COGITATIO

Urban Planning (ISSN: 2183–7635)
2018, Volume 3, Issue 1, Pages 58–74
DOI: 10.17645/up.v3i1.1269

Article

Social and Physical Characterization of Urban Contexts: Techniques and Methods for Quantification, Classification and Purposive Sampling

Miguel Serra [1,*], Sophia Psarra [2] and Jamie O'Brien [3]

[1] Research Centre for Territory, Transports and Environment, University of Porto, 4200-465 Porto, Portugal;
E-Mail: mserra@fe.up.pt
[2] Space Syntax Laboratory, Bartlett School of Architecture, University College London, London, NW1 2BX, UK;
E-Mail: s.psarra@ucl.ac.uk
[3] The Bartlett Centre for Advanced Spatial Analysis, University College London, London, W1T 4TJ, UK;
E-Mail: jamie.o'brien@ucl.ac.uk

* Corresponding author

Submitted: 17 November 2017 | Accepted: 26 February 2018 | Published: 29 March 2018

Abstract
Robust quantitative descriptions of the social and physical characteristics of urban contexts are essential for assessing the impacts of urban environments on other, potentially dependent variables. Common methodologies used for that purpose, however, are either coarse or suffer from biasing effects. At the social level, the use of indicators encoded into pre-defined areal units, makes results prone to the Modifiable Areal Unit Problem. At the physical level, the adopted morphological indicators are usually highly aggregated descriptors of urban form. Moreover, there is a lack of explicit methodologies for the purposive sampling of urban contexts with specific combinations of social and physical characteristics, which—we argue—may be more effective than probabilistic sampling, when exploring phenomena as elusive as the effects of urban contextual factors. This article presents a set of GIS-based methods for addressing these issues, based on: a) local indicators of spatial association; b) detailed quantitative morphological descriptions, coupled with unsupervised classification techniques; and c) purposive sampling strategies carried out on the data generated by the proposed context characterization methods (a and b). The methods are illustrated through the characterization of the urban contexts of the 77 state-sector secondary schools in Liverpool, but are generalizable across all categories of urban objects and are independent of the geographical context of implementation.

Keywords
characterization; morphological; purposive sampling; socio-economic; urban context

Issue
This article is part of the issue "Crowdsourced Data and Social Media for Participatory Urban Planning", edited by Bernd Resch (University of Salzburg, Austria), Peter Zeile (Karlsruhe Institute of Technology, Germany) and Ourania Kounadi (University of Salzburg, Austria).

1. Introduction

Urban contextual factors, both social and physical, have impacts on the observed variabilities of a wide range of phenomena as, for example: the distributions of socio-spatial inequalities (Rae, 2012), the spatial incidences of public health problems such as obesity (Townshend & Lake, 2009) and mental health (Cutrona, Wallace, & Wesner, 2006; Miles, Coutts, & Mohamadi, 2011), local differences in patterns of physical activity (Timperio et al., 2010) and mobility (Crane, 2000), or still the spatial distributions of crime occurrences (Charron, 2009). In more general terms, the whole body of literature in the field of 'neighbourhood effects', departs from the hypothesis

that local urban contexts have significant impacts on the life of residents, and seeks to assess such hypothesis (van Ham & Manley, 2012).

Nevertheless, any research aiming at identifying potential links between urban contextual factors and other variables of interest, must necessarily face at least two initial problems. The first is how to quantify urban contexts at both social and physical levels. The methodologies currently employed for those purposes have several limitations, and the need for robust quantitative methods has been acknowledged by several authors (Cummins, Macintyre, Davidson, & Ellaway, 2005; Gambaro, Joshi, Lupton, Fenton, & Lennon, 2016; Lupton, 2003). The second problem concerns the criteria for generating context-informed samples of urban areas. As van Ham and Manley (2012) note, quantitative studies using large randomized probability samples have been shown to be far less effective than qualitative studies (i.e., focusing on the experiences and perceptions of residents), in detecting contextual or neighbourhood effects. Qualitative research, however, due to its laborious inquiry processes, demands sampling strategies aimed at creating small yet information-rich samples; that is, purposively selected samples. Purposive sampling strategies are quite different from those of probabilistic sampling, seeking not generalization or randomness, but the well-informed selection of very specific cases, capable of maximizing the chances of observing phenomena of interest. They are also less well-known and understood than probabilistic sampling strategies, even though they are more suitable in certain circumstances (e.g., when the studied population is small) and more effective when used in qualitative research (Patton, 1990).

This article develops a unified methodological framework for the quantification, measurement and sampling of urban contexts, based on both social and physical characteristics, using GIS. The framework was devised within the scope of a community-based, participative research project "Visualising Inequality in Community Networks to Enhance Participatory Planning" (O'Brien, Garcia Vélez, & Austwick, 2017; O'Brien et al., 2016), supported by Leverhulme Trust Research Project Grant, where it was used for characterizing the socio-economic and morphological contexts of all state-sector secondary schools in the Liverpool City Region (Merseyside, UK). However, the framework was designed from the outset so as to remain applicable in any research exercise demanding quantitative contextual characterizations of a given class of urban features (e.g., urban neighbourhoods, the surroundings of institutional buildings or designated public spaces). These quantitative characterizations may be then used to support the generation of purposive samples of local urban areas, informed by a rigorous and detailed characterization of their social and physical differences.

The objective of this article is, therefore, methodological. It aims at contributing to the fields of urban and neighbourhood studies, by providing enhanced context characterization and sampling tools, with a particular focus on purposive sampling strategies. Such tools can be used for several research purposes, namely: a) supporting qualitative, participatory urban research designs, by providing quantitative contextual characterizations of the studied areas, against which qualitative findings may be assessed and interpreted; b) for supporting studies on the complex interactions between social phenomena and the built environment, by enabling the purposive sampling of urban areas with specific combinations of social and physical characteristics; and c) generating samples of local urban areas controlling for their social and physical characteristics, in order to avoid potential confounding effects on the study of other variables of interest.

The article is organized as follows: in Section 2 we discuss the limitations of the current urban contextual characterization methods and how to overcome or mitigate such limitations. We then present a unified methodological framework for characterizing and sampling urban contexts. We end this section by briefly describing the abovementioned research project, as a background to illustrating the application of the methodology. The application of methods is discussed in the following three sections, which constitute the main body of the article. We conclude by summarizing the outputs of the proposed methods.

2. Methodological Framework

Quantitative characterizations of urban contexts are usually carried out at two different levels: social and physical. At the social level, such characterizations usually rely on statistical indicators, spanning several socioeconomic and demographic dimensions, commonly aggregated into individual administrative divisions of varying geographies (e.g., census tracts). Although of obvious convenience, due to the wide availability of census data and the pre-defined nature of administrative boundaries, such approach raises several methodological issues (Caughy, Leonard, Beron, & Murdoch, 2013; Kim, Ali, Sur, Khatib, & Wierzba, 2012; Lebel, Pampalon, & Villeneuve, 2007).

Firstly, there is no guarantee that the boundaries of extant administrative geographies will indeed correspond to meaningful spatial or social units of analysis, within the context of each study. However, given the 'off the shelf' availability of administrative boundaries, researchers often use those whose average size better approximates their research objectives. Secondly, because census data are aggregated into administrative units of varying sizes and boundaries, individual units' attributes are prone to be biased by the Modifiable Areal Unit Problem (MAUP; Openshaw, 1983), both through its scale effects (i.e., sensitivity to levels of aggregation) and zoning effects (i.e., sensitivity to the shapes of aggregation units). And thirdly, administrative units are usually characterized and sampled accordingly only to their individual attributes, without regard to their wider spatial em-

bedment (i.e., not taking into account their neighbours' characteristics). However, because individual attributes may be biased by the MAUP, their use as only characterization criterion may not be the best option.

At the physical level, urban contexts are commonly described through broad morphological indicators (e.g., residential density, functional diversity or the total area of green spaces). These indicators are measured and aggregated also at the level of pre-defined administrative units (Charron, 2009; Inoue, Stickley, Yazawa, & Shirai, 2016; MacDonald, Wise, & Harris, 2008; Miles et al., 2011; Townshend & Lake, 2009). Besides being also prone to MAUP (due to the zoning effect), such indicators are rather coarse descriptions of the built environment and may not be sufficiently detailed for detecting potential statistical associations between different urban morphologies and other variables. On the other hand, urban typomorphologies (Vernez-Moudon, 1994) can be of great interest for characterizing physical urban contexts, because they allow identifying comparable and/or contrasting built environments. But urban typomorphologies are commonly identified visually and described only through semantic and graphical means. Constructed in this way, they are difficult to generalise and use outside of their original observation setting. Nevertheless, there is also today a growing body of research dedicated to quantitative methods for the detailed description and classification of urban form, using geocomputation and algorithmic classification methods; see, for example, the work of Gil, Beirao, Montenegro and Duarte (2011) and of Hamaina, Leduc and Moreau (2012). However, to the best of our knowledge, these algorithmic methods have never been applied in neighbourhood or community-based studies, requiring the support of urban physical characterizations.

These quantification shortcomings can affect the identification and sampling of relevant cases, among the variability of social and physical urban contexts. An important decision in studies of local urban communities or neighbourhoods, particularly in the case of qualitative research, is the selection of the cases to be studied. In qualitative research, as opposed to purely quantitative research, the generation of samples is often done purposefully, i.e., in a non-random manner, identifying information-rich cases, in the light of the specific phenomena under investigation (Patton, 1990). Such samples should not include too many cases (for logistical and financial reasons) and should meet defined conditions determined by the phenomena under study or by the question being asked. However, if the quantitative methods adopted for characterizing specific cases are biased in the first place, purposive sample generation based on their results may obviously be jeopardized.

The methodological framework proposed in this article (Figure 1) tries to overcome the abovementioned problems. The framework is divided into two analytical tracks—concerning social and physical urban contexts—which are organised into three steps: 'data preparation', 'quantification' and 'characterization'. The data produced by the two tracks are joined at the end, supporting the generation of purposive samples of local urban areas based on their contextual characteristics.

