New Horizons: Appendix A

Features of the local environment and educational attainment - methodology note, October 2022

This appendix sets out the methods and sources used to conduct the quantitative analysis as featured in CPP's report; New Horizons – Transforming educational opportunities to support inclusive growth. To explore the relationship between places' physical and economic characteristics and educational attainment, CPP have analysed a range of local authority level data. Due to the numerous variables of interest, we conducted a principal component analysis (PCA) to consolidate our variables into and reduce the scope for correlation between them. We then used a multivariate regression model to assess the impact of place characteristics on the educational attainment (KS4 results) of disadvantaged children, aggregated at the local authority (LA) level.

Purpose of the model

The purpose of our model is to explore the extent to which a relationship exists between the environmental characteristics of a place, and the educational attainment of disadvantaged young people that live there. The presence and nature of the relationship found, informs the theoretical framework as set out in the report, which in turn guides the qualitative interviews being undertaken.

Geography

This analysis is done at lower tier local authority level, reflecting our aim to compare and contrast the relative differences between places and the broader environments that shape them. Several datapoints that we use within this analysis are not available at a more granular geographic level (e.g. LSOA, MSOA), particularly relating to our labour market variables this is the level.

Our analysis will not address the variance between the schools and individuals within each LA. Yet given our research is concerned with how the wider social determinants of education differ across places, an LA-level model remains appropriate for this analysis. .

Unfortunately, not all of the variables we use in this analysis were available for all authorities due to a combination of missing data, and local authority mergers, and so 20 observations were dropped from our data, with our final sample consisting of 292 local authorities.[1] Our research also only covers England, given differences in the availability of some of our datapoints across the devolved nations.

Step 1: Principal Component Analysis

Given the large quantity of different indicators that may be indicative of the physical and economic environments of a place, the application of statistical methods that categorise indicators into related groups based upon overlaps in the data can be helpful in reducing the total number of indicators, while retaining information about the most relevant indicators. To this end, we applied principal component analysis (PCA) to our dataset, a statistical process that reduces the complexity of a dataset by organising indicators into a reduced number of unique components based upon the correlation of different indicators.

For example, the conditions of a local area's labour market cannot realistically be quantified with just one metric, whether it be earnings, employment rate, or the job quality – all of these factors are reflective of the strength of a local labour market. Yet through the PCA, we can reduce this number of variables into a singular principal component, which enables us to capture the effects of labour market conditions on attainment more broadly.

Given that components are based upon the identification of high levels of correlation between different indicators, they also provide a theoretical premise to which distinct themes that reflect the different characteristics of local areas can be drawn out.

To compute the PCA, we broadly follow the series of steps set out by the European Commission's "Composite Indicators Research Group"[2]. To summarise, these include:

1. **Selecting the variables**: We selected variables that broadly reflect the differing physical and economic environments between places. For our final model we compiled several of the English Indices of Multiple Deprivation (IMD) subcomponents, local labour market data from NOMIS, and data from other sources as listed in *Table* 1, for all lower-tier local authorities in England.
1. **Standardisation:** The variables use many different units of measurement. All variables are therefore standardised by subtracting the mean and dividing by the standard deviation.
2. **Multivariate analysis**: We then conducted the PCA. This grouped together variables that are strongly associated with each other, while rarely associated with others, so that the final groups refer to different local authority characteristics.
3. **Normalisation:** Since the values derived through PCA are hard to interpret, we rank each local authority based on its scores between 0 and 1.

The PCA process returned three separate components, which can theoretically be interpreted as:

1. High spatial proximity to key services and educational institutions
2. Strong labour markets (high wages, good jobs, high employment rate)

[1] The omitted local authorities are: Barking and Dagenham; Bournemouth, Christchurch and Pool; Buckinghamshire; St Albans; Stevenage; Welwyn Hatfield; West Suffolk; Somerset West and Taunton; East Hertfordshire; Gateshead; Dorset; East Suffolk; Northumberland; Kettering; East Northamptonshire; South Northamptonshire; Wellingborough; Daventry; Northampton; Corby

[2] Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). Tools for Composite Indicators Building: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC31473/EUR%2021682%20EN.pdf

3. Good housing conditions and access to employment opportunities (high proximity to employment centres and high job density)

Table 1: Principal Component Indicator List

Principal Component	**List of indicators contained within**	**Status within our final model**
1. High spatial proximity to key services and educational institutions	IMD Geographical Barriers (proximity to key services) Average number of FE institutions within a 45 minute commute, public transport/walking (Department for Transport data) Average number of secondary schools within a 45 minute commute, public transport/walking (Department for Transport data) Average number of median sized employment centres within a 45 minute commute, public transport/walking (Department for Transport data) IMD Income Affecting Children	Omitted from final model as not a significant predictor of average attainment for disadvantaged pupils
2. Strong labour markets (high wages, good jobs, high employment rate)	Economic Activity Rate (NOMIS data) Low Pay - % of jobs per local area that are low-paying (Health Foundation data) IMD Income Affecting Children Median Earnings (NOMIS data)	Included in final model
3. Good housing conditions and access to employment opportunities (high proximity to employment centres and high job density)	IMD Barriers to Services IMD Indoor Living Environment Average number of median sized employment centres within a 45 minute commute, public transport/walking (Department for Transport data) Job Density (NOMIS data)	Included in final model

Additionally, there were several variables that we tested within our PCA model that did not demonstrate any strong correlation with our component variables. These variables can be found in *Table 2*.

