Assigment 1: Regional Inequality

Levels deprivation in the North West cities of Liverpool and Manchester using Census 2021

Student ID:

11464537, 11395438

11345006 , 10806957

April 2024

Topological Data Analysis

DATA70302

Contents

List of Figures

Chapter 1

Introduction

11395438

Topological Data Analysis (TDA) is a data science methodology that focuses on understanding the shape and structure of complex datasets (Dlotko et al., 2022). Ball Mapper (BM) is a specific TDA tool that creates a visual map of high-dimensional data, helping uncover hidden patterns and relationships within datasets. This paper demonstrates the use of BM in exploring deprivation within 14 wards within the areas of Merseyside and Greater Manchester, comparing the contributing factors associated with higher levels of deprivation. The robustness of the results are assessed through changing two empirical choices: variables used, and radius chosen.

Chapter 2

Literature Review

11395438

Liverpool and Manchester are highly comparable cities in terms of deprivation. The cities share similar pasts in terms of industrialisation and, more importantly, deindustrialisation, commonly associated with higher poverty levels, lower life expectancy and other related deprivation factors (Walsh et al, 2010). Both Liverpool and Manchester emerged as global cities in the 18th Century, boasting international trade and grand architectural landscapes (Sykes et al., 2013). Liverpool emerged as a global port base for international trade, with Manchester connected through the vast network of canals created, both cities experiencing an economic boom from the trade of salt, slaves and raw materials (Brown, 2009). However, much like other cities in the Northwest, favouring Britain's south and east coast ports, transport links and maritime containerisation led to a serious decline in the cities' economic circumstances during the late 20th Century. Liverpool's population halved to just 430,000 by 2001 from a peak in the 1930s (Sykes et al., 2013). The effects of industrialisation led to widespread deprivation with Walsh et al., (2010) finding Liverpool and Manchester to be almost identical, with almost 25% classed as income deprived. Whilst various regeneration and re-urbanisation projects have aimed to reverse the cities' declines, spatial inequalities remain dramatic and exacerbated by these recent developments (Kelly et al., 2023). For instance, the difference between 'Healthy Life Expectancy' between Liverpool's wealthiest and poorest regions vary by almost 30 years (Sykest et al., 2013)

Deprivation is multi-faceted and many factors contribute to deprived areas which will be analysed within this study. As such, the Index of Multiple Deprivation (IMD) was created to rep-

resent a combination of these factors and used to separate 'more-deprived' from 'less-deprived' neighbourhoods. This index includes a range of indicators, including income, employment, health and disability, education skills and training, barriers to housing and other services, crime and living environment. The indicators are weighted and combined to generate a domain score, again weighted and combined to form a composite measure of deprivation for each neighbourhood (Robson et al., 2008; Schultz-Baing and Wong, 2012)

Other, single factors have been identified in literature as influential to deprivation in neighbourhoods. Charlton et al, (2013) conducted an epidemiological review, finding individuals living in deprived circumstances are disproportionately affected by complex health issues, such as chronic diseases and mental health disorders. Limited access to healthcare services in deprived areas compounds this disparity, leading to delayed diagnosis and treatment of illnesses. Aoki and Santiago (2018) found that English proficiency significantly impacts the socioeconomic outcomes of UK immigrants, linked to language barriers in accessing essential services, such as healthcare, education and welfare services, exacerbating economic and social disparities between areas. Aoki and Santiago (2018) also pointed to barriers to stable jobs, with language barriers limiting employment opportunities, leading to higher unemployment rates within these communities and contributing to economic deprivation. Car dependance, described by Mattioli (2021) contributed to transport poverty, spatially restricting those with lower incomes due to the provision of public transport, limited to time schedules or location of transport routes (Schwanen et al., 2015). This contributes to social isolation and exclusion through reduced access to services such as shops, healthcare and education locations, as well as restricting employment opportunities (Mattioli, 2021). Lower education qualifications are also linked to increased levels of deprivation, limiting socio-economic advancement through employment, limited to lower paid jobs and more likely to face job instability and unemployment (Ilie et al., 2021). This relationship is cyclical, leading to generational cycles of deprivation, with clear patterns between children living in deprived areas with lower educational outcomes (McCartney et al., 2017).

Chapter 3

Methodology

11464537

The methodology used in this report follows Rudkin and Webber (2023) Topological Data Analysis – Ball Mapper (TDA-BM). With it, we aim to map joint factors that contribute to higher levels of deprivation on wards in North West cities of Liverpool and Manchester.

First we select the data from the Census 2021 to pre-process for the statistical analysis (mean, variance, correlation, and OLS regression). The Census 2021 will provide variables to analyze the relationship between levels of deprivation in wards.

Considering the Anscombe's quartet (Anscombe, 1973), we can't rely solely on statistical properties when analyzing data. Anscombe's quartet demonstrates that different datasets can yield the same regression model but exhibit vastly different patterns when plotted, therefore mere existence of an OLS model explaining deprivation levels does not guarantee that the appearance of deprivation across wards will conform to the expectations.

That's why after the statistical analysis we will map using Topological Data Analysis BallMapper (TDABM) algorithm of Dłotko (2019). The TDA-BM graph allows us a visualization of the following aspects:

The ball's size indicates the quantity of wards with similarities. The color of the balls indicates the average value of the selected deprivation-related outcome variable across all included wards Connections between balls, which indicate that at least one ward is a member of both balls Outcomes from the TDABM algorithm are dependent solely on the choice of the radius Rudkin

et al. (2023), Well chose a size for the ball radius, which is the parameter that determines the size of the balls in the TDA-BM graph. Adjusting this radius can impact the granularity of the analysis. A small radius can lead to detailed pictures. However, it runs the risk of over-focusing on these local aspects, potentially obscuring the broader context or bigger picture. Conversely, a too large radius might overlook local details, leading to a lack of granularity and potentially missing critical insights. Balancing the radius parameter is crucial to strike meaningful insights.

Then we'll validate the robustness of any inference derived from TDA-BM analysis through two methods. Firstly, by changing the radius parameter, one can assess how sensitive the results are to changes in the size of the balls and their corresponding groupings. Secondly, experimenting with a different explanatory variable. These strategies collectively provide a comprehensive approach to assessing the reliability and robustness of insights gained through TDA-BM analysis.

