

Reference Data Selection

Technical Note D3

Krasen Samardzhiev, Barbara Metzler, Martin Fleischmann, Dani Arribas-Bel

Table of contents

1	Executive summary	1
2	Moprhological data	1
2.1	Ground truth classification	1
2.2	Building footprints	3
2.2.1	Available sources	4
2.2.2	Microsoft building footprints	5
2.3	Street data	6
2.3.1	Overture transportation layer	6
3	AI data	7
3.0.1	Satellite imagery	7
3.0.2	Urban fabric classes	8

1 Executive summary

Summary goes here.

2 Moprhological data

2.1 Ground truth classification

The main aim of this project is to create a predictive model that is capable of classifying urban fabric. To achieve this we need to use existing urban fabric classifications as ground truth data for model training. We have chosen the (Fleischmann and Samardzhiev Forthcoming) classification to act as the ground truth, since it the only avaiable classification that fully

covers our study area and is based on the highest quality data available - building cadastre data from official sources. Since (Fleischmann and Samardzhiev Forthcoming) is upcoming work we describe its methodology and results in detail in this and in the Algorithm Design and Theoretical Basis Description note.

(Fleischmann and Samardzhiev Forthcoming) extends the work of (Fleischmann et al. 2022), which has been used in numerous subsequent studies and official planning documents (Calafiore et al. 2023; Arribas-Bel and Fleischmann 2022; K. Samardzhiev et al. 2022) . It uses a combination of buildings and streets to generate Enclosed Tessellation Cells (ETCs), which are spatial units that are centred around buildings, and encapsulate the surrounding space that is nearest to them, rather than other buildings. ETCs boundaries are also limited by streets or a preset buffer distance from the centre of the building, which acts as the centre of the ETC. It is from this combination of buildings, streets and space, that the ETC can combine morphological characteristics of both streets, street intersections, buildings and blocks. The next step in Fleischmann and Samardzhiev (Forthcoming) is to measure different characteristics (or chars) of the ETCs. A literature review carried out by (Measuring urban form) suggests that there are around 74 unique ones that take into account the two dimensional nature of the data. (Fleischmann et al. 2022) use the full array of characters since they don't know a priori which will be the important variables. (Fleischmann and Samardzhiev Forthcoming) uses a subset of 54 and the detailed list of characters is presented in the Algorithm design technical note, since both our method and (Fleischmann and Samardzhiev Forthcoming) share the same characters. The next stage in the (Fleischmann and Samardzhiev Forthcoming) process is applying two clustering algorithms to generate a classification of ETCs into urban Similar to (Fleischmann et al. 2022) the Primus clustering approach is split into two parts - defining morphotopes and generating a taxonomy of morphotopes based on their similarity.

(**citation?**) defines morphotopes as the ‘smallest urban localities obtaining distinctive character among their neighbours from their particular combination of characteristics of plan type, building type and land use’. Primus follows this defining and operationalises a morphotope as a set of spatially contiguous ETCs with a minimum of 100 elements, where the internal distance calculated based on the morphological characters is on average lower within the set, than to other sets. The morphotopes for every region are calculated in isolation and in parallel. The definition of the subregions guarantees that all elements that are spatially contiguous, or have any spatial leakage effects, i.e. adjacency or being part of the same street segment, are processed together in the same region. Once the morphotopes are delineated, the mean of the characters of their constituent ETCs are assigned them. Therefore, every morphotope has the same dimensionality as the ETCs that form it.

The second step is the creation of a taxonomy of morphotypes for the whole of Central Europe. This is achieved through the application of non-spatially restrained ward clustering at two stages - one regional and one global. The resulting hierarchy can be cut at different levels depending on the application. For example at low values, there is some regional information still preserved in the hierarchy - morphotopes from one city are merged to identical morphotopes from the same city. Whereas at higher levels, different types of urban fabric, emerge

across countries.

The figure below shows a specific cutoff value, which separates different types of houses; from heterogenous historical urbanised areas; from recent modern urban developments such as apartment blocks and commercial areas; from large industrial areas.

