There is any relation between unemployment rates and socio-economic indicators at LSOA level in the London area?

Bogdan Rugina

12 December 2018

## Introduction

#### Urban unemployment rates seem to be shaped by different social problems that are nested behind poverty, ghettos and segregation (Zenou, 2000). The advantages of educated workers are the subject of an immense literature and could be translated in higher wages, greater mobility and employment stability (Mincer, 1991). Higher education among workers leads under this light to lower rate of unemployment, which results to human capital accumulation and high productivity (Nickell, 1979).

#### In this study the main objective is to identify to which extent different levels of education and median income of each London small areas contribute to unemployment for the year 2011. The first part relates to the inferential statistical analysis, developed on R environment, starting with a bi-variate linear regression and based on first results other indicators will be inserted one by one into a multiple regression model. The second part will consist on a geographical weighted regression (GWR) performed on ArcGIS software in order to visualize local aspects of the multiple regression model, some predictions will be inspected as part of the overall objective.

## Data

#### Data used to develop this research comes entirely from the UK Office for National Statistics, it refers to the London Lower Super Output Areas (LSOA) which are smallest geographically statistic areas containing data from the official 2011 census (OSN, 2018). The London Datastore website releases as well the ONS based “LSOA Atlas”" which “provides summary of demographic” and other related data from population segmentation to housing, land use, employment, etc. (London Datastore, 2018). In this study the London LSOA Atlas will be the direct source of data for the following analysis, the extracted values of interest are shown in Figure 1.

#### Figure 1. London Datastore data

library(readr)  
clean\_data <-read\_csv("C:\\Users\\Bogdan\\Documents\\UCL\\MRes\\GIS\\week9\\Assignment3\\clean\_data.csv")

## Parsed with column specification:  
## cols(  
## Codes = col\_character(),  
## Names = col\_character(),  
## Unemployment\_Rate = col\_double(),  
## No\_qualifications = col\_integer(),  
## Level\_1\_qualifications = col\_integer(),  
## Level\_4\_qualifications\_and\_above = col\_integer(),  
## Med\_Househ\_Inc\_estim = col\_double(),  
## Working\_age\_population = col\_integer()  
## )

head(clean\_data)

## # A tibble: 6 x 8  
## Codes Names Unemployment\_Ra~ No\_qualificatio~ Level\_1\_qualifi~  
## <chr> <chr> <dbl> <int> <int>  
## 1 E010~ Camd~ 9.9 227 104  
## 2 E010~ Camd~ 8 193 99  
## 3 E010~ Camd~ 4.9 132 70  
## 4 E010~ Camd~ 8.8 239 124  
## 5 E010~ Camd~ 2.1 70 70  
## 6 E010~ Camd~ 3.6 48 31  
## # ... with 3 more variables: Level\_4\_qualifications\_and\_above <int>,  
## # Med\_Househ\_Inc\_estim <dbl>, Working\_age\_population <int>

#### As a further step, social data extracted from the LSOA Atlas will be subjected to normalization, each variable will be divided by the working-age population resulting in 3 socio-economic indicators: primary education, higher education and median household income.

#### The first set of variables to be extracted from the databased are the “Unemployment Rates” that accordingly to the OSN are calculated as the total number of unemployed people divided by the active population (ONS, 2018). Unemployment rates will represent in this study the dependent value based on which all the indicators will be related. Primary education and higher education indicators were calculated respectively as the working aged population between 16-65 that completed the first level of qualification or lower (summing unqualified population and level 1 education) from one side and the working population who obtained level 4 qualifications and above from the other, both were then divided by the total population aged 16-25. Finally, the fourth variable resulted from the household median income annual estimates for the year 2011-2012 and divided by the working aged population. Median values were chose instead of mean ones, due to the fact that eventual outliers may raise the mean values in the social context of London.

#### Figure 3. Social indicators

## A new object called "social\_indicators"" will contain the indicators  
social\_indicators <- clean\_data  
# First we sum the two primary education levels:  
social\_indicators$Primary\_education <- clean\_data$No\_qualifications+clean\_data$Level\_1\_qualifications  
  
# Secondly we divide it by the working population:  
social\_indicators$Primary\_education\_Indicator <- social\_indicators$Primary\_education/clean\_data$Working\_age\_population  
  
# We divide as well the High education by working population  
social\_indicators$Higher\_education\_Indicator <- clean\_data$Level\_4\_qualifications\_and\_above/clean\_data$Working\_age\_population  
  
# Income indicator finally  
social\_indicators$Median\_income\_Indicator <- clean\_data$Med\_Househ\_Inc\_estim/clean\_data$Working\_age\_population  
  
head(social\_indicators[,c(1:3, 10:12)])

## # A tibble: 6 x 6  
## Codes Names Unemployment\_Ra~ Primary\_educati~ Higher\_educatio~  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 E010~ Camd~ 9.9 0.341 0.516  
## 2 E010~ Camd~ 8 0.271 0.603  
## 3 E010~ Camd~ 4.9 0.187 0.784  
## 4 E010~ Camd~ 8.8 0.329 0.568  
## 5 E010~ Camd~ 2.1 0.129 0.853  
## 6 E010~ Camd~ 3.6 0.0745 0.890  
## # ... with 1 more variable: Median\_income\_Indicator <dbl>

#### In order to have a first view of how the four variables are distributed, some basic statistics were calculated in Figure 5, respectively density distribution in Figure 4 and box plots in Figure 6a and 6b. The quick statistics show that unemployment rate has a mean value of 7.4% in the whole London area, working population is represented by 33% of unqualified individuals and higher education covers only 43% of the working-age population.