By acknowledging the limitations of purely statistical approaches and leveraging the visual insights provided by TDA-BM, we aim to uncover nuanced patterns and relationships that may not be apparent through traditional analysis alone.

Chapter 4

DATA

4.1 Data Collection

Data collection for this study was conducted using the 2021 census database accessible through https://www.nomisweb.co.uk/sources/census_2021. The analysis focuses on deprivation, examined at a ward level for the The Greater Manchester and Merseyside area.

From the census data the "Households by deprivation dimensions" topic was chosen as a dependent variable, offering estimates that categorize households across four dimensions of deprivation. Following the literature, we select as independent variables data on individuals' main language, their highest level of qualification, age categorized by single years, general health assessments ranging from very good to very bad, and data pertaining to car or van availability within households.

4.2 Data Preparation and Preprocessing

All datasets sourced from the Census were cleansed and standardized using Excel. Then using Python we initiated data transformation by converting numeric variables that were wrongly labeled as objects into floats, making them compatible for subsequent analysis.

Each topic was its own dataset, so after transforming the date we merged all six datasets, using ward codes to combine the data into a single dataset, facilitating the comprehensive analysis.

To ensure there is consistency in scale throughout the dataset, we mainly worked with the percentage column of each variable.

We created three new columns, the first one aged 65 and above, by adding age categories spanning from 65 to 74 years, 75 to 84 years, and 85 years and over. Similarly, within the health data, a new column titled "h_bad_verybad" was generated, adding the "health bad" and "health very bad" columns, thereby providing a comprehensive health assessment metric. The last column introduced was area classification, a categorical that distinguish wards between Greater Manchester and Merseyside.

Then we created a dummy variable called "Most_Depr" indicating households deprived on all four dimensions and surpassing the average using the value column of deprivation level 4 from the original dataset, were 1 was for true and 0 for false.

Finally a choice was made about which variables within each topic would best help us understand the levels of deprivation those being:

Variable	**Description**	**Abbreviation**
No deprivation	Deprivation at no level	`dep_noneP`
Deprivation at 4 levels	Deprivation found in all levels	`dep_fourP`
Highest Deprivation	dep_fourP is greater than the average	`Most_Depr`
No qualifications	No qualifications obtained	`no_qualifP`
Level 4 qualifications	College leve qualifications obtained	`level4_qualifP`
Main language not English	Main language not English	`main_not_englishP`
Age over 65	Age over 65 years	`65_overP`
Health really bad	Bad and very bad Health	`h_bad_verybadP`
No vehicles owned	No cars or vans owned	`no_cars_vansP`
Area	Greater Manchester and Merseyside	`area`

Table 1: Description of Variables

4.3 Exploratory Data Analysis

4.3.1 Comparasion of the Areas

As a first step, in order to corroborate the comparability of the areas, we ran a pairplot matrix where we use area as hue. The distributions of each variable in the matrix supports our literature review about the comparability of the areas and how we can use them together for the analysis.

Figure 1: Pairplot Manchester and Meyseyside

4.3.2 Summary Statistics

Then we ran a summary statistics (Figure 2) for the variables within our reduced dataset, we can see from the difference between means that some of the distributions are skewed. Variables like "Deprivation level 4", "Level 4 qualification", "Main not English", and "No cars or vans" seem to have positive skewness. This suggests that observations tend to cluster towards lower values. Which is supported by the distributions seen in figure 1.

	dep_noneP	dep_fourP	Most_Depr	no_qualifP	level4_qualifP	main_not_englishP	65_overP	h_bad_verybadP	no_cars_vansP	area
count	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Greater Manchester
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	158
mean	45.215152	0.261953	0.383838	20.297643	31.397306	7.719529	17.400337	6.462290	27.787205	NaN
std	8.816838	0.190838	0.487140	6.379112	10.337105	7.429203	6.131975	1.898913	11.855899	NaN
min	23.500000	0.000000	0.000000	3.300000	16.600000	0.500000	1.800000	1.100000	8.300000	NaN
25%	38.600000	0.100000	0.000000	15.900000	23.700000	2.400000	13.500000	5.200000	18.800000	NaN
50%	44.900000	0.200000	0.000000	20.100000	28.900000	4.900000	17.400000	6.300000	25.800000	NaN
75%	51.000000	0.400000	1.000000	24.300000	36.700000	9.800000	21.200000	7.600000	35.300000	NaN
max	70.300000	0.900000	1.000000	36.500000	71.000000	33.800000	37.100000	12.600000	68.100000	NaN

Figure 2: Summary Statistics

4.3.3 Correlation and Pairplot

Then we created both a correlation and a pariplot matrix to have a comprehensive visual summary to identify, Linear Relationships, strength of correlations, outliers, distribution and potential Collinearity, but this time we used High Deprivation variable as hue.

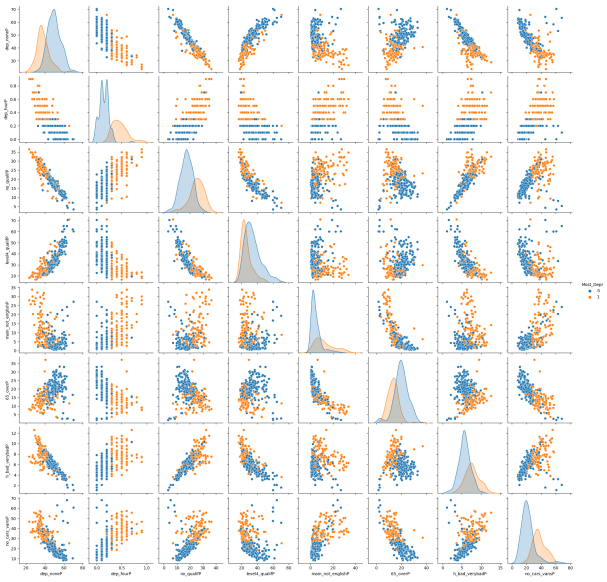

Figure 3: Paiplot of data using High Deprivation as hue

The new hue changes the distribution within some variables, like no qualification, by looking at the graph it can be inferred that the mean of having and not having high deprivation is not the same for the no qualification variable.

