Venerandi et al. *EPJ Data Science* (2018) 7:4
https://doi.org/10.1140/epjds/s13688-018-0132-1

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

A scalable method to quantify the relationship between urban form and socio-economic indexes

Alessandro Venerandi[1]*, Giovanni Quattrone[2] and Licia Capra[2]

*Correspondence:
alessandro.venerandi.12@ucl.ac.uk
[1] Department of Civil, Environmental & Geomatic Engineering, UCL, London, UK
Full list of author information is available at the end of the article

Abstract

The world is undergoing a process of fast and unprecedented urbanisation. It is reported that by 2050 66% of the entire world population will live in cities. Although this phenomenon is generally considered beneficial, it is also causing housing crises and more inequality worldwide. In the past, the relationship between design features of cities and socio-economic levels of their residents has been investigated using both qualitative and quantitative methods. However, both sets of works had significant limitations as the former lacked generalizability and replicability, while the latter had a too narrow focus, since they tended to analyse single aspects of the urban environment rather than a more complex set of metrics. This might have been caused by the lack of data availability. Nowadays, though, larger and freely accessible repositories of data can be used for this purpose. In this paper, we propose a scalable method that delves deeper into the relationship between features of cities and socio-economics. The method uses openly accessible datasets to extract multiple metrics of urban form and then models the relationship between urban form and socio-economic levels through spatial regression analysis. We applied this method to the six major conurbations (i.e., London, Manchester, Birmingham, Liverpool, Leeds, and Newcastle) of the United Kingdom (UK) and found that urban form could explain up to 70% of the variance of the English official socio-economic index, the Index of Multiple Deprivation (IMD). In particular, results suggest that more deprived UK neighbourhoods are characterised by higher population density, larger portions of unbuilt land, more dead-end roads, and a more regular street pattern.

Keywords: Urban form; Socio-economics; Spatial analysis; Open data; OpenStreetMap

1 Introduction

Cities are growing faster than ever before. In 1950, only 30% of the total world population was living in cities. Today, this datum stands around 54%. By 2050, the estimates project that 66% of the total world population will be urban, with cities in developing countries attracting the greatest number of new city dwellers [1]. Urbanisation is regarded by institutions and governments as a positive phenomenon as it brings, for example, better and less costly public services and improved living standards due to the concentration of economic activities [2]. However, this very same phenomenon is also reported to bring more inequality worldwide, with some areas benefiting more from public investments and economic growth than others [3]. It is thus necessary to develop a better understanding of the relationship between physical features of the urban environment and socio-economic levels of city dwellers, to inform urbanists and city planners.

Urban form had been investigated extensively in the past, in relation to socio-economics and well-being. Jacobs, for example, observed different parts of her city (i.e., New York) and reached the conclusion that the traditional, compact, pedestrian-friendly city form would have ensured the overall well-being of city dwellers [4]. Swiss architect Le Corbusier formulated a theory based on his personal perspective and reached an opposite conclusion, instead. In his view, the optimal city form was dispersed and more car-oriented [5]. The main limitation of these works lies in the use of qualitative methodologies which render studies difficult to repeat, and outcomes to generalise. More recently, researchers adopted quantitative methods to study the relationship between features of the urban environment and socio-economic aspects (see, for example, [6] and [7]). The main limitations of these works is that they analysed relatively small geographic areas (e.g., single neighbourhoods) and focused on single aspects of the urban environment (e.g., place accessibility) despite the fact that urban form is, by definition, the interplay of multiple elements and these should be thus studied together to best capture the socio-economics of city neighbourhoods.

The choice of using qualitative methods or analysing single metrics might have been dictated by the lack of available data. In the last decade, though, large data repositories and new techniques of data collection (e.g., crowd-sourcing) have become readily available (i.e., "open data revolution" [8, 9]). Researchers have recently started to take advantage of this and study cities through quantitative methods. For example, they analysed crowd-sourced visual perceptions of different urban environments in relation to socio-demographic factors [10] and urban qualities, such as beauty [11].

Inspired by this set of works, which analysed urban form in a more comprehensive manner, and by taking advantage of the "open data revolution" too, we propose a quantitative method that uses openly accessible datasets and spatial regression analysis to study the relationship between multiple features of urban form and socio-economic indexes. Unlike previous works, our method (i) relies on multiple descriptors of the urban environment and (ii) can be applied to cities of different sizes and repeatedly over time, at almost no cost.

To test the proposed method, we applied it to the UK major urban areas as identified by an official document (i.e., London, Manchester, Birmingham, Liverpool, Leeds, and Newcastle) [12]. Both features of urban form and information on the socio-economic levels of their communities were extracted from openly accessible datasets. The former were extracted from Ordnance Survey (OS) VectorMap District and OpenStreetMap (OSM), while the latter was obtained from the English IMD database. Outcomes of the spatial models could explain between 27% and 70% of the variance of IMD, confirming the existence of a relationship between urban form and socio-economics across the six cities. Furthermore, we found some aspects of urban form to have similar behaviours across the case studies thus highlighting some common patterns. In particular, more deprived neighbourhoods of urban UK were found to be characterised by higher population density, more unbuilt land, a higher presence of dead-end roads, and a more regular street pattern. The method proposed in this paper and its outcomes can be helpful in the current urbanisation age as they constitute a data driven basis for reasoning on possible design schemes and urban policies.

The remainder of this paper is structured as follows. We firstly illustrate related works. We then present our method, starting from the metrics of urban form and the concept of socio-economic index, and by then following with the type of analysis conducted. We discuss the results of its application to the six UK cities under study, before offering our interpretations. We conclude with final remarks and limitations.

2 Related works

As we presented in the Introduction, the relationship between urban form and socio-economic outcomes has been investigated in a variety of ways. Several authors used qualitative methods based on observation or personal views. These authors can be subdivided in two schools of thought: those supporting the compact, pedestrian city form and those favouring more spread out and car-oriented urban environments. Jacobs belongs to the former group as she supported the features of the traditional city, that is medium to high built density, perimeter blocks, walk-ability, and mixed-use [4]. Similarly, Whyte praised human scale streets, walk-ability, and argued that subtle urban details, such as shop widows, porticoes, steps, and doorways, were indispensable for city liveability [13]. Gehl is on the same page and favours pedestrian mixed-use streets, as well as active *city edges* (i.e., block frontages), which, in his view, can promote social interactions as well as stimulate commercial activities [14]. The other school of thought was openly against the traditional city form instead. Le Corbusier, for example, deemed it as disordered, chaotic, and unhealthy [5]. He proposed a city based on super blocks (i.e., urban blocks of big dimension) delimited by multi-lane highways and residential tower blocks retracted from the sidewalks and laid out in open space (the so-called "towers in the park"). Similarly, Hilberseimer despised the traditional city form and proposed plans characterised by a grid of highways and repetitive, tall residential blocks aligned to them [15]. Although these works presented insightful perspectives, they lack generalizability as they were based on qualitative methods. Moreover, they are hardly replicable as mostly based on personal views and observations.

