The relationship between childhood obesity and socioeconomic factors and its geographical variation per London borough

Programme: Smart Cities and Urban Analytics

Department: Centre for Advanced Spatial Analysis (CASA), UCL

GitHub repository URL: https://github.com/elenikalantzi/GIS_Coursework

Rpubs Website: https://rpubs.com/ekalantzi/Childhood-Obesity-London

Introduction

In recent decades there has been an increasing trend in childhood obesity, both in developing and developed countries (James, 2008). Specifically, in England almost 2/3 (63%) of adults are overweight/obese, while one in three children leave primary school being overweight/obese (Department of Health and Social Care, 2020). However, the problem is more serious in London compared to the rest of the country (Trust for London, 2020), but also with respective cities such as Madrid, Paris, Hong Kong and Toronto (London Health Commission, 2014).

It is important to fight childhood obesity, as most children are likely to become obese after adulthood, showing other serious problems such as diabetes, hypertension, cancer, etc. (Department of Health and Social Care, 2020). In fact, recent data show that obese people are twice as likely to die from COVID-19 (Phe, 2020). Meanwhile, obesity affects psychology and social interactions, but it's also affected by a number of socio-spatial, economic cultural and environmental factors (Stamatakis, Wardle and Cole, 2010), which should be identified to help combat it.

Therefore, it should be investigated which factors, how and to what extent affect it. Thus, this study will address the following questions: (a) is childhood obesity related to socioeconomic factors such as median annual income, education level, population density, access to green spaces and the density of fast-food outlets and b) whether it is related, varies spatially per borough or remains constant throughout London?

Literature Review

Overcoming childhood obesity is an important goal for England and especially for London. It's worth mentioning that during the 2017-2018 school year 22.6% of children aged 10-11 were overweight in London, while the country average was 20.1%. It is estimated that diseases associated with overweight and obesity cost in the NHS, 6.1 billion annually (Scarborough *et al.*, 2011). Several national policies have been adopted to address the problem and halve childhood obesity by 2030, with no apparent results (Liu *et al.*, 2019). The latest data show that in 2018-2019 hospital admissions related to obesity as well as its prevalence in children were 2.5 times more likely in the more degraded areas, compared to the less ones (NHS, 2020). This means that in order to ensure public health, local authorities should adopt effective and more targeted interventions, which can arise from the study of the spatial relationships between childhood obesity and the particular characteristics and needs of each neighborhood (Procter *et al.*, 2008).

Studies in various areas have shown that poor living conditions and low socioeconomic status negatively affect obesity or increase the risk of its prevalence (O'Dea and Dibley, 2010). Income is a key factor influencing childhood obesity, as the quality of food and access to health and sports infrastructure depend on it (Edwards *et al.*, 2010). At the same time, obesity rates are higher in areas with higher illiteracy rates (Devaux *et al.*, 2011), reduced access to open/green spaces (Nielsen and Hansen, 2007), higher population density (Gordon-Larsen *et al.*, 2006) and increased number of fast-food stores (Phe, 2018). Most studies that examined the effect of these variables used mainly stationary (such as OLS regression) rather than a spatial method. However, subsequent studies have shown that spatial inequalities are directly related to BMI (Krzyzanowska and Mascie-Taylor, 2012). Therefore, there is a need to explore not only the stationary, but also the spatial variations between childhood obesity and socioeconomic conditions in London, using local regression models.

Methodology

i. Data Collection

The data that will be used for childhood obesity is its prevalence percentage and concerns the school year 2017-2018 for children aged 10-11 years. The average income for the tax year 2017 - 2018 was used as an economic variable. The variable of population density refers to the number of inhabitants per km^2 for 2018. The level of education refers to the percentage of the population aged 25-64 that hasn't qualifications for 2018. The percentage of households with deficient access to nature in 2012 was selected as an environmental variable. The density of fast-food outlets was measured as the number of fast-food outlets per 100 people (snapshot on 02/07/2018), but the measurements may change slightly over time as fast-food stores open and close. In total, the above data and shapefiles for the 33 different boroughs of London were collected by the London Datastore and processed at RStudio.

ii. Regression methods

Ordinary least squares (OLS) is the most common regression technique and has been applied in various geographical areas and scales (Han, 2011). The present study attempts to examine the relationship between childhood obesity (dependent variable) and the factors that affect it, applying an OLS multiple linear regression model (confidence interval $\alpha = 0.05$), which is based on the assumption that all relationships are static in study area. The existence of a linear dependence of Y on a subset of the controlled variables is ascertained through the statistical control of the hypotheses:

- H0: $\beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = 0$
- H1: $\beta 1 * \beta 2 * \beta 3 * \beta 4 * \beta 5 \neq 0$

However, a global model cannot always accurately describe the spatial heterogeneity of the relationship between variables. For this reason, three more models will be applied: the Spatial Lagged Y Model, the Spatial Error Model and the local Geographically Weighted Regression Model.

The Spatial lagged y model (SLYM) is appropriate when we think that the values of y in a unit i are directly affected by the values of y located in the "neighbours" of i. Thus, the variable y is converted to a spatially lagged version, versus the independent variables, and the regression equation is converted as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 x_i + \beta_4 x_i + \beta_5 x_i + \rho w_i y_i + \varepsilon_i$$

where w_i is the vector of all neighbouring boroughs for each i borough. If $\rho w_i y_i$ parameter is positive, it indicates that the percentage of obese children is expected to be higher in a borough, if neighbouring boroughs also have high values (Ward and Gleditsch, 2011).

Another way of examining the spatial dependence in regression models is the spatial error model (SEM) which assumes that the errors of a model are spatially related. Its equation is as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 x_i + \beta_4 x_i + \beta_5 x_i + \lambda w_i \xi_i + \varepsilon_i$$

where λ is a measure of the spatial autocorrelation of the residuals for neighbouring residuals and ξ is the spatial component of the error term (Ward and Gleditsch, 2011). If there is no spatial autocorrelation in the residuals, we would expect the parameter $\lambda=0$ and the other parameters in the equation to be the same as the OLS regression.

The Geographically Weighted Regression Model is a local model that based on non – stationary variables and explores the geographical variation of the relationship between the dependent variable and independent ones. The general GWR equation is:

$$y_i(u) = \beta_{0i}\,(u) + \beta_{1i}(u)x_{1i} + \beta_{2i}(u)x_{2i} + \ldots + \beta_{ni}(u)x_{ni}$$

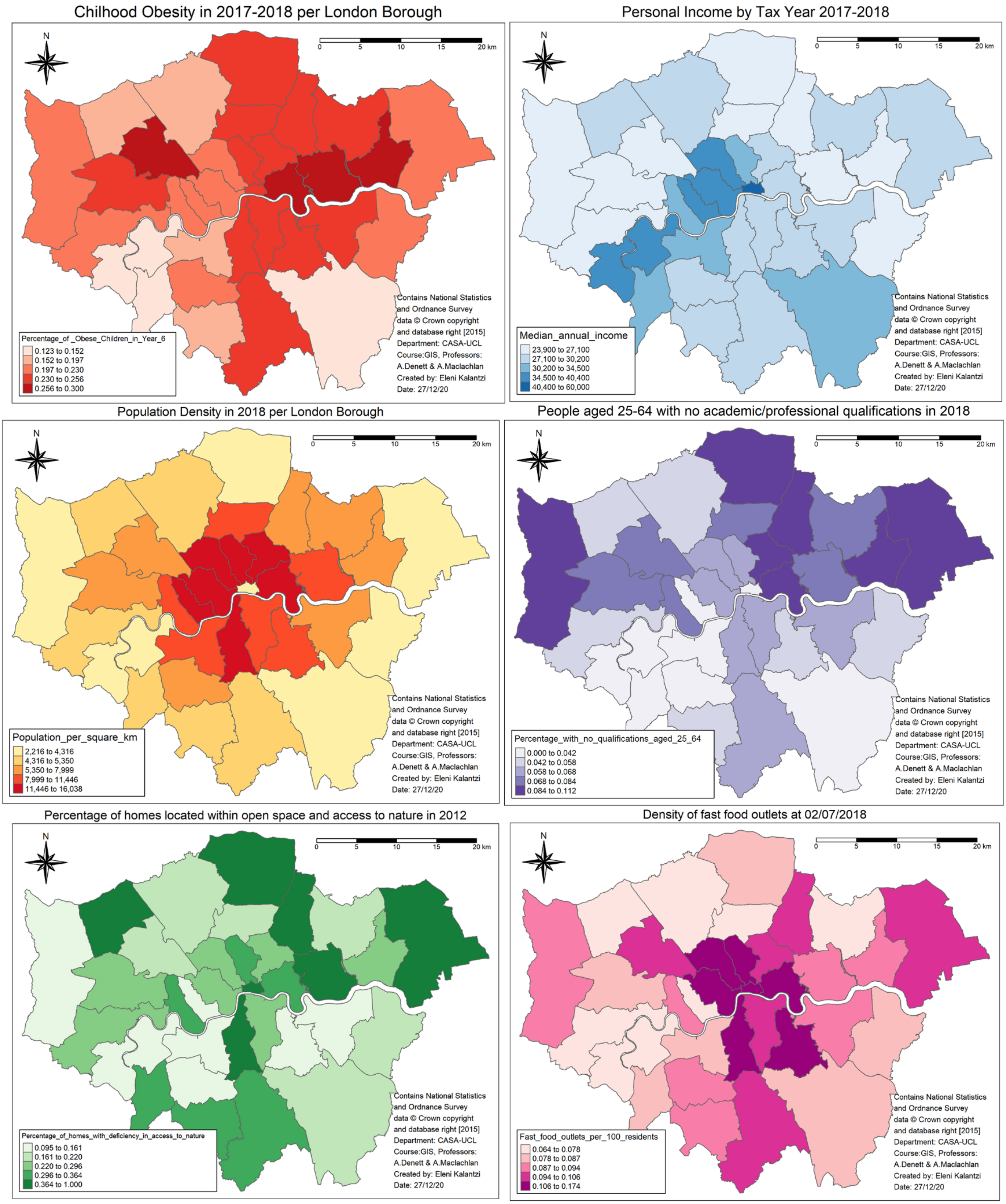
where y is childhood obesity in a location (u) and a set (n) of independent variables x in the same location and β describes a relation around location (u) and is specific to that position (Fotheringham, Brunsdon and Charlton, 2002). In this study, a fixed Gaussian kernel will be used. Core bandwidth was determined by minimizing the Akaike Information Criterion (AIC) for the GWR model. The model with the smallest AIC provides the closest approach to reality.

