Singleton and Spielman *EPJ Data Science* (2024) 13:34
https://doi.org/10.1140/epjds/s13688-024-00466-1

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Segmentation using large language models: A new typology of American neighborhoods

Alex D. Singleton[1*] and Seth Spielman[2]

*Correspondence:
alex.singleton@liverpool.ac.uk
[1]Geographic Data Science Lab, University of Liverpool, Roxby Building, 74 Bedford St South, L69 7ZT, Liverpool, UK
Full list of author information is available at the end of the article

Abstract

In the United States, recent changes to the National Statistical System have amplified the geographic-demographic resolution trade-off. That is, when working with demographic and economic data from the American Community Survey, as one zooms in geographically one loses resolution demographically due to very large margins of error. In this paper, we present a solution to this problem in the form of an AI based open and reproducible geodemographic classification system for the United States using small area estimates from the American Community Survey (ACS). We employ a partitioning clustering algorithm to a range of socio-economic, demographic, and built environment variables. Our approach utilizes an open source software pipeline that ensures adaptability to future data updates. A key innovation is the integration of GPT4, a state-of-the-art large language model, to generate intuitive cluster descriptions and names. This represents a novel application of natural language processing in geodemographic research and showcases the potential for human-AI collaboration within the geospatial domain.

Keywords: Geodemographics; Large Language Model (LLM); American Community Survey; Segmentation; Neighborhoods; Artificial Intelligence (AI); Demographics; Retreival Augmented Generation (RAG)

1 Introduction

For most of the past century, there has been an abundance of data available about places in the United States. The US Census Bureau's American Community Survey (ACS) publishes hundreds of tables each year for nearly 300,000 different geographic units. The ACS is the primary source of social and economic data in the US but the abundance of data that it provides masks a fundamental problem- much of data is low quality, especially at the highest levels of geographic detail (Block Groups). This creates a paradox whereby as one zooms-in geographically the resolution of social and economic data decreases; this is counter-intuitive to most data users who are used to zooming in on maps and seeing more, not less detail. This occurs because the ACS, as the name implies, is a survey, and as one zooms in geographically the sample is diluted. Spielman and Folch [1] document that in 72% of all areas reported by the census the margin of error around the estimated number of children under 5 in poverty is greater than the estimate, for example, 25 +/− 30 (implying that the true number of children in poverty is somewhere between zero and 55). The

Springer

problem is so acute that the US Census Bureau advises against the use of the most detailed levels of ACS geography, noting "For the ACS, block groups are not designed to be used individually, rather they provide a smaller geographic area than census tracts that allow data users to combine them to create larger geographic areas that may be more meaningful for their specific use" [2]. That is, the US Census Bureau advises data users to build lower resolution geographies out of Census Block Groups and *not* to use them directly.

This pattern of having a high volume of low quality data is not unique to the American Community Survey. The new operating environment for many social scientists is one in which large volumes of "organic" data of dubious quality and/or for which provenance are unavailable. In a world where data are abundant, but individual variables are of low quality, identifying groups or patterns in a data set can be a critical way to find meaningful insights. Online data about people often follows this pattern of large volumes of low quality data being available, in this setting segmentation (or profiles) are the norm, as a recent leak from one of the largest online advertising platforms makes clear.[1]

Although combining large numbers of variables via classification tends to be the norm in the private sector, arguably fewer options are available for scientists and policy makers. The lack of options are partially driven by demand, as their dominant mode of inquiry leans on individual variables ([3]). However, another driver of the reliance on low quality variables is the lack of alternatives in the public domain. We believe that this gap is due to a failure of the research community to adapt to the changing nature of the data landscape, while the US National statistical system has undergone some profound changes, such as a 50% reduction in the sample size used to produce small area estimates.[2] Spielman and Singleton [4] argue that a solution to the decline in data quality is the aggregation of noisy signals to produce a more precise picture of a place than any single variable would allow. The logic behind this approach is that if data are unbiased but noisy, as one adds noisy but unbiased variables to an ensemble it becomes possible to make reliable inferences about cluster means because the expected value of the error across many variables is zero. However, the Spielman and Singleton approach also requires a fundamental change in practise, a shift from a "Variables Approach" in which one focuses on individual variables, describing places by race or income, to a "contextual approach" in which one describes places holistically using many variables at once.

Social Scientific practise has not fully adapted to this new operating environment where data are more abundant but lower quality. However, the growing availability of spatially referenced data about populations and their local contexts has led to an upsurge of interest in these high-dimensional descriptive models of places [5, 6]. Commonly referred to as segmentation or geodemographic classification [7, 8], they represent a collection of methodological approaches that utilise unsupervised classification techniques to group geographic areas based on the similarity of their characteristics in the socioeconomic, demographic, and built environment [9]. Such systems have proven to be valuable tools for understanding complex spatial patterns in urban areas [10], targeting public policy interventions [11], and guiding business decisions [12]. While several commercial geodemographic classification systems for the United States exist (e.g. Esri's Tapestry Segmenta-

[1] https://www.theguardian.com/australia-news/2023/aug/15/tab-gamblers-betting-australia-targeted-microsoft-xandr-advertising-database.

[2] Details of the survey can be found here—https://www.census.gov/programs-surveys/acs/.

tion,[3] Nielsen's PRIZM[4]), these systems are closed source and proprietary, which complicates their use for science, policy, and governance because the underlying data and methods cannot be independently verified. There remains a need for a transparent, reproducible and updateable classification system that is grounded in open scientific principles and can evolve its representations alongside changes in data [13].

One of the persistent challenges to the use of high-dimensional typologies in the social sciences is that it can be difficult to understand such systems. Given hundreds of variables and hundreds of groups, creating descriptions of clusters used to require a manual process of reading many tables of data and writing descriptions by hand. This process of description is often error prone; [14] noted that a major research endeavor in the 1970s was undermined by misleading interpretation of multivariate groupings of data. This paper addresses these challenges by developing an open and reproducible geodemographic classification at the block group scale for the United States using the ACS, and further demonstrates how by coupling this classification with Generative Pre-trained Transformer 4 (GPT4[5]) one can generate intuitive descriptions and names for high dimensional clusters. Integration of GPT4 into geodemographic typologies represents a novel application of natural language processing techniques in the interpretation of high-dimensional cluster analyses.

To create a robust and reproducible geodemographic classification, we utilise a partitioning clustering algorithm to identify areas with similar characteristics, employing a range of measures describing people, their behaviours and residential context. By utilising a pipeline of open-source software and sharing our code, data, and methods, we ensure that this classification is transparent, reproducible, and adaptable to future updates of the ACS. Our final classification organises the United States into 7 Groups which are further broken down into 39 Types; each accompanied by descriptive pen portraits.

Our paper proceeds as follows: Sect. 2 provides a brief overview of the history and evolution of geodemographics; Sect. 3 provides an overview of the ACS and pre-processing steps, including a rationale for the selection and evaluation of variables aligned to a conceptual framework; Sect. 4 describes the clustering algorithm implementation; Sect. 5 presents the results of the geodemographic classification including their spatial distribution, socio-economic and demographic characteristics. Section 6 details the integration of GPT4 for generating descriptions and names for the clusters, with a discussion on the benefits and limitations of this approach. Section 7 presents an evaluation of our classification system; and finally, Sect. 8 concludes the paper, highlighting the unique contributions of our work and outlines some potential avenues for future research.