More recently, thanks to the diffusion of computers and Geographic Information Systems (GIS), the study of cities has become also quantitative. Vaughan *et al.*, for example, adopted a renowned technique for the analysis of street networks (i.e., Space Syntax [16]) to study the relationship between *integration* (i.e., a measure of spatial accessibility) and socio-economic levels of East London residents [6]. The researchers reported that more accessible places, such as main streets, were associated with more affluent residents while less accessible ones, such as back streets and interstitial spaces, were related to less advantaged citizens. Other scholars focused on urban form and criminal activity and studied the relationship between dwelling typologies and crime occurrences in a London borough [17]. They found the flat to be the safest house type. Researchers also separately investigated density in relation to crime; however, they found discordant outcomes. Some reported absence of any relationship [18, 19], while a more recent work found that density was overall beneficial against crime [17]. Hillier studied whether a specific configuration of the street network, that of cul-de-sac, was associated with more or less crimes in a London neighbourhood. Results suggested that cul-de-sac did not attract more crimes than other spatial configurations when integrated in a street network with significant through movement and many properties facing the streets [7]. Other researchers focused on the relationship between social aspects of urban life and the configuration of cities. Similar to density, outcomes were contrasting. Some scholars reported that higher densities improved social interactions [20] and that lower densities not only decreased social ties, but also favoured more car-oriented behaviours [21]. Contrary to this, other researchers found that higher densities diminished the will of people to socialize and increased stress [22–24]. Although these studies were based on quantitative methods rather than qualitative ones, findings are still hardly generalizable as the geographic contexts under study were limited (e.g., a single neighbourhood, a specific city). This was mainly due to a lack of data as most of it was proprietary and thus hardly accessible. Furthermore, these works mainly studied separate aspects of urban form in relation to socio-economics (e.g., spatial accessibility and social class in the work by Vaughan *et al.* [6]). However, cities are constituted by many interrelated components, which act in synergy rather than in isolation [25, 26].

A more recent line of research exploited the increased presence of data to investigate the relationship between the population size of multiple cities across the world and several socio-economic indicators. For example, researchers analysed the relationship between population size and wages, GDP, number of patents produced, number of research centres, in multiple cities across the world and found that these indicators all increased by around 15% more than the expected linear growth [27], with some deviations due to local, city-specific dynamics [28]. These studies clearly offered important insights on the scaling laws of urban settlements. However, they considered cities as macro-entities. Our aim, which is aligned with the Urban Morphology discipline [25, 26], is to analyse the relations between urban form and socio-economics at a much smaller scale instead, that of city neighbourhoods and their multiple basic components (e.g., connectivity of streets, population density). Moreover, our focus is on dynamics within cities rather than across cities.

In the last decade, given the increased availability of open data repositories and the rise of new data collection techniques (e.g. crowd-sourcing), the limitations associated with paucity of data have almost disappeared. Indeed, researchers have started to analyse cities in more comprehensive manners, for example, through the analysis of pictures, which, by default, comprise of multiple elements of the built environment. Quercia *et al.* used crowd-sourcing to ask more than 3000 respondents whether photos of different urban environments transmitted beauty, quietness, and happiness [11]. The researchers then used this information to understand what visual features (i.e., colours, textures) best correlated with these qualities and found that the colour green was positively correlated, while wide roads and faceless buildings were inversely correlated. Selasses *et al.* used the same data collection technique (i.e., crowd-sourcing) to ask more than 7000 people to rate their visual perceptions of street views in terms of safety, social status, and uniqueness [10]. The group of researchers then compared the responses to socio-demographic data and found that spatial dissimilarities in the perception of safety and social status better correlated with violent crimes than their absolute values. Furthermore, they found that, safety perception being equal, these crimes were more related to areas that looked more upper class. Other researchers used a computer vision algorithm, that analysed time-series street-view imagery of five US cities, to study the relationship between physical changes of city neighbourhoods and demographic data (i.e., population density and share of residents with college education) [29]. Results from this study suggested that: (i) neighbourhoods more densely populated by adults with college education were more likely to undergo processes of urban change; (ii) neighbourhoods that had better initial appearances tended to undergo larger positive improvements; and (iii) positive neighbourhood change was associated with proximity to the central business district and to other aesthetically attractive neighbourhoods. Although these works analysed multiple aspects of urban form at the same time, they focused on "point-based" data (e.g., the characteristics of the urban space which can be seen in a single picture). What is still missing is a method that enables the analysis of multiple aspects of urban form at an "area-based" level. We present next the details of a method that permits this type of spatial analysis, in a scalable manner.

3 Method

The methodology we propose mainly consists of two parts: (i) computation of metrics of urban form and socio-economics aggregated at areal level and (ii) quantitative analysis based on spatial linear regression to understand the relationship between the features of urban form and socio-economic levels of neighbourhoods. We present next the metrics and the procedural steps for carrying out the analysis, before describing how we applied it in practice.

3.1 Metrics

3.1.1 Urban form

Our method requires the computation of nine different metrics, capturing different aspects of the built environment. Five of these were derived from previous works, while four are being proposed in this paper. The five metrics derived from previous works are:

- Connected Node Ratio (CNR). This measures the level of connectivity and walk-ability of a street network and was derived from the work by Garrick and Marshall [30]. CNR is computed as the ratio between the number of street intersections that are not cul-de-sac (i.e., node degree equals to 1) and the total number of street intersections in an area:

  $$\mathrm{CNR}(a_k) = \frac{\sum_{i\neq 1} \mathrm{count}(n_i, a_k)}{\sum_i \mathrm{count}(n_i, a_k)},$$

  where $\mathrm{CNR}(a_k)$ is the Connected Node Ratio of the area a_k; $\mathrm{count}(n_i, a_k)$ represents the number of street intersections having a degree equal to i in the area a_k. We include this metric as several different authors, including Jacobs [4] and Gehl [14], considered connectivity and walk-ability fundamental aspects for thriving neighbourhoods. These not only positively affected the health conditions of citizens, but also improved social interactions, commercial activities, and provided informal protection against crime.
- Intersection Density (ID). This metric quantifies the density of street intersections in city areas. As for CNR, also ID has been extracted from the work by Garrick and Marshall [30]. Intersection Density is computed as the ratio between the total number of intersections lying in an area and the area of such region:

  $$\mathrm{ID}(a_k) = \frac{\sum_i \mathrm{count}(n_i, a_k)}{\mathrm{area}(a_k)},$$

  where $\mathrm{ID}(a_k)$ is the Intersection Density of the area a_k; $\mathrm{count}(n_i, a_k)$ represents the number of street intersections having a degree equal to i in the area a_k; $\mathrm{area}(a_k)$

denotes the total area, in square meters, of the city area a_k. The reason for including this metric is similar to the one stated above as ID and CNR are closely related. Generally, a dense street network tends also to be more connected and walk-able and thus be associated with the positive aspects mentioned above (i.e., better health conditions, more social and economic benefits, more control against crime [4, 14]).

- Percentage of Unbuilt Land (PUL). It measures the amount of land that is not covered by buildings and was derived from previous work by Banister *et al.* [31]. PUL is calculated by dividing the amount of unbuilt land in a city area, by the overall area of such region, and by then multiplying this value by 100:

$$\mathrm{PUL}(a_k) = \frac{A_u(a_k)}{A_T(a_k)},$$

  where PUL(a_k) is the Percentage of Unbuilt Land of the area a_k; $A_u(a_k)$ is the unbuilt land, in square meters, in the city area a_k; $A_T(a_k)$ represents the total area, in square meters, of the city area a_k. PUL provides information on the occupancy ratio of land. Smaller values of PUL mean that there are fewer buildings on the area considered and thus there is more unbuilt land. Conversely, greater values of PUL mean that most of the area is occupied by buildings and a small area is left unbuilt. We include this metric as the occupancy ratio of land was considered a relevant design aspect by different researchers. Modernist planners, for example, favoured few buildings surrounded by open space (i.e., "towers in the park") [5, 15]. Authors supportive of the compact city form favoured a more continuous urban fabric with less unbuilt area instead [4, 13, 14].
- Population Density (PD). It quantifies how densely populated is a city area. PD is a common statistical datum used by institutions and governments. It is computed as the ratio between the number of residents living in an area and the area of such region:

$$\mathrm{PD}(a_k) = \frac{R(a_k)}{A_T(a_k)},$$

  where PD(a_k) is the Population Density of the city area a_k; $R(a_k)$ is the number of residents of the city area a_k; $A_T(a_k)$ represents the total area, in hectares, of the city area a_k. PD together with PUL provides information on how built density is distributed across an area. For example, a neighbourhood with a big portion of unbuilt land and a high population density is likely to be characterised by residential towers. Conversely, a neighbourhood with small unbuilt surface and a high population density is likely to be characterised by perimeter blocks.
- Betweenness Centrality (BC). It measures the level of centrality of streets. To be more specific, BC is based on the concept that a street is central if it is included in many of the shortest paths linking pairs of nodes (street intersections) in a street network. The formula for computing BC can be found in the work by Porta *et al.* [32]. BC is usually computed for street segments. However, since our method focuses on areas rather than streets, we aggregate BC by considering the maximum value for each area. We argue that the maximum value of BC can be representative of the level of spatial accessibility of different areas with respect to the overall urban region. The formula

for aggregating BC for areas is as follows:

$$\mathrm{BC}(a_k) = \max(\mathrm{BC}_\alpha), \quad \text{with } \alpha \text{ a street segment contained in } a_k,$$

where $\mathrm{BC}(a_k)$ is the BC value of the area a_k and $\max(\mathrm{BC}_\alpha)$ represents the maximum value of BC of a street segment contained in a_k. We include BC as previous works showed it to be associated with positive aspects of cities such as employment density [33], agglomeration of economic activities [34], and street quality [35].

