London Output Area Classification

Technical Manual

28/08/23

Consumer Data Research Centre
Authored by: Alex Singleton, Paul Longley

Consumer Data Research Centre

Technical Manual

Introduction

The purpose of this document is to provide an overview of how the London Output Area Classification 2021 (hence forth, LOAC 2021) was created. This supplements full documentation in the code associated with the classification which can be found on our github repository: https://github.com/alexsingleton/LOAC_2021.
The overall process of creating LOAC 2021 broadly mirrors the methodology adopted in the creation of the national Output Area Classification (developed in partnership with the Office for National Statistics). The classification was created in partnership with the Greater London Authority (GLA) and with the support of the LOAC 2021 advisory group. The LOAC advisory group was assembled by the GLA to represent constituent future stakeholder users of the classification[1].

Data and Input Preparation

Inputs were all sourced from the 2021 Census at Output Area zonal geography. The input data were across a series of domains, representing key areas of influence on residential differentiation. The input variables mirror those used to create the national classification, apart from the industry variables (V61-v68) that were added as additional LOAC 2021 inputs at the request of the GLA to better capture London's complex employment structure. Most of the input measures were created as proportions, with the exception of v01, calculated as the usual residents per square kilometre; and v42, which was an age standardised disability ratio.
To improve the normality of the raw input data an inverse hyperbolic sine transformation was applied, and to ensure all variables are comparable and measured on the same scale, range standardisation was applied. Both choices mirror the England and Wales classification and are also consistent with the methods used in both the 2011 London and the 2011 UK national Output Area Classifications.

[1] Members included: Richard Cameron (GLA); Thomas Morgan (Hounslow); M Sinclair (Lambeth); R May (Lambeth); Elena Hido (Royal Borough of Kensington and Chelsea); Chris Sale (Newham); James Rapkin (Barnet); Katherine Prior (Enfield); Hubert Senyah (Waltham Forest); Martin Dittus (Lewisham); Llkka Sipila (Lewisham); William Cocodia (Lewisham); Keira Chapman (Southwark); Harry Vater (Metropolitan Police); Amy Springett (Royal Borough of Kensington and Chelsea); Neil Storer (Camden); Katherine Blair (Transport for London); Katharine Prior (Enfield).

Table 1: LOAC 2021 Input Variables

Domain	No.	Variable Name
Demographic	v01	Usual residents per square kilometre
	v02	Aged 4 years and under
	v03	Aged 5 to 14 years
	v04	Aged 25 to 44 years
	v05	Aged 45 to 64 years
	v06	Aged 65 to 84 years
	v07	Aged 85 years and over
Ethnicity and Origins	v08	Country of birth: Europe: United Kingdom
	v09	Country of birth: Europe: EU countries
	v10	Country of birth: Europe: Non-EU countries
	v11	Country of birth: Africa
	v12	Ethnic group: Bangladeshi
	v13	Ethnic group: Chinese
	v14	Ethnic group: Indian
	v15	Ethnic group: Pakistani
	v16	Ethnic group: Other Asian
	v17	Ethnic group: Black
	v18	Ethnic group: Mixed or Multiple ethnic groups
	v19	Ethnic group: White
	v20	Cannot speak English well or at all
	v21	No religion
	v22	Christian
	v23	Other religion
Living Arrangements	v24	Never married and never registered a civil partnership
	v25	Married or in a registered civil partnership
	v26	Separated or divorced
	v27	One-person household
	v28	Families with no children
	v29	Families with dependent children
	v30	All household members have the same ethnic group
Usual Residence	v31	Lives in a communal establishment
	v32	Address one year ago is the same as the address of enumeration
	v33	Detached house or bungalow
	v34	Semi-detached house or bungalow
	v35	Terraced (including end-terrace) house
	v36	Flat, maisonette or apartment
	v37	Ownership or shared ownership
	v38	Social rented
	v39	Private rented
	v40	Occupancy rating of rooms: +1 or more
	v41	Occupancy rating of rooms: -1 or less
Health and Education	v42	Standardised Disability Ratio
	v43	Provides no unpaid care
	v44	2 or more cars or vans in household
	v45	Highest level of qualification: Level 1, 2 or Apprenticeship
	v46	Highest level of qualification: Level 3 qualifications
	v47	Highest level of qualification: Level 4 qualifications or above
Employment	v48	Hours worked: Part-time
	v49	Hours worked: Full-time
	v50	NS-SeC: L15 Full-time students
	v51	SOC: 1. Managers, directors and senior officials
	v52	SOC: 2. Professional occupations

