SCHOOL OF GEOGRAPHY UNIVERSITY OF LEEDS

UNIVERSITY OF LEEDS

Geodemographic Classification of Health Determinants to Analyse Health Inequalities in Greater Manchester

Full Name: Kang, Zuwei
Student ID: 201902538
Date of Submission: August 21, 2025
Word Count: 13069
Character Count: 9586

A dissertation submitted in partial fulfilment of the requirements of the Masters Degree in Geographic Information Systems of the University of Leeds

Abstract

Health inequalities remain a pressing challenge, shaped by the spatial interaction of socio-economic, environmental, and healthcare factors. Understanding these disparities requires approaches that can integrate multidimensional health determinants into spatially interpretable units. This study applies a geodemographic classification framework to examine health inequalities at the LSOA level in Greater Manchester. Using socio-economic, housing, and environmental variables, K-means clustering was employed to classify neighbourhoods into distinct health-related area types. The optimal four-cluster solution reveals clear spatial heterogeneity: clusters differ significantly in socio-economic characteristics, environmental exposures, and accessibility to health services, with disadvantaged communities facing compounded health risks. Validation using health status and disability prevalence indicators further confirms the robustness of the classification. The findings demonstrate that geodemographic classification can effectively uncover fine-scale spatial health inequalities and provide evidence to inform place-based public health interventions. Methodologically, this study illustrates the integration of clustering techniques in health geography, while practically, it offers insights to guide policy-making and resource allocation in urban health planning.

Acknowledgements

I am deeply grateful to my supervisor, Dr. Paul Norman, for his invaluable guidance and support throughout this dissertation. I would also like to thank the University of Leeds Library for providing a productive study environment, and my classmates and friends for their encouragement during this journey. Most importantly, I am sincerely thankful to my family, whose love, understanding, and sacrifices have given me the opportunity to pursue this degree.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Research Background and Context

Health inequality, as one of the most prominent public health challenges globally, has attracted increasing attention. Marmot et al. (2020) pointed out that in England, although health equity has been promoted through policies for over a decade, the health gap between regions and social groups has not narrowed. In some socioeconomically disadvantaged communities, it has even widened. This highlights the urgency and practical significance of conducting continuous and in-depth research on health inequality.

It is worth noting that health inequality is not accidental; it stems not only from individual differences, but more deeply from the unequal positions different groups occupy within the social structure (WHO, 2008). The formation of health inequalities is the result of a complex interplay of various Determinants of Health (DoH)—including socioeconomic status (SES), environmental exposures (such as air pollution and green space accessibility), housing conditions and healthcare accessibility (Rad, 2025). These variables are also spatially correlated, creating regional health advantages or vulnerabilities. Therefore, identifying the combinations and distributions of these structural factors across different areas from a spatial perspective is of great theoretical and practical value for promoting more targeted and equitable health policies.

In recent years, geographic health research has increasingly adopted area classification methods to capture this complexity. Among them, geodemographic classification, especially unsupervised clustering techniques such as k-means, has become an important tool for revealing the spatial structures of health determinants. This method can uncover hidden spatial patterns that traditional deprivation indices may overlook, and supports place-based health interventions (Rizvi et al., 2021). Therefore, geodemographics holds inherent advantages when applied in health research, providing a robust framework for informing targeted policy-making and resource allocation.

This study focuses on Greater Manchester, a metropolitan county in northern England where health inequalities are particularly pronounced. Using the Lower Layer Super Output Area (LSOA) as the unit of analysis, this research integrates multiple health determinants to develop a geodemographic classification model, identifying distinct types of health-related population groups across the region. The study aims to investigate the spatial interplay between social deprivation and environmental exposure within Greater Manchester, providing both theoretical insights and practical guidance for spatial governance and targeted interventions aimed at reducing health inequalities.

1.2 Research Problem and Significance

Health inequality is a significant challenge in public health, particularly pronounced in urban areas. Greater Manchester, as a major metropolitan county in northern England, exhibits substantial disparities in health outcomes across communities, closely linked to socioeconomic status and environmental exposures. Although previous studies have examined these influencing factors, there remains a lack of systematic analysis on the spatial interplay of multiple health determinants, which constrains the development and implementation of targeted intervention strategies.

Current composite indices such as the Index of Multiple Deprivation (IMD) provide concise and effective measures of area-level socioeconomic deprivation (Noble et al., 2006). However, these indices primarily focus on socioeconomic variables such as income, unemployment, and education, often overlooking the role of environmental exposures like air pollution and green space coverage (Briggs, 2005; Bloemsma et al.,2019). This limitation reduces the ability of single deprivation indices to comprehensively capture compound health risks arising from the interaction of social and environmental factors.

In practical public health governance and policy-making, decision-makers face core questions: “Which areas face the most severe health risks?” and “What combination of factors drive these risks?” Therefore, there is a need for a spatial classification approach that integrates multidimensional health determinants and reflects the composite characteristics of areas, to support more precise interventions and resource allocation. (Kindig and Stoddart, 2003; Diez Roux and Mair, 2010)

Accordingly, this study aims to develop a geodemographic classification model that integrates socioeconomic and environmental factors to identify population groups with distinct health characteristics within Greater Manchester, and to reveal the spatial relationships between these groups and health outcomes. Unlike traditional research focusing solely on “the most deprived areas,” this study explores the diversity of health profiles across communities at a fine spatial scale (Lower Layer Super Output Areas, LSOAs), providing new insights into the spatial coupling of health determinants.

The significance of this research lies in proposing a systematic health geography classification method to assist urban managers and public health policymakers in moving beyond single deprivation indices, enabling the design of more precise and differentiated health promotion strategies based on “population types.” The spatial classification system constructed by this study will offer more detailed zoning and resource allocation references for urban health governance, enrich the spatial dimension perspective in health inequality research, and advance health policy from “intervention targeting high-risk areas” toward “differentiated interventions targeting diverse community types.”

1.3 Research Objectives and Questions

This study aims to identify and analyze different types of health-related populations in Greater Manchester, UK, using geodemographic classification based on multiple health determinants. It seeks to explore the spatial coupling of social deprivation and environmental exposure, and to evaluate how the classification results relate to health outcomes and the accessibility of healthcare resources.

Objectives:

- To apply unsupervised clustering algorithms such as k-means for geodemographic classification to identify types of health-related groups.
- To analyze the spatial distribution patterns of different group types in terms of socioeconomic deprivation and environmental exposure.
- To evaluate the relationships between classification results, health outcomes, and healthcare accessibility.
- To investigate the spatial structure underlying regional health inequalities and provide data support for policy-making.

Research Questions:

- How are health determinants spatially distributed across different communities in Manchester?
- Can an effective population classification be constructed based on multidimensional health determinants?
- What spatial relationships exist between different population types and healthcare accessibility, and health outcomes?

This study aims to provide a more detailed and precise spatial classification tool for urban public health management, supporting differentiated health interventions and resource allocation.

1.4 Study Area Description

Greater Manchester is a metropolitan county located in the North West of England, encompassing ten local authorities: Manchester, Salford, Bolton, Bury, Oldham, Rochdale, Stockport, Tameside, Trafford, and Wigan. The geographical extent and administrative boundaries of Greater Manchester are illustrated in **Figure 1. 1**.Covering an area of approximately 1,276 square kilometers, it is home to a population of around 2.8 million people, making it one of the largest and most densely populated urban regions in the UK. The region has a diverse socio-economic landscape, with affluent suburbs existing alongside some of the most deprived neighborhoods in England (Lake et al., 2023).

Historically, Greater Manchester has been a hub of industrial activity, which has left a lasting legacy on its urban environment and public health. The area exhibits significant disparities in health outcomes, with life expectancy and prevalence of chronic diseases, varying widely across different communities (Purdam, 2017; Richardson, 2022). These disparities are closely linked to spatial patterns of social deprivation, environmental exposures such as air pollution, and access to green spaces and healthcare services.

The spatial heterogeneity of health determinants in Greater Manchester provides an ideal setting for examining the complex interplay between social and environmental factors influencing population health. By using the Lower Layer Super Output Area (LSOA) as the unit of analysis, this study will capture fine-scale spatial variations, allowing for a nuanced understanding of local health vulnerabilities. The findings are expected to inform targeted interventions and support evidence-based public health planning at both local and regional levels.

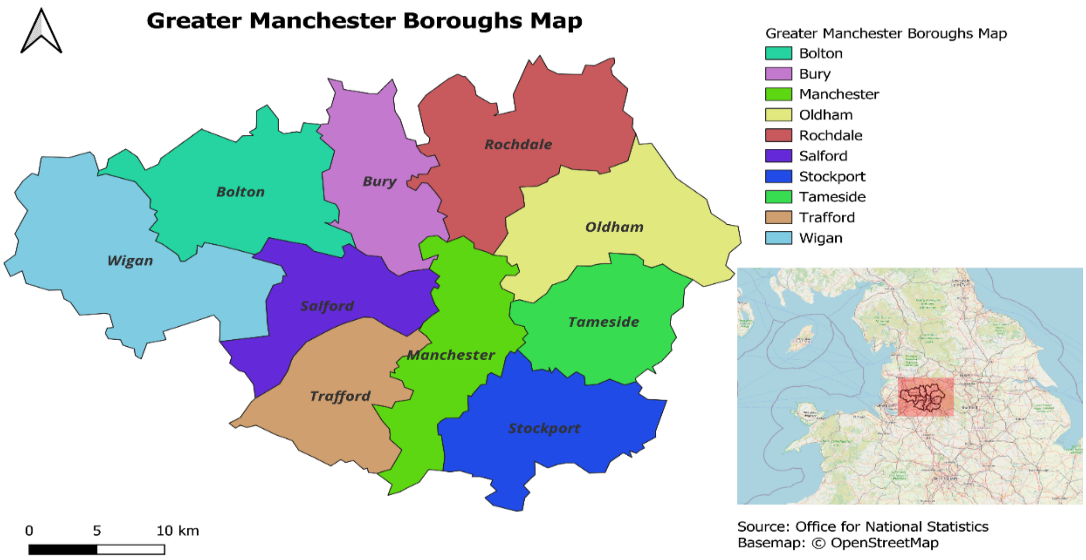

Figure 1. 1 Administrative boundaries of Greater Manchester (LAD level)

1.5 Dissertation Structure Overview

The thesis is structured into six chapters. **Chapter 1** introduces the research background, objectives, and significance, and provides an overview of the selected study area—the Greater Manchester region—and its health inequalities. **Chapter 2** reviews relevant literature, covering the application of Geographic Information Systems (GIS) in health research, studies on health inequalities, and the use of geodemographic classification in health studies, clarifying the research focus and identifying gaps in existing research. **Chapter 3** outlines the data sources, variable selection, and methodological framework, including k-means clustering and statistical analysis methods. **Chapter 4** presents the results and analysis of the study, including the characterization of geodemographic clusters, the distribution and comparison of health indicators across clusters, evaluation of healthcare accessibility, and findings from regression analyses on socioeconomic and environmental determinants of health. **Chapter 5** discusses the underlying reasons behind the findings and their academic and practical implications. **Chapter 6** concludes the dissertation by summarizing the main findings, emphasizing the health conditions and influencing factors across different clusters, proposing strategies to address health inequalities, and highlighting research limitations and directions for future studies.

Chapter 2. Literature Review

This chapter reviews the core theories and literature relevant to this study, focusing on four key aspects. First, it outlines the applications and development of Geographic Information Systems (GIS) in health research. Second, it summarizes the major findings and limitations of existing studies on the drivers of health inequalities. Third, it examines the role and advantages of geodemographic methods in revealing health disparities, while also discussing their limitations in terms of spatial scale, variable integration, and methodological application. By systematically synthesizing these strands of literature, this chapter provides a theoretical foundation for constructing a classification model of health determinants at a micro-spatial scale, clarifies the research entry point, and highlights the study’s innovation and practical significance.

2.1 GIS in Health Studies

Geographic Information Systems (GIS), as an integrated technological platform for spatial data collection, management, analysis, and visualization, have transcended their initial cartographic function to become indispensable tools in public health and health research. Early health geography studies primarily utilized GIS to produce disease maps that visually display the spatial distribution of diseases or health indicators, revealing health disparities across different regions (Rushton, 2003). With advancements in data acquisition capabilities and computational technologies, the applications of GIS have significantly expanded, encompassing medical resource allocation assessment, environmental exposure analysis, and health service accessibility evaluation, among others (Wang , 2020). GIS has evolved beyond a mere tool for spatial data presentation into a powerful platform for spatial analysis, capable of supporting in-depth data insights and public health decision-making.