Another useful indicator of spatial data regression modelling problems is the spatial autocorrelation analysis shown by the model residuals (Schuurman, Peters and Oliver, 2009). Therefore, we will compare the spatial correlation by calculating the Moran's I index. Moran's I values range from +1 to -1 and indicate whether we have clustered (close to 1) or dispersed values (close to -1) or no relationship (0).

Results

i. OLS Regression Models

The research data show that the mean percentage of obese children in year 6 in London was 22.67%, with minimum and maximum values of 12.26% and 30% respectively. Map Collage 1 displays the spatial variation in the rate of childhood obesity and the distribution of the explanatory variables. These maps present that there is a wide range of values for each variable in the study area and allow us to make comparisons.



Map Collage 1: Visualisation of the dependent (Childhood Obesity) and independent variables

From the results of ANOVA (Table 1) we observe that R_{adj}^2, which is a common criterion of the adequacy of the model as it doesn't necessarily increase with the addition of independent variables, equals to 42.3%, which is relatively small. Additionally, from the independent variables, only the percentage of houses with deficient access to nature seems to be statistically significant (p-value = 0.044 <0.05). This means that we should remove or modify some variables, so that there is a better fit of the model to the data.

Table 1: ANOVA results for regression model 1

```
Residuals:
    Min      1Q  Median      3Q     Max 
-5.4430 -1.9171  0.1383  1.5530  5.3451 

Coefficients:
                                                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                               18.1614443  4.5100724   4.027 0.000412 ***
LonBoroughs$Median_annual_income                          -0.0001944  0.0001307  -1.488 0.148405    
LonBoroughs$Percentage_with_no_qualifications_aged_25_64   0.4980465  0.3108712   1.602 0.120770    
LonBoroughs$Population_per_square_km                       0.0002054  0.0001539   1.334 0.193285    
LonBoroughs$Percentage_of_homes_with_deficiency_in_access_to_nature  0.0895941  0.0424278   2.112 0.044110 *  
LonBoroughs$Fast_food_outlets_per_100_residents           35.3446291 29.3973482   1.202 0.239683    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.865 on 27 degrees of freedom
Multiple R-squared:  0.5131,	Adjusted R-squared:  0.423 
F-statistic: 5.691 on 5 and 27 DF,  p-value: 0.001039
```

To check if the used variables need modifications, we should examine whether they follow the normal distribution. Starting with the dependent variable we notice that although is slightly shifted it seems to follow the normal distribution (Figure 2), so we choose not to modify it, since no power in Figure 1 seems to be optimal.

Figure 1: Boxplots for transformations of the dependent variable

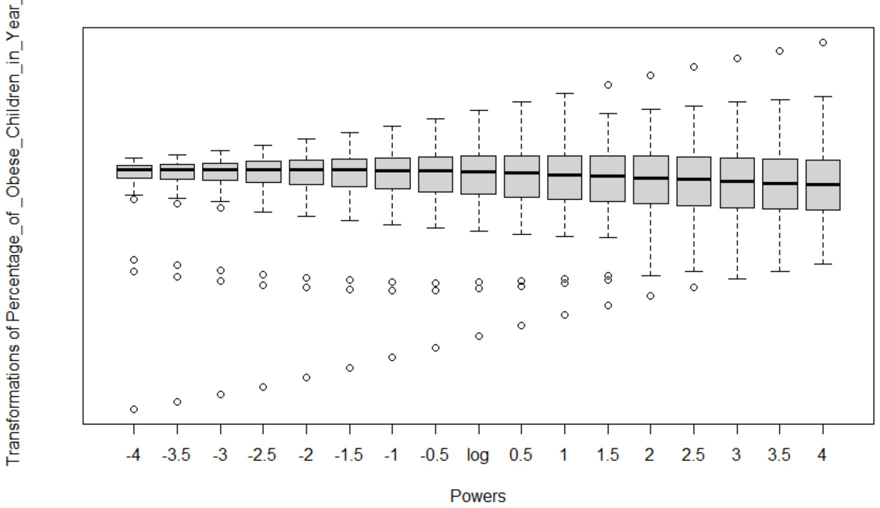
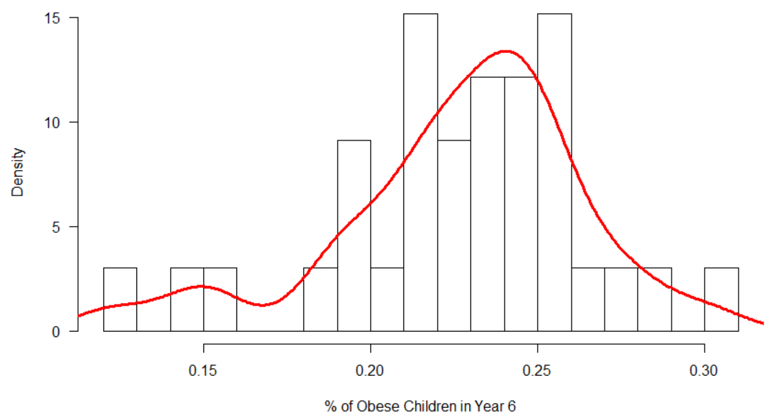


Figure 2: A histogram of the distribution of the Childhood Obesity variable



Regarding the median annual income variable, we observe that it doesn't follow the normal distribution (Figure 3) and the optimal modification is to raise it to -2.5 (Figure 4). From Figure 5, we see that neither the modified variable follows the normal distribution, but it is less skewed, so we will choose it for our model.

Figure 3: A histogram of median annual income variable

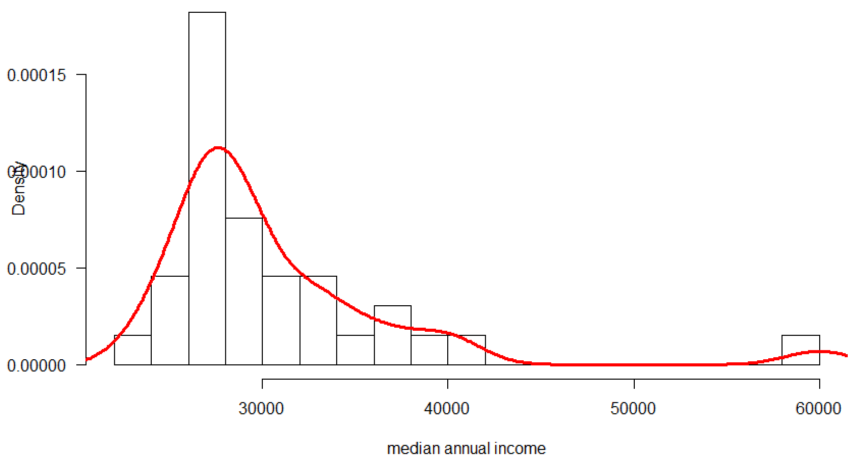


Figure 4: Boxplots for transformations of income variable

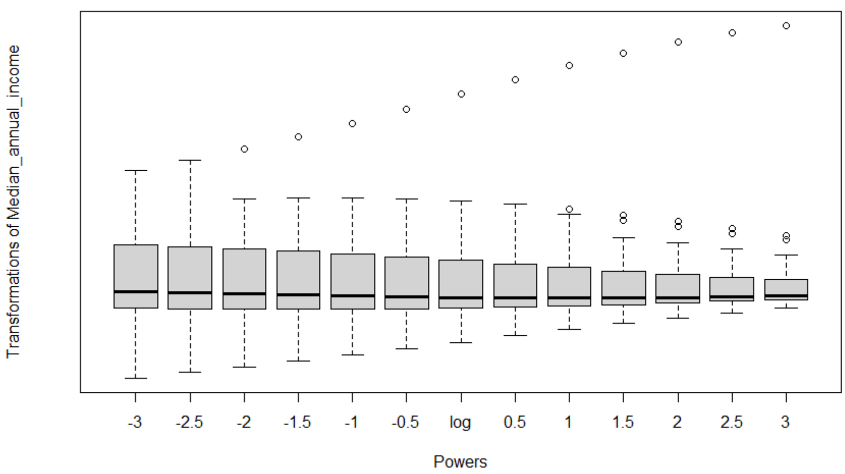
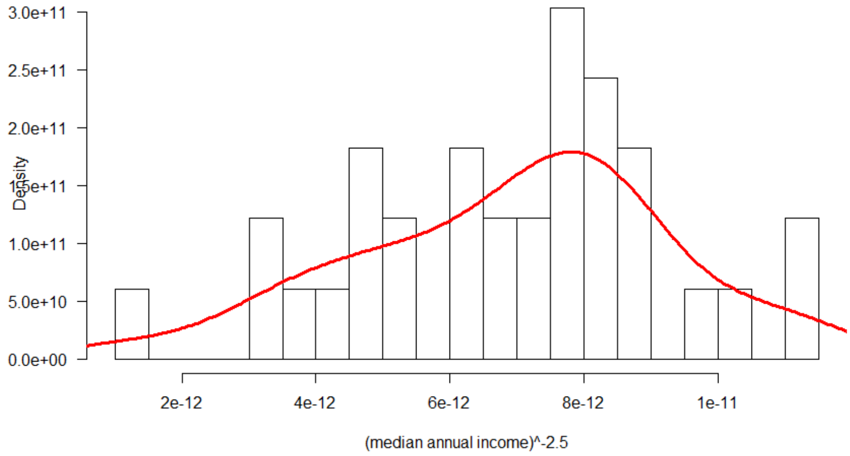


Figure 5: A histogram of (median annual income)^-2.5 variable



As for the population density variable, it doesn’t follow the normal distribution either (Figure 6), while the optimal modification is to log it (Figure 7). In Figure 8, we see that the logarithmic variable is “closer” to the normal distribution, so we will include it in the model.

