Transport, Infrastructure, and Territory Research
Group (t-GIS), Geography Department
Universidad Complutense de Madrid (UCM), C/ Profesor Aranguren, s/n, 28040 Madrid, Spain.

Integrated modelling of transport scenarios from stakeholders for air quality and emissions (INTRANCES).
Project Number 886050

WORKING PAPER NO. 1

Understanding the spatial variation of air pollution and its impacts through Geographically Weighted Regression (GWR).

Richard J. Hewitt[1,2†], Eduardo Caramés[1] and Rafael Borge[3]

1. Observatorio para una Cultura del Territorio, 28012, Madrid, Spain.
2. Transport and Territory Research Group (t-GIS), Department of Geography, Madrid Complutense University, 28040, Madrid, Spain
3. Escuela Técnica Superior de Ingenieros Industriales, Departamento de Ingeniería Química Industrial y del Medio Ambiente, C/ José Gutiérrez Abascal, 2, 28006, Madrid, Spain

June 2022

[†] Corresponding author.
Email address: rhewitt@ucm.es

Abstract

We explore the potential correlation between income and exposure to air pollution for the city of Madrid, Spain, and its neighboring municipalities. Statistical analyses were carried out using electoral district level data on gross household income, and NO_2 and $PM_{2.5}$ concentrations in air obtained from a mesoscale air quality model for the study area. Household income data were summarized at the grid level through a zonal statistics operation, carried out at four different cell resolutions for both 1 x 1 km and 2 x 2 km grids in order to account for modifiable aerial units and data conversion uncertainty. Our findings point to a clear negative correlation between level of household income and exposure to both pollutants, which was clearly present at all resolutions and both grid sizes, though with a high degree of variability depending on the resolution and grid size chosen. The strongest association between income and air pollution was found for minimum gross household income (MGHI) and $NO_{2,}$ and MGHI and $PM_{2.5}$. The global regression models explained between 10% and 20% of the variance for MGHI and NO_2, and between 12% and 19% of the variance for MGHI and $PM_{2.5}$, depending on resolution and grid size chosen. Standard residual error varied between 0.55-0.58 for MGHI and NO_2 and between 0.28-0.30 for MGHI and $PM_{2.5}$. To address the high degree of variability in the models we explored the spatial heterogeneity of the correlation effect, finding stronger decrease in contamination exposure as minimum rent increases in the north of the city of Madrid and the municipalities of Tres Cantos and Colmenar Viejo. In the centre and south of the metropolitan area, the slope of the regression line is shallower. This may be partly due to the fact that contaminant concentrations in the centre of the study area are uniformly higher than elsewhere, offering less opportunity to reduce exposure within this area. In the east of the metropolitan area no relationship between the variables was detected at the scale of the analysis. Our results suggest income-based inequality in exposure to air pollution in Madrid. They highlight the usefulness of electoral district level income data and simulated concentrations from Eulerian photochemical air quality models (a Community Multiscale Air Quality model – CMAQ, in this case) for understanding environmental inequality. Further work would be needed to explore the patterns observed at the district and neighborhood level.

Keywords: Environmental Justice; Air Pollution**;** CMAQ model**;** Inequality; GWR

1 Introduction

Outdoor air pollution is known to cause serious health impacts and to increase mortality, and is considered by the World Health Organization (WHO) as the world's greatest threat to environmental health (Izquierdo et al 2020, WHO 2016). The proportion of the global population living in urban areas is on the rise and expected to increase further in future (ref). Population exposure, and consequently increased mortality in urban areas, is therefore likely to rise unless action is taken. Improving air quality in cities is therefore an important priority for international bodies like the WHO and the European Union (EU). Increasingly the issue is being taken up by city authorities, e.g. (ref, ref, ref). Specifically, air pollution refers to high ambient air concentration levels of specific contaminants, usually from motorized vehicles and the burning of fossil fuels for energy and heat (Gurjar et al., 2014.; World Health Organization (WHO), 2013). Particulate matter (PM), nitrogen dioxide (NO2) and ground-level ozone (O3), are regarded as especially dangerous for human health (EEA 2018). Given the great diversity of heating and transport systems in cities, as well as climatic factors, ambient air concentration levels vary widely in time and space. Air quality models, such as Community Multiscale Air Quality Models (CMAQ), offer an opportunity to examine this variability at medium-high spatial resolution (ref).

At the same time, most cities today tend to be spatially segregated by socioeconomic group. Property prices tend to be higher in well-connected areas with high quality amenities (ref), and lower income residents tend to be forced out of these areas to the urban periphery (ref). Poor air quality and its related drivers (proximity to roads and industry) is likely to be a factor in residents' neighborhood choice, with the least well-off being least able to choose (ref). In this sense, many studies have sought to investigate whether there is a relationship between socioeconomic factors and exposure to air pollution, reflected by various indicators, e.g. immigration status, race, age profile, or educational level (Moreno Jiménez & Cañada Torrecilla, 2007, Giang & Castellani, 2019, Prieto-Flores et al., 2021). In this study we investigate whether any correlation can be found between lower incomes and higher levels of air pollution in the city of Madrid and its neighboring municipalities. To do this, we use simulated concentrations output from a CMAQ model for two airborne pollutants, NO_2 and $PM_{2.5}$ together with freely-available data on gross household income at the level of the census district.

2 Study Area

The city of Madrid is an urban municipality in centre of Spain with an area of 604.3km^2 and over 3 million inhabitants (Figure 1).

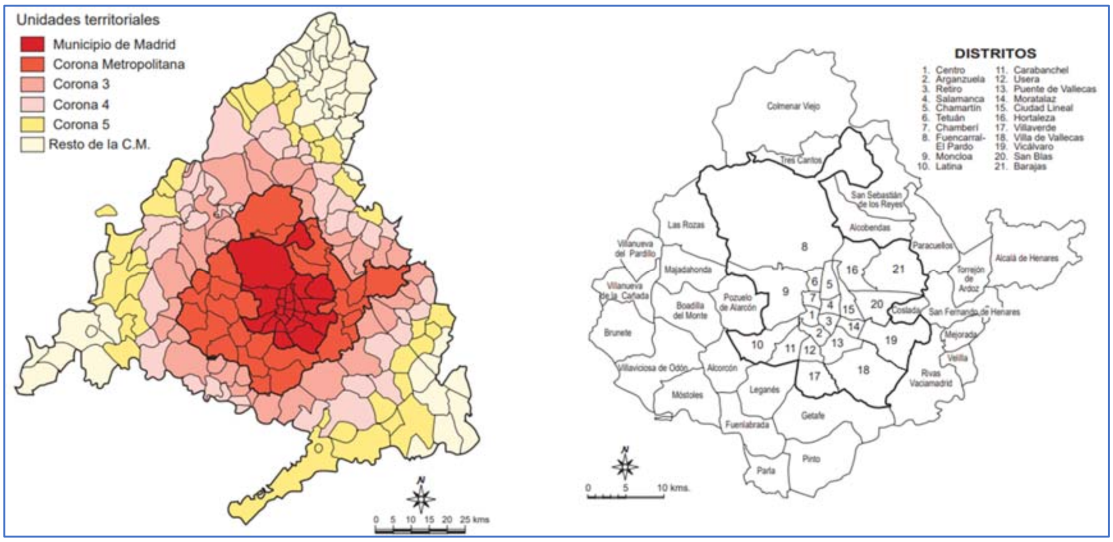


Figure 1. The municipality of Madrid (red) and suburban municipalities comprising the metropolitan area. Source: Own work based on data from Spanish National Geographical Institute; classifications according to Lozano (2002)