It is worth noting that although GIS has shown great potential in health research, its application still faces many challenges. Firstly, issues related to the quality and timeliness of geographic data may lead to biases in analysis results, especially when spatial units (such as LSOAs) undergo boundary changes across different years. Such changes introduce spatiotemporal inconsistencies, which can affect the accuracy of studies on the spatial distribution of health outcomes (Norman et al., 2024). Additionally, privacy protection and ethical concerns regarding individual health data restrict the implementation of some high-resolution spatial analyses (Boulos et al., 2009). The choice of spatial scales and inconsistencies in data resolution may significantly impact results; these technical challenges remain unresolved, and related debates continue in academic circles (Lee et al., 2014).

Driven by the rapid development of big data and mobile technologies, GIS technology is witnessing new trends, with health applications based on mobile GIS emerging continuously. Particularly, platforms combining augmented reality (AR) technologies offer new technical avenues for disease prevention and control (Bazlan et al., 2020). Such systems enhance the timeliness and precision of public health interventions through real-time data collection and spatial visualization, opening new possibilities for dynamic management of health spatial analysis and health services. However, effectively integrating and managing these multi-source heterogeneous data while ensuring data privacy and security remain critical challenges for GIS applications in the health domain (Boulos et al., 2009).

In summary, GIS, as a core technological platform for health research, not only provides spatial visualization and analysis functions but also offers crucial support for revealing the geographic and socio-environmental mechanisms underlying health inequalities. By continuously improving data quality, enhancing privacy protection mechanisms, and applying emerging mobile and real-time data technologies, GIS will play an increasingly important role in public health policy formulation and health resource optimization, facilitating precision health interventions and equitable resource allocation.

2.2 Review of Existing Studies on the Drivers of Health Inequalities

Health inequalities refer to systematic, avoidable, and unjust differences in health status, disease burden, life expectancy, and access to healthcare across different social groups (Marmot, 2005; Whitehead, 1992). These disparities are not only reflected in individual health outcomes but also reveal deeper structural injustices shaped by the interplay of social, economic, and environmental factors (Braveman et al., 2011).

A large body of reviews and studies has demonstrated that socioeconomic status (SES) is one of the core drivers of health inequality. Adler and Newman (2002), summarizing extensive literature, argued that SES affects health through multiple mechanisms such as income constraints, psychological stress, unhealthy behavioral patterns, and environmental exposures. The WHO’s (2008) global report, integrating cross-national data, revealed that socially disadvantaged groups consistently face greater health risks and unequal access to healthcare. Pickett and Wilkinson (2015), adopting a social-psychological perspective, emphasized that income inequality affects population health through mechanisms such as social comparison and erosion of trust, highlighting that health inequalities stem not only from material deprivation but also from long-term structural stress. Nevertheless, the universality of these mechanisms across different national and cultural contexts remains contested (Coburn, 2015). In addition, Bambra et al. (2010), through a systematic review of interventions addressing broader social determinants, found that measures such as improving housing quality can reduce disease risks and alleviate health inequalities, though the strength of evidence across different interventions varies considerably.

Empirical studies have further reinforced the significant influence of SES and related factors on health. Paul and Moser (2009), through a meta-analysis, showed that unemployment substantially increases the risk of depression and anxiety, underscoring the critical role of employment status in mental health. Montez and Zajacova (2013), using longitudinal data on U.S. white women, demonstrated the widening education gap as a key contributor to rising mortality rates, illustrating the far-reaching impact of education on health outcomes. Braveman et al. (2011), synthesizing multiple socioeconomic indicators, systematically highlighted the associations of income, education, and occupation with health. Mason et al. (2013) examined the relationship between housing affordability and mental health, finding that housing stress negatively affects mental well-being, with particularly pronounced effects among renters. Collectively, these studies underscore the central role of SES and its various manifestations in health inequalities. However, most rely on correlational analyses, and causal mechanisms remain to be fully established, pointing to directions for future research and policy development. It is also noteworthy that socioeconomic factors rarely act in isolation; their actual impact is shaped by the interplay of cultural context, lifestyle, and policy environments. This complexity produces significant variation in health outcomes across groups and regions, reminding us that diverse contextual factors must be accounted for when researching or designing policy interventions.

Beyond socioeconomic factors, environmental exposures constitute another critical dimension of health inequalities. Recent large-scale epidemiological studies have demonstrated that long-term exposure to air pollutants such as NO_2 and $PM_{2.5}$ significantly increases the incidence and mortality of respiratory diseases, cardiovascular diseases, and lung cancer (Chen et al., 2024). Such environmental risks are disproportionately concentrated in low-income and marginalized communities, meaning that health inequalities manifest not only socioeconomically but also spatially in the geographic distribution of risks. By contrast, green spaces are widely recognized as important positive environmental factors that mitigate health inequalities. Urban greenery improves living environments and air quality, while also fostering physical activity, psychological restoration, and community cohesion (Twohig-Bennett and Jones, 2018). However, the distribution of green space is highly unequal. Socioeconomically disadvantaged groups are more likely to live in areas with limited access to greenery, and disparities exist not only in accessibility but also in size, ecological quality, and amenity provision of green spaces (Rigolon, 2016;Zhang and Luo, 2024), further intensifying spatial inequalities in access to health-promoting resources.

Healthcare accessibility has long been considered another key dimension of health inequalities. Studies suggest that factors such as the spatial distribution of healthcare facilities, transportation accessibility, and service availability influence the ease with which individuals access healthcare, thereby affecting health outcomes to some degree (Guagliardo, 2004; Penchansky and Thomas, 1981). However, existing literature also indicates that the association between healthcare accessibility and broader population health outcomes is relatively weak, particularly when sociodemographic factors are taken into account (Alexandrov et al., 2024). This suggests that healthcare accessibility may reflect service availability rather than serving as a direct determinant of overall health levels.

In sum, health inequalities exhibit pronounced spatial characteristics, with intertwined socioeconomic, environmental, and healthcare mechanisms contributing to regional variations in health outcomes (Pearce, 2012). For instance, urban disadvantaged communities often face multiple environmental burdens such as high air pollution, limited green space, and traffic noise (Mueller et al., 2018). These “spatial social disadvantages” not only exacerbate chronic disease prevalence and healthcare access disparities but also contribute to the intergenerational transmission of health inequalities in specific areas (Diez Roux, 2001). Poor communities, due to spatial segregation and environmental degradation, are persistently exposed to unhealthy environments and may also suffer from social stigmatization (the “blemish of place”), further amplifying health disparities (Pearce, 2012). This phenomenon underscores the importance of GIS in providing spatial analytical capabilities for health research and highlights the theoretical and practical significance of geodemographic classification in uncovering health inequalities.

2.3 Geodemographics in Health Research

Geodemographics is a method that clusters geographic units based on multidimensional attributes, such as population, socio-economic, and environmental characteristics, combining “population traits” with “spatial location” to form representative “area types.” By integrating complex datasets into interpretable spatial units, this approach can reveal the interaction between “who people are” and “where they live,” making it increasingly relevant in health research, particularly when analyzing the spatial mechanisms of health inequalities.

Existing studies indicate that geodemographics provides a powerful tool for identifying spatial patterns of health disparities. For example, Ortiz-Prado et al. (2017) examined the spatial distribution of suicide rates in Ecuador under varying social and environmental contexts, while Grekousis et al. (2021) used geodemographic classification to analyze disparities in COVID-19 mortality in the United States, highlighting the spatial coupling of race, economic conditions, and health resources. Beyond traditional geodemographic methods, recent research has incorporated multi-source environmental data into health prediction models. MedGNN (Zhang et al., 2025), for instance, integrates socio-economic, environmental (e.g., green space coverage, air pollution), and geographic information through graph neural networks to spatially predict health outcomes such as antidepressant prescriptions. Although MedGNN is not a classic geodemographic method, its approach of combining multidimensional features with spatial information resonates with geodemographics, providing insights for identifying health-vulnerable areas and conducting multidimensional health impact analyses.

From a methodological perspective, different clustering algorithms exhibit significant variations in application due to their underlying principles and designs. K-means is widely used for its computational efficiency and suitability for large datasets (Jain, 2010), while hierarchical clustering is preferred for its interpretability, especially when exploring hierarchical relationships between groups. Fuzzy clustering and DBSCAN have also been gradually applied to health-related spatial research, though their parameter sensitivity and complexity increase the difficulty of interpreting results (Xu and Wunsch, 2005). The literature suggests that algorithm selection should align with research objectives and data characteristics while considering the stability and reproducibility of outcomes.

However, geodemographics has limitations. First, variable selection is often subjective, with different studies emphasizing different indicators, which may introduce bias and affect the representativeness and applicability of classification results. Second, key parameters in clustering analyses, such as the number of clusters (K), lack standardized criteria and are often determined empirically or heuristically, leading to high variability and challenges in interpretation. Moreover, most studies rely on static data and fail to adequately capture spatiotemporal changes in population structure and environmental factors (Bailey, 2009), limiting the ability to dynamically track health trends.

Nonetheless, the advantages of geodemographics cannot be overlooked. First, this method can effectively reveal spatial heterogeneity (Anselin, 1995), reflecting significant differences in health status and its social determinants across regions, thereby addressing the limitations of macro-level statistics in capturing local health inequalities. Second, geodemographics classifies micro-spatial units into “area types” with similar characteristics, facilitating a deeper understanding of the spatial distribution of health disparities and their socio-economic, environmental, and other relevant driving factors. Moreover, some studies have attempted to apply geodemographic classifications to health service utilization and public health policy planning, promoting spatially targeted interventions (Petersen et al., 2011). As a bridge between statistical analysis and policy-making, it demonstrates considerable potential and room for development.

In summary, geodemographics provides an effective framework for integrating multidimensional factors—spatial, socio-economic, and environmental—in health research, revealing micro-level spatial heterogeneity and supporting evidence-based health equity policies. However, uncertainty in variable selection and parameter settings limits the generalizability of results. Future research that integrates multi-source spatiotemporal data and interdisciplinary theory could further enhance its application in studying health inequalities and informing public health decision-making.

2.4 Research Gaps and Innovations

Although existing studies on health inequalities have achieved substantial results, several limitations and research gaps remain, primarily reflected in the following aspects:

1. **Spatial scale limitations**: Most studies on health inequalities focus on regional or city-level scales, whereas analyses of health determinants at micro-spatial units (e.g., LSOA or community level) are relatively scarce. This limits the ability to fully capture spatial health disparities and their driving factors at small scales.

2. **Insufficient variable integration**: Existing research often focuses on socioeconomic or environmental factors, lacking studies that systematically integrate multidimensional health determinants, such as socioeconomic status, environmental exposures (air pollution, green space), and accessibility to healthcare resources. The limited dimensionality of variables restricts a deeper understanding of the mechanisms underlying health inequalities.

3. **Methodological limitations**: Traditional statistical methods struggle to handle complex, multi-source data and nonlinear relationships. Moreover, applications of geodemographic classification in the health field remain limited, particularly studies combining spatial analysis and multidimensional health variables at micro-spatial scales.

Based on these gaps, the present study offers the following innovations:

1. **Micro-spatial scale analysis**: By constructing a classification model of health determinants at the LSOA level, the study captures health disparities at the community or neighborhood level, providing a basis for fine-grained health interventions.

2. **Integration of multidimensional variables**: Socioeconomic, environmental, and other multidimensional variables are systematically integrated to form a classification framework of health determinants. Combined with health indicators and accessibility to healthcare resources, this enables a multi-angle analysis of the mechanisms underlying health inequalities.

3. **Methodological innovation**: Geodemographic classification is integrated with GIS spatial analysis, using K-means clustering to process multidimensional data. Spatial visualization is then employed to explore the spatial patterns of health inequalities, enhancing both the interpretability and practical value of the model.

4. **Practical significance**: The results can provide micro-spatial references for public health policy, enabling targeted health interventions and optimized resource allocation, while also enriching the research applications of health geography and health social science.

Chapter 3. Methodology and Data

This chapter provides a detailed explanation of the methodological framework and data sources adopted in this study, with the aim of conducting a geodemographic classification of Greater Manchester based on key health determinants. It offers a comprehensive description of the datasets, variable definitions, rationale for selection, research scale, and analytical methods. The structured workflow ensures research reproducibility while facilitating a clear presentation of the application of geospatial and clustering methods.

3.1 Research Framework and Workflow Diagram

This study follows an eight-step research workflow (Figure 3.1), designed to systematically progress from research questions to the interpretation of results:

Define research objectives – Establish the research questions and the goal of spatial health classification.

Variable selection – Select indicators from socio-economic, environmental, and accessibility domains based on the literature.

Data collection – Obtain data from authoritative sources such as ONS and CDRC.

Data preprocessing – Clean, integrate, and standardize data at the LSOA level.

K-means clustering – Classify LSOAs based on standardized health determinant indicators.

Mapping and visualization – Generate spatial maps showing the distribution of clusters.

Statistical testing – Use ANOVA or Kruskal–Wallis tests to assess the statistical significance of differences in health outcomes and resource accessibility across health determinant categories (mapped to geographic areas), and visualize the distribution, median/mean, and outliers of each category using boxplots.