We present next four metrics of urban form that we propose in this paper to complement the previous ones:

- Percentage of Green Areas (PGA). This metric quantifies the amount of public green space available in a city area. It is computed by dividing the amount of green space in an area by the total area of such region and by then multiplying this value by 100:

  $$\mathrm{PGA}(a_k) = \frac{A_g(a_k)}{A_T(a_k)},$$

  where $\mathrm{PGA}(a_k)$ is the Percentage of Green Areas of the city area a_k; $A_g(a_k)$ is the amount of green areas, in square meters, in the city area a_k; $A_T(a_k)$ represents the total area, in square meters, of city area a_k. We include this metric as several authors deemed the presence of greenery an important aspect for city neighbourhoods. Jacobs, for example, argued that parks and garden positively affected city liveability. However, she also pointed out that they could potentially have negative effects, particularly in terms of safety, if these were relegated to peripheral areas with low densities [4].
- Irregularity of the Street Network (ISN). It measures to what extent the street network of a specific city area is irregular. ISN is computed by dividing the standard deviation of the node degrees associated with the intersections lying within an area by the average node degree relative to the same intersections:

  $$\mathrm{ISN}(a_k) = \frac{\sigma(a_k)}{\mu(a_k)},$$

  where $\mathrm{ISN}(a_k)$ represents the Irregularity of the Street Network in the city area a_k; $\sigma(a_k)$ is the standard deviation of the node degrees, in the area a_k; $\mu(a_k)$ represents the average of the node degrees, in the area a_k. Intuitively, a small ISN reflects an area with a small variation in node degrees, for example, an area characterised by a grid layout, where the majority of street intersections are four-way ones. Conversely, a great ISN corresponds to an area with a greater variation in node degrees, for example, that of an area characterised by a mix of different street intersections (e.g., cul-de-sac, three-way intersections, four-way intersections, six-way intersections). We consider this metric as the configuration of the street network was another aspect deemed important for city liveability by several authors. For example, Jacobs generally favoured the grid layout [4]. However, she also argued that this had to be interrupted by squares or diagonal roads as, in her view, these urban elements would have offered "visual interruptions" that enhanced urban life.

- Dead-end Density (DD). It measures the density of dead-end roads (cul-de-sac) in a specific city area. It is calculated as the ratio between the number of cul-de-sac lying in an area and the area of such region:

$$\mathrm{DD}(a_k) = \frac{\mathrm{count}(n_1, a_k)}{\mathrm{area}(a_k)},$$

  where $\mathrm{DD}(a_k)$ is the Dead-end Density of the city area a_k; $\mathrm{count}(n_1, a_k)$ represents the total number of nodes with degree 1 (i.e., cul-de-sac), in the area a_k. We chose this metric as the presence of dead-end roads in neighbourhoods attracted the attention of several authors. On the one hand, Jacobs argued that cul-de-sac negatively affected urban liveability as they diminished connectivity and thus the positive effects linked to it (e.g., better health and socio-economic outcomes, more safety against crime) [4]. On the other, Newman supported cul-de-sac as a reduced connectivity would have had positive outcomes, especially in terms of safety, as fewer strangers and more locals would have walked in neighbourhoods [36].
- Offering Advantage of Historic Properties (OAHP). It quantifies whether a city area offers more or less residential properties that are historic (i.e., built pre-1900), compared to the average city area. 1900 is selected as temporal threshold as, from roughly that point on, the modernist style started to unfold [37] and thus properties built in the period following this year cannot be considered historic. Note that this threshold is mainly valid for the European context. To compute this metric, we use a formula adopted in a previous work (i.e., Offering Advantage) [38] that showed to be effective in capturing variations in the offering of urban elements across a city. In this paper, Offering Advantage is adapted to reflect to what extent a city area a_k offers more historic property h_i, compared to the average area. More specifically:

$$\mathrm{OA}(h_i, a_k) = \frac{\mathrm{count}(h_i, a_k)}{\sum_{j=1}^{N} \mathrm{count}(h_j, a_k)} \cdot \frac{\sum_{j=1}^{N} \mathrm{count}(h_j)}{\mathrm{count}(h_i)},$$

  where $\mathrm{OA}(h_i, a_k)$ corresponds to the Offering Advantage of historic property h_i in the area a_k; $\mathrm{count}(h_i, a_k)$ represents the number of historic property h_i in the area a_k; N is the total number of historic properties; $\mathrm{count}(h_i)$ is the number of historic property h_i across a whole city. OAHP can be considered a proxy for the traditional urban form. The more an area offers historic properties, the more likely is that such area is characterised by features of the traditional compact city form (e.g., density, connectivity, perimeter blocks). We include OAHP as compactness and distribution of built density were deemed fundamental aspects that affected urban life by several authors. Jacobs, for example, supported the traditional compact city form as, in her view, it enhanced social tights, commercial activities, and safety [4]. Modernist architects, such as Le Corbusier, on the other hand, despised such city form as they saw it as overly dense and unhealthy, and proposed more dispersed urban plans [5].

3.1.2 Socio-economic indexes

Apart from quantitatively capturing urban form by means of the nine metrics above mentioned, our method also requires access to an index that captures the socio-economic levels of the area under study. Such indexes are ready available for many countries around the world. Although they differ in how they are computed, most of them are based on the concept that wealth or poverty are not caused only by economic factors (e.g., income, employment) but also by other aspects of life (e.g., education, health) and are thus composite. Socio-economics indexes are usually computed at a fine level of spatial granularity in developed countries (though, they tend to be coarser in developing ones). In England and Wales, there exists, for example, the Index of Multiple Deprivation (IMD), [39] which is computed by weighting seven different domains (i.e., income, employment, education, health, crime, barriers to housing and services, living environment), for small census areas of approximately 1500 residents. In developing countries, there exists the Multidimensional Poverty Index (MPI), [40] which is calculated as a weighted mean of three macro domains (i.e., health, education, and standard of living) at household level.

3.2 Analytical approach

Having illustrated the metrics that our method requires, we now present the analytical approach. This is based on spatial linear regression, as it allows to directly compare and interpret regression coefficients, and thus to measure to what extent our metrics of urban form can explain of the variance of the socio-economic index relative to one another. The analytical approach consists of a four-step process: (i) selection of the areal unit of analysis and computation of the metrics of urban form and socio-economics for such unit; (ii) normalisation and scaling of the metrics to meet the assumption of linear regression and obtain comparable regression coefficients; (iii) test for collinearity to avoid overinflated regression coefficients and unexpected signs; (iv) test for spatial autocorrelation to check for this issue and use of spatial regression in case the phenomenon is present. We follow with more details next.