Figure 6: A histogram of population density variable

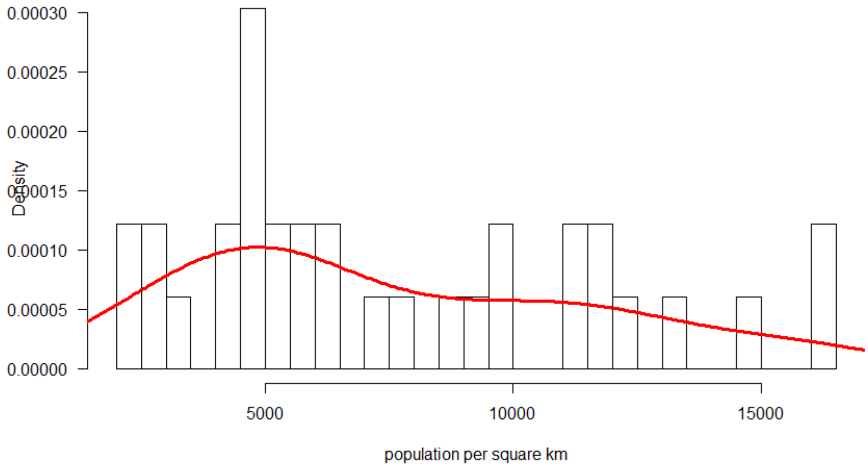


Figure 7: Boxplots for transformations of population density variable

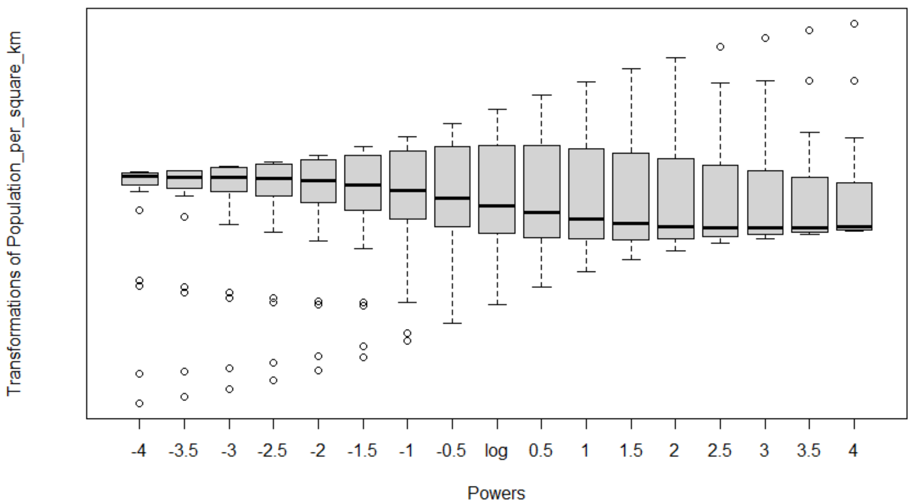
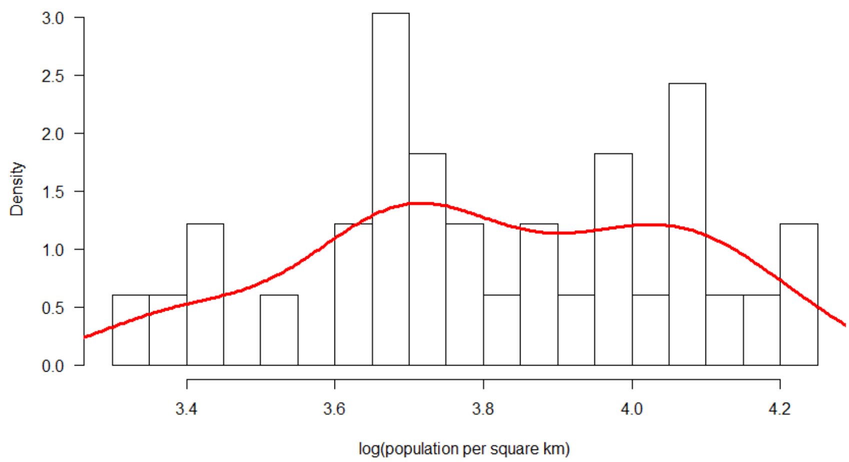
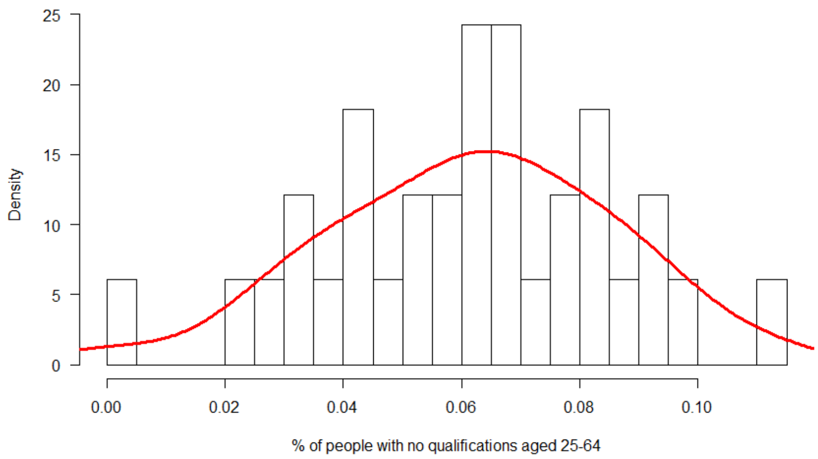


Figure 8: A histogram of log population density variable



Regarding the residents without qualifications variable, we observe that it follows the normal distribution (Figure 9) and therefore no modification is needed.

Figure 9: A histogram of people without qualifications variable



The environmental variable is right skewed (Figure 10), so according to Figure 11 we will raise it to -0.5. However, there is no significant improvement (Figure 12), so we won't modify it.

Figure 10: A histogram of access to nature variable

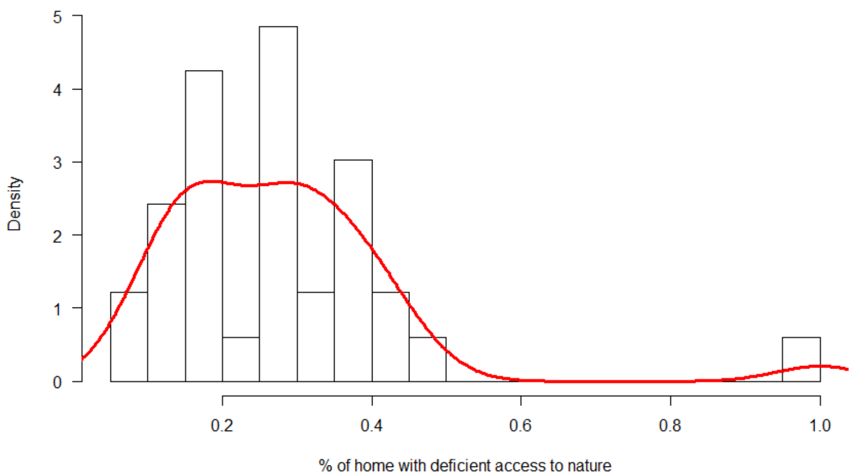


Figure 11: Boxplot for transformations of access to nature variable

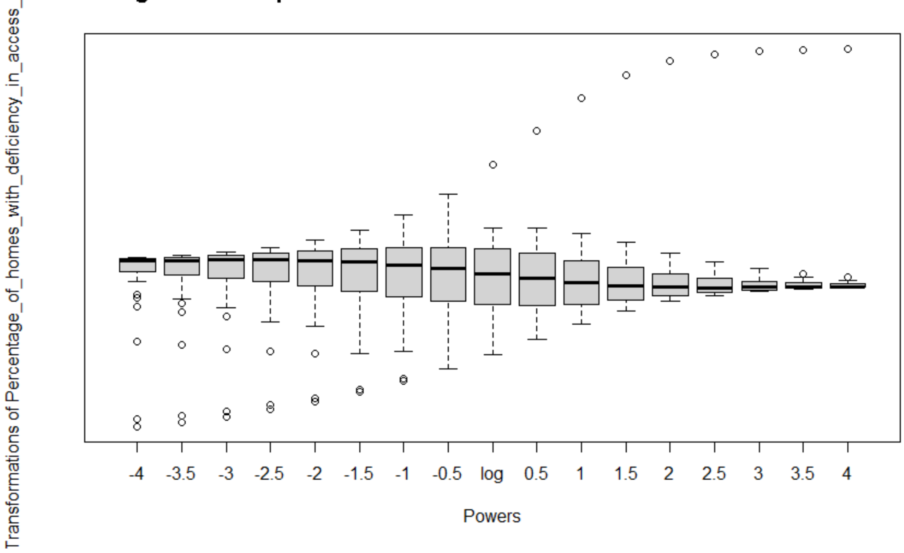
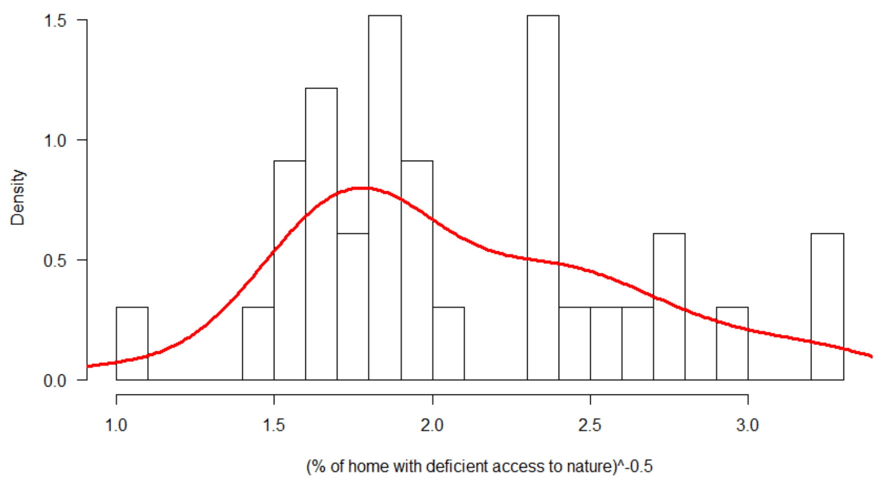