At the social level and in order to avoid or mitigate the effects of the MAUP, we propose to change focus from the specific values of the variables at each spatial unit (which is the level at which MAUP occurs) to another type of quantitative property, namely the degree of local spatial autocorrelation (LSA) of each spatial unit (regarding the variable under study). This approach has two advantages. Firstly, it mitigates the zoning effects of the MAUP because only the degree of LSA is taken into account for characterizing each spatial unit, to which is associated a probability of it being observed simply by chance. Thus, the selection of spatial units based on the significance of their degree of LSA on a given socioeconomic variable, ensures that the unit selected is actually embedded in a (non-random) spatial cluster of similar values. Secondly, it also avoids the scale effects of the MAUP, because LSA may be assessed across several spatial scales, allowing the selection of units that show consistent behaviours across scales. In order to characterize spatial units by their degree of spatial autocorrelation, we use Local Indicators of Spatial Association (LISA; Anselin, 1995), which are geo-statistical methods devised for that purpose. LISA methods and their application are detailed in Section 3.

Regarding the characterization of physical contexts, we move from the coarse descriptions of urban form mentioned previously, to more discriminant methods of quantitative morphological characterization. In the field of urban morphology there has been recently increasing interest in the development of quantitative methods for measuring and classifying urban forms (Barthelemy, 2015; Dibble et al., 2017; Gil et al., 2011; Hamaina et al., 2012; Marshall, 2005; Pont & Haupt, 2010; Serra, Gil, & Pinho, 2016), using available vector datasets of street networks and building footprints, analysing their morphological information through geocomputation and subjecting it to unsupervised classification algorithms. This algorithmic approach produces consistent and quantitatively defined morphological classifications, which are automatically derived only from the morphological data, providing objective criteria for accurately describing similarities and differences in local urban morphologies.

The proposed method adopts a number of morphometric indicators, describing three dimensions of urban form: the network of open space, the geometry of urban blocks and that of building footprints. These indicators are quantified in GIS, using open source vector datasets for each study area. The resulting morphological data are subsequently subjected to unsupervised hierarchical classification, in order to reduce their variability to a manageable number of clusters, representing actual and measurable morphological cleavages between the studied areas. These methods are described in Section 4.

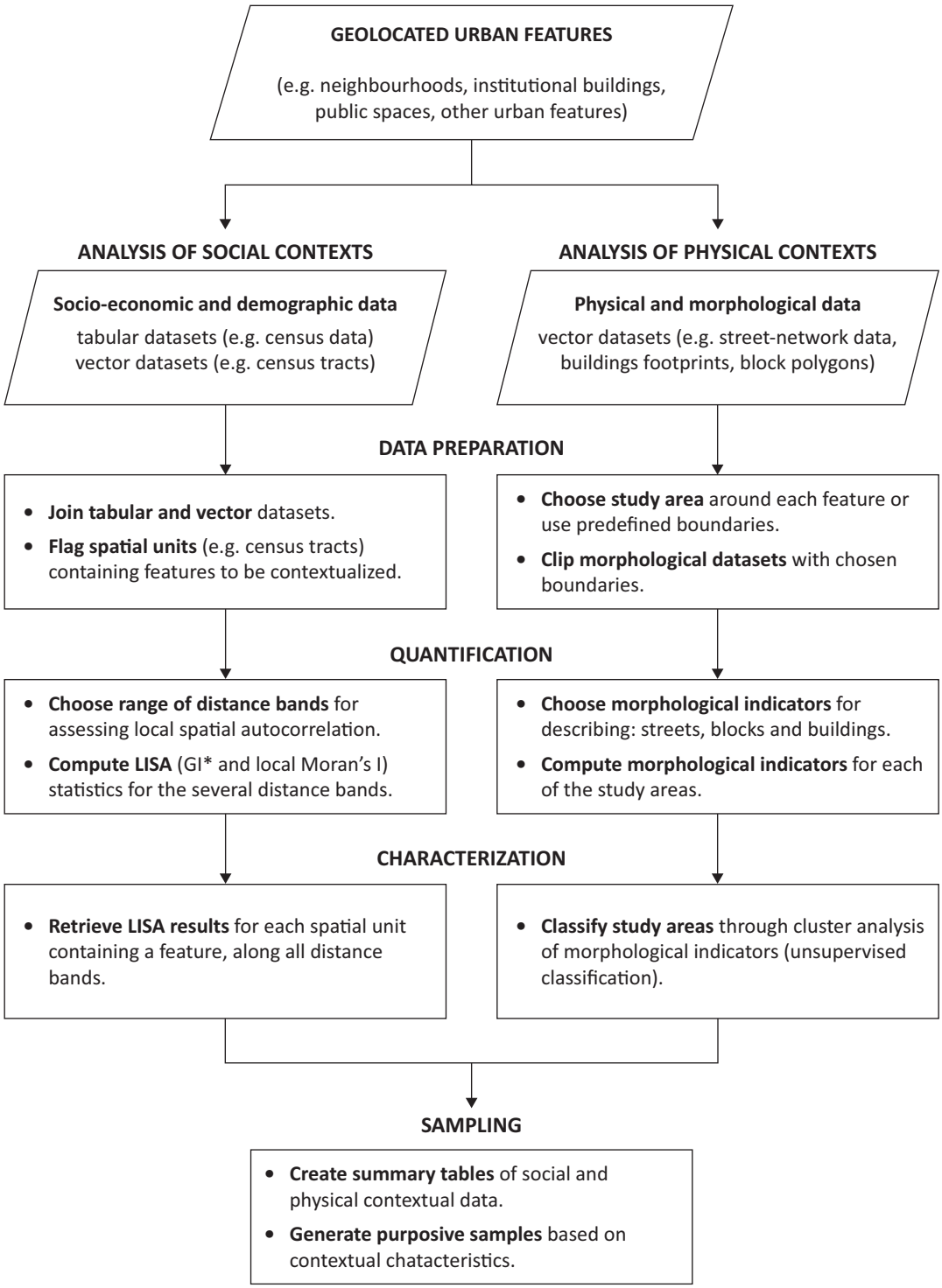

Figure 1. Proposed methodological framework, divided into two analysis tracks, each of which is organised into three stages and a synthesis in order to derive purposive samples.

Finally, the results of both characterization methods are summarized through simple data visualization schemes (i.e., summary tables), allowing for the quick identification of relevant cases, according to several purposive sampling strategies (Patton, 1990), aimed at answering specific research questions. These sampling strategies and the corresponding samples are described in detail in Section 5.

We end this section by providing a succinct description of the research project "Visualizing Inequalities in Community Networks", in the context of which the proposed methodological framework was developed, and whose data we use here for illustration purposes. The project consisted of community-based, qualitative research, in part carried out in Liverpool, Merseyside, UK (Figure 2). The main objective was to gain an understand-

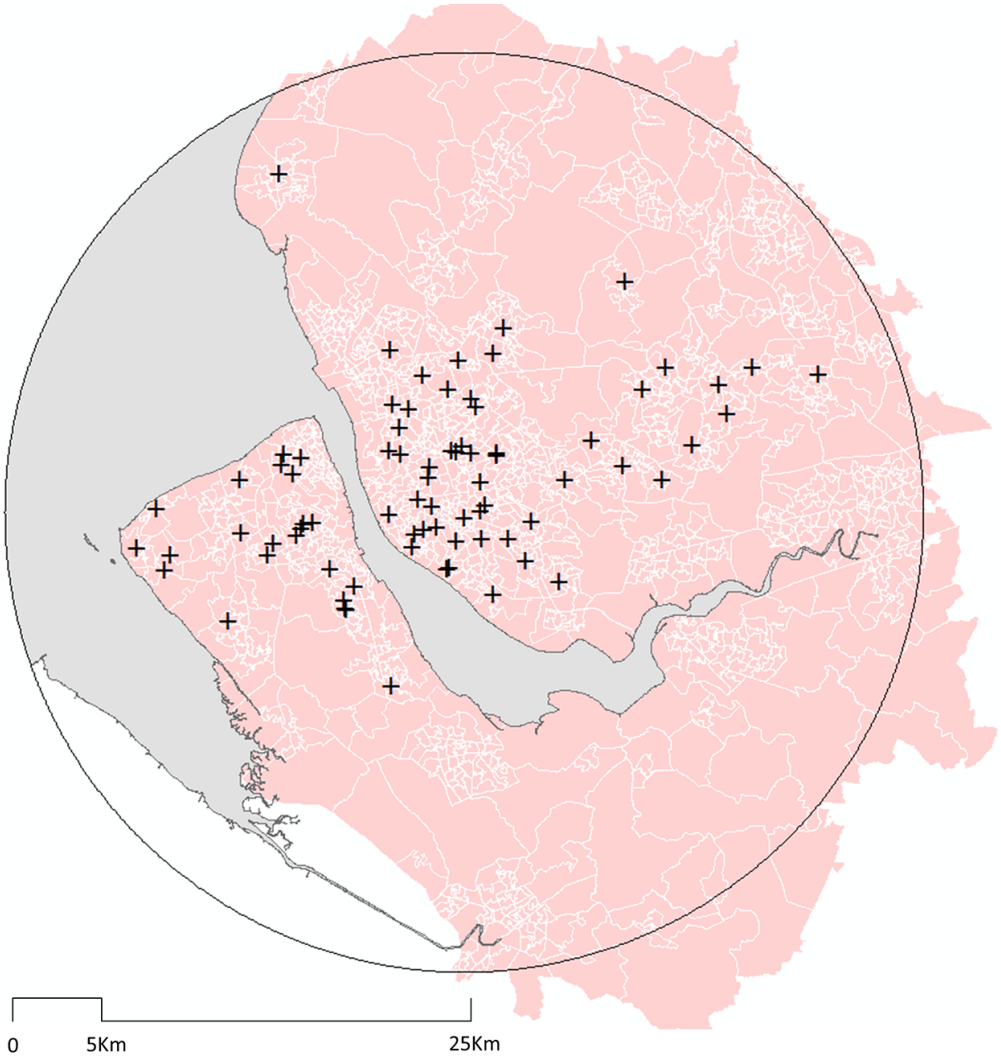

Figure 2. Distribution of the 77 secondary schools in the study area, over LSOA geography.

ing of how people make use of spatial assets in their vicinities, in order to conceptualize their local community formations. In addition, the project sought to understand how the characteristics of the built environment enabled or hindered individual community conceptualizations, and how this might vary across different social and physical urban contexts.

The study focused on the 77 state-sector secondary schools within Liverpool City Region (Figure 1), a region which presents among the UK's widest socio-economic inequalities (LCC, 2015). From the 77 schools, a small sample (n = 16) was chosen (23% of the total) based on the following criteria (by order of relevance): a) the responsiveness of secondary school teachers to our invitation to participate in the study; and b) the inclusion of schools with contrasting social and physical urban contexts. Qualitative data for the project's research purposes, described in detail elsewhere (O'Brien et al., 2017; O'Brien et al., 2016), was gathered through participatory workshops carried out in the 16 selected schools, involving 246 secondary school-age children, aged from 11 to 19 years.

