INEQUALITY IN MORTALITY FROM RESPIRATORY DISEASE IN ENGLAND

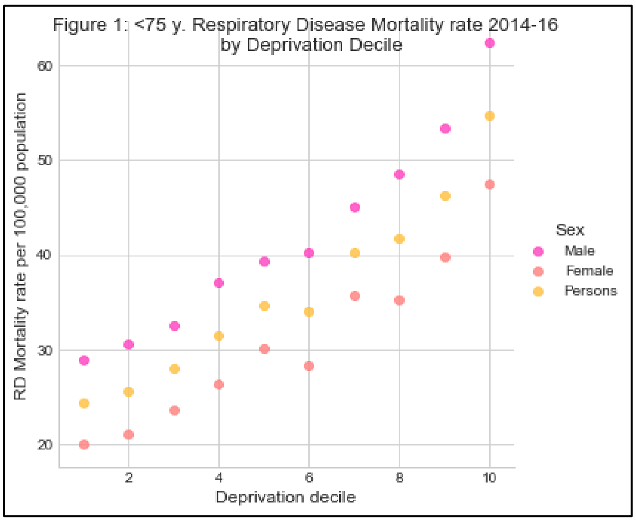
AN EXPLORATORY ANALYSIS OF PUBLIC HEALTH AND DEPRIVATION DATA

Veronica Tuffrey

Report submitted in partial fulfilment[1] of the requirements for the module
"INM430 Principles of Data Science"
MSc Data Science - City, University of London
December 2018

1. INTRODUCTION

Global patterns of disease have changed over the past few decades. Chronic non-communicable diseases (NCDs), including cancer, heart disease, diabetes, obesity and respiratory disease have emerged and greatly increased in prevalence. NCDs cost lives and place an economic burden on individuals and society.



There are pronounced health inequalities in England between the least and most deprived areas, in mortality and morbidity due to chronic diseases [1] e.g. see Figure 1. The economic cost to the NHS of behavioural risk factors for chronic disease is huge. Poor diet is the factor with highest impact on the NHS budget, followed by alcohol consumption, tobacco smoking and physical inactivity [2]. Such risk factors can be unevenly distributed between socioeconomic groups, and so may be important determinants of health inequalities [3].

While mortality rates have declined in England, recent NHS data show an increase in socio-economic inequality in mortality from respiratory disease [4], in contrast to other chronic diseases. It is important to examine what is driving the increase, to feed into policy and interventions.

1.1 Analysis Domain: Health and socio-economic.

1.2 Research Questions

- What are the patterns of variation in <75 y. mortality rates of respiratory disease in England with respect to deprivation, behavioural risk factors and pollution?
- How do the patterns for respiratory disease mortality differ from mortality patterns of cancer and cardiovascular disease?

[1] The other required item to be submitted is a Jupyter notebook, accessible at https://smcse.city.ac.uk/student/aczg664/VT_PDS%20coursework_161218.html

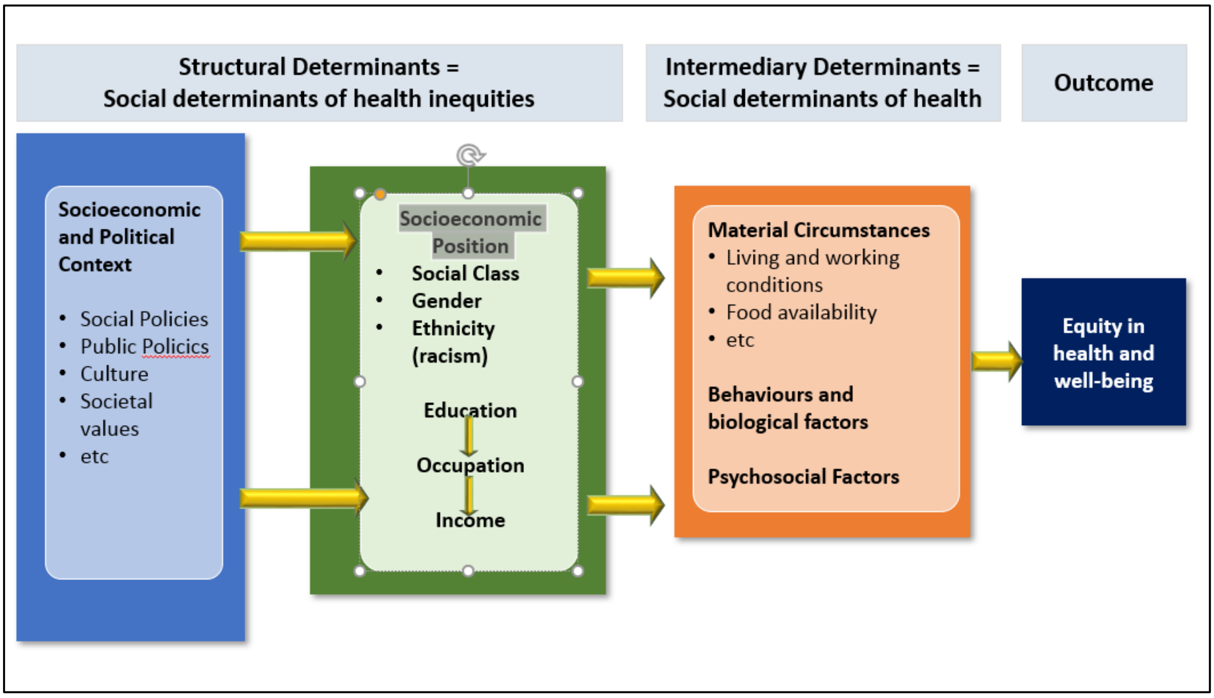


Figure 2: Conceptual Framework for Analysis, simplified from Solar and Irvin, 2010 [3]

It is difficult to disentangle the effects of individual determinants on health, since they work through causal chains, and interact with each other. Figure 2 shows the framework underlying analysis in this study, whereby social determinants of health inequalities impact on behavioural factors, which in turn impact on health outcomes[2].

1.3 Analysis Strategy

All data analysis was undertaken using the Python programming language (Python Software Foundation https://www.python.org/) while mapping used Tableau (https://www.tableau.com/).

1.3.1 Sourcing of suitable datasets (Please refer to Appendix 1 for details of each variable)

DEPENDENT VARIABLES

<75 mortality rates for respiratory disease, CVD and cancer: Data were obtained from the Public Health Outcomes Framework (PHOF) website https://fingertips.phe.org.uk/profile/public-health-outcomes-framework

PHOF are indicators designed to help stakeholders, especially local authorities, to understand how well public health is being improved and protected. Three year smoothed averages in < 75 y. mortality rates from four chronic diseases are available, with the most recent data from 2014-16.

INDEPENDENT VARIABLES

Lifestyle factors: Data on the behavioural factors of diet, smoking, and physical activity were obtained from the same source as the dependent variables (PHOF).

[2] Feedback in the reverse direction between elements of this framework has been omitted for simplicity. The leftmost element has been left in, because findings from studies such as this can feed into policy.

Deprivation indices: Data on relative deprivation are provided by the Department for Communities and Local Government and were obtained from the following website: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

The Indices of Deprivation 2015 are relative measures of deprivation for small areas (Lower-layer Super Output Areas) across England, based on seven domains covering a range of economic, social and housing issues. The domains are Income Deprivation; Employment Deprivation; Education, Skills and Training Deprivation; Health Deprivation and Disability; Crime; Barriers to Housing and Services, and Living Environment Deprivation. The domains are combined to produce one overall "Index of Multiple Deprivation". The most recent data are from 2015 and relate to tax year 2012/13 [5].

Averaged data across local authority areas on the IMD, the seven domains, and the supplementary index, "Income Deprivation Affecting Older People" were downloaded.

Pollution Data: Data on "Fraction of Mortality Attributable to Pollution" was obtained from PHOF[3].

Geographical data: The geographical data callable online from within the visualisation software were used, from https://www.tableaumapping.bi/wdc/

1.3.2 Extraction and combination of variables

Datasets were merged, variable names edited, and datapoints with missing values deleted.

1.3.3 Production of descriptive statistics and manipulation of datasets to normalise

Descriptive statistics and shape of variables' distributions were checked, and variables transformed if necessary.

1.3.4 Simple bivariate exploratory analysis

Scatterplot matrices were used to examine cross-sectional bivariate associations between variables

1.3.5 Data derivation

Principal Components Analysis analysis (via the "SciKit Learn" package of Python) was used to derive

- Three new variables from the exposure variables using data from 2015/6
- Three new deprivation variables from the seven deprivation domains

1.3.6 Model building

Separate models were developed for each of the three mortality outcome variables using stepwise multiple regression. Two approaches "Enter" and "Forward Stepwise" were used, with varied combination of predictor variables available for selection. Finally, residuals from regression models with the PCA deprivation component as predictors were used as outcome variables, to examine which lifestyle variables best explained departures from the expected variation due to deprivation. Code developed by Jordan Crouser was used for the regression analysis[4]. Adjusted R squared values were used to compare the effectiveness of the models together with the error criteria Omnibus, AIC and BIC. QQ plots were derived to evaluate if assumptions underlying the method were met.

[3] Ideally data would have been obtained from the Department for Environment, Food and Rural Affairs (https://uk-air.defra.gov.uk/), but their data had a higher level of geographical granularity than the other datasets being used in this project.