The city council has estimated that approximately 2.5 million vehicles start or finish in the city of Madrid on a typical weekday, with around 40 million km being driven on a typical day in the city (Izquierdo et al 2020); this inevitably generates major concentrations of airborne contaminants. Air pollution in Madrid causes an estimated 88 deaths per year from particulate matter (PM) and 519 from nitrogen dioxide (NO^2), which is equivalent to 4 deaths per 100,000 inhabitants in the first case, and 23 deaths per 100,000 cases in the second (Izquierdo et al 2020). Within the municipality of Madrid it has been estimated that 74.4% of all local NO^2 emissions (i.e. those arising from within the city itself), are attributable to road traffic (Madrid City Council 2019). Like many modern cities, the core urban area comprising the municipality of Madrid is surrounded by adjacent commuter towns which are home to large populations attracted by easy access to the city centre and comparatively lower property prices. These adjacent municipalities comprise the metropolitan periphery (*Corona Metropolitana* on Figure 1) (Ballesteros & Berzal, 2002), and are important centers of population and industry, but also strongly residential in character, and highly unequal socioeconomically. Two such municipalities, Pozuelo de Alarcón and Boadilla del Monte, to the west of the city, contain several census districts with gross household income levels among the highest in Spain (all 129,750€) (INE 2022, data from 2019). Four others, Alcalá de Henares to the east of the city, and Parla, Getafe and Leganes to the south each contain a census district with household incomes among the lowest in Spain (all < 25,000€) (INE 2022, data from 2019).

The high population density in the metropolitan periphery together with the high volume of motorized vehicle journeys at peak times leads to traffic congestion at key entrances, exits and intersection points into the city (traffic-related hotspots: see, e.g. Quaassdorff et al 2016) NO_2 and $PM_{2.5}$ concentrations tend to be higher at these points, as well as on the roads and nearby areas, showing that road traffic is clearly the main contributor to air pollution levels in Madrid (Borge et al 2014) However, the typical daily meteorological cycle also plays an important role in resulting ambient concentration levels, especially the late evening NO_2 peak (Artiñano et al 2003, Borge et al 2018)

For these reasons, we chose a study area including the whole metropolitan area, incorporating both the municipality of Madrid and the metropolitan periphery described above (Figure 1). The study area therefore includes income disparities that are among the largest in Spain, as well as key urban pollution hotspots.

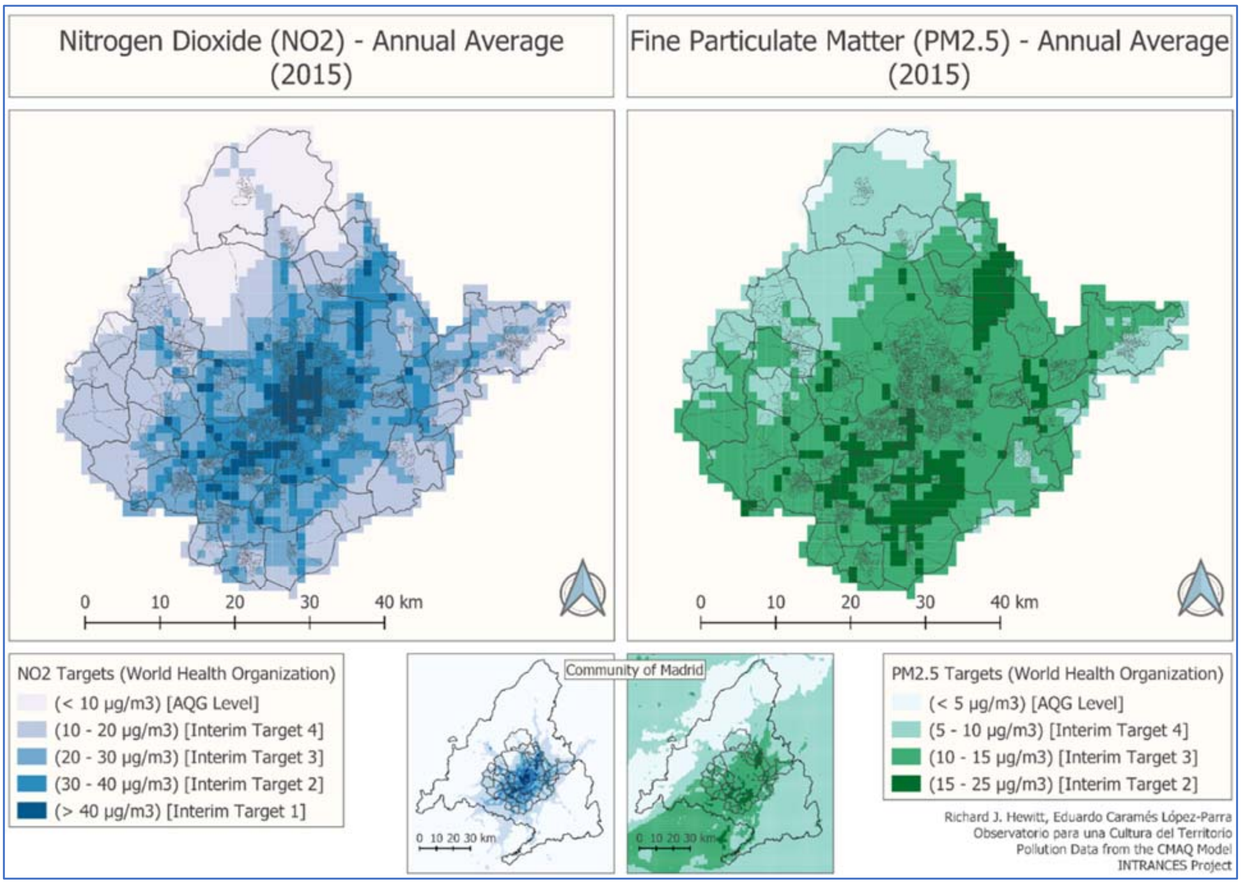
3 Data and Methods

Data comprised two different groups: 1) air pollution concentrations, and 2) gross household income data in Euros collected between 2015 and 2018.

3.1 Air pollution data

For air pollution, we used spatially-resolved simulated concentrations of nitrogen dioxide (NO_2) and particulate matter less than 2.5 micrometers in diameter ($PM_{2.5}$), output from the Community Multiscale Air Quality Model (CMAQ) for Madrid. These two pollutants have already been used in other articles to measure social inequality (Rosofsky et al., 2018) and are the most relevant regarding health impacts of air pollution specifically in Madrid (Izquierdo et al., 2020). The CMAQ model estimates the concentration in $\mu g/m^3$ of NO2 and PM2.5 for the year 2015 for the whole of the Community of Madrid. These data take the form of a square grid of $1km^2$ cells, in which each cell in the grid is a unique georeferenced polygon object in vector GIS format with attached attribute containing the estimated concentration value. For these pollutants, the World Health Organisation (WHO) establishes a series of limit values according to their danger to humans, with the maximum values to which a person can be exposed in a 1 year time period being 40 $\mu g/m^3$ for NO_2 and and 25 $\mu g/m^3$ for $PM_{2.5}$ (World Health Organisation (WHO), 2021). European Directive 2008/50/EC establishes annual maximum permitted exposure levels of 40 $\mu g/m^3$ for NO_2 and 20 $\mu g/m^3$ for $PM_{2.5}$, both of which limits are replicated in Spanish law (RD 102/2011).