The methods described in this article had a twofold purpose. Firstly, to provide quantitative information of the social and physical urban contexts of all 77 schools, so that that responded positively to the invitation could be evaluated regarding criterion b), above. Secondly, to use the quantitative data as benchmark against which the qualitative data gathered through participatory research activities could be interpreted.

3. Characterizing Urban Socio-Economic Contexts

In order to characterize the socio-economic contexts of all 77 Merseyside secondary schools, we started from a convenience definition of boundaries, namely those of the lower super output areas (LSOAs) where each school is located (Figure 2). LSOAs are geo-located units devised by the UK's Office of National Statistics (ONS) to represent population aggregations by place of residence of around 1500 inhabitants. LSOAs do not represent any meaningful definition of 'neighbourhoods' or of 'urban communities', but they do allow for relatively stable local area analyses because they are designed to have similar

population sizes and be as socially homogenous as possible (ONS, 2011).

We load all LSOAs into a GIS within and intersected by a 25 km radius circular boundary, centered on the Liverpool City Region polygonal centroid (Figure 1). As an indicator of their socio-economic composition, we associate each LSOA with its corresponding score in the 'Income Deprivation' domain of the Index of Multiple Deprivation (IMD), the official measure of relative deprivation in England (DCLG, 2015). IMD is a composite index, constructed by weighting indicators for seven domains of deprivation, of which 'Income' and 'Employment' carry the greatest weight (22.5% each). We use the 'Income' domain scores instead of IMD scores, because they are meaningful and interpretable (corresponding to the percentage of the income-deprived population in each LSOA). In contrast, IMD scores are highly transformed and not comparable (IMD scores should be ranked or classified in quantiles) (DCLG, 2015). IMD 'Income' scores, as well as LSOAs boundaries, are provided as open datasets by the ONS.

As previously explained, we are specifically interested in the spatial embeddedness of each LSOA within potentially larger geographical patterns of income deprivation or lack thereof. For this purpose, we use a set of spatial statistics based on the concept of spatial autocorrelation, known under the broad designation of LISA (Anselin, 1995). LISA methods determine the degree to which a geographical feature (e.g., a given LSOA) has a particularly high or low score, according to the attribute itself *and* to the location of the feature in question. In this way, we can evaluate the degree to which a school located within a high-, medium- or low-deprivation LSOA is also located within a larger area of relative deprivation or affluence. Moreover, spatial embeddedness may be assessed for neighbourhoods of varying sizes around each feature. As argued in Section 2, shifting the focus from individual scores to the LSA of those scores, mitigates both the zoning and scale effects of the MAUP.

We apply two LISA methods: the Gi* statistic (Ord & Getis, 1995) and the Local Moran's I statistic (Anselin, 1995), also known as 'hot spot analysis' and 'cluster and outlier analysis', after their respective implementations in ArcGIS 10 (ESRI, 2011). These methods allow comparing each LSOA's 'Income' score with those of its neighbours. When those scores are similarly high or low (i.e., when there is local spatial autocorrelation between a given feature and its neighbours) and when that similitude attains a given degree of statistical significance (i.e., a low probability of occurring by chance), both methods retrieve a signal of spatial clustering (i.e., a significantly high or low Z-score, associated with a certain p-value). Both methods also require the choosing of a given distance band in order to define which neighbouring features are included in the calculations.

The main difference between the two methods concerns the inclusion of the value of the feature under analysis in the calculations of the local mean. In the case of hot/cold spot analysis (Gi*), the local mean is calculated by taking the values of the feature under analysis and those of the features within its neighbourhood; this local mean is then compared to the global mean (i.e., of the entire study area). If significantly different (i.e., yielding a large positive or negative Z-score), the feature is categorized as being part of a hot or cold spot. The output, for each distance band, is therefore a set of hot and/or cold spots, with varying confidence levels (CL, 90%, 95% and 99%) and/or a set of areas without significant clustering (i.e., CL $<$ 90%) of both high or low scores.

In cluster and outlier analysis (local Moran's I) the process is similar, but only the values of the neighbouring features, and not the value of the feature under analysis, are considered. Again, this local mean is compared to the global one, in order to ascertain if they differ significantly. However, the resulting value of the statistic (I) indicates if the feature under consideration also differs from the local mean or not. If the value of I is positive, the feature under analysis has neighbouring features with similar values and is therefore part of a cluster (of high or low values). If the value of I is negative, the feature under analysis has a value that is dissimilar from the local mean and is therefore a local outlier.

We use both LISA methods simultaneously, because they produce slightly different, but complementary outputs. Analysis was run at a range of increasing fixed-distance bands, namely 0.5 km, 1 km, 1.5 km, 2 km, 2.5 km and 3 km. This approach allows us to study the clustering of 'Income' scores across several spatial scales, without the need to define some fixed, discretionary size for the neighbourhood of each feature. The output maps of both methods (Figures 3 and 4) show similar patterns, with high and low 'Income' scores clustering very clearly, revealing the existence of strong socio-economic cleavages within the study area. Central urban areas tend to display high values of deprivation, while peripheral urban centres and rural areas show concentrations of low deprivation. Deprived areas show significant clustering immediately at 0.5 km, whereas non-deprived areas start to cluster at larger scales (from 1 km on), reflecting the different sizes of central and peripheral LSOAs (deprivation is predominant in smaller and more central LSOAs).

Figure 5 shows the socio-economic contexts of each of the 77 schools. For each LSOA containing a secondary school (the columns, in Figure 5), we record its 'Income' score and the rank of that score, regarding the set of 77 secondary schools. We recall that low 'Income' scores mean low deprivation, thus the LSOA ranked first is also the most deprived. Furthermore, we also record the status of each LSOA regarding the spatial clustering of 'Income' scores, as categorized by the two LISA methods across the distance bands mentioned before.

Regarding the results of hot/cold spots analysis (Gi*) and for each distance band (the lines, in Figure 5), each LSOA containing a school may: 1) be part of an area with non-significant clustering of 'Income' scores (white cells); 2) be part of a cold spot of 'Income' scores (blue cells: light blue 90% CL, medium blue 95% CL and dark blue

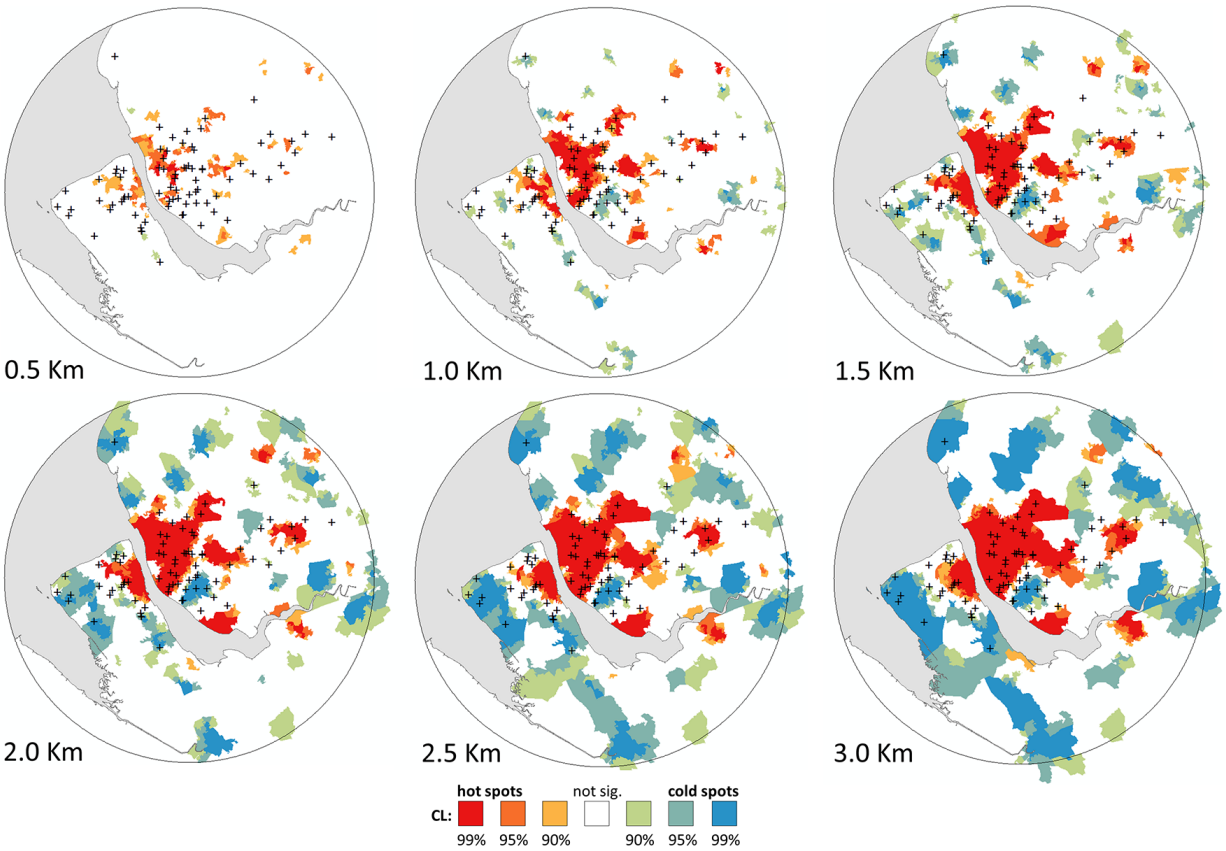

Figure 3. Hot/cold spot analysis of Income scores for the study area.