[4] http://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-py.html

2 ANALYTICAL PROCESS

2.1 Data processing (wrangling/merging)

In order to prepare a useful and robust data set to work with the following steps were taken -

- Datasets were merged by matching on the code for local authority area
- Names of areas (e.g. Bristol, City of) with commas were edited to avoid problems reading files
- Data for two local authority areas were missing in some datasets. Because the number of missing values was so small, the areas (Isles of Scilly and City of London) were deleted rather than imputing values for them
- Shape of variables' distributions were checked using histograms (see Figure 3 for an example) and QQ plots, and variables were transformed if necessary.

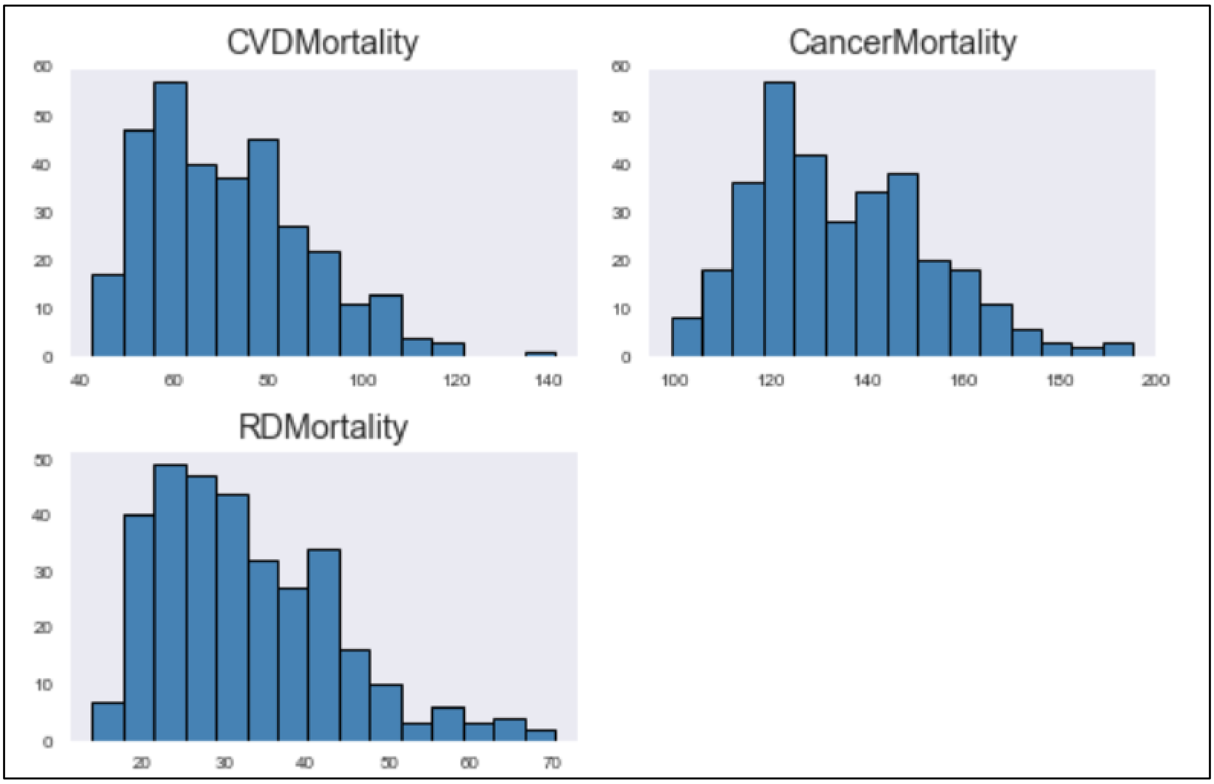


Figure 3: Histograms of Outcome variables, showing positive skewed shape of distributions

  - Respiratory Disease Mortality, CVD Mortality; Cancer Mortality; Employment; IDAOPI; IMD; Income, and Living_Environment were transformed using the logarithmic distribution to correct for positive skew
  - Percent Inactive was transformed using the square root transformation to correct for weaker positive skew
  - Pollution was transformed using the square root transformation to correct for negative skew
- Scatterplots matrices and heatmaps were used to examine cross-sectional bivariate associations between variables, and the existence of outliers was checked, see Figure 4 for example. The datapoints were local authorities.

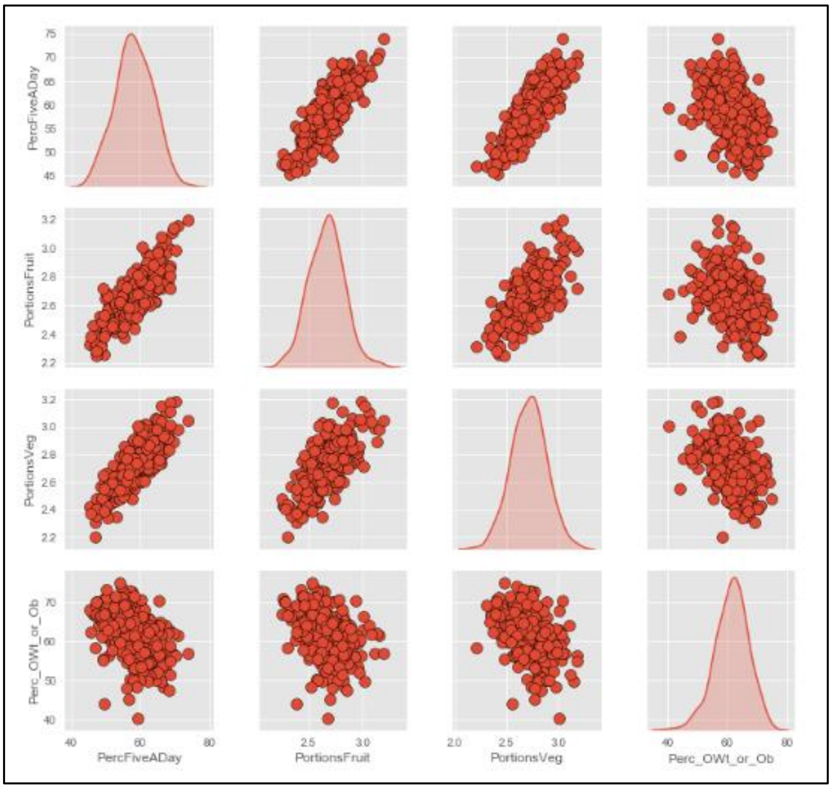


Figure 4: Bivariate scatterplots of selected predictor variables

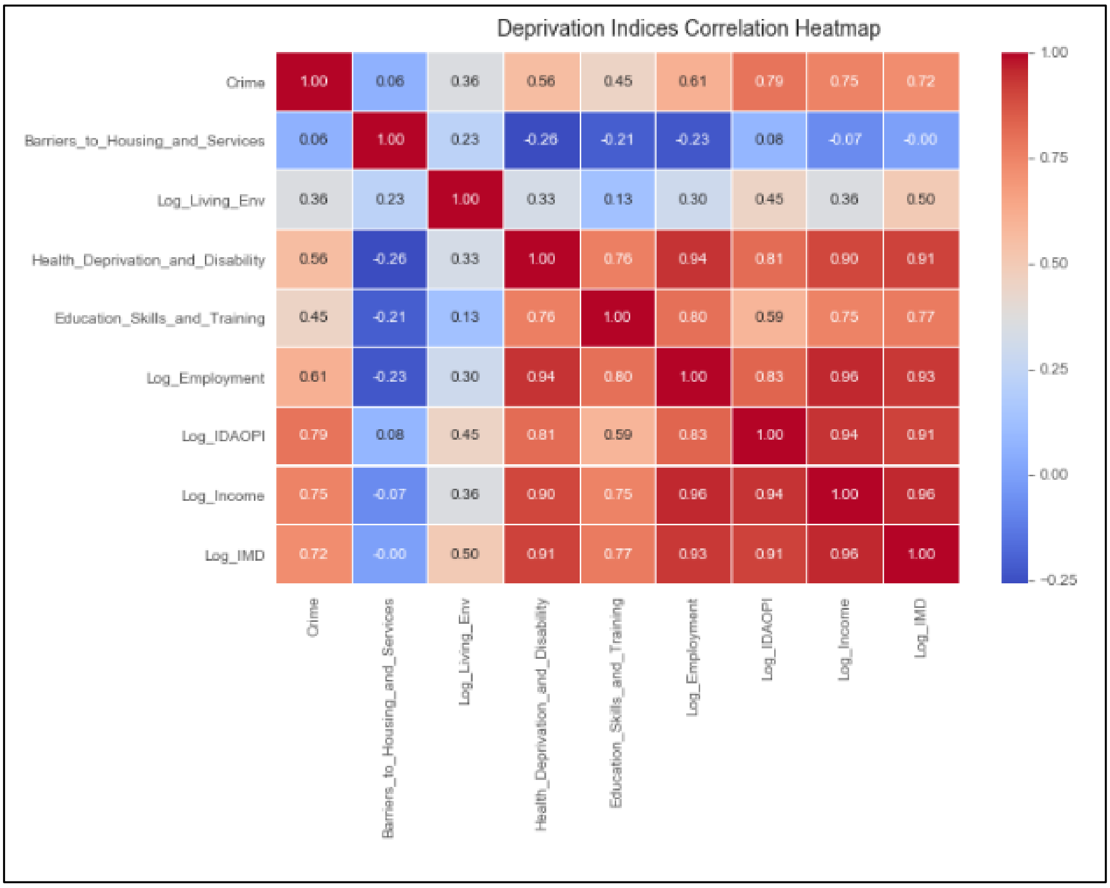


Figure 5: Heatmap showing correlation *between deprivation variables*

2.2 Data derivation

- Using Principal Components Analysis, three new “diet” variables were derived from the five variables related to diet, overweight/obesity, and physical activity, using data from 2015/6.
- Three new deprivation variables were derived from the seven deprivation domains.