2 The geodemographic characterisation of neighbourhoods

The process of building a geodemographic classification can be considered both art and science [8], as although many of the methods used in their creation are driven by algorithms, there are nuances to their implementation that mean that they are only partially automated. This includes the selection of variables that may have theoretical and empirical rationale [15], the choice of variable pre-processing, including normalisation and

[3] https://www.esri.com/en-us/arcgis/products/data/data-portfolio/tapestry-segmentation.

[4] https://claritas.com/prizm-premier/.

[5] https://openai.com/research/gpt-4.

standardisation, and the selection of cluster frequencies for the classification, including whether these should appear as a hierarchy. There are a variety of perspectives around the most effective approach [4, 8], all of which have the overarching objective of creating a geodemographic classification with descriptive validity for an intended application.

For most geodemographic classifications, after the creation of clusters, there is a more explicitly qualitative stage in the analysis that aims to provide descriptive portraits of the resulting groups. The process will usually involve the development of a table describing each input variable and its representation within each cluster relative to some base (e.g. national). The classification builder manually reads these tables to write "pen portrait" descriptions and give the clusters names. This process is time-consuming and, depending on the experience of the classification builder or the team of people working on these descriptions, may be of variable quality. We make the case here that through the integration of large language models, such as GPT4, these provide an alternative and efficient method for the creation of an interface to a segmentation.

3 Data and design

The US Census Bureau conducts the ACS which is a large and representative survey of the US population, comprising around two million interviews annually. The ACS provides hundreds of estimates for the more than 200,000 block groups in the United States. Block Groups are geographic units designed by the Census Bureau that contain between 600 and 3000 people, they vary in size (based on population density) but generally consist of a few city blocks.

Some simple maths tells you that it is difficult to produce annual estimates for 200,000 places from a survey of 2 million people, there simply are not enough people per block group to produce reliable estimates. So the "annual" estimates provided for block groups actually combine 5 years of data. The data used in this study were accessed through US Census Bureau's API via the TidyCensus R package [16], with count variables converted to proportions using the appropriate denominator variable (provided by the Census Bureau). There are numerous ways in which the input variables used to create a geodemographic classification can be selected. Most of the time, these blend empirical evaluation of their descriptive utility, experience of classification builders informed by previous work, and theoretical rationale for measures linked to drivers of residential differentiation.

The inputs for most geodemographic classifications are designed around an organising framework. These are often structured hierarchically and comprise broad "concepts" and more specific nested "domains". Within such frameworks, the specific labels for the groupings of input variables might change, and there are no fixed structures for how groupings of input measures could be presented. All these choices vary manually depending on the contexts for which a classification is being created. The purpose of such frameworks are to establish a greater theoretical basis for the choice of input variables, which is particularly important within the context of the social sciences, and follows practice that has been implemented since the first classifications were created in the 1970s [17]. The conceptual framework for this study is presented in Table 1 with three organising concepts that include economy, environment, and population. Then these concepts are broken down into 25 domains. This input taxonomy was manually designed to broadly represent key facets that are known to either differentiate between residential areas or are important outcomes of these patterns. The ACS has available over 18,000 potential variables, many of which

Table 1 Conceptual Framework for the Geodemographic

Concept	Domain
Economy	Digital Connectivity
	Digital Devices
	Employment
	Mobility
	Start Work Time
	Transport
	Wealth
Environment	Accommodation size
	Building Age
	Housing Cost
	Occupancy
	Rent
	Structure Type
	Tenure
	Unit Size
	Vacant Units
Population	Children
	Education
	Family
	Family Structure
	Household
	Housing
	Language
	Marriage
	Race / Ancestry / Citizenship

are not relevant for the context of this study, or as these are predominantly cross tabulations of different variables, for this geographic scale of analysis have very wide margins of error or large amounts of suppressed values as a result of disclosure controls. Based on a breadth of past classification builder experience and use of these data [4, 18] an initial selection of 697 variables encompassing all concepts were assessed based on their geographic coverage, correlation, and variance. The evaluation initially prioritised variables that offered near-complete coverage across the Block Groups of the United States and exhibited greater variance, which is a useful indicator of how effective a variable would be at differentiating between areas.

For variables deemed non-problematic in terms of coverage and variance, correlation was then considered. Ensuring that correlation is minimised (or understood) within cluster analysis inputs is important to prevent an over emphasis of a particular dimensions within the classification. It is common practice to limit these effects by excluding variables with high correlation, commonly defined by a Pearson correlation coefficient of +ve / −ve 0.7 [19]. The correlation coefficients between all variables were calculated and used to build a dense network from the associations of the variables. Each variable was represented as a node and the correlation coefficients linking the nodes were considered edges, with a weight assigned as the correlation coefficient value. Because correlation coefficients are calculated for all pairs of variables, the full network is dense, however, to make the graph more interpretable, edges where the correlation weights were +ve / −ve 0.7 were deleted, leaving a multi-part graph. A graph visualisation was then created where the nodes were coloured by the three Concepts; and the edges by the strength of their correlation. The advantage of a graph representation are that those variables with high correlation to multiple variables were immediately visible; enabling these groupings to be considered conceptu-

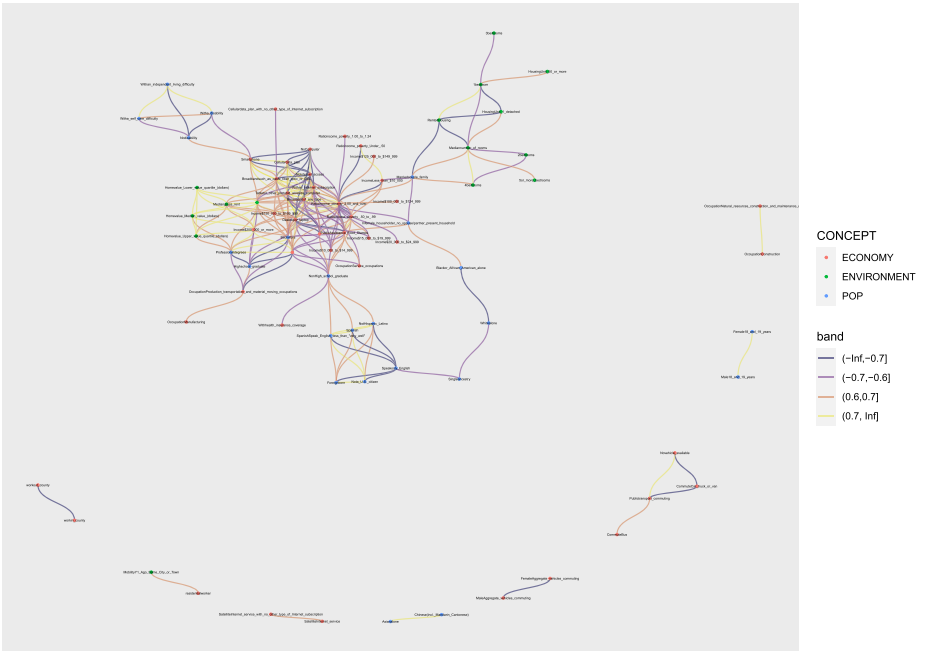

Figure 1 A Correlation Graph Showing Variables with High Correlation

ally and to ensure that changes in variable selection (due to high correlation) would not cause a loss of conceptual coverage. The graph guides the identification of variables to be removed or combined into new measures (Fig. 1). For example, the variables for males aged 18 to 19 and females aged 18 to 19 were highly correlated: with a decision taken to merge these variables into a new variable, aged 18 to 19. After this process of input refinement, a total of 247 input variables remained.