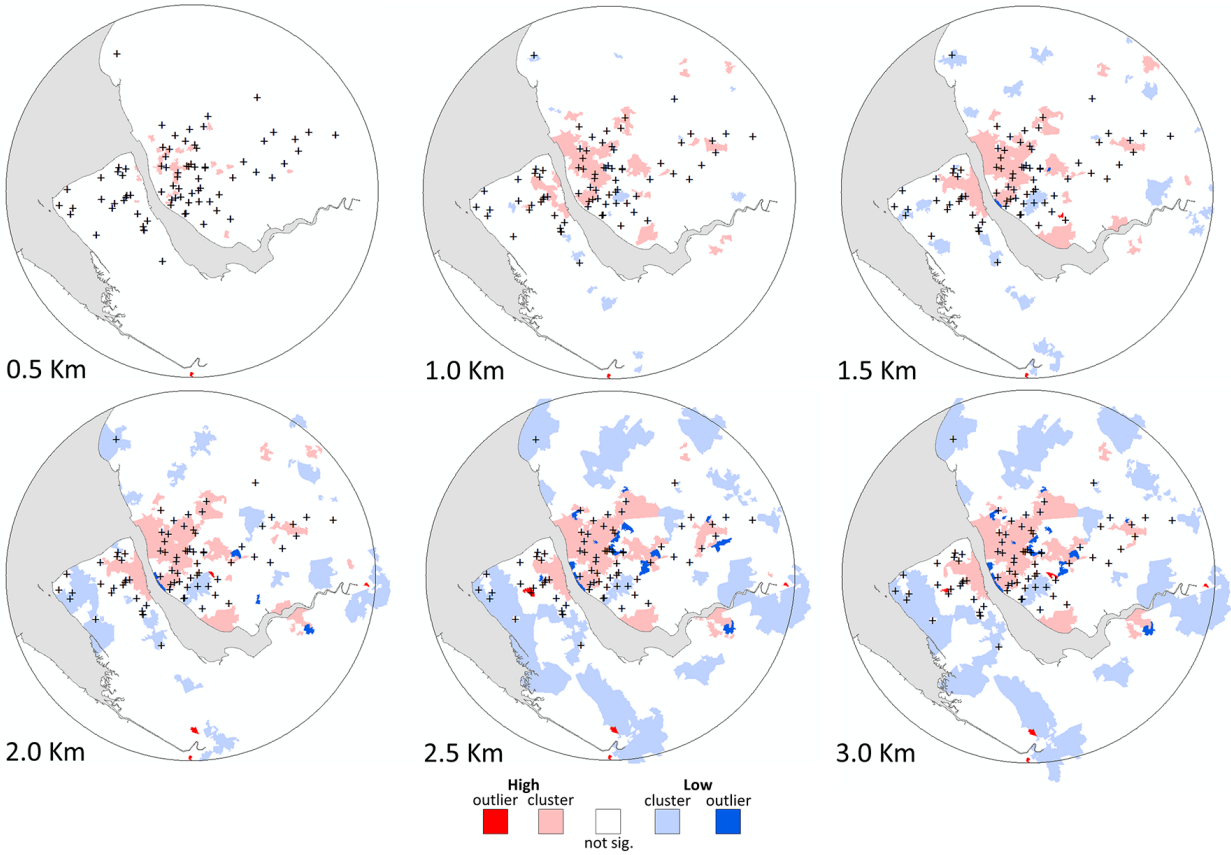

Figure 4. Cluster/outlier analysis of Income scores for the study area.

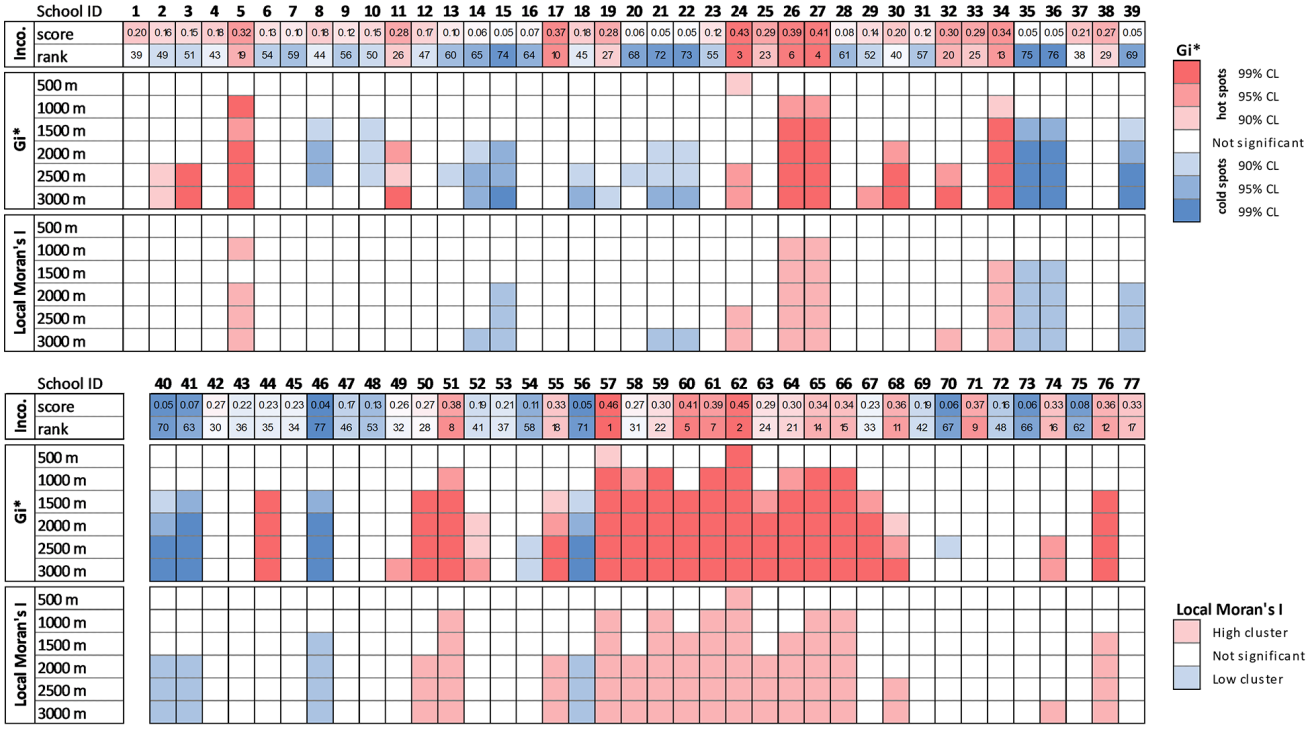

Figure 5. Socio-economic contexts of the 77 secondary schools.

99% CL); and 3) be part of a hot spot of 'Income' scores (red cells: light red 90% CL, medium red 95% CL and dark red 99% CL).

Similarly, regarding the results of cluster/outlier analysis (local Moran's I) and for each distance band, each LSOA containing a school may: 1) be part of an area with non-significant clustering of 'Income' scores (white cells); 2) be part of a cluster of low 'Income' scores (light blue cells, 95% CL); 3) be part of a cluster of high 'Income' scores (light red cells, 95% CL). In the maps of Figure 4 there are several LSOAs that are spatial outliers, but none of the 77 schools is located in these.

Figure 5 shows that there is indeed strong variability in the socio-economic contexts of the secondary schools in the Merseyside. In a number of cases, the school's socio-economic context does not show any significant clustering of either high or low 'Income' scores (e.g., cases 1, 4 or 6) and could be deemed uncharacteristic study targets. Nevertheless, some of these cases do correspond to LSOA that have either high individual scores (e.g., cases 17 and 71) or low individual scores (e.g., cases 7 and 16), in spite of not being part of spatial clusters of neither high nor low scores. We also observe a number of schools within areas where there is a very consistent clustering of high (e.g., cases 5, 27 or 34) and of low (e.g., cases 35, 39 or 46) scores, at almost all spatial scales. These schools are deeply embedded into consistent areas of high and low deprivation and would, therefore, constitute good study targets.

We also noted cases where the categorization produced by the two LISA methods is not consistent. For example, cases 8, 11 or 44, are part of cold/hot spots as defined by Gi*, but not of clusters of low/high scores, as defined by local Moran's I. These differences happen because of the different calculations used in the two methods, but they also speak to a lesser consistency in the characterization of these cases. There are therefore advantages in employing both methods, because they allow assessing the consistency of results and the eventual rejection of inconsistent cases.

4. Characterizing Urban Morphological Contexts

In order to characterize the urban morphological contexts of the 77 schools, we start by defining circular areas of 1 km radius, centred on each school's postcode centre point. This definition of boundaries is discretionary, pertaining to what in the context of the project was considered an adequate extent for the study area around each school. A boundary of this kind could be replaced by any other, without calling into question the adopted method. A number of morphological indicators were computed for the built environments within these areas, based on an open access vector dataset (OS, 2015) describing the full road network hierarchy and the footprints of all buildings.

The chosen morphological indicators cover three fundamental aspects of urban structure, which change significantly across urban areas and historical periods (Figure 6). These are: the geometry and topology of the street network (i.e., streets and junctions), the geometry and topology of urban blocks, and the geometry, density and grain of buildings. We note however, that a larger number of morphological attributes could be

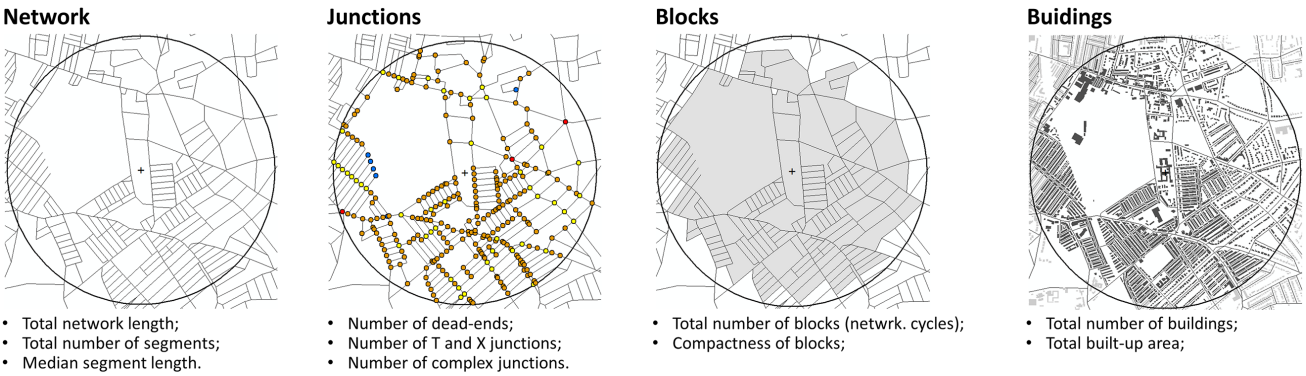

Figure 6. Graphic depiction of the adopted morphological indicators.

considered without jeopardizing the workings of the proposed method. The full list of the morphological indicators used in this study is as follows:

a) Network attributes: total network length (meters), total number of segments (straight street stretches) and ratio between the number of junctions and the number of segments.
b) Junction attributes: total number of junctions; number of T-junctions (three segments junction); number of X-junctions (four segments junction); number of complex junctions (more than four segments) and number of dead-ends.
c) Block attributes: total number of blocks (i.e., regions of space bounded by streets) and compactness of blocks (standardized area/perimeter ratio, yielding 1 for a circle and values close to 0 for thin, elongated shapes).
d) Building attributes: total number of buildings (i.e., count of buildings footprints polygons), total built-up area (i.e., sum of buildings footprints) and ratio between the built-up area and the number of buildings (contiguous buildings are represented as single polygons, being counted as a single feature; therefore, the denser and continuous a built tissue is, the greater the value of this ratio).

