

Charles Booth's poverty maps and modern patterns of affluence: understanding their relationship through vectorisation and spatial regression

Candidate name	William David Low
Date	30 th August 2019
Module title and code	Smart Cities and Urban Analytics, CASA0010
Supervisor	Sarah Wise
Word count	10,167

This dissertation is submitted in part requirement for the MSc in the Centre for Advanced Spatial Analysis, Bartlett Faculty of the Build Environment, UCL.

Abstract

120 years after their publication, Charles Booth's poverty maps of London remain a valuable and under-utilised source of spatial data on historical patterns of poverty in the city. This data holds potential insight into present day arrangements of poverty and wealth. The purpose of this study is two-fold: to vectorise Booth's maps to allow area-based analysis, and to determine the relationship between patterns of poverty in 1898-9 and modern patterns of affluence.

Manual and automatic options for the vectorisation of the Booth maps are considered, before a manual vectorisation process is described, culminating in set of 11,539 polygons covering a contiguous section of inner London. The data is aggregated to 49 intersecting Medium layer Super Output Areas, summarised, and subjected to cluster and outlier analysis. An aggregation of council tax band data is chosen from candidate sources of data to describe modern affluence, and is similarly spatially aggregated, summarised and subjected to analysis.

The relationship between both data sets is investigated, and subjected to Ordinary Least Squares and Spatial Lag Model regressions. A strong relationship between the aggregate Booth and modern affluence data is identified, suggesting that the patterns of poverty and wealth depicted in Booth's maps are a strong predictor of modern patterns of affluence. The geographic areas associated with two notable data groups are subjected to a qualitative review to verify accuracy of data and analysis.

Declaration

I hereby declare that this dissertation is all my own original work and that all sources have been acknowledged. It is 10,167 words in length.

William Low

Table of contents

Section	Heading	Page
	List of figures	4
	List of tables	5
	List of acronyms and abbreviations	6
	Acknowledgements	7
1	Introduction	8
1.1	Historical GIS	8
1.2	Charles Booth and the Inquiry into Life and Labour in London	9
1.3	Booth's work today	11
1.4	Software and data	14
1.5	Ethics	14
2	Vectorisation of the maps	15
2.1	Reviewing the maps	15
2.2	Reproducing the Booth symbology	16
2.3	Manual vectorisation	18
2.4	Automatic vectorisation	19
2.5	The vectorisation process	21
3	Creating the Booth poverty rating score	25
3.1	Applying the MSOA	25
3.2	Booth data summary	26
4	Assessing modern comparison indicators	30
4.1	Alternative modern indicators	30
4.2	Modern affluence data summary	33
5	Data comparison and regression	35
5.1	Booth score vs. modern affluence score	35
5.2	Ordinary Least Squares regression	38
5.3	Spatial Lag Model Regression	40
5.4	Excluding Westminster MSOA	43
6	Ground truthing our findings	45
6.1	Eastern outlier cluster	45
6.2	High performing Westminster group	47
7	Conclusions	50

Lists of figures

Number	Title	Page
1	Extract from 1898-9 Booth poverty map, sheet 6.	10
2	Legend from 1898-9 Booth poverty map, sheet 6.	10
3	Photograph of the Booth poverty map room at the Museum of London.	12
4	Extract from 1898-9 Booth poverty map, sheet 6, with annotation of issues.	15
5	Extract from Booth map sheet 4, showing varied colouring approaches.	16
6	Recreation of point-based digitisation process.	17
7	Extract from 1898-9 Booth poverty map, sheet 6 with first third unedited, second blurred, third pixelated.	20
8	Example input and output from the map-vectorizer project.	20
9	MSOA digitised, shown in reference to constituent boroughs.	22
10	Vectorised Booth data polygons.	24
11	Process for calculation of Booth polygon area within MSOA.	26
12	Equation for aggregation of Booth scores.	26
13	Box plot of Booth scores.	27
14	Choropleth Booth scores by MSOA.	27
15	Moran's I scatterplot for Booth score.	29
16	Moran's I reference distribution for Booth score.	29
17	Local Moran's cluster and outlier map for Booth score.	29
18	Local Moran's significance map for Booth score.	29
19	Equation for aggregation of MA scores.	32
20	Box plot of MA scores.	33
21	Choropleth MA scores by MSOA.	33
22	Local Moran's cluster and outlier map for MA score.	34
23	Local Moran's significance map for MA score.	34
24	Scatter plot of Booth and Modern affluence scores.	35
25	Slope chart of Booth to modern affluence rank change.	37
26	Diagram of score relationship trend.	37
27	Standard deviation map of OLS regression residuals.	39
28	Local Moran's cluster and outlier map for OLS regression residuals.	40
29	Local Moran's significance map for OLS regression residuals.	40
30	Quantile map for predicted values of SLM regression model.	42
31	Standard deviation map of SLM regression residuals.	42
32	Local Moran's cluster and outlier map for SLM regression residuals.	42
33	Local Moran's significance map for SLM regression residuals.	42
34	Eastern MSOA cluster of interest, with ward boundaries.	45
35	Localities map focused on eastern MSOA cluster of interest.	45
36	Western MSOA group of interest, with ward boundaries.	48
37	Localities map focused on western MSOA group of interest.	48

Lists of tables

Number	Title	Page
1	Number of MSOA vectorised, by borough.	22
2	Five top and bottom ranked MSOA for Booth score.	27
3	Council tax bands in England (based on 1 April 1991 values).	32
4	Five top and bottom ranked MSOA for MA score.	33
5	Top and bottom 20% MSOA by Booth score, with rank comparison.	36
6	Results for Ordinary Least Squares regression.	38
7	Results for spatial lag regression.	41
8	Comparison of Ordinary Least Squares and Spatial Lag Model regressions without Westminster MSOAs.	43
9	Summary of eastern outlier cluster.	46
10	Summary of western group.	47

List of acronyms and abbreviations

GIS	Geographic Information Systems
IMD	Index of Multiple Deprivation
LISA	Local Indicators of Spatial Association
LSOA	Lower layer Super Output Area
MA	Modern Affluence
MSOA	Medium layer Super Output Area
OA	Output Area
OLS	Ordinary Least Squares
OSM	Open Street Map (stylised as OpenStreetMap)
QGIS	Quantum Geographic Information System
R2V	Raster to Vector
SLM	Spatial Lag Model

Acknowledgements

My thanks to Charles Booth: an excellent collaborator, despite his death over 100 years ago.

1. Introduction

This project has two broad objectives: the vectorization of Charles Booth's 1898-9 poverty map of London, and the comparison of the data it contains with modern, equivalent data. Under the first objective we will evaluate options for vectorization of scanned maps of this type, and subsequently produce an academic resource that does not currently exist in any publicly accessible format.

Under the second, we will establish to what degree patterns of wealth and poverty in the original Booth map are able to predict those of London today. A suitable measure of affluence will be determined and its relationship with the Booth data investigated. Notable clusters or outliers within that relationship will be identified, and subjected to a brief qualitative review in order to test the validity of our findings.

1.1 Historical GIS

This project can be considered as fitting into the field of historical Geographic Information Systems (GIS). GIS describes systems with a broad range of functions related to the storage, manipulation, visualisation and analysis of geographic and spatial data (Batty and Longley, 2003). Historical GIS is simply the application of GIS to historical data and geographies, and should not be seen as a strictly delineated field separate from any other uses of GIS: many of these implementations make use of data of varying historical status.

Nor are the challenges frequently encountered in historical GIS in any way unique. Changing administrative borders, incompleteness and ambiguity haunt all those working with spatial data of any era, but are particularly common when working with historical sources (Gregory and Healey, 2007; Knowles, 2005).

Where historical GIS does differ is in the speed and extent of its adoption into existing spheres of research: existing trends of spatial analysis and quantitative thinking in geography, geology and the earth sciences made for ready adoption and incorporation of GIS (Batty and Longley, 2003). In the sphere of historical research however, GIS was particularly vulnerable to criticism over its presumed objectivism and value-neutrality (Gregory and Healey, 2007).

The resulting slower uptake of GIS into the social sciences may be partly why there is less spatial research into Charles Booth's maps than might be expected (Vaughan, 2018). This existing literature will be reviewed following an introduction to Booth and his work on poverty in London.

1.2 Charles Booth and the Inquiry into Life and Labour in London

Charles Booth lived from 1840 to 1916, and remains a well-known figure in the social sciences. A prosperous businessman, he was an adherent to positivist notions of empiricism and believed that the social sciences should be subject to the exactitude of the natural sciences (Orford et al., 2002). First-hand experiences of poverty on the campaign trail for the 1865 national elections seem to have shaped his determination to address it, but his work also took place in a time with increased focus on and concern over the perceived breakdown of urban society (Vaughan, 2018).

Concerned by the perceived sensationalism with which poverty was reported, Booth asserted the need for an accurate accounting of its extent and depth, in order to shape suitable policy responses. In 1885 the Social Democratic Federation, the UK's first organized socialist party, published its own accounting of poverty in London and claimed 25% of the population lived in conditions of extreme poverty (Vaughan, 2018). Booth felt that this was overstated and began his own investigation the following year.

That investigation finally came to an end in 1903 with the publication of the final volume of the third edition of his research *Life and Labour of the People of London*, at a personal cost to Booth of the equivalent to three million dollars today (O'Day and Englander, 2005; Orford et al., 2002). Amongst other findings, it revealed that the true figure for those living in poverty was around 33%.








We will not seek to review at length the methodology employed by Booth. His methods of data collection changed over the 17 years of the enquiry, and were based primarily on interviews and consultation with experts with experience in each area of London. Information was subject to

review against census data, and corroborated with other experts, and with field work by Booth and his team (O'Day and Englander, 2005).

Figure 1: Extract from 1898-9 Booth poverty map, sheet 6 (London School of Economics & Political Science, 2016).



Figure 2: Legend from 1898-9 Booth poverty map, sheet 6 (London School of Economics & Political Science, 2016).

	Lowest class. Vicious, semi-criminal.		Very poor, casual. Chronic want.		Poor. 18s. to 21s. a week for a moderate family.
	Mixed. Some comfortable, others poor.		Fairly comfortable. Good ordinary earnings.		Middle class. Well-to-do.
	Upper-middle and Upper classes. Wealthy.				

Booth's pioneering maps showed the houses and streets of London colour-coded with seven categories. Figure 1 shows an extract from one of the maps, and figure 2 shows the Booth's visual scale. The range the maps symbology portrays means that despite their name Booth's maps do not display just poverty, but rather relative wealth at both ends of the scale.

Additionally, though the language employed in the category labels is summary and makes frequent mention of class, Booth's categorisation was multi-dimensional and based on a variety of factors including employment status, regularity of income and type of occupation (Vaughan, 2018).

Booth is not the original progenitor of the style of and approach to poverty mapping he is famous for; in this he was preceded at least by Abraham Hume, who used similar methodology to explore the relationship between churchgoing and poverty (Pickering, 1972; Vaughan, 2018). However, in terms of scale and influence, Booth's contribution is unmistakably greater.

Booth began his work more than 130 years ago, and so inevitably his methodology will suffer in comparison with modern standards. Criticism, such as that weighed by Spicker in 1990, Bales in 1994, and Topalov in 2007, frequently relates to subjectivity and imprecision in classification of poverty. However, as Vaughan (2018) notes, through his self-conceived combination of direct observation and statistics, and quantitative and qualitative methodologies, Booth approached his survey with genuine rigour. Indeed, modern research into the consistency of his classification of data have found it to be internally consistent (Bales, 1994).

It is this latter point that is of most importance to this investigation. Despite any shortcomings that can be identified in Booth's methodology, we can rely at least on his study's internal consistency. This forms the first major assumption on which this investigation will operate: that Booth's work forms an accurate record of relative levels of wealth, and of their geographic distribution.

1.3 Booth's work today

Of Booth's work, it is the poverty maps that are probably most familiar to the general public. They appear regularly in exhibition, such as 2006's *London: A life in Maps* at the British Library, 2018's *Living With Buildings* at the Wellcome Collection, and a permanent walk-in Booth poverty map room at the Museum of London, shown in figure 3 (British Library, n.d.; Wellcome Collection, n.d.). In 2012 the BBC also produced a documentary series investigating how six London streets had changed since their appearance on Booth's maps (BBC, n.d.).

Figure 3: Photograph of the Booth poverty map room at the Museum of London. Author's own photograph.