4 Clustering and model specification

Prior to cluster analysis each input measure was converted to a constrained logit, which shifted the measures onto a scale of −6 to +6 and adjusted the shape of their distribution. This approach was selected because, relative to other commonly used normalisation processes in geodemographic classification including taking a log [20], or inverse hyperbolic sine [19], a logit provides a more appropriate treatment for proportions and controls for extreme outliers. The constrained logit applied to the proportions is defined in Algorithm 1.

Geodemographic classifications typically use unsupervised clustering methods to find groups of areas with shared characteristics. Publicly available classifications follow two main design patterns, “top-down” or “bottom-up”. A top-down approach clusters all areas into a number of large aggregate groups. These groups are then used to partition the data, with separate cluster analysis conducted on the subsets to build a hierarchy from the top (most aggregate) downwards. Bottom-up associates areas sharing similar characteristics into a larger number of initial clusters and then bringing these clusters together into aggregate groups, commonly using hierarchical methods such as wards clustering.

Algorithm 1 Constrained Logit

```
function ConstrainedLogit(p, m = 6)
    if p is NA then
        return NA
    else if p ≤ 0 then
        return −m
    else if p ≥ 1 then
        return m
    else
        Calculate logit as log(p/(1−p))
        return min(max(logit, −6), 6)
    end if
end function
```

In both top-down and bottom-up approaches, k-means is the most prevalent unsupervised partitioning algorithm in geodemographic classification, although there are examples of other methods [18, 21]. For this application, a top-down approach was selected and implemented using a k-means which initially partitions by assigning k random seed locations within the multidimensional data array. For each data point, the Euclidean distance to each centroid is calculated, then observations are assigned to the closest centroid. A cluster mean is then computed from the data points assigned to that cluster, producing a new centroid. This process is repeated until there are no more changes in the cluster assignments or the centroid locations, or a predefined number of iterations have been reached. This process converges when the algorithm has found a local minimum for the objective function—the within-cluster sum of squares (WCSS). WCSS is a measure of how much variation exists within each cluster. These scores can be summed over all clusters to give the total within sum of squares and is typically used as a measure of the quality of a clustering solution, with lower values indicating tighter and more distinct clusters.

Let C_i be the cluster containing data points $x_1, x_2, \ldots, x_{n_i}$, and let $\bar{x_i}$ be the centroid of cluster C_i. Then, the WCSS for cluster C_i can be calculated as:

$$WCSS_i = \sum_{j=1}^{n_i} \|x_j - \bar{x_i}\|^2$$

The total WCSS (TWCSS) for a clustering solution with k clusters is the sum of the WCSS for each cluster:

$$TWCSS = \sum_{i=1}^{k} WCSS_i$$

In geodemographic classification, there is no a priori knowledge of which value of k is most appropriate, and as such, this requires a set of heuristics to guide the selection. The most prevalent method to estimate k includes the use of an elbow criterion, that plots the change in TWCSS for different values of k, although other methods such as silhouette scores and gap statistics have also been used. However, for this example, we utilise a Clustergram to select the initial and then subsequent partitioning of the data. For each cluster,

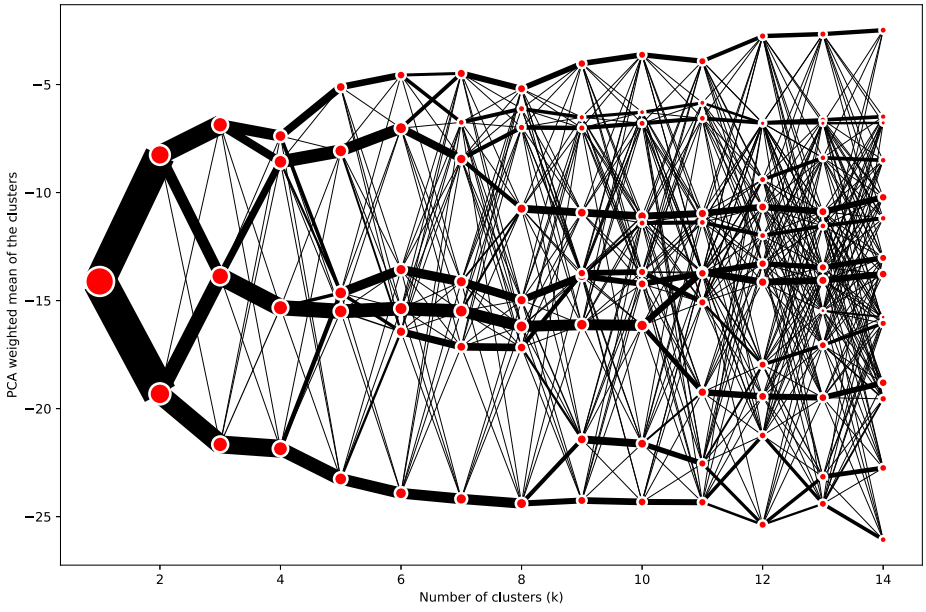

Figure 2 A Clustergram Illustrating k Potential Initial Data Partitions

Table 2 k Subsequent Data Partitions

Cluster	*Splits*
1	5
2	5
3	7
4	5
5	5
6	6
7	6

this visual tool plots cluster centroids along the y axis based on a weighted first component of a principal component analysis. The thickness of the lines between solutions relates to the number of areas assigned to clusters across the different values of k, indicating large, small or stable splits. An indicator of the most appropriate value of k is where the cluster centroids are well separated on the y axis. A Clustergram is shown in Fig. 2 and a value of 7 was selected.

A further feature of k means is that the initial location of the seeds impacts the final quality of the clustering result. Following previous geodemographic applications, random seed locations were allocated for 10,000 runs of the algorithm, selecting the result that minimised the TCWSS, representing more compact clusters. The input data were then partitioned by the seven cluster solution into separate datasets, and further Clustergrams estimated appropriate splits. The application of k means to these data subsets, enabled a second hierarchical tier of the classification to be created (See Table 2). For space we do not include these here but the plots are available on the code repository associated with this project. The development of a two tier model follows practices well established elsewhere in the geodemographic literature [19, 22].