Figure 12: A histogram of (access to nature)^-0.5 variable



The variable of Fast-food outlets is also right skewed (Figure 13), so according to Figure 14 we will raise it to -1. Based on Figure 15 it seems that the new variable follows the normal distribution, so we will select it for the model.

Figure 13: A histogram of density of fast food outlets variable

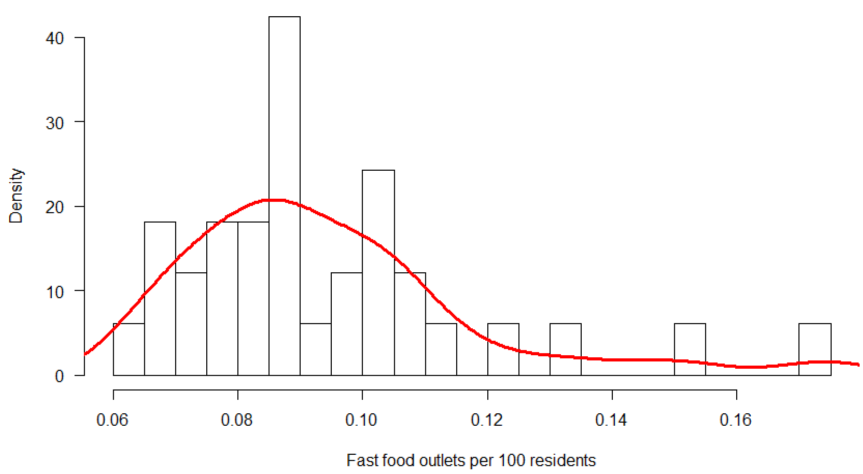


Figure 14: Boxplots for transformations of Fast Food outlets variable

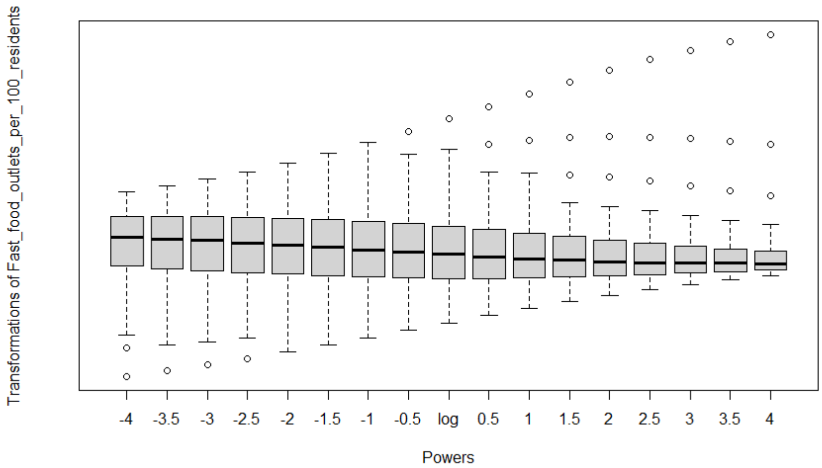
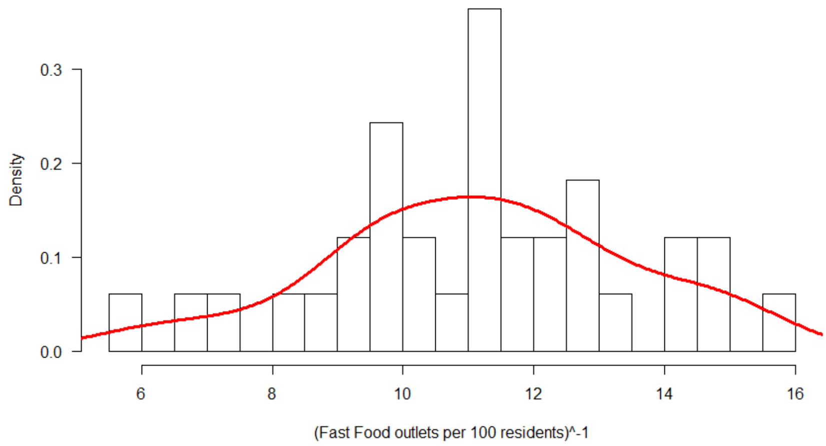


Figure 15: A histogram of density of (Fast Food outlets)^-1 variable



After the modifications, we perform multiple regression again and from the results (Table 2) we observe that R_{adj}^2=63,74%, higher than that of the first model, so the data modifications were effective. Also, the p-value of the education variable equals to 0.32> 0.05, so it isn't statistically significant and will be subtracted.

Table 2: ANOVA results for regression model 2

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.7962 -1.1960 -0.0369  0.7922  4.5922 

Coefficients:
                                                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                            -2.503e+00  9.309e+00  -0.269 0.790093
Data_reg_trans$`Median_annual_income _^-2.5`                            8.969e+11  2.285e+11   3.926 0.000538
Data_reg_trans$Percentage_with_no_qualifications_aged_25_64             2.238e-01  2.214e-01   1.011 0.321110
Data_reg_trans$log_Population_per_square_km                             5.375e+00  1.963e+00   2.738 0.010815
Data_reg_trans$`Percentage_of_homes_with_deficiency_in_access_to_nature_^-0.5`  7.519e-02  3.007e-02   2.501 0.018761
Data_reg_trans$`Fast_food_outlets_per_100_residents_^-1`               -4.507e-01  2.249e-01  -2.004 0.055202

(Intercept)                                                                       
Data_reg_trans$`Median_annual_income _^-2.5`                                   ***
Data_reg_trans$Percentage_with_no_qualifications_aged_25_64                       
Data_reg_trans$log_Population_per_square_km                                    *  
Data_reg_trans$`Percentage_of_homes_with_deficiency_in_access_to_nature_^-0.5` *  
Data_reg_trans$`Fast_food_outlets_per_100_residents_^-1`                       .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.271 on 27 degrees of freedom
Multiple R-squared:  0.694,	Adjusted R-squared:  0.6374 
F-statistic: 12.25 on 5 and 27 DF,  p-value: 2.935e-06
```

After the variable is subtracted, we observe that F_0=15.04> $F_{0,95,4,28}$ =2.70 (table 3), whereupon the null hypothesis H0 is rejected at a significance level of 5% and Y depends linearly on at least one independent variable. It's worth commenting that the p-value values for all variables (except intercept) are less than 5%, which also defines the rejection of H0. The fact that R_{adj}^2=0.6371 shows that 63.71% of the independent variables explain the variation in the value of obesity and the remaining 36.29% is interpreted by chance.

Table 3: ANOVA results for regression model 3

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.050502 -0.012735  0.000638  0.011500  0.050048

Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                             -3.911e-02  9.208e-02  -0.425  0.67426
Data_reg_trans$`Median_annual_income _^-2.5`             1.044e+10  1.762e+09   5.925 2.24e-06 ***
Data_reg_trans$log_Population_per_square_km              5.943e-02  1.882e-02   3.158  0.00379 **
Data_reg_trans$Percentage_of_homes_with_deficiency_in_access_to_nature  7.544e-02  3.008e-02   2.508  0.01820 *
Data_reg_trans$`Fast_food_outlets_per_100_residents_^-1` -4.873e-03  2.220e-03  -2.195  0.03662 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02272 on 28 degrees of freedom
Multiple R-squared:  0.6825,    Adjusted R-squared:  0.6371
F-statistic: 15.04 on 4 and 28 DF,  p-value: 1.118e-06
```

If the null hypothesis was true, then there would be no pattern in the cloud of points in Figure 16. We observe that as the number of houses with insufficient access to nature, the population density and the number of fast-food outlets increase, childhood obesity also increases. Conversely, as the average annual income increases, obesity decreases[1].

[1] The variables of income and fast food are raised to a negative power, so in the figure their trend is reversed.

