Characterisation of space in Great Britain using the Spatial Signatures model

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.

Please note: Abbreviations should be introduced at the first mention in the main text – no abbreviations lists or tables should be included. Structure of the main text is provided below.

Background & Summary

(700 words maximum) An overview of the study design, the assay(s) performed, and the created data, including any background information needed to put this study in the context of previous work and the literature. The section should also briefly outline the broader goals that motivated the creation of this dataset and the potential reuse value. We also encourage authors to include a figure that provides a schematic overview of the study and assay(s) design. The Background & Summary should not include subheadings. This section and the other main body sections of the manuscript should include citations to the literature as needed.

The best paper ever is[1].

Methods

The Methods should include detailed text describing any steps or procedures used in producing the data, including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). See the detailed section in our submission guidelines for advice on writing a transparent and reproducible methods section. Related methods should be grouped under corresponding subheadings where possible, and methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study. Specific data outputs should be explicitly referenced via data citation (see Data Records and Citing Data, below).

Authors should cite previous descriptions of the methods under use, but ideally the method descriptions should be complete enough for others to understand and reproduce the methods and processing steps without referring to associated publications. There is no limit to the length of the Methods section. Subheadings should not be numbered.

Subsection

Example text under a subsection. Bulleted lists may be used where appropriate, e.g.

- First item
- Second item

Third-level section

Topical subheadings are allowed.

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, wards or census output areas, that are convenient to obatain and can be easily linked to auxiliarry data. However, those rarely reflect the morphological composition of urban space and in some cases may even “obscure morphologic reality” REF Taubenbock 2019. At the same time, most of them are divisible and larger units are not alwyas internally consistent. Another group is based on arbitrary uniform grids linked either to spatial indexing method like H3 REF or OS National Grid REF, or to anciliarry data of remote sensing or other origins like a WorldPop grid REF. 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 REF, blocks REF buildings or plots 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 builiding 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 (REF Kropf).

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 enviroment but overcomes the drawbacks of conventionally used units. Its geometry is generated in three steps illustrated on a Figure . 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 REF, a polygon-based adaptation of Voronoi tessellation. Resulting geometries are indivisible as they contain, at most, a single achor building, internally consistent due to their granularity and link to morphological elements composing urban fabric, and geogrpahically exhasutive 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 (REF) representing simplified road centrelines cleaned of road segments under the ground. Railways are retrieved from OS OpenMap - Local ("RailwayTrack" layer) which captures surfance railway tracks. Rivers are extracted from OS OpenRivers (REF) representing river network of GB as centrelines, and a coastline is retrieved from OS Strategi® (2016) REF, capturing coastline as a continous 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 adajcent (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. quantitave 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. 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 table XXX.

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 discontinous 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. Furhtermore, each value is weigthed 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

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 preceeding the final analysis is the selection of an optimal number of clusters. We use exploratory clustergram method (REF) 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 furhter accompanies by measures of internal validation measures - the Silhouette score diagram, Calinski-Harabasz index and Davies-Bouldin index. 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 furhter 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

The Data Records section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats. Each external data record should be cited numerically in the text of this section, for example[2], and included in the main reference list as described below. A data citation should also be placed in the subsection of the Methods containing the data-collection or analytical procedure(s) used to derive the corresponding record. Providing a direct link to the dataset may also be helpful to readers (https://doi.org/10.6084/m9.figshare.853801).

Tables should be used to support the data records, and should clearly indicate the samples and subjects (study inputs), their provenance, and the experimental manipulations performed on each (please see 'Tables' below). They should also specify the data output resulting from each data-collection or analytical step, should these form part of the archived record.

Technical Validation

This section presents any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by figures and tables, as needed. This is a required section; authors must present information justifying the reliability of their data.

Usage Notes

The Usage Notes should contain brief instructions to assist other researchers with reuse of the data. This may include discussion of software packages that are suitable for analysing the assay data files, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Please see our code availability policy for advice on supplying custom code alongside Data Descriptor manuscripts.

For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.