Table 3 The Distribution of Block Groups by their Group and Type Assignments within the US

Group	*Type*	*N*	*%*
A	A1	5489	2.3
	A2	8212	3.5
	A3	5237	2.2
	A4	4826	2.0
	A5	7455	3.1
B	B1	4119	1.7
	B2	3001	1.3
	B3	1743	0.7
	B4	2259	1.0
	B5	4014	1.7
C	C1	9440	4.0
	C2	8086	3.4
	C3	5452	2.3
	C4	6159	2.6
	C5	7177	3.0
	C6	4422	1.9
	C7	9605	4.1
D	D1	9304	3.9
	D2	8122	3.4
	D3	8484	3.6
	D4	8473	3.6
	D5	5538	2.3
E	E1	8446	3.6
	E2	9036	3.8
	E3	7911	3.3
	E4	3907	1.7
	E5	9738	4.1
F	F1	4503	1.9
	F2	8071	3.4
	F3	4564	1.9
	F4	6472	2.7
	F5	6269	2.6
	F6	4596	1.9
G	G1	7313	3.1
	G2	1914	0.8
	G3	3510	1.5
	G4	4691	2.0
	G5	4832	2.0
	G6	4385	1.9

5 Clustering results

The clustering process returns a typology for the US, with the assignment of each area (Census Block Group) into a hierarchical and nested cluster. The initial and more aggregate 7 cluster partitioning will be referred to henceforth as a Group, and labelled sequentially A-G; and the 39 sub-clusters that were created through further splitting of each Group will be referred to as Type, and labeled numerically and sequentially within each Group. The national distribution of Block Group assignments by the created typology are shown in Table 3, with assignments ranging from 0.7% to 4.1% of the total frequency of Block Groups.

To begin to understand the geography of the clusters, maps were generated for the entirety of the US, with illustrative examples of the Groups presented for a number of US cities and regions in Fig. 3. The Groups have all been assigned a separate color, and those areas in light grey were areas not included within the cluster analysis, as a result of

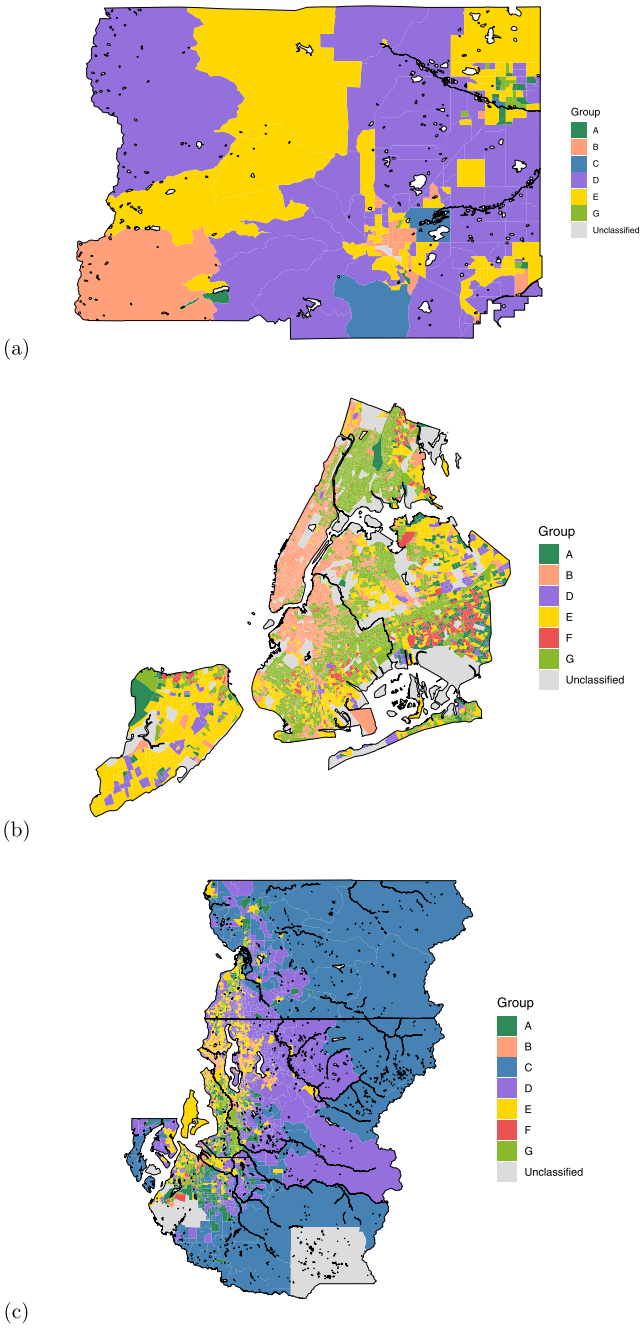

Figure 3 Group level maps of three selected areas of the US (a) Boulder County, CO (b) The 5 Boroughs of NYC, NY (c) Seattle and Surrounding Counties, WA

small population counts or a lack of comprehensive measures for all of the input variables. The US maps and these illustrations show clear patterns of spatial autocorrelation, where group values correspond to the values of their neighbouring areas. Furthermore, through the variegation of the clusters across these different areas, it can be seen how they are effective at picking out differences in the underlying distribution of populations and those places in which they live. Boulder is a small county in CO with compact form and has

quite different Group distribution than New York or the densest area of Seattle (excluding its wider hinterlands). New York, which is predominantly all urban features no Block Groups within Group C, which is found in lower density areas; and likewise, a very urban Group F is not found within Boulder. Mapping such differences aim to give reassurance that broadly expected geographic patterns have emerged; however, further descriptive detail is required to develop a fuller understanding of the Group and Type characteristics representing these spatial patterns.

6 Large language models for cluster descriptions

When building a geodemographic classification, their utility as descriptors of different neighbourhoods is linked to their ability to provide concise, faithful and usable descriptions of places. Key to the utility of Geodemographic classifications is the creation of an interface for the clusters. The most common way to do this is to create short textual labels for each cluster supported by longer "pen portrait" descriptions. Historically, this has been a manual process, with scores showing the relative distribution of variables by cluster compiled into a table (which are often but not exclusively the input variables to the classification), then the classification builder or team will interpret these scores and qualitatively develop labels and descriptions. These will likely go through a range of reviews with stakeholders. There is significant creative skill to the generation of such content, ensuring that it is effectively pitched to an audience, factually accurate, and isn't inflammatory or discriminatory in any way. The process is very time consuming.

Large Language Models (LLM) such as GPT4 have the ability to both interpret data and develop articulate descriptions of data sets, and we argue are useful to support the process of cluster naming and development of pen portraits. We develop a "human in the loop" solution where GPT4 was used to generate a series of thematic descriptions for each cluster using a pattern called retrieval augmented generation or "RAG". In retrieval augmented generation authoritative data or documents are presented to a Large Language Model to ensure its results are informed by the best available data. However, RAG is not infalliable so we manually validate and review all machine-generated content. These reviews initially were used to refine prompts and later used to finalise pen-portraits and labels.

The process involved creating hundreds of prompts, each containing text and data for one aspect of the cluster. Variables were renamed to simplify Census Bureau description into a human readable form, and inference on these prompts was run using the OpenAI GPT4 API. The responses were stored and manually reviewed. Manual checks of the accuracy of the results relative to the input data were completed to explore if there is any problematic supposition or spurious content and, finally, if there are other more general quality issues such as repetition or poor sentence construction. We would always adopt such checks from both an ethical and practical stance when the geodemographic outputs and created labels/descriptions may have real-world applications or impacts.

We developed standardized index scores for the classification input variables and a prompt that comprises of a number of sections. The prompt first introduces the task to create descriptions using the provided data, and then provides examples of how the data table is structured. Finally, the correct use of index scores are explained, and how they should be interpreted.

The following was therefore common across all prompts:

A geodemographics company is trying to explain the characteristics of a neighborhood to a new customer. They present data comparing this neighborhood to the national average.