Result interpretation – Derive spatial distribution characteristics and policy implications based on clustering patterns and their relationship with health outcomes and healthcare accessibility.

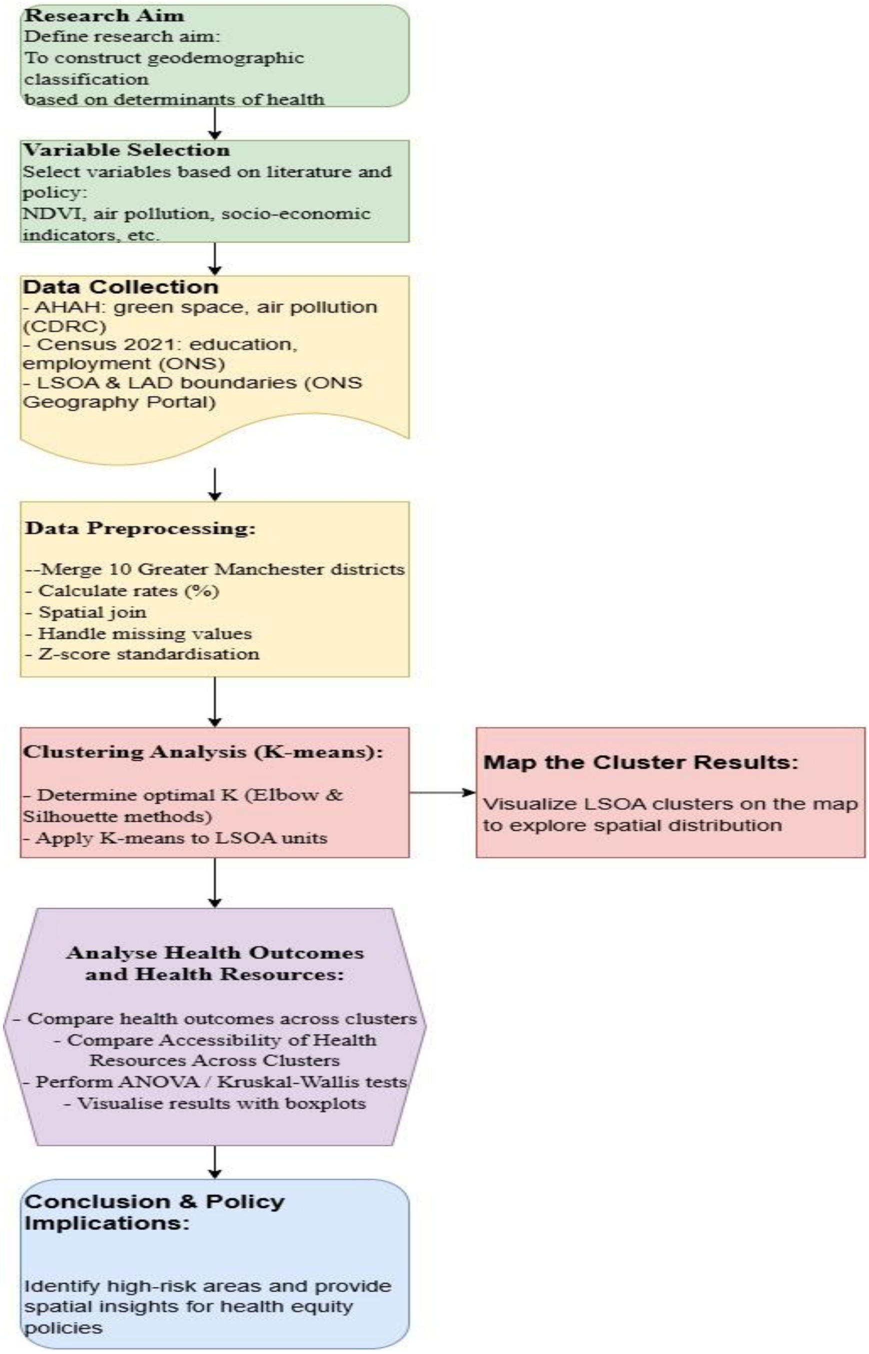

Figure 3. 1 Research framework and workflow diagram.

3.2 Variable Selection and Data Sources

This study selected variables that reflect the major determinants of community health, covering socio-economic, environmental, and natural environment dimensions. These variables were used for the geodemographic classification analysis, while additional indicators of healthcare accessibility and health outcomes were applied to validate the classification results. Table 3. 1 summarizes the selected variables, their definitions, and data sources. The study area is defined at the LSOA (Lower Layer Super Output Area) level in Greater Manchester, UK. This spatial scale captures local variations in social and environmental conditions while ensuring data availability and comparability.

3.2.1 Health Determinant Variables (for Clustering)

The selection of clustering variables was guided by theoretical relevance and empirical evidence, with a focus on factors closely associated with health outcomes. Socio-economic variables were primarily drawn from the 2021 UK Census (ONS, 2021), while environmental and natural environment indicators were obtained from the Access to Healthy Assets and Hazards (AHAH, CDRC) dataset. This dataset integrates multiple sources, including DEFRA air quality data and Sentinel satellite-derived NDVI data.

Socio-economic variables include the unemployment rate, proportion of residents with Level 4+ (higher education) qualifications, proportions of different housing types (detached houses, terraced houses), rental rate, and housing overcrowding rate. These variables reflect economic activity, education level, housing conditions, and living environments. For instance, higher unemployment rates are often associated with an increased risk of mental health problems and chronic diseases (Milner et al., 2014), while education level is strongly linked to health literacy and health behaviors (Cutler and Lleras-Muney, 2006). Housing type and overcrowding serve as proxies for socio-economic status and residential quality, with implications for both physical and mental health (Evans et al., 2003; WHO, 2018). In addition, higher rental rates may indicate increased housing instability, thereby contributing to elevated psychological stress and health risks (Bentley et al., 2011; Padgett, 2020).

The environmental exposure dimension includes variables on air pollution and green space coverage. Specifically, air pollution indicators consist of annual mean concentrations of nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and particulate matter (PM_{10}). Long-term exposure to these pollutants has been widely documented to correlate with cardiovascular diseases, respiratory illnesses, and overall health conditions (Huang et al., 2021; Orellano et al., 2021; Wilkie et al., 2025), thereby reflecting the potential health risks of the living environment. On the other hand, the NDVI vegetation index was used as a proxy for green space coverage, as greener environments are often associated with increased physical activity, improved mental health, and overall better health outcomes (Markevych et al., 2017).

3.2.2 Health Outcomes and Healthcare Accessibility Variables (for Validation)

To validate the clustering results, this study employed health outcome variables from the 2021 Census, including general health status (TS037) and disability prevalence (TS038). Based on the literature review (see earlier section), healthcare accessibility shows a relatively weak association with overall health levels once socio-economic and environmental factors are controlled for. However, its policy significance remains substantial. Therefore, healthcare accessibility was included as a validation variable, using AHAH V4 (CDRC) data, which provides estimated drive times to the nearest GP, dentist, hospital, and pharmacy. These indicators enable the assessment of differences in health status and service accessibility across community types, thereby testing whether the clustering results genuinely capture spatial health disparities of policy and planning relevance.

Table 3. 1 Overview of Variables, Definitions, Sources, and Health Relevance

Variable	Description	Source	Health Relevance
Unemployment rate	Percentage of population unemployed	ONS, 2021	Risks to mental and physical health
Higher education	Percentage of population with Level 4 or higher qualifications	ONS, 2021	Health literacy, chronic disease prevention
Detached/Terraced housing	Proportion of detached or terraced housing	ONS, 2021	Socioeconomic and living environment
Rental housing	Proportion of rented households	ONS, 2021	Housing instability, psychological stress
Overcrowded housing	Proportion of overcrowded households	ONS, 2021	Infectious diseases, mental health
NO_2, SO_2, PM10	Annual mean concentration (μg/m³)	DEFRA / AHAHV4	Respiratory and cardiovascular health
NDVI (green space)	Normalized Difference Vegetation Index	Sentinel / AHAHV4	Physical activity, mental health
Health Outcomes (for validation)	General Health (TS037) Disability (TS038)	ONS Census 2021	Captures self-assessed overall health, widely used in population health surveys. Indicates prevalence of long-term illness or health limitation.
Healthcare Accessibility (for validation)	Travel time to GP, Dentist, Hospital, Pharmacy	AHAHV4 (CDRC)	Measures ease of access to healthcare services;

3.3 Data Pre-processing and Standardisation

All variables were spatially aligned and aggregated at the Lower Layer Super Output Area (LSOA) level to ensure consistency across datasets. This spatial matching allowed for accurate integration of diverse data sources into a unified analytical framework. A total of 1702 LSOAs were included in the analysis. No missing data were detected, so no imputation or record deletion was required

To address potential multicollinearity among variables, Pearson correlation coefficients and Variance Inflation Factor (VIF) analyses were performed. The variable Household Carlessness Rate initially considered as a socioeconomic indicator, showed high collinearity with other socioeconomic variables ($r > 0.85$, $VIF > 10$) and was therefore excluded from subsequent analyses to improve model stability and interpretability.

To eliminate scale differences among variables and to make them comparable for clustering, all continuous variables were standardized using Z-score standardisation. This transformation converts each variable to have a mean of zero and a standard deviation of one, facilitating balanced contributions of each variable in the clustering process. The standardization was performed as:

$$Z = \frac{X - \mu}{\sigma}$$

where X is the original value, μ is the mean, and σ is the standard deviation of the variable.

After pre-processing, all variables were merged into a single feature matrix, where each row represents an individual LSOA and each column corresponds to a standardized variable. This feature matrix serves as the input for the subsequent k-means clustering analysis, providing a consistent and comparable dataset for classifying communities based on health determinants. The workflow is summarized in Figure 3. 2

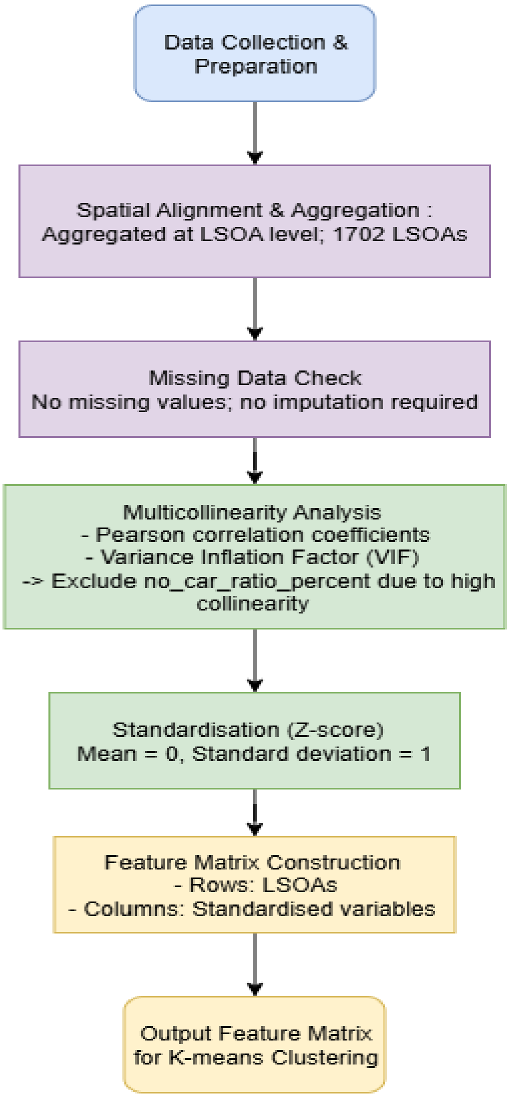

Figure 3. 2 Workflow for data preprocessing and standardization

3.4 Clustering Analysis using K-Means

This section introduces the clustering analysis method based on the K-means algorithm. It first provides an overview of the basic principles of K-means and explains the reasons for selecting this method in the study. It then describes how the optimal number of clusters was determined using both the elbow method and the silhouette coefficient. Finally, the implementation process of clustering is detailed, including the classification of LSOAs, the generation of static maps in QGIS, and the development of an interactive web-based map to enhance spatial visualization and analytical capabilities.

3.4.1 Brief Introduction to K-means Method

K-means clustering is an unsupervised machine learning algorithm that iteratively partitions samples into a predetermined number of non-overlapping clusters, K. In simple terms, it is a method for grouping data into several clusters. The algorithm starts by randomly selecting several “centroids” and then assigns each data point to the cluster whose centroid is closest to it. Next, the centroids are recalculated by computing the mean position of all points within each cluster. This process repeats until the centroids no longer change positions. Through this iterative procedure, the data are divided into clusters such that similarity within each cluster is maximized, while differences between clusters are maximized.