An archive of Booth's inquiry is held by the London School of Economics, and contains the maps, notebooks and various papers associated with the inquiry. The 1898-9 maps have been subjected to high quality scanning and the resulting files hosted on a website dedicated to the archive, licensed under a public domain mark with no rights reserved (London School of Economics & Political Science, 2016). The files are raster image files, edited only to stitch multiple maps together in order to produce a single seamless image, and allowing simple overlay with modern maps of London for visual comparison.

Despite the fame of Booth and his work, Vaughan notes, 'few of the books about Booth written in the past quarter century devote more than a passing mention to the maps' (2018, p.61), and suggests that this may be in part due to an extant division between social and spatial science. Similarly, notwithstanding the growth of historical GIS, there is a surprising lack of modern spatial analysis of Booth's maps. The few examples will be summarised, and our own project contextualised against them.

Vaughan (2007) has conducted space syntax research into the Booth maps. Space syntax analysis focuses on the spatial pattern of physical urban systems, and on using this analysis to gain insight into other aspects of those systems – in this case the spatial patterns that influence levels of poverty on Booth's maps. In 2009, with Geddes, Vaughan established a modern equivalent measure to Booth's and subjected it to the same form of space syntax analysis for the

Islington area, in order to compare this relationship between space and poverty to that found previously (Vaughan and Geddes, 2009).

In this latter study Vaughan considered the modern measure that our project makes use of, but ultimately selected an alternative. The reasoning for that and our own decision will be discussed later.

A 2002 study digitised the Booth maps in order to determine the relationship between poverty in the late 19th century and mortality ratios, and to compare it to the relationship between current poverty patterns and the same ratio (Orford et al., 2002). This study made some comparison of Booth and modern poverty at ward level, but did not attempt to directly quantify the relationship between the two.

Orford et al's study was a useful proof of concept for our project, though their approach to digitisation differs to that selected here. This will also be discussed later.

A 2008 pilot study sought to determine the continuity of poverty since Booth's study, focused on a selected area in the inner east end of London (Lindsay, 2008). Overcrowding was used as a modern measure of poverty, and systematic grid sampling was used for comparison. It is unclear how the Booth data was digitised in this study.

In his conclusions on the pilot, Lindsay acknowledged issues with the sampling process, and suggested the use of more advanced forms of analysis to improve accuracy. Our project avoids these issues by aggregating data into polygons representing modern administrative areas, though this introduces other issues that will be discussed.

Finally, The Economist magazine published an article in May 2006 comparing the Booth data for the Chelsea area with that of the 2001 census (The Economist, 2006). While its methodology was less clear, this illustrates that Booth's inquiry continues to be used as a backdrop to discussion of wealth and poverty in London. This, along with the various exhibitions and popular references to his work, shows a continued interest in Booth and his maps.

1.4 Software and data

Our study will make use of QGIS for the georeferencing of Booth map sheets, and the vectorisation of data. QGIS (previously known as Quantum GIS) is an open source desktop geographic information system, allowing manipulation, analysis and plotting of spatial data.

RStudio and Microsoft Excel were used to manage and prepare data for analysis, to perform the analysis itself, and to produce graphs used in this work. RStudio is an integrated development environment for the open source programming language R. Excel is well-known proprietary spreadsheet software.

GeoDa was also used for data analysis and graphs, particularly around regression analytics. GeoDa is open source software focused on spatial analysis and visualisation.

Extensive use was made of London administrative boundaries, sourced from the London Datastore under the Open Government Licence version 2.

1.5 Ethics

Only publicly available data was used in this project, and so no specific ethical considerations were observed.

2. Vectorisation of the maps

2.1 Reviewing the maps

The maps currently hosted by LSE are direct scans of the original maps. As such they are raster files, where the value of pixels, in this case the colour, describe the category of an area.

However, in their current form they are intended only for visual analysis and cannot be subjected to our intended GIS analysis.

Figure 4: Extract from 1898-9 Booth poverty map, sheet 6, with annotation of issues.



Figure 4 shows some of the issues preventing this. The maps are covered with annotation of streets, buildings, and neighbourhoods (1). Administrative areas are marked with dotted lines that sometimes pass through buildings (2). The colour of areas labelled with category symbology is not uniform, with overlapping text and building outlines from the basemap altering the colouring (3).

Categories three and five are labelled with diagonal hatching made up of two different colours – each with pixel level variety in colouring (4). The lowest income category uses black symbology (5), identical in colour to the dashed borders around administrative areas. Finally, since the maps are scanned copies of analogue products, the category colouring itself is not suitable for raster analysis: categories with single colour symbology are actually made up of a variety of slightly different colours at the pixel level (visible throughout).

2.2 Reproducing the Booth symbology

The intent of this first part of this project is to vectorise selected areas of the Booth maps in such a way that the existing category symbology is reproduced in a manner conducive to spatial analysis. This approach is not without issue. Colouring of the map is done with what appears to be varying attention to the actual layout of buildings beneath it. Figure 5 illustrates this problem. Sometimes the exact outline of a building is followed; in other instances a series of buildings with space between them are grouped in a single swathe of colour.

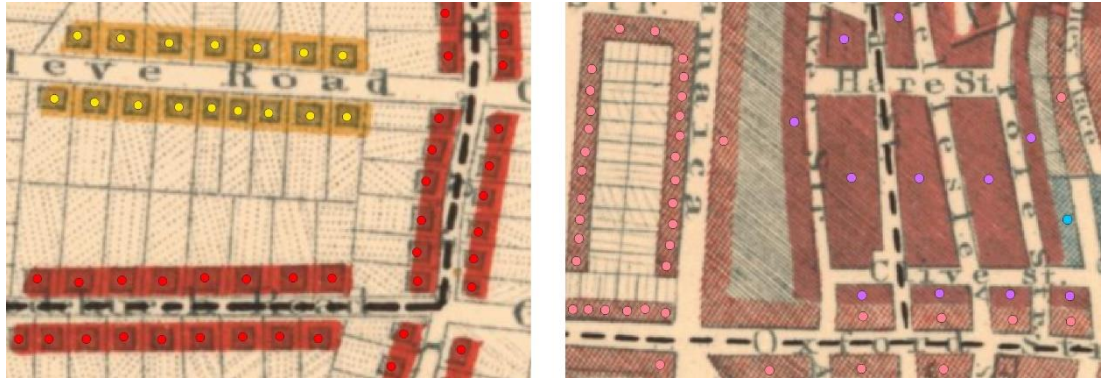
Figure 5: Extract from Booth map sheet 4, showing varied colouring approaches.



There are other approaches to digitisation of the Booth maps: in Orford et al's digitisation process (2002) each house identified was given a point vector with an attribute value that described its category. Buildings with more than one category were given multiple points. The number of points in a given area was then used to create aggregate statistics for those areas.

Figure 6 shows a recreation of this approach, but also highlights an issue with it. While certainly more accurate where houses are clearly delineated, this is regularly not the case on Booth's maps. Within larger buildings it is unclear how many subdivisions there are, and so how many households are present. Often entire streets of houses are indicated by a single rectangular polygon – the presence of gardens can sometimes allow differentiation, but these are not always present (and indeed are less so in poorer areas).

Figure 6: Recreation of point-based digitisation process.



There is no perfect solution to this issue when drawing from only Booth's maps, though it is possible that in combination with additional sources the final product might avoid these compromises. For example, the original source data in Booth's archived notebooks could be used to attempt verification of digitised data. Alternatively, if more detailed maps with clearly delineated internal housing boundaries and data on residential arrangements was identified, these could be combined with the Booth maps to create far more exact models.

For our purposes, we will take the approach of fundamentally recreating the colouring of the Booth maps. This forms the second major assumption of this project, that distortions to our statistical data due to basing our vectors on direct recreation of the Booth map colouring will be even between categories and areas, and not unduly impact on our ability to compare them.

We are seeking to summarise Booth's data for selected geographical areas, and so first will vectorise the data. In this context, vectorisation describes the process of converting raster data into vector data. Vector data differs from that of rasters in that data is stored as a series of points, which can be joined to form lines or polygons. These shapes can then be assigned attribute data.

In fact, both vector and raster data types (presuming the above issues were resolved) would be suitable for generating summary statistics for intersecting polygons. However, our preference is for vector data.

Vector data provides higher geographical accuracy, as it is not dependent on the resolution of an underlying pixel grid. It allows more complicated operations (such as buffering), and encodes

topology more efficiently than raster data, and so is more efficient at related operations (such as proximity analysis). It is easier to edit and manipulate shapes stored in the data, to allow corrections or refinements. Finally, it generates more aesthetically-pleasing lines and shapes. Since our intention is to generate a resource for a broad variety of potential future uses, vector data best suits our needs.

Since one of our aims is to generate a resource for use beyond this work, storing our Booth data in vector form ensures it will be more useful to others in future. However, the accurate vectorisation of scanned maps is a far from solved problem.

2.3 Manual vectorisation

Manual vectorisation describes the process by which a human user generates vector data, in reference to original data. Typically this involves use of digitising tools within GIS software, referring to the features of an overlaid basemap. This may be done with dedicated digitisation hardware, or with a standard desktop computer arrangement. This is the simplest and most common means of generating vector data, and such functionality is a core element of GIS software with graphical user interfaces such as ArcMap and QGIS (Heywood, 2006).

Besides the accessibility of required software and equipment, the advantage of manual digitisation is its use of human pattern recognition. All of the issues noted previously with the Booth maps are largely non-problematic when building vectors are manually identified and created by human users.

Manual digitisation introduces its own sources of error however, which Jenks categorised into two types: psychological and physiological (Jenks, 1981). In the former, the user is unable to discern the intended nature of the feature being digitised, such as the true centre point of a line. In the latter, an unsteady hand or similar physical constraint or malfunction can inadvertently introduce inaccuracies. Similarly, the quality of equipment being used can affect digitisation (Heywood, 2006).

Despite advancements in technology and efforts to reduce the requirement for manual input, it remains a common and reliable means of gathering vector data. For example, OpenStreetMap

(OSM) is a collaborative project building a map of the world entirely from public data, the majority of which is collected through manual digitisation by volunteers (Mooney and Minghini, 2017; OpenStreetMap Foundation, n.d.).

2.4 Automatic vectorisation

In an automatic vectorisation process, the data on the scanned map is digitally detected and extracted into a vector format. This functionality is less common in GIS, and so may require supplementary applications (Heywood, 2006). The same obstacles that prevent us from using the current Booth scanned maps for raster analysis also prevent easy conversion to vector data, such as text and boundary artifacts interfering with polygon shape detection, and colouring preventing accurate classification of symbology.

There are various examples of Raster to Vector (R2V) conversion packages, in two broad categories: line and polygon vectorisation. Given the nature of the Booth data, our interest is in the latter. GIS applications like ArcMap and QGIS have R2V functionality, but without capacity to deal with artifacts and variable category colours. These functions, like much of the body of work on raster to polygon vector conversion is focused around land classification of remote sensing imagery (Liao et al., 2012; Lou et al., 2005; Teng et al., 2008).

Preparation and editing of scanned material is a usual part of a digitisation process (Heywood, 2006), and one possible solution considered was the editing of the Booth poverty map imagery in order to better meet the limitations of available functionality. The open source raster graphics editor GNU Image Manipulation Program (GIMP) was used to explore this.

Figure 7: Extract from 1898-9 Booth poverty map, sheet 6 with first third unedited, second blurred, third pixelated.



Blurring or pixelating of the images (shown in figure 7) resolved the issue of diagonally hatched colouring and helped smooth some of the colouring issues caused by basemap elements underneath category symbology, though with a loss of fine detail. It could not however resolve the issue of larger artifacts and the colour overlap between the lowest income category and boundary and label markings.

One example of successful automatic vectorisation of scanned maps is the New York Public Library Lab's map-vectorizer project. The code created by the project is able to extract polygons and categorise them according to the closest reference value provided to their background colour (Arteaga, 2013). Figure 8 shows examples of input and outputs from the project.

Figure 8: Example input and output from the map-vectorizer project. Images downloaded from <https://github.com/nysl-spacetime/map-vectorizer>, August 2019.



This project shows the potential for fully automated vectorisation – but also its complexity. This was a multi-year project with several contributors lending expertise, and involving the use of Python, ImageMagick (image editing software), R, GIMP, and GDAL tools (a translator library for raster geospatial data) (Arteaga, 2013; NYPL Labs, 2019). In addition, the maps on which this project focused are considerably clearer and less problematic than the Booth maps.

