Geographical Characterisation of British Urban Form and Function using the Spatial Signatures Framework

Martin Fleischmann[1,*] and Daniel Arribas-Bel[1]

[1]Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Roxby Building , 74 Bedford St S , Liverpool , L69 7ZT, United Kingdom
[*]corresponding author(s): Martin Fleischmann (m.fleischmann@liverpool.ac.uk)

ABSTRACT

This is a manuscript template for Data Descriptor submissions to *Scientific Data* (http://www.nature.com/scientificdata). The abstract must be no longer than 170 words, and should succinctly describe the study, the assay(s) performed, the resulting data, and the reuse potential, but should not make any claims regarding new scientific findings. No references are allowed in this section.

Background & Summary

How the buiding blocks that make up cities are spatially arranged is worth quantifying and understanding. By "building blocks", we mean both the activities and agents that inhabit cities, as well as the (infra)structure that supports them. The former can be conceptualised as *urban function*, while the latter falls under the study of *urban form*. Understanding urban form and function is important for two main reasons. First, the combination of both *encodes* much information about the history, character and evolution of cities. For example, the shape and properties of the street network encode the technology of the time (e.g., automobile); while the degree of mix in land uses can reflect cultural values. Second, the spatial pattern of urban form and function also acts as a frame that *influences* a variety of outcomes, from economic productivity to socio-economic cohesion to environmental sustainability.

In this paper, we use the Spatial Signatures framework[1,2], which develops a "characterisation of space based on form and function designed to understand urban environments"[1]. Spatial Signatures are theory-informed, data-driven computable classes that describe the form and function of a consistent patch of geography. Figure 1 presents an overview of the development of a spatial signature classification. We build a series of enclosures that we combine with building footprints to further subdivide geographical space. We then attach form and function characters to each of these subdivisions, and use those to group them into consistent and differentiated classes we call signatures. Each phase is expanded in detail in the next section.

We introduce an open data product (ODP[3]) containing a classification of spatial signatures for Great Britain. In doing so, we provide an analysis-ready layer that brings urban form and function consistently, in detail, and at national scale. To the best of our knowledge, this is the first dataset capturing urban form and function published both with a degree of detail and scale as ours. Our results are based on the analysis of more than 14 million of ET cells, to each of which we attach more than 300 characters capturing a wide range of aspects relating to urban form and function. We provide access to both granular geographical boundaries of the delineated spatial signatures as well as measurements for each character at the signature level. The ODP also includes a web map that allows exploration without any technical requirement other than a web browser, and we have open sourced all the code, including details on the computational backend. The uniquenes of our ODP makes setting up a technical validation as a comparison with existing datasets challenging. Nevertheless, we relate our signatures to a few well-established data products that capture each a subset of the form and function dimensions we consider. Our results are encouraging in that they show broad agreement in expected areas, but also highlight aspects that can only be discovered when considering form and function in tandem.

The approach and outputs presented bring several benefits to a range of stakeholders interested in cities. This spatial signatures ODP provides insight generated from detailed, comprehensive and computationally intensive data analysis and presents it in a way that is easy to access, work with and integrate into larger projects. Together with the importance of form and function discussed above, we anticipate the output will be relevant to both academic researchers as well as policymakers and practitioners. As a conceptual framework, the spatial signatures provide a flexible yet generalisable approach to understand, characterise and quantify urban form and function. One way to understand our results is as an implementation of a more general way of thinking about the spatial dimension of cities. In this context, it can be useful to researchers and practitioners who, even if not specifically interested in Great Britain, would like to implement a similar approach. In this respect, we hope the present