Domain	No.	Variable Name
	v53	SOC: 3. Associate professional and technical occupations
	v54	SOC: 4. Administrative and secretarial occupations
	v55	SOC: 5. Skilled trades occupations
	v56	SOC: 6. Caring, leisure and other service occupations
	v57	SOC: 7. Sales and customer service occupations
	v58	SOC: 8. Process, plant and machine operatives
	v59	SOC: 9. Elementary occupations
	v60	Economically active: Unemployed
	v61	Agriculture, energy and water
	v62	Manufacturing
	v63	Construction
	v64	Distribution, hotels and restaurants
	v65	Transport and communication
	v66	Financial, real estate, professional and administrative activities
	v67	Public administration, education and health
	v68	Other

Cluster Analysis

Following the transformation of the input data, the next stage was to apply cluster analysis to identify groupings of Output Areas within London that shared the greatest similarity. The objective was to create LOAC 2021 as a two-tier nested classification, in which 2021 Output Areas were organised into larger and more aggregate Supergroups, and then split into nested and more detailed Groups. In this instance the classification was created from the "top-down", that is, the most aggregate Supergroups were created first, followed by the nested Groups. The practical implementation was that after applying the cluster analysis to create identify a convenient number of Supergroups, the input data for Output Areas assigned to each Supergroup were then partitioned into convenient numbers of Groups.

This analysis was performed using a k-means clustering algorithm, which is a standard approach used to create geodemographic classifications. This also mirrors the methods implemented in the 2011 London and 2011 UK and 2021 England & Wales Output Area Classifications. One feature of k-means clustering is that the number of clusters used at Supergroup and Group levels must be pre-specified. There are many ways to identify an ideal number of clusters, however for this application we applied a Clustergram. A Clustergram is a visualisation technique that plots the weighted first component of a principal components analysis by cluster, for a series of different cluster solutions: in this instance, from two to fourteen clusters. Lines

widths are scaled by the number of Output Areas that are assigned to the clusters as the number of clusters increases. These diagrams are interpreted by looking for a cluster value that offer a good separation of the red dots (representing cluster means) on the y axis. In this instance six cluster Supergroup solution was selected. When creating Groups, Clustergrams were also used to select the cluster frequencies for the data subsets. Supergroups were split into either two or three clusters, making a total of 16 Groups.