	dep_noneP	dep_fourP	Most_Depr	no_qualifP	level4_qualifP	main_not_englishP	65_overP	h_bad_verybadP	no_cars_vansP
dep_noneP	1.000000	-0.776070	-0.647532	-0.939607	0.831416	-0.435202	0.422665	-0.813606	-0.630270
dep_fourP	-0.776070	1.000000	0.793579	0.707112	-0.488233	0.604775	-0.557551	0.546901	0.671606
Most_Depr	-0.647532	0.793579	1.000000	0.591056	-0.377442	0.507610	-0.499822	0.508011	0.639445
no_qualifP	-0.939607	0.707112	0.591056	1.000000	-0.838928	0.338002	-0.321513	0.820827	0.517954
level4_qualifP	0.831416	-0.488233	-0.377442	-0.838928	1.000000	-0.010778	0.098695	-0.748409	-0.280403
main_not_englishP	-0.435202	0.604775	0.507610	0.338002	-0.010778	1.000000	-0.756217	0.001899	0.667195
65_overP	0.422665	-0.557551	-0.499822	-0.321513	0.098695	-0.756217	1.000000	-0.111458	-0.748001
h_bad_verybadP	-0.813606	0.546901	0.508011	0.820827	-0.748409	0.001899	-0.111458	1.000000	0.508266
no_cars_vansP	-0.630270	0.671606	0.639445	0.517954	-0.280403	0.667195	-0.748001	0.508266	1.000000

Figure 4: Correlation Matrix

We found a positive correlation between “No deprivation” and “Level 4 education qualification” suggests that areas with lower deprivation tend to have a higher proportion of individuals with level 4 education qualifications.

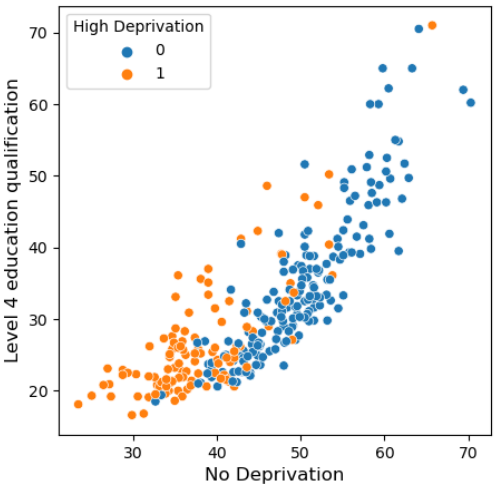

Figure 5: Paiplot of Level 4 education qualification using High Deprivation as hue

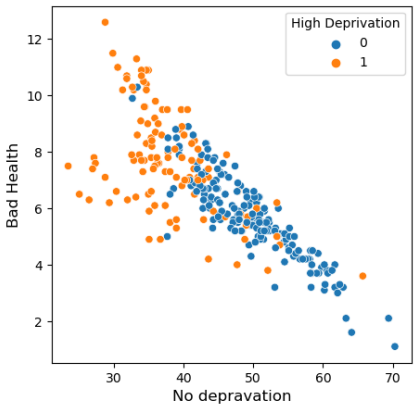

Figure 6: Paiplot of Bad Health using High Deprivation as hue

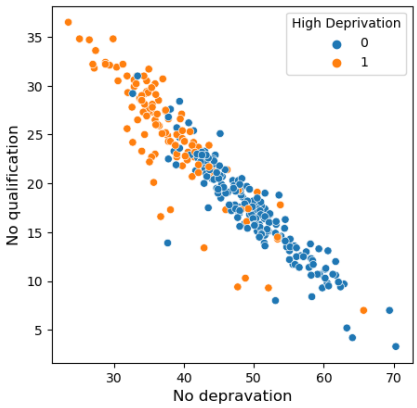

Figure 7: Paiplot of No qualification using High Deprivation as hue

Still Conversely, a strong negative correlation between “No deprivation” and ”No education qualification; “No deprivation” and ”Bad Health” and “No deprivation” and “Not having a vehicle” respectively indicates that areas with lower deprivation tend to have a lower proportion of individuals with no education qualifications, bad health and no cars.

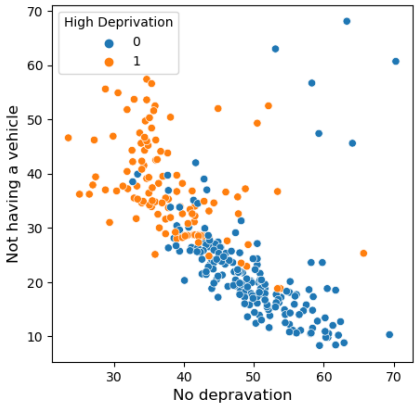

Figure 8: Paiplot of Not having a vehicle using High Deprivation as hue

While here we don't see a correlation, still from the graph one can inferred that in areas where there are a small percentage of older people and there is presence of bad health, there is a high chance that there is high deprivation.

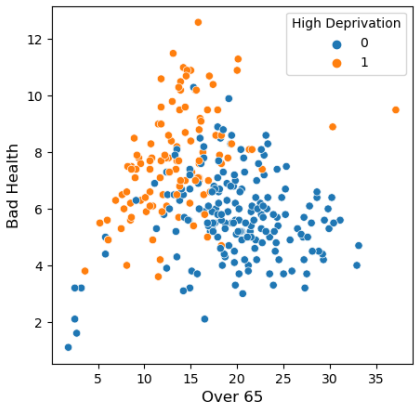

Figure 9: Paiplot of Bad Health and age over 65 using High Deprivation as hue

This pairplot confirms the strong impact that education has on depravation and the low impact of you English not being your main language.

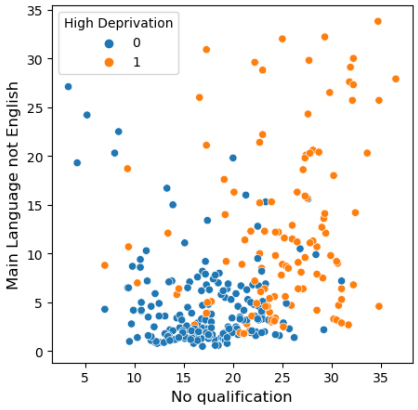

Figure 10: Paiplot of Main Language not English and No qualification using High Deprivation as hue

All of the graphs presented above suggest relationships that might be expected, but still we should think what lies behind the data.