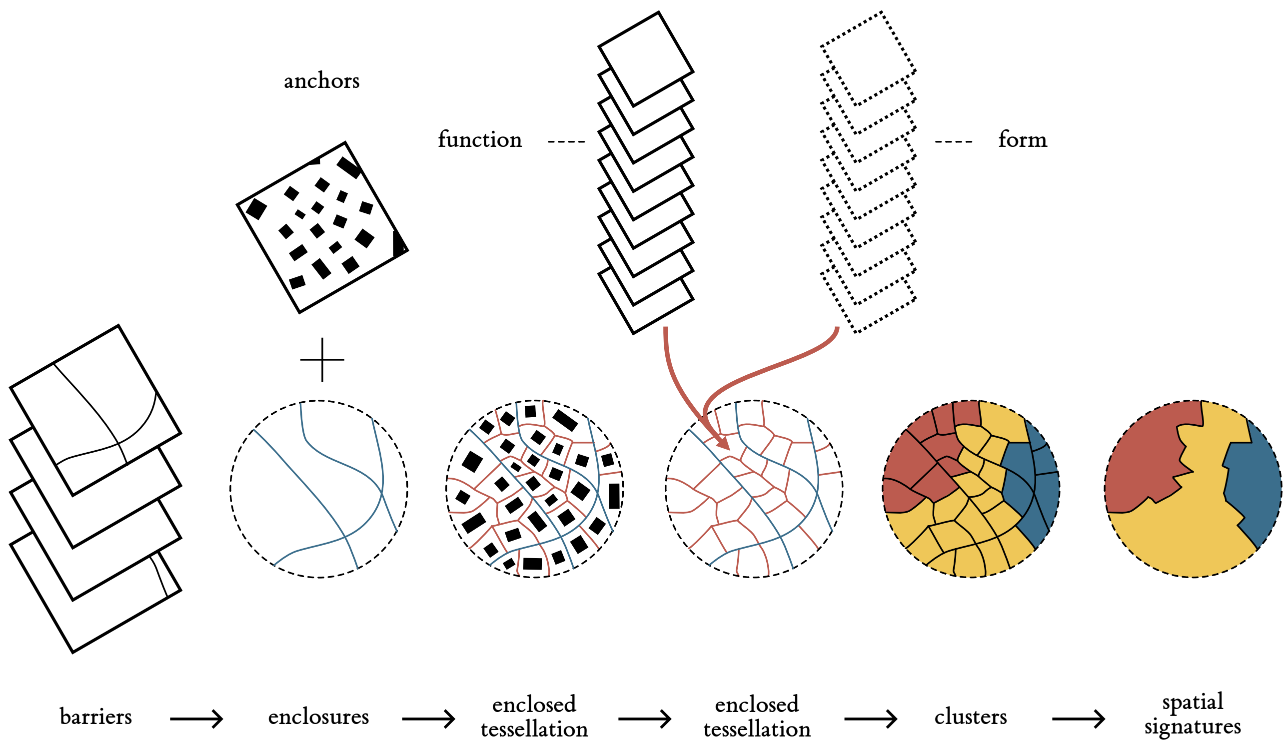

Figure 1. Diagram illustrating the sequential steps leading to the delineation of spatial signatures. From a series of enclosing components, to enclosures, enclosed tessellation (ET), the addition of form and function characters to ET cells, and the development of spatial signatures.

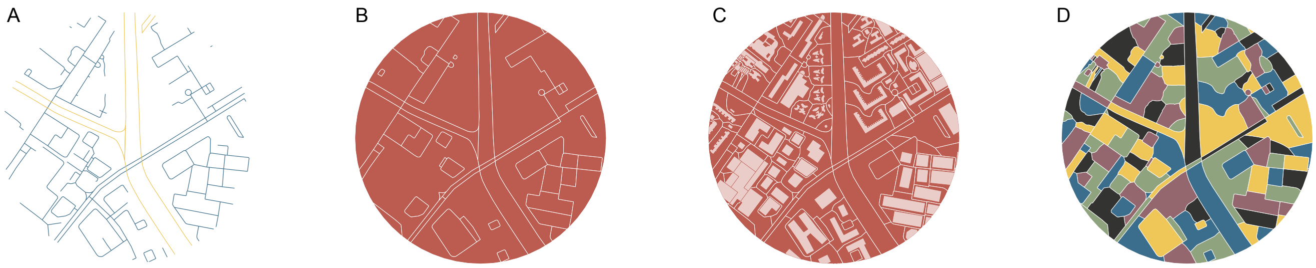

Figure 2. Diagram illustrating the sequential steps leading to the delineation of enclosed tessellation. From a series of enclosing components, where blue are streets and yellow river banks (A), to enclosures (B), incorporation of buildings as anchors (C) to final tessellation cells (D).

paper serves not only to document our own work but to inspire future efforts aimed at urban form and function.

Methods

The method of identification of spatial signatures consists of three top level steps. First, we need to delineate spatial unit of analysis, one that reflects the structure of urban phenomena on a very granular level. Then we charactercterise each of them according to the form and function capturing the nature of each unit and its spatial context. Finally, we use cluster analysis to derive a typology of our spatial units that, once combined into contiguous areas, forms a typology of spatial signatures.

Spatial unit

The first major methodological decision needs to be taken on the definition of the spatial unit. As mentioned, it needs to reflect space in a granular manner and we argue that it should fulfil three conditions. First, it should be *indivisible*, meaning that when such a unit would be subdivided into smaller parts, none of them would be enough to capture the nature of spatial signature. Second, it needs to be *internally consistent* - it should always reflect only a single signature type. Last, it should be geographically *exhaustive*, covering entirety of the study area.