3.2.1 Spatial unit of analysis

The spatial unit of analysis is the basic geographic entity for which the metrics presented above are computed. While there is no systematic method to select such unit, two considerations should be taken into account when choosing one. First, given that some of the metrics are based on the count of street intersections, the spatial unit should be big enough to contain some of these elements. For example, a unit of analysis comprising of a 500 meters by 500 meters block might not be suited for the approach proposed, as it might not have any intersection within its boundary. Second, official units that existed for a long period of time are generally better suited than, for example, more grid-shaped ones. Historical boundaries, in fact, tend to keep the morphological unity of neighbourhoods, for example, by not cutting buildings or blocks. This provides metrics that better reflect what exists in the real world. Once the spatial unit of analysis is selected, metrics of urban form and socio-economics should be computed and aggregated for such unit. Note that the selection of any unit of analysis is affected by the Modifiable Areal Unit Problem (MAUP), [41] that is values of spatial variables might change based on the unit of analysis selected for the study. This is a common issue that affects spatial analyses at areal level. We will come back to this in the Limitations.

3.2.2 Normalisation and scaling

Linear regression requires that candidate variables are normally distributed. Normalising the metrics is thus necessary to meet this assumption. This can be achieved in different ways, depending on how the values of the metrics are distributed. Common normalisation techniques are exponentiation and logarithmic transformation. Scaling is required because the metrics have different magnitudes: some are measures of density (e.g., Population Density, Dead-end Density), some others are percentages (e.g., Percentage of Unbuilt Land, Percentage of Green Areas). If these were regressed untransformed against the socio-economic index, their relative regression coefficients would be hard to compare and interpret. To avoid this, our method requires the computation of the standard scores (or z scores) associated with the normalised metrics. This can be achieved through the following formula:

$$z = \frac{X - \mu}{\sigma},$$

where X represents the metric raw value, μ is the mean value of such metric, and σ its standard deviation.

3.2.3 Test for collinearity and linear model

It is possible that two or more candidate variables (the metrics of urban form) show collinearity, that is they are strongly correlated. If strongly collinear variables are used in a regression, it is likely that their relative regression coefficients would be inflated or show unexpected signs. Since this approach is based on the interpretation of such coefficients, it is necessary to detect and discard strongly collinear variables. This can be achieved through the computation of the Variance Inflation Factors (VIFs) associated with each candidate variable. Let *reg* be a regression model with predictor variables $v_1, v_i, \ldots, v_n$. The VIF of the variable v_i is obtained by, first, performing linear regression with v_i as dependent variable and the other variables as independent ones $v_1, v_i - 1, v_i + 1, \ldots, v_n$, and, second, by using the overall model fit (i.e., R^2 value) obtained at the previous step in the following formula:

$$\mathrm{VIF} = \frac{1}{1 - R^2}.$$

If a variable has a strong linear relation with at least another one, its correlation coefficient is likely to be close to 1 and the VIF related to that variable large. A VIF equal to or greater than 10 is a sign of a collinearity issue [42]. If the candidate variables show VIFs smaller than 10, they can be regressed against the socio-economic index. Conversely, if the candidate variables have VIFs equal to or greater than 10, it is necessary to implement a stepwise procedure that, first, excludes the candidate variable with the highest VIF and, second, repeats the same process until none of the variables has a VIF equal to or greater than 10. At the end of this procedure, the candidate variables should be devoid of collinearity and can be regressed against the socio-economic index.

Once obtained the linear model, our method requires to check for spatial autocorrelation. This phenomenon occurs when observations located near one another are correlated or, as Tobler put it: 'everything is related to everything else, but near things are more related than distant things' [43]. Not considering this special dependency in linear models can cause over-inflated regression coefficients or unexpected signs. To check for this issue, our method relies on a renowned technique in spatial studies called Moran's test [44]. This checks whether the residuals of a regression analysis are spatially autocorrelated. The outputs of the Moran's test are an index I and a p-value. The former varies between -1 and 1, and can be interpreted similarly to a Pearson's correlation coefficient. The latter measures the statistical significance of the test. A negative Moran's I means that dissimilar values cluster together thus forming a dispersed pattern. Conversely, a positive Moran's I means that similar values are located near one another thus forming a clustered pattern. If the Moran's test is not significant (i.e., there is no statistical evidence that the residuals are spatially autocorrelated), the linear model can be trusted and interpreted. Conversely, if the Moran's test were to be significant (i.e., there is statistical evidence of the presence of spatial autocorrelation in the residuals), our method requires the use of a spatial model that incorporates the overlooked spatial information. We propose the use of the Spatial Autoregressive (SAR) model, a type of spatial model that accounts for the proximity of observations in space by including a spatial weighting matrix in the equation [45]. To ascertain that the SAR model accounts for all the spatial autocorrelation present in the data, our method requires to perform again the Moran's test. If the outputs are negative, the SAR model can be accepted and interpreted. Conversely, if they are positive, it should be rejected.

We present next the application of the proposed method to six UK cities.

4 Application of the method

We applied the method presented above to UK urban areas. To do so, we first identified what areas are considered urban. We extracted such information from an official document called *Rural Urban Classification* [12]. In this document, areas were classified in ten classes depending on their level of "urbanity", with the most rural category being *Hamlets and Isolated Dwellings in a Sparse Setting* and the most urban being *Major conurbation*. We chose the areas classified as *Major conurbation* for this analysis. The resulting urban areas corresponded to the cities of London, Manchester, Liverpool, Birmingham, Newcastle, and Leeds. These vary quite substantially both in socio-economic, historic, and cultural terms and in size. London covers the vastest surface and is the most populated, while Newcastle is the smallest and least populated of the set. We provide more information on the six urban areas under study in Table 1 and a map with their locations in Figure 1.

The first step of our method consists in the computation of the nine metrics of urban form and the socio-economic index at a *suitable* spatial unit of analysis. We selected the *ward* as spatial unit of analysis for the present study, for two reasons. First, areas of wards were never too small to cause issues in the computation of the metrics, yet it was small enough to offer city planners fine grained units of analysis and possible intervention. Second, wards are long standing administrative boundaries defined by the UK government, which have both electoral and ceremonial functions, and were first implemented in the

Table 1 Population and area of the six cities under study. Source: UK Census 2011 [46]

City	Population	Surface (ha)
London	8,173,941	229,546
Birmingham	2,736,460	94,661
Manchester	2,682,528	144,284
Leeds	2,226,058	94,315
Liverpool	1,381,189	52,267
Newcastle	1,104,825	57,127

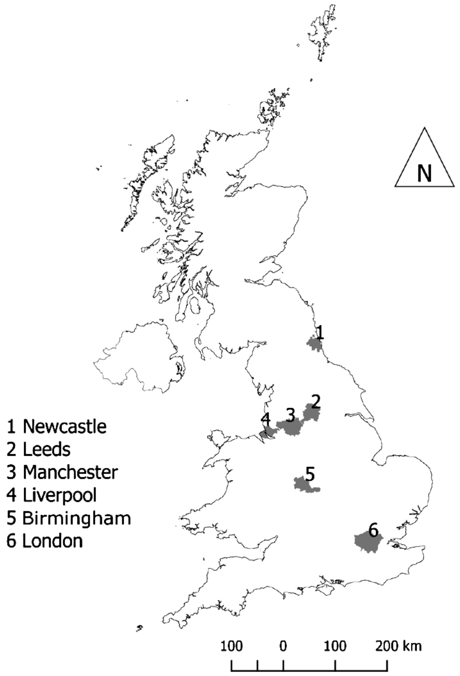

Figure 1 Location of the six urban areas under study

Middle Ages [47]. We identified 847 wards for London, 238 for Manchester, 183 for Birmingham, 119 for Newcastle, 100 for Liverpool, and 93 for Leeds.

Having chosen the spatial unit at which to perform our analysis, we then needed access to openly accessible datasets from which to compute the metrics of urban form and socioeconomics. We used five of such datasets: OS VectorMap District, Dwellings by Property Build Period and Type, the 2011 UK Census, OSM, and IMD. We present these datasets next.