Each morphological indicator results in a single figure for each area surrounding a school. After standardization of all indicators as Z-scores, a first screening for potential collinearities reduced their number to just three, non-correlated variables, namely: the ratio between the number of junctions and street segments, or [JunctSegs]; the compactness of urban blocks, or [CycCompct]; and the ratio between the built-up area and the number of buildings, or [AreaNBuild]. Multicollinearity between variables used for unsupervised classification exercises should be avoided (Tan, Steinbach, & Kumar, 2005). Still, the three remaining variables are capable of describing the connectivity of the street network (through the density of junctions per street segment, [JunctSegs]), the general geometric shape of urban blocks (through their compactness, [CycCompct]) and the density and continuity of buildings (through the ratio [AreaNbuild], by the reasons mentioned before). These variables are then subjected to hierarchical cluster analysis.

Cluster analysis is a family of unsupervised classification methods aimed at dividing data into homogeneous classes (or clusters), so that the objects in a given class are more similar among themselves than to the objects in the other classes. It differs from supervised classification techniques (i.e., classification with a previous model or classification label), deriving the resulting classes only from the data itself; that is, as the result of intrinsic cleavages and associations between the data points and not of pre-defined classification criteria. Here we used Ward's minimum variance method (Tan et al., 2005), one of the most common hierarchical classification algorithms.

Ward's method starts with all objects separated and each object being a cluster; at each iteration, the two clusters the merging of which would lead to the minimum increase in total within-cluster variance are joined, becoming a single cluster. The process continues until all objects are merged into a single cluster. Figure 7 shows the resulting dendrogram; the length of the vertical lines represents the value of the inter-cluster dissimilarity between each cluster's two predecessors. Thus, one should look for a cutting level of the dendrogram where the vertical lines are all long and at which the number of clusters is parsimonious. Figure 7 also shows that, in our case, a division into four clusters seems optimal and this is indeed confirmed by the cubic clustering criterion (CCC; Sarle, 1983), whose value peaks at four clusters.

Having extracted these four clusters, we inspect their profiles on the three morphological variables (i.e., [JunctSegs], [CycCompct] and [AreaNBuild]), as well as the urban tissues to which they correspond. Figure 8 shows, for each cluster, an image of the case that is closer to the cluster's centroid (which may be considered its 'archetype') and also a chart depicting the values of each cluster's members on the three morphological variables (where the archetypal cases are represented by a thicker line). Visual inspection of the maps of each cluster's archetype reveals evident differences between the four urban tissues they describe. Also, the values of the members of each cluster on the three morphological variables

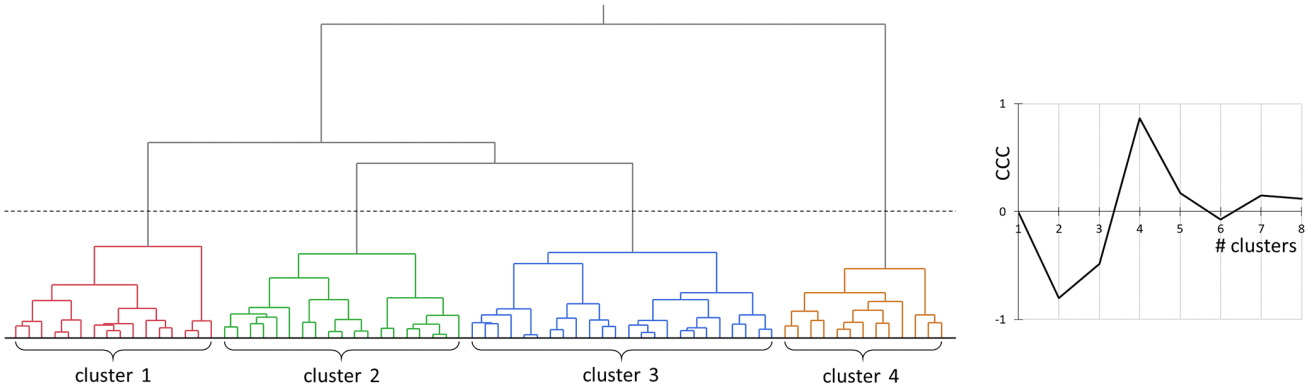

Figure 7. Hierarchical classification results.

(see lower charts on Figure 8) are quite similar within each cluster and clearly different between clusters. We can thus semantically characterize the four resulting morphological clusters in the following way:

Cluster 1 (n = 16): "Modern planned areas 1", composed of large and geometrically regular blocks (high average [CycCompct]). There are free-standing small buildings, creating a sparse urbanscape (low average [AreaNBuild]). There are very sparse street grids, with long street segments and few intersections (low average [JunctSegs]).

Cluster 2 (n = 20): "Modern planned areas 2", with blocks similar to cluster 1 (same average [CycCompct]). However, buildings (even if also small and separated) are more numerous in relation to Cluster 1 (higher [AreaNBuild]). The main difference between the two clusters is that in this one the street grid is denser, with more frequent junctions and shorter street segments (significantly higher average [JunctSegs]).

Cluster 3 (n = 25): "Early suburban developments", with more irregular and smaller urban blocks (lower [CycCompct]), and a more organic street grid. Buildings, although still mostly separated, are more densely organized (higher average [AreaNBuild], with greater variance). The average junction density ([JunctSegs]) is slightly lower than cluster 2, but its variance is higher.

Cluster 4 (n = 16): "Central historical fabric", with small elongated blocks (low [CycCompct]). There are densely packed, contiguous buildings (high [AreaNBuild]) in an organic and very dense street grid (high [JunctSegs]).

These morphological differences between the four clusters should correspond to also different epochs of urban expansion, with cluster 4 representing the older urban tissues and cluster 1 contemporary ones. This is indeed confirmed by looking at the spatial distribution of the four morphological clusters over the Liverpool City Region. Figure 9 shows that the members of cluster 4 are all located in the central area of the City of Liverpool and only on the east bank of the Mersey (which is the area of oldest occupation). Cluster 3 members immediately surround this central area, while appearing also on the west bank of the Mersey (i.e., on the Wirral Peninsula). On the right bank, Cluster 2 and 1 have mainly peripheral locations, with a greater incidence of cluster 1 on the farthest areas. On the Wirral peninsula this pattern is not so clear, because urbanization there is more recent and not so intensive.

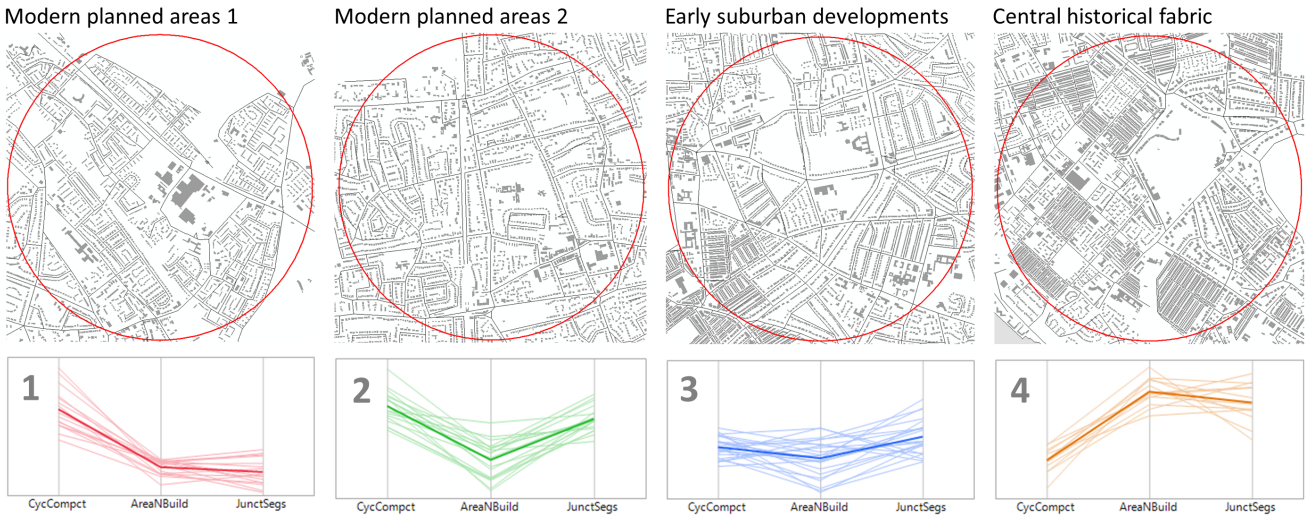

Figure 8. Archetypes and numerical profiles of the four morphological clusters.