Figure 2 shows the spatial distribution of both pollutants in the study area as estimated by the CMAQ model. For NO_2 (Fig 2a) concentrations tend to be found in the centre of the study area in the city of Madrid and decrease as the distance from the centre increases. The lowest NO_2 concentration levels are found in the northern part of the Madrid metropolitan area and in the east of the municipality of Alcalá de Henares. For PM2.5 (Fig 2b), the central area shows concentrations of between 10 and 15 $\mu g/m^3$, with some areas in the northeast and south showing higher concentrations, between 15 and 25 $\mu g/m^3$. The high estimated concentration values in the south correspond to the Madrid district of Villa de Vallecas, and the adjacent metropolitan municipalities of Getafe, Pinto and Leganés. Although these concentrations are not found in the census districts, which may be because these emissions come from road traffic or from industries located in these areas. As with NO_2 concentrations, the lowest values are found in the northwest and east of the municipality of Alcalá de Henares.


Nitrogen Dioxide (NO2) - Annual Average (2015)

Fine Particulate Matter (PM2.5) - Annual Average (2015)

0 10 20 30 40 km
0 10 20 30 40 km
NO2 Targets (World Health Organization)
(< 10 µg/m3) [AQG Level]
(10 - 20 µg/m3) [Interim Target 4]
(20 - 30 µg/m3) [Interim Target 3]
(30 - 40 µg/m3) [Interim Target 2]
(> 40 µg/m3) [Interim Target 1]
Community of Madrid
0 10 20 30 km
0 10 20 30 km
PM2.5 Targets (World Health Organization)
(< 5 µg/m3) [AQG Level]
(5 - 10 µg/m3) [Interim Target 4]
(10 - 15 µg/m3) [Interim Target 3]
(15 - 25 µg/m3) [Interim Target 2]

Richard J. Hewitt, Eduardo Caramés López-Parra
Observatorio para una Cultura del Territorio
Pollution Data from the CMAQ Model
INTRANCES Project

Figure 2. Map of concentrations of NO_2 (left) and $PM_{2.5}$ (right) pollutants according to World Health Organisation (WHO) limit values (Source; CMAQ).

In view of the limited coverage provided by air quality monitoring stations, these simulated concentrations represent the best available information on the spatial extent of these two pollutants. Nonetheless, as with any simulation model output, these data sources should be approached with caution, as a result of the unavoidable uncertainty and error arising from the modelling process.

3.2 Household Income data

For household incomes, we used the Atlas of Household Income of the Spanish National Statistics Institute (INE 2022). We extracted gross household income (GHI) data at the level of the census tract, the highest resolution data currently publicly available for the year 2015, to match the date for which simulated air pollution data were available (Figure 3). Where data were not available for the year 2015, the nearest available date was chosen (2016,2017 or 2018). Census tracts are the statistical unit inferior to the municipality that is the basis of the statistical operations of National population censuses. Every municipality is divided into one or more census tracts and there is no part of any municipality that does not belong to a census tract. Census tracts vary in size according to the number of inhabitants, with the most populous municipalities having many more census tracts than those with few inhabitants. Madrid municipality has over 2400 census tracts, the peripheral municipality of Boadilla del Monte (Figure 1) has 27, and Zarzalejo, a rural municipality, has just one. In our study area, the smallest census tract is just 0.6 ha, the largest 1762 ha, though large census tracts are very uncommon (median approx 4.4 ha).

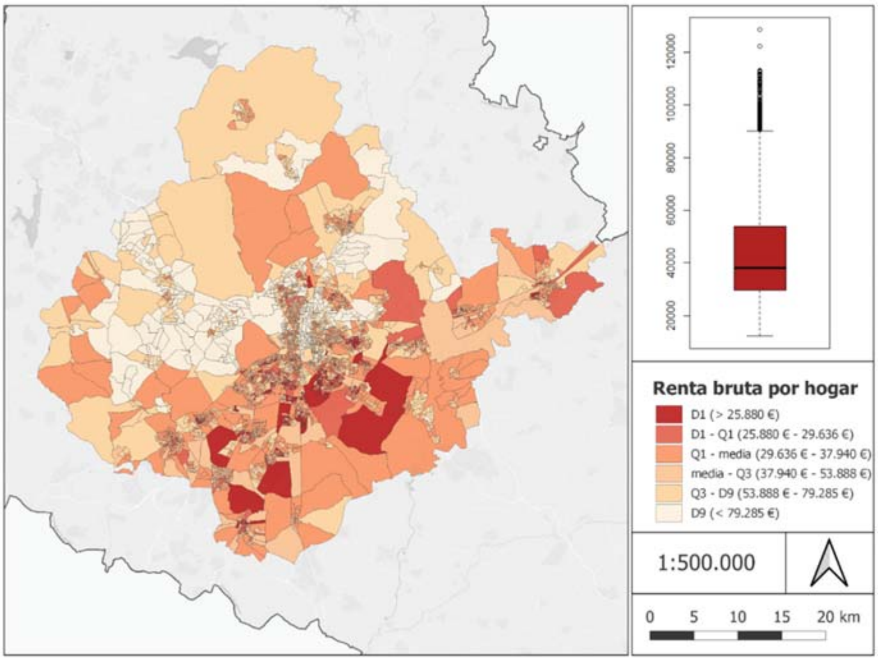

120000
100000
80000
60000
40000
20000
Renta bruta por hogar
🟥 D1 (> 25.880 €)
🟥 D1 - Q1 (25.880 € - 29.636 €)
🟧 Q1 - media (29.636 € - 37.940 €)
🟧 media - Q3 (37.940 € - 53.888 €)
🟧 Q3 - D9 (53.888 € - 79.285 €)
⬜ D9 (< 79.285 €)
1:500.000
0 5 10 15 20 km

Figure 3. A (left) Distribution of annual gross income per household (GHI) in the metropolitan area and the city of Madrid (Source; Instituto Nacional de Estadística [INE]). Census tracts without information on gross household income (e.g. Retiro Park) marked in dark grey. B (right) Boxplot of GHI in the study area.

GHI in the study area varies between 12,153€ and 128,571€, with median 37,940€. With SD=22,282€ and mean=45,574€ this gives a coefficient of variation of 48.89%. Figure 3 shows the high degree of variability between incomes in the study area, with a noticeable dividing line from lower left to upper right. The highest incomes (upper 10%, lightest colour) are uniformly located to the north of this line, while the lowest incomes (lower 10%, dark red) are all found to the south of this line. In the core city at the centre of the study area, census tracts are smaller, reflecting higher population density, and the pattern is more heterogenous, with a mixture of lower and higher income neighbourhoods.