Spatial units used in literature can be split into three groups. One is using administrative boundaries like city regions[4], wards or census output areas[5], that are convenient to obtain and can be easily linked to auxiliary data. However, those rarely reflect the morphological composition of urban space and in some cases may even “obscure morphologic reality”[6]. At the same time, most of them are divisible and larger units are not always internally consistent. Another group is based on arbitrary uniform grids linked either to spatial indexing method like H3[7] or Ordnance Survey National Grid, or to ancillary data of remote sensing or other origins like a WorldPop grid[8]. The issue is that grids cannot be considered internally consistent as they have no relation to the real-life spatial pattern. Finally, urban morphology tends to use morphological elements as street segments[9], blocks[10], buildings[11] or plots[12] as a unit of analysis. Some of those could be seen as indivisible and internally consistent but since they are largely based on built-up fabric, they are not exhaustive. When there is no building or street, there is no spatial unit to work with. Plots could be theoretically considered as exhaustive, consistent and indivisible but there is no accepted conceptual definition and unified geometric representation[13].

We are, therefore, proposing an application of an alternative spatial unit called *enclosed tessellation cell* (ETC), defined as:

A characterisation of space based on form and function designed to understand urban environments

ETC follows the morphological tradition in a sense that is it based on the physical elements of an environment but overcomes the drawbacks of conventionally used units. Its geometry is generated in three steps illustrated on a Figure 2. First, a set of features representing physical barriers subdividing space, in our case composed of street network, railways, rivers and a coastline, is combined together, generating a layer of boundaries. These then partition space into smaller enclosed geometries called *enclosures*, which can be very granular or very coarse depending on the geographic context. In dense city centres where a single enclosure represent a single block is a high frequency of small enclosures, while in the countryside, we can observe very few large enclosures as their delimiters are far away from each other. Enclosures are then combined with building footprints, posing as anchors in the space and are subdivided into enclosed tessellation cells using the morphological tessellation algorithm[14], a polygon-based adaptation of Voronoi tessellation. Resulting geometries are indivisible as they contain, at most, a single anchor building, internally consistent due to their granularity and link to morphological elements composing urban fabric, and geographically exhaustive as they cover entire area limited by specified boundaries.

In the case of classification of Great Britain, street networks are extracted from OS Open Roads datasets[15] representing simplified road centrelines cleaned of road segments under the ground. Railways are retrieved from OS OpenMap - Local[16] ("RailwayTrack" layer) which captures surface railway tracks. Rivers are extracted from OS OpenRivers[17] representing river network of GB as centrelines, and a coastline is retrieved from OS Strategi®[18], capturing coastline as a continuos line geometry. Building geometry is extracted, again, from OS OpenMap - Local ("Building" layer) and represents generalised building footprint polygons. Note that the dataset does not distinguish between individual buildings when they are adjacent (e.g. perimeter block composed of multiple buildings is represent by a single polygon).

Characterisation of space

Spatial signatures are capturing the character of the built and unbuilt environment based on two components - form and function. Each of them is quantified on the level of individual ETCs using methods appropriate for the specific datasets. While form component is described using urban morphometrics (i.e. quantitative analysis of urban form), function is a composite of a variety of data inputs outlined in detail below.

Form

Morphometric characterisation of urban form is based on the numerical description of four elements capturing the built environment - buildings, streets, ETCs, and enclosures, and reflects their patterns based on six categories of characters - dimensions, shapes, spatial distribution, intensity, connectivity and diversity[19]. Each element is considered across different scales, from the measurements of individual geometries, relations of neighbouring geometries to graph-based analysis of street network. The combination of elements, categories and scales results in a set of 59 individual morphometric characters listed in the online table 1.

However, measuring individual characters is not enough to understand the predominant spatial patterns as for some types of urban form is typical high heterogeneity. That means that using, for example, areas of building footprints would in most cases resulted in largely discontinuous clusters. We are, therefore, representing each of the morphometric characters using three proxy variables reflecting statistical distributions of measured data within a spatial context of each ETC. Context is defined as 10^{th} order of contiguity based on the mesh composed of contiguous ETCs. Furthermore, each value is weighted by the inverse distance between so called poles of inaccessibility (defined as a centre of a maximum inscribed circle) of each ETC. Three proxy variables then capture the first, the second and the third quartile of the resulting weighted distribution. Such a characterisation is able to capture the contextual tendency of each morphometric character and hence identify contiguous clusters in both homogenous and heterogenous urban tissues.

Function

Characterisation of the function component uses a different approach. While data describing urban form are not generally available in a processed format, hence we have to employ morphometric approaches, different aspects of function are often available as open data products. Therefore, the main goal of our characterisation of ETCs based on function is to develop appropriate transfer methods to link data published as grids or linked to administrative boundaries to enclosed tessellation.