Code availability

For all studies using custom code in the generation or processing of datasets, a statement must be included under the heading "Code availability", indicating whether and how the code can be accessed, including any restrictions to access. This section should also include information on the versions of any software used, if relevant, and any specific variables or parameters used to generate, test, or process the current dataset.

References

1. Singleton, A. & Arribas-Bel, D. Geographic data science. *Geogr. Analysis* **53**, 61–75 (2021).

2. Hao, Z., AghaKouchak, A., Nakhjiri, N. & Farahmand, A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. *figshare* https://doi.org/10.6084/m9.figshare.853801 (2014).

3. Kaufman, D. *et al.* A global database of holocene paleotemperature records. *Sci. Data* **7**, 115, https://doi.org/10.1038/s41597-020-0445-3 (2020).

4. Figueredo, A. J. & Wolf, P. S. A. Assortative pairing and life history strategy – a cross-cultural study. *Hum. Nat.* **20**, 317–330, https://doi.org/10.1007/s12110-009-9068-2 (2009).

5. Babichev, S. A., Ries, J. & Lvovsky, A. I. Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon. Preprint at https://arxiv.org/abs/quant-ph/0208066 (2002).

6. Behringer, R. *Manipulating the mouse embryo: a laboratory manual* (Cold Spring Harbor Laboratory Press, New York, 2014).

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g.[3–6]. For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file. For journal articles, DOIs should be included for works in press that do not yet have volume or page numbers. For other journal articles, DOIs should be included uniformly for all articles or not at all. We recommend that you encode all DOIs in your bibtex database as full URLs, e.g. https://doi.org/10.1007/s12110-009-9068-2.

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

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Competing interests

(mandatory statement)

The corresponding author is responsible for providing a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Figures & Tables

Figures, tables, and their legends, should be included at the end of the document. Figures and tables can be referenced in LaTeX using the ref command, e.g. Figure 1 and Table 1.

Authors are encouraged to provide one or more tables that provide basic information on the main 'inputs' to the study (e.g. samples, participants, or information sources) and the main data outputs of the study. Tables in the manuscript should generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an appropriate data repository.

Tables may be provided within the LaTeX document or as separate files (tab-delimited text or Excel files). Legends, where needed, should be included here. Generally, a Data Descriptor should have fewer than ten Tables, but more may be allowed when needed. Tables may be of any size, but only Tables which fit onto a single printed page will be included in the PDF version of the article (up to a maximum of three).

Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the article and will be made available in the online version only. Any such tables must be labelled in the text as ‘Online-only’ tables and numbered separately from the main table list e.g. ‘Table 1, Table 2, Online-only Table 1’ etc.

Figure 1. Legend (350 words max). Example legend text.

Condition	n	p
A	5	0.1
B	10	0.01

Table 1. Legend (350 words max). Example legend text.

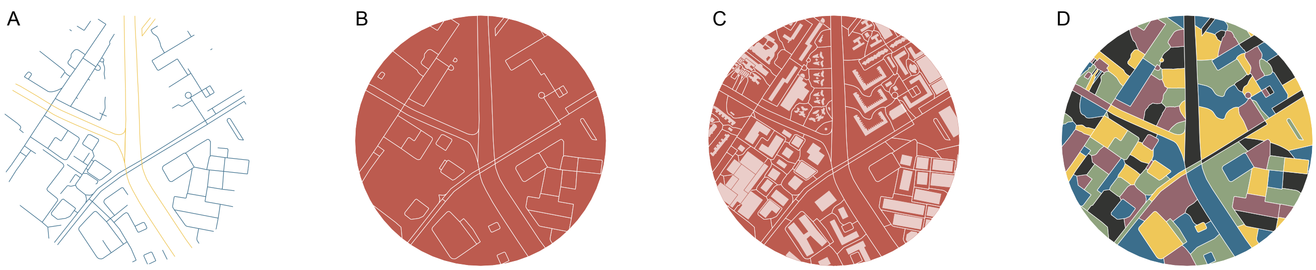

Figure 2. Diagram illustrating the sequential steps leading to the delineation of enclosed tessellation. From a series of enclosing components, where blue are streets and yellow river banks (A), to enclosures (B), incorporation of buildings as anchors (C) to final tessellation cells (D).