3.3 Data preparation and statistical analysis

First, we transformed the vector polygon coverage of GHI by census tract obtained from INE into raster format (rasterization). We used a zonal statistics operation in GIS to extract and summarize the pixel level GHI data obtained from the rasterization operation for each km^2 of the CMAQ model grid. The zonal statistics operation produced four outputs for each km^2: 1) mean GHI within each km^2; 2) maximum GHI within each km^2; 3) minimum GHI within each km^2; total GHI within each km^2; 4) minimum GHI within each km^2.

We then carried out ordinary least squares regression (OLS) using simulated mean annual NO_2 concentration for the year 2015 as the dependent variable (y), and each of the four variants of GHI as the independent variable (x). The process was repeated using simulated mean annual $PM_{2.5}$ concentration for the year 2015 as the dependent variable (y). Since only minimum GHI (minGHI) produced a significant response for $p < 0.01$ with > 10% of the variance explained (for both NO_2 and and $PM_{2.5}$) and the GHI variants are clearly not independent from each other, a multiple regression model was not used. To ensure normal distribution of the residuals – a key assumption of linear regression – a log transformation was performed on both dependent and independent variables.

To account for the Modifiable Aerial Unit Problem (Openshaw 1984), where statistical information can be shown to depend on the size of the zone in which it is sampled or reported, as well as uncertainty derived from the rasterization operation (Diaz Pacheco et al 2018), we: 1) rasterized the GHI vector layer at four different resolutions – 48m, being the most appropriate cell size for the GHI vector layer according to Piwowar's rule (reported by Congleton et al), 100m, 200m and 500m; 2) we summarized the zonal statistics from GHI raster maps at 48m, 100m, 200m and 500m resolutions using a larger vector grid obtained by grouping pairs of individual km^2 together create a 4km^2 grid. In this way, both the effect of cell resolution of the rasterization of the GHI data as well as the effect of the size of the reporting units were tested.

As noted above, for both the simulated concentrations of NO_2 and $PM_{2.5}$ (Figure 2) and GHI (Figure 3) values appeared to cluster together in particular locations. This phenomenon, known as spatial autocorrelation, is virtually ubiquitous in geographical data, but can be problematic if not accounted for in regression models. In particular, empirical studies have shown that OLS regression on spatially autocorrelated data leads to low precision (high variance, giving poor model fit) and Type 1 errors (claiming a correlation where no such correlation exists, or claiming no correlation where a correlation does exist) (Hazigüzeller 2020). Spatial autocorrelation was formally confirmed for NO2 and PM2.5 and min, max, mean and sum GHI using a Moran's I test (Gimond 2019: appendix I). To understand the implications of spatial autocorrelation across the study area the relationship between NO_2 and min GHI and $PM_{2.5}$ and min GHI were explored using Geographically Weighted Regression (GWmodel package in R). GWR is a well-known technique designed to overcome the limitations of global regression approaches where variables are highly spatially autocorrelated, and has been used in many comparable studies, especially in public heath. For example: Su et al (2017) examined the spatial variation of the relationship at district scale between indicators of social deprivation and non-communicable chronic diseases; Cheng and Truong (2012), used GWR to explore the spatial variation of place-level disadvantages and obesity in Taiwan, Morrissey et al (2015) explored the spatial variability in the relationship between long-term limiting illness and area-level deprivation at the city scale. In this study, the GWR approach described by Brunsdon (2015) was followed. The approach involves dividing the study area up into local circular windows known as kernels, in which the diameter of the circle is known as the bandwidth, and carrying out individual local regressions within each kernel. Data points with the kernel are weighted according to their distance away from the centre of the kernel, giving them declining influence in the regression equation as distance increases (Brunsdon 2015). The smaller the kernel bandwidth, the smaller the dataset included in the regression equation. Clearly, there is a tradeoff between bandwidth and reliability, since a kernel bandwidth equal to or greater than the size of the study area may add no information that cannot be obtained through the global regression model, while a very small bandwidth may give unreliable results due to the small number of data points included in the regression equation. The problem can be addressed by estimating the standard error of the analysis at different bandwidths. GWR analysis was carried out for 10,000, 5,000 and 2,500 metre bandwidths, and standard error was computed for each set of results using the bootstrapping technique (Brunsdon 2015).

4 Results

4.1 Ordinary least squares regression

Global level OLS regression revealed a clear negative correlation between level of household income and exposure to both NO_2, and $PM_{2.5}$. Although the model goodness of fit did vary depending on the resolution and grid size chosen, the correlation was clearly present at all resolutions and both grid sizes (Table 1). The strongest association (highest coefficient of determination R^2 and lowest residual standard error RSE) between income and air pollution was found for minimum gross household income (MGHI) and NO_2, and MGHI and $PM_{2.5}$ (Figure 4). The global regression models explained between 10% and 20% of the variance for MGHI and NO_2, and between 12% and 19% of the variance for MGHI and $PM_{2.5}$, depending on resolution and grid size chosen (Table 1). Standard residual error varied between 0.55-0.58 for MGHI and NO_2 and between 0.28-0.30 for MGHI and $PM_{2.5}$. Though the 2km x 2km grids do show an apparent improvement in model fit, it should be remembered that the original simulated concentrations from the CMAQ model were generated for a 1x1km grid. Greater spatial generalization in this case departs from the original data but improves fit by eliminating outliers. Having established that the correlations are reliable at all resolutions and both grid sizes, the 1km x

1km grid was used with the 48 x 48m resolution GHI data for the GWR analysis, being the original model grid size and the recommended resolution according to Piwowar's rule (Congleton 1997). As the best performing variable in the OLS analysis, only MGHI was retained as the independent variable in the GWR analysis.

Table 1: Results of the OLS regression for minimum household income (MGHI) for different resolutions and grid sizes. All $p < 0.0001$. The name convention in the first column "NO2renta1_1_100clip" refers to the pollutant used (e.g. NO_2) as the dependent variable (yvar), followed by the grid dimensions (1_1 or 2_2), followed lastly by the cell resolution of the rasterized income map from which data were extracted (48,100,200 or 500).

	Analysis	Model	Coefficient	R^2	RSE	dfree	F-statistic
1	NO2renta1_1_100	yvar ~ xmin	-0.482	0.137	0.569	1878	297.759
5	NO2renta1_1_200	yvar ~ xmin	-0.464	0.127	0.572	1878	273.417
9	NO2renta1_1_48	yvar ~ xmin	-0.489	0.14	0.568	1878	306.373
13	NO2renta1_1_500	yvar ~ xmin	-0.428	0.103	0.58	1878	216.674
17	NO2renta2_2_100	yvar ~ xmin	-0.607	0.196	0.548	502	122.353
21	NO2renta2_2_200	yvar ~ xmin	-0.585	0.18	0.553	502	110.192
25	NO2renta2_2_48	yvar ~ xmin	-0.592	0.188	0.551	502	116.45
29	NO2renta2_2_500	yvar ~ xmin	-0.58	0.17	0.557	502	102.872
33	PM25renta1_1_100	yvar ~ xmin	-0.251	0.139	0.294	1878	303.087
37	PM25renta1_1_200	yvar ~ xmin	-0.248	0.135	0.295	1878	293.933
41	PM25renta1_1_48	yvar ~ xmin	-0.254	0.141	0.294	1878	308.491
45	PM25renta1_1_500	yvar ~ xmin	-0.24	0.121	0.297	1878	258.155
49	PM25renta2_2_100	yvar ~ xmin	-0.306	0.185	0.286	502	114.268
53	PM25renta2_2_200	yvar ~ xmin	-0.3	0.176	0.288	502	107.085
57	PM25renta2_2_48	yvar ~ xmin	-0.3	0.18	0.287	502	110.171
61	PM25renta2_2_500	yvar ~ xmin	-0.292	0.161	0.29	502	96.005