The main reason for selecting the K-means method in this study is that it is a commonly used clustering approach with high computational efficiency, easy implementation, and suitability for handling large-scale geographic units (such as thousands of LSOAs). Additionally, since K-means is sensitive to the scale of variables, all variables were standardized using Z-score normalization prior to clustering to ensure that each variable contributes equally to the clustering process.

3.4.2 Determine the number of clusters

When applying K-means clustering, determining a reasonable number of clusters is a key step to ensure the reliability and interpretability of the analysis results. To this end, this study employed two complementary methods to identify the optimal number of clusters: the Elbow method and the Silhouette coefficient.

The Elbow method evaluates model performance by analyzing the changes in within-cluster sum of squares (WCSS) for different numbers of clusters. As the number of clusters increases, WCSS gradually decreases, but the rate of decrease shows a noticeable inflection point. As shown in Figure 3. 3 when the number of clusters is 4, the reduction in WCSS begins to level off, indicating that the model achieves a good balance between clustering accuracy and complexity at this point.

The Silhouette coefficient measures how similar each sample is to its own cluster relative to other clusters. A higher Silhouette coefficient indicates tighter similarity within clusters and clearer differences between clusters. Figure 3. 3 shows that when the number of clusters is 3, the average Silhouette coefficient reaches its maximum, suggesting that the clustering structure is most distinct at this point.

Considering the results from both methods, clustering into three or four groups performs well across different indicators. Considering the actual characteristics of the study area and the requirements for subsequent analysis, selecting four clusters better reflects the spatial heterogeneity of Greater Manchester, achieving a balance between statistical metrics and practical application value. Ultimately, this study based subsequent analyses on four clusters, providing a reliable basis for revealing the distribution and differences of health determinants across different areas.

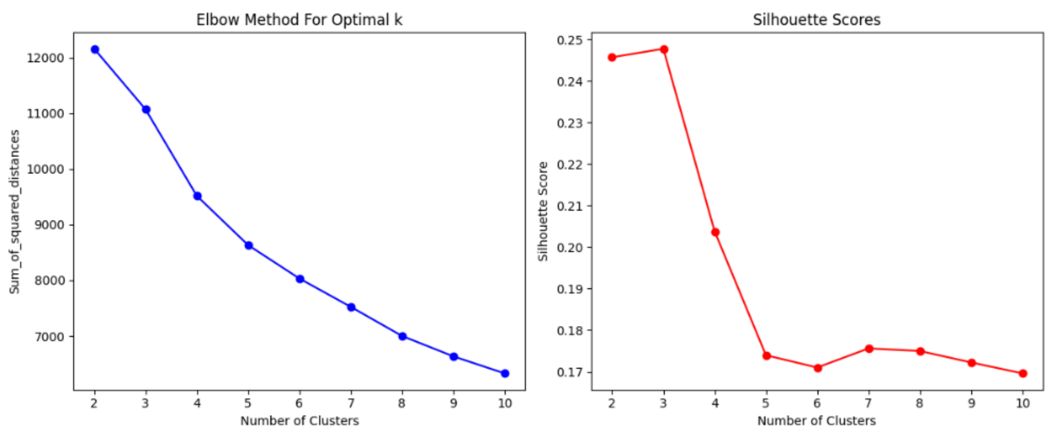

Figure 3. 3 Elbow Method Plot, and Silhouette Coefficient Plot.

3.4.3 Implementation of Clusters

In terms of clustering implementation, standardized variable data for each LSOA were analyzed, with each LSOA's data treated as a single sample input into the K-means algorithm. Each LSOA was assigned a cluster ID, forming representative spatial categories. For visualization, Python was first used to visualize the clustering results to examine spatial distribution patterns and the validity of the clusters. Subsequently, the data were imported into QGIS to generate high-quality static maps, presenting more detailed spatial information. To enhance the usability and flexibility of the analysis, this study further developed an interactive web map using Python's Folium library, combining the QGIS basemap with the clustering information to enable dynamic zooming, hover prompts, and cluster queries, allowing researchers to explore the spatial patterns of each area in an intuitive and detailed manner.

Through the above methods, the study not only ensured the scientific rigor and practicability of the clustering analysis but also provided a reliable basis for revealing the spatial differences in health determinants across different LSOAs in Greater Manchester, thereby supporting subsequent policy-making and intervention strategies.

3.5 Spatial Cluster Analysis of Health Outcomes and Healthcare Access

This section aims to explore the relationships between the spatial clusters derived from the aforementioned K-means clustering and health outcomes as well as healthcare accessibility indicators. By leveraging the defined spatial clusters, it is possible to systematically assess differences in health status and access to medical resources across clusters, thereby revealing potential spatial health inequalities.

Methodologically, health outcome indicators and healthcare accessibility metrics were first combined with each LSOA's cluster label to form an integrated dataset for analysis, which characterizes the health and accessibility features of each spatial cluster.

Descriptive statistical analyses were then conducted by calculating the mean, median, and standard deviation of indicators within each cluster group, accompanied by boxplots to visually illustrate differences and distribution patterns across clusters. To further evaluate the significance of inter-group differences, the normality of each indicator was tested using the Shapiro-Wilk test (Shapiro and Wilk, 1965). Based on the normality results, one-way ANOVA was applied to normally distributed indicators, while non-normally distributed indicators were examined using the non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952).

Furthermore, to assess the relationship between health outcomes and healthcare accessibility, Spearman correlation coefficients were calculated to analyze their rank-based associations. Spearman correlation was chosen because some indicators may not follow a normal distribution, and relationships between variables may be monotonic rather than strictly linear. This method measures monotonic associations based on ranks, avoiding the assumptions of normality and linearity (Spearman, 1904).

By integrating descriptive statistics, significance testing, and correlation analysis, this study systematically interprets the health status and healthcare accessibility across different clusters, discusses potential spatial health inequalities, and examines the possible influence of healthcare accessibility on health outcomes, providing evidence to inform relevant policy-making.

3.6 Software and Tools

This study primarily employed the Python programming language and

Geographic Information System (GIS) software for data processing, spatial analysis, and visualization. Python served as the core tool for data manipulation, statistical analysis, and clustering, using the following libraries:

- pandas: Efficient handling and management of tabular data;
- scikit-learn (sklearn): Implementation of machine learning algorithms, such as K-means clustering and data standardization;
- geopandas: Processing and analysis of spatial vector data;
- matplotlib: Generation of statistical charts and visualizations;
- folium: Creation of interactive maps and spatial visualizations;
- SciPy (scipy.stats): Statistical analysis, including normality tests (Shapiro-Wilk), one-way ANOVA, and Kruskal-Wallis tests;
- statsmodels.api: Statistical modeling and analysis, such as regression and correlation analysis.

These libraries provided a flexible and powerful environment for preprocessing large-scale datasets, conducting descriptive and inferential statistical analyses, and performing spatial classification.

For spatial data visualization, QGIS was employed. QGIS supports multiple geospatial data formats, facilitating integration of spatial layers, thematic map production, and spatial validation of clustering results. Its user-friendly interface makes it suitable for intuitively presenting spatial distribution patterns.

By combining the Python ecosystem with QGIS tools, this study ensured a comprehensive analysis workflow with high reproducibility.

Chapter 4.Results and Analysis

This chapter presents the results of a geodemographic classification of Greater Manchester LSOAs based on health determinants. It first describes the spatial distribution and key characteristics of the four clusters, highlighting their differences. Next, disparities in health indicators and their statistical significance across clusters are compared. Differences in healthcare accessibility and associations with health outcomes are then analyzed. Finally, regression analyses examine the effects of socioeconomic and environmental variables, notably finding that overcrowding shows an unexpected pattern, providing a basis for the subsequent discussion of local health disparities.

4.1 Spatial Distribution of Health Determinant Clusters

In this section, the Greater Manchester LSOAs were classified based on health determinant-related variables using the k-means clustering method, identifying four spatial types with distinct health determinant characteristics. To present the cluster features and their spatial distribution more intuitively, the main variables' average levels for each cluster are first summarized in bar charts, followed by maps showing the geographical distribution of the clusters. Finally, detailed interpretations of each cluster are provided through tables combined with textual explanations.

4.1.1 Cluster Feature Presentation

To provide a more intuitive understanding of the characteristics of each cluster, the mean values of key variables were first calculated and displayed using a bar chart (Figure 4.1). The chart includes air pollution indicators (NO_2, SO_2, PM10), green space coverage (NDVI), and socio-economic variables (unemployment rate, proportion of residents with Level 4 or higher qualifications, proportion of detached houses, proportion of terraced houses, proportion of rented housing, and overcrowding rate). Comparisons reveal that

Cluster 1 has the highest education level; Cluster 2 exhibits the greatest green space coverage and the lowest air pollution levels; Cluster 3's indicators are close to the overall averages; and Cluster 4 shows the highest unemployment rate and overcrowding. These findings indicate clear differences in environmental and socio-economic characteristics across clusters.

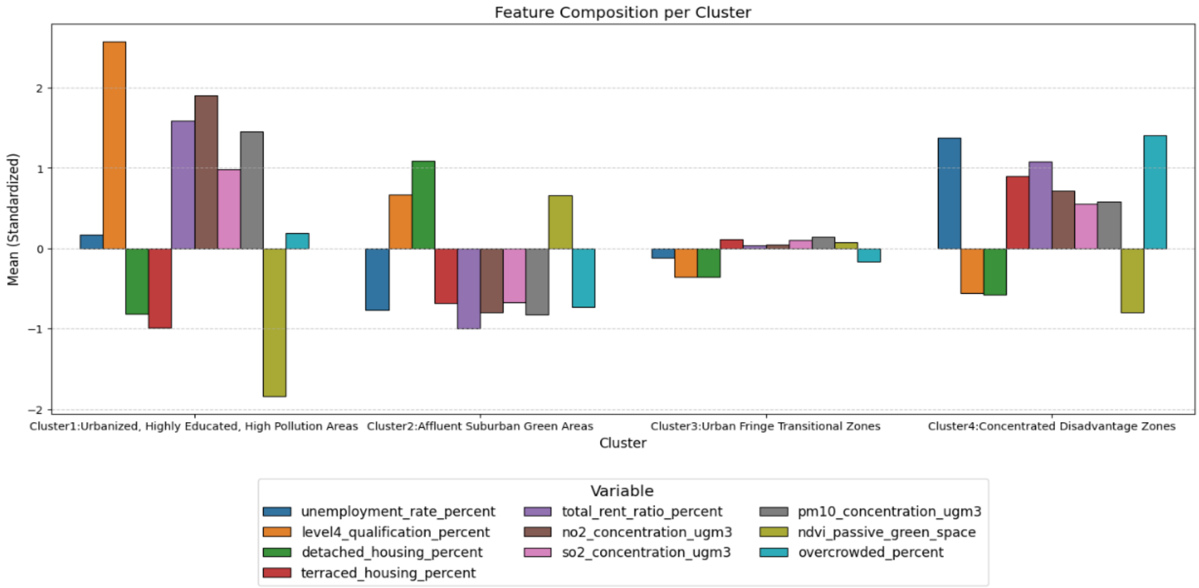

Figure 4.1 Key characteristic bar chart of four health determinant clusters

4.1.2 Cluster Spatial Distribution

After examining the numerical characteristics of each cluster, Figure 4. 2 illustrates their geographic distribution across Greater Manchester. The map differentiates the four cluster types using distinct colors:

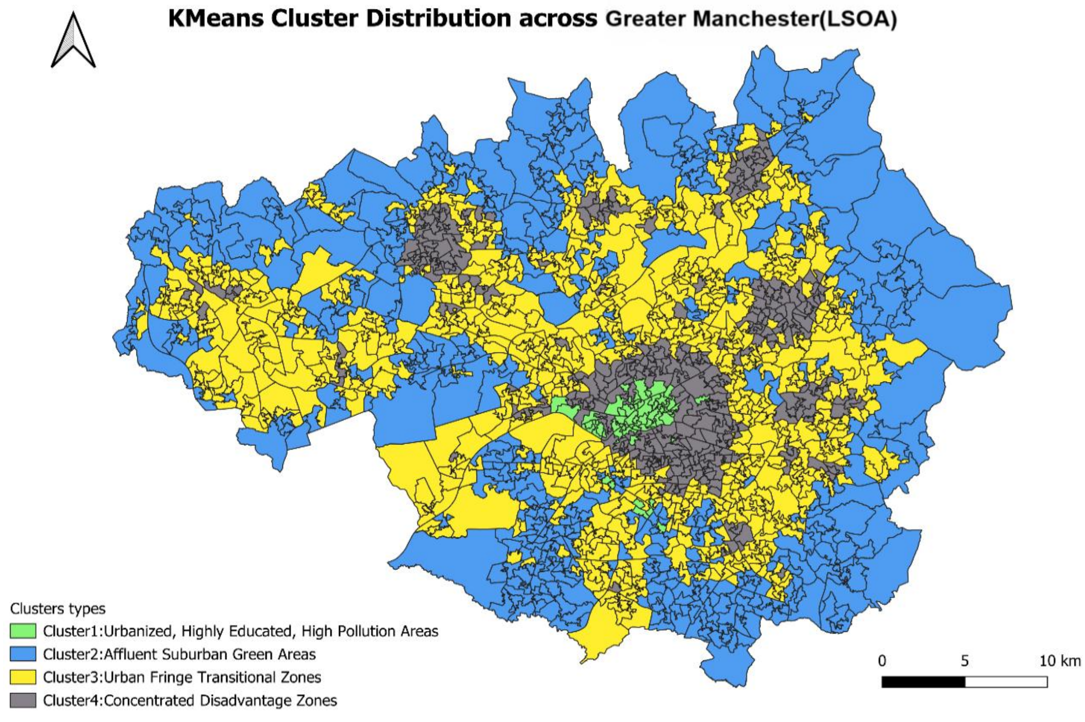

Figure 4. 2. Spatial distribution of four health determinant clusters

Cluster 1:Urbanized, Highly Educated, High Pollution Areas (Green)
Representing highly urbanized areas with a well-educated population but high pollution, mainly concentrated in the city center and its immediate surroundings.

