualizes the average accessible population per street segment within different catchment areas (160 m, 400 m, 800 m, and 1200 m) in Lille, France. As the catchment area increases, the estimated population coverage expands, leading to denser clusters of high-population areas. At 160 m, population concentration is highly localized. By 400 m, populated zones emerge, capturing areas with higher pedestrian accessibility in the compact city-centres. At 800 m and 1200 m, all urban and village centers as well as semi-central neighborhoods become distinctly highlighted. These results illustrate how walkable accessibility varies with scale.

GHS is a global population dataset, whose reliability depends on the census data used to populate remote imagery-derived built-up areas. In a

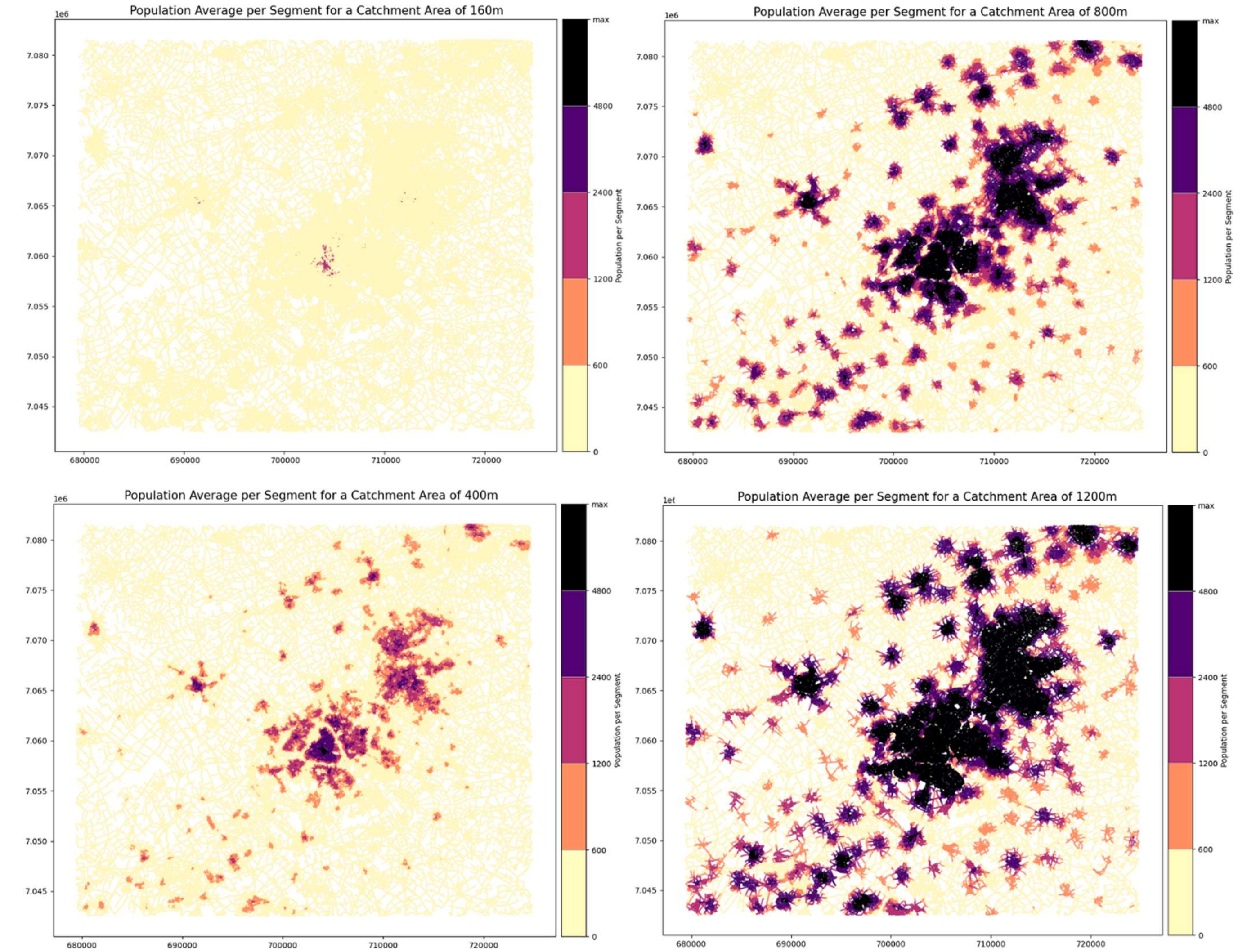

Fig. 7. Population Potential per Street Segment across Different Catchment Areas. Lille, France.

European context, it could be considered as relatively reliable. To assess this, we reran the entire PPCA protocol using Filosofi data as the population input and compared the results. Filosofi is an extremely reliable population dataset produced by INSEE (the French National Institute of Statistics and Economic Studies), offering demographic estimates on a 200 m × 200 m grid based on administrative and tax data. A comparative analysis between GHS population estimates and Filosofi data [10] has been carried out for Lille and Nice across different catchment areas (160 m, 400 m, 800 m, and 1200 m). The R^2 values increase with larger catchment areas, from 0.56 at 160 m to 0.91 at 1200m: Filosofi and GHS estimates align more closely at broader spatial scales. At smaller catchment areas, greater dispersion is observed, reflecting localized differences between the datasets. These results suggest that while GHS provides a reliable approximation starting from 400 m, its accuracy at finer resolutions requires careful interpretation, particularly when applied to high-resolution urban analyses.