In each case three components were derived, because

- Nearly 92% of the variance in the five diet variables had already been explained by the three components. Having four components would not have offered a great advantage over having five variables. Three were chosen rather than two because the third explained a high proportion of the variance (11%).
- More than 88% of the variance in the seven deprivation variables had already been explained by the three components. Three were chosen rather than two because the third explained 9%.

Figure 6 below shows the second principal component plotted against the first, for the diet and lifestyle variables (upper) and deprivation (lower) coloured by values of the Respiratory Disease

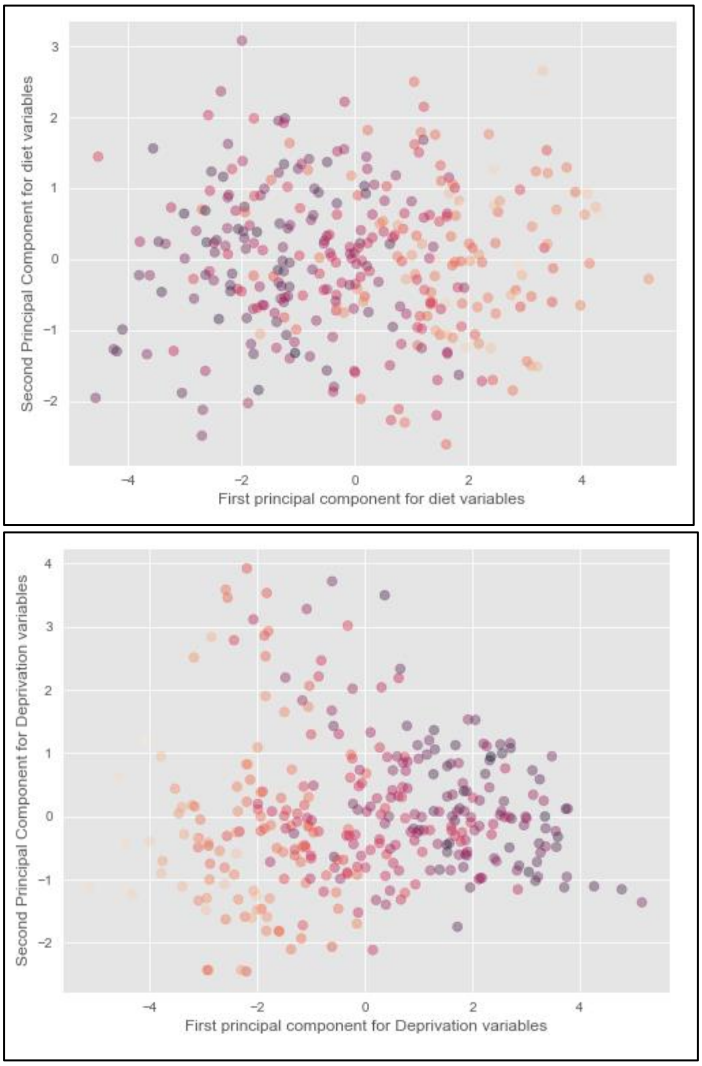


Figure 6: Second Principal Component plotted against First Principal Component, for Diet and Lifestyle Variables (Upper plot) and Deprivation Variables (Lower plot), coloured by values of Respiratory Disease Mortality

Mortality rate. These show that the mortality rates are highly correlated with the first component in each case and not visibly so for the second. Similar patterns were obtained for CVD and cancer mortality rates.

2.3 Model Building

A series of regression models were tested to explore the underlying reasons for the patterns of mortality in the three diseases. Separate models were run for each disease.

First the best subset of predictors was selected by consecutively increasing the number of predictors included, where "best" was quantified by minimising the residual sum of squares (RSS).

Second the "forward stepwise" method was used, as this is a good choice when multicollinearity is a problem (this had been identified in the scatterplots and heatmaps earlier), and also to see if the same "best" variables would be selected as for the enter method.

Plots were created for RSS, adjusted RSquared, AIC, and BIC[5] for all of the sets of models, in order to see what difference it made including more or fewer predictors, and how stable the models were. An example is given in Figure 7.

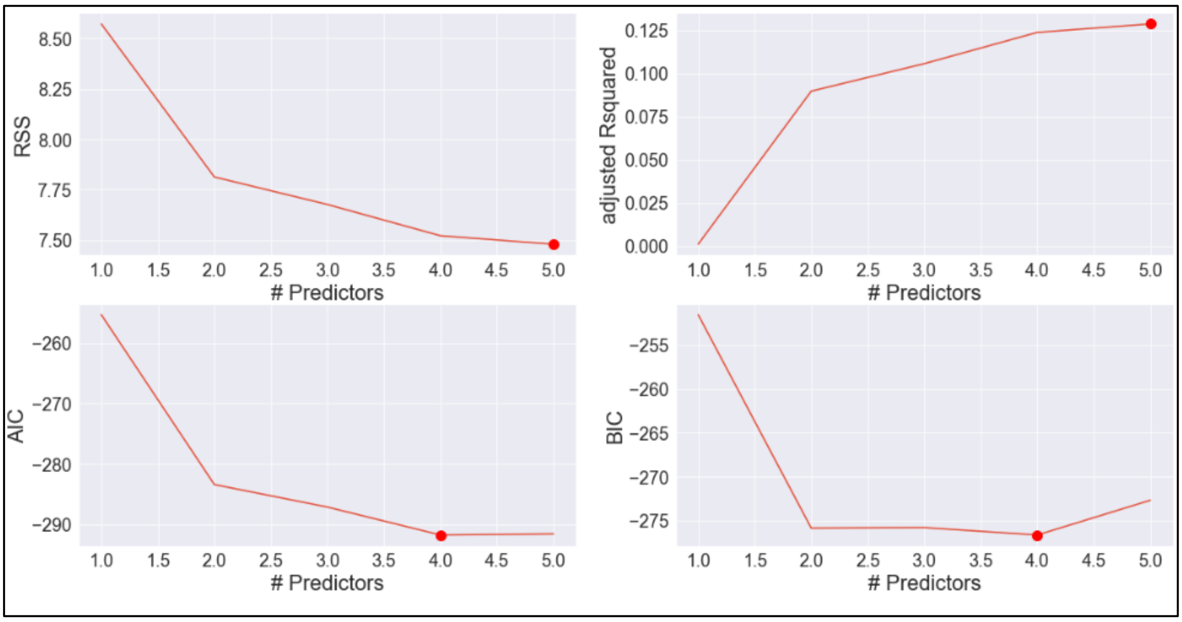


Figure 7: Plots to check performance of regression modelling, for method "Enter", RD Mortality

3 FINDINGS Please refer to Appendix 2 for details of each model)

3.1 Models 1 – 6 were intended to evaluate whether the PCA components would be included in addition to the original variables, or instead of them. The PCA components were available as predictors as well as the original variables from which they were derived.

- (Models 1 and 2) Around 81% of the variance in RD Mortality was explained by 5 predictors, that included a mix of PCA components and original variables. The selected variables differed

[5] AIC and BIC are both "penalized-likelihood criteria" used for choosing best predictor subsets in regression

between the two methods used. With the enter method, when the number of predictors was restricted to one (plus constant), IMD was chosen over the deprivation PCA components.

- (Models 3 and 4) Around 85% variance in CVD Mortality was explained by 5 predictors, that included a mix of PCA components and original variables. For both methods, the first selected variable was the first PCA deprivation component, but the other variables differed
- (Models 5 and 6) Around 78% variance in Cancer Mortality was explained by 5 predictors, and these were more heavily weighted towards the PCA components than the other two mortality types. For both methods, the first selected variable was the first PCA deprivation component, but the other variables selected differed between the two methods.

In each case the plots showed that including more than around 5 predictors would not reduce R squared value. With respect to the question being explored, the PCA components did explain a large proportion of the variance but original variables were also selected for inclusion with them.

3.2 Models 7 – 9 were intended to evaluate whether the PCA components were sufficiently good predictors alone. Only the PCA components were available to be selected, together with pollution and smoking (since those variables had not been used to derive the principal components).

- (Model 7) The first two lifestyle components together with the first two deprivation components explained 78% of RD Mortality variance.
- (Model 8) The first and third lifestyle components together with the first two deprivation components explained 83% of CVD Mortality variance.
- (Model 9) The first three deprivation components together with the second lifestyle component explained nearly 78% of Cancer Mortality variance.

With respect to the question being explored, the finding was that when only the PCA components were used as predictors, they did explain nearly as much variance as the unrestricted models (1 – 6).