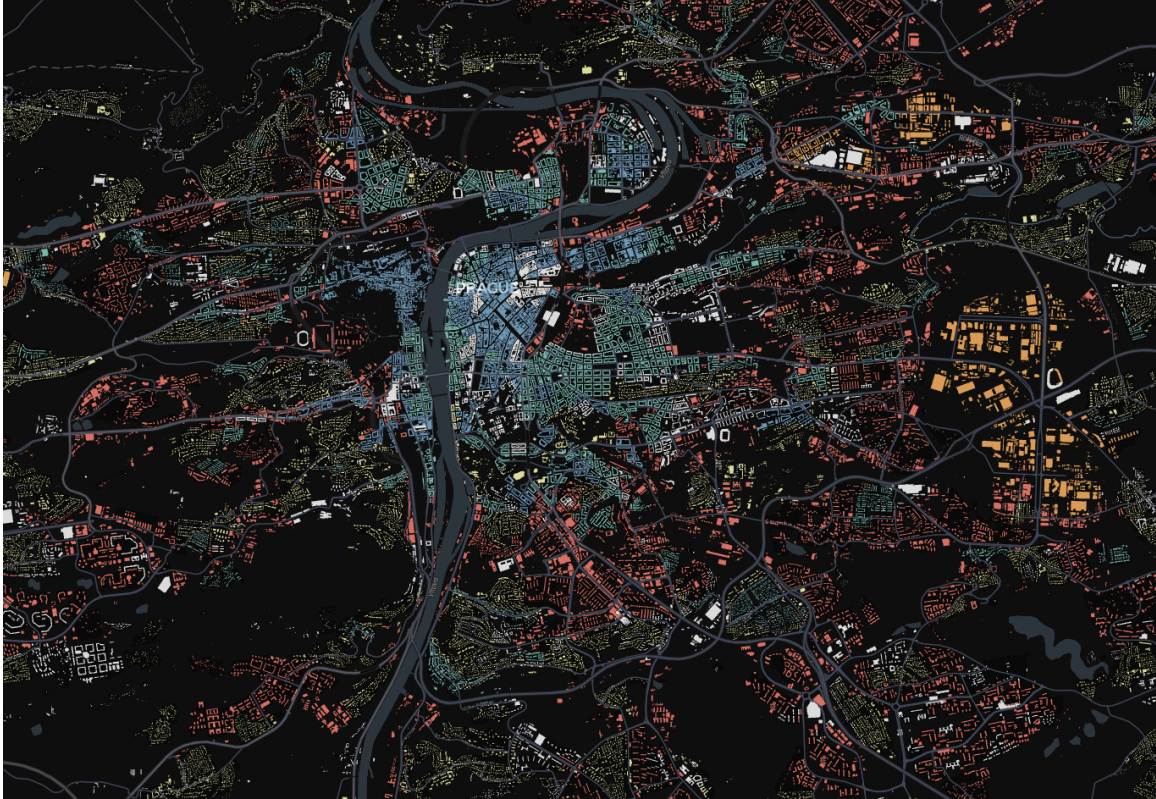


Figure 1: High-level urban fabrics

More detailed information about the primus methodology is available in the algorithm-design document, as we aim to follow it as closely as possible in this project.

2.2 Building footprints

Building footprints are one of the two core elements needed to generate morphometric classifications. Measurements of aspects of buildings provide the bulk of the information about the morphological character of a place. Many characters in the literature (Fleischmann, Romice, and Porta 2021), for example the ratio of perimeter shared with other connected buildings, require very high quality building footprints to accurately reflect the reality of the underlying topology. There are numerous available source of building footprints - official cadastre

polygons; volunteer made OpenStreetMap data; Satellite-derived building footprints; or combinations of all of these such as Overture or EUBUCCO. Each of these data sources come with their own drawbacks and advantages.

2.2.1 Available sources

The highest quality available data is from official cadastre building footprints. These provide the ground truth information for official planning and administrative applications in many countries. However, the data is not always readily available and even when it is, it requires a lot of preprocessing. For example, there is no readily available cadastre data to download for EU countries such as Hungary and Bulgaria. Even though Data on Germany is available, it is split across 16 Bundesregions, each of which has its own interface and data format. Furthermore, the data is not always comparative between countries reflective of the physical reality - for example in Czechia the official cadastre reflects shared administration in some modernist apartment blocks, rather than the physical structure of the buildings.

Another source is OpenStreetMap map, which is free and open geographical database generated by volunteers. In many cases the cadastre data is processed and put into OpenStreetMap, alongside volunteer contributions where official data is missing. The issues with open streetmap data are similar to cadastre data where the cadastre data is used as the basis and is even harder to process since the schema used for buildings is more generic. Furthermore, the quality is unequal - for example the data in Germany is very high quality, but not in other countries like Italy or Bulgaria.