A score of 100 means the neighborhood is equivalent to the national average, a score of 150 means the neighborhood one and a half times the national average, a score of 200 means the neighborhood is twice the national average, a score of 50 means the neighborhood is half of the national average, a score of 300 means the neighborhood is three times the national average. Their neighborhood has the following characteristics, described in #DATA# below. Data are presented for each characteristic followed by a colon, and then a score. The description of the neighborhood should focus on characteristics that have scores which are greater than 120 or less than 80. If 'housing units that are Renter occupied' have scores that are less than 80, then don't describe any details about the rent prices.

The second element of the prompt was to provide the index scores for each of the variables. The variables were each given reasonably descriptive names to provide context for the scores. The following is an example of how a subset of a data table was presented to the prompt:

#DATA# commutes to work using public transit public transportation (excluding taxicab): 681
commute length Less than 5 minutes: 90
commute length 5 to 9 minutes: 106
commute length 10 to 14 minutes: 123
commute length 15 to 19 minutes: 135
commute length 20 to 24 minutes: 153
commute length 25 to 29 minutes: 157
commute length 30 to 34 minutes: 158
commute length 35 to 39 minutes: 160

Finally, an instruction was given to the prompt concerning what the desired output should comprise. This was also common for all prompts assessed and is presented as follows:

In the third person, write a description of the neighborhood in no more than 500 words. Don't mention the specific scores from the #DATA#, and instead use descriptive words to illustrate rates that are above or below the national average.

It was necessary to add prompt context to not cite specific scores, as without this additional instruction the outputs tended to include scores by default which were deemed to be undesirable. As an illustrative example of an output, the following description was created by GPT 4 for Cluster A2.

The neighborhood is a diverse and family-oriented community, with a significant presence of Black or African American residents, as well as a notable mix of individuals of two or more races and three or more races. The area is predominantly composed of households living with biological, adopted, or stepchildren, and a considerable number of households with grandchildren. Marriage is quite common, with a higher-than-average number of married couple families.

The age distribution in the area is fairly average, with a slightly larger population of women aged 55 to 69 years. The neighborhood is also characterized by a high number of seniors who speak Asian and Pacific Island languages.

When it comes to work and commuting, residents of the neighborhood are more likely to work in the same county as their residence and have a higher propensity for traveling to work by car, truck, or van. Many residents have early start times for work, particularly between

midnight and 5:59 a.m. The majority of the working population experiences commuting times of 15 to 29 minutes.

In terms of education, the area has a higher number of residents who have completed some college, less than 1 year, and those with an associate's degree. However, the number of individuals with advanced degrees, such as master's, professional, or doctorate, is lower compared to the national average.

The neighborhood's income distribution is mostly focused around the middle range, with an above-average number of households earning between $60,000 and $124,999 annually. The prevalence of low-income households is lower in the area, and the number of residents with retirement income is higher than the national average.

Housing in the neighborhood is primarily owner-occupied, with a considerably lower number of renter-occupied housing units. The area features a significant amount of detached single-unit buildings, with a substantial prevalence of homes built in the 1980s and 1970s. The number of vacant houses is considerably lower than the national average.

The local workforce is employed in a variety of industries, with a higher-than-average number of men working in transportation and warehousing, utilities, and administrative and support services. Women are more likely to be employed in transportation and warehousing, utilities, finance and insurance, real estate, rental and leasing, and management of companies and enterprises.

Overall, the neighborhood is characterized by its diverse population, family-oriented nature, strong preference for owner-occupied housing, and a workforce that is engaged in various industries, particularly transportation and warehousing, utilities, and administrative and support services.

6.1 Manual review of GPT4 output

Outputs were manually assessed by checking the number of different attributes/facts referenced and establishing the number that were factually inaccurate, those which made spurious supposition, contained repetition or were missing features that should probably have been considered notable. The results from the assessment for each cluster are shown in Table 4.

On average the returned descriptions contained around 25 facts for each cluster, and included 1 factual inaccuracy, very low levels of supposition or repetition and also around 1 missing feature. The quality of the returned written descriptions were overall very accurate, and written in an accessible and clear style. Although each description contained a very small level of error, we would still advocate that these require human intervention to check the results, following similar processes as building traditional geodemographic descriptive profiles [19]. Where factual inaccuracies occurred these were reasonably unpredictable, and did not focus on any particular area of description. At their most extreme, these took the antipode of the desired outcome. For example, a cluster with homogeneous commuting might be described as "commuting patterns in the neighborhood are diverse". Missing features tended to be very minor, and usually missed an element of a range in the data table. For example "with most people traveling 10 to 19 minutes to work", should have been reported as "with most people traveling 10 to 34 minutes to work": thus, although not classed as inaccurate as the statement is true, there is some missing detail. A further issue observable in some descriptions were elements of value judgement or sentiment ascribed to some of the patterns. This mostly occurred in the final paragraph of the

Table 4 Accuracy Checks of GPT 4 Generated Geodemographic Descriptions

Type	*Factual Inaccuracy*	*Supposition*	*Repetition*	*Missing Features*	*Not Notable Feature*	*Attributes N*
A1	1	0	0	1	0	22
A2	0	0	0	0	0	23
A3	2	0	0	1	0	23
A4	1	0	0	0	0	22
A5	1	0	0	1	0	27
B1	1	0	0	2	0	25
B2	0	1	0	0	0	24
B3	0	0	0	1	0	27
B4	1	0	0	0	0	28
B5	2	0	0	1	1	28
C1	0	1	0	0	0	25
C2	1	0	0	2	0	23
C3	2	0	0	1	0	22
C4	1	0	0	0	0	17
C5	0	0	1	2	0	28
C6	2	0	1	0	0	24
C7	0	0	0	1	0	25
D1	4	0	0	1	0	28
D2	0	0	0	1	0	27
D3	1	0	0	0	0	27
D4	0	0	0	1	0	19
D5	1	0	0	2	0	28
E1	0	0	1	0	0	24
E2	0	0	0	2	0	19
E3	0	0	0	0	0	28
E4	1	0	0	0	1	28
E5	2	0	1	3	0	26
F1	3	0	0	1	0	23
F2	5	1	1	1	0	27
F3	1	1	0	0	0	22
F4	0	0	0	3	0	23
F5	0	0	1	1	0	19
F6	1	0	0	1	0	22
G1	1	0	0	2	0	28
G2	1	0	0	1	0	27
G3	1	0	1	1	0	24
G4	1	0	0	2	0	27
G5	2	0	0	0	0	25
G6	1	0	0	0	0	26

profile description (as illustrated in the presented example), which GPT 4 tends to produce on longer outputs as a summary of the text generated. We would argue such phrases are semantically problematic as imply a degree of ecological fallacy; that is area characteristics being erroneously applied to individuals, and acutely in this case where drivers of behaviour or preference are not known. Prospectively, such issues can likely be addressed through additional refinement to the prompt engineering technique.

After the returned descriptions for Types were manually cleaned to remove inaccuracies, a second application of GPT 4 was used to generate summarising names for each cluster. The following prompt was used, followed by each Type description. In this instance, we manually ran this a number of times to develop a range of labels, selecting what was most acceptable. A qualitative evaluation of outputs was made to ensure these represented the salient characteristics of the cluster.

In the style of a commercial geodemographic classification; create a cluster name that would summarise the following geodemographic pen portrait. The name should capture

as many different characteristics contained within the description as possible. The cluster name should be no more than 3 words.