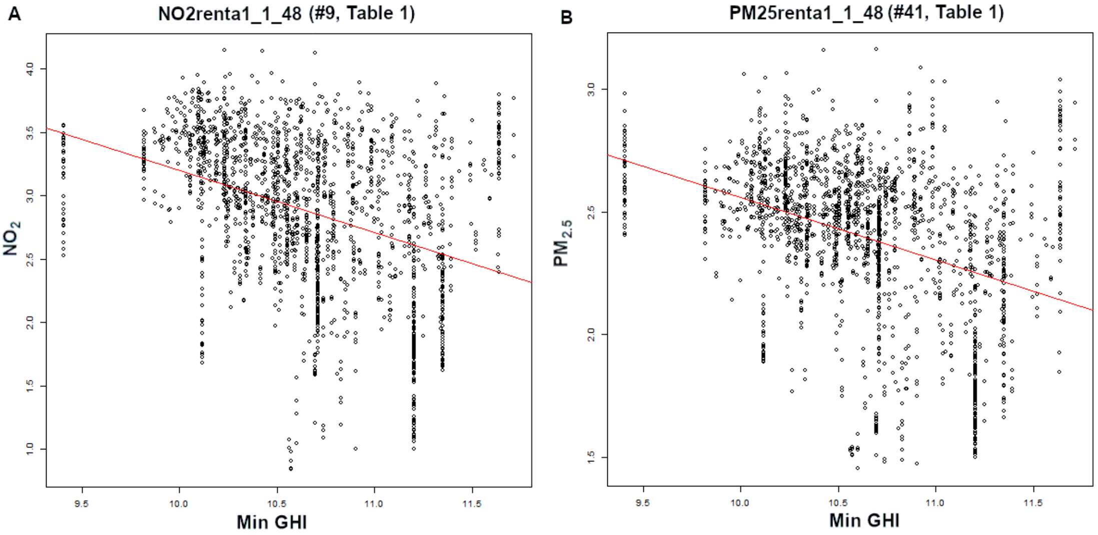


Figure 4. A (left): Scatter plot of Minimum GHI (x-axis) against NO2 (y-axis) for analysis of 1km x 1km grid

with GHI data from 48m x 48m raster (Table 1, #9), with OLS regression line fit (red). B (right): As A, for PM2.5 (y-axis) (Table 1, #41).

3.2 Testing for spatial autocorrelation

Moran's I statistic was applied using the Monte Carlo simulation method described by Gimond (2019) and all variables were found to be highly spatially autocorrelated at all cell sizes and resolutions ($p < 0.01$). The low precision of the global regression models (Table 1) is likely to be at least partly due to the spatial autocorrelation phenomenon.

3.3 Geographically Weighted Regression (GWR)

To explore the degree of variation across the study area implied by the spatial autocorrelation test, we used the coplot function described by Brunsdon (2015) to split the study area into equal sized panels and visualize the relationships in each part of the study area using the separate panels. The coplots (Figure 5) show a steeper regression line in the central northern part of the study area (top centre), and a much shallower one elsewhere, with the tendency being flat or even slightly reversed at the easternmost extreme (right centre). This indicates a steeper rate of decrease in concentrations of contaminants as minimum income increases in the north of the study area, and a relationship which is either absent or undetectable in the east. There is little difference in the pattern between either of the two contaminants. Also notable in both plots is the smaller spread of the y-axis data, corresponding to NO_2 (Fig 5A) and $PM_{2.5}$ (Fig 5B), in the centre of the study area where nearly all contamination concentrations are high, compared especially to the upper panels where there is more variation in contamination concentrations, with a larger number of lower values indicating cleaner air to the north of the study area.

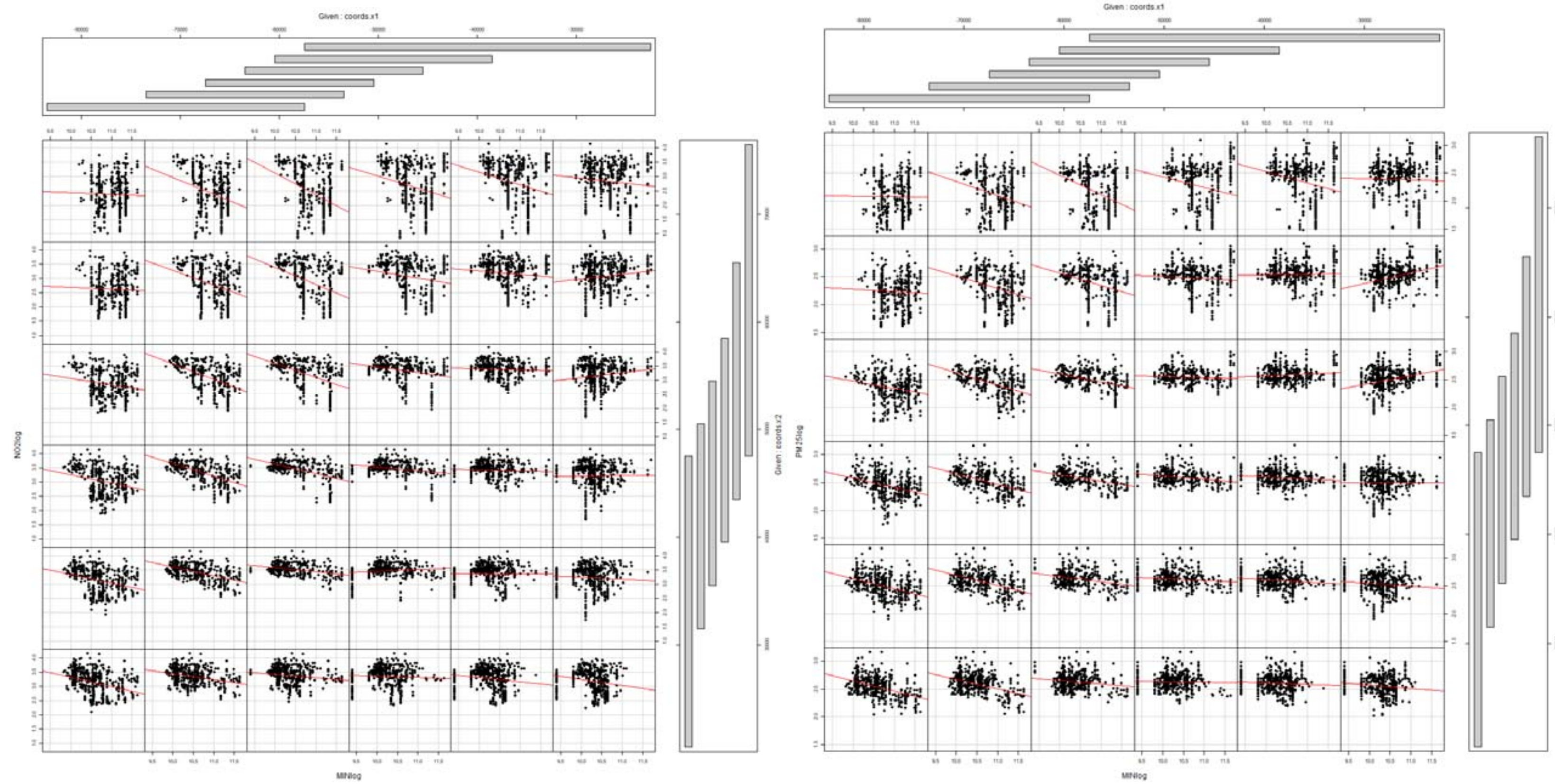


Figure 5. *Left:* A: coplot for NO_2 indicating variation in regression relationships across the study area*: Right:* B: coplot for $PM_{2.5}$ indicating variation in regression relationships across the study area.