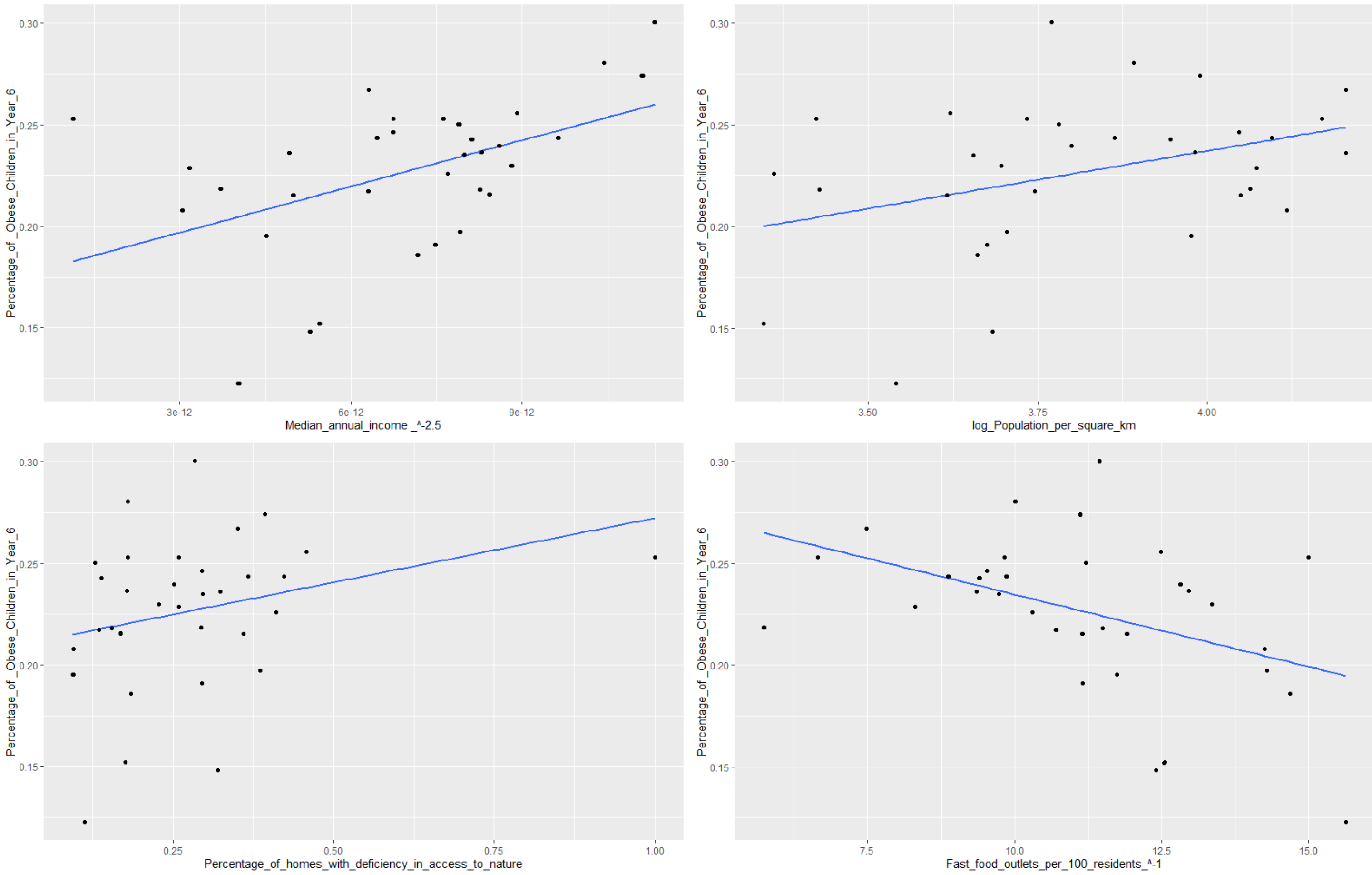


Figure 16: Line fit scatterplots of dependent variable versus independent variables

Afterwards, we examine whether the residuals of model 3 follow the normal distribution, which is confirmed by figure 17.

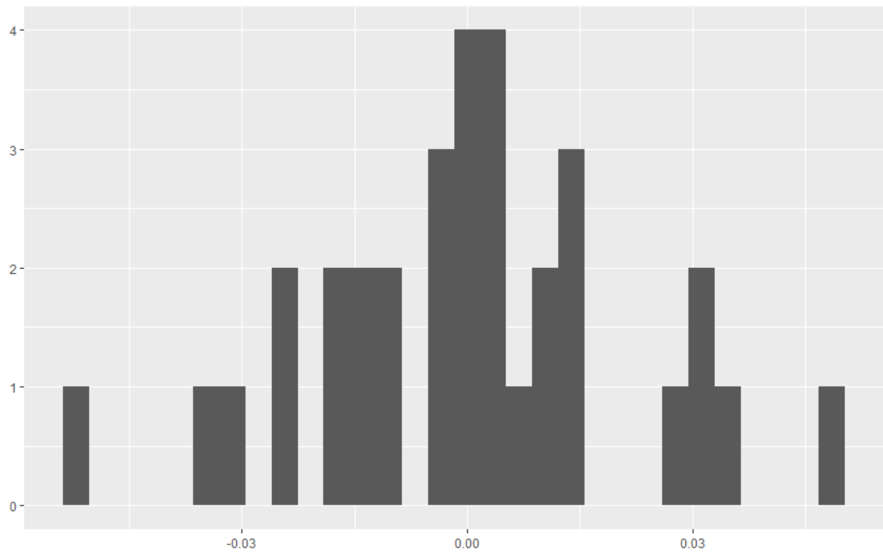


Figure 17: Histogram of residuals of model 3

Then, we will check if the four explanatory variables satisfy the no–multicollinearity assumptions. Figure 18 reveals there is low correlation between the variables (highest value is – 0.47). Additionally, the VIF values are lower than 10, so the multiple regression model meets the assumptions around multicollinearity.

```
> vif(model3)
                                  Data_reg_trans$`Median_annual_income _^-2.5`
                                                                      1.076685
                                      Data_reg_trans$log_Population_per_square_km
                                                                      1.344984
Data_reg_trans$`Percentage_of_homes_with_deficiency_in_access_to_nature_^-0.5`
                                                                      1.537056
                        Data_reg_trans$`Fast_food_outlets_per_100_residents_^-1`
                                                                      1.686878
```

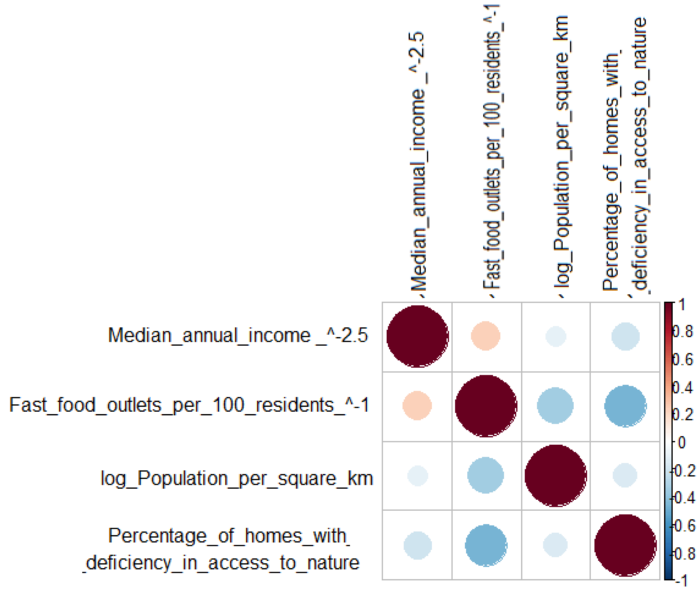


Figure 18: Correlation plot and VIF values of explanatory variables

Next, we will check for homoscedasticity by plotting the residuals in the model against the predicted values. In the first and third plot of Figure 19, there is a random cloud of points with no apparent patterning, so there is heteroscedasticity.