#### Figure 5. Statistic summary

summary(social\_indicators[,c(1:3, 10:12)])

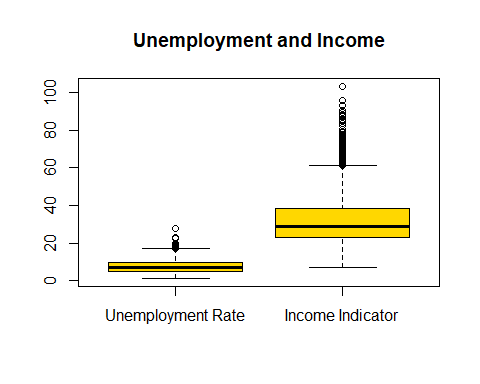
## Codes Names Unemployment\_Rate  
## Length:4835 Length:4835 Min. : 1.200   
## Class :character Class :character 1st Qu.: 4.700   
## Mode :character Mode :character Median : 6.800   
## Mean : 7.429   
## 3rd Qu.: 9.600   
## Max. :27.600   
## Primary\_education\_Indicator Higher\_education\_Indicator  
## Min. :0.02899 Min. :0.1023   
## 1st Qu.:0.24217 1st Qu.:0.3023   
## Median :0.33969 Median :0.4027   
## Mean :0.33799 Mean :0.4334   
## 3rd Qu.:0.42966 3rd Qu.:0.5511   
## Max. :0.85314 Max. :1.0000   
## Median\_income\_Indicator  
## Min. : 6.788   
## 1st Qu.: 22.797   
## Median : 28.532   
## Mean : 31.831   
## 3rd Qu.: 38.251   
## Max. :103.229

#### 

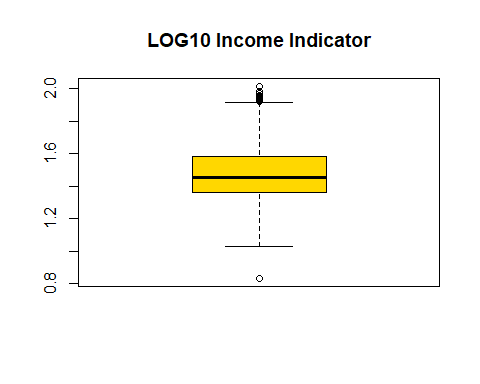
#### Median household income indicator shows a mean value of 31.8, this indicates that in London the divergence between economically rich and poor LSOA areas is very large, density distribution and box plots identify also a skewed conformation and an important range of outliers, in this case a log10 transformation seems to be a right maneuver.

#### Figure 6a. Unemployment rates and median households income boxplots

D1 <- social\_indicators$Unemployment\_Rate  
I2 <- social\_indicators$Primary\_education\_Indicator  
I3 <- social\_indicators$Higher\_education\_Indicator  
I4 <- social\_indicators$Median\_income\_Indicator  
  
boxplot(D1,I4, vertical=TRUE,names = c( "Unemployment Rate","Income Indicator"), col="gold", main="Unemployment and Income" )

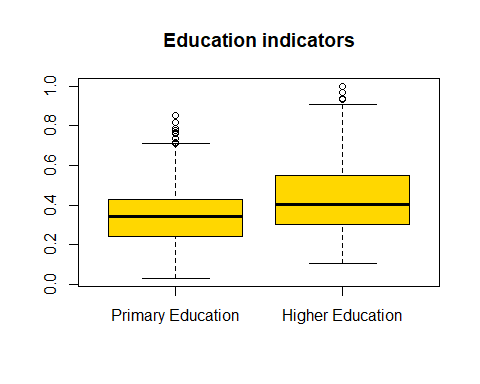


social\_indicators$Median\_Incom\_LOG10 <- log10(social\_indicators$Median\_income\_Indicator)  
boxplot(social\_indicators$Median\_Incom\_LOG10, vertical=TRUE, name = c("Income Indicator"), col="gold", main="LOG10 Income Indicator" )



#### Figure 6b. Primary education and Higher education boxplots

boxplot(I2,I3, vertical=TRUE, names = c( "Primary Education","Higher Education"), col="gold", main="Education indicators" )



#### Figure 4. Density distribution among the socio-economic indicators

##https://rpubs.com/adam\_dennett/334459   
LSOA\_of\_interest <- social\_indicators[,c(1:3, 10:12)]  
LSOA\_of\_interest

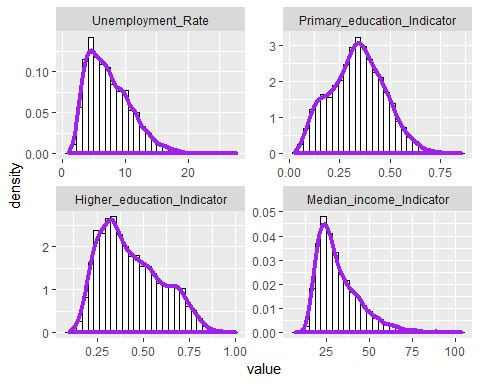
## # A tibble: 4,835 x 6  
## Codes Names Unemployment\_Ra~ Primary\_educati~ Higher\_educatio~  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 E010~ Camd~ 9.9 0.341 0.516  
## 2 E010~ Camd~ 8 0.271 0.603  
## 3 E010~ Camd~ 4.9 0.187 0.784  
## 4 E010~ Camd~ 8.8 0.329 0.568  
## 5 E010~ Camd~ 2.1 0.129 0.853  
## 6 E010~ Camd~ 3.6 0.0745 0.890  
## 7 E010~ Camd~ 3.6 0.0838 0.889  
## 8 E010~ Camd~ 3 0.179 0.788  
## 9 E010~ Camd~ 3.9 0.104 0.938  
## 10 E010~ Camd~ 2.5 0.144 0.801  
## # ... with 4,825 more rows, and 1 more variable:  
## # Median\_income\_Indicator <dbl>

library(reshape2)  
LSOA\_melt <- melt(LSOA\_of\_interest)

## Using Codes, Names as id variables

attach(LSOA\_melt)  
  
  
library(ggplot2)  
hist\_All <- ggplot(LSOA\_melt, aes(x=value)) + geom\_histogram(aes(y = ..density..), colour="black",   
 fill="white") + geom\_density(colour="orange", size=1, adjust=1)+  
 geom\_density(colour="purple",size=1.5,adjust=1)+  
 scale\_colour\_manual(name = 'Legend',   
 guide = 'legend')  
  
hist\_All + facet\_wrap(~ variable, scales="free")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Analysis

#### This study intends to understand the extent of socio-economic indicators as primary education, higher education, median households’ income and their relationships to unemployment rate in the London area for the year 2011. After pre-analyzing the data for possible skewed distributions and correlation between variables, a multiple linear regression is performed in order to test statistical significance of the findings.