Ordnance Survey (OS) VectorMap district This is one of the official digital maps of the UK [48]. It contains information on various geographic objects such as roads, building footprints, and natural areas. The content of this dataset is generated and kept updated by Ordnance Survey, the UK official mapping agency, and was made freely accessible for the first time in 2010. The geographic information is provided in vectorial format for tiles of 100 km by 100 km. We selected the tiles corresponding to the six urban areas under study and extracted the information needed for the computation of the metrics. We then computed the degrees of each node (i.e., street intersection) in the street networks, discarded the nodes of degree 2 (as they were not intersections but inaccuracies), and calculated the areas occupied by buildings, in the six cities under study. Note that roundabouts have been kept in their original shapes. This information was then used to compute six of the nine metrics of urban form at ward level: Connected Node Ratio (CNR), Intersection Density (ID), Dead-end Density (DD), Irregularity of the Street Network (ISN), Percentage of Unbuilt Land (PUL), and Betweenness Centrality (BC).

Dwellings by property build period and type This database contains the count of properties in England and Wales for several build periods and housing types [49]. To be more specific, the information on build periods is subdivided in twelve classes of around ten years each, with the first being the class with properties built before 1900 and the last being the one with properties built between 2010 and 2015. This information is provided for official census areas of about 1500 residents, the Lower-layer Super Output Areas (LSOAs). However, LSOAs are much smaller than wards (indeed too small for our analysis). The information on build periods was thus aggregated at the level of wards by summing the values associated with the LSOAs contained in each ward. This information was then used to compute the metric Offering Advantage of Historic Properties (OAHP).

2011 census: population and household estimates for wards and output areas in England and Wales This database contains information on the population density (i.e., persons per hectare) of each ward in England and Wales [46]. This data was used to compute the metric Population Density (PD).

OpenStreetMap With more than three million users, OSM is probably the best known example of geographic crowd-sourcing [50]. OSM contributors are collectively building and keeping updated the first free and editable map of the world. Many studies have been carried out to ascertain the quality of its content in different parts of the world, for example in the UK [51], France [52], and Germany [53], and reported an overall good level of spatial accuracy, especially in urban areas. For the purpose of this analysis, we used the OSM dataset as source of information for public green areas. Once these were identified, we assigned them to each of the wards of the urban regions under study and computed their areas. This information was then used to compute the metric Percentage of Green Areas (PGA).

Index of multiple deprivation This dataset includes information on the socio-economic deprivation of communities of England and Wales [39]. It is computed for LSOAs by weighting seven different domains: income deprivation, employment deprivation, health deprivation, education deprivation, barriers to housing and services, crime levels, and living environment deprivation. The higher the IMD score, the more deprived the area, and vice versa. IMD values are normally distributed [54]. Since the spatial unit adopted in this analysis was the ward, we aggregated IMD values at the level of such unit by averaging the IMD values associated with the LSOAs contained in each ward. This was possible since the variation of IMD scores of the LSOAs contained in each ward was small (i.e., their standard deviation values were small and always smaller than their average values). This data constituted the metric of socio-economic deprivation (IMD).

We present a summary table with metrics, relative brief descriptions, means, and standard deviations in Table 2. Choropleth maps of the metrics of urban form and IMD, for the six cities under study, are presented in the Additional file 1. The actual values of the metrics can be found in the Additional file 4.

5 Results

In this section, we present first some preliminary results drawn from the observation of the density distribution plots of the metrics of urban form and IMD. Second, we illustrate the outcomes of the regression analyses performed for each city. Third, we offer interpretations for the behaviours of the regression coefficients.

Table 2 Metrics of urban form and IMD with relative descriptions, means, and standard deviations

Variable name	Description	Mean	SD
Connected Node Ratio (CNR)	Level of connectivity of the street network	0.74	0.08
Intersection Density (ID)	Density of street intersections	0.62	0.33
Percentage of Unbuilt Land (PUL)	Proportion of land left unbuilt	73.23	12.30
Population Density (PD)	Density of city dwellers	62.76	43.15
Betweenness Centrality (BC)	Level of accessibility	0.03	0.04
Percentage of Green Areas (PGA)	Amount of green areas	7.93	8.19
Irregularity of the Street Network (ISN)	Level of irregularity of the street network	0.37	0.06
Dead-end Density (DD)	Density of dead-end roads (cul-de-sac)	0.20	0.11
Offering Advantage of Historic Properties (OAHP)	Weighted offering of historic properties	0.90	0.99
Index of Multiple Deprivation (IMD)	Socio-economic deprivation	24.16	13.07

5.1 Preliminary results

To explore the nature of the data considered and what it meant in terms of urban form, we produced density distribution plots for the metrics under study by applying a Gaussian smoothing function to the distributions. We present these plots, for each city under study, in Figure 2. As shown, none of the metrics were normally distributed, with only Connected Node Ratio (CNR) showing a distribution close to the normal. The majority of the metrics (six out of nine) showed positive skews, having most of their values close to their respective first quartiles. These were: Percentage of Green Areas (PGA), Intersection Density (ID), Dead-end Density (DD), Population Density (PD), Betweenness Centrality (BC), and Offering Advantage of Historic Properties (OAHP). Irregularity of the Street Network (ISN) and Percentage of Unbuilt Land (PUL) showed negative skews, instead, with most of their values concentrating around their respective third quartiles. IMD also presented a non-normal distribution (i.e., positive skew).

We drew some observations from these preliminary results. Most metrics of urban form had similar distributions across the six cities under study. This was not surprising since the metrics were computed for regions within the same country, which thus had been subject to similar historic, social, and cultural phenomena. However, there were also some unique variations. First, the urban form of London seemed to be denser in terms of built form and population, compared to the other cities (i.e., more low values of PUL, more high values of PD). Moreover, it seemed to be better connected (i.e., more high values of CNR, more low values of DD) and less deprived (i.e., more low values of IMD). These findings seemed to be in line with the research carried out by Bettencourt and West [27], who discovered a super-linear relationship between the population size of cities and a set of urban indicators such as income, GDP, and crime. Second, Birmingham's urban features showed peaks of values rather than more varied distributions. In particular, most of its neighbourhoods seemed to have low values of ID (around 0.5), moderately high values of PUL (around 70%), and low values of OAHP (close to 0). Third, Newcastle seemed to offer less green areas than the other cities. The majority of its neighbourhoods, in fact, showed values of PGA close to 0. Finally, Leeds seemed to be more sparsely built (i.e., more high values of PUL) and offer more historic properties (i.e., more high values of OAHP).

5.2 Linear models

After having normalised and standardised the variables of urban form, we checked whether they were collinear through the VIF test. Outcomes of this test indeed showed that some of the variables presented collinearity. In particular, CNR and ID were strongly collinear in all cities, with VIFs significantly greater than 10. PUL was found to be collinear