In this work we are using five different transfer methods:

- Areal interpolation
- Building-based dasymetric areal interpolation
- Network-constrained accessibility
- Euclidean accessibility
- Zonal statistics

Areal interpolation is used when the functional data covers the entirety of space in a form of polygon geometry and when there is no assumption that the phenomena it captures is linked directly to the human population, for example land cover data. When there is an assumption of relation to the population, building-based dasymmetric areal interpolation is used instead. The main difference is that instead of ETC polygons, building footprint polygons linked to individual ETCs are used as a target of interpolation. That ensures that data like population estimates are linked to ETCs proportionally to their ability to provide accommodation, rather than by their area. Network-constrained accessibility is used when the input data represent points of interest like locations of supermarkets. Points are then snapped to the nearest node on the street network and linked to the ETCs as a number of observations within 15 minutes walking distance (1200m on the street network) and a distance to the nearest point. In some cases, Euclidean (as-crow-flies) accessibility is measured instead to accommodate for phenomena that are often reached outside of drivable network like water bodies. Final method, zonal statistics, is used to transfer data originally stored in a raster format to ETC polygons as a mean value of raster pixels intersecting each polygon geometry. Finally, characters based on interpolation and zonal statistics are expressed using their contextual versions following the method used for form characters to, again, reflect the pattern of measured values. The selection of datasets and the chosen transfer method are listed in the table 1.

data	input geometry	transfer method
Population estimates	Vector (output area polygon)	Building-based dasymetric areal interpolation
Retail POIs (supermarkets)	Vector (point)	Network-constrained accessibility
Water bodies	Vector (water body polygon)	Euclidean accessibility
Listed Buildings	Vector (point)	Network-constrained accessibility
Night Lights	Raster (500m)	Zonal statistics
Food Hygiene Rating Scheme Ratings	Vector (point)	Network-constrained accessibility
Workplace population	Vector (output area polygon)	Building-based dasymetric areal interpolation
Culture (theatres, cinemas)	Vector (point)	Network-constrained accessibility
Corine land cover	Vector (land cover zone polygon)	Areal interpolation
NDVI	Raster (10m)	Zonal statistics
CDRC Retail centres	Vector (retail centre polygon)	Euclidean accessibility

Table 1. Functional characters used to describe the function component of spatial signatures. For details of the implementation, refer to the reproducible Jupyter notebooks available at urbangrammarai.github.io.

Cluster analysis

When combined, contextual proxies of form and function characters (or characters themselves when they are reflecting the context by definition) compose a dataset describing each ETC by 331 variables (177 for form and 154 for function.) We treat all of them equally (there is no weighting involved), standardize each variable applying Z-score normalization and use them as an input for K-Means cluster analysis.

Due to the nature of the selected K-Means clustering the step preceding the final analysis is the selection of an optimal number of clusters. We use exploratory clustergram method[20] reflecting the behaviour of different options, the relationship between clustering solutions regarding the allocation of individual observations to classes and the separation between the clusters within each tested solution. Clustergram is further accompanies by measures of internal validation measures - the Silhouette score diagram, Calinski-Harabasz index[21] and Davies-Bouldin index[22]. The optimal number of classes is selected based on the interpretation of clustergram supported by additional measures aiming at a balance between cluster separation and an appropriate detail of resulting classification.

The results of the top level clustering capture the first layer of a national signature classification. However, since the classified ETCs cover entirety of space from vast natural open spaces to dense city centres, it may result in only a few class representing urban areas. While that is caused by the variable heterogeneity of our dataset in combination with K-Means clustering, the measured characters have the ability to further distinguish sub-classes of already identified clusters. As spatial signatures are focused on urban environment, we further subdivide those clusters covering substantial portion of urban areas using another iteration of K-Means clustering. Resulting classification then provide two hierarchical levels capturing the typology of spatial signatures with a detailed focus on urban development.

Finally, individual spatial signature geometries are generated as a combination of adjacent ETCs belonging to the same signature class.

Data Records

Technical Validation

Spatial signatures are unique as a classification method, limiting the potential validation methods to only indirect methods using ancillary datasets capturing conceptually similar aspects of environment. We compare the signatures with three of such datasets, each focusing on a different classification perspective, but all related to our classification to a degree when we can assume there will be measurable level of association between the two:

- WorldPop settlement patterns of building footprints (2021)[8]
- Classification of Multidimensional Open Data of Urban Morphology (MODUM) (2015)[5]
- Copernicus Urban Atlas (2018)[23]

Validation method

All datasets, spatial signatures as well as those selected as validation contain categorical classification of space linked to their unique geometry. The first task, to make each pair comparable is to transfer data to the same geometry. That can be interpolation