Cluster 2:Affluent Suburban Green Areas (Blue)
Representing affluent suburban areas with substantial green space, mainly located in the outer suburbs, forming large blue zones on the map.

Cluster 3:Urban Fringe Transitional Zones (Yellow)
Representing transitional zones on the urban fringe, these areas are widely distributed, surrounding the central green urban areas and the outer blue suburban areas.

Cluster 4:Concentrated Disadvantage Zones (Gray)
Representing disadvantaged areas primarily around the city center, showing a scattered but discernible distribution pattern.

4.1.3 Detailed Cluster Interpretation

This section systematically presents the characteristics and spatial distribution of health-determinant clusters across different LSOAs in Greater Manchester, integrating both numerical features and spatial patterns. It aids in understanding the spatial structure of the urban health environment and its potential health implications, providing a foundation for subsequent analyses of health outcomes and healthcare resources. Table 4. 1 summarizes the demographic-geographic labels, representative areas, core characteristics, and potential health implications of each cluster. Specifically:

Table 4. 1Summary of Demographic Labels, Representative Areas, Key Characteristics, and Potential Health Implications of Each Cluster

Clusters	**Geodemographic label**	**Example areas**	**Characteristics**	**Potential Health Implications:**
1	**Urbanized, Highly Educated, High Pollution Areas**	Salford Quays; Salford Center ; Manchester City Centre	**Likely represents urban core areas or high-density employment centers:** • Populated by highly educated individuals (possibly young professionals) • High levels of air pollution (NO_2/SO_2/PM10) • Scarcity of green spaces (low NDVI values) • High rental occupancy, fewer detached/terraced houses → Residential structures likely dominated by apartments or high-rise buildings	Despite the high level of education, air pollution and limited green space may have adverse impacts on health.

2	**Affluent Suburban Green Areas**	Rochdale(Littleborough); Oldham(New Delph) Bury(Ramsbottom); Stockport(Marple,Cheadle Hulme)	**This cluster clearly exhibits characteristics of a "middle-to-high income suburban area"**: • High green space (NDVI) and low pollution • Mostly owner-occupied housing (low renting rate) • High education levels and low unemployment • Low overcrowding	It has high green space, low pollution, high education, and low unemployment—factors known to benefit health. Thus, better environmental and social conditions likely contribute positively to health.
3	**Urban Fringe Transitional Zones**	Rochdale(Middleton); Bury(Prestwich) Bolton(Kearsley); Salford(Swinton,Irlam); Wingan(Hindley);	**No significant high or low characteristics in the variables for this cluster** (or information may be incomplete). These areas often represent traditional working-class neighbourhoods that have not undergone significant gentrification. • Moderate to relatively high air pollution levels • High proportion of terraced housing • Likely old urban residential	Having health and environmental conditions between the more affluent/educated clusters (Cluster 1 and Cluster 2) and the more deprived cluster (Cluster 4) Representing a "middle-type" group

			areas or lower-middle-class zones on urban fringes	
4	**Concentrated Disadvantage Zones**	Wigan Town Centre; Oldham Town Centre; Bury Town Centre; Tameside(Ashton-under-lyne,);	**Clear Profile of "Multiple Disadvantaged Communities":** • High unemployment + high pollution • Low education level • Little green space (low NDVI) • High rental housing → high residential instability	This group has the poorest health status among all clusters.

4.2 Health Outcome Characteristics of Each Cluster

This section presents the distribution of health indicators across clusters using boxplots, providing a descriptive comparison of health differences among the four clusters.

1. **Figure 4.3** shows the differences in overall health status (TS037) among the four health determinant clusters in Manchester LSOAs. Overall health status is represented by two indicators: the **poor health rate** (combining "bad" and "very bad" self-reported health) and the **very good health rate** (proportion of respondents reporting "very good" health). To assess the statistical significance of differences between clusters, a non-parametric **Kruskal–Wallis test** was applied. The results indicate that both health indicators differ significantly across the four clusters (**$p < 0.001$**), suggesting

that the observed variations in health outcomes are not merely visual fluctuations but represent statistically meaningful differences.

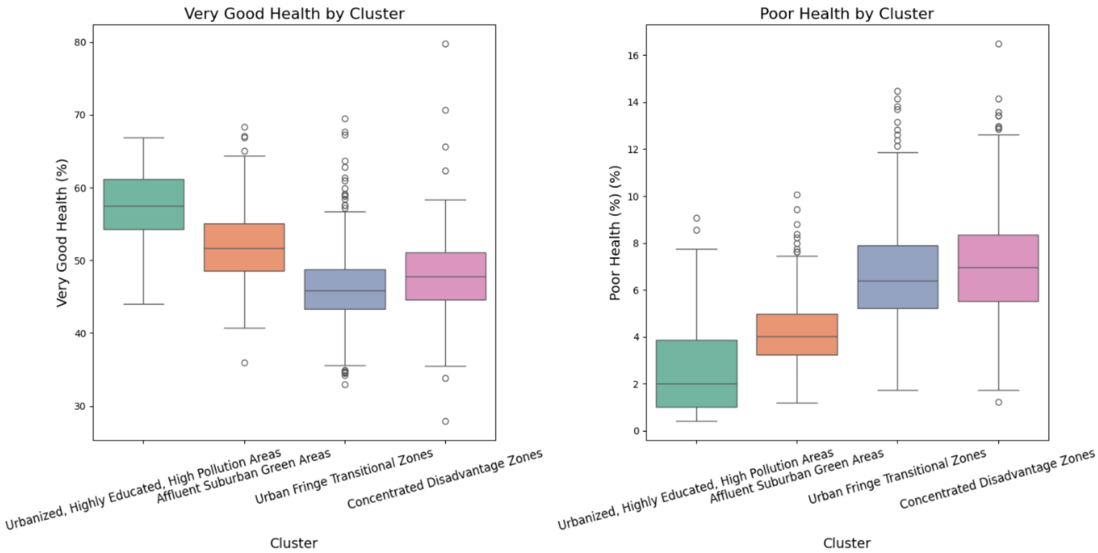

Figure 4. 3 Very Good Health and Poor Health by Cluster

Very Good Health: Cluster 1 (Urbanized, Highly Educated, High Pollution) has the highest proportion of residents reporting "very good" health, with a median near 60%; education and younger population likely enhance health awareness. Cluster 2 (Affluent Suburban Green Areas) is slightly lower but still high, reflecting the protective effects of a favorable living environment. Clusters 3 (Urban Fringe) and 4 (Concentrated Disadvantage) are significantly lower, with Cluster 3 showing a narrower distribution and lower median, indicating notable health disadvantages.

Poor Health: Cluster 4 has the highest proportion of "poor" health (~8%), with Cluster 3 showing similar levels, both indicating health disadvantages. Cluster 1 has the lowest proportion, suggesting education may mitigate pollution impacts, while Cluster 2 is intermediate.

2. To further investigate health disparities among clusters, disability status was included as an additional health indicator. **Figure 4.4** shows the differences in disability status (TS038) across the four LSOA health determinant clusters

in Greater Manchester. Disability status is represented by two metrics: **disability rate** and **health rate** (the proportion of residents without disabilities or long-term illnesses). A Kruskal–Wallis test was also applied, showing significant differences among the four clusters for both disability indicators ($p < 0.001$).

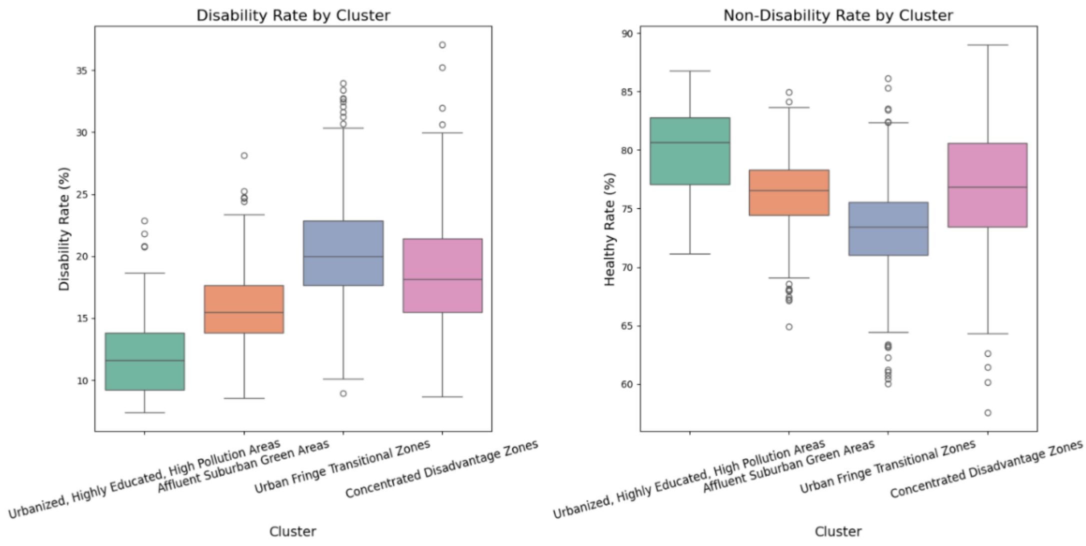

Figure 4. 4 Disability Rate and Non-Disability Rate by Cluster

Disability Rate by Cluster: Cluster 3 (Urban Fringe) shows the highest disability rate, with a median near 20%, indicating significant health challenges. Clusters 1 (Urban, Educated, High Pollution) and 2 (Affluent Suburban) report lower rates, reflecting better health and living conditions.

Good Health Rate by Cluster: Cluster 1 has the highest proportion without long-term illness or disability, suggesting education may mitigate pollution-related risks. Clusters 3 and 4 show lower rates, with Cluster 3 performing worse than Cluster 4, highlighting greater vulnerability in the urban fringe.

In summary, combining the boxplots and significance test results confirms that the four LSOA clusters in Greater Manchester exhibit significant differences in health indicators. Notably, Cluster 3 shows lower overall health and non-disability rates than Cluster 4, while its disability rate is higher, indicating a relatively more vulnerable health status for residents in this area. To further investigate this phenomenon, additional analyses were conducted using multiple regression models to examine the direction and strength of the associations between ten cluster-related variables and health outcome indicators (see Section 4.4 for details).

4.3 Accessibility of Healthcare Resources and Its Correlation with Health Outcomes Across Clusters

1. Based on the above four LSOA clusters, this section focuses on analyzing the spatial accessibility of major healthcare resources, including general practitioners (GPs), dental clinics, hospitals, and pharmacies. Accessibility is measured primarily by residents' travel time from their place of residence to each type of healthcare resource, providing a comprehensive reflection of healthcare service availability. The results (Figure 4. 5) show significant differences in accessibility across clusters: urban center areas (e.g., Cluster 1) have a higher concentration of healthcare resources and thus greater accessibility; suburban and peri-urban transitional areas (e.g., Cluster 2 and Cluster 3) face relatively longer travel times, which may be related to their geographical location. Interestingly, the multiple disadvantage areas (Cluster 4) show relatively shorter travel times to healthcare services, indicating comparatively better accessibility in these regions.