The coplots confirmed the impression of high spatial heterogeneity indicated by the Moran's I test. The GWR results allowed this phenomenon to be explored in more detail, indicating a much greater range of variation in the regression coefficients that could be seen from the global OLS results (Table 2).

Analysis	**Variables (y ~ x)**	**Coefficient**	**Coefficient Min**	**Coefficient median**	**Coefficient Max**
Global OLS NO_2	NO_2 ~ MinGHI	-0.489	_	_	_
GWR NO_2 BW 10000	NO_2 ~ MinGHI	_	-0.602	-0.226	0.396
GWR NO_2 BW 5000	NO_2 ~ MinGHI	_	-0.809	-0.072	0.895
GWR NO_2 BW 2500	NO_2 ~ MinGHI	_	-1.160	0.006	1.512
Global OLS PM2.5	$PM_{2.5}$ ~ MinGHI	-0.254	_	_	_
GWR $PM_{2.5}$ BW 10000	$PM_{2.5}$ ~ MinGHI	_	-0.291	-0.1	0.304
GWR $PM_{2.5}$ BW 5000	$PM_{2.5}$ ~ MinGHI	_	-0.455	0.011	0.594
GWR $PM_{2.5}$ BW 2500	$PM_{2.5}$ ~ MinGHI	_	-0.678	0.056	0.531

Table 2: Results of GWR analysis, in comparison with results of OLS regression.

While the global OLS regression equation estimated a coefficient value (m in the linear regression equation y=mx+c) across the whole study area of -0.489 (NO_2) and -0.254 ($PM_{2.5}$), GWR coefficient estimates unsurprisingly vary much more widely. For every one unit of change to the variable MinGHI, at bandwidth 10000m, the average increase in the response variable NO_2 or $PM_{2.5}$ varies from -0.602 to 0.396 (NO_2) and -0.291 to 0.304 ($PM_{2.5}$) (Table 2). As the table shows, the variation increases as bandwidth decreases. Though these values are not intuitively meaningful because of the log transformation, the change of sign indicates an important difference in the regression line depending on the specific locality investigated. This analysis provides more detail than we could obtain from the coplots, and reliable quantification of the degree of variation in the relationships explored.

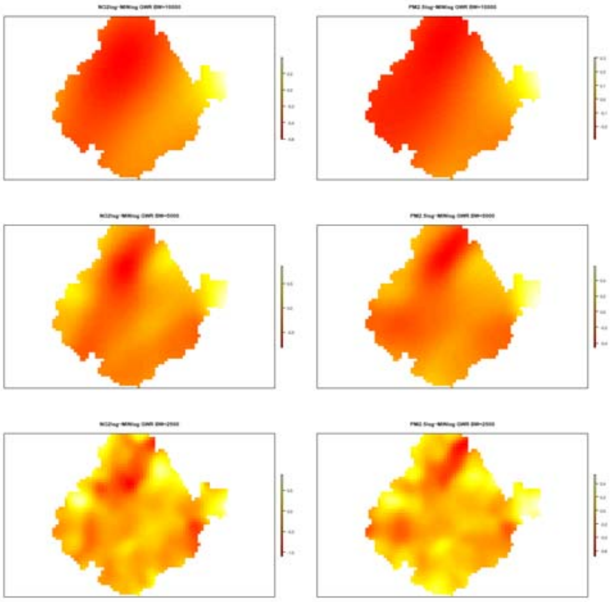


Figure 7. GWR results for NO2log~minlog (left) and PM25log~minlog (right), for each bandwidth. In both cases, the sharpest decline in contaminant concentrations as minimum income per household increases (steepest trendline) are found in the north and central parts of the study area. Reducing the bandwidth seems to produce more highly resolved patterns, however, the extreme variation in the coefficient estimates across bandwidths indicates a high level of uncertainty.

The plots of these results (Figure 7) clearly shows variation in the relationship between contaminant concentrations and minimum income, with increasing minimum income leading to a steeper fall off in contamination in the northern and central parts of the study area. Although smaller kernel bandwidths seem to produce more highly resolved patterns, the extreme variation in the coefficient estimates across bandwidths (Table 2) indicates a very high level of uncertainty. To investigate the reliability of the GWR analysis results, estimates of the standard error were computed using the bootstrapping technique (ref).

The results of the bootstrap estimation of model standard error show that reducing the kernel size increases error in the coefficient estimates to an unacceptable degree (Table 3, Figure 8). For the 10000m bandwidth models, estimated standard error is around 20% of the coefficient estimate values for three quarters of the data (18.82 for NO_2 and 22.37 for $PM_{2.5}$ at the 3rd quartile). Though already quite large, the boxplot (Fig 8) shows how this error seems to increase sharply as the kernel bandwidth is reduced. For the largest bandwidth, though the spatial variation of the relationship is rather general, we can conclude that the north-south differential is likely to be real, since the errors in the north and west of the study area, where the spatial variation is most evident are relatively low (5-20% of coefficient estimate) (Fig 9). Results from the eastern part of the study area are inconclusive, on the one hand because regression lines show very weak evidence of correlation (Fig 6), and on the other, because the estimated standard errors are unacceptably large in this area (Fig 9).

	Bandwidth	Min	1st Quartile	Median	Mean	3rd Quartile	Max
NO2	10000	5.93	8.02	12.17	20.63	18.82	1124.60
	5000	8.27	17.20	28.05	80.62	50.42	7980.16
	2500	7.24	26.73	45.12	321.19	96.28	78945.14
PM2.5	10000	6.01	8.28	13.04	35.59	22.37	8114.201
	5000	7.62	15.04	28.06	138.59	65.33	39806.03
	2500	7.95	24.81	44.14	251.66	99.37	83506.4

Table 3. Standard Error (SE) estimates for GWR as percentages of the coefficient estimates.