Chapter 5

Results

5.1 Ordinary Least Squares (OLS) regression

1345006

According to the EDA part, we can assume that these variables follow the OLS regression, where 'dep_noneP' is the dependent variable and 'no_qualifP', 'level4_qualifP', 'main_not_englishP', ' 65_overP' , ' h_bad_verybadP' and 'no_cars_vansP' are independent varibles.

We will be estimating the following model:

$$\text{dep\_noneP} = \alpha + \beta_1 \cdot \text{no\_qualifP} + \beta_2 \cdot \text{level4\_qualifP} + \beta_3 \cdot \text{main\_not\_englishP} + \beta_4 \cdot \text{no\_cars\_vansP} + \beta_5 \cdot \text{h\_bad\_verybadP} + \beta_6 \cdot \text{65\_overP} + \phi$$

Using python to generate OLS regression, the outcomes as below:

OLS Regression Results

Dep. Variable:	dep_noneP	R-squared:	0.962
Model:	OLS	Adj. R-squared:	0.961
Method:	Least Squares	F-statistic:	1215.
Date:	Tue, 09 Apr 2024	Prob (F-statistic):	3.00e-202
Time:	22:55:04	Log-Likelihood:	-582.74
No. Observations:	297	AIC:	1179.
Df Residuals:	290	BIC:	1205.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	51.6228	1.469	35.130	0.000	48.731	54.515
no_qualifP	-0.1669	0.055	-3.013	0.003	-0.276	-0.058
level4_qualifP	0.3572	0.022	16.570	0.000	0.315	0.400
main_not_englishP	-0.4801	0.033	-14.656	0.000	-0.545	-0.416
65_overP	0.0565	0.031	1.806	0.072	-0.005	0.118
h_bad_verybadP	-1.9980	0.165	-12.073	0.000	-2.324	-1.672
no_cars_vansP	0.0503	0.022	2.306	0.022	0.007	0.093

Omnibus:	38.937	Durbin-Watson:	1.839
Prob(Omnibus):	0.000	Jarque-Bera (JB):	219.582
Skew:	0.280	Prob(JB):	2.08e-48
Kurtosis:	7.175	Cond. No.	745.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 11: OLS 1

We can find that the model exhibits an exceptional fit with an R-squared value of 0.962, suggesting that 96.2% of the variance in the dependent variable is explained by the model. Additionally, the adjusted R-squared is 0.961 which means that the model is unlikely to be overfitting.

For each independent variable, we can also know the extent of the impact of different independent variables on the dependent variable. For instance, the coefficient for no_qualifP is -0.1669, suggesting that an increase of one unit in this variable would decrease the expected value of dep_noneP by 0.1669 units. And its P-value is less than 0.05, indicating that this effect is statistically significant. Conversely, the coefficient of level4_qualifP (0.3572) suggests a positive expected influence on the dependent variable, which is also statistically significant, because the p-value of it is also less than 0.05.

Another two variables, 'main_not_englishP' and 'h_bad_verybadP', exhibit negative coefficients, also suggesting a negative correlation with the dependent variable and effects of them on dependent variable are very significant, because the P-values of them are much less than 0.05). The 65_overP has a very small coefficient (0.0565) and a P-value of 0.072, indicating a less certain impact on the dependent variable with weaker statistical significance. Lastly, the coefficient for 'no_cars_vansP' (0.0503) suggests a slight positive effect on the dependent variable, but with a P-value of 0.093 which is higher than 0.05, indicating that this effect is not significant.

The F-statistic of the model is high (1215) and, combined with an exceedingly low P-value which is close to zero, suggests that at least one predictor is significantly influencing the dependent variable. The Durbin-Watson statistic is 1.839, which is close to 2, shows no autocorrelation be-

tween residuals. The Jarque-Bera statistic indicates that residuals are not normally distributed because the Prob (JB) is close to zero, which might affect the efficiency of the coefficient estimates. However, considering the high R-squared value and statistically significant variables, the model provides a good explanation for dep_noneP and could have practical implications for understanding the actual impact of these variables.

Next, we can see the outcomes of the VIF of these variables:

```
const                211.319193
no_qualifP            12.176615
level4_qualifP         4.843750
main_not_englishP      5.776249
65_overP               3.586836
h_bad_verybadP         9.631916
no_cars_vansP          6.531092
dtype: float64
```

Figure 12: VIF table 1

We can find that the value for no_qualifP is 12 above the 10 threshold it is not over by much and the value for h_bad_verybadP is close to 10, and these may indicate multilinearity among them. Then, we can try to remove no_qualifP.

```
const                122.316382
level4_qualifP         2.611344
main_not_englishP      2.938628
65_overP               3.524768
h_bad_verybadP         4.529249
no_cars_vansP          5.386320
dtype: float64
```

Figure 13: VIF table 2

OLS Regression Results

Dep. Variable:	dep_noneP	R-squared:	0.961
Model:	OLS	Adj. R-squared:	0.960
Method:	Least Squares	F-statistic:	1417.
Date:	Thu, 11 Apr 2024	Prob (F-statistic):	6.59e-202
Time:	17:25:54	Log-Likelihood:	-587.32
No. Observations:	297	AIC:	1187.
Df Residuals:	291	BIC:	1209.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	48.7494	1.133	43.012	0.000	46.519	50.980
level4_qualifP	0.4013	0.016	25.008	0.000	0.370	0.433
main_not_englishP	-0.5493	0.024	-23.189	0.000	-0.596	-0.503
65_overP	0.0689	0.031	2.191	0.029	0.007	0.131
h_bad_verybadP	-2.3609	0.115	-20.521	0.000	-2.587	-2.134
no_cars_vansP	0.0779	0.020	3.875	0.000	0.038	0.117

Omnibus:	40.967	Durbin-Watson:	1.842
Prob(Omnibus):	0.000	Jarque-Bera (JB):	201.054
Skew:	0.394	Prob(JB):	2.20e-44
Kurtosis:	6.953	Cond. No.	523.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 14: OLS 2

After removing no_qualifP, we can find the VIF of these variables decrease. However, the R-square,adjusted R-square and coefficient do not change a lot, which means that the statistical significance of the model and fit of the model do not appear to be significantly affected by this. As a result, we can keep no_qualifP in the model.