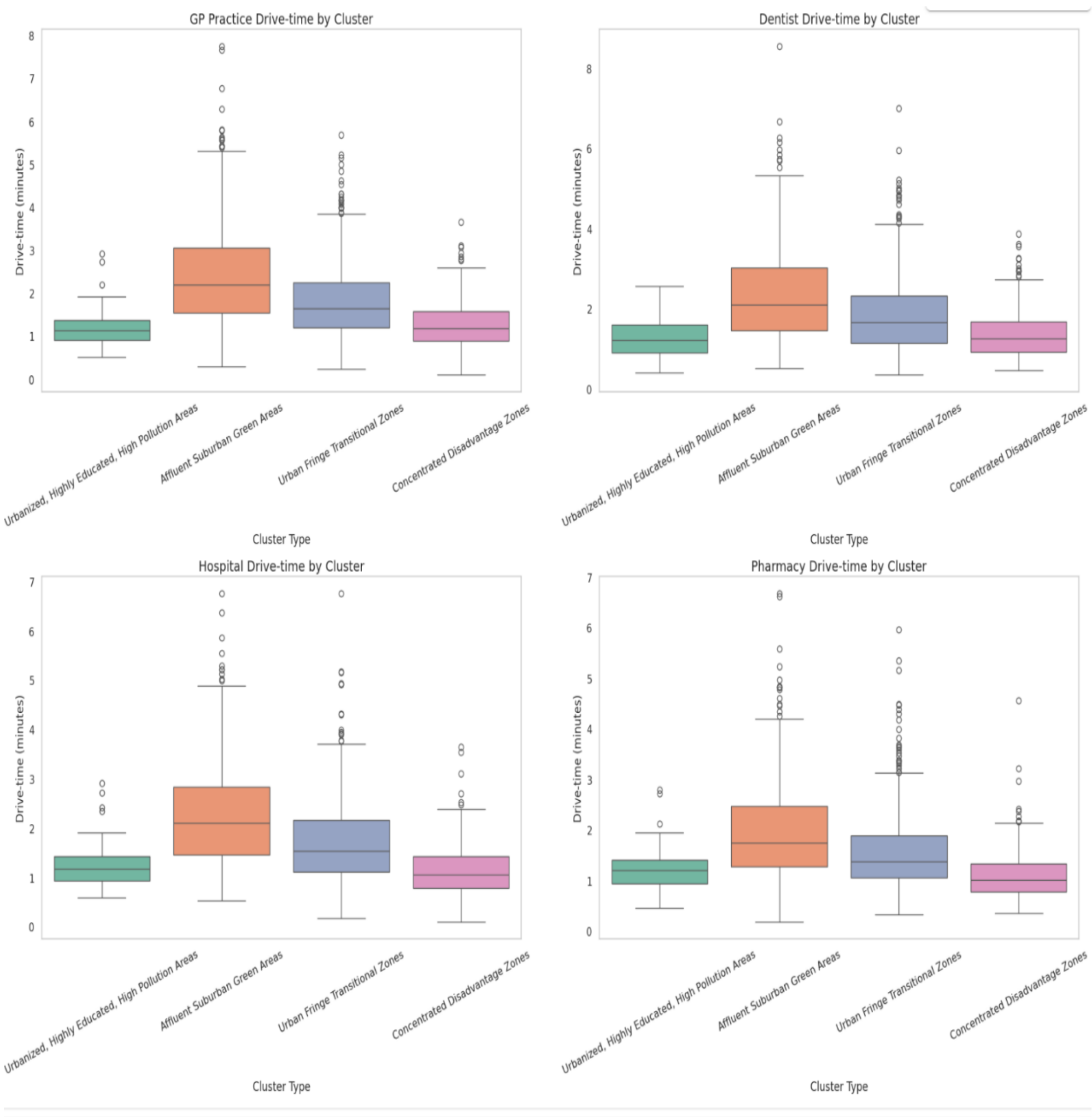

Figure 4. 5 Comparison of healthcare accessibility across different clusters

2.To further investigate the relationship between healthcare accessibility and residents' health status, this study employed Spearman's rank correlation analysis (Spearman, 1904). The Spearman correlation coefficient reflects the strength of monotonic relationships between variables; a p-value below 0.05 indicates statistical significance, and the larger the absolute value of the coefficient, the stronger the correlation. Using this method, the correlation between self-rated health indicators and travel time to healthcare resources (GPs, dental clinics, hospitals, and pharmacies) was examined.

Table 4. 2 Spearman correlation coefficients and p-values between health indicators and healthcare access measures across clusters

Health Indicator	Healthcare Access (travel time in minutes)	Spearman Correlation	p-value
Very good health_pct	GP	0.067	0.005
	Dentist	-0.010	0.688
	Hospital	0.160	0.000
	Pharmacy	0.061	0.012
Poor health_pct	GP	-0.163	0.000
	Dentist	-0.068	0.005
	Hospital	-0.223	0.000
	Pharmacy	-0.171	0.000
disability_rate	GP	-0.066	0.006
	Dentist	0.014	0.562
	Hospital	-0.115	0.000
	Pharmacy	-0.048	0.048
Non-disability_rate	GP	-0.007	0.765
	Dentist	-0.056	0.020
	Hospital	0.017	0.485
	Pharmacy	-0.035	0.154

The results (Table 4.2) show significant correlations between healthcare accessibility and certain health indicators. Travel time to GPs and pharmacies is weakly but significantly positively correlated with the proportion reporting "very good health" ($\rho = 0.067$ and 0.061, $p < 0.05$), while travel time to hospitals shows a stronger positive correlation ($\rho = 0.160$, $p < 0.001$). By contrast, the proportion reporting "poor health" is significantly negatively correlated with travel times to all types of healthcare resources, particularly hospitals ($\rho = -0.223$, $p < 0.001$) and GPs ($\rho = -0.163$, $p < 0.001$). This suggests that areas with longer travel times to healthcare services actually report lower proportions of poor health, which may reflect differences in population structure and health status between urban centers and suburban areas, requiring further analysis in conjunction with socioeconomic factors.

In addition, disability rates are significantly negatively correlated with travel times to hospitals and GPs ($\rho = -0.115$ and -0.066, $p < 0.01$), indicating that distance to healthcare resources may have a greater impact on the health of disabled populations. In contrast, correlations between travel times and the health of non-disabled populations are generally weak and often not statistically significant.

It should be noted that the overall correlation coefficients are relatively low, suggesting that while healthcare accessibility is one factor influencing health, residents' health outcomes are also shaped by multiple determinants, including socioeconomic conditions and environmental quality. Therefore, improving the spatial distribution of healthcare resources should be integrated with efforts to enhance socioeconomic conditions and environmental quality, to achieve more comprehensive health promotion.

4.4 Additional Analysis: Multivariate Regression Testing and Interpretation of Results

To investigate why the health levels of Cluster 3 (the transitional area) are lower than or comparable to those of Cluster 4 (the multiply disadvantaged area), this study constructed multiple linear regression models with four health indicators as dependent variables and ten socioeconomic and environmental variables as independent variables (see workflow in Figure 4. 6). Prior to modeling, all continuous variables were standardized, and multicollinearity was tested using the Variance Inflation Factor (VIF), confirming no multicollinearity issues (all VIF values < 5). To further verify the robustness of the results, two additional tests were conducted: (1) replacing "overcrowding rate" with "non-overcrowding rate" in the model, and (2) excluding the "overcrowding rate" entirely from the model. In both cases, the regression results remained consistent, strengthening the reliability of the findings.

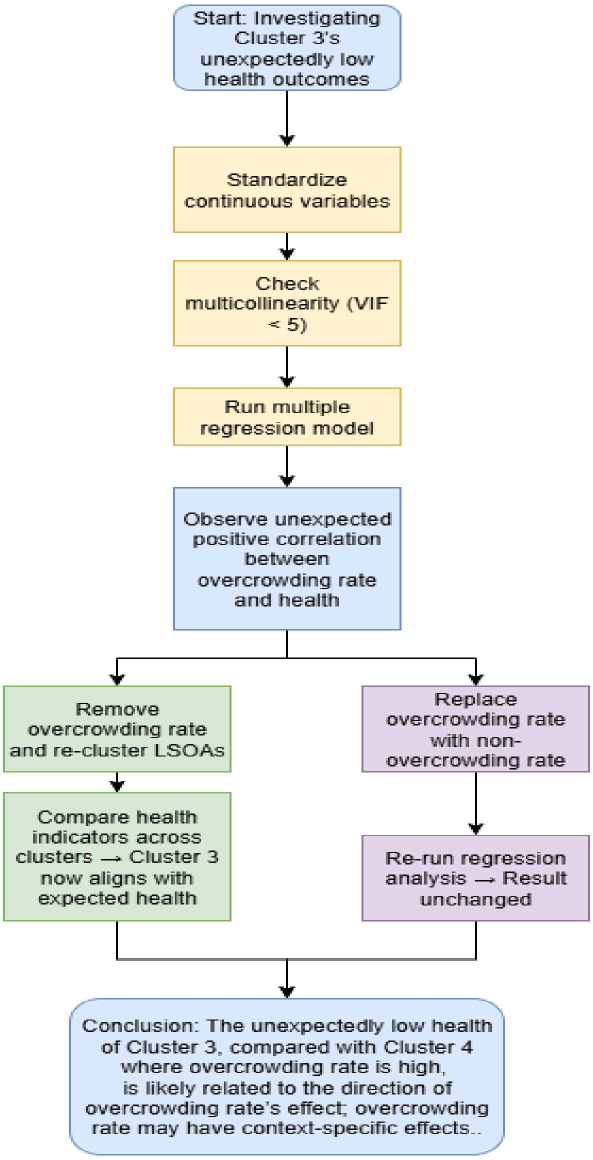

Figure 4. 6 Additional Analysis workflow

The regression results indicate that educational attainment, rental housing proportion, and overcrowding are the most influential factors. Among them, the effects of education and rental housing share align with theoretical expectations, whereas the effect of overcrowding diverges from conventional assumptions. Specifically (see **Table 4. 3**):

- For **Very Good Health_pct** and **Non-Disability_rate**, the coefficients of overcrowding are positive, implying that higher overcrowding is associated with higher proportions of favorable health outcomes.

- For **Poor Health_pct** and **Disability_rate**, the coefficients of overcrowding are negative, indicating that higher overcrowding is linked to lower proportions of adverse health outcomes.

Table 4. 3 Regression Coefficients and Confidence Intervals of Overcrowded Percent on Four Health Indicators

health indicator	coef	ci_lower	ci_upper	p_value
Very good health_pct	**1.4436**	**1.179**	**1.708**	**p < 0.001**
Poor health_pct	**-0.3484**	**-0.445**	**-0.252**	**p < 0.001**
disability_rate	**-1.7245**	**-1.913**	**-1.536**	**p < 0.001**
Non-disability_rate	**2.4366**	**2.221**	**2.652**	**p < 0.001**

This finding contradicts the conventional view that overcrowded living conditions harm health, suggesting that the relationship between overcrowding and health in the study area is complex and context-specific. As shown in the clustering results (Figure 4.1), Cluster 4 (Concentrated Disadvantage Zones) exhibits much higher overcrowding than Cluster 3 (Urban Fringe Transitional Zones), underscoring the distinctive role of overcrowding in explaining the unexpectedly low health levels in Cluster 3. Taken together, these results indicate that the effects of overcrowding may be shaped by local sociocultural factors, residential patterns, or data characteristics, calling for cautious interpretation and further research that considers regional context.

Chapter 5. Discussion

This study aims to conduct a geodemographic classification of Greater Manchester based on key determinants of health and to explore the spatial patterns of health outcomes and healthcare accessibility across different LSOAs. Using k-means clustering, four distinct clusters were identified, exhibiting significant differences in social and environmental determinants. Cluster 1 is characterized by a central urban location, high educational attainment, and relatively elevated pollution levels, showing the best overall health outcomes. Cluster 2 primarily represents affluent suburban areas with abundant green space and moderately high health outcomes. Cluster 4 comprises concentrated disadvantaged communities with notably poor health outcomes. Interestingly, Cluster 3 represents transitional suburban-peripheral areas, with moderate socioeconomic indicators but unexpectedly lower health outcomes (see Sections 4.1–4.3). These patterns highlight the complex interplay of social, environmental, and spatial factors in shaping population health (WHO, 2008; Marmot, 2020).

The clustering analysis reveals health disparities and their spatial distribution across Greater Manchester. Cluster 1, located in highly urbanized central areas, exhibits superior health outcomes relative to other groups. Official statistics indicate a high proportion of young working-age residents and students in this area, suggesting that a younger population structure may benefit overall health, while higher per capita income reflects economic conditions that potentially confer protective effects on health (Office for National Statistics, 2023). Additionally, high educational attainment in this cluster is associated with greater health literacy and the promotion of positive health behaviors (Cutler and Lleras-Muney, 2006). Overall, the health advantage of Cluster 1 likely arises from the combined effects of demographic composition, education, and economic conditions, consistent with prior urban health research.

Cluster 2 exhibits slightly lower health levels than Cluster 1, reflecting the protective influence of favorable socioeconomic conditions and environmental quality. Rich suburban green spaces, low pollution, and high educational levels may improve health by fostering positive health behaviors and enhancing mental well-being. This observation aligns with Mitchell and Popham's (2008) findings on the positive association between green space and resident health.