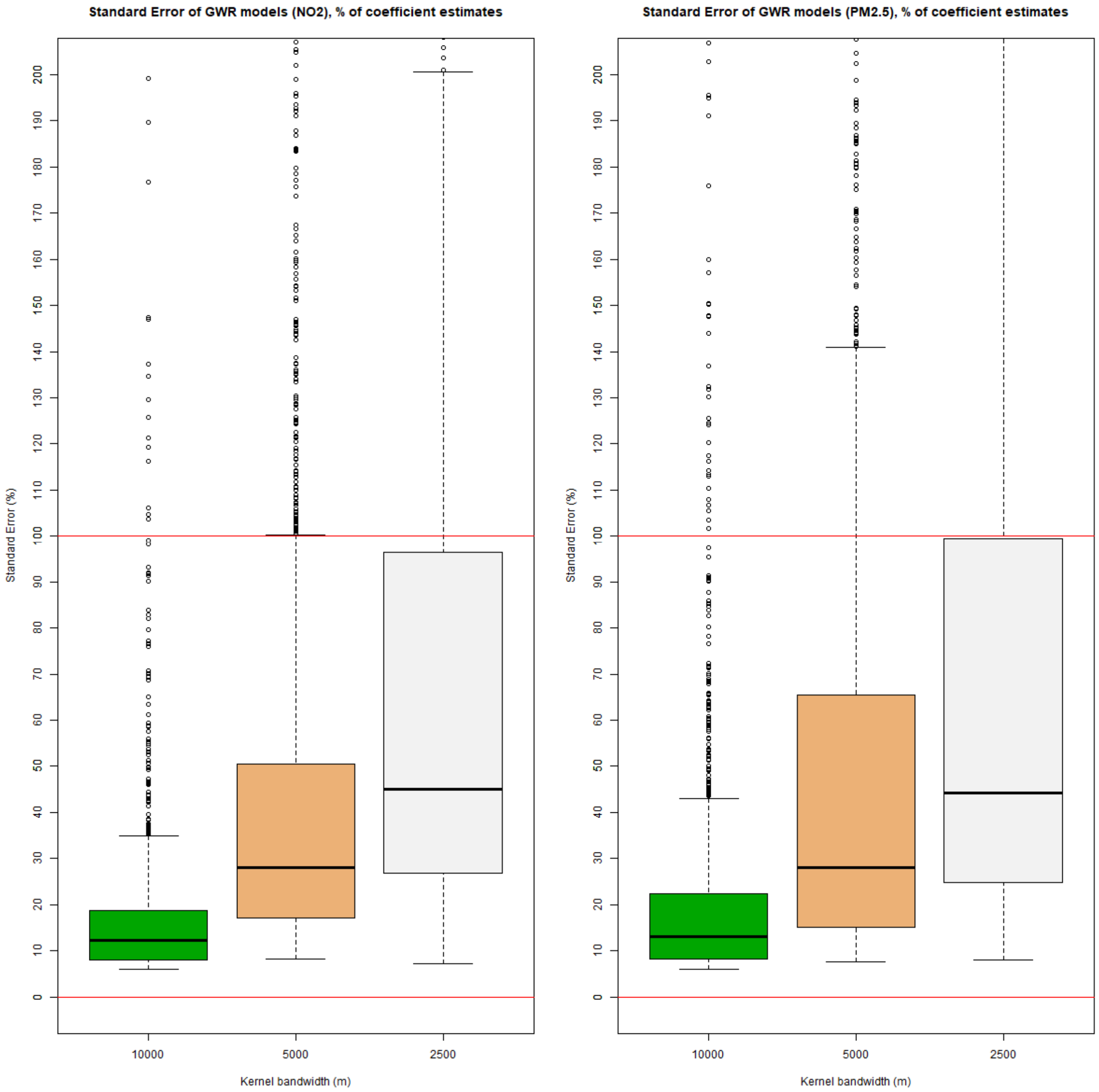


Figure 8. Standard Error (SE) estimates for GWR results for NO2 (left) and *$PM_{2.5}$* (right), for each bandwidth as a percentage of the coefficient estimates. Horizontal red lines indicate SE of 0% and 100% of coefficient estimates. Note that for the PM2.5 GWR analysis results, the percentage error is greater than for *NO_2* for both 10000m and 2500m bandwidths.

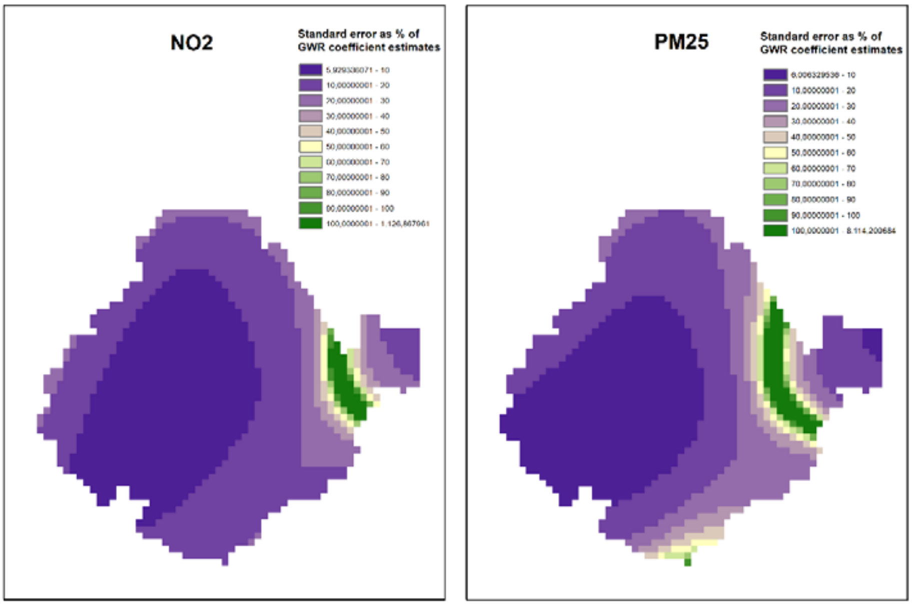


Figure 9. Spatial distribution of estimated Standard Error in GWR models for bandwidth 10000 for NO_2 (left) and $PM_{2.5}$ (right). Compare with Figure 7 (top row). The lowest SEs are found in the central northern and western parts of the study area, where the GWR indicates a stronger decline in contamination exposure as minimum household income increases.

4. Concluding discussion

Our analysis indicates a negative correlation between levels of exposure to contamination from NO_2 and $PM_{2.5}$ and gross household income. In general, within the metropolitan area of Madrid, comprising the city of Madrid and adjacent suburban municipalities, as minimum income increases, the concentrations of airborne contaminants as estimated by a Eulerian photochemical air quality model, decreases. It is notable that the most important income variable, by some margin, was *minimum*, rather than total, maximum or mean household income. This suggests that policy approaches that look to increase minimum incomes overall are likely to have positive impacts on quality of life.

The pattern shows considerable spatial heterogeneity, in particular a sharper decline in contamination concentrations with increasing minimum income in central and northern parts of the study area, corresponding, approximately to the municipalities of Colmenar Viejo and Tres Cantos in the north, and Las Rozas, Majadaonda, Boadilla de Monte, Pozuelo de Alarcón to the north-west of the city, and the northern part of the municipality of Madrid (Fuencarral-El Pardo). The results are not altogether surprising, since these areas enjoy low urban and transport network density and many parks and other green areas. The slope of the regression line between contamination and minimum income (Figure 6) was shallower in central and central-southern areas, mainly because there were fewer localities in this area where contamination concentrations were low, unlike in the north.