5.2 BallMapper

10806957

To understand the high-dimensional data in the Ball Mapper visualisation, each ball/node represents a collection of wards that indicate similar characteristics, the edges suggest an overlap of these wards, and the ball colours indicate the range of values of these variables.

The TDABM graph shows the prediction of 0% deprivation indicators at ward-level (Figure 1). There is higher prevalence of no deprivation seen on the smaller balls, specifically ball 13 and 14, implying that the least deprived populations are scattered throughout cities. Also, the closeness of the connected balls to the right indicate that many wards are influenced by the similar variable. This is then compared to the true value (Figure 2). To observe the difference between observed and actual values, a ball mapper of residuals is shown in Figure 3. Comparing the outcomes, worth noting is that there are only a few nodes that differ in colour intensity where the model predicted higher no deprivation levels for smaller nodes like 6 and 14, and underpredicted low deprivations in certain balls (7, 9, 13 ,16). This suggests slight inaccuracies in under- and over-estimating no deprivation proportions.

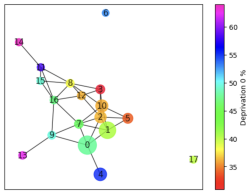

Figure 15: BallMapper Plots

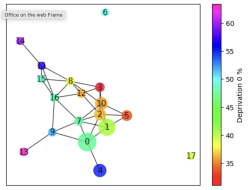

Figure 16: True value Graph

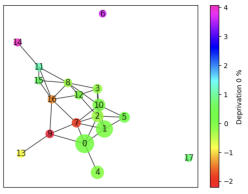

Figure 17: BallMapper of Residuals

5.2.1 No qualification variable

These proportions increase from left to right of the Figure 18 , demonstrated by the gradual colour intensity shift of the balls. The proportion of no qualification is highest in areas in ball 3 and lowest in ball 14.

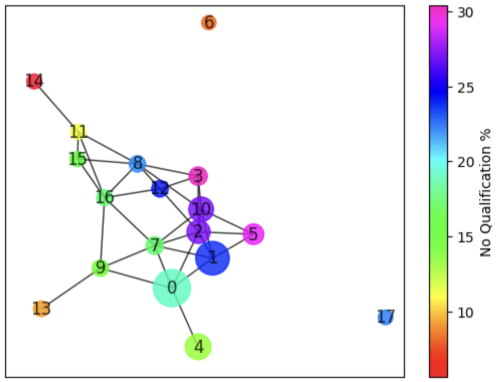

Figure 18: BallMapper plot of No Qualification Proportions

5.2.2 Level 4 qualification

Based on the colour intensity scale, this graph shows the opposite of the no qualification BallMapper (Figure 19), which is sensible assuming that the higher proportion of wards with no qualification would have a lower percentage of Level 4 qualifications. However, they are not directly comparable as the range of values for Level 4 qualification are much higher, implying that the proportions of having this qualification at ward-level are much bigger in the nodes than no qualifications.

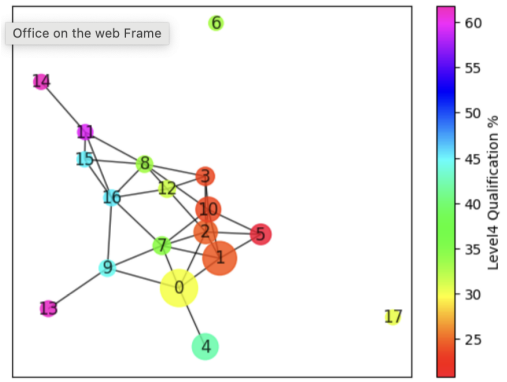

Figure 19: BallMapper plot of Level 4 Qualification

5.2.3 Main language not English

Based on the colour intensity scale, this graph shows the opposite of the no qualification BallMapper (Figure 20), which is sensible assuming that the higher proportion of wards with no qualification would have a lower percentage of Level 4 qualifications. However, they are not directly comparable as the range of values for Level 4 qualification are much higher, implying

that the proportions of having this qualification at ward-level are much bigger in the nodes than no qualifications.

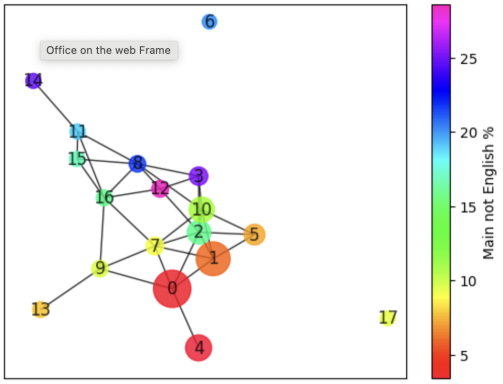

Figure 20: BallMapper plot of Main Language not English

5.2.4 Ages over 65

Notably, the colour intensity of the nodes increases from bottom-up, with the isolated nodes 6 and 17 being the most intense (Figure 21). This suggests that the outlier clusters have a significantly higher proportion of above 65s. Ball clusters most well-connected and densely grouped have an average percentage of above 65s.

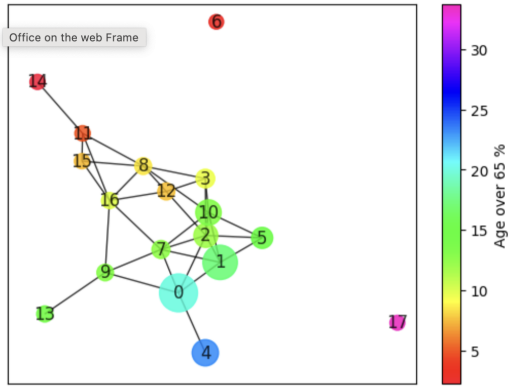

Figure 21: BallMapper plot of Ages over 65

5.2.5 Bad/Very Bad Health

The range of proportions for having poor health are relatively low (under 10%), but are on the higher end of the spectrum for the balls on the right side, specifically ball 5, followed by balls 17 and 10. This implies that areas that fall under these clusters reported poorer health compared to the rest of the areas which, based on the colour intensity, reported an average-low percentage

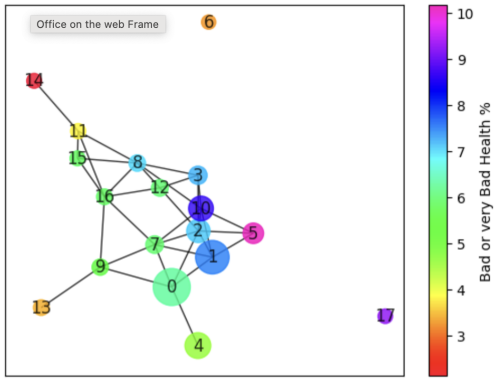

Figure 22: BallMapper plot of Poor Health

5.2.6 Car/van ownership

The percentage of populations who do not own these vehicles increase from the bottom up (Figure 23) - a pattern similarly shown in the graph of those whose main language is not English. The largest ball 0 particularly showed a lower intensity colour, indicating that the many collections of wards here are people who do not own these vehicles.