3.3 Models 10 - 15 were intended to explore how the PCA deprivation components compared to the government-derived index of deprivation. Values of R squared are in this table:

% variance in mortality explained by official deprivation index, and this study's first PCA component[6]				
Models	Type of Mortality	IMD	Deprivation PCA component 1	Deprivation PCA components 1, 2 and 3
10, 11	RD	**63.6%**	**71.9%**	74.7%
12, 13	CVD	**73.8%**	**79.0%**	79.2%[7]
14,15	Cancer	**62.6%**	**70.6%**	77.0%

In short, the first deprivation PCA component was a more efficient predictor of variation in the three mortalities compared to the government index IMD.

3.4 Models 16 – 21 were intended to explore to what extent the variation existing beyond that associated with deprivation, could be modelled. This approach is based on the conceptual framework in Figure 2, whereby behavioural factors would be expected to be associated with socio-economic status. The initial modelling of deprivation was based on the findings from the analysis described above under 3.3, in that deprivation PCA components were used in preference to the government index. Values of R squared from modelling of the residuals are given in this table:

[6] For interest, values are also included showing % variance explained by all three PCA components
[7] In fact inclusion of deprivation PCA component 2 actually reduced the adjusted R Squared values

% variance in residuals output from modelling using deprivation, explained by predictor variables					
Models	Type of Mortality	Method Enter		Method Forward Stepwise	
		R sq.	Main Predictors	R sq.	Main Predictors
16, 17	RD	**11.5%**	**Veg, % active, Fruit**	**10.2%**	**Pollution, 5aDay, %Active, Veg**
18, 19	CVD	**15.1%**	**Veg, % OverWt, Fruit**	**14.8%**	**5aDay, Veg, %OverWt**
20, 21	Cancer	**3.9%**	**Veg, %Active, %Inactive, 5aDay, Fruit**	**3.6%**	**Pollution, Smoking, %Active, Veg, Fruit**

In short, quite a low proportion of variation in residuals could be accounted for (maximum 15%). A much higher proportion of residual variation in CVD mortality could be accounted for compared to cancer mortality. This finding is in line with the findings in 3.1 above where for cancer the predictors selected were more heavily in favour of the deprivation PCA components.

4 DISCUSSION

With respect to the original research questions, findings are as follows:

1) Patterns of variation in <75 y. mortality rates of respiratory disease in England are strongly linked with deprivation. Deprivation explains up to 85% of the variation between mortality rates of local authority areas in England. Once deprivation has been accounted for, behavioural risk factors and pollution do not explain a high proportion of additional variation.

2) Determinants of respiratory disease mortality are closer to those of CVD than those of cancer.

These findings are of relevance to public health policy. The evidence for existence of health inequalities between local authority areas underlines the need for more attention to preventive public health measures. As Newton and colleagues stated "If levels of health in the worst performing regions in England matched the best performing ones, England would have one of the lowest burdens of disease of any advanced industrialised country" [1].

Limitations of this study

- The unit of analysis was local authority areas. Their size can mask significant health inequality within them. For example, the local authority area of Kensington and Chelsea includes extremes of rich and poor that live side by side.
- Pollution needs to be examined at higher levels of geographic granularity
- The conceptual framework was not followed completely in that living conditions are included as components of deprivation
- A lifestyle variable for alcohol consumption should have been included

5 CONCLUSIONS

Deprivation is an enormously important determinant of mortality from respiratory disease, as well as from cancer and CVD. The importance of lifestyle variables differs between the diseases. This study provides a useful exploration of the available data, as it indicates which variables should be focused on in future time series analysis[8].

[8] To be undertaken as part of the Visual Analytics coursework

REFERENCES

[1] Newton JN, Briggs AD, Murray CJ, Dicker D, Foreman KJ, Wang H, et al. Changes in health in England, with analysis by English regions and areas of deprivation, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* 2015; 386**:**2257-2274.

[2] Scarborough P, Bhatnagar P, Wickramasinghe KK, Allender S, Foster C, Rayner M. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006–07 NHS costs. *Journal of public health* 2011; 33**:**527-535.

[3] Solar O, Irwin A. *A conceptual framework for action on the social determinants of health.* In Social Determinants of Health Discussion Paper 2 (Policy and Practice). Geneva: World Health Organization; 2010.

[4] NHS Digital. *NHS Outcomes Framework 2017/18 Indicator and Domain Summary Tables.* 2018.

[5] Smith T, Noble M, Noble S, Wright G, McLennan D, Plunkett E. *The English Indices of Deprivation 2015, Technical report.* London: Department for Communities and Local Government; 2015.

Appendix 1 – Variables

Variable name	Original Source	Full name in Source	Notes
Dependent Variables			
RD mortality	Public Health England (based on Office for National Statistics (ONS) source data)	4.04i - Under 75 mortality rate from all cardiovascular diseases	Age-standardised rate of mortality from all cardiovascular diseases (including heart disease and stroke) in persons less than 75 years per 100,000 population
CVDMortality	Public Health England (based on ONS source data)	4.05i - Under 75 mortality rate from cancer	Age-standardised rate of mortality from all cancers in persons less than 75 years per 100,000 population
CancerMortality	Public Health England (based on ONS source data)	4.07i - Under 75 mortality rate from respiratory disease	Age-standardised rate of mortality from respiratory disease in persons less than 75 years per 100,000 population
Diet and Activity Independent Variables			
PercFiveADay	Public Health England (based on Active Lives, Sport England)	2.11i - Proportion of the adult population meeting the recommended 5-a-day on a 'usual day' (adults)	Proportion of the population who, when surveyed, reported that they had eaten the recommended 5 portions of fruit and vegetables on a usual day.
PortionsFruit	Public Health England (based on Active Lives, Sport England)	2.11ii - Average number of portions of fruit consumed daily (adults)	Average (mean) number of portions reported by survey respondents aged 16+ when asked how many portions of fruit they ate on the previous day.
PortionsVeg	Public Health England (based on Active Lives, Sport England)	2.11iii - Average number of portions of vegetables consumed daily (adults)	Average (mean) number of portions reported by survey respondents aged 16+ when asked how many portions of vegetables they ate on the previous day.
Perc_OWt_or_Ob	Public Health England (based on Active Lives, Sport England)	2.12 - Percentage of adults (aged 18+) classified as overweight or obese	Percentage of adults aged 18 and over classified as overweight or obese
PercActive	Public Health England (based on Active Lives, Sport England)	2.13i - Percentage of physically active adults	The number of respondents aged 19 and over, with valid responses to questions on physical activity, doing at least 150 moderate intensity equivalent (MIE) minutes physical activity per week in bouts of 10 minutes or more in the previous 28 days expressed as a percentage of the total number of respondents aged 19 and over.
PercInactive	Public Health England (based on Active Lives, Sport	2.13ii - Percentage of physically inactive adults	The number of respondents aged 19 and over, with valid responses to questions on physical activity, doing less than 30 moderate intensity equivalent (MIE)

	England)		minutes physical activity per week in bouts of 10 minutes or more in the previous 28 days expressed as a percentage of the total number of respondents aged 19 and over.
Deprivation Independent Variables			
IMD	Department for Communities and Local Government	Index of Multiple Deprivation	Population weighted average of the combined scores for the LSOAs of a local authority district. Domains are combined using weights as follows: • Income Deprivation (22.5%) • Employment Deprivation (22.5%) • Education, Skills and Training Deprivation (13.5%) • Health Deprivation and Disability (13.5%) • Crime (9.3%) • Barriers to Housing and Services (9.3%) • Living Environment Deprivation (9.3%)
Income	As above	Income Deprivation	Derived from indicators: • Adults and children in Income Support families • Adults and children in income-based Jobseeker’s Allowance families Adults and children in income-based Employment and Support Allowance families • Adults and children in Pension Credit (Guarantee) families • Adults and children in Child Tax Credit and Working Tax Credit families, below 60% median income not already counted • Asylum seekers in England in receipt of subsistence support, accommodation support, or both
Employment	As above	Employment Deprivation	Derived from indicators: • Claimants of Jobseeker’s Allowance, aged 18-59/64 • Claimants of Employment and Support Allowance, aged 18-59/64 Claimants of Incapacity Benefit, aged 18-59/64 • Claimants of Severe Disablement Allowance, aged 18-59/64 • Claimants of Carer’s Allowance, aged 18-59/64
Education_Skills_ and_Training	As above	Education, Skills and Training Deprivation	Derived from indicators: • Key stage 2 attainment: average points score • Key stage 4 attainment: average points score • Secondary school absence • Staying on in education post 16 • Entry to higher education

			• Adults with no or low qualifications, aged 25-59/64 • English language proficiency, aged 25-59/64
Health_ Deprivation_ and_Disability	As above	Health Deprivation and Disability	Derived from indicators: • Years of potential life lost • Comparative illness and disability ratio • Acute morbidity • Mood and anxiety disorders
Crime	As above	Crime	Recorded crime rates for: Violence; Burglary; Theft; Criminal damage
Barriers_to_ Housing_ and_Services	As above	Barriers to Housing and Services	Derived from indicators: • Road distance to: post office; primary school; general store / supermarket; GP surgery • Household overcrowding • Homelessness • Housing affordability
Living_ Environment	As above	Living Environment Deprivation	Derived from indicators: • Housing in poor condition • Houses without central heating • Air quality • Road traffic accidents
IDAOPI	As above	Income Deprivation Affecting Older People Index (IDAOPI)	Derived from indicator: Proportion of all those aged 60 or over who experience income deprivation. This includes adults aged 60 or over receiving Income Support or income-based Jobseekers Allowance or income-based Employment and Support Allowance or Pension Credit
Other Indpendent Variables			
PercSmoking	Annual Population Survey	2.14 - Smoking Prevalence in adults - current smokers	Prevalence of smoking among persons 18 years and over
Pollution	DEFRA/Air Pollution and Climate Change Group Public Health England	3.01 - Fraction of mortality attributable to particulate air pollution	Fraction of annual all-cause adult mortality attributable to anthropogenic (human-made) particulate air pollution (measured as fine particulate matter, PM2.5)