GPT 4 was then used to generate group-level descriptive summaries from the associated Type descriptions. The following prompt was used, followed by the descriptions for all the Types within the Group. The prompt also details how the Type descriptions will be presented to GPT 4.

The following text represent a series of cluster descriptions from a geodemographic classification. Create an aggregate description that summarises the key characteristics of all the clusters. This should describe the age profile, racial composition and language diversity, types of housing, commuting characteristics and modes of transport, and the types of employment.

This should be no more than 300 words. Each cluster description starts with a name, followed by the description. Each cluster description is separated by "##########".

Finally, once the Group descriptions were created and checked, the previously specified prompt that was used to generate names for the Types was then applied to each of Group descriptions. The final Group and Type hierarchy is presented Table 5.

7 Classification evaluation

Geodemographic classification systems require two types of evaluation to ensure their utility. The first is a check of the internal consistency, designed to make sure that the clusters are homogeneous, and have similar amounts of internal variability [4, 19, 23]. This provides potential users with insights into the levels of certainty associated with the assignment of Block Groups to specific Groups or Types, as well as identifying any potential geographical biases.

The second type of evaluation establishes the utility of the classification for describing social and economic phenomena in the real world. This type of evaluation is designed to test the ability of the resulting geodemographic classification to provide rich multivariate insights into some real world event, policy, or action.

7.1 Internal evaluation

A useful measure of cluster fit is the within sum of square distance between the input measures of each Block Group to the values of their assigned cluster centroid. Larger scores indicate areas that fit less well, whereas smaller scores represent areas with input profiles that are closer to the cluster average. These scores were calculated for all input areas and are presented by Groups and Types in Fig. 4 and also mapped to explore their geography in Fig. 5.

Some of the lowest error rates are found within Group C: Aging Agricultural Settlers, pertaining to areas that are predominantly more rural and have more stable aging populations. An interesting exception to this is Type C6: Mature Rural Dwellers, which has a tendency to be found in areas that are particularly remote and low density, and the error may be a result of wider margins of error associated with the ACS in these locations. Group D: Diverse Affluent Professionals have relatively low error that is reasonably uniform over the Types. These are mostly family areas that might be considered more stable, exhibiting features such as high income, larger housing and a well-educated and professional demographic. Group F: Economic Adversity has more mixed error across the nested Types, acutely in those areas with greater over representation of non-white racial groups.

Table 5 Group and Type Labels

Groups		Types
A: Commuting Families	A1	Multicultural Commuter Communities
	A2	Diverse Family Commuters
	A3	Diverse Early-Riser Families
	A4	Multicultural Sunrise Laborers
	A5	Senior Suburban Stability
B: Young Professionals	B1	Multicultural Urban Professionals
	B2	Diverse Multilingual Commuters
	B3	Youthful Academic Diversity
	B4	Young Urbanites
	B5	Multicultural Urban Achievers
C: Aging Agricultural Settlers	C1	Elderly Agricultural Enclave
	C2	Aging Rural Haven
	C3	Diverse Elders' Enclave
	C4	Agricultural Employment Hub
	C5	Modest Agricultural Settlers
	C6	Mature Rural Dwellers
	C7	Mature, Rural Homesteaders
D: Prosperous Professionals	D1	Affluent Mature Families
	D2	Diverse Family Professionals
	D3	Diverse Aging Affluents
	D4	Prosperous Multicultural Professionals
	D5	Multilingual Professional Community
E: Culturally Rich Achievers	E1	Affluent Educated Elders
	E2	Diverse Professional Enclave
	E3	Educated Commuting Professionals
	E4	Affluent Elderly Enclave
	E5	Mature Multicultural Settlement
F: Economic Adversity	F1	Lower-Income Agricultural Community
	F2	Young Multilingual Commuters
	F3	Lower-Income Ethnic Melting Pot
	F4	Diverse Cultural Mosaic
	F5	Struggling Service Workers
	F6	Urban Economic Strugglers
G: Urban Melting Pot	G1	Diverse Young Renters
	G2	Diverse Commuter Community
	G3	Low-Income Multilingual Renters
	G4	Economically Challenged Commuters
	G5	Multigenerational Latino Enclaves
	G6	Diverse Low-Income Community

The error map shown in Fig. 5 does not highlight any particular geographic areas of specific concern, but does reiterate that in more isolated areas (visible in large block group zones), there is a tendency for greater error, which as discussed, is picked up in some of the Type level patterns. The purpose of this internal validation is to provide additional guidance on the salience of the patterns and where (Geography, Groups or Types) robustness of applications might be given extra scrutiny.

7.2 External evaluation: evictions in the United States

The 2016 book *Evicted* by Matthew Desmond is an ethnographic look at eviction in Milwaukee, Wisconsin. The book shed light on an understudied aspect of the housing crisis in America and led to the creation of the Eviction Data Lab at Princeton University. This lab tracks eviction filings at the block group level for much of the United States. Eviction is a complex phenomenon rooted in poverty, race, built environment, housing markets, family structure, and more. Due to its multidimensional nature, it is an interesting way to