#### The final output of this research will consist on local regression predicting to which extent social indicators influence the unemployment rate in London Lower Super-Output Areas (LSOA) in 2011. The inspiration for this research comes from an international project called “INEQ-CITIES” which main aim was to provide evidences of health inequalities in cities across Europe (INEQ-CITIES, 2018 ). Different publications followed the project, interesting for this study, Borrell et al (2014) emphasize “a consistent pattern of socio-economic inequalities in mortality” in Europe emerging from socio-economic indicators such as unemployment and education levels.

#### Once cleaned and transformed, data in Figure 7 are ready to be analyzed as follow: first simple linear regression will be performed in order to inspect relation between unemployment and primary education indicator

#### Figure 7. Final input data

analysis\_indicators <- social\_indicators [,c(3, 10:11, 13)]  
head(analysis\_indicators)

## # A tibble: 6 x 4  
## Unemployment\_Rate Primary\_education~ Higher\_education\_~ Median\_Incom\_LO~  
## <dbl> <dbl> <dbl> <dbl>  
## 1 9.9 0.341 0.516 1.49  
## 2 8 0.271 0.603 1.47  
## 3 4.9 0.187 0.784 1.73  
## 4 8.8 0.329 0.568 1.46  
## 5 2.1 0.129 0.853 1.83  
## 6 3.6 0.0745 0.890 1.78

#### As an intermediate step towards regression analysis, Pearson’s correlation shows there is an evident positive correlation of 0.47 between unemployment and primary education, higher education and income from the other side are negatively corelated between -0,61 and -0.69, meaning they contribute in dropping down the unemployment.

#### Figure 8. Pearson’s Correlation matrix

library(corrplot)

## corrplot 0.84 loaded

cormat <- cor(analysis\_indicators, use = "complete.obs", method = "pearson")  
corrplot(cormat)



*Taking into consideration the unemployment rates of London LSOA areas in 2011, as the unemployed working-age population normalized by the total active population, the proposed analysis implies the null hypothesis:*

*H0: There is no statistically significant relation between unemployment rates and socio-economic indicators with significance level a >=0.05 and coefficients β1 = β2 = β3 = 0*

*and the alternative hypothesis:*

#### H1: There is relationship between unemployment with significance level a<=0,05 and at least one of the coefficients {β1, β2, β3} ≠ 0.

#### The analysis intents to prove that socio-economic indicators influence unemployment in the London area. The regression model used for this study consists on the following equations:

#### - simple regression y = β1x1 + β0

#### - multiple regression y = β1x1 + β2x2 +…+ βkxk+ β0,

#### where β are the gradients and β0 the intersect of the variables.

#### From the results in Figure 10. is evident that in terms of R-squared, relations between variables are explained on average only in 22% of the cases. P-values are strong enough to ensure the statistical significance. Beta coefficient has a positive value of 12.2, meaning each time we increase the primary education indicator by one unit, the unemployment rate will rise on average by 12.2%, in this case the model equation has the following coefficients:

#### y = 12.2(x) + 3.29

#### Figure 9. Linear regression - Primary education indicator vs Unemployment rate in LSOA

qplot(Unemployment\_Rate,   
 Primary\_education\_Indicator,   
 data = analysis\_indicators,  
 geom = "point",  
 size = I(0.1),  
 alpha =I (0.2)) +   
 stat\_smooth(method="lm",   
 se=FALSE,  
 col="red",  
 size=1)



#### Figure10. Coefficients analysis – Single linear regression

Bi\_variate\_model <- lm(analysis\_indicators$Unemployment\_Rate ~ analysis\_indicators$Primary\_education\_Indicator)  
summary(Bi\_variate\_model)

##   
## Call:  
## lm(formula = analysis\_indicators$Unemployment\_Rate ~ analysis\_indicators$Primary\_education\_Indicator)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.6155 -1.9652 -0.2512 1.8012 18.7830   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 3.2928 0.1181  
## analysis\_indicators$Primary\_education\_Indicator 12.2386 0.3254  
## t value Pr(>|t|)   
## (Intercept) 27.87 <2e-16 \*\*\*  
## analysis\_indicators$Primary\_education\_Indicator 37.61 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3 on 4833 degrees of freedom  
## Multiple R-squared: 0.2264, Adjusted R-squared: 0.2263   
## F-statistic: 1415 on 1 and 4833 DF, p-value: < 2.2e-16

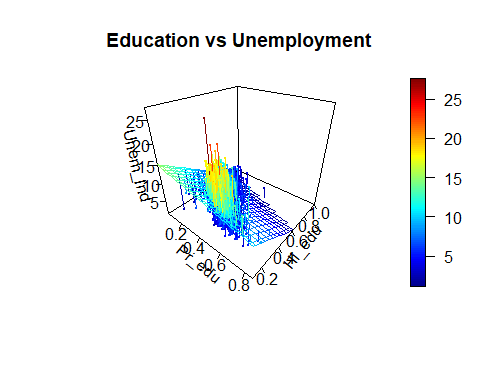
#### The next investigation consists on adding the higher education variable to the model that now shows a strong increase, peaking almost 40% of variation in unemployment rates (Figure 12). The coefficients indicate a negative value of -6.41 for primary education and -17.18 for higher education, meaning that in this model both contribute on bringing down the unemployment:

#### y = -6.41(x) - 17.18(x) + 17.04

#### The three variables are plotted on a 3D graph in order to facilitate their interpretation. The 3D regression model in Figure 11 shows clearly that the predicted plane is inclined when higher education rises, meaning that more qualified workers are in the system, less unemployment rates would be registered.