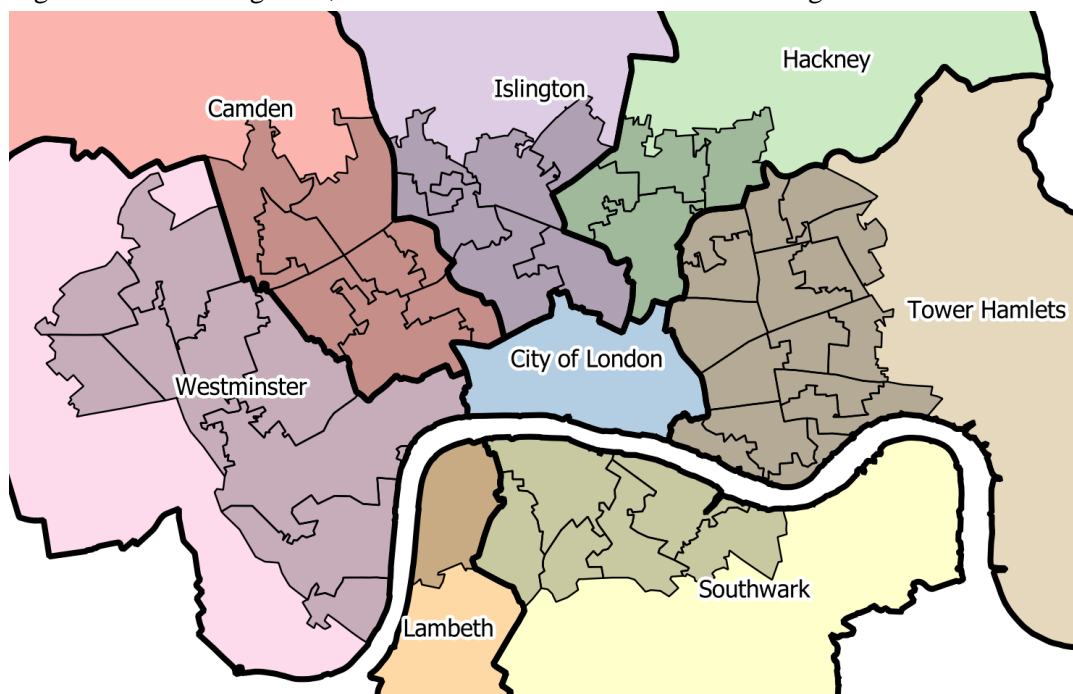
2.5 The vectorisation process

Given these findings, our vectorisation of the Booth data made use of a manual process. This was completed in QGIS, using the LSE Booth rasters as georeferenced basemaps.

Georeferencing, wherein the scanned maps were aligned with real world coordinates, was completed against OSM basemaps, and where possible surviving buildings and landmarks that appeared in the original maps were used to maximise accuracy.

Booth data was digitised within the area of 49 Medium layer Super Output Areas (MSOAs), loosely centred on central London, from seven different London boroughs. The MSOA is a geographical unit based on census data, and will be introduced more fully in the following chapter. Figure 9 shows the MSOAs in reference to the boroughs they are a part of. The City of London was excluded from Booth's inquiry due to a lack of significant numbers of residents (London School of Economics & Political Science, n.d.), so is also omitted from our own vectorisation.

Figure 9: MSOAs digitised, shown in reference to constituent boroughs.



Since the aim was to ensure a contiguous data set rather than to adhere to a balance of MSOAs from each borough, there is considerable variability in the representation of boroughs in the data (see table 1). This means that we must be careful to avoid making judgements based on the performance of larger boroughs in datasets, where that performance could simply be a reflection of greater representation in MSOAs.

Table 1: Number of MSOAs vectorised, by borough.

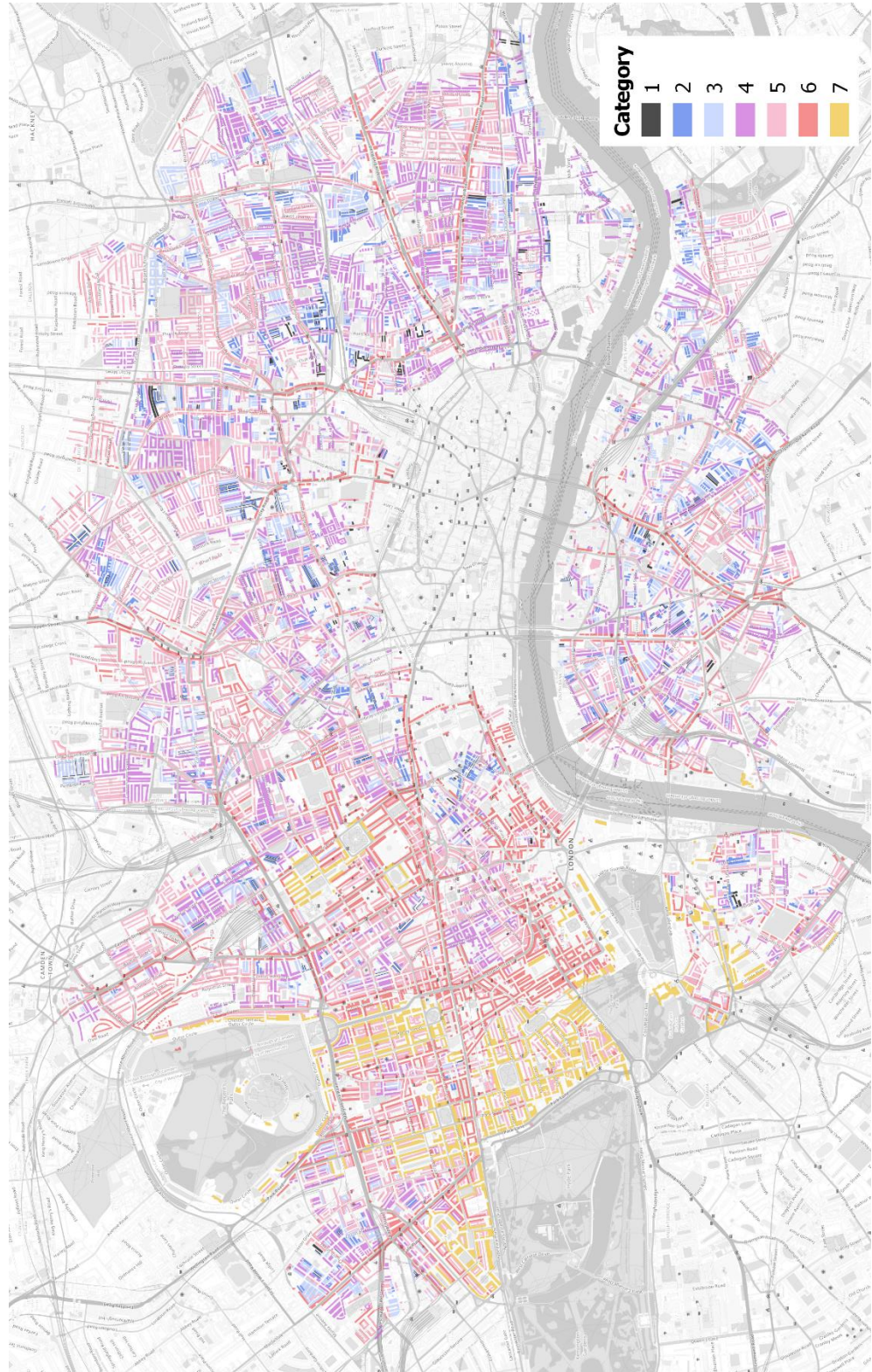
Borough	Number of MSOAs
Camden	8
Hackney	4
Islington	6
Lambeth	1
Southwark	6
Tower Hamlets	15
Westminster	9

The vectorisation process took approximately 45 hours to complete, and eight of the twelve Booth maps were used, for full or partial digitisation. The total dataset comprises 11,539 polygons, each with a category attribute based on its Booth colouring. Category labels are a 1 to 7 scale, where 1 represents the most poverty affected areas (coloured black on Booth's maps) and 7 the least (coloured yellow).

A number of checks were conducted to detect user error. Automated geometry checks were used to detect invalid polygon shapes. All polygon symbology was set to black and the map visually checked for any visible colour to identify any missed areas of colouring on the map. Finally, adjacent polygons with the same category attribute were detected and verified, since typically these would be vectorised as a single polygon.

Figure 10 shows a complete image of the vectorised polygons, with categorised symbology loosely following Booth's classification. An edited OSM basemap is provided for reference.

Figure 10: Vectorised Booth data polygons. Basemap © OpenStreetMap, used under the Open Database Licence.



3. Creating the Booth poverty rating score

3.1 Applying the MSOAs

With the selected portions of the Booth data fully vectorised, the polygons could be subjected to spatial analysis. The modern socio-economic measures we will consider are not publicly available to the level of detail of Booth's maps, and are instead aggregated to wider areas. As such we will aggregate our Booth data to the same areas to allow comparisons to be made.

Our administrative unit of choice is the Medium layer Super Output Area (MSOA). Output Areas (OA) were generated from the 2001 census, and were intended to form stable geographical units for reporting statistics over time (Office for National Statistics, 2012). Each output area consists of around 125 households, with a minimum of 40. This means output areas vary in size, and are smaller in densely populated areas.

Output areas were grouped to form super output areas: a Lower layer Super Output Area (LSOA) with a target of 600 households, and a medium super output area with a target of 3000 (Office for National Statistics, 2012). Both output areas and super output area groups do not attempt to follow the shape of existing neighbourhoods, but do seek to maximise social homogeneity through grouping households with similar characteristics, based on the census data for tenure and accommodation type (Vickers et al., n.d.).

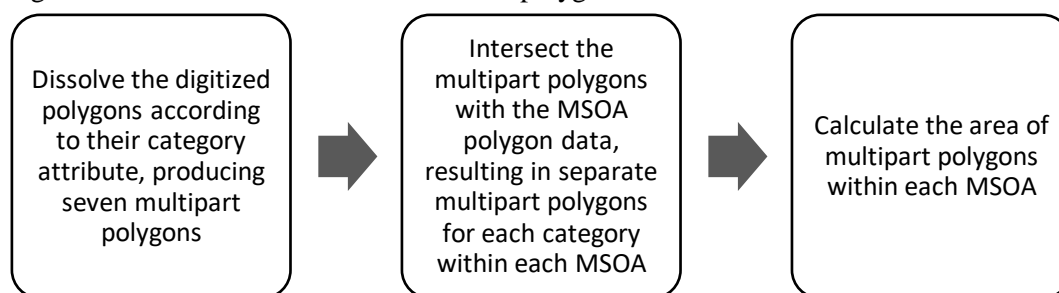
Limited changes were made to the output areas following the 2011 census data to preserve population sizes, reflect local authority boundary changes, and to maximise social homogeneity based on public consultation (Office for National Statistics, 2012). It is this version of the MSOAs that we will use.

The role of homogeneity in the design of OA means that they are well suited as units for aggregation for modern social data. Unfortunately this does not mean that they are necessarily well suited to our historical Booth data. Here we encounter the modifiable areal unit problem, wherein the size and shape of an area affects analysis of data within it (Manley, 2014; Openshaw, 1984).

Our options to deal with this issue are limited. The calculation of homogenous areas for aggregation of the Booth map would be an interesting process, but would leave us unable to compare it with our modern data. For the purposes of our aggregation the MSOA was favoured over the more granular LSOA, as using larger output areas ensures that sufficient Booth polygon data is available within each area to avoid small amounts of data creating erroneously high or low results on aggregate. As such we will continue under the assumption that this decision will suffice to allow accurate representation of the Booth data for our purposes.

Figure 11 describes the process by which the area of each category of polygon within each MSOA was calculated.

Figure 11: Process for calculation of Booth polygon area within MSOAs.



With the exact amount and proportion of every MSOA associated with each category determined, a single aggregate Booth score for each MSOA was required. This was built by assigning each category a 1 to 7 score (with 7 as the highest category). Scores were multiplied by the area value of that category in each MSOA, and then divided by the total area of all categories within the MSOA. Figure 12 shows the equation for this calculation, where α_1 is the area of category one polygons in the MSOA.