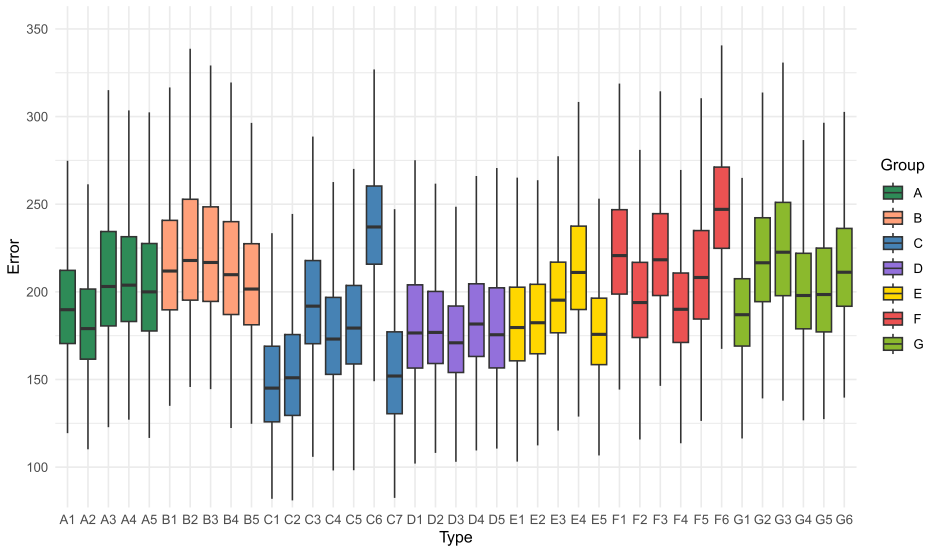

Figure 4 A Box and Whisker Plot of the Error Metric by Types and Groups

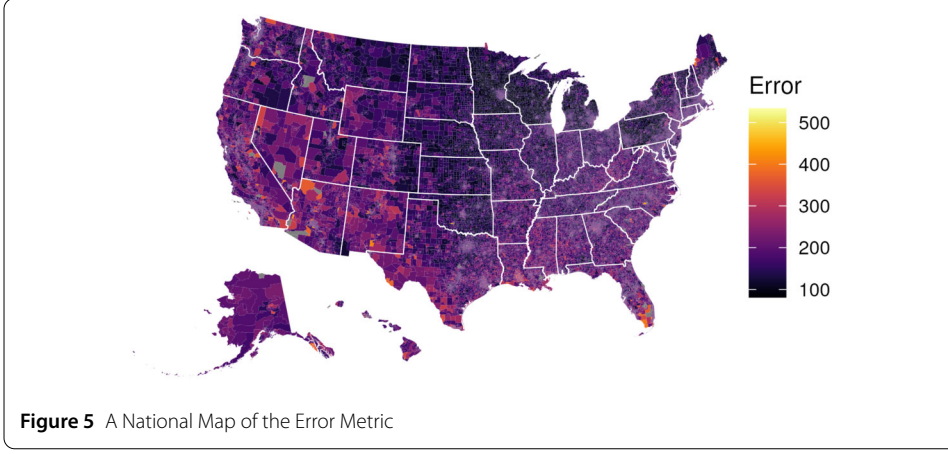

Figure 5 A National Map of the Error Metric

demonstrate the utility of a high-dimensional typology of neighbourhoods. The complexity of eviction and the imprecision of data from the American Community Survey at the block-group level make a variable-by-variable approach to eviction untenable. A holistic understanding of patterns in eviction filings can be difficult, a variable-by-variable analysis would fail to capture interactions, and a regression model would be unwieldy (due to those same interactions). This makes the geodemographic approach ideally suited to understanding these patterns.

Although national patterns clearly show that some types of neighbourhood have a much higher eviction rate than others 6, geodemographic classifications also allow the disaggregation of trends. For example, Fig. 7 shows trends in evictions over time for clusters with at least 5000 evictions between 2000 and 2018. The group "Diverse Commuter Community" is a highly racially diverse group, with 10x the national average public transit ridership long commutes, and many pre-war large (10 or more unit) multifamily buildings. This particular configuration of the built environment only really exists in larger, older American cities. The population has far fewer people with college and advanced degrees than

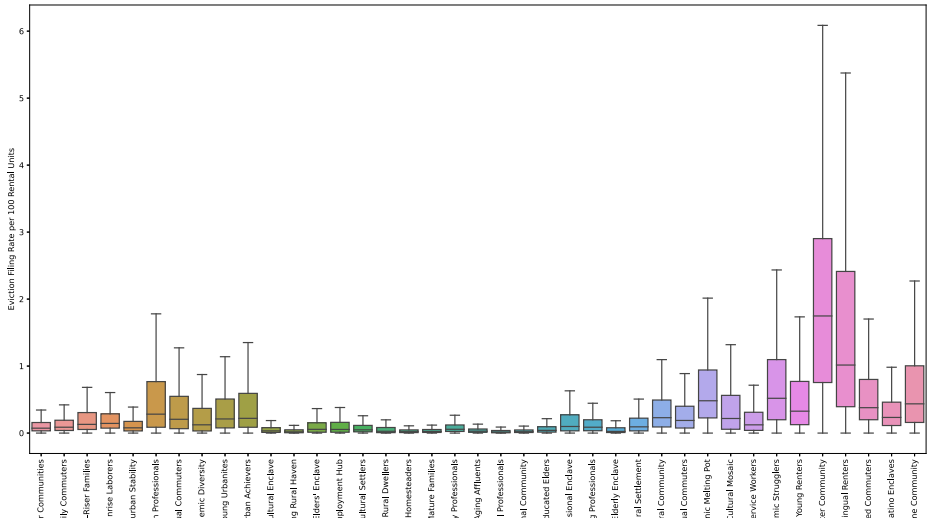

Figure 6 Eviction filing rate (evictions filed per 100 rental units) by cluster type

the national average. However, in New York another type of block group with high eviction rates is "Multicultural Urban Achievers", in that it has very high levels of educational attainment, and many more residents have advanced degrees than the national average. While both of these types of block group are diverse urban areas with an older dense housing stock and their built environment might look similar to the casual observer, they have very different populations.

Validating any geodemographic classification is difficult, but we would argue that our model's ability to separate classes of block groups with different eviction rates, in a way that provides insights is a useful validation of utility. There are of course some caveats, while eviction happens to people and families, our data describe places (block groups), and the characteristics of the block group might not apply to the people being evicted. Even if the classification only helps researchers ask better questions, we think it has value.

8 Discussion and conclusion

In this research, we addressed the challenges posed by the geographic-demographic resolution trade-off in the United States by creating a new way of working with the American Community Survey (ACS). Recognizing the limitations of existing geodemographic systems, we introduced an AI-based, open-source geodemographic classification system for the United States, leveraging small area estimates from the ACS. Our methodology employed a partitioning clustering algorithm, considering a range of socio-economic, demographic, and built environment variables.

A significant innovation in our approach was the integration of AI through the GPT4 large language model and retrieval augmented generation. This integration facilitated the automated generation of intuitive cluster descriptions and names, marking the first application of natural language processing in academic geodemographic research. However, we do not argue that such innovations should be implemented uncritically. The integration of AI into geodemographic research workflows in the way that is presented in this paper necessitates a comprehensive framework that not only leverages the efficiency and capabilities of the AI, but also safeguards end users against potential biases and ethical

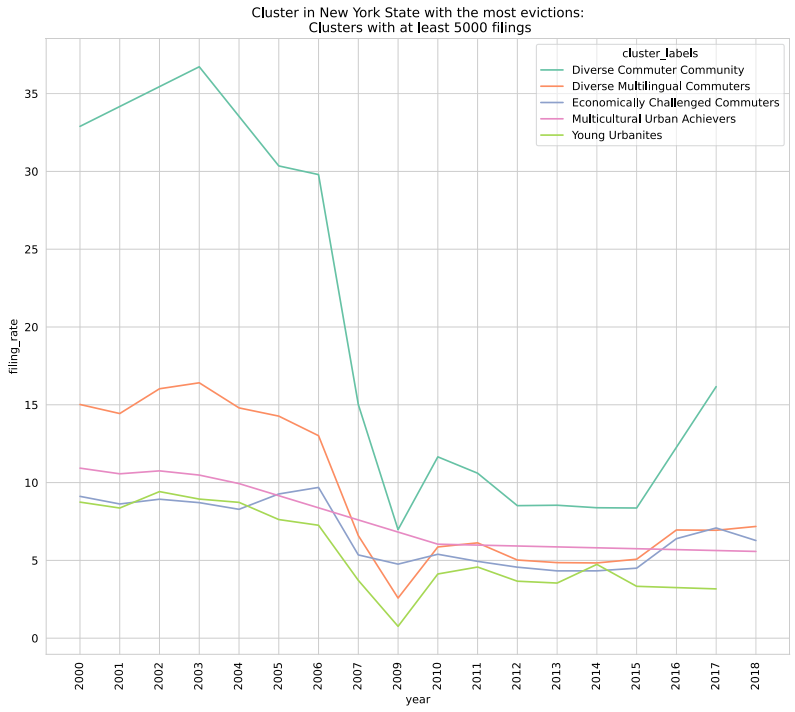

Figure 7 Eviction filing rate by year for New York State. Clusters with more than 5000 evictions filed from 2000-2018

issues. As we argued in this paper, having a "human in the loop" as part of the process of output creation and quality evaluation is critically important, and there are some interesting prospective research challenges around how such processes could be better supported through automation, both in terms of human computer interaction evaluation related to prospective interfaces to such models, and how the manual validation of GPT4 outputs as presented here might be supported through greater automation. Furthermore, the wider the operational use of any classification with generative AI outputs, the greater the potential for any inherent bias or negative effects to be realised. As with any geodemographic classification that has expected wide use, we would advocate that stakeholders are consulted as an integral part of the design, development and evaluation process. We would argue that this is good practice in general, and has been exemplified in the past three editions of the UK ONS Output Area Classification that have been compiled after each Census from 2001-2021 [19, 20, 22]. Elsewhere we also have argued for the development of public interfaces to geodemographic classifications enabling wider consultation and feedback to be gathered [24].