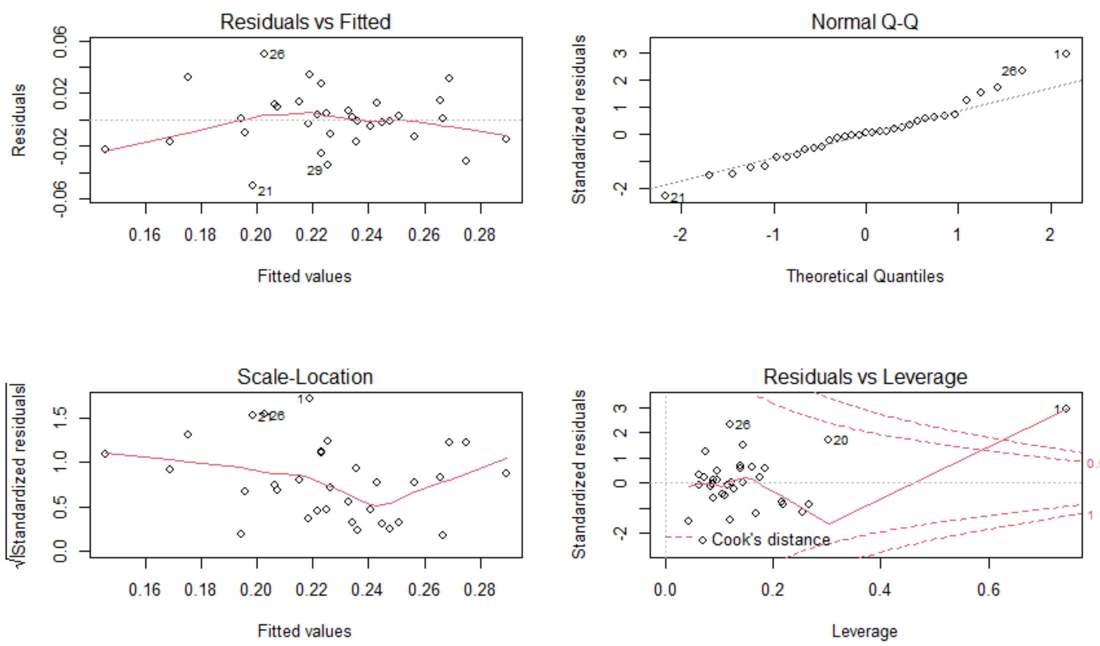


Figure 19: Model 3 diagnostics for homoscedasticity

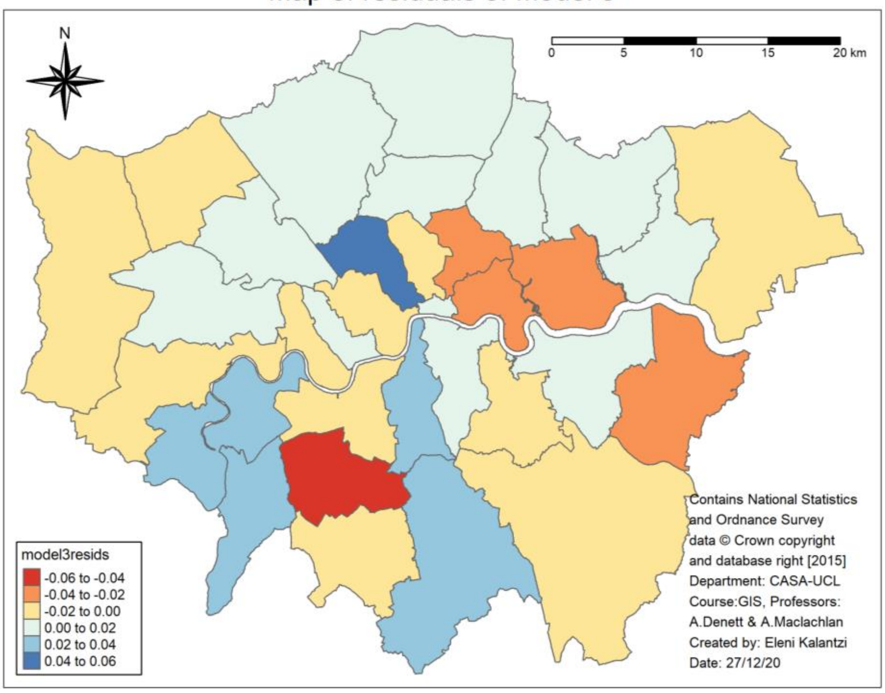
We will also check for autocorrelation of the residuals using the Durbin – Watson test. From the table 4, we observe that the DW value for our model is 2.19, which means that there is a slight negative autocorrelation that doesn't cause problems, but the fact that p-value is 0.65 >>0.05 leads us to accept there isn't any autocorrelation of the residuals.

Table 4: Durbin – Watson Test results for model 3

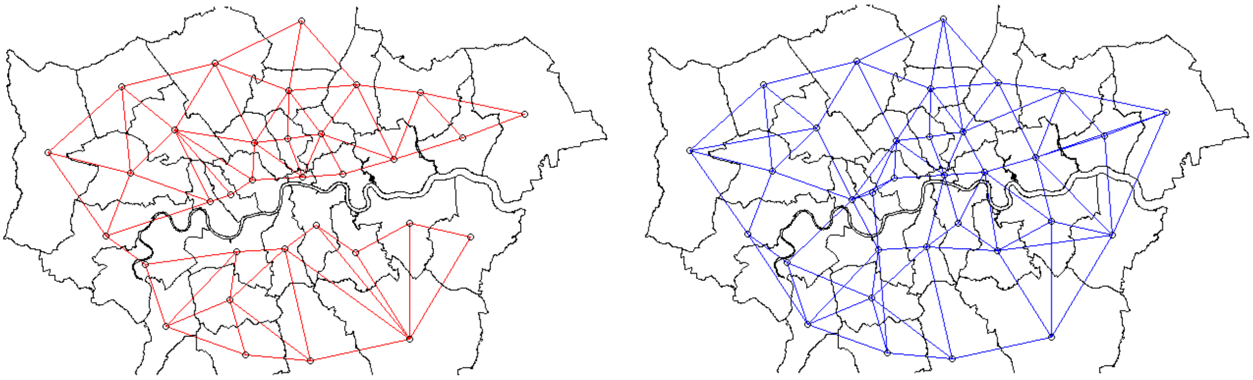
```
> tidy(DW)
# A tibble: 1 x 5
  statistic p.value autocorrelation method             alternative
      <dbl>   <dbl>           <dbl> <chr>              <chr>
1      2.19   0.652          -0.140 Durbin-Watson Test two.sided
```

As the data is spatial, a spatial autocorrelation test will be performed. On the residuals' map, some blue areas are next to others blue and some yellow and orange next to others. However, we will calculate the Moran's I index to double-check spatial autocorrelation. Two spatial weight tables will be used to calculate it: the Queen's case and the k - nearest neighbours (Map Collage 2). The appropriate number of neighbourhoods was set at 4 based on the average neighbours (4.12) found in the Queen's case.

Map of residuals of Model 3



Observing the Moran's I statistic for both Queen's case neighbours and 4-nearest neighbours, we can see that the Moran's I statistic is somewhere between -0.0613 and -0.0637 (Table 5), which is very close to 0 and p-value is 0.6>0.05. Thus, we can conclude that there isn't spatial autocorrelation in the residuals.



Map Collage 2: Queens case (left) and 4 – nearest neighbours (right)

Table 5: Moran's I test for Queen's case and 4 – nearest neighbours

```
> Queen
# A tibble: 1 x 7
  estimate1 estimate2 estimate3 statistic p.value method                            alternative
      <dbl>     <dbl>     <dbl>     <dbl>   <dbl> <chr>                             <chr>
1   -0.0613   -0.0312    0.0123    -0.271   0.607 Moran I test under randomisation greater
> Nearest_neighbour
# A tibble: 1 x 7
  estimate1 estimate2 estimate3 statistic p.value method                            alternative
      <dbl>     <dbl>     <dbl>     <dbl>   <dbl> <chr>                             <chr>
1   -0.0637   -0.0312    0.0109    -0.311   0.622 Moran I test under randomisation greater
```

ii. Spatial Regression Models

As we observed from the above Moran's I indicators there isn't spatial autocorrelation, so we expect that the results of the Spatial Lagged Y Model and the Spatial Error Model won't be statistically significant. Actually, the results in Table 6 confirm that the p-values for both spatial regression models for either the queen's case or the 4 nearest neighbours are greater than 0.05. Moreover, both λ/ρ parameters are smaller than standard errors (except SLYM of 4-nearest neighbours). This means that we must accept the null hypothesis and don't take spatial dependence into account.

Table 6: Summaries of Spatial Lagged Y Model and the Spatial Error Model

```
tidy(slag_dv_model3_queen)
A tibble: 6 x 5
term                                                 estimate std.error statistic  p.value
<chr>                                                   <dbl>     <dbl>     <dbl>    <dbl>
rho                                                  -9.60e- 3   5.03e-2    -0.191 8.49e- 1
(Intercept)                                          -4.03e- 2   8.49e-2    -0.475 6.35e- 1
`Median_annual_income _^-2.5`                         1.05e+10   1.63e+9     6.44  1.21e-10
log_Population_per_square_km                          6.02e- 2   1.77e-2     3.40  6.83e- 4
Percentage_of_homes_with_deficiency_in_access_to_nature 7.64e- 2 2.80e-2     2.73  6.35e- 3
`Fast_food_outlets_per_100_residents_^-1`            -4.88e- 3   2.04e-3    -2.39  1.69e- 2
glance(slag_dv_model3_queen)
A tibble: 1 x 6
r.squared   AIC   BIC deviance logLik  nobs
    <dbl> <dbl> <dbl>    <dbl>  <dbl> <int>
    0.683 -148. -137.   0.0144   80.8    33

tidy(SEM_model3_queen)
A tibble: 6 x 5
term                                                 estimate std.error statistic  p.value
<chr>                                                   <dbl>     <dbl>     <dbl>    <dbl>
(Intercept)                                          -3.93e- 2   8.48e-2  -0.463   6.43e- 1
`Median_annual_income _^-2.5`                         1.04e+10   1.62e+9   6.44    1.22e-10
log_Population_per_square_km                          5.95e- 2   1.73e-2   3.43    6.01e- 4
Percentage_of_homes_with_deficiency_in_access_to_nature 7.55e- 2 2.77e-2   2.73    6.41e- 3
`Fast_food_outlets_per_100_residents_^-1`            -4.88e- 3   2.04e-3  -2.38    1.71e- 2
lambda                                               -2.17e- 3   2.50e-1  -0.00866 9.93e- 1
glance(SEM_model3_queen)
A tibble: 1 x 6
r.squared   AIC   BIC deviance logLik  nobs
    <dbl> <dbl> <dbl>    <dbl>  <dbl> <int>
    0.682 -148. -137.   0.0145   80.8    33
```

```
> tidy(slag_dv_model3_knn4)
# A tibble: 6 x 5
  term                                                 estimate std.error statistic      p.value
  <chr>                                                   <dbl>     <dbl>     <dbl>        <dbl>
1 rho                                                   2.70e-1   1.43e-1      1.89 0.0586
2 (Intercept)                                          -8.73e-2   8.55e-2     -1.02 0.307
3 `Median_annual_income _^-2.5`                         9.61e+9   1.60e+9      6.02 0.00000000175
4 log_Population_per_square_km                          5.66e-2   1.64e-2      3.45 0.000554
5 Percentage_of_homes_with_deficiency_in_access_to_nature 6.66e-2 2.65e-2      2.51 0.0119
6 `Fast_food_outlets_per_100_residents_^-1`            -4.46e-3   1.94e-3     -2.30 0.0213
> glance(slag_dv_model3_knn4)
# A tibble: 1 x 6
  r.squared   AIC   BIC deviance logLik  nobs
      <dbl> <dbl> <dbl>    <dbl>  <dbl> <int>
1     0.716 -151. -140.   0.0129   82.4    33

tidy(SEM_model3_knn4)
A tibble: 6 x 5
term                                                 estimate std.error statistic  p.value
<chr>                                                   <dbl>     <dbl>     <dbl>    <dbl>
(Intercept)                                          -6.94e- 2   8.21e-2    -0.845 3.98e- 1
`Median_annual_income _^-2.5`                         1.11e+10   1.44e+9     7.69  1.53e-14
log_Population_per_square_km                          6.62e- 2   1.65e-2     4.00  6.25e- 5
Percentage_of_homes_with_deficiency_in_access_to_nature 8.50e- 2 2.72e-2     3.13  1.76e- 3
`Fast_food_outlets_per_100_residents_^-1`            -5.12e- 3   2.03e-3    -2.52  1.16e- 2
lambda                                               -2.63e- 1   2.82e-1    -0.935 3.50e- 1
glance(SEM_model3_knn4)
A tibble: 1 x 6
r.squared   AIC   BIC deviance logLik  nobs
    <dbl> <dbl> <dbl>    <dbl>  <dbl> <int>
    0.692 -148. -138.   0.0141   81.0    33
```