In contrast, Cluster 4 is concentrated in economically and socially disadvantaged areas, showing the poorest health outcomes. High unemployment, a high proportion of rented housing, and low educational attainment collectively explain this cluster's health disadvantage, consistent with evidence on health inequalities in deprived areas of the UK (Marmot et al., 2020).

The health status of Cluster 3 was unexpectedly lower than or comparable to that of Cluster 4, which had unfavorable overall health determinants. Referring to the regression analyses in Section 4.4, overcrowding rate was positively associated with health indicators such as "Very Good Health" and the "Non-disability Rate," suggesting that in this area, communities with higher population density do not necessarily exhibit poorer long-term health outcomes. This pattern may reflect multiple factors: (1) **Demographic factors**: high-density areas often concentrate younger families or ethnic minority groups, whose long-term health tends to be relatively favorable, contributing to a positive overall health trend; (2) **Socio-cultural factors**: multigenerational households and community support networks can provide psychosocial support and caregiving resources, which are protective for health (Hu et al., 2020); (3) **Housing and spatial environment factors**: while housing affordability may be lower in high-density areas, strong community networks and well-planned high-density environments may promote health by encouraging physical activity and improving transportation and air quality (Kleeman et al., n.d.); and (4) **Statistical and methodological factors**: variable distributions, cluster parameter selection, and spatial heterogeneity may lead to discrepancies between observed results and underlying conditions. Therefore, the observed positive relationship between overcrowding and long-term health indicators does not imply that high-density living inherently improves health, but rather reflects the combined influence of the factors described above. It is important to note that the census-based health indicators used in this study primarily capture long-term or structural health status rather than short-term disease or infectious risk. The commonly cited adverse effects of overcrowding on health are usually based on short-term infectious diseases (e.g., respiratory infections, influenza, COVID-19) or acute health risks, which are directly related to population density. Thus, even if high-density areas show relatively favorable long-term health outcomes, they may still face potential short-term health risks.

Healthcare accessibility and health outcomes display a complex relationship. Suburban areas in Cluster 2 have relatively limited healthcare access but exhibit higher health levels, indicating that socioeconomic and environmental advantages may compensate for accessibility limitations. Conversely, Cluster 4 shows poor health outcomes despite relatively good healthcare access, reflecting high baseline health needs and adverse social conditions. These findings suggest that accessibility alone cannot explain health disparities and must be considered alongside broader determinants of health.

Overall, the results align with UK studies on health inequalities, highlighting the significant influence of socioeconomic and environmental factors on population health. This study extends understanding of regional health heterogeneity through a multifactorial spatial classification approach. Unlike prior research focused on single socioeconomic disadvantages or environmental quality, this study identifies transitional areas with anomalous health outcomes (Cluster 3), underscoring the complexity of population health distribution. Notably, the health outcomes of Cluster 3 are worse than those of the more socioeconomically disadvantaged Cluster 4, suggesting that factors such as population density, household structure, and community support may moderate health outcomes.

These findings underscore the potential value of geodemographic classification in public health interventions and resource allocation. By identifying areas where health outcomes do not fully align with socioeconomic or environmental conditions, targeted health promotion strategies can be designed, healthcare resources optimized, and future research on health inequalities informed.

Chapter 6. Conclusion

This study applies a geodemographic classification of LSOAs in Greater Manchester based on social determinants of health, aiming to reveal the spatial features of health inequalities. It builds on the recognition that health is not solely shaped by individual behaviors or biological risks, but is deeply embedded in the socioeconomic, housing, and environmental conditions of communities. By employing k-means clustering at the LSOA level, combined with statistical analysis of health outcomes, evaluation of healthcare accessibility, and regression modeling, this research provides a spatialized understanding of health disparities in Greater Manchester.

The analysis shows that education level, rental burden, and household overcrowding are the most salient predictors of health. First, education level proves to be the central determinant of health: residents in areas with higher educational attainment display significantly better health outcomes than those in less-educated areas. Notably, this advantage persists even under environmental pressures such as higher air pollution or limited green space, a pattern particularly pronounced in Cluster 1. Second, housing affordability also exerts a significant influence: areas with higher rental rates generally show poorer health outcomes. However, this effect is mitigated in locations with strong educational resources and good accessibility to healthcare services (e.g., Cluster 1), suggesting that the impact of housing conditions on health is moderated by social and educational resources. At the same time, in this study, household overcrowding is positively associated with better health outcomes, indicating that in some areas, higher residential density may foster family or community-based social support, thereby improving health. These findings underscore that the influence of health determinants is context-dependent, shaped by the interplay of education, housing stress, environment, healthcare accessibility, and community support.

The findings yield three key insights. First, education remains the core determinant of health, and improving educational attainment continues to be the most effective long-term strategy for reducing health inequalities. Second, housing affordability is associated with health outcomes, but its effects are not isolated; rather, they are moderated by education and social resources. Third, the case of overcrowding highlights that the impacts of health determinants are not universally homogeneous, but may vary by local context.

Methodologically, this study integrates k-means clustering, descriptive statistics, group-difference testing, correlation analysis, and regression modeling into a multidimensional validation framework. This not only strengthens explanatory power but also highlights the potential of clustering outcomes for informing targeted policy interventions. Specifically, the results suggest that the urban core (Cluster 1) should prioritize improving air quality; suburban areas (Cluster 2), while showing overall higher health levels, require enhanced healthcare accessibility; transitional edge areas (Cluster 3) need attention to housing conditions and community support networks; and disadvantaged communities (Cluster 4) demand integrated improvements in socioeconomic conditions and environmental quality. The combination of clustering and statistical analysis thus reveals the multidimensional causes of health inequality and provides scientific evidence for precision-targeted interventions and resource allocation.

In summary, this study demonstrates that health inequalities in Greater Manchester are deeply embedded in the interplay of multiple factors, including educational attainment, housing affordability, environmental pressures, and resource accessibility, rather than being solely determined by individual behaviors or biological risks. Therefore, at the policy level, priority should be given to improving education and housing conditions while preserving the existing strengths of each area. More importantly, interventions must be tailored to the specific socioeconomic and environmental contexts of different regions, rather than adopting a uniform and universal approach.

Despite these contributions, the study has several limitations. First, the use of LSOA-level analysis may introduce the modifiable areal unit problem (MAUP), potentially obscuring within-area heterogeneity (Lee et al., 2014). Second, while the analysis covers ten variables across socioeconomic and environmental domains, data availability constraints limited inclusion of other potentially influential factors such as individual income, lifestyle behaviors, cultural practices, social support networks, and household structures, which may weaken explanatory power. Third, the overcrowding indicator may not fully capture subjective residential experiences. Finally, the cross-sectional design precludes causal inference, and correlational findings should therefore be interpreted with caution.

Future research could adopt longitudinal designs to capture the temporal dynamics of health inequalities and explore causal mechanisms. Beyond overall health outcomes, future studies could also focus on specific disease prevalence (e.g., cardiovascular diseases, respiratory conditions, mental health issues, or infectious diseases), thereby uncovering their spatial associations with socioeconomic, environmental, and healthcare access factors. Combining such analyses with qualitative surveys or household-level data would further enhance understanding of the mechanisms underlying atypical health patterns.

References

Adler, N.E. and Newman, K., 2002. Socioeconomic disparities in health: pathways and policies. *Health Affairs*, 21(2), pp.60–76. [online] DOI:10.1377/hlthaff.21.2.60 [Accessed 23 July 2025].

Alexandrov, N., Scott, E.S., Janssen, M.F., Lubetkin, E.I., Yfantopoulos, J.N., Bonsel, G.J. and Haagsma, J.A., 2024. The relationship between healthcare access and change in health-related quality-of-life among the general population of five countries during the COVID-19 pandemic. *Quality of Life Research*, 33, pp.2541–2552. https://doi.org/10.1007/s11136-024-03704-1 [Accessed 23 July 2025].

Anselin, L., 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2), pp.93–115. [online] Available at: https://doi.org/10.1111/j.1538-4632.1995.tb00338.x [Accessed 23 July 2025].

Bailey, A.J., 2009. Population geography: lifecourse matters. *Progress in Human Geography*, 33(3), pp.407–418. DOI:10.1177/0309132508096355 [Accessed 23 July 2025]

Bambra, C., Gibson, M., Sowden, A., Wright, K., Whitehead, M. and Petticrew, M., 2010. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. *Journal of Epidemiology and Community Health*, 64(4), pp.284–291. Available at: https://www.jstor.org/stable/20721190 [Accessed 2 August 2025].

Bazlan, M.J., Rasam, A.R.A. and Ghazali, R., 2020. Web-based augmented reality mobile GIS for disease prevention and control programme. *Test Engineering and Management*, 82(12018), pp.12018–12028. [PDF]

Bentley, R., Baker, E., Mason, K., Subramanian, S.V. and Kavanagh, A.M., 2011. Association between housing affordability and mental health: a longitudinal analysis of a nationally representative household survey in Australia. *American Journal of Epidemiology*, 174(7), pp.753–760. Available at: https://doi.org/10.1093/aje/kwr161

Bloemsma, L.D., Gehring, U., Klompmaker, J.O., Hoek, G., Janssen, N.A.H., Lebret, E., Brunekreef, B. and Wijga, A.H., 2019. Green space, air pollution, traffic noise and cardiometabolic health in adolescents: The PIAMA birth cohort. *Environment International*, 131, p.104991. Available at: https://doi.org/10.1016/j.envint.2019.104991 [Accessed 19 August 2025].

Boulos, M.N.K., Curtis, A.J. and AbdelMalik, P., 2009. Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics*, 8, article 46. Available at: https://doi.org/10.1186/1476-072X-8-46 [Accessed 23 July 2025].

Braveman, P., 2006. Health disparities and health equity: Concepts and measurement. *Annual Review of Public Health*, 27, pp.167–194. [online] Available at: https://doi.org/10.1146/annurev.publhealth.27.021405.102103 [Accessed 18 August 2025].

Braveman, P., Cubbin, C., Egerter, S., Williams, D.R. and Pamuk, E., 2011. Socioeconomic disparities in health in the United States: What the patterns tell us. *American Journal of Public Health*, 100(S1), pp.S186–S196. Available at: https://doi.org/10.2105/AJPH.2009.166082 [Accessed 3 August 2025].

Briggs, D., 2005. The role of GIS: Coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health, Part A*, 68(13–14), pp.1243–1261. Available at: https://doi.org/10.1080/15287390590936094 [Accessed 19 August 2025].

Chen, J., Atkinson, R.W., Andersen, Z.J., Oftedal, B., Stafoggia, M., Lim, Y.-H., Bekkevold, T., Krog, N.H., Renzi, M., Zhang, J., Bauwelinck, M., Janssen, N., Strak, M., Forastiere, F., de Hoogh, K., Rodopoulou, S., Katsouyanni, K., Raaschou-Nielsen, O., Samoli, E., Brunekreef, B., Hoek, G., Vienneau, D., 2024. Long-term exposure to ambient air pollution and risk of lung cancer – A comparative analysis of incidence and mortality in four administrative cohorts in the ELAPSE study. *Environmental Research*, 120236. DOI: 10.1016/j.envres.2024.120236[Accessed 1 August 2025].

Coburn, D., 2015. Income inequality, welfare, class and health: A comment on Pickett and Wilkinson, 2015. *Social Science and Medicine*, 146, pp.228–232. Available at: https://doi.org/10.1016/j.socscimed.2015.09.002 [Accessed 5 August 2025].

Cutler, D.M. and Lleras-Muney, A., 2006. *Education and health: evaluating theories and evidence*. Cambridge, Massachusetts: National Bureau of Economic Research. DOI: 10.3386/w12352 [Accessed 4 August 2025].

Diez Roux, A.V. and Mair, C., 2010. Neighborhoods and health. *Annals of the New York Academy of Sciences*, 1186(1), pp.125–145. https://doi.org/10.1111/j.1749-6632.2009.05333.x

Diez Roux, A.V., 2001. Investigating Neighborhood and Area Effects on Health. *American Journal of Public Health*, 91(11), pp.1783–1789. DOI: 10.2105/AJPH.91.11.1783 [Accessed 3 August 2025].

Evans, G.W., Wells, N.M. and Moch, A., 2003. Housing and mental health: A review of the evidence and a methodological and conceptual critique. *Journal of Social Issues*, 59(3), pp.475–500. https://doi.org/10.1111/1540-4560.00074 [Accessed 2 August 2025].

Grekousis, G., Wang, R. and Liu, Y., 2021. Mapping the geodemographics of racial, economic, health, and COVID-19 deaths inequalities in the conterminous US. *Applied Geography*, 135, p.102558. https://doi.org/10.1016/j.apgeog.2021.102558.

Guagliardo, M.F. 2004. Spatial accessibility of primary care: concepts, methods and challenges. *International journal of health geographics*. 3(1),3. Available at: https://doi.org/10.1186/1476-072X-3-3 [Accessed 23 July 2025].