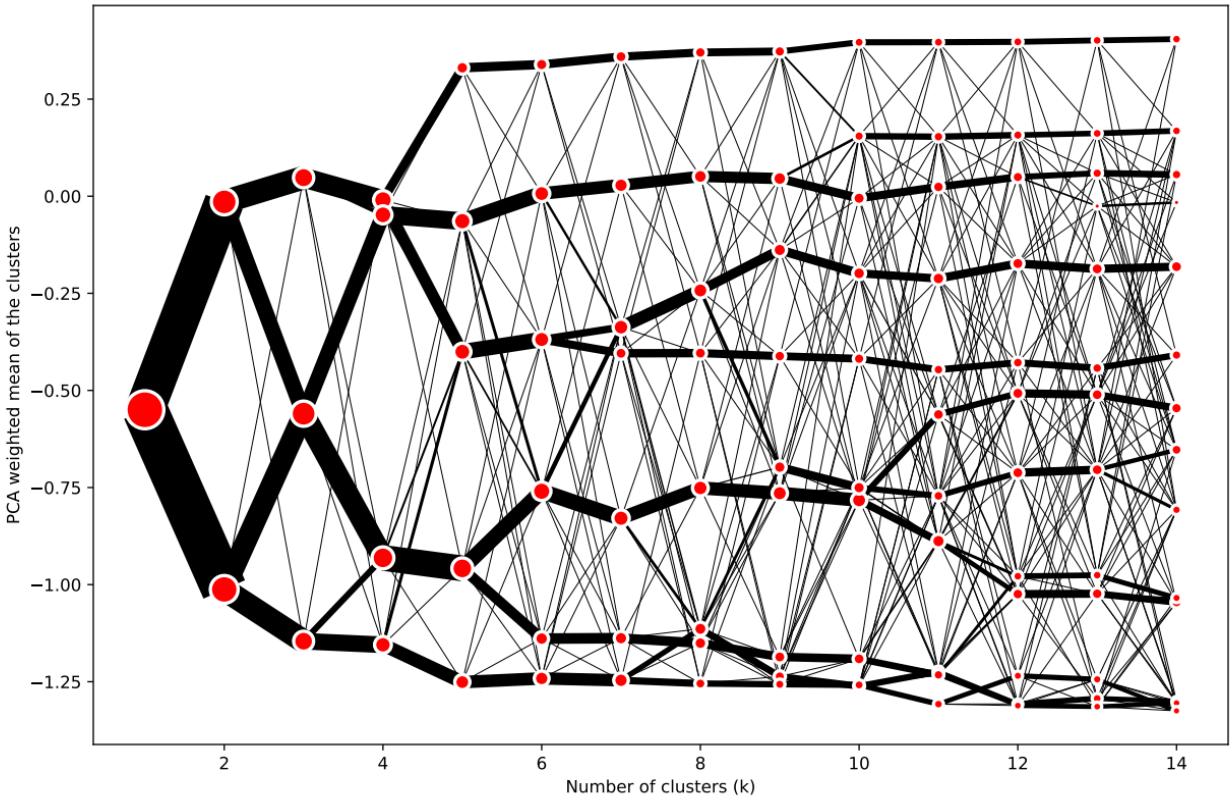


Figure 1: A Clustergram was used to select the number of clusters for the Supergroups