#### Figure 11. 3D Multilinear regression

## 3D Plot  
  
  
##http://www.sthda.com/english/wiki/impressive-package-for-3d-and-4d-graph-r-software-and-data-visualization   
  
library("plot3D")  
  
# x, y and z coordinates  
x <- I2  
y <- I3  
z <- D1  
# Compute the linear regression (z = ax + by + d)  
fit <- lm(z ~ x + y)  
# predict values on regular xy grid  
grid.lines = 20  
x.pred <- seq(min(x), max(x), length.out = grid.lines)  
y.pred <- seq(min(y), max(y), length.out = grid.lines)  
xy <- expand.grid( x = x.pred, y = y.pred)  
z.pred <- matrix(predict(fit, newdata = xy),   
 nrow = grid.lines, ncol = grid.lines)  
# fitted points for droplines to surface  
fitpoints <- predict(fit)  
  
# scatter plot with regression plane  
f<-list(  
 family="Courier New, monospace",  
 size=14,  
 color = "#7f7f7f")  
  
scatter3D(x, y, z, pch = 20, angle=45, cex = 0.1,   
 theta = 40, phi = 30, ticktype = "detailed",  
 xlab = "Pr\_edu", ylab = "Hi\_edu", zlab = "Unem\_Ind",   
 surf = list(x = x.pred, y = y.pred, z = z.pred,font=f,   
 facets = NA, fit = fitpoints),   
 main = "Education vs Unemployment",  
 axis.scale = FALSE)



#### Figure 12. Coefficients analysis – Multiple linear regression

Multi\_model2 <- lm(analysis\_indicators$Unemployment\_Rate ~ analysis\_indicators$Primary\_education\_Indicator + analysis\_indicators$Higher\_education\_Indicator)  
summary(Multi\_model2)

##   
## Call:  
## lm(formula = analysis\_indicators$Unemployment\_Rate ~ analysis\_indicators$Primary\_education\_Indicator +   
## analysis\_indicators$Higher\_education\_Indicator)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.5686 -1.6670 -0.0984 1.4823 16.9887   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 17.0419 0.3963  
## analysis\_indicators$Primary\_education\_Indicator -6.4108 0.5934  
## analysis\_indicators$Higher\_education\_Indicator -17.1819 0.4775  
## t value Pr(>|t|)   
## (Intercept) 43.01 <2e-16 \*\*\*  
## analysis\_indicators$Primary\_education\_Indicator -10.80 <2e-16 \*\*\*  
## analysis\_indicators$Higher\_education\_Indicator -35.98 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.664 on 4832 degrees of freedom  
## Multiple R-squared: 0.3899, Adjusted R-squared: 0.3896   
## F-statistic: 1544 on 2 and 4832 DF, p-value: < 2.2e-16

#### The final regression model comprises all variables and demonstrates an additional increase to the R-squared values reflecting 54% of fit goodness on variables. All three variables have statistical significance near to 100% as the p-value is well below the significance border for which a=0.05. The null hypothesis is rejected as all coefficients are different from zero and the alternative hypothesis is accepted:

#### y = 4.59(x) - 2.39(x) -12.14(x) + 24.81

#### The final multi-regression equation reflects a precise social model on which the unemployment grows on average by 4.59% each time unqualified workers percentage rises by 1%, drops down by 2,39% each time the higher specialized workers increase by 1% and falls again down by 12.14% each time on average, median income per workers increase by 1 unit.

#### Figure 13. Final social model – multiple linear regression with 3 independent variables

Final\_model <- lm(analysis\_indicators$Unemployment\_Rate ~ analysis\_indicators$Primary\_education\_Indicator + analysis\_indicators$Higher\_education\_Indicator + analysis\_indicators$Median\_Incom\_LOG10)  
summary(Final\_model)

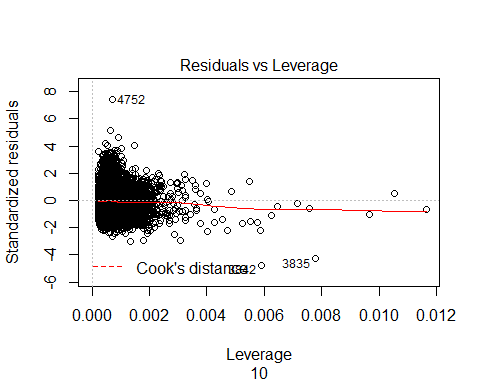
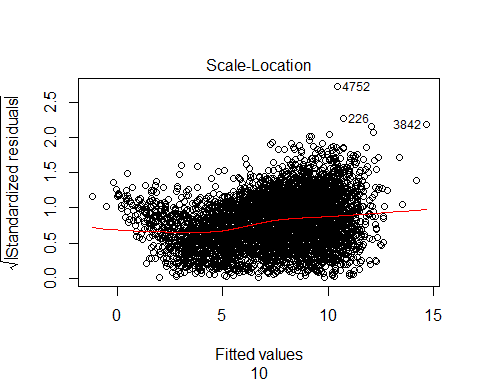
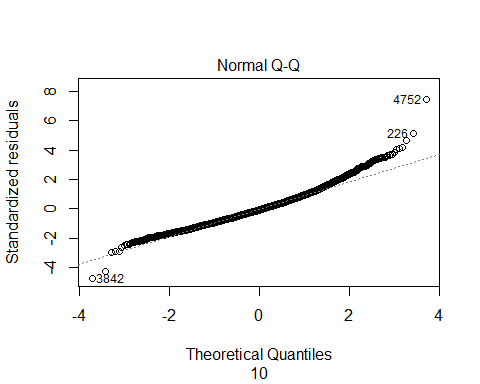
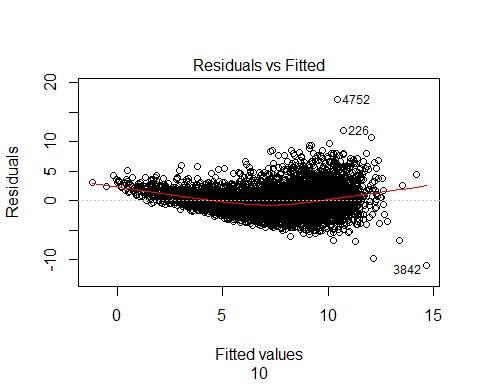
##   
## Call:  
## lm(formula = analysis\_indicators$Unemployment\_Rate ~ analysis\_indicators$Primary\_education\_Indicator +   
## analysis\_indicators$Higher\_education\_Indicator + analysis\_indicators$Median\_Incom\_LOG10)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.946 -1.579 -0.186 1.322 17.153   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) 24.8107 0.3952  
## analysis\_indicators$Primary\_education\_Indicator 4.5919 0.5842  
## analysis\_indicators$Higher\_education\_Indicator -2.3954 0.5561  
## analysis\_indicators$Median\_Incom\_LOG10 -12.1439 0.3047  
## t value Pr(>|t|)   
## (Intercept) 62.782 < 2e-16 \*\*\*  
## analysis\_indicators$Primary\_education\_Indicator 7.860 4.68e-15 \*\*\*  
## analysis\_indicators$Higher\_education\_Indicator -4.308 1.68e-05 \*\*\*  
## analysis\_indicators$Median\_Incom\_LOG10 -39.859 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.312 on 4831 degrees of freedom  
## Multiple R-squared: 0.5409, Adjusted R-squared: 0.5406   
## F-statistic: 1897 on 3 and 4831 DF, p-value: < 2.2e-16