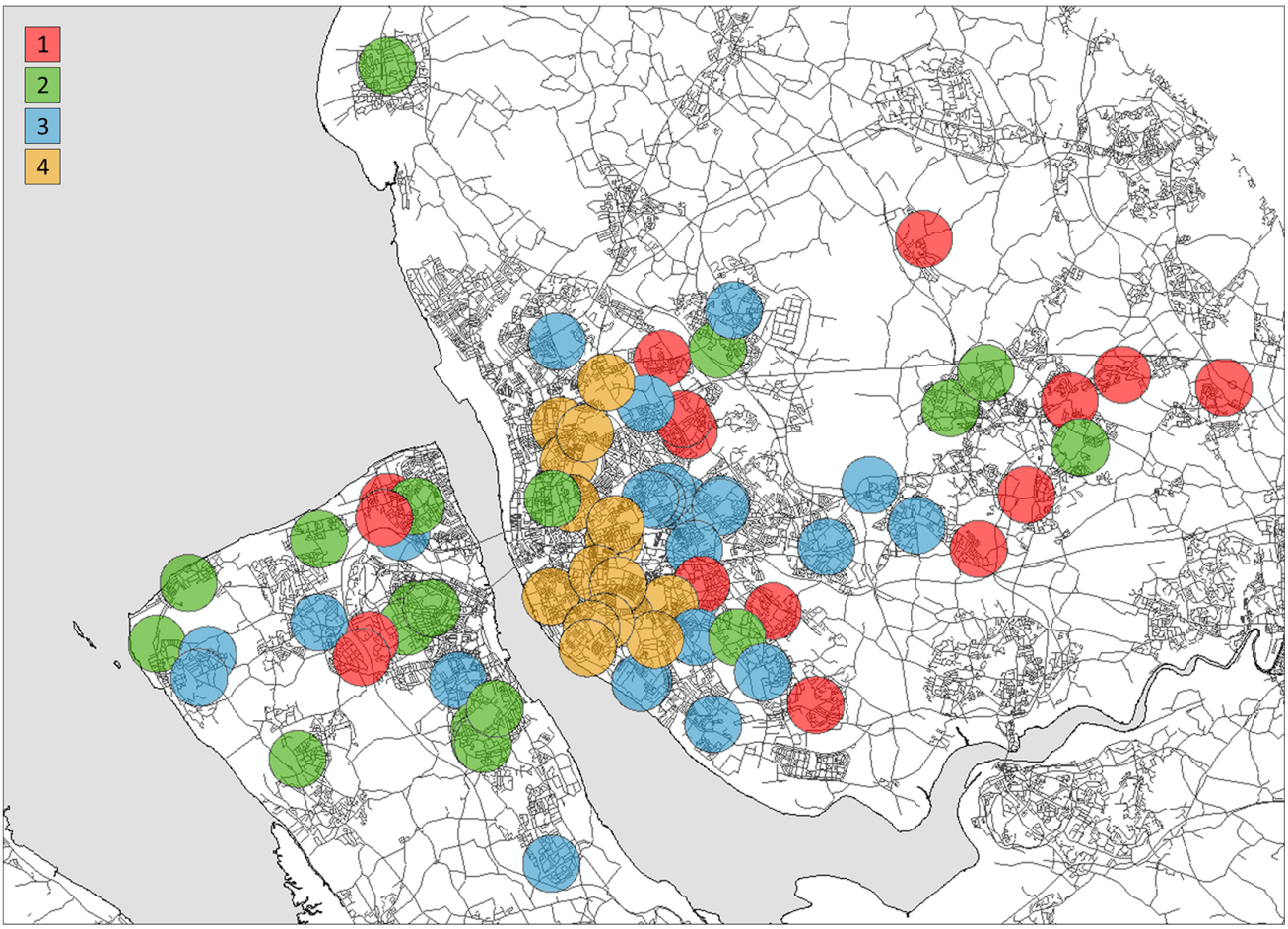

Figure 9. Geographical distribution of the four morphological clusters.

The proposed method for characterizing morphological urban contexts is therefore able to reduce the large initial variability of urban forms to a compact, yet meaningful categorization, of just four types of contexts. It takes into account detailed aspects of urban form, capable of detecting differences between urban tissues of different epochs and phases of development, which are otherwise difficult to classify objectively.

We summarize the morphological characterization of the 77 secondary schools by adding another layer to the socio-economic information displayed on Figure 5 (see Figure 10). For each school (i.e., each column of Figure 10) we record the morphological cluster to which it belongs, as well as its values (z-scores) on the three morphological variables that were used to define the clusters. With all the information produced by the proposed methods thus summarized, we can now use it to generate several context-informed purposive samples, aimed at different research questions and objectives.

5. Designing Context-Informed Purposive Samples of Urban Objects

In this section we use the data of the context characterization methods described before, in order to simulate four types of purposive samples proposed by Patton (1990), namely: *maximum variation sample*, *intensity sample* and two different types of *homogeneity samples* (Figure 11). Each of these illustrative samples is constituted by 16 observation units. This was the sample size that was used in the research project mentioned before, representing 23% of the whole population and being commensurate with the typical sample sizes of purposeful sampling.

When the population to be sampled is also small (as it is the case of the 77 Merseyside's secondary schools), random samples may not be an adequate way of achieving representativeness of the studied phenomena. In such cases a random sample has a non-negligible probability of not being representative at all, and a maximum variation sample may be a more efficient way of achieving representativeness. Maximum variation samples capture the extremes of a given set of characteristics, relevant for the problem under consideration. The logic behind this type of sampling is that any potential patterns found across the cases of such a sample, derive their significance from having emerged out of maximal heterogeneity (Patton, 1990).

In order to generate a maximum variation sample, we start by defining the dimensions along which varia-

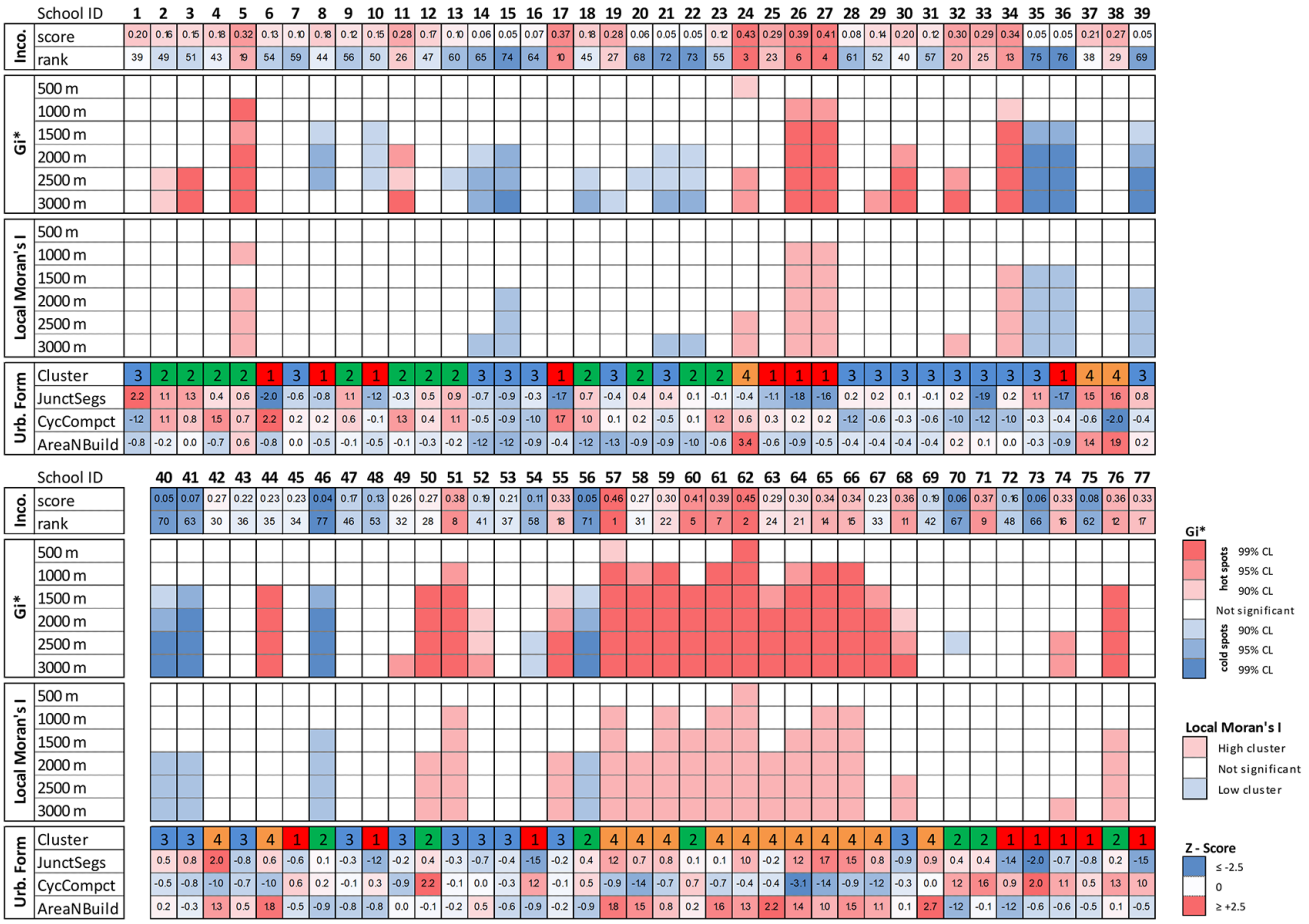

Figure 10. Social and physical contexts of the 77 secondary schools.

tion will be maximized; in our case, these are the social and physical composition of the urban contexts of the secondary schools. We have two extremes regarding social contexts: deprived and non-deprived, as defined in Section 3 (i.e., when such characteristics are verified across several spatial scales at once). Regarding material contexts, we have four possible types of variation, namely the four morphological clusters defined in Section 4. A maximum variation sample of urban contexts along these two dimensions with 16 observations, would therefore be composed by four cases of each morphological cluster, namely the two most deprived and the two least deprived (see Figure 11a).

Such a sample, covering the extremes of social and morphological variation observed in the population, would allow the investigation of the following research questions:

- Which regularities (if any) may be observed across all cases? These may be deemed general or transversal phenomena, independent of both social and morphological contexts.
- Which regularities (if any) may be observed across the four cases of each morphological cluster, independently of their social composition? These may be attributable to specific physical characteristics.
- Which regularities (if any) may be observed only on deprived and/or affluent cases, independently of their specific morphology? These may be attributable to specific social characteristics.

Intensity sampling aims at selecting cases in which the intensity of the phenomenon under investigation is maximized (Patton, 1990). In contrast to maximum variation, intensity sampling presupposes a previous observation or hypothesis to be further explored. For example, looking at Figure 10, it is clear that the majority of cases in cluster 4 are highly income-deprived (11 out of 16 cases), whereas the cases in cluster 1 tend to be rather less deprived (only 3 out of 16 cases showing high income deprivation). Independently of the causes behind such regularity, one may argue that high income deprivation is typical of the morphological contexts described by cluster 4 and atypical of those described by cluster 1. Furthermore, these two clusters are clearly the most separated in time and most dissimilar in morphological terms, with cluster 4 representing historical urban tissues and cluster 1 representing modern planned ones of the urban sprawl type. Thus, a sample composed of cluster's 4 deprived cases and cluster's 1 non-deprived cases would maximize the intensity of both social and morphological differences.