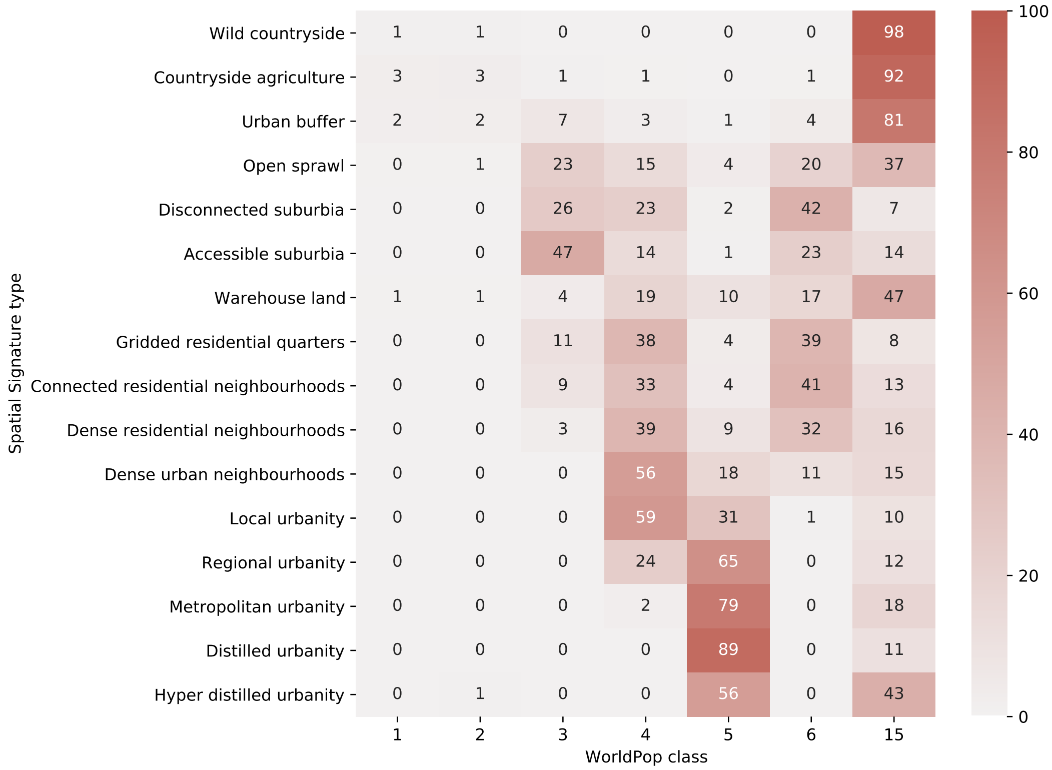

Figure 3. Contingency table showing frequencies (in %) of WorldPop classes within signature types.

of one set of polygon-based data to another (input to ETCs) or converting spatial signatures to the raster representation matching an input raster as the latter is computationally more feasible when one of the layers is already a raster. The second step is a statistical comparison of two sets of classification labels, one representing spatial signature typology and the other validation classes. We use contingency tables and a Pearson's χ^2 test to determine whether the frequencies of observed (signature types) and expected (validation types) labels significantly differ in one or more categories. Furthermore, we use Cramér's V statistics[24] to assess the strength of an association (assuming the Pearson's χ^2 test rejected the hypothesis of independence).

WorldPop settlement patterns of building footprints

WorldPop settlement patterns of building footprints aim to derive a typology of morphological patterns based on the gridded approach (spatial unit is a grid of a size 100x100m per cell) and building footprints. Authors measure 6 morphometric characters linked to the grid cells and use them as an input of unsupervised clustering leading to a 6 class typology. As the classification is dependent on the building footprint data, grid cells that do not contain any information on building-based pattern are treated as missing in the final data product. For the validation of spatial signatures, this *missing* category is treated as a single class. It is assumed that the top-level large scale patterns detected by the WorldPop method and spatial signatures will provide similar results. However, there will be differences caused by inclusion of function in spatial signatures, higher granularity of both initial spatial unit and the resulting classification (6 vs 19 classes).

Signature typology is rasterized and linked to the WorldPop grid. The resulting contingency table is shown on Figure 3. There is a significant relationship between two typologies, $\chi^2(114, N = 22993921) = 13341832, p < .001$. The strength of association measured as Cramér's V is 0.311, indicating moderate association.

MODUM

Multidimensional Open Data Urban Morphology (MODUM) classification describes a typology of neighbourhoods derived from 18 indicators capturing built environment as streets, railways or parks, linked to the Census Output Area geometry. The classification identifies 8 types of neighbourhoods. Compared to the WorldPop classification, MODUM takes into the account more features of built environment than building footprints, which makes it conceptually closer to the spatial signatures. However, it is still focusing predominantly on the form component, although there are some indicators that would be classified as function within the signatures framework (e.g. population). The MODUM method uses a different way of capturing the context compared to the signatures, which leads to some classes being determined predominantly by a single character. For