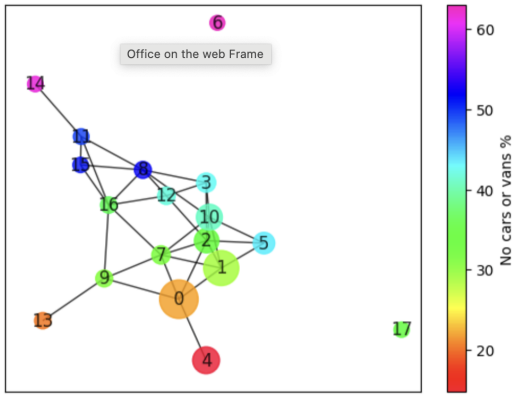

Figure 23: BallMapper plot of No Cars/vans

5.2.7 TDABM analysis to the ward-level map file enables

Summary statistics are calculated to describe the data points and how its average values potentially influence the deprivation levels. The output shows that on average, those with 0% qualifications were most strongly and negatively associated with no deprivation levels. Conversely, those with Level 4 qualifications had the strongest positive correlation with no deprivation. Observing the BallMapper, the same collection of wards with higher level 4 qualifications also had higher 0% deprivation levels, confirming the relationship shown in descriptives. The

deprivation levels explained by the language variable is moderate compared to others, though it shows a positive relationship. The age variable does not significantly explain deprivation, but would give a relationship where the higher the proportion of those aged above 65, the lower the deprivation levels.

Merging the TDABM analysis to the ward-level map file enables further understanding of the context of our results (Figure 24). Figure 25 shows how the balls are mapped.

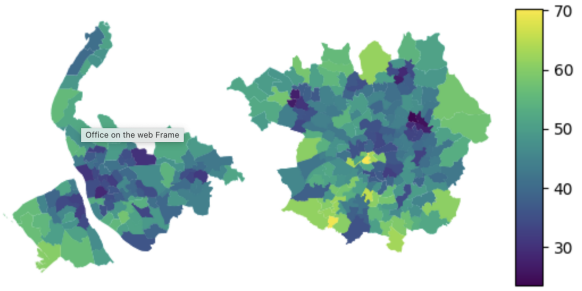

Figure 24: Mapping of 0% deprivation levels across wards

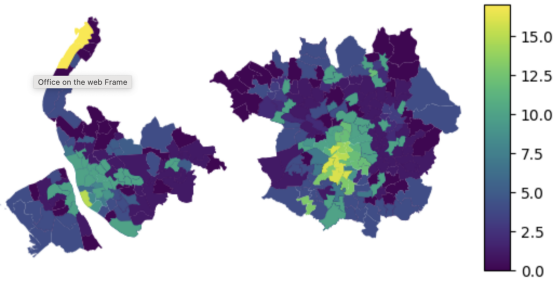

Figure 25: Mapping of deprivation levels by balls

Ball 17 is highlighted as it shows a prominently value whereby the higher deprivation levels are explained mostly by the variables (Figure 26).

Figure 26: Ball 17 Mapping

Chapter 6

Robustness

6.1 Radius assesment

11395438

This robustness assessment aims to validate the selection of radius ($\varepsilon = 15$).for the TDABM and assess its effectiveness in capturing the structure of the data. The choice of radius, , alters the coverage of the data points within each ball and the connections between balls, leading to variations in visualization and interpretation of the data. Smaller radii offer detailed insights into local phenomena but risk excess noise, whilst larger radii potentially oversimplify the data. By comparing the results obtained at four additional radius ($\varepsilon = 10, 13, 17, 20$), insights are gained into the impact of changing radii, focusing on consistency in the number of connections and coloration of balls, thereby validating the chosen radius against alternative options. Despite the absence of an algorithm to determine the optimal ball radius, this sensitivity to changes in radius serves as a robust test of the model's performance.

The assessment identifies the optimal radius as 15, which aligns with the radius used in our analysis. A radius of 10 includes excessive noise, with the balls pushed out into the circular, disconnected ring whilst a radius of 20 oversimplifies the structure with fewer balls all connected. Radius of 13 demonstrates a partial reduction of noise, balls 25, 6, 16, and 20 still create a partial outer ring whilst radius of 17 connects all the balls, including potential outliers. Radius 15 consistently demonstrates a balanced performance, effectively capturing the underlying topology of the dataset whilst minimising noiseas evident in balls 17 and 6, which remain disconnected

across each variable (see Appendix). These findings enhance the reliability of our analysis by providing confidence in the identified patterns and relationships within our results.

Overall, the model appears fairly consistent across the changing radii, but there is some concern regarding outliers. The analysis of ball 0 and 1 exemplifies a level of consistency across various radii. Ball 0 exhibits a consistent coloration, transitioning slightly from its initial blue-green colour in ε =10 to a light blue in ε =20. Similarly, ball 1 maintains a stable coloration, with a slight darkening of its blue colour observed across radii, with the slight change indicating subtle variations in cluster density, connectivity, and boundary effects across different radii. Both balls exhibit stable connectivity, remaining consistently connected across all radii, and maintain a relatively constant size compared to other balls in the dataset. This consistency indicates that the model can reliably capture and represent the underlying structure of the data, resulting in stable clusters with consistent characteristics. However, ball 6 exhibits notable inconsistencies in coloration and connectivity, suggesting its outlier status. Initially light blue and connected at ε =10, it shifts to yellow and orange when disconnected at ε=13 and ε=15, then transitions to pink and purple when reintegrated at ε=17 and ε=20, suggesting variability in its data structure. Outliers can significantly impact the performance and interpretation of clustering algorithms, challenging the model's ability to accurately identify and characterise clusters, potentially compromising its robustness.