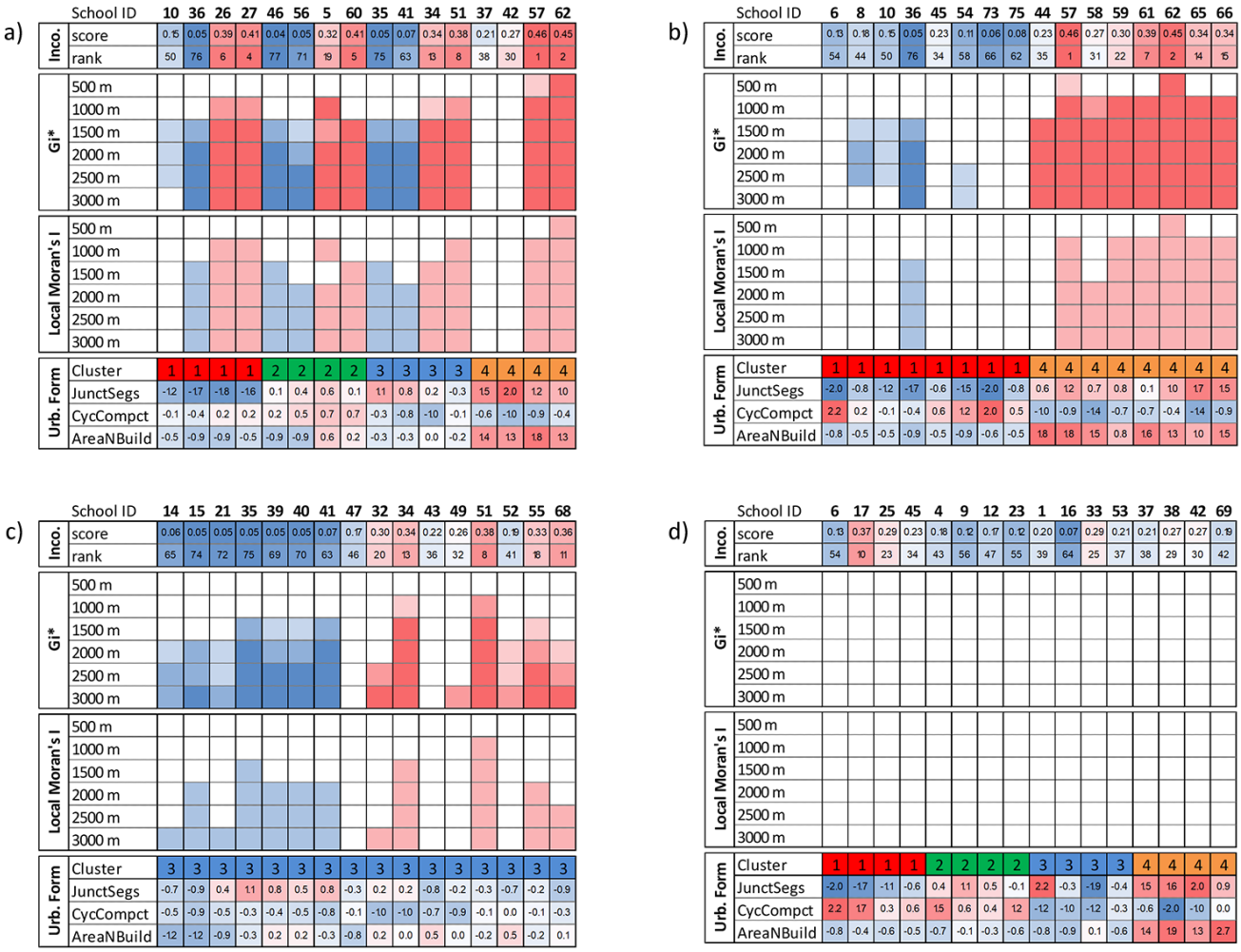

Figure 11. Four types of purposive samples: a) maximum variation sample, b) intensity, c) morphological homogeneity and d) socio-economic homogeneity samples.

An intensity sample with 16 observations would therefore be composed by the 8 most deprived cases of cluster 4, and the 8 least deprived cases of cluster 1 (see Figure 11, b). Such a sample would maximize both the probability of observing urban community inequalities and the intensity of their specific characteristics. It would allow the investigation of the following research questions:

- Which factors may explain the observed association between the two types of physical contexts and their specific social compositions?
- Which regularities (if any) are specific to each group of deprived and non-deprived cases?
- To which extent the specific material contexts of deprived and non-deprived cases are related to such regularities?

Finally, we propose two variations of 'homogeneity sampling' (Patton, 1990), a strategy which is the opposite of maximum variation sampling. Instead of maximizing variation, one tries to minimize it on one or several variables of interest. The purpose here is to study a given subgroup in depth, or to maintain the variability of a given dimension constant in order to reduce as far as possible its potential confounding effects.

We generate two different homogeneous samples: one in which we maintain the physical characteristics constant (Figure 11c); and another one in which we do the same regarding socio-economic characteristics (Figure 11d). In the first case, we select only cases belonging to cluster 3, half of them highly deprived and another half affluent. We choose to hold cluster 3 constant, because it is the most frequent morphological type (n = 25) with a high socio-economic variability, which we try to maximize by selecting only highly deprived and affluent cases. The objective is to study specifically the impacts of socio-economic composition of urban contexts, while maintaining urban form constant. Because all cases have similar physical contexts, we can be reasonably confident that any detected regularities would pertain to the socio-economic characteristics of the selected cases. Such a sample would allow the investigation of the following research question:

- Which are the specific impacts of deprived and affluent socio-economic contexts on urban communities, while controlling for urban form's potential confounding effects?

In the second case (Figure 11d), we select only cases showing not-significant spatial clustering of either deprivation or affluence, with income scores close to the mean, while choosing four cases of each morphological cluster. Conversely to the previous situation, the objective here is to study the specific impacts of physical characteristics, while maintaining socio-economic contexts constant and at an average level (i.e., neither particularly deprived nor affluent). Again, because all cases have similarly average socio-economic characteristics, but also quite different morphological contexts, one would expect that any regularities found would pertain to differences in physical context. This sample would allow the investigation of the following research question:

- Which are the specific impacts of different morphological contexts on urban communities, while controlling for potential socio-economic confounding effects?

We end this section by displaying the spatial distributions of the four sample simulations discussed above (Figure 12). Different samples result in also different spatial distributions, covering diverse parts of Liverpool City Region. Each sample serves different research objectives and none is a priori preferable over the others.

6. Conclusions

This article has proposed a set of GIS methods for quantifying, classifying and sampling the social and physical urban contexts of 77 secondary schools in Liverpool City Region, Merseyside, UK. The proposed methods overcome a number of shortcomings that current approaches to the characterization of urban contexts suffer from, namely: the exposure to the MAUP and its biasing effects; the rudimentary level at which urban form is commonly quantified and classified; and the lack of methodology in supporting purposive sampling, for exploring the complex relationships between urban contextual characteristics and other variables of interest.

Regarding the characterization of social urban contexts, and as the means to overcome the deleterious effects of MAUP, we make use of LISA, applied to avail-

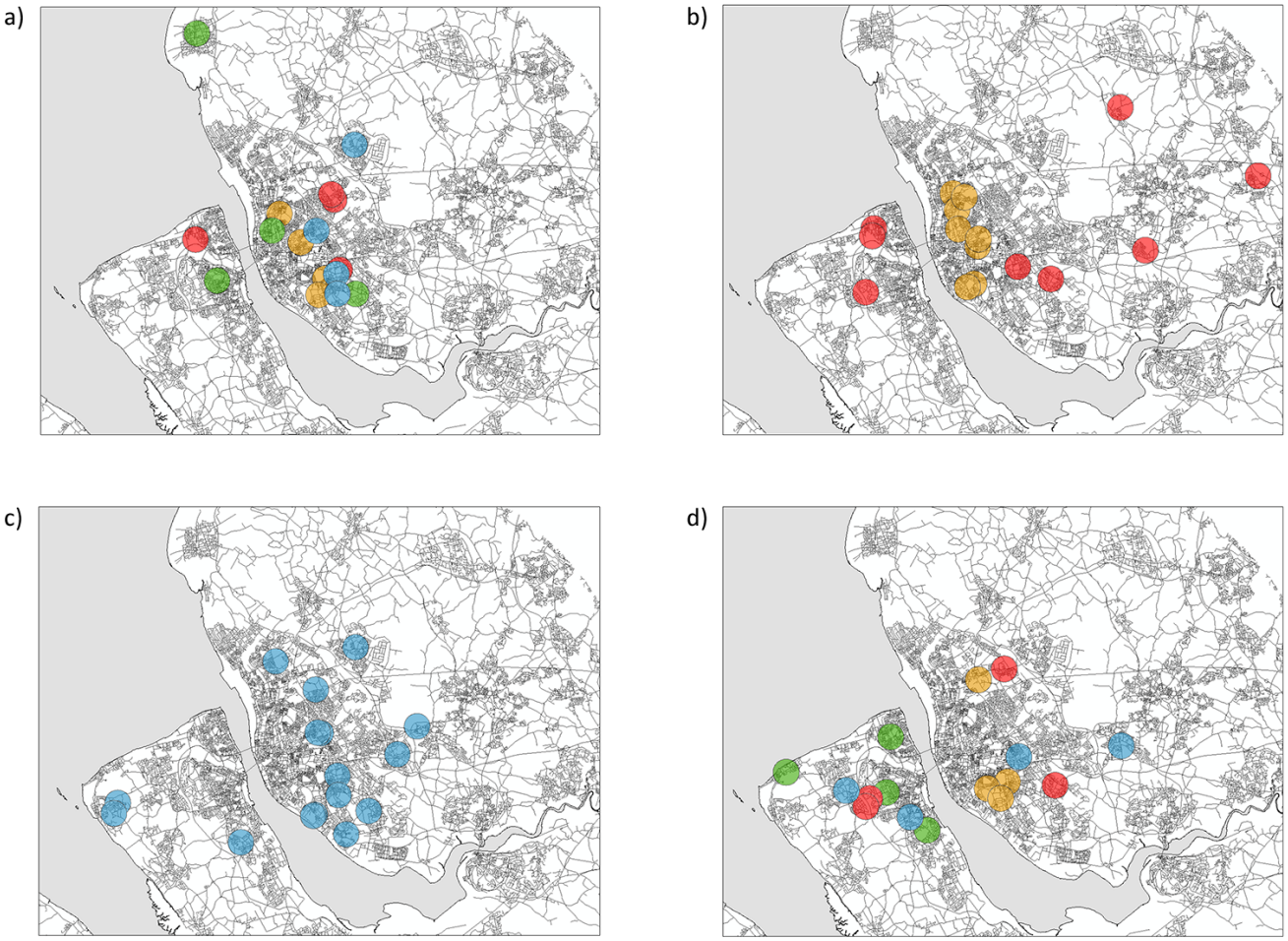

Figure 12. Geographical distributions of the four sample types; a) maximum variation, b) intensity, c) morphological homogeneity and d) socio-economic homogeneity.

able socio-economic indicators. We draw attention to the importance of evaluating the consistency of socio-economic indicators across several spatial scales, in order to identify accurately local areas where such indicators attain consistent high or low scores, as well as others where they do not. We apply simultaneously two LISA methods, namely the Gi* statistic and the Local Moran's I statistic, showing how their conjoint use is capable of providing detailed information about the specific social context of each school.