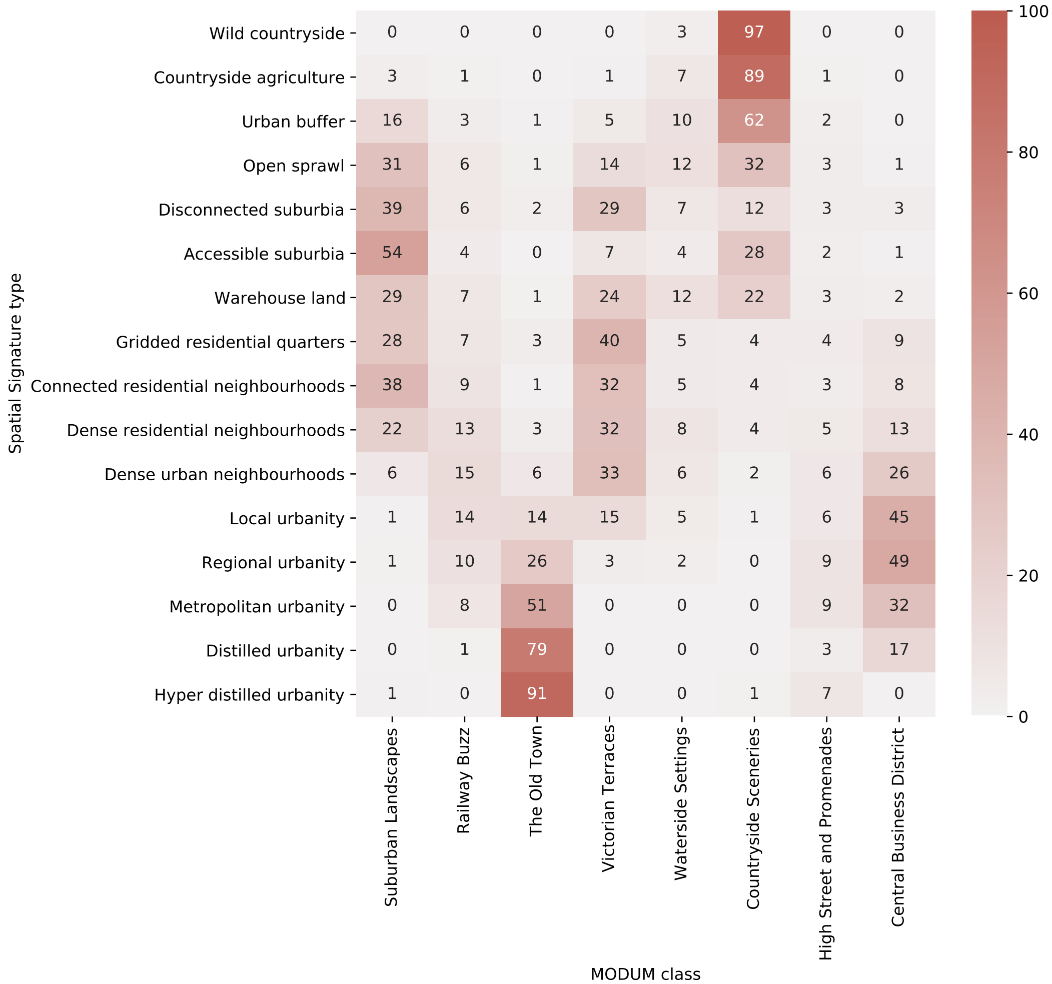

Figure 4. Contingency table showing frequencies (in %) of MODUM classes within signature types.

example, *Railway Buzz* type forms a narrow strip around the railway network, which is an effect signatures are trying to avoid. MODUM typology is available only for England and Wales, therefore the validation takes into the account only ETCs covering the same area. The classification is linked to the ETC geometry based on the proportion (the type covering the largest portion of ETC is assigned). The resulting contingency table is shown on Figure 4. There is a significant relationship between two typologies, $\chi^2(152, N = 13067584) = 13938867, p < .001$. The strength of association measured as Cramér's V is 0.300, indicating moderate association of a very similar levels we have seen above.

Copernicus Urban Atlas

Copernicus Urban Atlas is the least similar of the validation datasets. It is a high-resolution land use classification of functional urban areas derived primarily from Earth Observation date enriched by other reference data as OpenStreetMap or topographic maps. Its smallest spatial unit in urban areas is 0.25 ha and 1 ha in rural areas, defined primarily by physical barriers. The Urban Atlas classification, which identifies 27 classes predefined classes using supervised method. The majority of urban areas is classified as urban fabric further distinguished based on continuity and density resulting in 6 classes of urban fabric. The classification does not consider the type of the pattern or any other aspect. Furthermore, it does not take into account what signatures call the *context* as each spatial unit is classified independently, which in some cases leads to a high heterogeneity of classification within a small portion of land. Signatures take a different approach, therefore it is expected that the similarity between the two will be limited. Urban Atlas is available only for functional urban areas (FUA), leaving rural areas unclassified. Validation then works with FUAs only. The classification is linked to the ETC geometry based on the proportion (the type covering the largest portion of ETC is assigned). The resulting contingency table is shown on Figure **??**. There is a significant relationship between two typologies, $\chi^2(450, N = 8396642) = 5229900, p < .001$. The strength of association measured as Cramér's V is 0.186, indicating weak association.