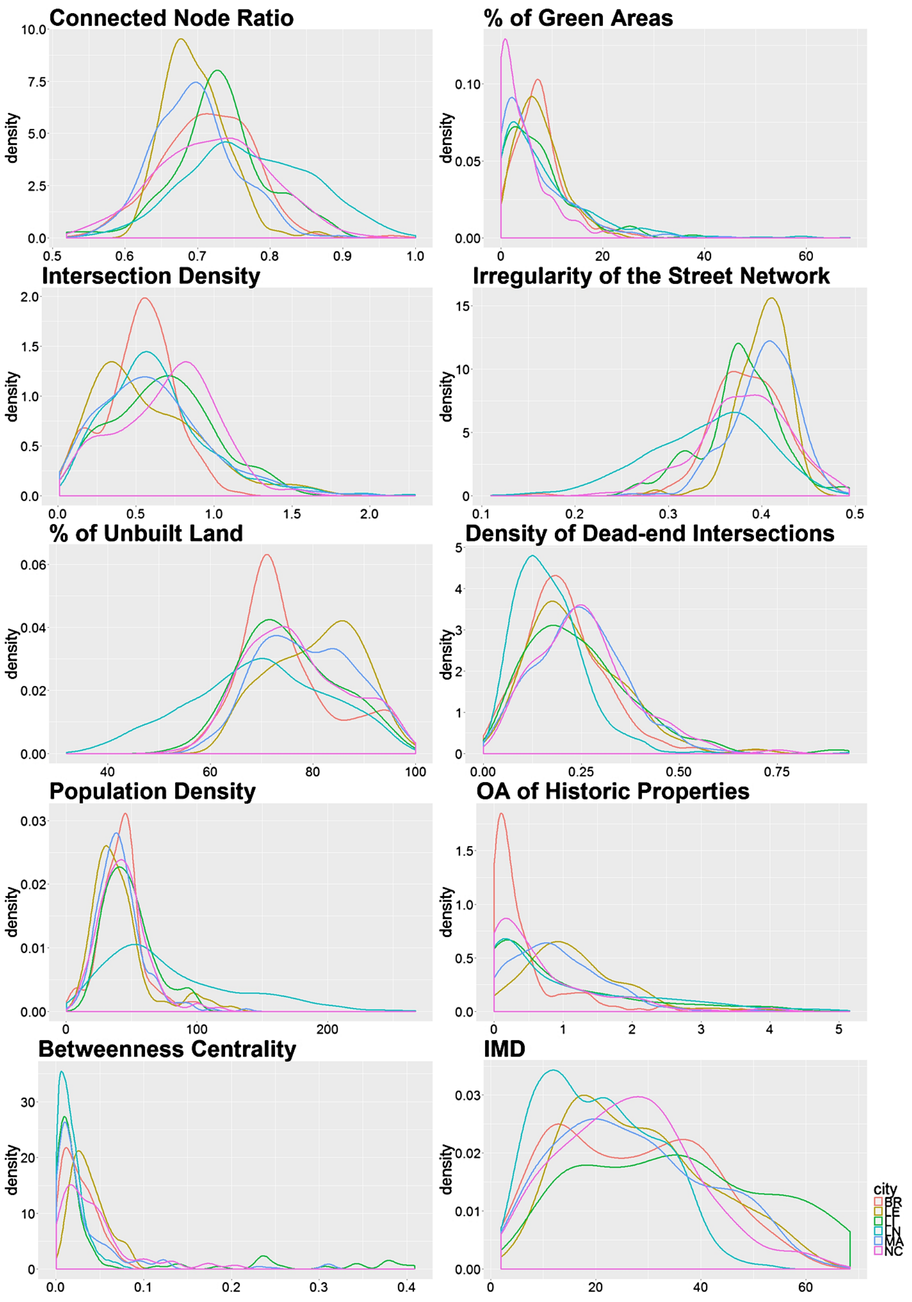

Figure 2 Density distribution plots for the metrics of urban form and IMD, for the six cities. BR: Birmingham. LN: London. MA: Manchester. LE: Leeds. LI: Liverpool. NC: Newcastle

only in Leeds. Such variables were thus discarded from the list of candidates for the regression analysis.

We then input the remaining variables in six regression models, one for each city, with IMD as dependent variable. Model outcomes suggested that multiple features of urban form were associated with deprivation. Models were all statistically significant, at 99% confidence level, and generally presented moderate fits, with four models out of six being able to explain around 50% of the variance of IMD. To be more specific, urban form could explain 50% of the variance of deprivation in Birmingham and Leeds; it could explain 49% of the variance in London and 48% in Manchester. The explanatory power of the models for Liverpool and Newcastle was lower, instead. Urban form could explain 25% of the variance of IMD in the first city and only 11% in the second.

To check whether spatial autocorrelation was not affecting these outcomes, we performed the Moran's test on the residuals. Outputs of such test showed statistical evidence of the presence of spatial autocorrelation in all models. Moran's I values were statistically significant, at 99% confidence level, with a minimum value of 0.16 (Manchester, Leeds, and Newcastle) to a maximum of 0.44 (London). Given the presence of spatial autocorrelation, which could have biased regression coefficients and overall model fit, we implemented the SAR technique to account for such phenomenon. SAR models were all statistically significant, at 99% confidence level, and showed greater explanatory powers and smaller regression coefficients, meaning that part of IMD was indeed explained by the spatial proximity of observations. To be more specific, the model for Birmingham could explain 67% of the variance of IMD, the one for London could explain 70%, the one for Manchester 56%, the one for Leeds 59%, the one for Liverpool 49%, and the one for Newcastle 27%. We performed a second Moran's test to ascertain whether spatial autocorrelation did not affect the residuals of the SAR models. Outputs of such test confirmed that there was no statistical evidence of the presence of the issue in none of the models (i.e., p-values > 0.05). Full results of the linear regression (LR) and SAR models can be found in Table 3. Frequency distribution plots of the residuals of the SAR models are presented in the Additional file 2. For informational purposes only, we also produced plots of single correlations between metrics of urban form and IMD, for the six cities under study. These are presented in the Additional file 3. Knowing that models were robust, we proceeded to investigate their relative regression coefficients. Common patterns are summarised below:

- Dead-end Density (DD) was statistically significant and positively associated with deprivation, in five cities out of six (i.e., Birmingham, London, Manchester, Liverpool, and Newcastle);
- Irregularity of the Street Network (ISN) was significant and negatively associated with deprivation, in four cities out of six (i.e., Birmingham, Manchester, Liverpool, and Newcastle);
- Percentage of Unbuilt Land (PUL) was significant and positively associated with deprivation, in four cities out of six (i.e., London, Manchester, Liverpool, and Newcastle);
- Population Density (PD) was significant and positively associated with deprivation, in four cities out of six (i.e., Birmingham, London, Manchester, and Leeds).

For what concerned the remaining coefficients, Betweenness Centrality (BC) was associated with deprivation in two cities out of six (i.e., London and Leeds), Offering Advantage of Historic Properties (OAHP) was related to advantaged areas in London only, while Percentage of Green Areas (PGA) was negatively associated with deprivation in Newcastle only. We elaborate more on these findings next.

5.3 Interpretations

As mentioned above, several regression coefficients (i.e., DD, ISN, PUL, and PD) showed similar patterns across the cities under study. This meant that we could identify an urban form that was associated with deprivation, in the majority of the cities considered. In four cities out of six, deprived English neighbourhoods appeared to be characterised by

Table 3 LR and SAR models for the six urban areas under study with outcomes of the Moran's test. *P*-value symbols correspond to: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	London				Manchester			
	LR		SAR		LR		SAR	
	p-val.	β	p-val.	β	p-val.	β	p-val.	β
(intercept)		0.00		-0.01		0.00		-0.02
CNR		-		-		-		-
ID		-		-		-		-
PUL	***	0.36	***	0.27	***	0.91	***	0.78
PD	***	1.06	***	0.60	***	0.35	***	0.27
BC	**	0.10	.	0.04		-0.05		-0.07
PGA		0.01		-0.01		0.06		0.04
ISN	*	0.11		0.03	***	-0.53	***	-0.46
DD		0.02	*	0.08	***	0.96	***	0.85
$OAHP$	***	-0.17	***	-0.14		-0.06		-0.05
adj. R^2	**0.49**		**0.70**		**0.48**		**0.56**	
p-value	0.00		0.00		0.00		0.00	
Moran's	0.04		0.03		0.16		0.02	
p-value	0.17		0.08		0.00		0.26	

	Birmingham				Leeds			
	LR		SAR		LR		SAR	
	p-val.	β	p-val.	β	p-val.	β	p-val.	β
(intercept)		0.00		-0.05		0.00		-0.01
CNR		-		-		-		-
ID		-		-		-		-
PUL		0.12		0.13		-		-
PD	***	0.45	***	0.35	**	0.44	**	0.37
BC	**	0.19		0.06	**	0.21	*	0.17
PGA	*	0.14		0.07		-0.10		-0.07
ISN	*	-0.23	*	-0.17		-0.10		-0.10
DD	*	0.24	*	0.18	.	0.23		0.19
$OAHP$	.	0.11		0.06		-0.04		-0.05
adj. R^2	**0.50**		**0.67**		**0.50**		**0.59**	
p-value	0.00		0.00		0.00		0.00	
Moran's	0.30		0.03		0.16		-0.02	
p-value	0.00		0.21		0.00		0.59	