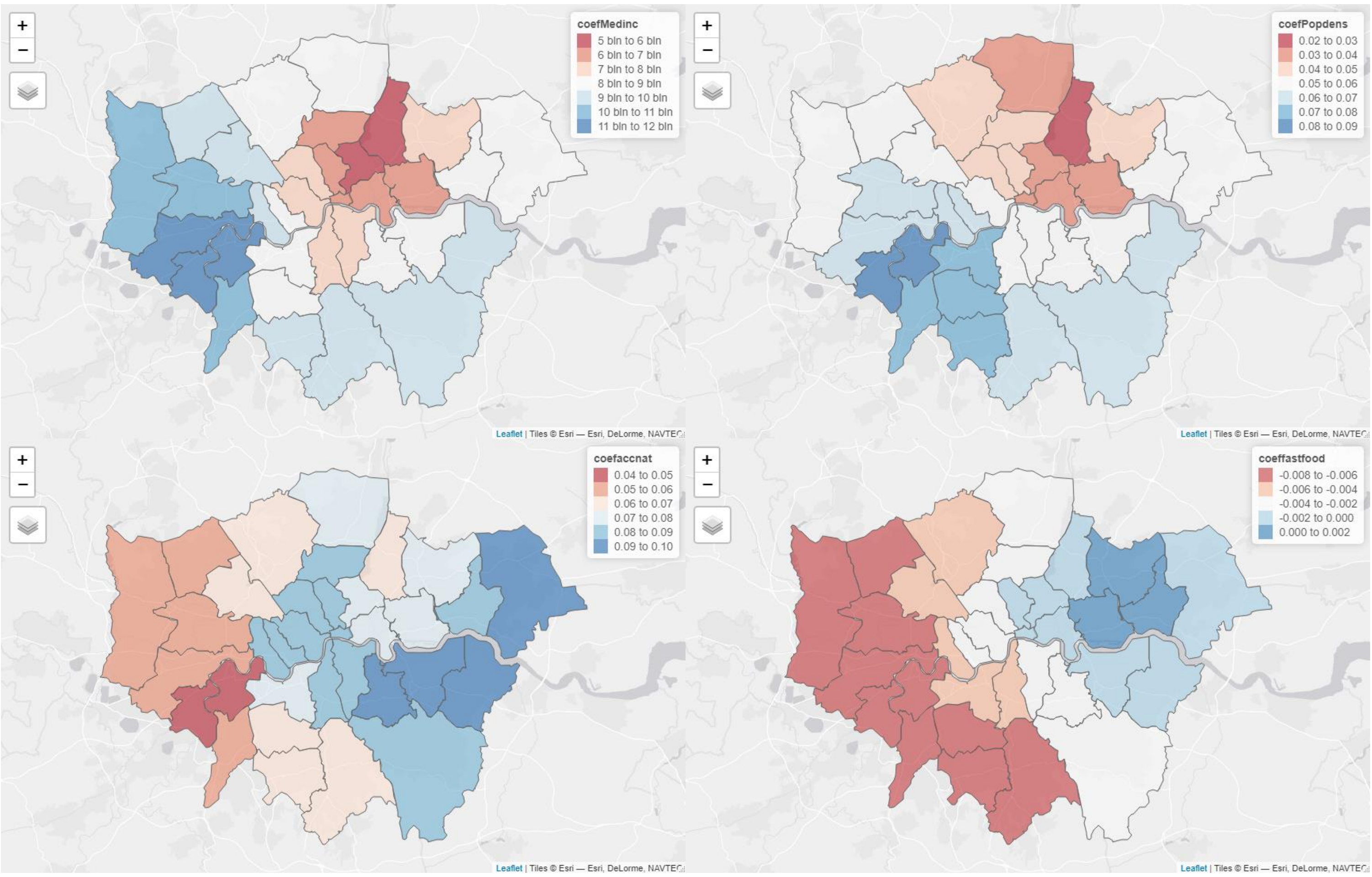
iii. Geographically Weighted Regression Model (GWR)

The output from the GWR model uncovers how the coefficients vary across the 33 Boroughs in London. We observe that global coefficients are exactly the same as them OLS model 3. For example, if we take the environmental variable, we'll see that 1-unit change in homes with deficient access to nature results in a rise in childhood obesity of 0.048 to 0.098. The R^2 increased from 0.63 to 0.83, but that doesn't necessarily mean it's definitely better than the global model. Furthermore, the Akaike Information Criterion (AIC) was lower for the GWR (-171.59) than the other models. The residuals from the GWR model showed no spatial auto-correlation (Moran's I=-0.12, p-value=0.39). Therefore, this pattern seems to occur due to random chance. This finding suggests that the GWR method is beneficial when dealing with spatial non-stationarity.

Table 7: GWR model summary

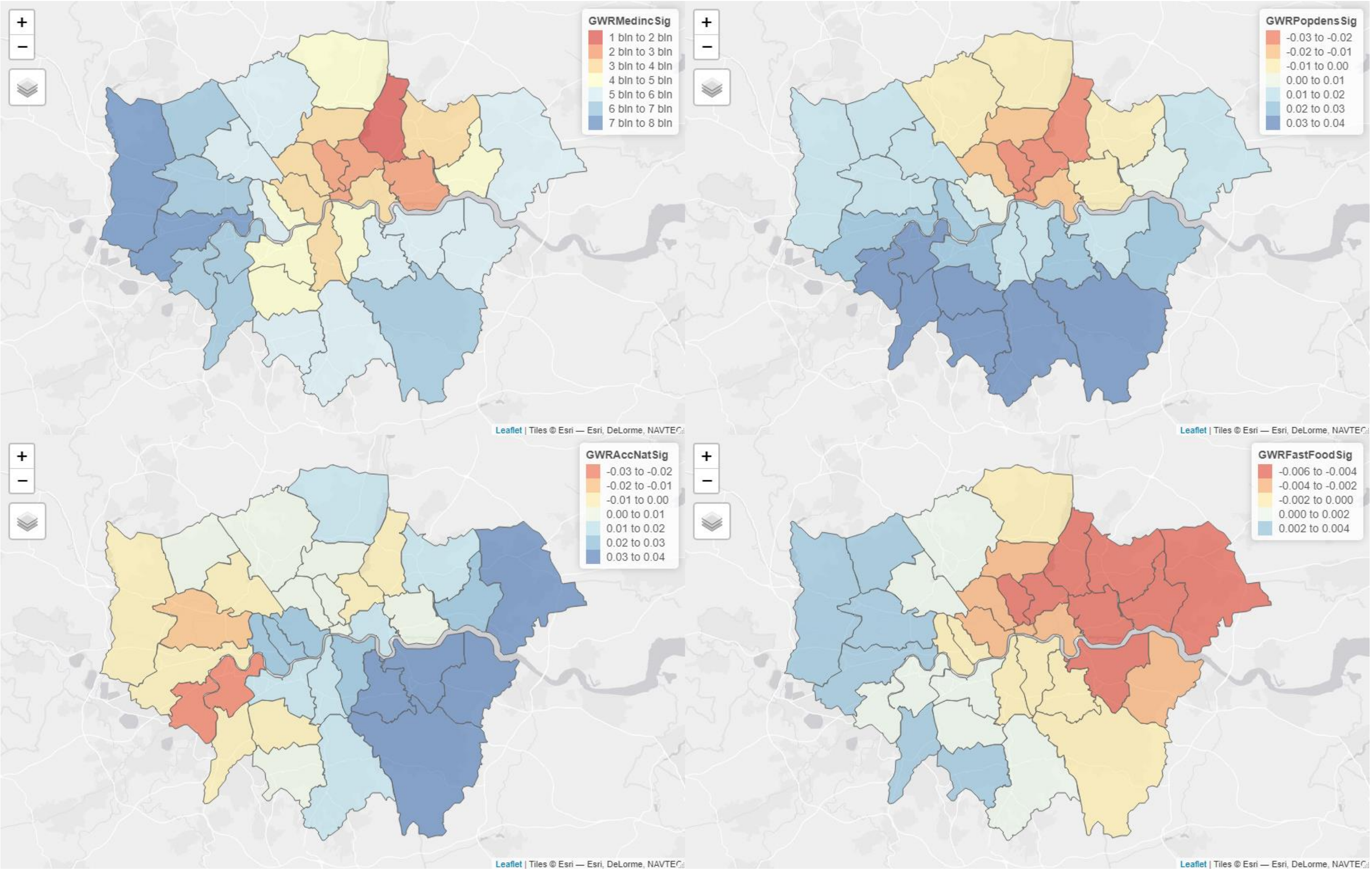
```
> gwr.modeltrans
Call:
gwr(formula = Percentage_of._Obese_Children_in_Year_6 ~ Median_annual_income._..2.5 +
    log_Population_per_square_km + Percentage_of_homes_with_deficiency_in_access_to_nature +
    Fast_food_outlets_per_100_residents_..1, data = LonBoroughsSPtrans,
    coords = coordsBSP, adapt = GWRbandwidthtrans, hatmatrix = TRUE,
    se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.1818512 (about 6 of 33 data points)
Summary of GWR coefficient estimates at data points:
                                                              Min.     1st Qu.      Median     3rd Qu.        Max.
X.Intercept.                                           -1.1758e-01 -5.8218e-02 -1.8931e-02  1.4746e-02  8.4562e-02
Median_annual_income._..2.5                             5.8826e+09  7.1845e+09  8.5226e+09  9.4882e+09  1.1447e+10
log_Population_per_square_km                            2.6732e-02  4.3915e-02  5.5738e-02  6.6814e-02  8.3678e-02
Percentage_of_homes_with_deficiency_in_access_to_nature 4.8259e-02  6.3068e-02  7.6677e-02  8.8804e-02  9.8878e-02
Fast_food_outlets_per_100_residents_..1                -6.8343e-03 -6.2540e-03 -3.2608e-03 -1.4918e-03  6.6187e-04
                                                            Global
X.Intercept.                                           -3.910e-02
Median_annual_income._..2.5                             1.044e+10
log_Population_per_square_km                            5.940e-02
Percentage_of_homes_with_deficiency_in_access_to_nature 7.540e-02
Fast_food_outlets_per_100_residents_..1                -4.900e-03
Number of data points: 33
Effective number of parameters (residual: 2traceS - traceS'S): 14.50372
Effective degrees of freedom (residual: 2traceS - traceS'S): 18.49628
Sigma (residual: 2traceS - traceS'S): 0.02021397
Effective number of parameters (model: traceS): 11.35138
Effective degrees of freedom (model: traceS): 21.64862
Sigma (model: traceS): 0.01868438
Sigma (ML): 0.01513341
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): -141.4577
AIC (GWR p. 96, eq. 4.22): -171.5948
Residual sum of squares: 0.007557667
Quasi-global R2: 0.8340129
```