Figure 12: Equation for aggregation of Booth scores.

$$\text{Booth score} = \frac{\alpha_1 1 + \alpha_2 2 + \alpha_3 3 + \dots + \alpha_7 7}{\alpha_1 + \alpha_2 + \alpha_3 + \dots + \alpha_7}$$

3.2 Booth data summary

Table 2 shows the structure of the resulting data, and the five top and bottom ranked MSOAs. Our two largest MSOA groups are immediately visible, with Westminster dominating the top ranks and Tower Hamlets the bottom. Figure 13 shows a box plot of the data, showing a slight

skew to the data, and the presence of a number of outliers. The outlying nature of any data points are of interest, so none will be removed.

Table 2: Five top and bottom ranked MSOAs for Booth score.

MSOA code	MSOA name	Booth rating	Booth rank
E02000970	Westminster 011	5.93	1
E02000977	Westminster 018	5.83	2
E02000974	Westminster 015	5.70	3
E02000971	Westminster 012	5.48	4
E02000967	Westminster 008	5.35	5
39 other records			
E02000812	Southwark 006	3.88	45
E02000876	Tower Hamlets 013	3.81	46
E02000889	Tower Hamlets 026	3.80	47
E02000872	Tower Hamlets 009	3.56	48
E02000888	Tower Hamlets 025	3.49	49

Figure 14 shows the data visually represented as a choropleth map. A pattern already appears to have emerged, with most higher ranked MSOAs found in Westminster borough adjoining Hyde Park, and lower ranked MSOAs in Tower Hamlets, to the east.

Figure 13: Box plot of Booth scores

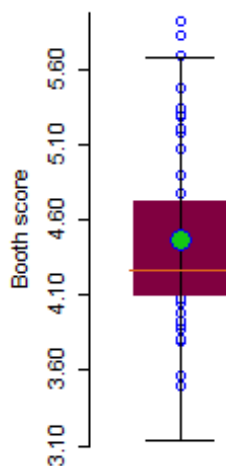
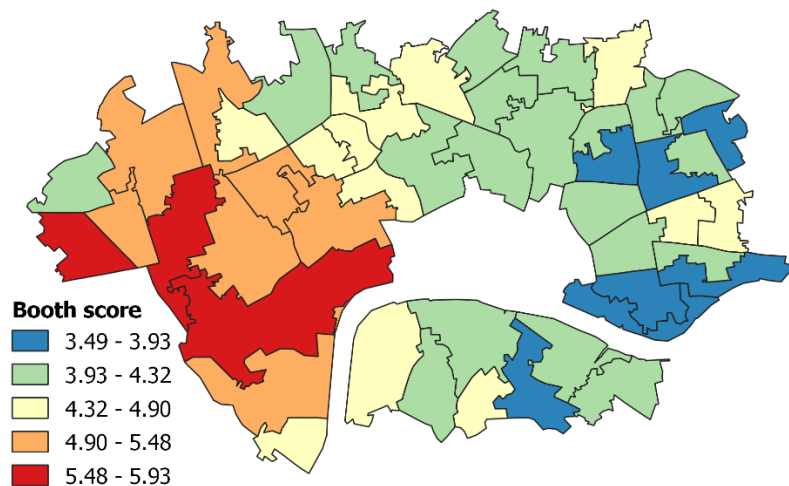


Figure 14: Choropleth Booth scores by MSOA



In order to accurately detect meaningful geographic clustering in the dataset, we will make use of Moran's I measure of spatial autocorrelation. Spatial autocorrelation describes the correlation of variables across space: the relationship between a variable shared between a number of geographical units and a measure of their geographical proximity (Cliff and Ord, 1973).

Moran's I is a popular indicator of spatial autocorrelation, providing a measure between -1 and 1, where 0 describes complete randomness, -1 describes perfect dispersion and 1 describes

perfect clustering of similar values (Moran, 1948). This measure additionally requires hypothesis testing to ensure statistical significance.

The Moran's I statistic is a single global measure for the dataset as a whole, allowing us to assess the overall degree of spatial autocorrelation, but not to locate it geographically when it is found to be present. In order to identify local clusters and outliers, we will also use the Local Moran, one of a group of Local Indicators of Spatial Association (LISA). The Local Moran replicates the function of the global measure for each geographical unit, producing local Moran's I statistics allowing detection of clusters and outliers and generating local measures of significance (Anselin, 1995).

In both cases, the geographical relationship between our area polygons is passed in the form of a spatial weights matrix, asserting in this instance which polygons are neighbours, with contiguous borders. Since our MSOA polygons are divided by the Thames river, the spatial weights matrix had to be adjusted to ensure that those polygons south of the river are not made inaccessible to clusters north of it. MSOA connected by bridges were considered neighbours. River crossings by boat were not given the same consideration, given the multitude of connections that would afford and the relative ease of bridge crossing by comparison.

As a result of these changes, neighbour status was established between Tower Hamlets 027 and Southwark 003, and between Lambeth 036 and Westminster 18 and 20 (the latter of which are already neighbours through contiguity). It should be noted that this approach gives equal weighting to the spatial relationship afforded by a connecting bridge and a direct contiguous border with several interconnecting streets. A future refinement to this model could be to model the relative strength of these relationships and reflect this in the spatial weight allocations.

Figures 15 through 18 show the output of this clustering detection. Figure 15 shows the Global Moran's I scatter plot and Global Moran's I value of 0.573, suggesting considerable autocorrelation. Figure 16 is a reference distribution based on 999 random permutations of the data, reporting a pseudo p-value of 0.001; the minimum possible, confirming statistical significance.

Figure 15: Moran's I scatterplot for Booth score.

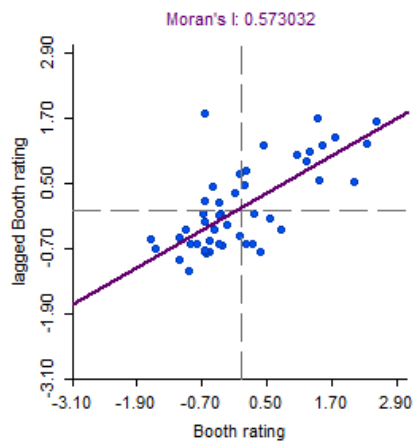


Figure 16: Moran's I reference distribution for Booth score.

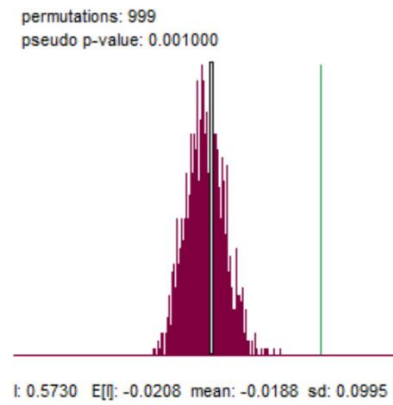
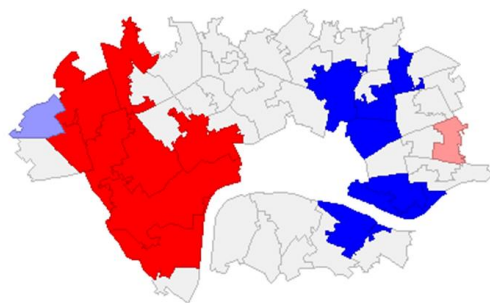


Figure 17 shows the results of the Local Moran cluster and outlier mapping. This confirms the existence of clusters initially identified in the choropleth map, and refines our understanding of their distribution. At this point we can reliably assert clustering of high Booth scores in the west of the selected area, and low scores in the east. Drawing from figure 18, which shows relative significance levels within those MSOA categorised as clusters or outliers, we can see that our western cluster has greatest significance, with much more varied levels in the eastern cluster.

Figure 17: Local Moran cluster and outlier map for Booth score.



Cluster and outlier groupings

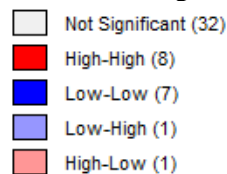
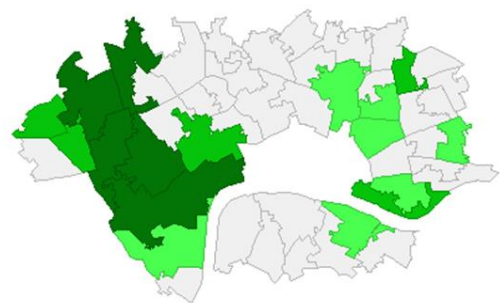
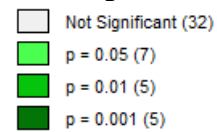


Figure 18: Local Moran significance map for Booth score



LISA significance



4. Assessing modern comparison indicators

4.1 Alternative modern indicators

In order to determine to what extent the patterns of wealth and poverty shown in the Booth data persist, modern data allowing comparison is required. Since it is not possible to reproduce data directly equivalent to that recorded in the Booth poverty maps (and so our Booth poverty score), various available alternative indicators were considered around poverty, income and house values.

Firstly, poverty-focused indicators were evaluated. The Index of Multiple Deprivation (IMD) is the official measure of relative deprivation in England, and is published at MSOA level (Dept. for Communities and Local Government, 2015). However, as previously discussed, Booth's poverty maps do not solely focus on communicating the presence of poverty in the capital, but the full spectrum of affluence.

The measurement of affluence in an area cannot be achieved with the IMD. Indeed, the official guidance of the IMD specifically notes that it should not be used to describe the affluence of a given area, noting that 'an area with a relatively small proportion of people (or indeed no people) on low incomes may also have relatively few or no people on high incomes. Such an area may be ranked among the least deprived in the country, but it is not necessarily among the most affluent.' The degree to which poverty is absent from an area does not necessarily correlate with its level of affluence.

Given the nature of Booth's categorization, with an inherent focus on patterns of employment and income, income would initially seem the most appropriate comparative measure. The MSOA atlas published by the Greater London Authority (GLA, n.d.) provides data on mean and median annual household income. Of the two, median income can be considered the stronger measure due to its greater resistance to outliers.

Income as a measure of affluence is by no means perfect: surveys of household income have been found to frequently fail to accurately capture the wealth of the richest households (Atkinson and Piketty, 2007; Goebel, 2007), leading to adjusted approaches like incorporation

of real estate data (Weide et al, 2017). That latter data is the final category considered, and was ultimately selected.

The MSOA atlas also contains the median price and number of annual sales of houses in each area for a six-year period. However, there was considerable variety in the number of sales per MSOA – meaning that some of the aggregate values listed were based on very few data points, making them prone to distortion. In some cases zero houses were sold in an annual period, providing no data at all.

A more robust data set is the annual stock of properties published by the Valuation Office Agency (VOA). This data contains the of number of properties by council tax band at various geographic levels, including the MSOA level (VOA, n.d.). The VOA is responsible for placing every home in England into one of eight council tax bands based on its value. The relative value of a home is then used to determine that household's tax contribution. This means that the council tax dataset is consistently vetted for consistency, draws on a highly detailed dataset compared to the numbers of sales in the MSOA atlas, and also measures the full spectrum of affluence.

In Vaughan and Geddes's *Urban form and deprivation: A contemporary proxy for Charles Booth's analysis of poverty* (2009), council tax data was passed over for the titular role, in favour of data on benefit claimants. Like Orford et al (2002), this meant a focus on the lower end of the spectrum of wealth. The authors noted that, unlike Booth's data, using council tax band information meant focusing on a household's property, rather than the people within that household. This is a fair criticism, as the condition, value or nature otherwise of housing did not feature in the methodology Booth used to categorise his findings (Lindsay, 2008).

However, studying the value of housing through the council tax data has an advantage in that it allows the accumulated impact of the 120 years since the Booth maps were published to more readily manifest in our data. Where levels of income or poverty are doubtless themselves a result of many contributing factors, many historical in nature, they are less temporally bound than the value of London's buildings: some of which marked on Booth's maps still exist today and are imparted varying value as a result. It is the social and structural changes associated with

the passage of time that we are interested in capturing in our data, and so the value of houses seems to more directly demonstrate this.

This choice forms our third major assumption, and our most controversial: that this data is internally consistent, and sufficiently accurate to make comparison with our Booth score data meaningful. A refinement of this approach might be an index of multiple indicators, foremost income and property values. Ideally this would require access to household level data for both.

Table 3 shows the eight council tax bands. Houses are assigned to a band on the basis of their value on April 1st, 1991, following a valuation in accordance with regulations and methodology established in the 1992 Local Government Finance Act.