#### 

#### As a final proof, in Figure 14, residuals standardized errors are plotted with predicted values in order to underpin possible suspect clustering around variables. The overall distribution seems to be casual and no patterns are happening.

plot(Final\_model, s=10)

#### Figure 14. Standardized values vs Fitted values Plot



## Local regression

#### The second phase of the study consists on joining the findings of the multi-linear regression model and conduct a statistical spatial analysis with “local regression”. The shapefile used for the last part of this study derives from the data.gov.uk website, it covers all LSOA areas in England and Wales. Once clipped on London perimeter, the shapefile was merged with the data analyzed to this point (data.gov.uk, 2018).

library(rgdal)

## Loading required package: sp

## rgdal: version: 1.3-4, (SVN revision 766)  
## Geospatial Data Abstraction Library extensions to R successfully loaded  
## Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20  
## Path to GDAL shared files: C:/Users/Bogdan/Documents/R/win-library/3.5/rgdal/gdal  
## GDAL binary built with GEOS: TRUE   
## Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ\_VERSION: 493]  
## Path to PROJ.4 shared files: C:/Users/Bogdan/Documents/R/win-library/3.5/rgdal/proj  
## Linking to sp version: 1.3-1

LondonLSOA <- readOGR("C:\\Users\\Bogdan\\Documents\\UCL\\MRes\\GIS\\Week9\\Assignment3\\Lower\_Layer\_Super\_Output\_Areas\_December\_2011\_Full\_Extent\_\_Boundaries\_in\_England\_and\_Wales\\london\_lsoa.shp", "london\_lsoa")

## OGR data source with driver: ESRI Shapefile   
## Source: "C:\Users\Bogdan\Documents\UCL\MRes\GIS\Week9\Assignment3\Lower\_Layer\_Super\_Output\_Areas\_December\_2011\_Full\_Extent\_\_Boundaries\_in\_England\_and\_Wales\london\_lsoa.shp", layer: "london\_lsoa"  
## with 4835 features  
## It has 6 fields  
## Integer64 fields read as strings: objectid

##In order to Print a more accurate map, we need to add the lon and lat columns to the shapefile  
  
C <- coordinates(LondonLSOA)  
  
LondonLSOA$lon <- C[,1]   
LondonLSOA$lat <- C[,2]

# The csv file is appended to the shapefile by the codes column  
  
library(tmaptools)  
LondonLSOA\_Joined <- append\_data(LondonLSOA,clean\_data, key.shp = "lsoa11cd", key.data ="Codes", ignore.duplicates = TRUE)

## Keys match perfectly.

LondonLSOA\_Joined

## class : SpatialPolygonsDataFrame   
## features : 4835   
## extent : 503568.1, 561957.4, 155850.8, 200933.9 (xmin, xmax, ymin, ymax)  
## coord. ref. : +proj=tmerc +lat\_0=49 +lon\_0=-2 +k=0.9996012717 +x\_0=400000 +y\_0=-100000 +datum=OSGB36 +units=m +no\_defs +ellps=airy +towgs84=446.448,-125.157,542.060,0.1502,0.2470,0.8421,-20.4894   
## variables : 15  
## names : objectid, lsoa11cd, lsoa11nm, lsoa11nmw, st\_areasha, st\_lengths, lon, lat, Names, Unemployment\_Rate, No\_qualifications, Level\_1\_qualifications, Level\_4\_qualifications\_and\_above, Med\_Househ\_Inc\_estim, Working\_age\_population   
## min values : 1, E01000001, Barking and Dagenham 001A, Barking and Dagenham 001A, 18362.02, 689.2264, 504715.5, 156926.6, Barking and Dagenham 001A, 1.2, 11, 20, 88, 16166.85, 608   
## max values : 999, E01033746, Westminster 024F, Westminster 024F, 15797244.73, 23801.5004, 559643.0, 199816.4, Westminster 024F, 27.6, 665, 326, 1366, 92431.29, 4235

library(sf)