Table 2: Additional Tested PCA Indicators

Indicator Name

Access to greenspace (average amount of public greenspace available within a 1km radius, per local area IMD Crime KS2 Obesity Rate Average child hospital admissions relating to mental health per 1000 population, per local area

Step 2: Multivariate Regression

Following the PCA, we built a multivariate regression model to explore how our components impact the educational attainment of disadvantaged children. We tested multiple model variations containing both our principal components and a selection of demographic and regional control variables to identify the most robust model variation. For a full list of control variables see Table 4. Given our methodology, it is important to note that the relationships we show are purely descriptive; however, they provide an indication of the role that the physical and economic environments of a place may have on the educational attainment of disadvantaged students, at the local authority levelThe final model output can be seen below in *Table 3*.

Table 3: **Linear regression output**

KS4 Attainment (Disadvantaged Children)	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
Labour Market	2.045	.6	3.41	.001	.863	3.226	***
Housing and Jobs Proximity	1.732	.589	2.94	.004	.572	2.892	***
KS2 Attainment (Disadvantaged)	.799	.212	3.77	0	.382	1.215	***
Non UK-Born	.837	.242	3.46	.001	.361	1.314	***
London	3.28	.852	3.85	0	1.602	4.958	***
North West	.307	.48	0.64	.523	-.638	1.253	
Constant	36.988	.486	76.10	0	36.031	37.945	***
Mean dependent var		39.256	SD dependent var			3.743	
R-squared		0.525	Number of obs			292	
F-test		52.529	Prob > F			0.000	
Akaike crit. (AIC)		1395.067	Bayesian crit. (BIC)			1420.804	

**** p<.01, ** p<.05, * p<.1*

Note: the dependent variable used for our model is KS4 Attainment rate (Attainment 8) of disadvantaged students, local authority average

Assumptions

By conducting our analysis at local authority level, we assume that variation in place-based characteristics can meaningfully explain differences in the average attainment of disadvantaged pupils across local authority areas and that this relationship can be described using a linear model. Our analysis will not capture differences between smaller geographies (e.g. LSOA, MSOA), between schools, or between individuals.

Our model treats our variables as continuous and considers the relative difference between local authorities through rankings, therefore we assume that spacing between the rankings of each local authority are equal. Given also that the 292 local authorities that make it into our final model represents the overwhelming majority of people and places England, we also assume that our findings are reflective of trends that are relevant to the nation at large.

Interpretation

The relationships shown in our model are descriptive and do not prove causation; however, they provide an indication of the role that the physical and economic environments of a place may have on the educational attainment of disadvantaged students, at the local authority level.

The coefficients within our regression model show the relationship between our independent place variables and average Attainment 8 scores for disadvantaged pupils. Given that our PCA variables are ranked between 0 and 1, an increase of 1 unit can be interpreted as the difference between the highest and lowest ranking local authorities.

Attainment 8

Attainment 8 scores are a measure of academic performance. They are calculated by adding together a pupil's GCSE grades on the new 1-9 grading system (where 1 represents the lowest grade, and 9 the highest), across eight government approved subjects. The eight subjects are English and Maths, the top three scores from English Baccalaureate (EBacc) subjects, and the top three scores from remaining EBacc subjects or other qualifications. English and Maths are also double weighted for the overall score, whereas for the remaining 6 subjects the GCSE grade, between 1-9, is taken on its own.[3] For the dataset that we use, Attainment 8 scores are representative of the average of all secondary schools within a local authority.

Control Variables

In addition to our principal components, we tested several demographic variables that, according to the literature, may account for some of the differences in attainment rates between places. Controlling for demographic determinants of educational success at place level was challenging and many variables

[3] For a further explanation on Attainment 8 scores, please see the following: https://www.locrating.com/Blog/attainment-8-and-progress-8-explained.aspx

subject to high levels of collinearity and therefore not included in the final model. See Table 4 for full details.

We also sought to control for prior attainment (KS2 attainment) and included several regional dummy variables.

Table 4: Control Variable List

Variable Name	**Description**	**Status within our final model**	**Reason for status**
KS2 Attainment	The average attainment of disadvantaged students per local authority, taken from the year that our cohort took their KS2 exams	Included in the final model	Appropriate fit for the final model
Degree coverage	% of population with an NVQ4+ qualification	Omitted	Insignificance due to collinearity
Median age	The median age of the local population	Omitted	Insignificance
Population size	The size of the local population	Omitted	Insignificance
White British %	The % of the local population who are white British	Omitted	Insignificance due to collinearity
Non UK Born	The % of the local population who are non-UK born	Included in the final model	Appropriate fit for the final model
Population Density	The density of a local population, which considers the size of a population relative to its area size	Omitted	Insignificance due to collinearity

Regional dummies	Dummy variables used to quantify the effect that living within a specific region has on attainment	London and North West dummies included	London dummy included due to significance and large coefficient, whereas other dummy variables for each region were insignificant. North West dummy included due to research question and focus being concerned with this particular region
Free school meal coverage	The % of children in receipt of free school meals per local area	Omitted	Insignificance due to collinearity