Appendix 2 – Summary Findings from Linear Regression

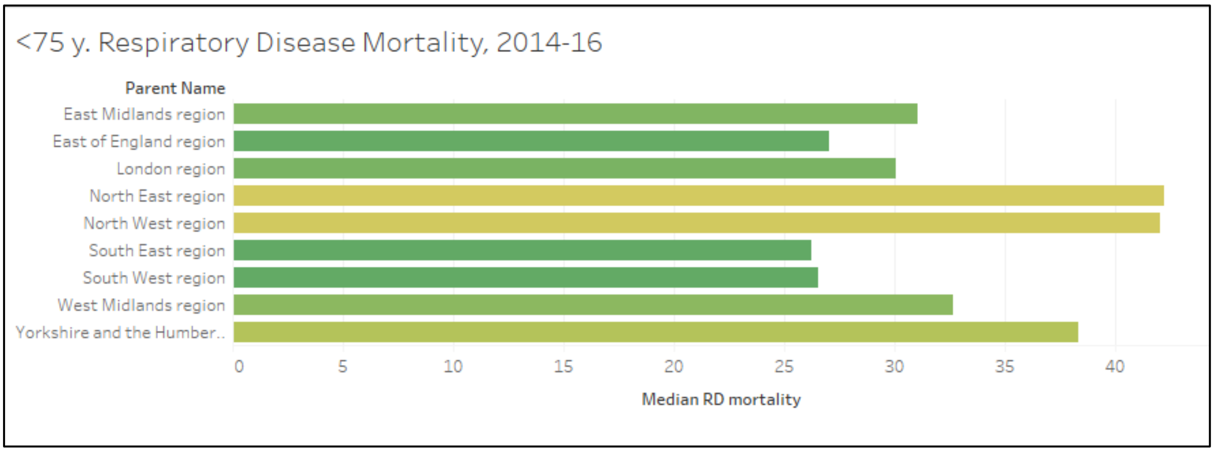
Model	Dependent Variable	Method	Independent Variables included	Adjusted R Squared	Change in R Squared
			ALL PREDICTORS AVAILABLE TO CHOOSE FROM		
1	Log_RDMortality	Enter	Health	72.8%	
			Crime, Health	77.3%	4.5%
			DepPCAComp2, Crime, Health	79.8%	2.5%
			DepPCAComp2, Crime, Health, Veg	81.2%	1.4%
2	Log_RDMortality	Forward Stepwise	Log_IMD	63.6%	
			Log_IMD, DepPCAComp1	74.5%	10.9%
			Log_IMD, DepPCAComp1, Fruit	77.2%	2.7%
			Log_IMD, DepPCAComp1, Fruit, Health	79.5%	2.3%
			Log_IMD, DepPCAComp1, Fruit, Health, Crime	80.7%	1.2%
3	Log_CVDMortality	Enter	DepPCAComp1	79.0%	
			DepPCAComp1, Veg	82.9%	3.9%
			DietPCAComp1, Crime, Health	85.0%	2.1%
			DietPCAComp1, Crime, Health, IDAOPI	85.2%	0.2%
4	Log_CVDMortality	Forward Stepwise	DepPCAComp1	79.0%	
			DepPCAComp1, Veg	82.9%	3.9%
			DepPCAComp1, Veg, IDAOPI	83.6%	0.7%
			DepPCAComp1, Veg, IDAOPI, Income	84.4%	0.8%
			DepPCAComp1, Veg, IDAOPI, Income, Living_Env	84.9%	0.5%
5	Log_CancerMortality	Enter	DepPCAComp1	70.6%	
			DepPCAComp1, DepPCAComp2	76.5%	5.9%
			DepPCAComp1, Living_Env, Income	77.1%	0.6%
			DepPCAComp3, Living_Env, Housing, Health	77.7%	0.6%
6	Log_CancerMortality	Forward Stepwise	DepPCAComp1	70.6%	
			DepPCAComp1, DepPCAComp2	76.5%	5.9%
			DepPCAComp1, DepPCAComp2, DepPCAComp3	77.0%	0.5%
			DepPCAComp1, DepPCAComp2, DepPCAComp3, DietPCAComp2	77.4%	0.4%
			DepPCAComp1, DepPCAComp2, DepPCAComp3, DietPCAComp2, Health	77.7%	0.3%

Model	Dependent Variable	Method	Independent Variables included	Adjusted R Squared	Change in R Squared
			RESTRICTED SET OF PREDICTORS AVAILABLE = only PCA components, Smoking and Pollution		
7	Log_RDMortality	Forward Stepwise	DepPCAComp1	71.9%	
			DepPCAComp1, DietPCAComp1	75.1%	3.2%
			DepPCAComp1, DietPCAComp1, DepPCAComp2	76.5%	1.4%
			DepPCAComp1, DietPCAComp1, DepPCAComp2, DietPCAComp2	78.0%	1.5%
			DepPCAComp1, DietPCAComp1, DepPCAComp2, DietPCAComp2, Smoking	77.9%[9]	-0.1%
8	Log_CVDMortality	Forward Stepwise	DepPCAComp1	79.0%	
			DepPCAComp1, DietPCAComp1, DietPCAComp2	83.0%	4.0%
			DepPCAComp1, DietPCAComp1, DepPCAComp2, DietPCAComp3	83.1%	0.1%
			DepPCAComp1, DietPCAComp1, DepPCAComp2, DietPCAComp3, Smoking	83.1%	0.0%
9	Log_CancerMortality	Forward Stepwise	DepPCAComp1	70.6%	
			DepPCAComp1, DepPCAComp2	76.5%	5.9%
			DepPCAComp1, DepPCAComp2, DepPCAComp3	77.0%	0.5%
			DepPCAComp1, DepPCAComp2, DepPCAComp3, DietPCAComp2	77.4%	0.4%
			DepPCAComp1, DepPCAComp2, DepPCAComp3, DietPCAComp2, Smoking	77.6%	0.2%
10	Log_RDMortality	Enter	ONLY IMD AVAILABLE	63.6%	
			Only PCAComponents Available:		
11	Log_RDMortality	Forward Stepwise	DepPCAComp1	71.9%	
			DepPCAComp1, DepPCAComp2	74.6%	2.7%
			DepPCAComp1, DepPCAComp2, DepPCAComp3	74.7%	0.1%
12	Log_CVDMortality	Enter	ONLY IMD AVAILABLE	73.8%	
			Only PCAComponents Available:		
13	Log_CVDMortality	Forward Stepwise	DepPCAComp1	79.0%	5.2%
			DepPCAComp1, DepPCAComp3	79.3%	0.3%
			DepPCAComp1, DepPCAComp2, DepPCAComp3	79.2%	-0.1%
14	Log_CancerMortality	Enter	ONLY IMD AVAILABLE	62.6%	
			Only PCAComponents Available:		
15	Log_CancerMortality	Forward Stepwise	DepPCAComp1	70.6%	

[9] Yellow highlighting indicates no change or a decrease in R Squared value

Model	Dependent Variable	Method	Independent Variables included	Adjusted R Squared	Change in R Squared
			DepPCAComp1, DepPCAComp2	76.5%	5.9%
			DepPCAComp1, DepPCAComp2, DepPCAComp3	77.0%	0.5%
			ONLY LIFESTYLE VARIABLES AND POLLUTION AVAILABLE		
16	RD_Resids	Enter	Veg	8.7%	
			Veg, PercActive	10.0%	1.3%
			Veg, PercActive, Fruit	11.5%	1.5%
			Veg, PercActive, Fruit, PercOverwt	11.8%	0.3%
17	RD_Resids	Forward Stepwise	Pollution, 5 a day (no Constant)	2.5%	
			Pollution, 5 a day, PercActive (no Constant)	7.2%	4.7%
			Pollution, 5 a day, PercActive, Constant	8.0%	0.8%
			Pollution, 5 a day, PercActive, Veg, Constant	10.2%	2.2%
			Pollution, 5 a day, PercActive, Veg, Fruit, Constant	11.4%	1.2%
18	CVD_Resids	Enter	Veg	13.7%	
			Veg, PercOverWt	15.0%	1.3%
			Veg, PercOverWt, Fruit	15.1%	0.1%
			Veg, PercOverWt, Fruit, Smoking	15.2%	0.1%
19	CVD_Resids	Forward Stepwise	5 a day	10.1%	
			5 a day, Veg	14.0%	3.9%
			5 a day, Veg, PerOverWt	14.8%	0.8%
			5 a day, Veg, PerOverWt, Smoking	14.9%	0.1%
			5 a day, Veg, PerOverWt, Smoking, Fruit	14.9%	0.0%
20	Cancer_Resids	Enter	Veg, PercActive (no Constant)	2.4%	
			Veg, PercActive, PercInactive (no Constant)	2.5%	0.1%
			Veg, PercActive, PercInactive, 5 a day (no Constant)	2.8%	0.3%
			Veg, PercActive, PercInactive, 5 a day, Fruit (no Constant)	3.9%	1.1%
21	Cancer_Resids	Forward Stepwise	Pollution, Smoking (no Constant)	-0.4%	
			Pollution, Smoking, PercActive (no Constant)	0.2%	0.6%
			Pollution, Smoking, PercActive, Veg (no Constant)	2.0%	1.8%
			Pollution, Smoking, PercActive, Veg, Constant	2.6%	0.6%
			Pollution, Smoking, PercActive, Veg, Fruit, Constant	3.6%	1.0%