	Liverpool				Newcastle			
	LR		SAR		LR		SAR	
	p-val.	β	p-val.	β	p-val.	β	p-val.	β
(intercept)		0.00		-0.02		0.00		0.01
CNR		-		-		-		-
ID		-		-		-		-
PUL	.	0.40	*	0.36	*	0.60	*	0.50
PD	.	0.33		0.18		0.10		0.07
BC		0.14		0.07	.	0.16		0.13
PGA		-0.06		-0.02	*	-0.20	.	-0.14
ISN	.	-0.28	.	-0.26	.	-0.38	.	-0.34
DD	**	0.66	**	0.58	**	0.62	**	0.53
$OAHP$		-0.11		-0.12		0.07		0.05
adj. R^2	**0.25**		**0.49**		**0.11**		**0.27**	
p-value	0.00		0.00		0.01		0.00	
Moran's	0.25		0.01		0.16		-0.02	
p-value	0.00		0.37		0.00		0.56	

high population density, vast portions of unbuilt land, numerous cul-de-sac, and regular street patterns. This seemed to closely resemble the modernist "towers in the park" design scheme, which saw the concentration of residents in small portions of land (i.e., residential towers laid out in open space), conspicuous presence of dead-end roads, and regular street patterns of major roads [5, 15].

The link between these urban features and deprivation thus seemed to discredit the modernist theories and support the ones of the compact city form [4, 13, 14], for the UK context. Jacobs, for example, was in favour of perimeter blocks rather than isolated residential towers as the retraction of buildings from side walks diminished social interactions, as fewer points of exchange (e.g., doors, windows, porticoes) between buildings and streets were present [4]. Furthermore, they reduced commercial activity, as there was no physical space on the sides of streets for amenities. Finally, isolated tower blocks also reduced safety, as streets were not informally controlled by windows facing them (the so-called "eyes on the street" effect). Similarly, she also supported well-connected streets rather than cul-de-sac, as the latter diminished connectivity and thus also the ability of pedestrians to move in the urban space. This aspect, in her view, was not only fundamental for the vitality of urban spaces but also for their economic prosperity and safety. She argued that fewer pedestrians corresponded to fewer social interactions, smaller use of amenities, and fewer "eyes on the street" to prevent crime. Finally, she supported street networks with irregularities (e.g., diagonal roads and squares) rather than overly regular grids. This aspect, in her view, would have provided "visual interruptions", which enhanced the perception of space and, ultimately, urban life.

The link between presence of cul-de-sac and deprivation seemed also to discredit the theory proposed by Newman (i.e., cul-de-sac were beneficial against crime as they reduced passage of people thus making urban spaces more controllable [36]). Although we did not test a pure measure of crime, a domain associated with such topic was included in the computation of IMD.

Three more associations were found to be significant, although not across all cities. The relationship between BC and deprivation seemed to be linked to the detrimental effects of too much accessibility, which a recent study found to be associated with more road traffic [55], on people's well-being. These negative effects, in fact, could be associated with more air and noise pollution, more congestion, and higher levels of stress. The inverse relationship between OAHP, our proxy for the traditional urban form, and deprivation seemed to be backed up by theories of the compact city form [4, 13, 14]. As we mentioned earlier, aspects of the traditional urban form (e.g., density, connectivity, perimeter blocks) were deemed fundamental for social, economic, and safety reasons. Finally, the negative relationship between PGA and deprivation in Newcastle could be linked to the beneficial effects of urban greenery on the well-being of residents. This was supported, for example, by the study of Maas *et al.* who found that higher percentages of urban green were associated with higher scores of perceived health [56].

6 Limitations

We ought to acknowledge some limitations for this work. First, urban form is not the only factor influencing the socio-economic levels of city areas. Many other aspects are at stake, for example, specific housing policies, economic interventions, gentrification. While it would be impossible to account for all of these different factors in one model, research outcomes from other fields (e.g., demography, econometrics) can be used to contextualise and interpret the results provided by the application of our method. Second, the proposed approach and relative outcomes do not imply causation. For instance, this means that, although one may find that connectivity is associated with better socio-economic outcomes, increasing the connectivity of a neighbourhood might not necessarily bring actual improvements of the socio-economic conditions of the resident population. Third, the results found for the British cities cannot be generalised as they only hold for the specific geographic regions (i.e., the six cities under study) and time frame (i.e., 2015) investigated in this work. Nonetheless, one can use the very same method to test larger areas and different time frames and thus extend generalisability. Fourth, the selection of the spatial unit of analysis inevitably comes with the issue of the Modifiable Areal Unit Problem

[41]. This states that values of metrics can vary quite substantially if computed for different spatial units. This can clearly bias the outcomes of a spatial analysis. While there is no systematic method to address this issue, one should be aware of this problem and eventually test whether metrics keep similar values when tested for different units of analysis. Fifth, this study is carried out by focusing on the relationship between urban form and only one socio-economic indicator (i.e., IMD). This because, at the time of this study, IMD was the only socio-economic index freely accessible at country level and thus available for multiple cities. Alternative indexes might be available. However, they might be available for London only (e.g., the London Ward Well-being Score[a]). We would like to stress that the very same method presented in this paper can be applied to these indexes to gain additional insights into the relationship between urban form and socio-economics. Sixth, although researchers found that OSM provides high quality spatial information, especially in urban contexts [51–53], it might still suffer from data completeness biases due to its crowd-sourced nature that tends to overlook poorer or "less attractive" neighbourhoods. Finally, the proposed method models the relation between aspects of urban form and socio-economics in a linear fashion. However, it is possible that such relationships are not linear. For example, interactions between metrics of urban form might be better suited to explain socio-economic levels than the metrics taken separately. This warrants a separate future investigation.

7 Conclusions

With the world undergoing a process of fast urbanisation, inequality is on the rise as some areas are benefiting more than others of public fundings and international investments. Analysing the relationship between urban form and socio-economics has thus become urgent as it can assist planners and policy makers when debating how to design cities and where to allocate resources. In this paper, we proposed a quantitative method to analyse such relationship at scale through spatial linear regression. More specifically, the method extracts metrics of urban form and socio-economics from openly accessible datasets. It then identifies, through regression analysis, what set of urban features are associated with socio-economic levels of city areas. When applied to the major UK cities, the method found that urban form could explain up to 70% of the variance of IMD, an official deprivation index. We also observed that some specific regression coefficients showed common patterns across the cities under study: high population density, vast portions of unbuilt land, presence of cul-de-sac, and regular street patterns were all related to deprivation. By connecting these findings to previous works, we argued that the relationship between this specific combination of urban features and deprivation discredited modernist theories and supported theories of the traditional city form, in the UK case.

Additional material

Additional file 1: Additional file 1. Here we provide choropleth maps of the metrics of urban form and IMD, for the six cities under study. (pdf)
Additional file 2: Additional file 2. This represents the frequency distribution plots of the residuals of the SAR models, for the six cities considered. (pdf)
Additional file 3: Additional file 3. Here we provide correlation plots for each metric of urban form and IMD, for the six cities under study. (pdf)
Additional file 4: Additional file 4. This compressed folder contains the tables with metrics of urban form and IMD, for the six cities under study. (zip)

Acknowledgements
Not applicable.

Funding
The research presented in this paper has been developed thanks to the Engineering and Physical Sciences Research Council (EPSRC) grant number EP/G037698/1.

Abbreviations
UK, United Kingdom; IMD, Index of Multiple Deprivation; OSM, OpenStreetMap; OS, Ordnance Survey; GIS, Geographic Information System; CNR, Connected Node Ratio; ID, Intersection Density; PUL, Percentage of Unbuilt Land; BC, Betweenness Centrality; PGA, Percentage of Green Areas; DD, Dead-end Density; OAHP, Offering Advantage of Historic Properties; MPI, Multidimensional Poverty Index; LSOA, Lower-layer Super Output Area; VIF, Variance Inflation Factor; SAR, Spatial Autoregressive model; LR, Linear Regression.