## Linking to GEOS 3.6.1, GDAL 2.2.3, proj.4 4.9.3

LondonLSOA\_SF <- st\_as\_sf(LondonLSOA\_Joined)  
LondonLSOA\_SF

## Simple feature collection with 4835 features and 15 fields  
## geometry type: MULTIPOLYGON  
## dimension: XY  
## bbox: xmin: 503568.1 ymin: 155850.8 xmax: 561957.4 ymax: 200933.9  
## epsg (SRID): NA  
## proj4string: +proj=tmerc +lat\_0=49 +lon\_0=-2 +k=0.9996012717 +x\_0=400000 +y\_0=-100000 +ellps=airy +towgs84=446.448,-125.157,542.06,0.1502,0.247,0.8421,-20.4894 +units=m +no\_defs  
## First 10 features:  
## objectid lsoa11cd lsoa11nm lsoa11nmw  
## 0 1 E01000001 City of London 001A City of London 001A  
## 1 2 E01000002 City of London 001B City of London 001B  
## 2 3 E01000003 City of London 001C City of London 001C  
## 3 4 E01000005 City of London 001E City of London 001E  
## 4 5 E01000006 Barking and Dagenham 016A Barking and Dagenham 016A  
## 5 6 E01000007 Barking and Dagenham 015A Barking and Dagenham 015A  
## 6 7 E01000008 Barking and Dagenham 015B Barking and Dagenham 015B  
## 7 8 E01000009 Barking and Dagenham 016B Barking and Dagenham 016B  
## 8 9 E01000010 Barking and Dagenham 015C Barking and Dagenham 015C  
## 9 10 E01000011 Barking and Dagenham 016C Barking and Dagenham 016C  
## st\_areasha st\_lengths lon lat Names  
## 0 129865.34 2635.781 532150.8 181617.5 City of London 001A  
## 1 228419.33 2708.052 532443.4 181643.6 City of London 001B  
## 2 59054.01 1224.771 532205.5 182030.0 City of London 001C  
## 3 189577.17 2275.832 533620.0 181152.8 City of London 001E  
## 4 146536.52 1966.162 544933.4 184296.5 Barking and Dagenham 016A  
## 5 200093.95 2949.773 544142.0 184437.7 Barking and Dagenham 015A  
## 6 216901.70 2933.076 543627.4 184387.6 Barking and Dagenham 015B  
## 7 127961.81 2159.015 544601.7 184456.5 Barking and Dagenham 016B  
## 8 352068.28 3245.349 544228.2 184047.2 Barking and Dagenham 015C  
## 9 91630.55 1543.255 544384.2 184738.4 Barking and Dagenham 016C  
## Unemployment\_Rate No\_qualifications Level\_1\_qualifications  
## 0 3.5 25 35  
## 1 1.7 30 41  
## 2 4.7 161 81  
## 3 8.7 200 67  
## 4 9.5 238 155  
## 5 11.9 161 145  
## 6 18.0 220 155  
## 7 10.0 272 149  
## 8 9.7 268 169  
## 9 10.2 229 157  
## Level\_4\_qualifications\_and\_above Med\_Househ\_Inc\_estim  
## 0 1047 58347.32  
## 1 1024 57159.50  
## 2 706 34386.56  
## 3 283 23999.85  
## 4 431 39140.64  
## 5 342 26228.95  
## 6 252 24203.85  
## 7 511 32464.43  
## 8 982 31630.20  
## 9 375 28128.96  
## Working\_age\_population geometry  
## 0 1082 MULTIPOLYGON (((532104.9 18...  
## 1 1024 MULTIPOLYGON (((532746.8 18...  
## 2 988 MULTIPOLYGON (((532293.1 18...  
## 3 694 MULTIPOLYGON (((533499.1 18...  
## 4 1176 MULTIPOLYGON (((545271.9 18...  
## 5 1000 MULTIPOLYGON (((544180.3 18...  
## 6 1002 MULTIPOLYGON (((543610.2 18...  
## 7 1281 MULTIPOLYGON (((544667.1 18...  
## 8 2105 MULTIPOLYGON (((544440.4 18...  
## 9 1179 MULTIPOLYGON (((544667.1 18...

## Copy the residuals to the joined dataframe  
LondonLSOA\_Joined@data$residuals <- Final\_model$residuals  
  
  
##Tranform data to Dataframe format for GGPLOT usage  
  
LondonLSOA\_Joined@data$id <- rownames(LondonLSOA\_Joined@data)  
  
  
social\_unemployment <- fortify(LondonLSOA\_Joined, region = "id")  
LondonLSOA\_DF <- merge(social\_unemployment, LondonLSOA\_Joined@data, by = "id")  
  
cbind(lapply(LondonLSOA\_DF, class))

## [,1]   
## id "character"  
## long "numeric"   
## lat.x "numeric"   
## order "integer"   
## hole "logical"   
## piece "factor"   
## group "factor"   
## objectid "factor"   
## lsoa11cd "factor"   
## lsoa11nm "factor"   
## lsoa11nmw "factor"   
## st\_areasha "numeric"   
## st\_lengths "numeric"   
## lon "numeric"   
## lat.y "numeric"   
## Names "character"  
## Unemployment\_Rate "numeric"   
## No\_qualifications "integer"   
## Level\_1\_qualifications "integer"   
## Level\_4\_qualifications\_and\_above "integer"   
## Med\_Househ\_Inc\_estim "numeric"   
## Working\_age\_population "integer"   
## residuals "numeric"

head(LondonLSOA\_DF)

## id long lat.x order hole piece group objectid lsoa11cd  
## 1 0 532104.9 182011.9 1 FALSE 1 0.1 1 E01000001  
## 2 0 532105.3 182010.6 2 FALSE 1 0.1 1 E01000001  
## 3 0 532125.7 181949.8 3 FALSE 1 0.1 1 E01000001  
## 4 0 532126.3 181948.0 4 FALSE 1 0.1 1 E01000001  
## 5 0 532127.2 181943.3 5 FALSE 1 0.1 1 E01000001  
## 6 0 532129.3 181936.7 6 FALSE 1 0.1 1 E01000001  
## lsoa11nm lsoa11nmw st\_areasha st\_lengths lon  
## 1 City of London 001A City of London 001A 129865.3 2635.781 532150.8  
## 2 City of London 001A City of London 001A 129865.3 2635.781 532150.8  
## 3 City of London 001A City of London 001A 129865.3 2635.781 532150.8  
## 4 City of London 001A City of London 001A 129865.3 2635.781 532150.8  
## 5 City of London 001A City of London 001A 129865.3 2635.781 532150.8  
## 6 City of London 001A City of London 001A 129865.3 2635.781 532150.8  
## lat.y Names Unemployment\_Rate No\_qualifications  
## 1 181617.5 City of London 001A 3.5 25  
## 2 181617.5 City of London 001A 3.5 25  
## 3 181617.5 City of London 001A 3.5 25  
## 4 181617.5 City of London 001A 3.5 25  
## 5 181617.5 City of London 001A 3.5 25  
## 6 181617.5 City of London 001A 3.5 25  
## Level\_1\_qualifications Level\_4\_qualifications\_and\_above  
## 1 35 1047  
## 2 35 1047  
## 3 35 1047  
## 4 35 1047  
## 5 35 1047  
## 6 35 1047  
## Med\_Househ\_Inc\_estim Working\_age\_population residuals  
## 1 58347.32 1082 2.889527  
## 2 58347.32 1082 2.889527  
## 3 58347.32 1082 2.889527  
## 4 58347.32 1082 2.889527  
## 5 58347.32 1082 2.889527  
## 6 58347.32 1082 2.889527

#### Before starting the spatial regression analysis, is important to calculate the spatial autocorrelation between variables. Moran’s I test calculates the extent of similarity between LSOA variables and their surroundings as possible spatial autocorrelation may invalid the model. Applied to spatial analysis, Moran’s I is built on a weighted matrix that in this case is created from each LSOA center and associated with the residual errors calculated within the model.The folowing codes are adapted from Dennett (2018).