Appendix 3 – Maps and barcharts showing geographical variation in outcome variables



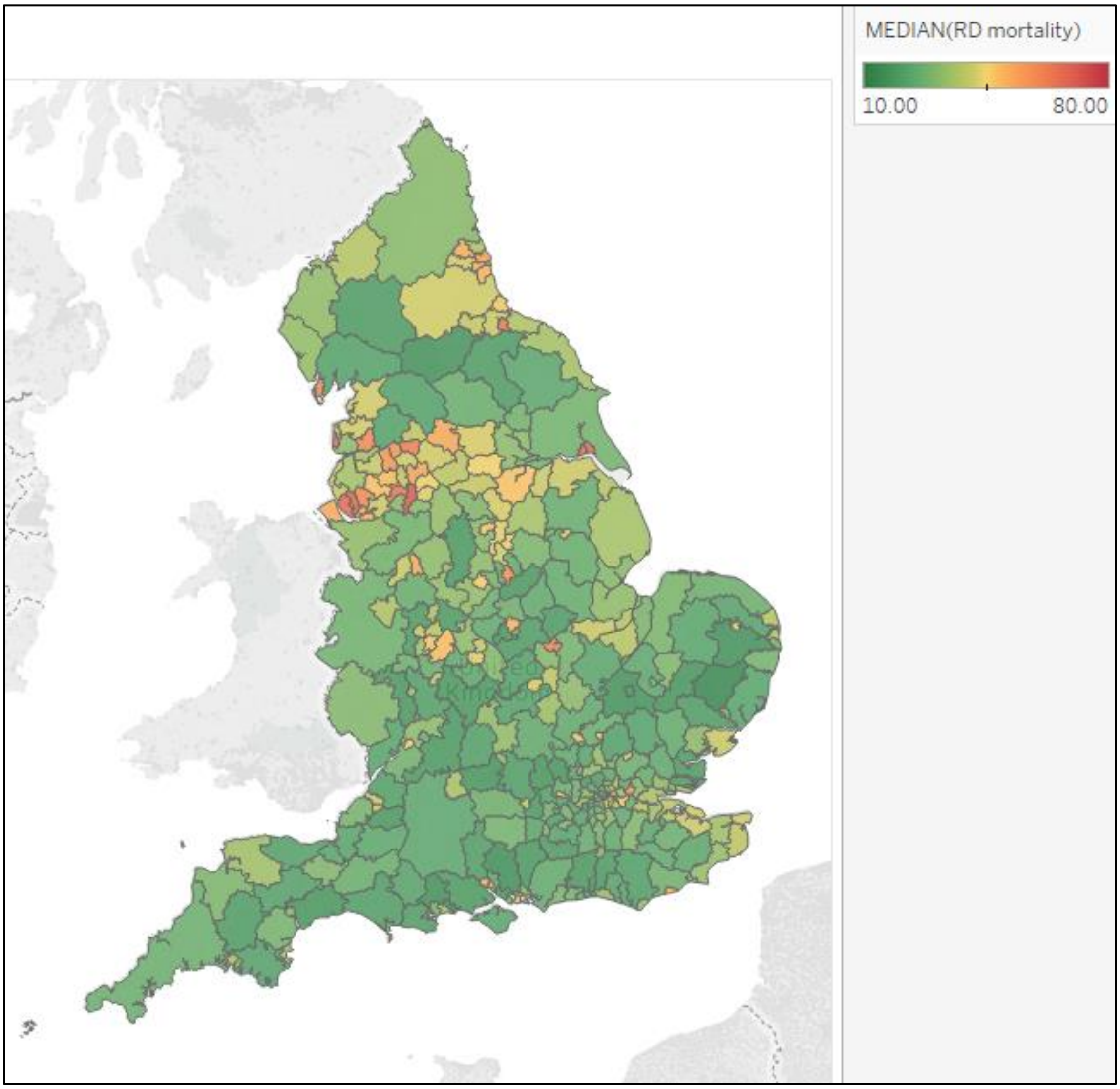
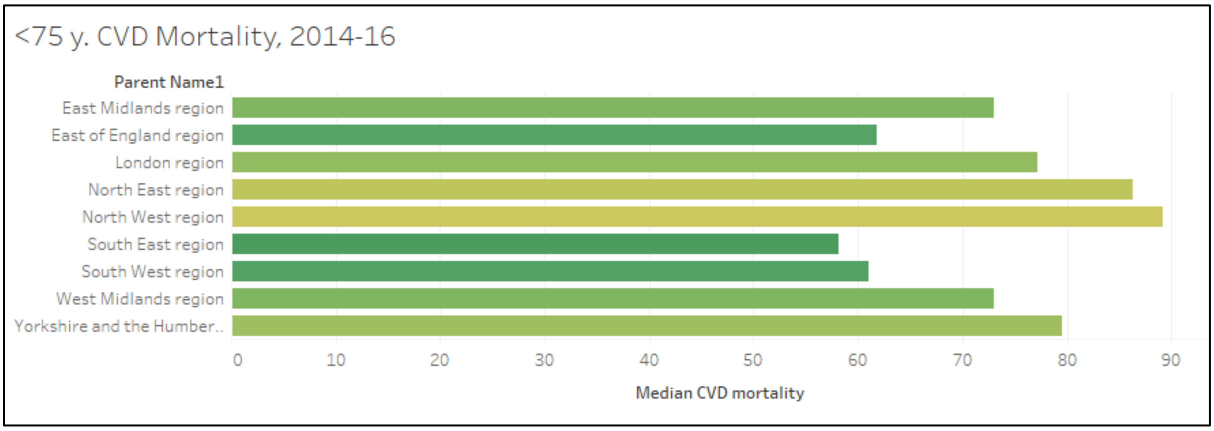


Figure a: Geographical distribution of <75 y. Respiratory Disease Mortality, England 2014-6



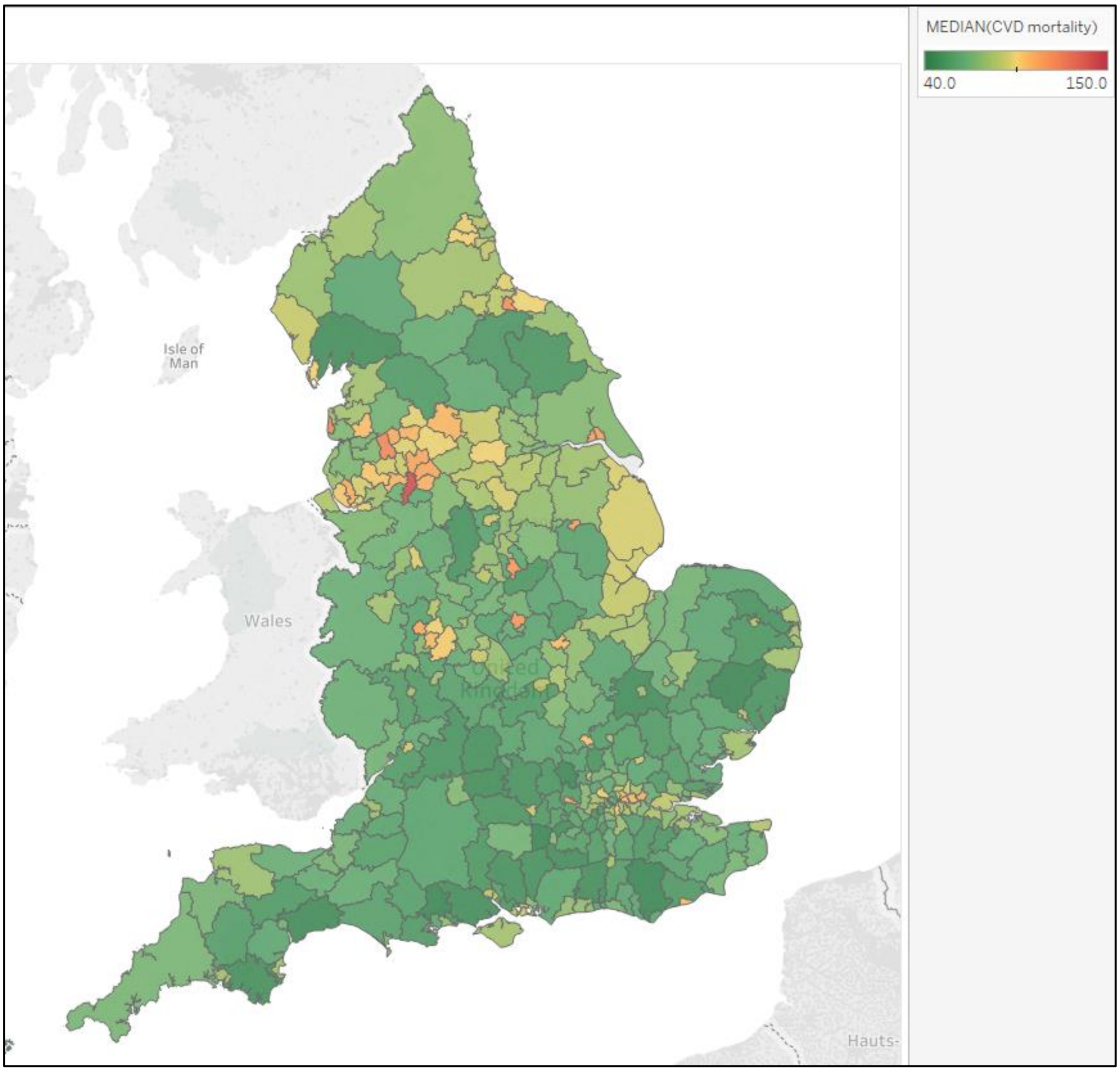
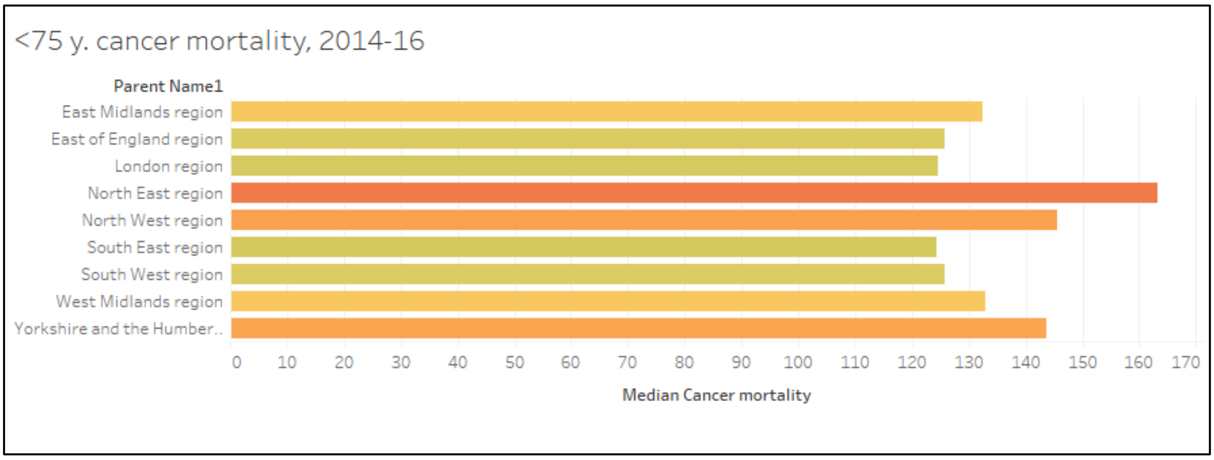