Table 3: Council tax bands in England
(based on 1 April 1991 values) (gov.uk)

Band	Value at 1 April 1991
A	up to £40,000
B	£40,001 to £52,000
C	£52,001 to £68,000
D	£68,001 to £88,000
E	£88,001 to £120,000
F	£120,001 to £160,000
G	£160,001 to £320,000
H	more than £320,000

A single summary value was required for each MSOA, and so similar to our handling of the seven Booth categories, each band was assigned a 1 to 8 score. Scores were multiplied by the number of houses in that band in each MSOA, and then divided by the total number of houses within the MSOA. Figure 19 shows the equation for this calculation, where b_1 is the number of houses in the first council tax band. This score will be referred to as the Modern Affluence (MA) score.

Figure 19: Equation for aggregation of MA scores.

$$\text{Modern affluence score} = \frac{b_1 1 + b_2 2 + b_3 3 + \dots + b_8 8}{b_1 + b_2 + b_3 + \dots + b_8}$$

4.2 Modern affluence data summary

As with our Booth score data, we will perform a brief summary analysis of the modern affluence score data. Table 4 shows the structure of the data and the five top and bottom ranked data.

Table 4: Five top and bottom ranked MSOA for MA score.

MSOA code	MSOA name	MA score	MA score rank
E02000970	Westminster 011	6.491045	1
E02000977	Westminster 018	6.102601	2
E02000974	Westminster 015	6	3
E02000967	Westminster 008	5.887689	4
E02000971	Westminster 012	5.601103	5
39 other records			
E02000874	Tower Hamlets 011	2.92	45
E02000885	Tower Hamlets 022	2.914498141	46
E02000880	Tower Hamlets 017	2.823529412	47
E02000368	Hackney 024	2.802850356	48
E02000371	Hackney 027	2.764139591	49

Figure 20, a box plot of the MA data, shows a similar skew to the Booth score data, with no outliers. Figure 21 shows the data visually represented as a choropleth map. The western MSOAs seem again to be high scoring, with lower values in the east – though the distribution of the latter is further north.

Figure 20: Box plot of MA scores.

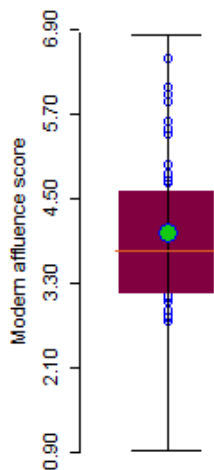
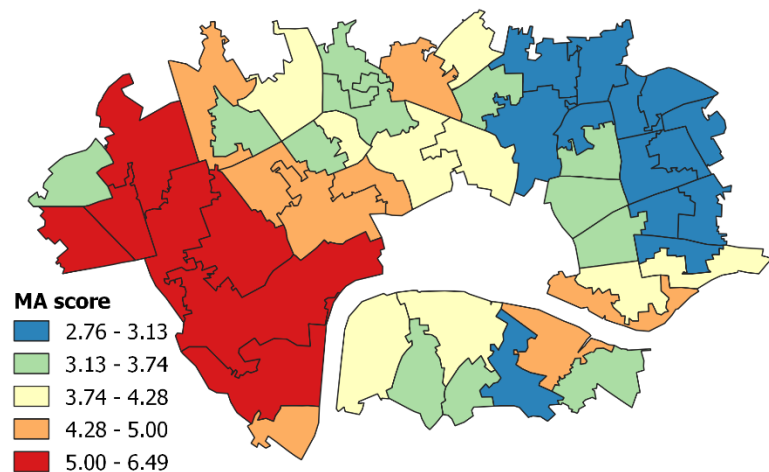


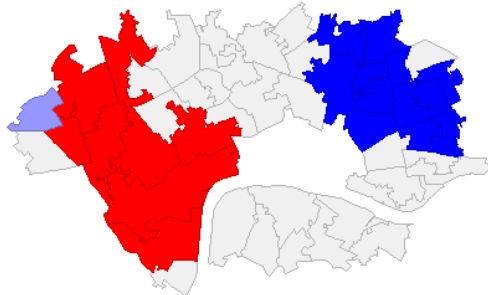
Figure 21: Choropleth MA scores by MSOA.



Figures 22 and 23 show the results of Local Moran clustering detection, following a Global Moran's I value of 0.622, with a pseudo p-value of 0.001. No changes were required to the

spatial weight matrix, as no bridges built since the Booth dataset was collected have linked any previously unconnected MSOA in our selection.

Figure 22: Local Moran cluster and outlier map for MA score.



Cluster and outlier groupings

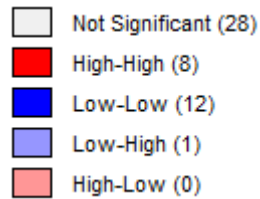
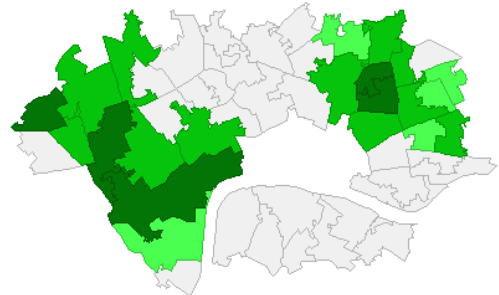
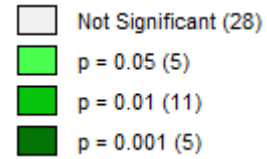


Figure 23: Local Moran significance map for MA score.



LISA significance



The cluster map confirms that the cluster pattern of modern affluence differs from that of the Booth score data in the east, but follows a near identical pattern in the west. Interestingly, the outlier at the north-west side of the map is also present in the modern data. This might hint at the existence of a larger cluster of low values west of the area chosen for study.

5. Data comparison and regression

5.1 Booth score vs. modern affluence score

We will now investigate the relationship between the Booth and modern affluence scores. Our intent is not just to understand and quantify that relationship, but also to identify interesting cases and outliers that we can review, in order to ground-truth our data and verify our findings.

Figure 24: Scatter plot of Booth and Modern affluence scores

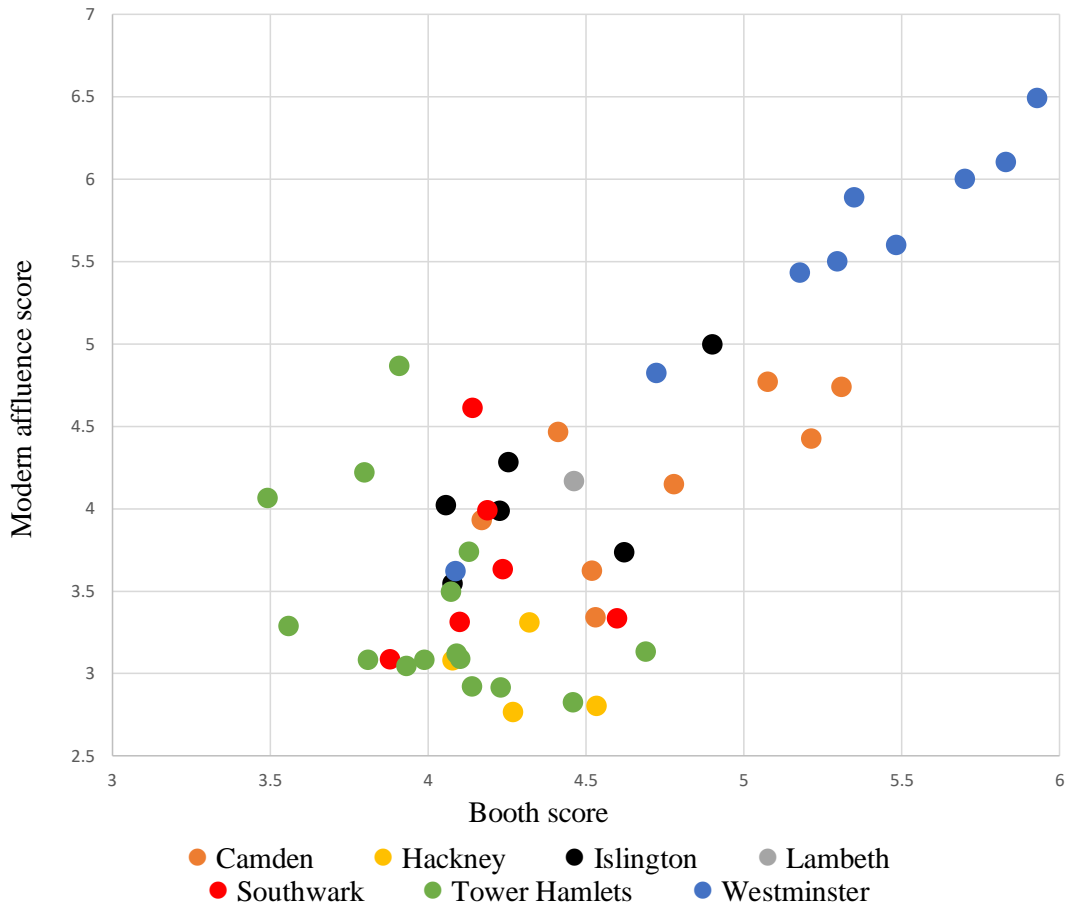


Figure 24 is a plot of the each MSOA's Booth score and its modern council tax score. Two clusters are immediately evident: between 3.5 and 4.75 on the x axis, the data is noisy and does not display a clear relationship. After the Booth score reaches 4.75, the relationship appears much more linear.

At this stage some initial observations can be stated. Firstly, high performing areas in 1900 are clearly more likely to be high performing in our modern affluence measure. Table 5 shows the top and bottom 20% of MSOAs in terms of their Booth score, and their associated rank in the data. Alongside this is their new rank in the modern affluence, and the percentile that rank

represents. MSOA names have been abbreviated: Westminster to WM, Camden to CM, Tower Hamlets to TH, Islington to IS, and Southward to SW.

Table 5: Top and bottom 20% MSOA by Booth score, with rank comparison.

Top 20% MSOA by Booth score				Bottom 20% by Booth score			
MSOA name	Booth rank	MA rank	New percentile	MSOA name	Booth rank	MA rank	New percentile
WM 011	1	1	2%	TH 015	40	31	63%
WM 018	2	2	4%	IS 023	41	21	43%
WM 015	3	3	6%	TH 006	42	41	84%
WM 012	4	5	10%	TH 007	43	44	90%
WM 008	5	4	8%	TH 027	44	9	18%
CM 026	6	12	24%	SW 006	45	40	82%
WM 013	7	6	12%	TH 013	46	42	86%
CM 021	8	15	31%	TH 026	47	17	35%
WM 020	9	7	14%	TH 009	48	36	73%
CM 028	10	11	22%	TH 025	49	20	41%

No MSOA that placed in the top 20% in 1900 has fallen out of the upper third of modern rankings. Indeed, the five highest ranking areas remain the five highest scoring today, with only a single minor change in sequence between fourth and fifth place. On the other hand, the bottom 20% shows far more variability in modern rankings. Only half of these are in the bottom third of our modern affluence measure, and one outlier has actually moved into the top 20% (marked in bold typeface).

Figure 25: Slope chart of Booth to modern affluence rank change.