library(spdep)

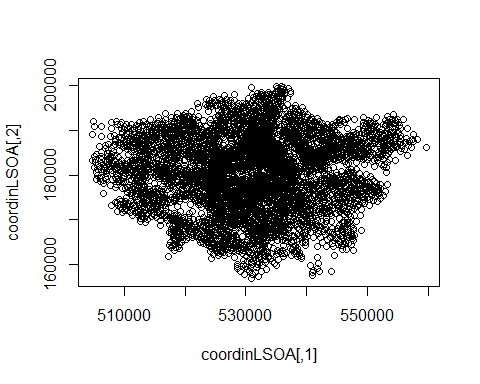
## Loading required package: Matrix

## Loading required package: spData

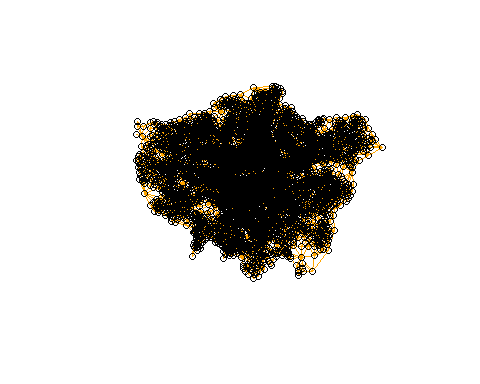
##   
## Attaching package: 'spData'

## The following objects are masked \_by\_ '.GlobalEnv':  
##   
## x, y

# The centroids of each LSOA areas are used for the matrix construction  
coordinLSOA <- coordinates(LondonLSOA\_Joined)  
plot(coordinLSOA)+  
coord\_equal()



#The spatial matrix is constructed using the original shapefile  
LSOA\_nb <- poly2nb(LondonLSOA\_Joined, queen=T)  
  
  
plot(LSOA\_nb, coordinates(coordinLSOA), col="orange")



#A spatial weighted object is created  
LSOA.lw <- nb2listw(LSOA\_nb, style="C")  
head(LSOA.lw$neighbours)

## [[1]]  
## [1] 2 3 4675 4676 4751  
##   
## [[2]]  
## [1] 1 3 2831 4675  
##   
## [[3]]  
## [1] 1 2 2831 2832 4751  
##   
## [[4]]  
## [1] 4300 4301 4313 4316 4675 4684  
##   
## [[5]]  
## [1] 8 42 65 68  
##   
## [[6]]  
## [1] 7 8 9 10 11

#### Moran’s I test in Figure 16 indicates a value of -0.25 which means LSOA variables after regression are dispersed enough to form clusters. Residuals plot in the Figure 17 show no particular autocorrelation between LSOA areas.

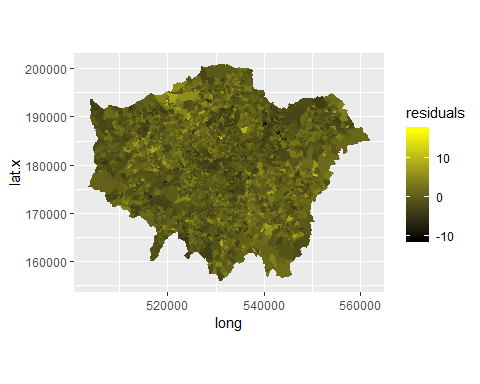
#### Figure 16. Moran’s I test

I\_LSOA\_Global\_Density <- moran.test(LondonLSOA\_Joined@data$residuals, LSOA.lw)  
  
I\_LSOA\_Global\_Density

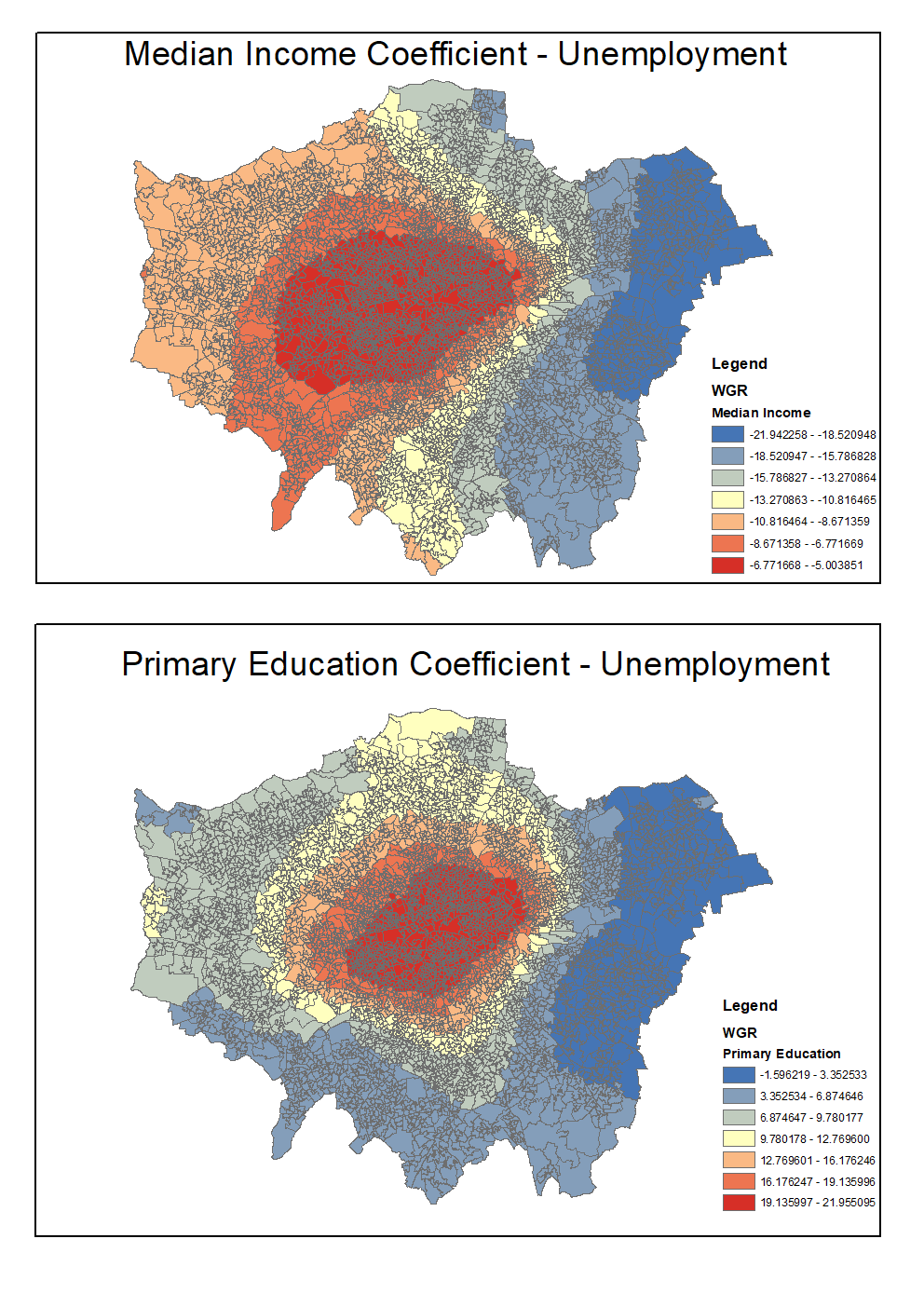
##   
## Moran I test under randomisation  
##   
## data: LondonLSOA\_Joined@data$residuals   
## weights: LSOA.lw   
##   
## Moran I statistic standard deviate = 30.931, p-value < 2.2e-16  
## alternative hypothesis: greater  
## sample estimates:  
## Moran I statistic Expectation Variance   
## 2.569726e-01 -2.068680e-04 6.913171e-05