Our classification organized the US into 7 distinct Groups, further divided into 39 Types, each accompanied by descriptive pen portraits. This structure provides a comprehensive and nuanced understanding of the diverse geodemographic landscape of the US. Validation of our classification system was a crucial step in ensuring its reliability and applicability. Through internal and external evaluation we ensured that our geodemographic clusters accurately represent the underlying data and could be effectively used in real-world applications through a study of national eviction rates.

This paper contributes significantly to the field of geodemographics by offering a transparent, reproducible and an adaptable classification system; with the integration and evaluation of advanced AI models like GPT4 demonstrating applicability for future geodemographic research.

Abbreviations
ACS,American Community Survey; API,Application Programming Interface; GPT 4,Generative Pre-trained Transformer 4; LLM,Large Language Model; WCSS,Within Cluster Sum of Squares.

Author contributions
Authors were equal participants in the conception, methodological development, implementation and coding, analysis of results and the preparation of the manuscript. All authors read and approved the final manuscript.

Funding
This work was funded by ESRC grant ES/L011840/1.

Data availability
The geodemographic segmentation of American Block Groups described in this paper is available from the authors or at https://github.com/alexsingleton/US_Geodemographic. All data used in this study are open and available from the US Census Bureau (https://www.census.gov/data.html). This paper accessed these data through: https://walker-data.com/tidycensus/index.html which links to the US Census Bureau API.

Code availability
Code for this study is available on Github: https://github.com/alexsingleton/US_Geodemographic.

Declarations

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

Author details
[1]Geographic Data Science Lab, University of Liverpool, Roxby Building, 74 Bedford St South, L69 7ZT, Liverpool, UK.
[2]Microsoft, 1650 Canyon Blvd., CO 80302, Boulder, USA.

Received: 27 October 2023 Accepted: 18 March 2024 Published online: 22 April 2024

References

1. Spielman SE, Folch DC (2015) Reducing uncertainty in the American community survey through data-driven regionalization. PLoS ONE 10(2):0115626. https://doi.org/10.1371/journal.pone.0115626
2. Bureau UC (2018) Block groups for the 2020 census-proposed criteria. https://www.federalregister.gov/documents/2018/02/15/2018-02624/block-groups-for-the-2020-census-proposed-criteria. Accessed 2023-10-27
3. Abbott A (1997) Of time and space: the contemporary relevance of the Chicago school. Soc Forces 75(4):1149–1182
4. Spielman SE, Singleton A (2015) Studying neighborhoods using uncertain data from the American community survey: a contextual approach. Ann Assoc Am Geogr 105(5):1003–1025. https://doi.org/10.1080/00045608.2015.1052335
5. Su T, Sun M, Fan Z, Noyman A, Pentland A, Moro E (2022) Rhythm of the streets: a street classification framework based on street activity patterns. EPJ Data Sci 11:43. https://doi.org/10.1140/epjds/s13688-022-00355-5
6. Zignani M, Quadri C, Gaito S, Rossi GP (2019) Urban groups: behavior and dynamics of social groups in urban space. EPJ Data Sci 8:8. https://doi.org/10.1140/epjds/s13688-019-0187-7
7. Webber R, Burrows R (2018) The predictive postcode: the geodemographic classification of British society, 1st edn. SAGE, Los Angeles. OCLC: 1028809809
8. Harris R, Sleight P, Webber R (2005) Geodemographics, GIS and neighbourhood targeting. Mastering GIS. Wiley, West Sussex
9. Clark S, Lomax N, Morris MA (2017) Classification of westminster parliamentary constituencies using e-petition data. EPJ Data Sci 6:16. https://doi.org/10.1140/epjds/s13688-017-0113-9
10. Reibel M (2011) Classification approaches in neighborhood research: introduction and review. Urban Geogr 32(3):305–316. https://doi.org/10.2747/0272-3638.32.3.305
11. Batey P, Brown P, Pemberton S (2008) Methods for the spatial targeting of urban policy in the UK: a comparative analysis. Appl Spat Anal Policy 1(2):117–132. https://doi.org/10.1007/s12061-008-9007-3
12. Weiss MJ (2000) The clustered world: how we live, what we buy, and what it all means about who we are. Little, Brown and Company, Boston
13. Singleton AD, Spielman S, Brunsdon C (2016) Establishing a framework for open geographic information science. Int J Geogr Inf Sci 30(8):1507–1521. https://doi.org/10.1080/13658816.2015.1137579
14. Palm R, Caruso D (1972) Factor labelling in factorial ecology. Ann Assoc Am Geogr 62(1):122–133
15. Liu Y, Singleton A, Arribas-Bel D (2019) A Principal Component Analysis (PCA)-based framework for automated variable selection in geodemographic classification. Geo-Spat Inf Sci 22(4):251–264. https://doi.org/10.1080/10095020.2019.1621549

16. Walker K, Herman M (2023) Tidycensus: load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 1.3.2. https://walker-data.com/tidycensus/
17. Singleton AD, Spielman SE (2014) The past, present, and future of geodemographic research in the United States and United Kingdom. Prof Geogr 66(4):558–567. https://doi.org/10.1080/00330124.2013.848764
18. Spielman SE, Thill J-C (2008) Social area analysis, data mining, and GIS. Comput Environ Urban Syst 32(2):110–122. https://doi.org/10.1016/j.compenvurbsys.2007.11.004
19. Gale CG, Singleton AD, Bates AG, Longley PA (2016) Creating the 2011 area classification for output areas (2011 OAC). J Spat Inf Sci 12(2016):1–27. https://doi.org/10.5311/JOSIS.2016.12.232
20. Vickers D, Rees P (2007) Creating the UK national statistics 2001 output area classification. J R Stat Soc, Ser A, Stat Soc 170(2):379
21. Yazgi Walsh B, Brunsdon C, Charlton M (2021) Open geodemographics: classification of small areas, Ireland 2016. Appl Spat Anal Policy 14(1):51–79. https://doi.org/10.1007/s12061-020-09343-6
22. Wyszomierski J, Longley PA, Singleton A, Gale C, O'Brien O (2023) A neighbourhood output area classification from the 2021 and 2022 UK censuses. Geogr J (in press). eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/geoj.12550. Accessed 2023-11-22. https://doi.org/10.1111/geoj.12550
23. Singleton AD, Longley P (2015) The internal structure of Greater London: a comparison of national and regional geodemographic models. Geo: Geogr Environ 2(1):69–87. https://doi.org/10.1002/geo2.7
24. Longley PA, Singleton AD (2009) Classification through consultation: public views of the geography of the E-society. Int J Geogr Inf Sci 23(6):737–763

Publisher's Note