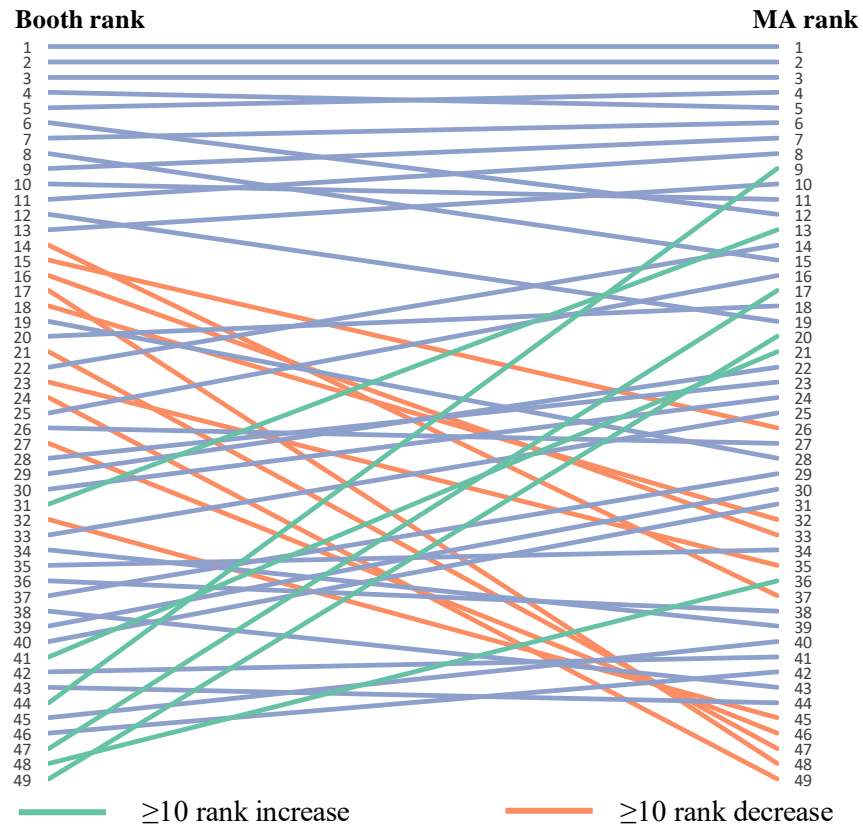
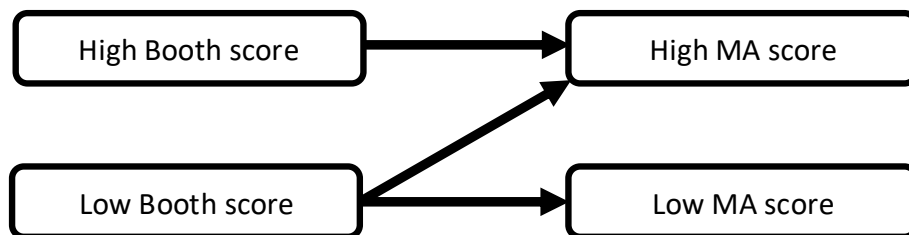


Figure 25 shows a slope chart comparing all ranks of both sets of data. This visualization makes readily apparent the different experience of higher ranking areas and moderate and lower ranking areas between our data sets.

Based purely on observation of the data, we can therefore state the general rule depicted in figure 26: in the intervening 120 years between Booth’s work and our modern data, relatively high performing areas have generally remained so. Lower performing areas exhibit far more varied outcomes, with some remaining relatively less relatively affluent, and others becoming more so.

Figure 26: Diagram of score relationship trend.



While we can make this assertion based on the distribution of the data, it is not clear to what extent the Booth score of an area is actually able to predict modern day affluence. In order to investigate this, we will make use of regression analysis, first using the Ordinary Least Square (OLS) model, and then an appropriate spatial regression model.

5.2 Ordinary Least Squares regression

Initially a linear regression was completed in order to quantitatively define this relationship, with modern affluence as the dependent variable. This first regression was a simple regression based on the Ordinary Least Squares model, under which the sum of the squares of the difference between the dependent variable and those predicted are minimised. Both modern affluence and the Booth scores were normalised to fix their values between 0 and 1, allowing for easier comparison. Table 6 shows the results.

Table 6: Results for Ordinary Least Squares regression

R-squared	0.575
Probability (F-statistic)	<0.001
Log likelihood	17.285
Akaike info criterion	-30.57
Schwarz criterion	-26.787

Variable	Coefficient	t-Statistic	Probability
Booth score (normalised)	0.833	7.967	<0.001
Jarque-Bera probability	0.89		
Test for spatial dependence	Value	Probability	
Moran's I (error)	0.327	<0.001	
Lagrange Multiplier (lag)	13.803	<0.001	
Robust LM (lag)	3.956	0.047	
Lagrange Multiplier (error)	10.337	0.001	
Robust LM (error)	0.489	0.484	

The initial regression seems to confirm the existence of a convincing relationship, with an R-squared value of 0.575 and an unambiguous p-value. However, these findings can only be considered reliable if the assumptions of the linear regression model are met. Most of these are met by our data; the Jarque-Bera test suggests the regression residuals are not problematically distributed. However, there is evidence of spatial auto-correlation within our data – unsurprising

given its presence in both the dependant and independent variable, and the visual similarity in its distribution.

Spatial auto-correlation breaks the assumption of OLS regression that error terms are independent. When the value assigned to a geographical unit depends on the values assigned to its neighbours, the data and the error terms produced by the regression contain a spatial dependence.

Our OLS regression includes some tests for spatial dependency, including a highly significant Moran's I score of 0.327. We can further investigate this issue using the residuals from the regression. Figure 27 shows a map of standard deviation of our regression residuals, showing where the model has under and overpredicted values. We can additionally pass these residuals to a Local Moran test, shown in figures 28 and 29. This shows clear and problematic evidence of significant clustering in the regression model.

Figure 27: Standard deviation map of OLS regression residuals.

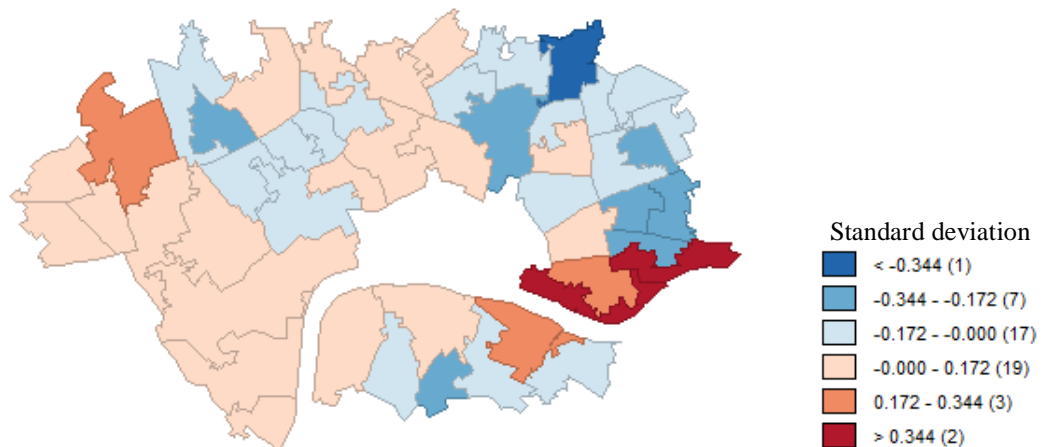
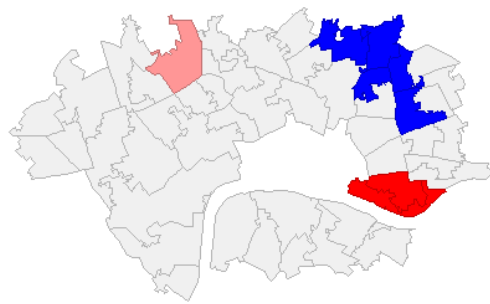


Figure 28: Local Moran cluster and outlier map for OLS regression residuals.



Cluster and outlier groupings

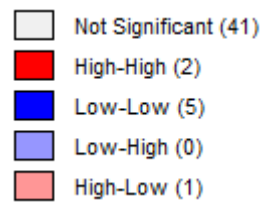
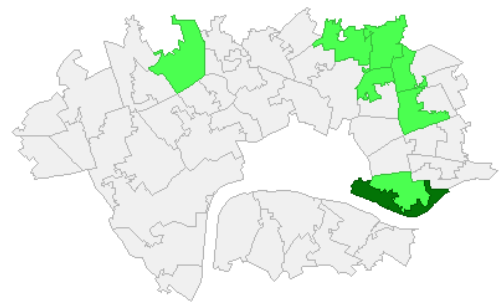
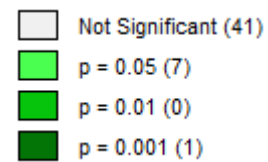


Figure 29: Local Moran significance map for OLS regression residuals.



LISA significance



In order to compensate for this, we can apply specialised spatial regression models. The nature of the best suited model depends on the nature of the spatial dependency identified: spatial error models when only the assumption of uncorrelated errors is broken, spatial lag models when both this and the assumption of independent observations are broken. The latter is the case when the dependent variable in a given location is affected by the variables of other spatially relevant locations.

The spatial dependency tests in our regression included Lagrange Multiplier tests for both spatial error and lag, and robust variants that test for these dependencies in the presence of the other dependency (Anselin, 1988). In our case, the significance of the robust test for lag suggests that this model is best suited. The robust test for error is no longer significant, showing that when the dependent variable is corrected for lag, the error dependency disappears.

5.3 Spatial Lag Model regression

A second regression was run on the variables, this time using the Spatial Lag Model (SLM). The model compensates for the diagnosed dependency by introducing an additional variable: for each geographical area the new variable's value is the average of the dependent variable in areas connected to it (by the spatial weight matrix). This means that the spatial relationship between neighbouring points is acknowledged within the regression itself.

The relevant results of the regression are in table 7.

Table 7: Results for spatial lag regression.

R-squared 0.699

Log likelihood 24.229

Akaike info criterion -42.457

Schwarz criterion -36.782

Variable	Coefficient	z-value	Probability
Booth score (normalised)	0.555	5.067	<0.001
Spatial lag adjustment	0.474	4.265	<0.001
Test for spatial dependence		Value	Probability
Likelihood Ratio Test		13.887	<0.001

The R-squared has now increased to 0.699, which we can interpret as meaning that it is able to explain almost 70% of our modern affluence score. The new variable introduced by the spatial lag model is also visible with a significant coefficient of 0.474, though our Booth score coefficient has fallen.

The Log likelihood, Akaike info and Schwarz criterion are used for model selection, allowing us to compare the quality of the model and its fit relative to other models. All three measures are improved in the SLM (smaller values are superior for the latter two), allowing us to confidently state that this model is superior to the OLS.

However, the SLM includes a Likelihood Ratio Test for the continued presence of spatial lag dependence, which has returned with a significant result. As such, though we have improved the model, we have not been able to completely account for the presence of spatial auto-correlation.

Figure 30: Quantile map for predicted values of SLM regression model.

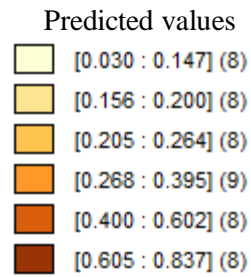
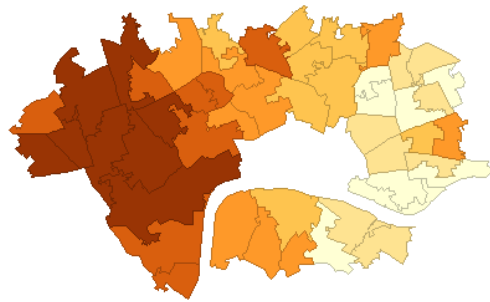
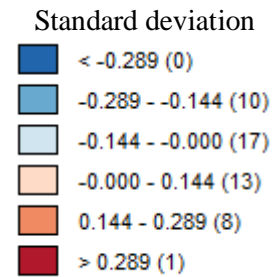
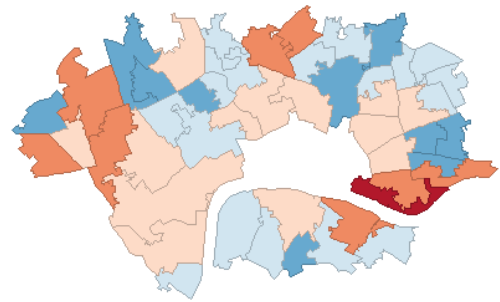


Figure 31: Standard deviation map of SLM regression residuals.



In order to investigate this further figure 30 shows the model's predicted values, and figure 31 is another standard deviation map of our residuals, in order to determine where the problematic areas are. Comparing the two allows us to see the shape of the model's predictions, and where it is underperforming. It should be noted that the symbology of figure 31 and the previous OLS residual map cannot be directly compared: the range of the second dataset is smaller due to improved fit.

Figure 32: Local Moran cluster and outlier map for SLM regression residuals.