Hu, Y., Ruiz, M., Bobak, M. and Martikainen, P., 2020. Do multigenerational living arrangements influence depressive symptoms in mid-late life? Cross-national findings from China and England. *Journal of Affective Disorders*, 277, pp.584–591. Available at:https://doi.org/10.1016/j.jad.2020.07.142

Huang, S., Li, H., Wang, M., Qian, Y., Steenland, K., Caudle, W.M., Liu, Y., Sarnat, J., Papatheodorou, S. & Shi, L., 2021. Long-term exposure to nitrogen dioxide and mortality: A systematic review and meta-analysis. *Science of The Total Environment*, 776, p.145968. Available at:https://doi.org/10.1016/j.scitotenv.2021.145968

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp.651-666. https://doi.org/10.1016/j.patrec.2009.09.011[Accessed 23 July 2025].

Kindig, D. and Stoddart, G., 2003. What is population health? *American*

Journal of Public Health, 93(3), pp.380–383. Available at: What Is Population Health? | AJPH | Vol. 93 Issue 3 [Accessed 19 August 2025].

Kleeman, A., Gunn, L. and Giles-Corti, B., n.d. Density and other health benefits. National Heart Foundation of Australia. Available at: Health Benefits of Housing Diversity [Accessed 13 August 2025].

Kruskal, W.H. and Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), pp.583–621. Available at: https://doi.org/10.1080/01621459.1952.10483441 [Accessed 19 August 2025].

Lake, C., Mishra, A. and Conroy, M., 2023. Embedding a multiple health condition approach across NHS Greater Manchester. *Future Healthcare Journal*, 10(3), pp.211–218. doi: 10.7861/fhj.2023-0081

Lee, J., Alnasrallah, M., Wong, D., Beaird, H. and Logue, E., 2014. Impacts of scale on geographic analysis of health data: An example of obesity prevalence. *ISPRS International Journal of Geo-Information*, 3(4), pp.1198–1210. Available at: https://doi.org/10.3390/ijgi3041198 [Accessed 23 July 2025].

Markevych, I., Schoierer, J., Hartig, T., Chudnovsky, A., Hystad, P., Dzhambov, A.M., de Vries, S., Triguero-Mas, M., Brauer, M., Nieuwenhuijsen, M.J., Lupp, G., Richardson, E.A., Astell-Burt, T., Dimitrova, D., Feng, X., Sadeh, M., Standl, M., Heinrich, J. and Fuertes, E., 2017. Exploring pathways linking greenspace to health: Theoretical and methodological guidance. *Environmental Research*, 158, pp.301–317. https://doi.org/10.1016/j.envres.2017.06.028

Marmot, M., 2005. Social determinants of health inequalities. *The Lancet*, 365(9464), pp.1099–1104. [online] Available at: https://doi.org/10.1016/S0140-6736(05)71146-6 [Accessed 4 August 2025].

Marmot, M., 2020. Health equity in England: the Marmot review 10 years on. *BMJ*, 368, p.m693. Available at: https://doi.org/10.1136/bmj.m693

Mason, K.E., Baker, E., Blakely, T. and Bentley, R.J., 2013. Housing affordability and mental health: Does the relationship differ for renters and home purchasers? *Social Science and Medicine*, 94, pp.91–97. DOI: 10.1016/j.socscimed.2013.06.023 [Accessed 18 August 2025].

Milner, A., Page, A. and LaMontagne, A.D., 2014. Cause and effect in studies on unemployment, mental health and suicide: a meta-analytic and conceptual review. *Psychological Medicine*, 44(5), pp.909–917. Doi:10.1017/S0033291713001621 [Accessed 1 August 2025].

Mitchell, R. and Popham, F., 2008. Effect of exposure to natural environment on health inequalities: an observational population study. *The Lancet*, 372(9650), pp.1655–1660. Available at: Effect of exposure to natural environment on health inequalities: an observational population study - The Lancet [Accessed 3 June 2025].

Montez, J.K. and Zajacova, A., 2013. Trends in mortality risk by education level and cause of death among US white women from 1986 to 2006. American Journal of Public Health, 103(3), pp.473–479. Available at: DOI: 10.2105/AJPH.2012.301128 [Accessed 2 August 2025].

Mueller, N., Rojas-Rueda, D., Khreis, H., Cirach, M., Milà, C., Espinosa, A., Foraster, M., McEachan, R.R.C., Kelly, B., Wright, J. and Nieuwenhuijsen, M., 2018. Socioeconomic inequalities in urban and transport planning related exposures and mortality: A health impact assessment study for Bradford, UK. *Environment International*, 121(Pt 1), pp.931–941. Available at: https://doi.org/10.1016/j.envint.2018.10.017 [Accessed 2 August 2025].

Noble, M., Wright, G., Smith, G. and Dibben, C., 2006. Measuring multiple deprivation at the small-area level. *Environment and Planning A*, 38(1), pp.169–185. Available at: https://doi.org/10.1068/a37168[Accessed 22 June 2025]

Norman, P., Lloyd, C., McLennan, D., Ferguson, S. and Catney, G., 2024. 50-year deprivation trajectories: Local area change in England and Wales, 1971–2021. *Applied Spatial Analysis and Policy*, 17, pp.1183–1208. Available at: https://doi.org/10.1007/s12061-024-09583-w [Accessed 22 July 2025].

Office for National Statistics, 2023. Population and migration statistics transformation, Manchester case study: Understanding the effectiveness of our administrative data in telling the story of Manchester. 27 June. Available at: Population and migration statistics transformation, Manchester case study - Office for National Statistics.

Orellano, P., Reynoso, J. and Quaranta, N., 2021. Short-term exposure to sulphur dioxide (SO_2) and all-cause and respiratory mortality: A systematic review and meta-analysis. *Environment International*, 150, p.106434. Available at: https://doi.org/10.1016/j.envint.2021.106434

Ortiz-Prado, E., Simbaña, K., Gómez, L., Henriquez-Trujillo, A.R., Cornejo-Leon, F., Vasconez, E., Castillo, D. and Viscor, G., 2017. The disease burden of suicide in Ecuador: a 15 years' geodemographic cross-sectional study (2001–2015). *BMC Psychiatry*, 17(1), pp.1-10. DOI: 10.1186/s12888-017-1502-0 [Accessed 24 July 2025].

Padgett, D.K., 2020. Homelessness, housing instability and mental health: making the connections. *BJPsych Bulletin*, 44(5), pp.197–201. Available at: https://doi.org/10.1192/bjb.2020.49

Paul, K.I. and Moser, K., 2009. Unemployment impairs mental health: Meta-analyses. *Journal of Vocational Behavior*, 74(3), pp.264–282. Available at: https://doi.org/10.1016/j.jvb.2009.01.001 [Accessed 2 August 2025].

Pearce, J., 2012. The 'blemish of place': Stigma, geography and health inequalities. A commentary on Tabuchi, Fukuhara and Iso. *Social Science and Medicine*, 75(11), pp.1921–1924. Available at: https://doi.org/10.1016/j.socscimed.2012.07.033 [Accessed 18 August 2025].

Penchansky, R. and Thomas, J.W., 1981. The concept of access: Definition and relationship to consumer satisfaction. *Medical Care*, 19(2), pp.127–140. Available at: https://www.jstor.org/stable/3764310 [Accessed 17 August 2025].

Petersen, J., Gibin, M., Longley, P., Mateos, P., Atkinson, P. and Ashby, D., 2011. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems*, 13, pp.173–192. [online] Available at: https://doi.org/10.1007/s10109-010-0113-9 [Accessed 18 August 2025].

Pickett, K.E. and Wilkinson, R.G., 2015. Income inequality and health: A causal review. *Social Science and Medicine*, 128, pp.316–326. Available at: https://doi.org/10.1016/j.socscimed.2014.12.031 [Accessed 1 August 2025].

Purdam, K., 2017. The devolution of health funding in Greater Manchester in the UK: A travel map of life expectancy. *Environment and Planning A: Economy and Space*, 49(7), pp.1562–1564. doi:10.1177/0308518X17697701.

Python-visualization (2025) *Folium: Python Data Visualisation library*. Available at: https://python-visualization.github.io/folium/ [Accessed: 9 August 2025]

Rad, J. ,2025. Health inequities: a persistent global challenge from past to future, *International Journal for Equity in Health*, 24, p. 148. Available at: https://doi.org/10.1186/s12939-025-02526-y

Richardson, A., 2022. Cancer deaths, diabetes figures and obesity rates show shocking health inequalities in Trafford. *Manchester Evening News*, 17 January. Available at:Cancer deaths, diabetes figures and obesity rates show shocking health inequalities in Trafford - Manchester Evening News [Accessed 30 July 2025].

Rigolon, A., 2016. A complex landscape of inequity in access to urban parks: A literature review. *Landscape and Urban Planning*, 153, pp.160–169. Available at: https://doi.org/10.1016/j.landurbplan.2016.05.017 [Accessed 1 August 2025].

Rizvi, S.A., Umair, M. and Cheema, M.A., 2021. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons and Fractals*, 151, p.111240. Available at: https://doi.org/10.1016/j.chaos.2021.111240 [Accessed 3 June 2025].

Roelfs, D.J., Shor, E., Davidson, K.W. and Schwartz, J.E., 2011. Losing life and livelihood: A systematic review and meta-analysis of unemployment and all-cause mortality. *Social Science and Medicine*, 72(6), pp.840–854. Available at: https://doi.org/10.1016/j.socscimed.2011.01.005 [Accessed 8 August 2025].

Rushton, G., 2003. Public health, GIS, and spatial analytic tools. *Annual Review of Public Health*, 24(1), pp.43–56. Available at: https://doi.org/10.1146/annurev.publhealth.24.012902.140843 [Accessed 22 July 2025]

Shapiro, S.S. and Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), pp.591–611. Available at: https://doi.org/10.1093/biomet/52.3-4.591 [Accessed 19 August 2025].

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), pp.72–101.

Available at: https://www.jstor.org/stable/1412159 [Accessed 9 August 2025].

Twohig-Bennett, C. and Jones, A., 2018. The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. *Environmental Research*, 166, pp.628–637. Available at: https://doi.org/10.1016/j.envres.2018.06.030 [Accessed 1 August 2025].

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., et al., 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), pp.261–272. Available at: https://doi.org/10.1038/s41592-019-0686-2 [Accessed 1 August 2025].

Wang, F., 2020. Why public health needs GIS: a methodological overview. *Annals of GIS*, 26(1), pp.1–12. Available at: https://doi.org/10.1080/19475683.2019.1702099[Accessed 22 July 2025]

Whitehead, M., 1992. The concepts and principles of equity and health. *International Journal of Health Services*, 22(3), pp.429–445. [online] Available at: https://www.jstor.org/stable/4513105 [Accessed 4 August 2025].

Wilkie, A.A., Krajewski, A.K., Njie, F., Park, K., Zelasky, S., Rappazzo, K.M. and Luben, T.J., 2025. Long-term exposure to criteria air pollutants and infant mortality: a systematic review and meta-analysis. *International Journal of Hygiene and Environmental Health*, [online] Available at: https://doi.org/10.1016/j.ijheh.2025.114587 [Accessed 7 Aug. 2025].

World Health Organization, 2008. *Closing the gap in a generation: health equity through action on the social determinants of health*. Geneva: World Health Organization. Available at: https://www.who.int/publications/i/item/WHO-IER-CSDH-08.1 [Accessed 25 July 2025]

World Health Organization, 2018. *WHO housing and health guidelines*. Geneva: World Health Organization. ISBN 978-92-4-155037-6. [online] Available at: WHO Housing and Health Guidelines - NCBI Bookshelf[Accessed 26 July 2025]

Xu, R. and Wunsch, D., 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), pp.645-678. DOI: 10.1109/TNN.2005.845141[Accessed 23 July 2025]

Zhang, Y. and Luo, F., 2024. Linkages among socio-economic status, green space accessibility, and health outcomes: An environmental justice perspective in Australia. *Sustainable Cities and Society*, 114, 105784. Available at: https://doi.org/10.1016/j.scs.2024.105784 [Accessed 1 August 2025].

Zhao, M., Scepanovic, S., Law, S., Obadic, I., Wu, C. and Quercia, D. 2025. MedGNN: Capturing the Links Between Urban Characteristics and Medical Prescriptions. *arXiv preprint arXiv:2504.04739*. Available at: https://doi.org/10.48550/arXiv.2504.04739[Accessed 24 July 2025]

Appendix

- The code for this dissertation is available in my GitHub repository: https://github.com/kzw333/5160
  or can be run directly on Google Colab: https://colab.research.google.com/drive/1wHHuuEcmY2ExPCdW7gjPGbJFBuNquT4A#scrollTo=seFZKlSdXbW7
- The OneDrive folder contains all the original raw data, processing steps, and other related materials 5160
- An interactive map for exploring the clustered areas in detail: new-boroughs-greater_manchester_clusters_map.html