Another available data source is satellite-derived building footprints, which are becoming more widely adopted due to advances in satellite technology and computer vision algorithms. They have successfully been used in applications for city boundary delineation, segregating population and others (K. P. Samardzhiev 2023). There are multiple sources of global building footprints available - Google building footprints, Microsoft building footprints, ESRI community data, Overture maps, as well as other proprietary data. The advantage of satellite-derived building footprints data is that it is all derived in a consistent manner and covers large geographical areas. However, the quality is generally worse than other sources and the geometric topology is typically broken.

The newest available data, such as EUBUCCO or OvertureMaps combine different types of other sources. Overture is based off of OpenStreetMap, but adds a hierarchical structure of sources on top of it. The buildings from Openstreetmap are considered the ground truth and buildings from additional sources such as satellite data are added where there is no open-streetmap data. This means that in general it has the highest number of available buildings. However, it comes with the same issues that OpenStreetMap data has, and it even further exacerbates the data heterogeneity issues - in different places different data sources with different characteristics are merged. Thus, working with the data requires more preprocessing than other resources.

Similarly, EUBUCCO is another data source that combines multiple datasets - official cadastre data and OpenStreetMap. Therefore it has the same advantages and disadvantages of its core sources. Furthermore, it has missing data issues that stem from the specific fusion approach of these sources. For example, there is a whole missing neighbourhood in central Prague and, in general, it has less buildings than the official cadastre data for the Central European study area.



Figure 2: Missing data in central Prague

2.2.2 Microsoft building footprints

In this project, we use the Microsoft building footprints since they fully cover our study area - Poland, Czechia, Germany, Austria, Slovakia (i.e Central Europe). The dataset also has good worldwide coverage - they contain around 1.5 billion building polygons - and the footprints are derived in a consistent way, which ensures scalability of our method to different countries. It should be noted that our approach can also work for other consistently derived building polygons, Google Footprints for example, so long as all the data is generated from the same source.

However, the data does not come without issues. For example, in dense urban centres entire blocks can be delineated as individual buildings. Given that morphology calculations rely on precise local topological relations between neighbours, such as two buildings touching, this problem renders a whole number of possible measurements described in (Fleischmann, Romice, and Porta 2021) meaningless. Furthermore, this issue affects even simpler calculations such as counting the number of buildings within a radius or topological neighbourhood. Other issues are that computer vision techniques sometimes miss entire buildings or misidentify building boundaries. Therefore, any approach that uses satellite-derived building footprints should be able to account for these three and potentially other problems.

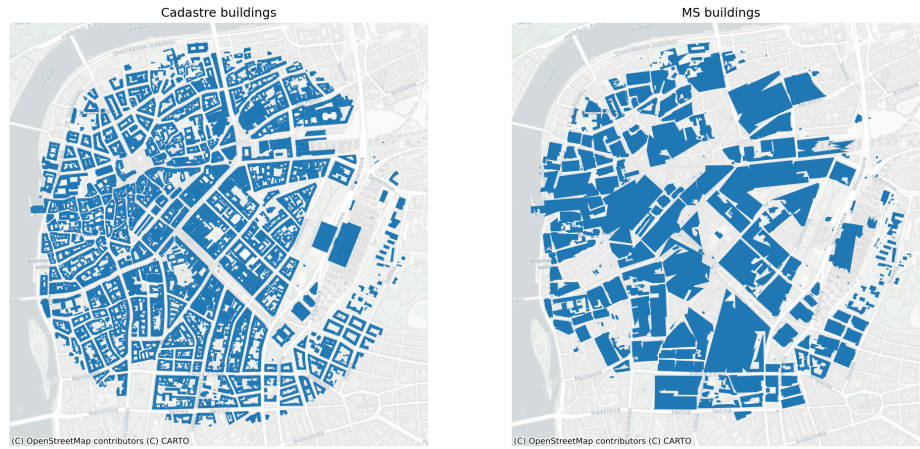


Figure 3: Comparison between MS buildings and cadastre level buildings in central Prague

2.3 Street data

Street network data is the other major data source used in numerical morphometric analysis. It is used to identify spatial relationships, delineate boundaries for other elements and calculate morphological features of the streets themselves, such as length or width. There are numerous sources for street network data, however the major one is OpenStreetMap and its derivative - OvertureMaps.

2.3.1 Overture transportation layer

The street network used in this study is a direct download from Overture maps, a processed subset of data from OpenStreetMap with some additions. OpenStreetMap has global coverage

and high quality data, especially for vehicular roads, on actual physical layouts of the roads, as well as additional information on street type. We use Overture directly due to the fact that its more accessible and easier to process, however the underlying geometry and information of the two data sources is identical.