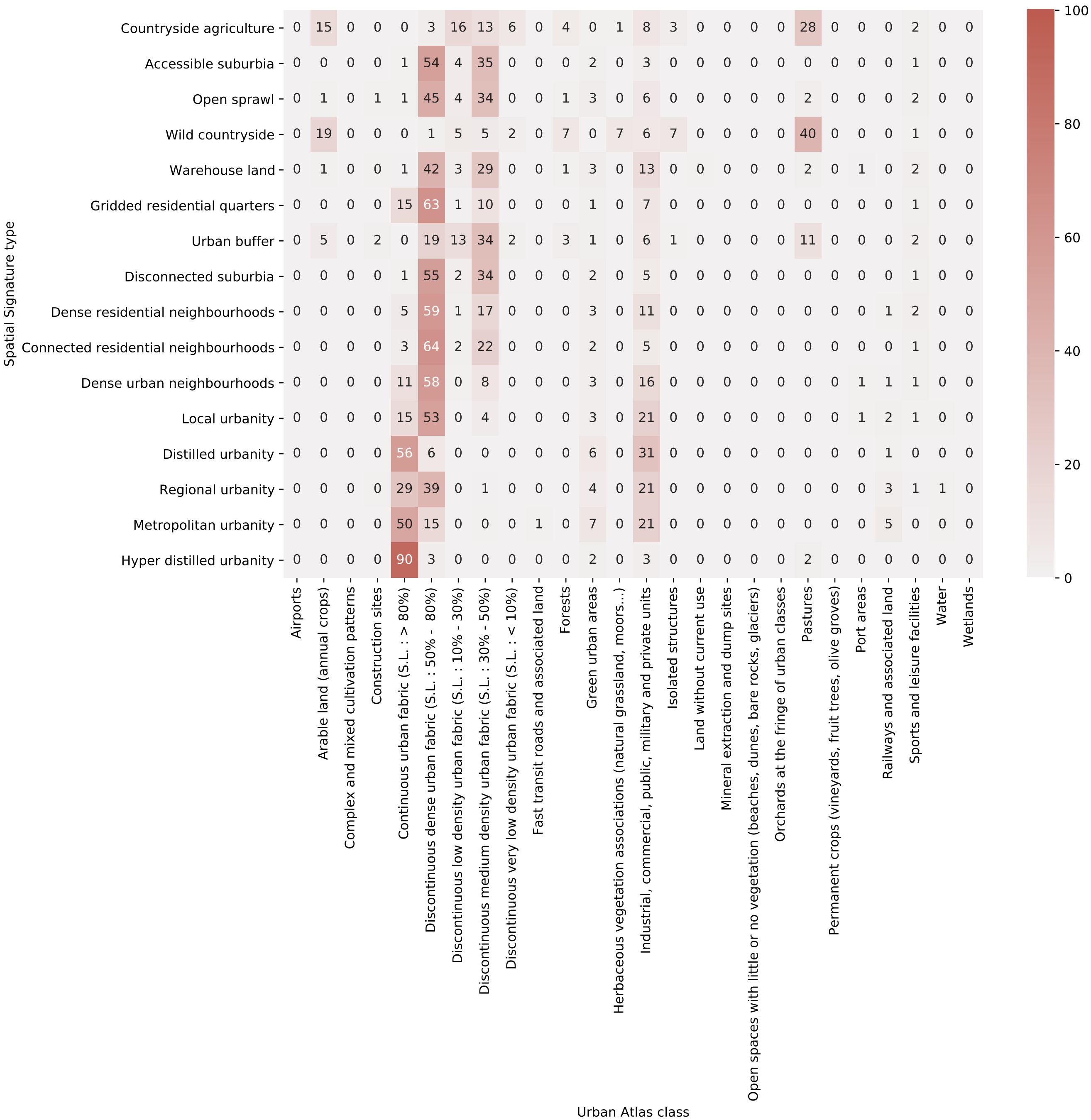

Figure 5. Contingency table showing frequencies (in %) of Urban Atlas classes within signature types.

Summary

None of the comparisons shows better than a moderate association but since none of the validation datasets is aiming to capture the same conceptualization of space as spatial signatures do, such a result is expected. The moderate association with both WorldPop settlements patterns and MODUM is reassuring as both are conceptually closer to signatures than the Urban Atlas (especially in their unsupervised design). Urban Atlas, though very different in its aims and methods still shows measurable association, indicating that the key structural aspects forming cities are captured by both. The validation exercise suggests that general patterns forming cities are shared among signatures and existing typologies. Quantitatively assessing the specificity of spatial signature typology from the perspective of validation further is unfeasible due do uniqueness of the presented typology.