The coefficients for the median income and population density parameters were larger in the western and southern part and smaller in the northern and central part of the study area (Map Collage 3). On the other hand, the coefficient for the deficient access to nature variable was lower in the western and higher in the eastern and central part. The coefficient for Fast Food outlets density was slightly positive in Newham, Redbridge and Barking-Dagenham and negative in the other areas, with highest values in western parts.



Map Collage 3: Visualisation of the coefficients of explanatory variables

To examine if the results are statistically significant for all London boroughs (Map collage 4), we will compare coefficients with standard errors. For example, for the income variable only the Waltham Forest area is not 2 standard errors from the coefficient. In contrast, for the population density variable the statistically significant results are located southern and central, for access to nature eastern and in the borough of Richmond upon Thames, while the variable fast-food mainly in the eastern and western boroughs. The coefficients from GWR indicate that median annual income variable has the strongest relationship with childhood obesity, confirming the findings from the global OLS regression model.



Map Collage 4: Visualisation of the significance test of explanatory variables

Discussion

In Table 8, we can compare all the models used to examine the relationship of childhood obesity with the explanatory variables and ascertain that the GWR will be selected as the most appropriate for our research, as it has the lowest AIC and the greatest R_{adj}^2.

Table 8: Comparison of Regression Models

Models/Criteria	R_{adj}^2	AIC
OLS Model 3	0.637	-150
Spatial Lagged Y Model Queen's Case	0.683	-148
Spatial Lagged Y Model 4NN	0.716	-151
Spatial Error Model Queen's case	0.682	-148
Spatial Error Model 4NN	0.692	-148
GWR Model	0.834	-172

In this study there are some limitations that should be taken into account. Firstly, although it reflects the local conditions that affect childhood obesity, it doesn't include features such as geographical disparities or the number of children in each borough, creating shortcomings. Secondly, the GWR model may reveal spatial variations of variables, but doesn't explain what causes variability between boroughs. Thirdly, regarding the explanatory variable of income, we cannot be sure of its accuracy, as we cannot detect cases of tax evasion that could possibly change the results. Fourthly, with regard to the variable access to green spaces, unfortunately the most recently available data were for the year 2012, so it’s likely that the percentages have changed since then. Finally, in this paper only some of the variables that could affect childhood obesity were selected. However, there are other variables (e.g. exercise rate, degree of problem’s awareness, etc.), which if added to later research could improve the performance of the model and add a better explanation.

Conclusions

The analysis of the problem of childhood obesity in London using different models of global and local regression, led us to conclude that childhood obesity is to some extent related to the following socioeconomic factors: median annual income, population density, access to green spaces, density of fast-food outlets and that this relationship doesn’t remain stable, but varies spatially per borough. The results show that GWR model has stronger explanatory power compared to other models, so it was the most suitable for this kind of research. Thus, urban planning and policies should be adjusted in different areas based on the spatially varying coefficients from GWR model.

Our findings show that the median annual income is the one that most affects childhood obesity. Therefore, local authorities should implement measures and strategies that are more supportive of low-income boroughs. For example, they could include more free physical activity in schools, provide health food discount coupons, inform parents about how to properly distribute their income, increase local jobs, and so on. For all the other variables, the relationship with childhood obesity was weaker overall.

References

Department of Health and Social Care (2020) ‘Tackling obesity: empowering adults and children to live healthier lives’. Available at: https://www.gov.uk/government/publications/tackling-obesity-government-strategy/tackling-obesity-empowering-adults-and-children-to-live-healthier-lives#fn:1.

Devaux, M. *et al.* (2011) ‘Exploring the Relationship Between Education and Obesity’, *OECD Journal: Economic Studies*, 2011. doi: 10.1787/19952856.

Edwards, K. L. *et al.* (2010) ‘The neighbourhood matters: Studying exposures relevant to childhood obesity and the policy implications in Leeds, UK’, *Journal of Epidemiology and Community Health*. doi: 10.1136/jech.2009.088906.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002) *Geographically Weighted Regression the analysis of spatially varying relationships*. Wiley.

Gordon-Larsen, P. *et al.* (2006) ‘Inequality in the Built Environment Underlies Key Health Disparities in Physical Activity and Obesity.’, *Pediatrics*, 117, pp. 417–424. doi: 10.1542/peds.2005-0058.

Han, J. (2011) ‘Variance least squares estimators for multivariate linear mixed model’, *Journal of Statistical Planning and Inference*, 141(10), pp. 3345–3355. doi: 10.1016/j.jspi.2011.04.018.

James, W. P. T. (2008) ‘The epidemiology of obesity: The size of the problem’, *Journal of Internal Medicine*. doi: 10.1111/j.1365-2796.2008.01922.x.

Krzyzanowska, M. and Mascie-Taylor, C. G. N. (2012) ‘Impact of social mobility and geographical migration on variation in male height, weight and body mass index in a british cohort’, *Journal of Biosocial Science*. doi: 10.1017/S0021932011000551.

Liu, D. *et al.* (2019) ‘Effects of local authority expenditure on childhood obesity’, *European Journal of Public Health*, 29(4), pp. 785–790. doi: 10.1093/eurpub/cky252.

London Health Commission (2014) *Global City Comparisons Overview*.

NHS (2020) *Part 1: Obesity related hospital admissions - NHS Digital*. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-obesity-physical-activity-and-diet/england-2020/part-1-obesity-related-hospital-admissions-copy (Accessed: 6 January 2021).

Nielsen, T. S. and Hansen, K. B. (2007) ‘Do green areas affect health? Results from a Danish survey on the use of green areas and health indicators’, *Health and Place*, 13(4), pp. 839–850. doi: 10.1016/j.healthplace.2007.02.001.

O’Dea, J. A. and Dibley, M. J. (2010) ‘Obesity increase among low SES Australian schoolchildren between 2000 and 2006: Time for preventive interventions to target children from low income schools?’, in *International Journal of Public Health*. doi: 10.1007/s00038-009-0079-x.

Phe (2018) ‘Obesity and the environment - Density of fast food outlets at 31/12/2017’. doi: 10.1136/archdischild-2017-312981.

Phe (2020) *Excess Weight and COVID-19*. Available at: https://www.gov.uk/government/publications/excess-weight-and-covid-19-insights-from-new-evidence (Accessed: 23 December 2020).

Procter, K. L. *et al.* (2008) 'Micro-level analysis of childhood obesity, diet, physical activity, residential socioeconomic and social capital variables: where are the obesogenic environments in Leeds?', *Area*, 40(3), pp. 323–340. doi: 10.1111/j.1475-4762.2008.00822.x.

Scarborough, P. *et al.* (2011) '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*, 33(4), pp. 527–535. doi: 10.1093/pubmed/fdr033.

Schuurman, N., Peters, P. A. and Oliver, L. N. (2009) 'Are Obesity and Physical Activity Clustered? A Spatial Analysis Linked to Residential Density', *Obesity*, 17(12), pp. 2202–2209. doi: 10.1038/oby.2009.119.

Stamatakis, E., Wardle, J. and Cole, T. J. (2010) 'Childhood obesity and overweight prevalence trends in England: Evidence for growing socioeconomic disparities', *International Journal of Obesity*. doi: 10.1038/ijo.2009.217.

Trust for London (2020) *Childhood obesity by London borough*. Available at: https://www.trustforlondon.org.uk/data/child-obesity/ (Accessed: 21 December 2020).

Ward, M. and Gleditsch, K. (2011) *Spatial Regression Models*, *SAGE Publications, Inc.* doi: 10.4135/9781412985888.

Declaration of Authorship

I, Eleni Kalantzi, confirm that the work presented in this assessment is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Eleni Kalantzi

Date of signature: 11/01/21

Assessment due date: 11/01/21