The types of street segments used in the analysis are limited to the ones listed in Table 1.

<i>Types of streets</i>
living_street
motorway
motorway_link
pedestrian
primary
primary_link
residential
secondary
secondary_link
tertiary
tertiary_link
trunk
trunk_link
unclassified

Another type of segment that is filtered out are tunnels , since the analysis strictly focuses on two dimensions and therefore undergrounds structures adversely affect the calculation of boundaries and characters.

3 AI data

3.0.1 Satellite imagery

We use satellite images as input into the AI model to predict urban fabric. The satellite image data is sourced from the GHS-composite-S2 R2020A dataset (Corbane et al. 2020). This dataset is a global, cloud-free composite derived from Sentinel-2 L1C data, covering the period from January 2017 to December 2018. The composite images have a resolution of 10 meters per pixel and include three visible bands (red, green, and blue), which are well-suited for urban analysis.

Sentinel-2 data is particularly suitable for this application for several reasons: - Coverage: It provides consistent, openly available data for the entire Earth, making it ideal for projects requiring cross-regional comparisons. - Temporal flexibility: The composite includes data from

multiple time points, enabling us to analyze urban patterns over time. - Resolution: At 10 meters per pixel, it offers the highest resolution among openly available datasets, sufficient to distinguish urban features at the neighborhood scale. - Model compatibility: Many machine learning models in geospatial AI are pre-trained on Sentinel-2 data, ensuring compatibility with existing methods and facilitating transfer learning.

3.0.2 Urban fabric classes

The urban fabric predictions from our AI model are based on labels from the Spatial Signatures Framework (Arribas-Bel and Fleischmann 2022; Fleischmann and Arribas-Bel 2022). This framework provides a typology of British urban environments, characterized by both form (physical structure) and function (usage). It captures the complexity of urban areas, offering insights into how different spaces look and operate. While our primary interest is in urban fabric classification focused on form — an approach we expect may be simpler since form is visible in imagery — this classification scheme is still under development. Therefore, we currently use the more comprehensive Spatial Signatures Framework as a proxy, as it aligns with the goals of urban characterization in our project.

- Arribas-Bel, Daniel, and Martin Fleischmann. 2022. “Spatial Signatures-Understanding (Urban) Spaces Through Form and Function.” *Habitat International* 128: 102641.
- Calafiore, Alessia, Krasen Samardzhiev, Francisco Rowe, Martin Fleischmann, and Daniel Arribas-Bel. 2023. “Inequalities in Experiencing Urban Functions. An Exploration of Human Digital (Geo-) Footprints.” *Environment and Planning B: Urban Analytics and City Science*, 23998083231208507.
- Corbane, C., P. Politis, P. Kempeneers, D. Simonetti, P. Soille, A. Burger, M. Pesaresi, F. Sabo, V. Syrris, and T. Kemper. 2020. “A Global Cloud Free Pixel- Based Image Composite from Sentinel-2 Data.” *Data in Brief* 31 (August): 105737. <https://doi.org/10.1016/j.dib.2020.105737>.
- Fleischmann, Martin, and Daniel Arribas-Bel. 2022. “Geographical Characterisation of British Urban Form and Function Using the Spatial Signatures Framework.” *Scientific Data* 9 (1): 546.
- Fleischmann, Martin, Alessandra Feliciotti, Ombretta Romice, and Sergio Porta. 2022. “Methodological Foundation of a Numerical Taxonomy of Urban Form.” *Environment and Planning B: Urban Analytics and City Science* 49 (4): 1283–99.
- Fleischmann, Martin, Ombretta Romice, and Sergio Porta. 2021. “Measuring Urban Form: Overcoming Terminological Inconsistencies for a Quantitative and Comprehensive Morphologic Analysis of Cities.” *Environment and Planning B: Urban Analytics and City Science* 48 (8): 2133–50.
- Fleischmann, Martin, and Krasen Samardzhiev. Forthcoming. “Numerical Taxonomy of Urban Fabric in Central Europe,” Forthcoming.
- Samardzhiev, Krasen Petrov. 2023. “Characterising Urban Processes Using New Forms of Data and Analysis.” PhD thesis, University of Liverpool United Kingdom.

Samardzhiev, Krasen, Martin Fleischmann, Daniel Arribas-Bel, Alessia Calafiore, and Francisco Rowe. 2022. “Functional Signatures in Great Britain: A Dataset.” *Data in Brief* 43: 108335.