The physical characteristics of urban contexts are quantified through three morphological variables measured in GIS, namely the ratio between the number of junctions and street segments, the general geometric shape of urban blocks, and the density and continuity of buildings. We then use cluster analysis to objectively classify the physical context of each school, into a compact, yet meaningful categorization, of just four types of contexts: "modern planned areas 1", "modern planned areas 2", "early suburban developments" and "central historical fabric", each corresponding to different periods of urban expansion and types of geographical distribution in the history of Liverpool. By dividing data into classes that are derived by algorithmic means from the data themselves, this method overcomes the potential bias of pre-existing semantic classifications, while resulting in a high level of morphological detail.

Finally, the data generated by these methods is summarized into visualization schemes, revealing the relative variation of the social and physical contexts of the 77 schools. We use such schemes to produce four types of purposive samples, illustrating the design of context-informed samples of urban objects, aimed at different potential research questions in community and neighbourhood studies. We note that purposive sampling strategies, even though generally overlooked, can be extremely useful for exploring the inherently complex relationships between urban context and other variables of interest. The current focus on probabilistic sampling techniques, in its endeavour to find generalizable effects, is perhaps not the best initial approach to such intricate and elusive phenomena. We suggest that purposive sampling strategies, by virtue of selecting specific information-rich cases, may be more fruitful for exploring the potential impacts of different urban contexts, whose generality may subsequently be tested with larger probabilistic samples.

This work responded to the research objectives of visualising and measuring social inequalities in Liverpool's urban environments as part of a specific research project. However, the proposed methods are not limited to the chosen variables or urban objects (schools), do not depend on geographical context and can address a larger range of dimensions without loss of consistency. On the contrary, they can provide a robust and efficient methodology on comparative profiling and sampling of a wide range of socio-economic factors and urban forms, across time, scale and contexts.

Acknowledgements

The research project "Visualizing Inequalities in Community Networks to Enhance Participatory Planning" was supported by the Leverhulme Trust Research Project Grant RPG-2014-169. Miguel Serra is supported by the FCT grant SFRH/BPD/111260/2015.

Conflict of Interests

The authors declare no conflict of interests.

References

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115.

Barthelemy, M. (2015). From paths to blocks: New measures for street patterns. *Environment and Planning B: Urban Analytics and City Science*, *44*(2), 256–271.

Caughy, M., Leonard, T., Beron, K., & Murdoch, J. (2013). Defining neighborhood boundaries in studies of spatial dependence in child behaviour problems. *International Journal of Health Geographics*, *11*(24), 1–12.

Charron, M. (2009). Neighbourhood characteristics and the distribution of police-reported crime in the city of Toronto. *Crime and Justice Research Paper Series*. Ottawa: Statistics Canada.

Crane, R. (2000). The influence of urban form on travel: An interpretive review. *Journal of Planning Literature*, *15*(1), 3–23.

Cummins, S., Macintyre, S., Davidson, S., & Ellaway, A. (2005). Measuring neighbourhood social and material context: Generation and interpretation of ecological data from routine and non-routine sources. *Health & Place*, *11*, 249–260.

Cutrona, C., Wallace, G., & Wesner, K. (2006). Neighborhood characteristics and depression: An examination of stress processes. *Current Directions in Psychological Science*, *15*(4), 188–192.

DCLG (2015). *The English indices of deprivation 2015—Technical report*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/464485/English_Indices_of_Deprivation_2015_-_Technical-Report.pdf

Dibble, J., Prelorendjos, A., Romice, O., Zanella, M., Strano, E., Pagel, M., & Porta, S. (2017). On the origin of spaces: Morphometric foundations of urban form evolution. *Environment and Planning B: Urban Analytics and City Science*. https://doi.org/10.1177/2399808317725075

ESRI. (2011). ArcGIS Desktop: Release 10.3: Redlands, CA: Environmental Systems Research Institute.

Gambaro, L., Joshi, H., Lupton, R., Fenton, A., & Lennon, M. C. (2016). Developing better measures of neighbourhood characteristics and change for use in studies of residential mobility: A case study of Britain in the early 2000s. *Applied Spatial Analysis*, *9*, 569–590.

Gil, J., Beirao, J., Montenegro, N., & Duarte, J. (2011). On the discovery of urban typologies: Data mining the many dimensions of urban form. *Urban Morphology*, *16*(1), 27–40.

Hamaina, R., Leduc, T., & Moreau, G. (2012). Towards urban fabrics characterization based on buildings footprints. In J. Gensel, D. Josselin, & D. Vandenbroucke (Eds.), *Bridging the Geographic Information Sciences, Lecture Notes in Geoinformation and Cartography (LNG&C)* (pp. 327–346). Berlin and Heidelberg: Springer.

Inoue, Y., Stickley, A., Yazawa, A., & Shirai, K. (2016). Neighbourhood Characteristics and Cardiovascular Risk among Older People in Japan: Findings from the JAGES Project. *PloSONE*, *11*(10), 1–16.

Kim, R., Ali, M., Sur, D., Khatib, A., & Wierzba, T. (2012). Determining optimal neighborhood size for ecological studies using leaving-one-out cross validation. *International Journal of Health Geographics*, *11*(10), 1–6.

LCC. (2015). *The index of multiple deprivation 2015: A Liverpool analysis*. Liverpool: Liverpool City Council.

Lebel, A., Pampalon, R., & Villeneuve, P. (2007). A multiperspective approach for defining neighbourhood units in the context of a study on health inequalities in the Quebec City region. *International Journal of Health Geographics*, *6*(27), 1–15.

Lupton, R. (2003). *Neighbourhood effects: Can we measure them and does it matter?* (CASE paper 73). London: Centre for Analysis of Social Exclusion, London School of Economics.

MacDonald, J., Wise, M., & Harris, P. (2008). *The health impacts of the urban form: A review of reviews*. Retrieved from http://hiaconnect.edu.au/old/files/Health_Impact_of_Urban_Form.pdf

Marshall, S. (2005). *Streets & patterns*. London and New York: Spon Press.

Miles, R., Coutts, C., & Mohamadi, A. (2011). Neighborhood urban form, social environment and depression. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, *89*(1), 1–18.

O'Brien, J., Serra, M., Hudson-Smith, A., Psarra, S., Hunter, A., & Zaltz Austwick, M. (2016). Ensuring VGI credibility in urban-community data generation: A methodological research design. *Urban Planning*, *1*(2), 88–100.

O'Brien, J., Garcia Vélez, L., & Zaltz Austwick, M. (2017). Visualizing the impacts of movement infrastructures on social inclusion: Graph-based methods for observing community formations in contrasting geographic contexts. *Social Inclusion*, *5*(4), 132–146.

Office of National Statistics. (2011). Census geography: An overview of the various geographies used in the production of statistics collected via the UK census. Retrieved from https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography

Openshaw, S. (1983). *The modifiable areal unit problem*. Norwick: Geo Books.

Ord, J., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, *27*(4), 286–306.

OS. (2015). VectorMap Local. Dataset. Retrieved from https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-map-local.html

Patton, M. (1990). *Qualitative evaluation and research methods*. Beverly Hills, CA: Sage.

Pont, M. B., & Haupt, P. (2010). *Spacematrix: Space, density and urban form*. Roterdam: NAi Publishers.

Rae, A. (2012). Spatially concentrated deprivation in England: An empirical assessment. *Regional Studies*, *46*(9), 1183–1199.

Sarle, W. (1983). *Cubic clustering criterion* (SAS Technical Report, 108). Cary, NC: Statistical Analysis System Institute.

Serra, M., Gil, J., & Pinho, P. (2016). Towards an understanding of morphogenesis in metropolitan street-networks. *Environment and Planning B: Urban Analytics and City Science*, *44*(2), 272–293.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (1st ed.). Boston, MA: Addison-Wesley Longman Publishing.

Timperio, A., Jeffery, R., Crawford, D., Roberts, R., Giles-Conti, B., & Ball, K. (2010). Neighbourhood physical activity environments and adiposity in children and mothers: A three-year longitudinal study. *International Journal of Behavioral Nutrition and Physical Activity*, *7*(18), 1–8.

Townshend, T., & Lake, A. (2009). Obesogenic urban form: Theory, policy and practice. *Health & Place*, *15*, 909–916.

van Ham, M., & Manley, D. (2012). Neighbourhood effects research at a crossroads: Ten challenges for future research. *Environment and Planning A*, *44*, 2787–2793.

Vernez-Moudon, A. (1994). Getting to know the built landscape: Typomorphology. In A. Frank (Ed.), *Ordering space: Types in architecture and design* (pp. 289–314). New York: Van Nostrand Reinhold.

About the Authors

Miguel Serra is Senior Researcher at CITTA (Research Centre for Territory, Transports and Environment, University of Porto) and Honorary Research Associate at the Bartlett School of Architecture (University College London, UCL). His current research focuses on large spatial networks, quantitative urban morphological descriptions and classifications, geographic information systems and urban data analysis and visualization. He was Senior Research Associate at UCL (2015–2016) in the research project "Visualizing Inequality in Community Networks to Enhance Participatory Planning", funded by the Leverhulme Trust.

Sophia Psarra is Reader at UCL. Her latest book, *The Venice Variations* (UCL Press, 2018), explores spatial and social networks as multi-authored processes of formation, alongside the ways in which they interact with urban design based on individual intention. Psarra was co-investigator in the research project "Visualising Inequality in Community Networks to Enhance Participatory Planning" (UCL) funded by the Leverhulme Trust; and co-principal investigator in a project exploring spatial, social networks and organisational innovation, funded by the NSF (the University of Michigan).

Jamie O'Brien is Hon. Research Associate at Bartlett Centre for Advanced Spatial Analysis (CASA), University College London, and is a consultant in GIS data visualization. He was Senior Research Associate at CASA (2014–2017), where he focused on communities' uses of local spatial assets in achieving their sense of place. He holds a Doctorate in Engineering from UCL's Bartlett School of Architecture (supported by an EPSRC studentship; 2002–2006), and is Fellow of the Higher Education Academy.