Overall, the robustness assessment suggests that the model demonstrates a level of consistency, particularly in the behavior of balls 0 and 1 across different radius. Additionally, the decision to base the results on the optimal radius (ε = 15) further reinforces the model's consistency, as it showcases a deliberate effort to select parameters that yield reliable and interpretable results. However, the fluctuations observed in ball 6 raise concerns about the model's consistency, potentially due to its outlier status. Overall, while the assessment highlights certain areas of concern, such as outlier handling, the model demonstrates a degree of consistency in capturing the underlying structure of the data.

Health Variable Ball mapper across different radii:Using variable Very bad Health

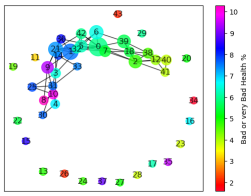

Figure 27: $\varepsilon = 10$

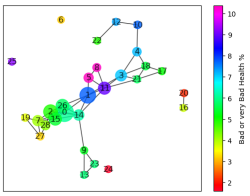

Figure 28: $\varepsilon = 17$

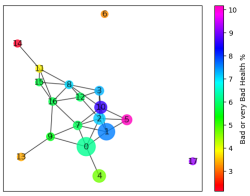

Figure 29: $\varepsilon = 15$

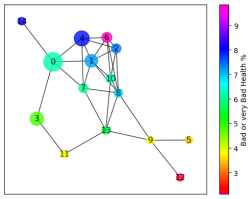

Figure 30: $\varepsilon = 17$

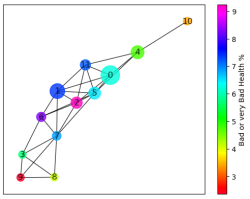

Figure 31: $\varepsilon = 20$

6.2 Changing an explanatory variable assesment

10806957 Another robustness check was analysed by changing an explanatory variable from language to the ethnicity variable. This specific variable was selected as it can be assumed that the relationship between language and ethnicity is the strongest, demonstrated in the correlation analysis. First, the percentage of all ethnic minority categories are merged into one before changing it with the proportion of those whose English is not their main language.

The updated variable visualises a different analysis on the BallMapper. The distribution of the multivariate BallMapper plots in the updated model shows a higher number of clusters, higher connections, and denser balls in the middle at 0% deprivation levels, compared to having English not as the main language (Figure 32). This demonstrates that ethnicity has a different relationship with 0% deprivation levels than having English not as the main language, whereby the former has better explanatory power. When compared to its true value and residuals (Figure 33), the adjusted ethnicity variable showed better performance than the language model. Observed in Figure 34, the lower intensity colours in bigger balls of the adjusted variable suggests lower proportion of ethnic minorities in bigger collections of wards - similar to the pattern seen in language where lower proportions of non-English are observed in bigger-sized nodes.

Robustness Check Graphs for Changing Variables:

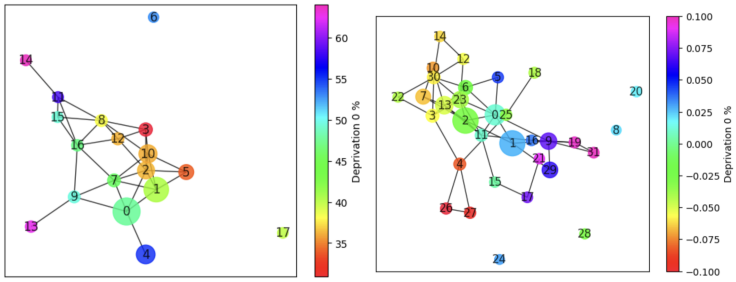

Figure 32: Multivariate BallMapper for Language (left) vs Ethnic Minority (right)

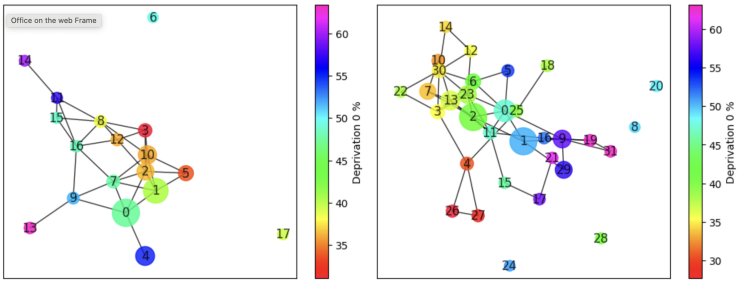

Figure 33: True Values for Language vs Ethnic Minority

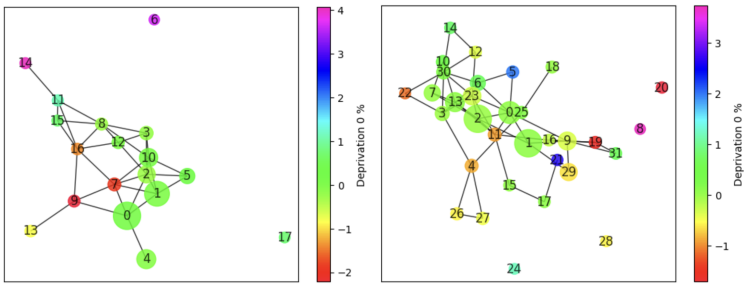

Figure 34: Residuals plot for Language vs Ethnic Minority

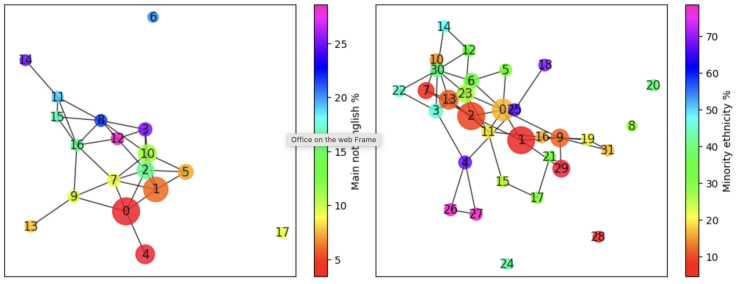

Figure 35: TDABM for Language vs Ethnic Minority

Overall, robustness checks indicate that changing the variable does, indeed, influence the inferences made from the TDABM analysis. Although literature points out the language barrier factor in predicting deprivation levels, our model suggests that being an ethnic minority explains

deprivation levels more although the two are highly positively correlated. This suggests a lack of robustness in the model.