4. Impact

PPCA is a fully open-source protocol, providing a powerful tool for urban analysts studying global urbanization and pedestrian accessibility within cities and designed to address the limitations of existing urban analysis tools by integrating population distribution with pedestrian accessibility in a unified framework. By applying it to five different European case studies, we demonstrated that its performance makes it suitable for metropolitan areas of varying scales, from smaller cities (e.g., the Versilia area, Italy) to larger urban regions (e.g., Vienna, Austria).

Supported by the European DUT initiative, the emc2 project applies the PPCA protocol to flexibly adapt the 15-minute city model for suburban areas. Emc2 focuses on transforming suburban corridors into compact, mixed-use networks that improve pedestrian connectivity and align services with population distribution [11]. By mapping accessibility at a fine spatial scale, the PPCA protocol helps in finding locations where a shift toward more sustainable suburban designs is required. However, PPCA can be adapted for other applications.

By reversing the accessibility logic, the protocol can evaluate people's access to urban resources such as retail, services, and green spaces within a given radius. Beyond pedestrian accessibility, cycling and car accessibility can be computed by modifying the input network and the corresponding speeds, while transit accessibility would require additional developments. This adaptability opens pathways for real-world applications, allowing policymakers to identify underserved areas, optimize public investments, and support data-driven urban planning decisions. For private developers, PPCA provides a reliable framework for assessing site accessibility, optimizing commercial site selection, and enhancing sustainable urban development projects.

5. Conclusions

The paper presented the specificities of the PPCA protocol and its relevance for analyzing pedestrian accessibility to populations within user-specified catchment areas in the context of international comparisons. By automating the integration of the globally available geo-databases GHS and OpenStreetMap, PPCA facilitates standardized, scalable assessments across different urban contexts, addressing the limitations of existing tools that often operate in isolation or require extensive preprocessing.

The implementation of PPCA in two French case studies demonstrated that GHS is a sufficiently reliable database for European contexts, particularly for catchment areas with a radius of 400 m or more. As for OSM, the PPCA protocol effectively compensated for incomplete building data by employing specifically trained decision tree classifiers.

Further enhancements could improve pedestrian accessibility modeling by incorporating private paths that connect buildings to the street network. This would be straightforward within PPCA, provided a more detailed pedestrian network is available. Alternatively, an estimated length for these paths could be integrated into the accessibility calculations in Step 4. Another valuable improvement would be the integration of a Digital Terrain Model (DTM) to account for street slopes and their impact on pedestrian speed. This would enhance the accuracy of accessibility calculations but would require specific software development and the use of a global DTM dataset such as Copernicus DEM or ALOS World 3D Additionally, the methodology developed for building classification and floor estimation using decision trees could be extended to fill other missing information from OSM.

CRediT authorship contribution statement

Joan Perez: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Giovanni Fusco:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization, Validation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Fusco Giovanni reports financial support was provided by Driving Urban Transition Partnership, co-funded by ANR (France), FFG (Austria), MUR (Italy), and Vinnova (Sweden) with contributions from the European Commission. Giovanni Fusco reports a relationship with Driving Urban Transition Partnership, co-funded by ANR (France), FFG (Austria), MUR (Italy), and Vinnova (Sweden) with contributions from the European Commission. that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the emc2 project under the Driving Urban Transition Partnership, co-funded by ANR (France), FFG (Austria), MUR (Italy), and Vinnova (Sweden) with contributions from the European Commission.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.softx.2025.102245.

References

[1] Melchiorri M, et al. Unveiling 25 years of planetary urbanization with remote sensing: perspectives from the global Human settlement layer. Remote Sens 2019; 11(12):296.
[2] Moreno C, et al. Introducing the "15-minute city": sustainability, resilience and place identity in future post-pandemic cities. Smart Cities 2021;4(1):93–111. https://doi.org/10.3390/smartcities4010006.
[3] OECD. Cities policy responses. OECD; 2020. Available at: https://www.oecd.org/coronavirus/policy-responses/cities-policy-responses-fd1053ff/.
[4] EIT Urban Mobility. ±15-Minute city: human-centred planning in action. TU München; 2022. www.eiturbanmobility.eu/wp-content/uploads/2022/11/EIT-UrbanMobilityNext9_15-min-City_144dpi.pdf.
[5] Fusco G, Berghauser Pont M, Cutini V, Psenner A. Guiding principles for the 15-minute city in peripheral areas: the emc2 model. dir.. In: Cremaschi M, editor. Game changer? Planning for just and sustainable urban regions, AESOP 2024 Proceedings, Paris, July 8th-12th 2024; 2024. p. 690–707. https://hal.science/hal-04798781.
[6] Gorelick N, et al. Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ 2017;202:18–27.
[7] Boeing G. OSMnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput Environ Urban Syst 2017;65:126–39.
[8] Fleischmann M. momepy: urban morphology measuring toolkit. J Open Source Softw 2019;4(43):1807.
[9] Simons GD. The cityseer Python package for pedestrian-scale network-based urban analysis. Environ Plann B 2023;50(5):1328–44.
[10] INSEE. Carroyé 2019 - Population Grid Data. 2023. Available at: https://www.insee.fr/fr/statistiques/7655475?sommaire=7655515 (Accessed 15 January 2025).
[11] Fusco G, et al. Guiding principles for the 15-minute city in peripheral areas: the emc2 model. dir.. In: Cremaschi M, editor. Proceedings of AESOP annual congress. Sciences Po – AESOP; 2024. p. 690–707. July 8th-12th 2024, https://hal.science/hal-04798781.