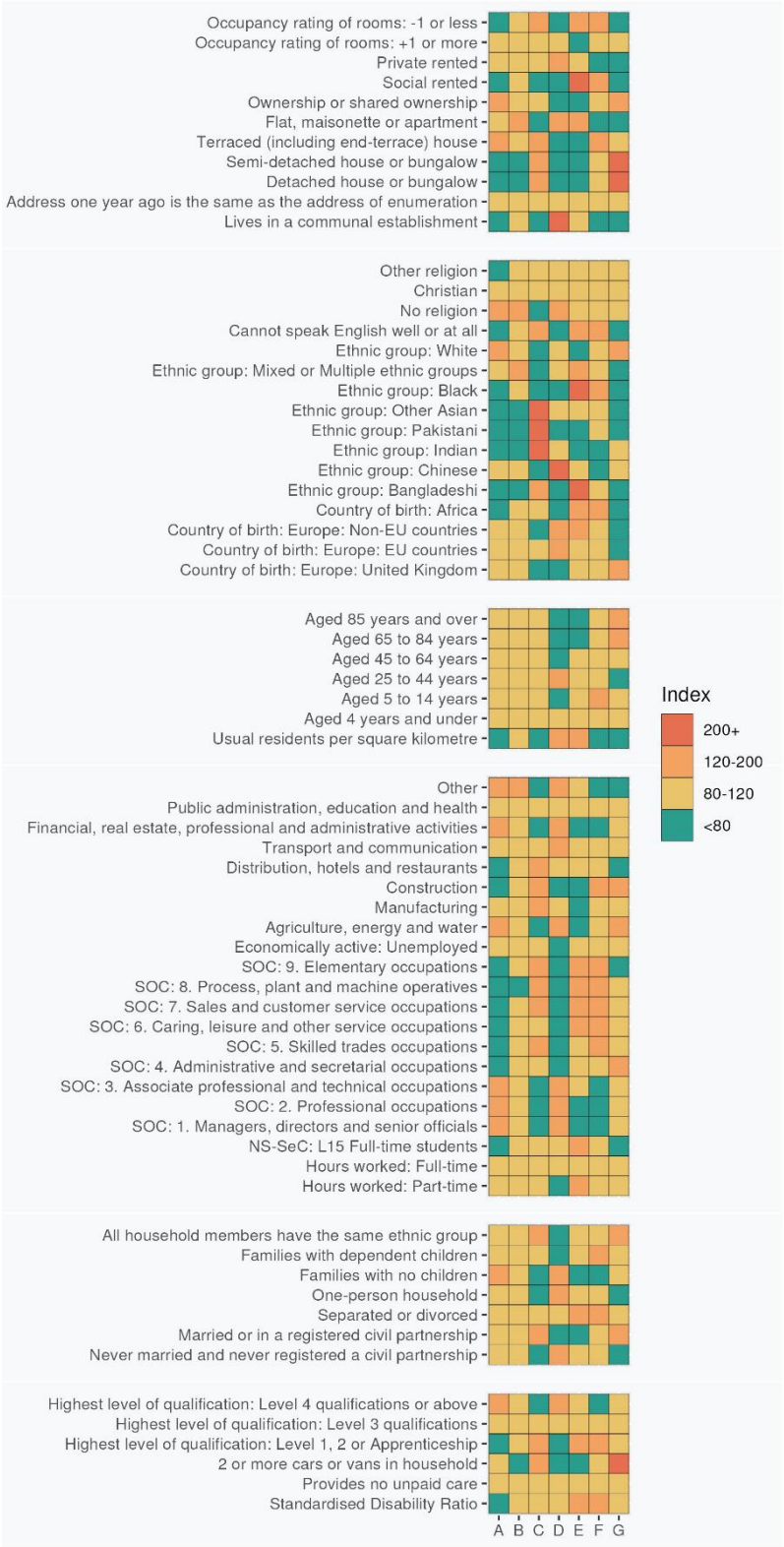
Cluster Descriptions

A standard way in which geodemographic classifications are described is to create a "grand index" of those variables used as input to the classification. These show whether each variable is over or under represented within the cluster and can be used to create written descriptions of the cluster and also create a name. The scores are typically scaled around 100, so in this instance a score of 100 would equate to a

variable demonstrating a rate within a cluster around the London average. A score of 200 would be double the London average and score of 50, a half. A 'rule of thumb' that has commonly been applied when describing geodemographic clusters, is that the most useful range to examine are those scores falling outside of the 80-120 range. A chart was created for both Supergroup and Group index scores (see

Figure 2), with the coloured bands highlighting score ranges. Supergroup descriptions are created to represent the main differentiating features within clusters relative to the London average. For Groups, these descriptions focus on those variables that have a different expression within the Group relative to the Supergroup. This aims to avoid repetition with the description for the nested Supergroup, focusing differentiating features.

As described in the next section, the LOAC 2021 Advisory Group was convened to ground truth the descriptions, both at the Supergroup and Group naming stage. Additional materials used to describe the clusters included maps.



a) Supergroup index scores

b) Group index scores

Figure 2: A Plot of the Index Scores for LOAC 2021 Supergroups and Groups

End User Consultation

During the creation of LOAC 2021, two consultation meetings were held with the Advisory Group. The initial meeting first presented an overview of the methodology to create LOAC 2021, informed by the process being implemented in the national classification alongside amendments specifically for London around input variable selection. This meeting also presented a pilot set of Supergroup clusters alongside maps and descriptions. The methods were approved, and the Supergroup clusters were explored in depth during the meeting, eliciting feedback. Further feedback was sought after the meeting by email and collated by the GLA. Finally, the meeting also presented suggested splits for the Supergroups, that would form the groups. After amendments to the Supergroups were made in response to the feedback, Group level partitions were then created alongside draft descriptions and labels. These were circulated alongside a written response to the meeting one feedback. The purpose of meeting two was to present the amended Supergroups and the draft Groups, again seeking feedback. A second discussion also enquired what types of deliverables would be of most utility to the advisory group to encourage applications, alongside additional discussion of ways in which the Economic and Social Research Council (ESRC) Consumer Data Research Centre will support LOAC 2021. A final set of Supergroup and Group labels and descriptions was circulated for approval by the GLA.