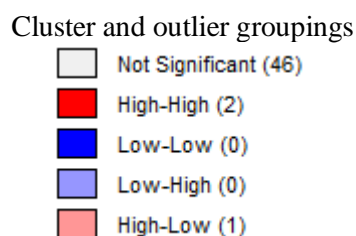
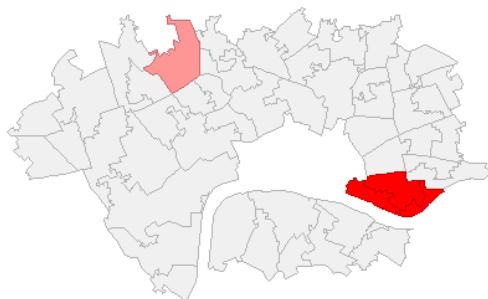
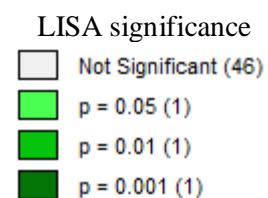
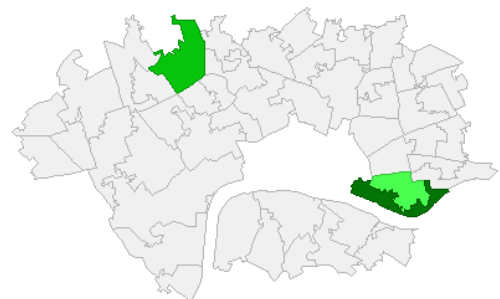


Figure 33: Local Moran significance map for SLM regression residuals.



A Local Moran test on the residuals shows that a section of Tower Hamlets in east London is the strongest source of disruption to the fit of our model, shown in figures 32 and 33. We will return to this outlier cluster in the next section. Before that we will consider one final question in regards to our regression analysis: to what extent are our Westminster MSOA controlling the results of our analysis?

5.4 Excluding Westminster MSOAs

The Westminster MSOAs have been repeatedly highlighted throughout this work: they dominate the highest ranks of both the Booth and modern affluence scores, exhibit strong clustering, and from our initial scatter plot in figure 24, appear to be relatively linear compared with other borough groups. In order to gauge the extent to which this is driving our regression results we will treat the entire Westminster group as outliers to be removed and will complete an OLS and spatial regression without them.

The results of this investigation should be considered only indicative, as the removal of the Westminster data will affect our results in other ways. For instance, the removal of these MSOAs (and the bridge connecting them to the south-west) means that some remaining MSOAs have lost one or more neighbours. Additionally, our overall selection size is reduced to 40 points of data.

Table 8 shows the outcome of the regression analyses. Unsurprisingly, the SLM was again found to be the most appropriate form of spatial regression.

Table 8: Comparison of Ordinary Least Squares and Spatial Lag Model regressions without Westminster MSOAs.

Regression model	OLS	SLM
R-squared	0.127100	0.421050
Coefficients and probability		
Booth score (normalised)	0.372477 (p 0.02394)	0.227704 (p 0.07671)
Spatial lag adjustment	-	0.589809 (p 0.00001)
Log likelihood	16.7533	22.8511
Akaike info criterion	-29.5065	-39.7023
Schwarz criterion	-26.1288	-34.6357

The removal of Westminster dramatically reduces the R-squared of the OLS regression to 0.127. However, with compensation for spatial dependency it rises to 0.421. This difference, though only indicative, does suggest that Westminster is exerting a strong influence on the overall relationship between Booth and modern affluence scores – but not so much that we should dismiss our overall model.

6. Ground-truthing our findings

Besides the overall relationship between our two scores, this analysis has also highlighted some areas exhibiting interesting characteristics: the Tower Hamlets outlier cluster that disrupted our main SLM model, and Westminster, exerting considerable influence over our model and dominating the upper ranks of both original datasets. In this section we will briefly examine these areas in order to confirm or dispute these findings.

6.1 Eastern outlier cluster

Figure 34 shows a map of the Tower Hamlets borough according to electoral wards. Our focal MSOAs are highlighted in green based on the significance in figure 33, showing that they correspond to the St Katherine and Wapping ward. However, ward names do not necessarily correspond to real world cohesive localities, the boundaries of which are fluid and sometimes contentious (Mills, 2013). Figure 35 shows an extract from a map showing identified localities of London, published by the Greater London Authority. Here the area is referred to solely as Wapping, which is the terminology also most familiar to this author.

Figure 34: Eastern MSOA cluster of interest, with ward boundaries.

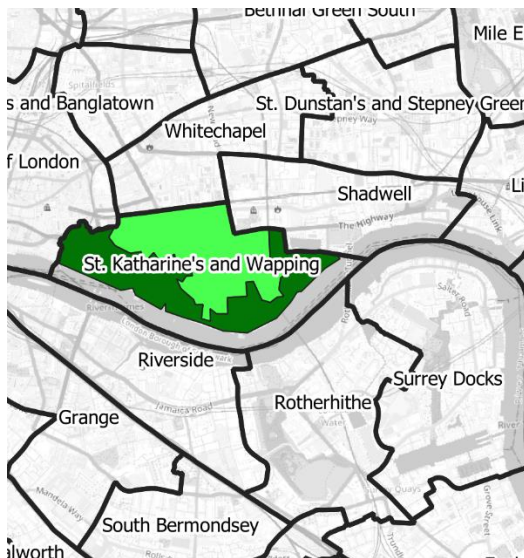


Figure 35: Localities map focused on eastern MSOA cluster of interest. Image downloaded from www.london.gov.uk/in-my-area, August 2019.



The outlier cluster detected in our regression is made up of two MSOAs, Tower Hamlets 026 and 027. Table 9 shows a summary of their data, including both Booth and modern affluence scores.

Table 9: Summary of eastern outlier cluster.

	Booth score (normalised)	Booth rank	MA score (normalised)	MA rank	Difference (normalised)	Rank of difference
TH26	0.126	47	0.39	17	0.264	9
TH27	0.171	44	0.563	9	0.392	3

Both MSOAs started with low Booth scores and have exhibited considerable change – TH27 in particular, with the third largest difference in normalised scores. We noticed this MSOA before in table 5, as the only area to move from the bottom 20% of Booth ranks to the top 20% of MA ranks. In order to determine if this data should be considered erroneous or otherwise remarkable, we will briefly consider the history of Wapping.

That history is defined by proximity to the river, and in Booth’s era Wapping was part of the docks of London, hosting London Docks in the east and St. Katherine’s docks in the west.

While the docks themselves were bustling, the income and living conditions of local labourers were bleak. Due to fluctuating need for labour – dependent on the arrival of ships – only a small proportion of labourers at the docks were permanently employed (Ackroyd, 2000). The overwhelmingly majority were hired for short term labour on a daily basis, selected from a crowd of hundreds who gathered at the gates of the docks (Thornbury, n.d.).

Walter George Bell wrote of a walk through Wapping in 1910, describing ‘reeking drink-shops, inexpressible in their squalor and dirt, the natural home for every kind of abomination’, and called it ‘the foulest, the most loathsome spot in all London’ (1919, pg. 114 and 124).

The docks themselves were heavily damaged in wartime bombing, and a short boom period afterwards ended with the advent of containerised cargo, which required larger ships than the docks could accommodate (London’s Royal Docks, 2018). By 1981 all docks had closed, and the government-owned London Docklands Development Corporation was established and made responsible for regenerating the area.

It is these regeneration efforts that likely explain the impressive change in the areas performance in its Booth and MA scores. Wapping itself became the new home of the News International media group (owners of the British newspapers The Sun and The Times) in 1986. The site of their headquarters has since been sold to make way for various residential and commercial developments, such as Clipper Wharf, with apartments starting from £660,000 (Barber, 2015). These prices are likely buoyed by the proximity to the financial district of Canary Wharf, also born out of the regeneration and which regularly appears on property management companies as a selling point for residing in the Wapping area (CBRE Limited, 2015; Movebubble Ltd, n.d.). Determining the validity and extent of this relationship would form an interesting follow on to this project, and extending the Booth vectorisation to include the full docklands area would allow its inclusion within our model.

High performing Westminster group

Our second area of interest encompasses the Westminster area – particularly those that dominate both the Booth and MA score ranks. Table 10 summarises the Westminster MSOAs, the upper seven MSOAs all rank in the top ten for both Booth and MA. These MSOAs have seemingly sustained their performance across both measures, 120 years apart.

Table 10: Summary of western group.

	Booth score (normalised)	Booth rank	MA score (normalised)	MA rank	Difference (normalised)	Rank of difference
WM 011	1	1	1	1	0	49
WM 018	0.959	2	0.896	2	0.064	33
WM 015	0.906	3	0.868	3	0.038	39
WM 012	0.817	4	0.761	5	0.055	34
WM 008	0.763	5	0.838	4	0.076	30
WM 013	0.74	7	0.735	6	0.006	47
WM 020	0.692	9	0.716	7	0.024	43
WM 021	0.505	13	0.552	10	0.047	35
WM 009	0.244	37	0.23	29	0.015	46

Figure 36 highlights these seven high performing MSOAs, and figure 37 shows a relevant extract of the map of localities. This group of high value MSOAs takes in a number of interesting localities: Mayfair, Marylebone, Soho, Convent Garden, St James, Westminster, and Paddington South.

Figure 36: Western MSOA group of interest, with ward boundaries.

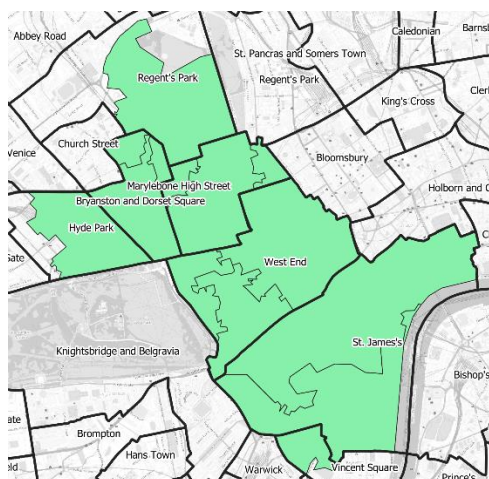


Figure 37: Localities map focused on western MSOA group of interest. Image downloaded from www.london.gov.uk/in-my-area, August 2019.



This is a broad area and one that defies easy summary. The Soho area was previously the home of the aristocracy of London, and various buildings from this period survive today. In Booth's time the area was undergoing transition, with theatres opening and restaurants gathering renown (Sheppard, 1966). Today it remains a hub for both gastronomy and entertainment, and is bordered to the north by Oxford Street, identified in 2017 as the busiest shopping street in Europe (BNP Paribas Real Estate, 2017).

It was Mayfair, and to a lesser extent Marylebone, which were in Booth's time the home of the aristocracy. Built and expanded on throughout the 18th century on the former estates of local gentry, many moved from older homes such as those in Soho to newer, more palatial mansions closer to the more historically aristocratic areas of St. James (Sheppard, 1966).

Housing in these areas has retained its relative value, and the area attracts international buyers looking for long term investments (Planet International UK, n.d.; Wetherell, n.d.). The Mayfair area is also well known as the most expensive square on the popular board game Monopoly, and was chosen as such in 1935. A 2013 study by Halifax Bank found that of all locations represented on the board, Mayfair remained the most expensive today (Property Wire, 2013).

Finally, the Westminster locality contains some of the most famous and important buildings in the United Kingdom, such as the Houses of Parliament, Westminster Abbey, and 10 Downing Street; home of the Prime Minister. It also contains Buckingham Palace, the London residence

of the monarch of the United Kingdom. On Booth's map the palace is diligently coloured yellow, the highest category, and unsurprisingly a review of council tax data for the palace today shows an allocation to the highest possible band.

These brief narrative excursions, while no means exhaustive, suggest narratives in alignment with our observations on the Westminster data. Further study and analysis of the experience of these, and other highlighted areas, would likely reveal greater insight into the correlation between conditions in these MSOAs during Booth's era and our own.

7. Conclusions

The first part of this project was concerned with vectorisation of the Booth poverty maps. A review of automatic and manual vectorisation options determined that the latter methodology was most suited to the Booth maps, given challenges associated with various artifacts and inconsistency in symbology. Eight of the twelve Booth maps were then at least partially vectorised, creating a polygon data set covering a contiguous region of central London, encompassing 49 MSOAs.

Though Booth's maps have previously been subjected to various forms of digitisation, most examples reviewed were considerably smaller in scope, focusing on selected neighbourhoods (Lindsay, 2008; The Economist, 2006; Vaughan, 2007; Vaughan and Geddes, 2009). Only Orford et al. attempted a similar larger scale analysis, and used a point-based digitisation process. As such, this project has produced a new resource that can be expanded upon, no doubt improved, and used for further analysis of the Booth poverty maps.