Chapter 7

Conclusion

11345006

In conclusion, this report presents a detailed analysis of deprivation levels among the wards of Liverpool and Manchester using data from the Census 2021. It integrates classical statistical methods with Topological Data Analysis (TDA) to explore and visualize the factors of deprivation.

The report confirms that deprivation is a complex issue with multiple contributing factors, including education, health, employment, language proficiency, and access to transport. The OLS regression models employed suggest that low education level (no_qualifP), poor health (h_bad_verybadP), and lack of car or van access (no_cars_vansP) significantly contribute to higher deprivation levels. While the level of English proficiency and age over 65 (65_overP) presenting weaker but noteworthy associations. The high R-squared value reinforces the robustness of the OLS regression model in capturing the variance in the dependent variable dep_noneP.

However, the Variance Inflation Factor (VIF) analysis raises concerns about multicollinearity, particularly for the no_qualifP. This means that the coefficient of this variable might influence the robustness of the model. Despite this, there is no evidence that it makes our model unstable or unreliable.

The study uses TDABM to visualize data relationships. TDABM is a useful tool to help us understand the relationship of data in different areas and reflects the wards of same features on geography map. The choice of radius in TDABM is crucial, with the study finding a balanced representation of the data at a radius of 15. This strikes a compromise between too much detail,

which may introduce noise, and oversimplification, which could overlook important subtleties.

The findings of this report can be used for policymaking. They highlight the need for targeted interventions to address the root causes of deprivation. For instance, policies aimed at increasing educational attainment, improving health outcomes, and enhancing transport connectivity could be effective. Additionally, the data of language proficiency suggests that policies to improve English skills among non-native speakers could help reduce socioeconomic disparities. Moreover, policymakers should consider the localized conditions of deprivation, as the data visualization tells us that the relationship among these factors can vary significantly across different wards and regions.

In addition, policymakers can also use similar topological data analysis methods to assess the impact of proposed interventions before their implementation. By visualizing the potential outcomes, they can ensure that resources are allocated efficiently and make different programs to the unique characteristics of each deprived area.

In summary, the study illustrates the complex interactions between various deprivation factors in Liverpool and Manchester. It shows the necessity of using topological methods to analyze the data of a large area, since data visualization can help policy makers understand the different conditions of each region. As cities continue to evolve, ongoing research using topological methods like TDABM will be critical in informing dynamic and responsive urban policy frameworks

Bibliography

1. Abbasi, A., Alalouch, C., & Bramley, G. (2016). Open space quality in deprived urban areas: user perspective and use pattern. *Procedia-Social and Behavioral Sciences*, 216, 194–205.

2. Aoki, Y., & Santiago, L. (2018). Deprivation, Segregation, and Socioeconomic Class of UK Immigrants: Does English Proficiency Matter?

3. Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17–21.

4. Brown, J. (2009). Liverpool betrayed: From post-war to Pathfinder. Triumph, Disaster and Decay–The SAVE Survey of Liverpool's Heritage, SAVE Britain's Heritage, London, 23–27.

5. Charlton, J., Rudisill, C., Bhattarai, N., & Gulliford, M. (2013). Impact of deprivation on occurrence, outcomes and health care costs of people with multiple morbidity. *Journal of Health Services Research & Policy*, 18(4), 215–223.

6. Dłotko, P. (2019a). Ball Mapper: a shape summary for topological data analysis. arXiv preprint arXiv:1901.07410.

7. Dlotko, P., Qiu, W., & Rudkin, S. (2022). Topological data analysis ball mapper for finance. arXiv preprint arXiv:2206.03622.

8. Ilie, S., Rose, P., & Vignoles, A. (2021). Understanding higher education access: Inequalities and early learning in low and lower-middle-income countries. *British Educational Research Journal*, 47(5), 1237–1258.

9. Kelly, A., Croasdale, S., Taylor, S., & Marshall, K. (2023). Putting communities and citizens at the heart of what we do - The development of a continuum of integration within a locality in the Northwest of the UK to enhance health well-being and reduce inequalities. *International Journal of Integrated Care*, 23(1), 171.

10. Margalit, Y. (2019). Economic Insecurity and the Causes of Populism, Reconsidered. *Journal of Economic Perspectives*, 33(4), 152–170.

11. Mattioli, G. (2021). Transport poverty and car dependence: A European perspective. In *Advances in Transport Policy and Planning* (Vol. 8, pp. 101–133). Academic Press.

12. McCartney, G., Popham, F., Katikireddi, S. V., Walsh, D., & Schofield, L. (2017). How do trends in mortality inequalities by deprivation and education in Scotland and England & Wales compare? A repeat cross-sectional study. *BMJ Open*, 7(7), e017590.

13. Robson, B., Lymperopoulou, K., & Rae, A. (2008). People on the move: exploring the functional roles of deprived neighbourhoods. *Environment and Planning A*, 40(11), 2693–2714.

14. Rudkin, S., & Webber, D. (2023). Regional growth paths and regional resilience. *SSRN Electronic Journal [Preprint]*. doi:10.2139/ssrn.4333276.

15. Rudkin, S., et al. (2023). An economic topology of the Brexit vote. *Regional Studies*, 58(3), 601–618. doi:10.1080/00343404.2023.2204123.

16. Schulze Bäing, A., & Wong, C. (2012). Brownfield residential development: what happens to the most deprived neighbourhoods in England? *Urban Studies*, 49(14), 2989–3008.

17. Schwanen, T., Lucas, K., Akyelken, N., Solsona, D. C., Carrasco, J. A., & Neutens, T. (2015). Rethinking the links between social exclusion and transport disadvantage through the lens of social capital. *Transportation Research Part A: Policy and Practice*, 74, 123–135.

18. Sykes, O., Brown, J., Cocks, M., Shaw, D., & Couch, C. (2013). A city profile of Liverpool. *Cities*, 35, 299–318.

19. Walsh, D., Bendel, N., Jones, R., & Hanlon, P. (2010). It's not 'just deprivation': Why do equally deprived UK cities experience different health outcomes? *Public Health*, 124(9), 487–495.