Availability of data and materials
The normalised and scaled version of the data used in this research is provided in the Additional file 4.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
The main idea of this paper and its development was carried out by AV. GQ performed all the steps of the proofs in this research. LC supervised the process. All authors read and approved the final manuscript.

Author details
[1]Department of Civil, Environmental & Geomatic Engineering, UCL, London, UK. [2]Department of Computer Science, UCL, London, UK.

Endnote
[a] https://data.london.gov.uk/dataset/london-ward-well-being-scores

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 1 June 2017 Accepted: 22 January 2018 Published online: 02 February 2018

References
1. Department of Economic and Social Affairs (2014) World urbanization prospects, the 2014 revision: highlights. Technical report, Population Division United Nations
2. Overseas Development Institute (2008) Briefing paper 44: Opportunity and exploitation in urban labour markets. Technical report
3. Moreno EL, Bazoglu N, Mboup G, Warah R (2008) State of the world's cities 2008/2009 – harmonious cities. Technical report, UN-HABITAT
4. Jacobs J (1961) The life and death of great American cities. Random House, New York
5. Corbusier L (1947) The city of tomorrow and its planning. Architectural Press, London
6. Vaughan L, Clark DLC, Sahbaz O, Haklay MM (2005) Space and exclusion: does urban morphology play a part in social deprivation? Area 37(4):402–412
7. Hillier B (2004) Can streets be made safe? Urban Des Int 9(1):31–45
8. Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Houghton, Boston
9. Kitchin R (2014) The data revolution: big data, open data, data infrastructures and their consequences. Sage, London
10. Salesses P, Schechtner K, Hidalgo CA (2013) The collaborative image of the city: mapping the inequality of urban perception. PLoS ONE 8(7):68400
11. Quercia D, O'Hare NK, Cramer H (2014) Aesthetic capital: what makes London look beautiful, quiet, and happy? In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing. ACM, New York, pp 945–955
12. Office for National Statistics (2013) Department for Environment, Food & Rural Affairs: The 2011 Rural Urban Classification For Small Area Geographies. Technical report
13. Whyte WH (1988) City: rediscovering its center. Doubleday, New York
14. Gehl J (2013) Cities for People. Island press, Washington DC
15. Hilberseimer L (1944) The new city: principles of planning. Theobald, Chicago
16. Hillier B (2007) Space is the machine: a configurational theory of architecture. Space Syntax, London
17. Hillier B, Sahbaz O (2009) An evidence based approach to crime and urban design, or, can we have vitality, sustainability and security all at once. In: Cooper R, Evans G, Boyko C (eds) Designing sustainable cities: decision-making tools and resources for design. Wiley, New York, pp 163–186

18. Harries K (2006) Property crimes and violence in United States: an analysis of the influence of population density. Int J Crim Justice Sci 1(2):24–34
19. Haughey RM (2005) Higher-density development: myth and fact. ULI–Urban Land Institute, London
20. Duany A, Plater-Zyberk E, Speck J (2001) Suburban nation: the rise of sprawl and the decline of the American dream. Macmillan & Co., London
21. Burchell RW, Shad NA, Listokin D, Phillips H, Downs A, Seskin S, Davis JS, Moore T, Helton D, Gall M (1998) The costs of sprawl-revisited vol. Project H-10 FY'95
22. Bridge G (2002) The neighbourhood and social networks. CNR paper 4, 1–32
23. Georg S (1995) The Metropolis and mental life. Metropolis: centre and symbol of our times. Macmillan & Co., London
24. Freeman L (2001) The effects of sprawl on neighborhood social ties: an explanatory analysis. J Am Plan Assoc 67(1):69–77
25. Muratori S (1959) Studi per Una Operante Storia Urbana di Venezia, vol 1. Instituto poligrafico dello Stato, Libreria dello Stato
26. Conzen MRG (1960) Alnwick , northumberland: a study in town-plan analysis. Trans Pap (Inst Br Geogr) 27:122
27. Bettencourt LM, Lobo J, Helbing D, Kühnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. Proc Natl Acad Sci 104(17):7301–7306
28. Bettencourt LM, Lobo J, Strumsky D, West GB (2010) Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. PLoS ONE 5(11):13541
29. Naik N, Kominers SD, Raskar R, Glaeser EL, Hidalgo CA (2017) Computer vision uncovers predictors of physical urban change. Proc Natl Acad Sci 114(29):7571–7576
30. Marshall W, Garrick N (2009) The shape of sustainable street networks for neighborhoods and cities. In: Congress for the new urbanism XVII, Denver
31. Banister D, Watson S, Wood C (1997) Sustainable cities: transport, energy, and urban form. Environ Plan B, Plan Des 24(1):125–143
32. Porta S, Crucitti P, Latora V (2006) The network analysis of urban streets: a primal approach. Environ Plan B, Plan Des 33(5):705–725
33. Wang F, Antipova A, Porta S (2011) Street centrality and land use intensity in Baton Rouge, Louisiana. J Transp Geogr 19(2):285–293
34. Timothée P, Nicolas L-B, Emanuele S, Sergio P, Stéphane J (2010) A network based kernel density estimator applied to Barcelona economic activities. In: International conference on computational science and its applications. Springer, Berlin, pp 32–45
35. Remali AM, Porta S, Romice O (2014) Correlating street quality, street life and street centrality in Tripoli. In: Libya. the past, present and future of high streets
36. Newman O (1972) Defensible space. Macmillan Co., New York
37. Curtis WJ (1996) Modern architecture since 1900, vol 2. Phaidon, London
38. Venerandi A, Quattrone G, Capra L, Quercia D, Saez-Trumper D (2015) Measuring urban deprivation from user generated content. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. ACM, New York, pp 254–264
39. Smith T, Noble M, Noble S, Wright G, McLennan D, Plunkett E (2015) The English indices of deprivation 2015. Department for Communities and Local Government, London.
40. Alkire S, Conconi A, Roche JM (2012) Multidimensional poverty index 2012: brief methodological note and results. University of Oxford, Department of International Development, Oxford Poverty and Human Development Initiative, Oxford
41. Openshaw S, Taylor PJ (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Stat Appl Spat Sci 21:127–144
42. Chatterjee S, Hadi AS (2015) Regression analysis by example. Wiley, Hoboken
43. Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234–240
44. Cliff AD, Ord JK (1981) Spatial processes: models and applications. Pion, London
45. Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge
46. 2011 Census: Population and Household Estimates for Wards and Output Areas in England and Wales. http://tinyurl.com/jwwcz95. Accessed 4 May 2017
47. Borer MIC (1978) The city of London: a history. D. McKay Co., New York
48. OS VectorMap District http://tinyurl.com/jxrhcqw. Accessed 4 May 2017
49. Dwellings by Property Build Period and Type, LSOA and MSOA. http://tinyurl.com/m3we5sn. Accessed 4 May 2017
50. OpenStreetMap. www.openstreetmap.org. Accessed 4 May 2017
51. Haklay M (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. Environ Plan B, Plan Des 37(4):682–703
52. Girres J-F, Touya G (2010) Quality assessment of the French OpenStreetMap dataset. Trans GIS 14(4):435–459
53. Ludwig I, Voss A, Krause-Traudes M (2011) A comparison of the street networks of navteq and OSM in Germany. In: Advancing geoinformation science for a changing world. Springer, Berlin, pp 65–84
54. Smith T, Noble M, Noble S, Wright G, McLennan D, Plunkett E (2015) The English indices of deprivation 2015: techincal report. Department for Communities and Local Government, London
55. Xiao Y (2012) Urban morphology and housing market. PhD thesis, University College London
56. Maas J, Verheij RA, Groenewegen PP, De Vries S, Spreeuwenberg P (2006) Green space, urbanity, and health: how strong is the relation? J Epidemiol Community Health 60(7):587–592