The second part of this project focused on analysis of the Booth data, our modern affluence measure based on council tax band data, and regression analysis of the two. From this we have established the existence of a strong relationship between the Booth and modern affluence data. Using a spatial lag model to account for spatial dependence in our data, we showed that the Booth poverty map data is a strong predictor for modern affluence.

Through briefly investigating the eastern cluster, where our regression model was underpredicting modern affluence, we identified some characteristics that suggested some initial explanations for this deviation from the model – primarily regeneration of the area and proximity to Canary Wharf.

The implications of these findings are complex, and invite follow up work to add detail, explanation and to challenge or corroborate. The strength of the predictive relationship between the Booth and MA scores suggest that patterns of relative affluence have remained largely intact over the 120 years since Booth's data was collected. Individual MSOAs have experienced varying changes in rank, but these are generally minor and have not disrupted overall patterns of

affluence. However, the example of the Wapping cluster suggests that regeneration efforts can disrupt the continuity of these patterns.

If the vectorised area of the Booth maps was extended to include the whole docklands area, and particularly Canary Wharf itself, we would gain greater insight into the impact of the regeneration on the Booth-MA relationship. However, though the change experienced by the docklands is remarkable, it is not the only area of the city that has been subject to regeneration. It would be useful to determine the location of noteworthy areas of regeneration within the study area, in order to better understand the impact on our model.

There are other interesting findings that suggest follow on work. Examination of the two data sets showed that while some poor areas had remained relatively less affluent and others had increased dramatically in affluence, richer areas had much lower tendency to experience significant decreases in affluence. Our brief overview of the Westminster area of interest suggested our data is accurate, but did not provide any insight into this resilience.

Through developing our understanding of how areas have experienced the intervening 120 years, we may then be able to improve and add complexity to the modelled relationship between Booth's London and our own. Of particular interest might be characteristics of the areas that provide other links between the two time periods, such as the presence of surviving buildings.

It would also be valuable to identify other historical resources describing patterns of affluence in London for other points in time between Booth's inquiry and our modern data. Building a timeline of data would allow us to better understand the impact of changes over time, such as regeneration efforts.

Finally, looking to the future, it would be interesting to periodically repeat this exercise with new affluence data, in order to track the longevity of the patterns of wealth and poverty observed in Booth's data, and how they might continue to influence successive generations of Londoners.

Bibliography

- Ackroyd, P., 2000. *London: The Biography*. Chatto & Windus.
- Anselin, L., 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis* **27**, 93–115. <https://doi.org/10.1111/j.1538-4632.1995.tb00338.x>
- Anselin, L., 1988. Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis* **20**, 1–17. <https://doi.org/10.1111/j.1538-4632.1988.tb00159.x> (accessed 28 August 2019)
- Arteaga, M.G., 2013. Historical Map Polygon and Feature Extractor, in: *Proceedings of the 1st ACM SIGSPATIAL International Workshop on MapInteraction, MapInteract '13*. ACM, New York, NY, USA, pp. 66–71. <https://doi.org/10.1145/2534931.2534932> (accessed 28 August 2019).
- Atkinson, A.B., Piketty, T. (Eds.), 2007. Top incomes over the twentieth century: a contrast between continental European and English-speaking countries. Oxford University Press, Oxford ; New York.
- Bales, K., 1994. Early innovations in social research: the Poverty Survey of Charles Booth (phd). The London School of Economics and Political Science (LSE).
- Barber, L., 2015. Clipper Wharf: Luxury flats at former News International HQ and offices of News of the World in Wapping go on sale with million pound price tags [WWW Document]. *CityAM*. URL <https://www.cityam.com/flats-former-news-international-hq-and-offices-news-world-go-sale-million-pound-price-tags/> (accessed 28 August 2019).
- Batty, M., Longley, P., 2003. *Advanced Spatial Analysis: The CASA Book of GIS*.
- BBC, n.d. *Mapping London* [WWW Document]. URL http://www.bbc.co.uk/london/content/articles/2006/12/04/kurt_london_maps_feature.shtml (accessed 28 August 19).
- BNP Paribas Real Estate, 2017. *Pan-European Footfall Analysis 2017-2018*.
- British Library, n.d. *London: A Life in Maps* [WWW Document]. URL <https://www.bl.uk/londoninmaps> (accessed 28 August 19).
- CBRE Limited, 2015. *Wapping | London Borough Area Guide | CBRE Residential UK* [WWW Document]. CBRE. URL <https://www.cbreresidential.com/uk/en-GB/area-guides/wapping> (accessed 28 August 2019).
- Cliff, A.D., Ord, J.K., 1973. *Spatial autocorrelation*. Pion, London.
- Dept. for Communities and Local Government, 2015. *English Index of Multiple Deprivation*. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/464430/English_Index_of_Multiple_Deprivation_2015_-_Guidance.pdf (accessed 28 August 2019).
- GLA, n.d. *MSOA Atlas - London Datastore*. URL <https://data.london.gov.uk/dataset/msoa-atlas> (accessed 28 August 19).
- Goebel, J., 2007. Methodological issues in the measurement of income and poverty. Doctorial dissertation, Technische Universität Berlin.
- Gregory, I.N., Healey, R.G., 2007. Historical GIS: structuring, mapping and analysing geographies of the past. *Progress in Human Geography* **31**, 638–653. <https://doi.org/10.1177/0309132507081495> (accessed 28 August 2019).
- Heywood, 2006. *An introduction to geographical information systems 3rd edition*. Pearson.

- Jenks, G.F., 1981. Lines, Computers, and Human Frailties*. *Annals of the Association of American Geographers* **71**, 1–10. <https://doi.org/10.1111/j.1467-8306.1981.tb01336.x> (accessed 28 August 2019).
- Knowles, A.K., 2005. Emerging trends in historical GIS. *Historical Geography* **33**, 7–13.
- Liao, S., Bai, Z., Bai, Y., 2012. Errors prediction for vector-to-raster conversion based on map load and cell size. *Chinese Geographical Science* **22**, 695–704. <https://doi.org/10.1007/s11769-012-0544-y> (accessed 28 August 2019).
- Lindsay, J., 2008. Measuring the persistence of poverty in East London. *Information, society and justice journal* **2**, 37–45.
- London School of Economics & Political Science, 2016. *Charles Booth's London* [WWW Document]. URL <https://booth.lse.ac.uk/> (accessed 28 August 19).
- London School of Economics & Political Science, n.d. *What were the poverty maps?* | Charles Booth's London [WWW Document]. URL <https://booth.lse.ac.uk/learn-more/what-were-the-poverty-maps> (accessed 25 August 19).
- London's Royal Docks, 2018. *London's Royal Docks History - Official Timeline* [WWW Document]. London's Royal Docks. URL <https://www.londonsroyaldocks.com/londons-royal-docks-history/> (accessed 28 August 2019).
- Lou, X., Huang, W., Shi, A., Teng, J., 2005. Raster to vector conversion of classified remote sensing image. *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS '05.* **5**, 3656–3658. <https://doi.org/10.1109/IGARSS.2005.152664> (accessed 28 August 2019).
- Manley, D., 2014. Scale, Aggregation, and the Modifiable Areal Unit Problem, in: Fischer, M.M., Nijkamp, P. (Eds.), *Handbook of Regional Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1157–1171. https://doi.org/10.1007/978-3-642-23430-9_69 (accessed 28 August 2019).
- Mills, E., 2013. Where does Dalston start and end? A study in place identity. URL http://ma3t.co.uk/euanmills/euanmills/tifd_files/tifd.pdf (accessed 28 August 19).
- Mooney, P., Minghini, M., 2017. A review of OpenStreetMap data. pp. 37–59. <https://doi.org/10.5334/bbf.c> (accessed 28 August 2019).
- Moran, P.A.P., 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)* **10**, 243–251.
- Movebubble Ltd, n.d. *Where to live in London if you work in the banking sector* [WWW Document]. URL <https://www.movebubble.com/london/renting-in/2017/08/the-best-places-to-live-in-london-if-you-work-in-the-banking-sector-new> (accessed 28 August 2019).
- NYPL Labs, 2019. *An open-source map vectorizer*. The New York Public Library - NYC Space/Time Directory. URL <https://github.com/nypl-spacetime/map-vectorizer> (accessed 28 August 2019).
- O'Day, R., Englander, D., 2005. *Mr. Charles Booth's Inquiry: Life and Labour of the People in London Reconsidered*. Bloomsbury Publishing PLC.
- Office for National Statistics, 2012. *Changes to Output Areas and Super Output Areas in England and Wales, 2001 to 2011*.
- Openshaw, S., 1984. *The modifiable areal unit problem*. GeoBooks, Norwich.
- OpenStreetMap Foundation, n.d. OpenStreetMap [WWW Document]. *OpenStreetMap*. URL <https://www.openstreetmap.org/about> (accessed 28 August 2019).

- Orford, S., Dorling, D., Mitchell, R., Shaw, M., Smith, G.D., 2002. Life and death of the people of London: a historical GIS of Charles Booth's inquiry. *Health & Place* **8**, 25–35. [https://doi.org/10.1016/S1353-8292\(01\)00033-8](https://doi.org/10.1016/S1353-8292(01)00033-8) (accessed 28 August 2019).
- Pickering, W.S.F., 1972. Abraham Hume (1814-1884): A Forgotten Pioneer in Religious Sociology. *Archives de Sciences Sociales des Religions* **33**, 33–48. <https://doi.org/10.3406/assr.1972.1876> (accessed 28 August 2019).
- Planet International UK, n.d. *Mayfair is at the very heart of the prime Central London property market* [WWW Document]. URL <http://www.planetinternationaluk.com/mayfair-is-at-the-very-heart-of-the-prime-central-london-property-market/> (accessed 28 August 2019).
- Property Wire, 2013. *Mayfair still the most expensive street on the Monopoly board* [WWW Document]. PropertyWire. URL <https://www.propertywire.com/news/europe/uk-properties-monopoly-prices/> (accessed 28 August 2019).
- Sheppard, F.H.W. (Ed.), 1966. “General Introduction”, in *Survey of London: Volumes 33 and 34, St Anne Soho*. pp. 1–19.
- Teng, J., Wang, F., Liu, Y., 2008. An Efficient Algorithm for Raster-to-Vector Data Conversion. *Annals of GIS* **14**, 54–62. <https://doi.org/10.1080/10824000809480639> (accessed 28 August 2019).
- The Economist, 2006. Booth redux. *The Economist*, 4 March 2006.
- Thornbury, W., n.d. *St. Katherine's Docks* | *British History Online* [WWW Document]. URL <https://www.british-history.ac.uk/old-new-london/vol2/pp117-121> (accessed 28 August 2019).
- Vaughan, L., 2018. Charles Booth and the mapping of poverty, in: *Mapping Society, The Spatial Dimensions of Social Cartography*. UCL Press, pp. 61–92. <https://doi.org/10.2307/j.ctv550dcj.8> (accessed 28 August 2019).
- Vaughan, L., 2007. The spatial form of poverty in Charles Booth's London, in: Vaughan, L. (Ed.), *The Spatial Syntax of Urban Segregation*. Elsevier, pp. 231–250.
- Vaughan, L., Geddes, I., 2009. Urban form and deprivation: A contemporary proxy for Charles Booth's analysis of poverty. *Radical Statistics* Issue 99.
- Vickers, D., Rees, P., Birkin, M., n.d. Creating the National Classification of Census Output Areas: Data, Methods and Results 81.
- VOA, n.d. *Council tax: stock of properties, 2018* [WWW Document]. GOV.UK. URL <https://www.gov.uk/government/statistics/council-tax-stock-of-properties-2018> (accessed 28 August 2019).
- Weide et al, 2017. Is Inequality Underestimated in Egypt? Evidence from House Prices [WWW Document]. URL <https://webcache.googleusercontent.com/search?q=cache:J9ZJuefKTh0J:documents.worldbank.org/curated/en/588041468195577344/pdf/WPS7727.pdf> (accessed 28 August 2019).
- Wellcome Collection, n.d. *Living with Buildings* | *Wellcome Collection* [WWW Document]. URL <https://wellcomecollection.org/exhibitions/Wk4sPSQAACcANwrX> (accessed 28 August 2019).
- Wetherell, n.d. *Who lives in Mayfair* [WWW Document]. Wetherell. URL <https://wetherell.co.uk/mayfair/lives-mayfair/> (accessed 28 August 2019).