#### Figure 17. Residuals Plot

## Plotting Residuals on Morans's I  
map3 <- ggplot(data = LondonLSOA\_DF,  
 aes(x=long,y=lat.x))+  
 geom\_polygon(aes(group=group, fill = residuals))+  
 scale\_fill\_gradient(low = "black", high = "yellow")+  
 coord\_equal()   
map3

*When proceeding with the GWR analysis in ArcGIS, the model faced some critical issues, running an Ordinary Least Square analysis revealed that the model has one variable indicating “multicollinearity” above 7.5 which prevents the GWR algorithm to run, in this case the variance inflation factor (VIF) indicates higher education variables to have a value of 7.59, it was necessary to exclude this variable from the model in order to further investigate the WGR.*

#### Figure 18. WGR Coefficients



#### The two maps illustrate in Figure 18 the respective coefficients having different impact in London LSOA areas. Primary Education has a positive impact on unemployment, especially in the center of London, for 1% increase on primary educated workers the unemployment rises between 19-21% on the red areas. East London faces an opposite direction, surprisingly, primary education drops down the unemployment, this may be the effect of high level of unqualified workers in that area.

#### Median income coefficients have an overall negative impact on unemployment, meaning that on average it contributes to decrease unemployment. Figure 19 illustrates in red those areas that pull down the unemployment between 5-6%. The impact of higher households’ income is even more evident in East London, where the unemployment percentage would be affected between -15% and -21% each time the income values rise by 1 unit.

## Conclusion and further directions

#### This study has analyzed different socio-economic factors that may influence unemployment. Two levels of education between workers indicate a high degree of relationship but both inserted on the regression model suggests important levels of multicollinearity which in turn had negatively interfered with individual predictors. Primary education among workers have a huge positive impact on unemployment in the geographical center of London, meaning indirectly that LSOA areas with high needs for specialized workers suffer the injection of low qualified workers. Rising median households’ income from the other side has a negative impact on every LSOA areas, west and central London clearly show a softer impact compared to east side, where poorer areas can lose unemployment with higher wages.

#### Findings reveal that unemployment in London’s LSOA areas is directly related to primary education incidence, but a more detailed dataset could give a wider picture about the negative impact on unemployment in east London. Median households’ income indicates a clear edge between west and east London, further analysis may also explain this division.

## Bibliography

#### Dennett, A. (2018). ‘Spatial Inferential Statistics’. Available at:<https://rpubs.com/adam_dennett/334459>. (Accessed: 18 Dicember 2018).

#### Borrell, C., Marí-Dell’olmo, M., Palència, L., Gotsens, M., Burström, B., Domínguez-Berjón, F., Rodríguez-Sanz, M., Dzúrová, D., Gandarillas, A., Hoffmann, R., Kovacs, K., Marinacci, C., Martikainen, P., Pikhart, H., Corman, D., Rosicova, K., Saez, M., Santana, P., Tarkiainen, L., Puigpinós, R., Morrison, J., Pasarín, M.I., Díez, È., (2014). ‘Socioeconomic inequalities in mortality in 16 European cities’. Scand J Public Health 42, pp. 245–254. doi: <https://doi.org/10.1177/1403494814522556>.

#### Data.gov.uk (2018). Lower Layer Super Output Area (LSOA) boundaries. Available at: <https://data.gov.uk/dataset/fa883558-22fb-4a1a-8529-cffdee47d500/lower-layer-super-output-area-lsoa-boundaries>. (Accessed: 10 November 2018).

#### INEQ-CITIES (2018). Available at: <https://www.ucl.ac.uk/ineqcities>. (Accessed: 20 November 2018).

#### London Datastore (2018) LSOA Atlas. Available at: <https://data.london.gov.uk/dataset/lsoa-atlas>. (Accessed: 10 November 2018).

#### Mincer, J., (1991). ‘Education and Unemployment’. National Bureau of Economic Research, Cambridge, MA. doi: <https://doi.org/10.3386/w3838>.

#### Nickell, S., (1979). ‘Education and Lifetime Patterns of Unemployment’. Journal of Political Economy 87, pp. S117–S131. doi: <https://doi.org/10.1086/260825>.

#### OSN (2018). Census geography. Available at: <https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography>. (Accessed: 17 November 2018).

#### Zenou, Y., (2000). ‘Urban unemployment, agglomeration and transportation policies’. Journal of Public Economics 77, pp. 97–133. doi: <https://doi.org/10.1016/S0047-2727(99)00083-3>.