Usage Notes

Released dataset is following the widespread standards for geographic data storage and should not pose a challenge for researches wanting to reuse it. Due to the density of signature geometry (resulting from the detailed ETCs), it may be needed to simplify the geometry for smoother interactive experience on weaker machines.

Replication of the analysis optimally requires at least a single computational node with large amount of RAM (100 GB+) due to the size of the input data and detail on which signature characterization is computed. It is also recommended to revisit the state of the development of related software packages, notably momepy, libpysal, tobler and dask-geopandas as they may in the near future offer more efficient drop-in replacements of the custom code used to produce this dataset.

Code availability

The source code used to produce this dataset is openly available in a GitHub repository at https://github.com/urbangrammarai/spatial_signatures and in the form of a website on https://urbangrammarai.github.io. Code is organized in a series of Jupyter notebooks and have been executed within the darribas:gds_env Docker container, unless specified otherwise in the individual notebooks. The specific version of the container is listed on top of each notebook.

References

1. Arribas-Bel, D. & Fleischmann, M. Spatial Signatures - Understanding (urban) spaces through form and function (2021). Mimeo.
2. Fleischmann, M. & Arribas-Bel, D. Classifying urban form at national scale - The British morphosignatures (2021). ISUF Proceedings.
3. when outXXX, X. Geographic data science. *Geogr. Analysis* **53**, 61–75 (2021).
4. Angel, S., Arango Franco, S., Liu, Y. & Blei, A. M. The shape compactness of urban footprints. *Prog. Plan.* **139**, 100429, 10/gg638j (2020).
5. Alexiou, A., Singleton, A. & Longley, P. A. A Classification of Multidimensional Open Data for Urban Morphology. *Built Environ.* **42**, 382–395, 10/gddwsn (2016).
6. Taubenböck, H. *et al.* A new ranking of the world's largest cities—do administrative units obscure morphological realities? *Remote. Sens. Environ.* **232**, 111353 (2019).
7. Brodsky, I. H3: Uber's hexagonal hierarchical spatial index. *Available from Uber Eng. website: https://eng. uber. com/h3/[22 June 2019]* (2018).
8. Jochem, W. C. & Tatem, A. J. Tools for mapping multi-scale settlement patterns of building footprints: An introduction to the R package foot. *PloS one* **16**, e0247535, 10/gh7sjr (2021).
9. Araldi, A. & Fusco, G. From the street to the metropolitan region: Pedestrian perspective in urban fabric analysis:. *Environ. Plan. B: Urban Anal. City Sci.* **46**, 1243–1263, 10.1177/2399808319832612 (2019).
10. Gil, J., Montenegro, N., Beirão, J. N. & Duarte, J. P. On the Discovery of Urban Typologies: Data Mining the Multi-dimensional Character of Neighbourhoods. *Urban Morphol.* **16**, 27–40 (2012).
11. Hamaina, R., Leduc, T. & Moreau, G. Towards Urban Fabrics Characterization Based on Buildings Footprints. In *Bridging the Geographic Information Sciences*, vol. 2, 327–346, 10.1007/978-3-642-29063-3_18 (Springer, Berlin, Heidelberg, Berlin, Heidelberg, 2012).
12. Bobkova, E., Berghauser Pont, M. & Marcus, L. Towards analytical typologies of plot systems: Quantitative profile of five European cities. *Environ. Plan. B: Urban Anal. City Sci.* 239980831988090, 10/ggbgsm (2019).
13. Kropf, K. Plots, property and behaviour. *Urban Morphol.* **22**, 5–14 (2018).

14. Fleischmann, M., Feliciotti, A., Romice, O. & Porta, S. Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale. *Comput. Environ. Urban Syst.* **80**, 101441, 10.1016/j.compenvurbsys.2019.101441 (2020).

15. Ordnance Survey. OS Open Roads (2020).

16. Ordnance Survey. OS OpenMap - Local (2020).

17. Ordnance Survey. OS Open Rivers (2020).

18. Ordnance Survey. Strategi (2016).

19. Fleischmann, M., Romice, O. & Porta, S. Measuring urban form: Overcoming terminological inconsistencies for a quantitative and comprehensive morphologic analysis of cities. *Environ. Plan. B: Urban Anal. City Sci.* 239980832091044, 10/ggngw6 (2020).

20. Schonlau, M. The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. *The Stata J.* **2**, 391–402, 10/ghh97z (2002).

21. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **3**, 1–27, 10.1080/03610927408827101 (1974). https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101.

22. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis machine intelligence* 224–227 (1979).

23. European environment agency (EEA). Urban Atlas (2018).

24. Cramér, H. *Mathematical Methods of Statistics (PMS-9), Volume 9* (Princeton university press, 2016).

Acknowledgements

(not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

M.F. and D.A. designed the method, M.F. conducted the experiments, M.F. and D.A. analysed the results. M.F. and D.A. written and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Figures & Tables