Figure b: Geographical distribution of <75 y. Cardio-Vascular Disease Mortality, England 2014-6



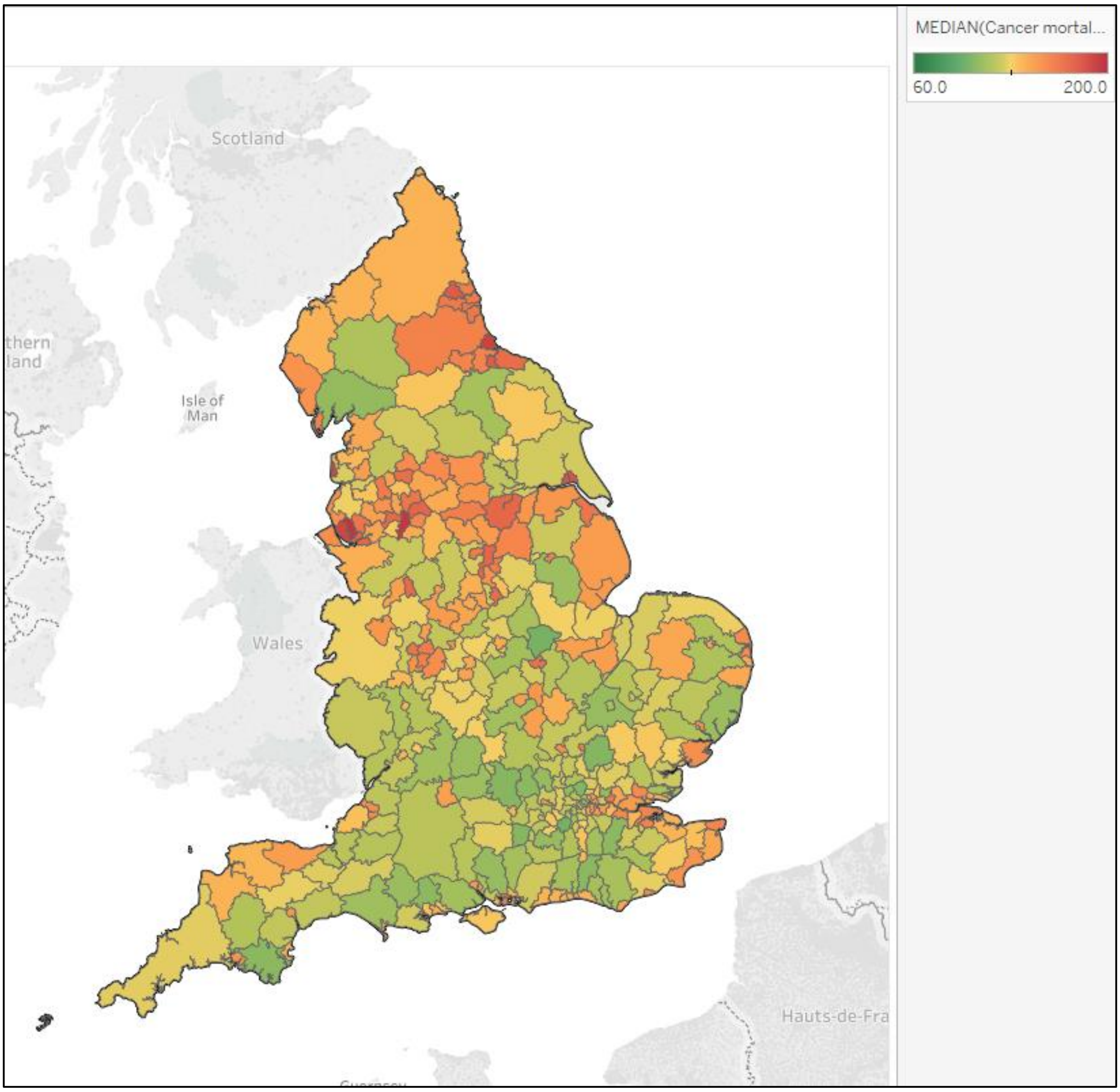


Figure c: Geographical distribution of <75 y. Cancer Mortality, England 2014-6

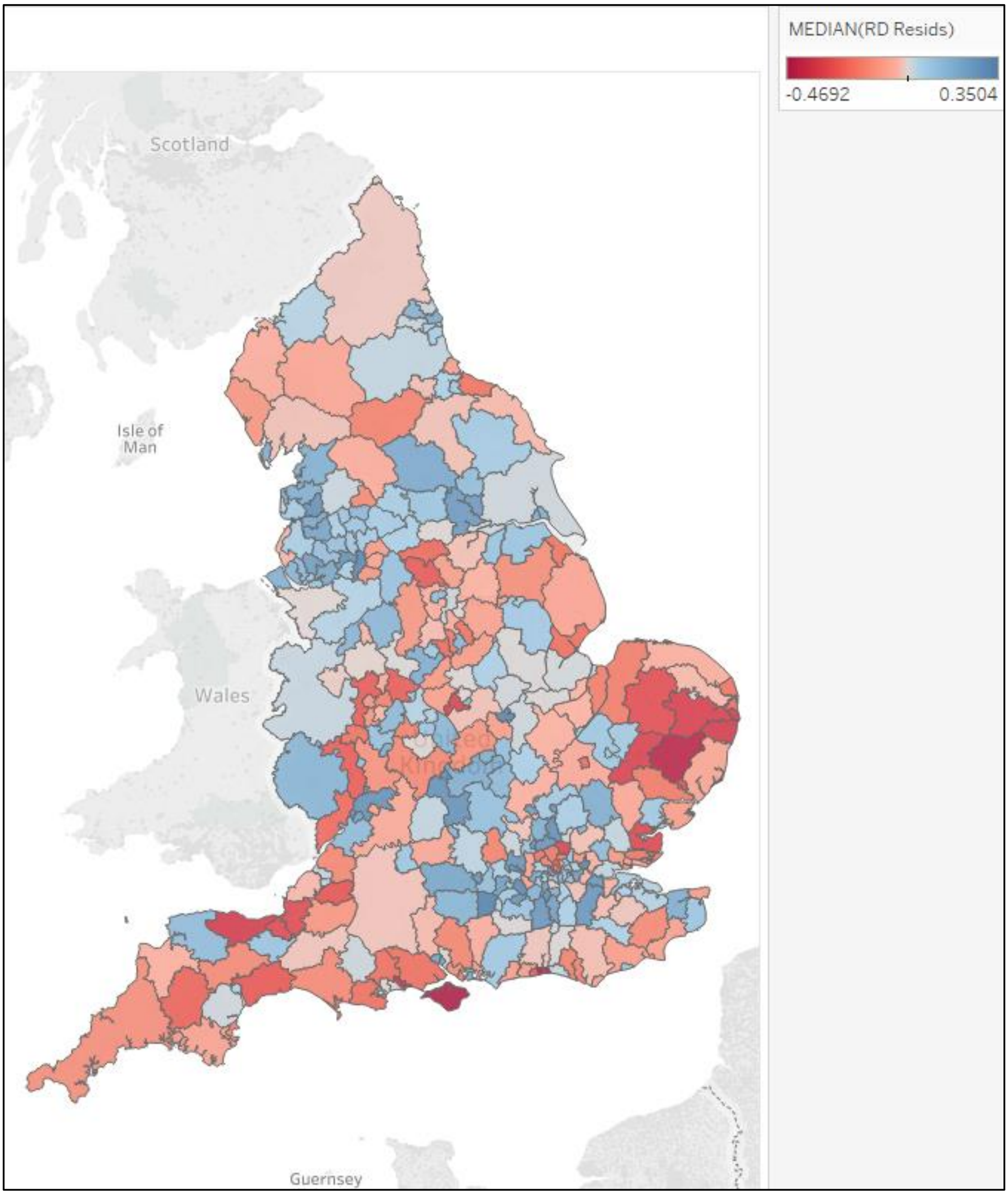


Figure d: Geographical distribution of Residuals from Linear regression of <75 y. Respiratory Disease Mortality as outcome, and Deprivation PCA Components as predictors

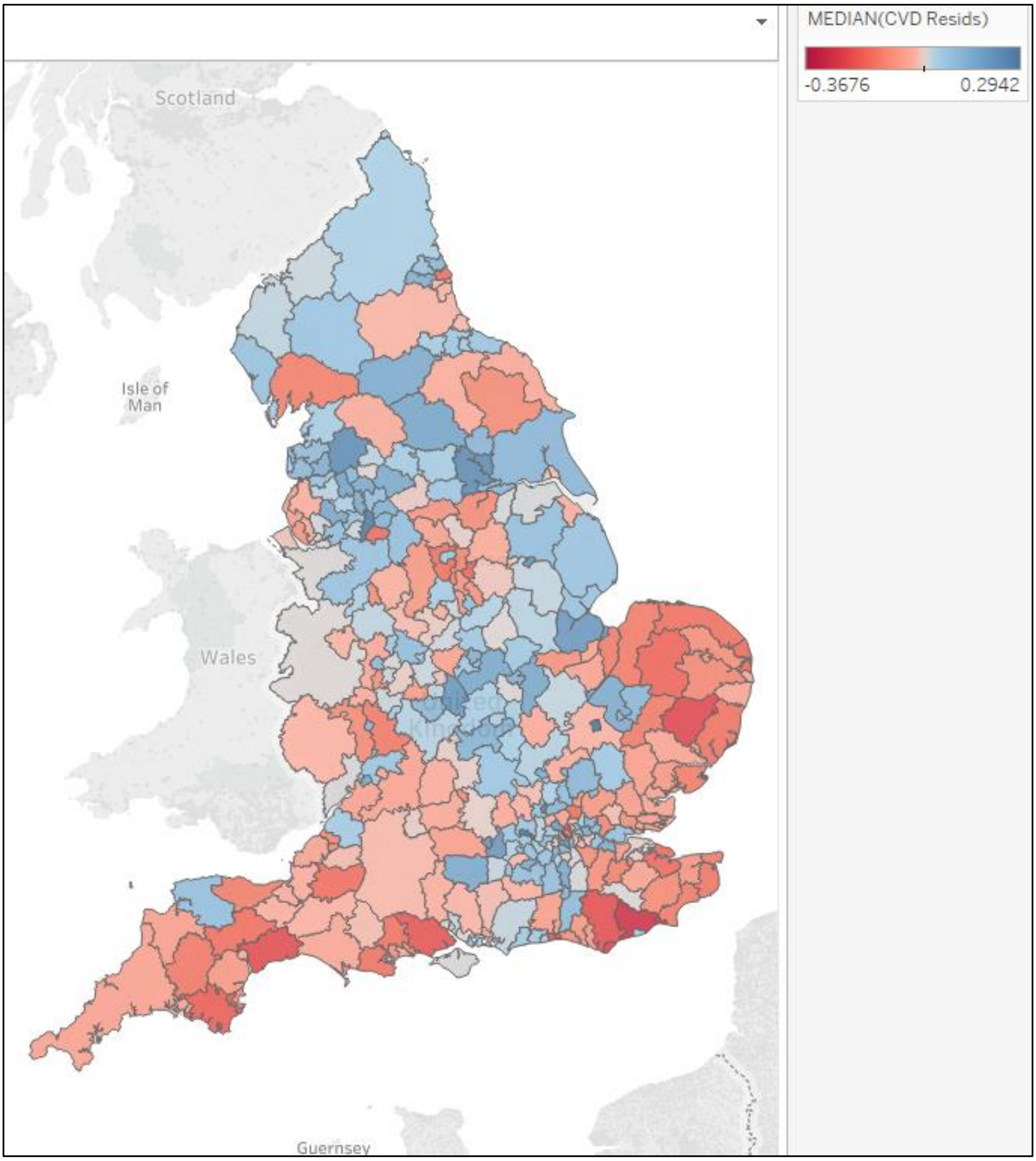


Figure e: Geographical distribution of Residuals from Linear regression of <75 y. Cardiovascular disease Mortality as outcome, and Deprivation PCA Components as predictors

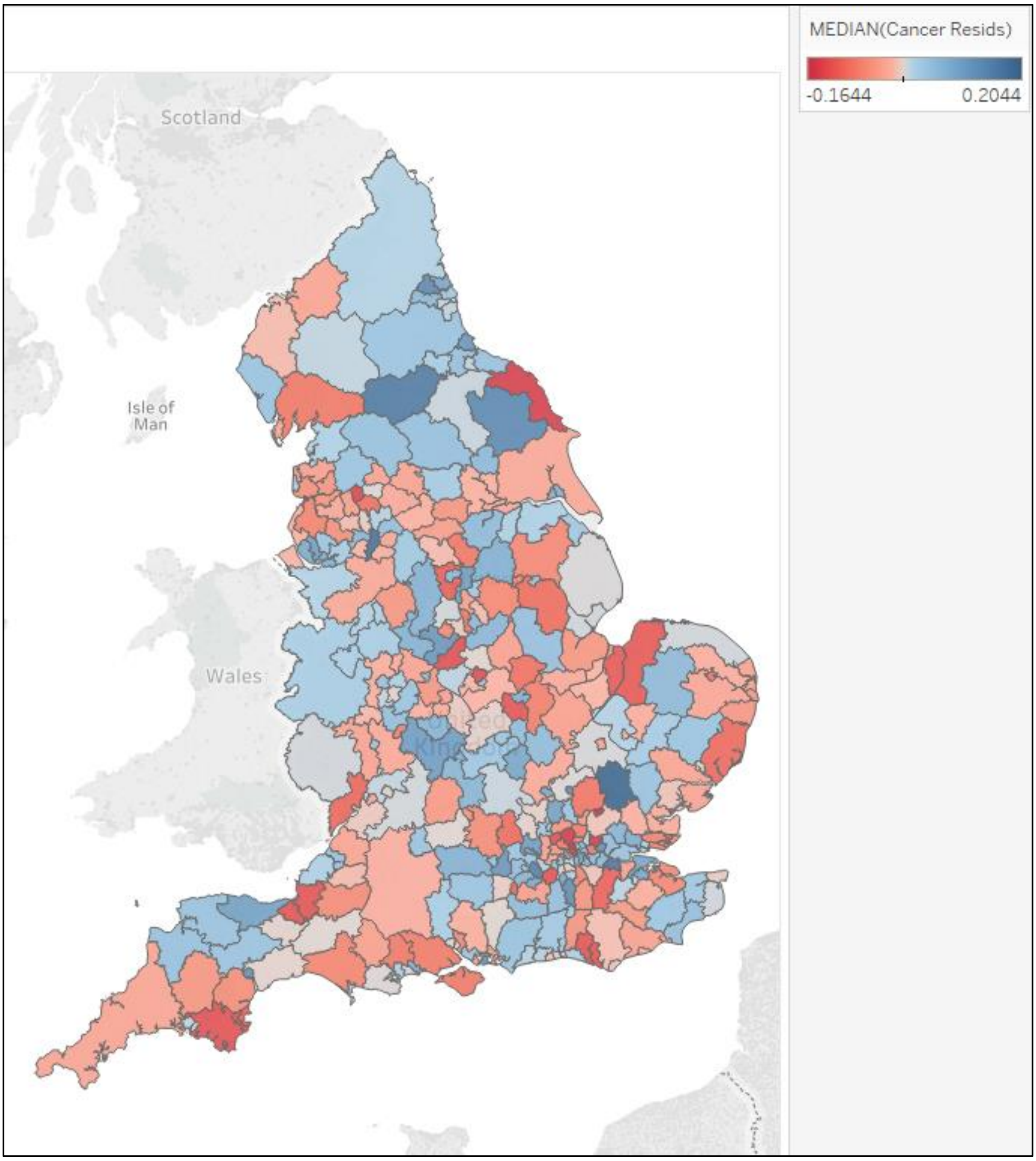


Figure f: Geographical distribution of Residuals from Linear regression of <75 y. Cancer Mortality as outcome, and Deprivation PCA Components as predictors