Some limitations should be noted, as follows. Our results are rather general, but attempts to spatially disaggregate further by decreasing the GWR bandwidth increased uncertainty to unacceptable levels (SE as % of coefficient estimates >20% for 75% of the data). To remedy this, and understand how exposure to poor air quality varies at the individual district level, estimated contaminant concentrations would be needed at a finer spatial resolution than currently available from the air quality model used ($1km^2$). This is unfortunately a rather difficult task. Though higher spatial resolution products are known to exist, their reliability can be questioned (ref). At the same time, a different division of the territory than that of the census tracts (i.e household or street level data) would likewise improve the level of detail obtainable. However, for reasons of data protection, such data are not normally freely available. In future studies, the analysis could be extended by application to different pollutants, such as Ozone (O_3), or Sulphur Dioxide (SO_2). Finally, although we have used gross household income, in future research it would be interesting to use other indicators of social deprivation, such as separating population by work activity, level of education, age range, and so on. Composite indicators of multiple deprivation are available for some countries and regions (e.g., Scotland), and may be obtainable for the case of Spain.

References

Ayuntamiento de Madrid (2019). Air quality and climate change plan for the city of Madrid. Gen. Sustain. Environ. Control. Disponible en: https://www.madrid.es/UnidadesDescentralizadas/Sostenibilidad/CalidadAire/Ficheros/PlanAire&CC_Eng.pdf

Borge, R., Lumbreras, J., Pérez, J., de la Paz, D., Vedrenne, M., de Andrés, J. M., & Rodríguez, M. E. (2014). Emission inventories and modeling requirements for the development of air quality plans. Application to Madrid (Spain). *Science of the Total Environment*, 466, 809-819.

Borge, R., Artíñano, B., Yagüe, C., Gomez-Moreno, F. J., Saiz-Lopez, A., Sastre, M., ... & Cristóbal, Á. (2018). Application of a short term air quality action plan in Madrid (Spain) under a high-pollution episode-Part I: Diagnostic and analysis from observations. *Science of the Total Environment*, 635, 1561-1573.

Brunsdon (2015) Geographically Weighted Regression. https://rpubs.com/chrisbrunsdon/101305

Congalton, R. G. (1997). Exploring and evaluating the consequences of vector-to-raster and raster-to-vector conversion. *Photogrammetric Engineering and Remote Sensing*, 63(4), 425-434.

Gimond, M (2019) A basic introduction to Moran's I analysis in R. https://mgimond.github.io/simple_moransI_example/

Hacıgüzeller, P. (2020). Spatial applications of correlation and linear regression. In *Archaeological Spatial Analysis* (pp. 135-154). Routledge.

Izquierdo, R., Dos Santos, S. G., Borge, R., de la Paz, D., Sarigiannis, D., Gotti, A., & Boldo, E. (2020). Health impact assessment by the implementation of Madrid City air-quality plan in 2020. *Environmental research*, 183, 109021.

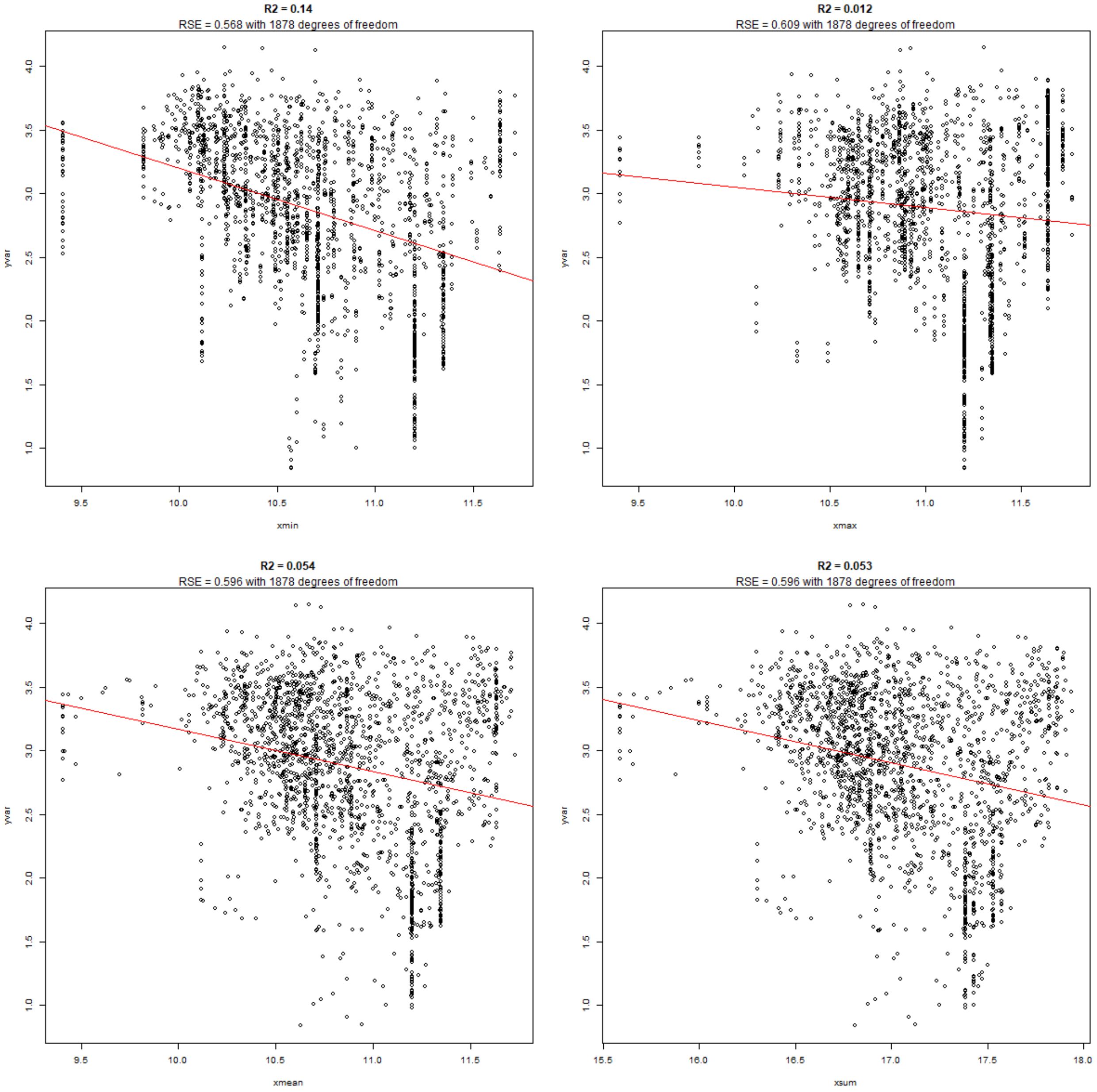
Quaassdorff, C., Borge, R., Pérez, J., Lumbreras, J., de la Paz, D., & de Andrés, J. M. (2016). Microscale traffic simulation and emission estimation in a heavily trafficked roundabout in Madrid (Spain). *Science of the Total Environment*, 566, 416-427.

Prieto-Flores, M. E., Gómez-Barroso, D., & Jiménez, A. M. (2021). Geographic health inequalities in Madrid City: exploring spatial patterns of respiratory disease mortality. *Human Geographies*, 15(1), 5-16.

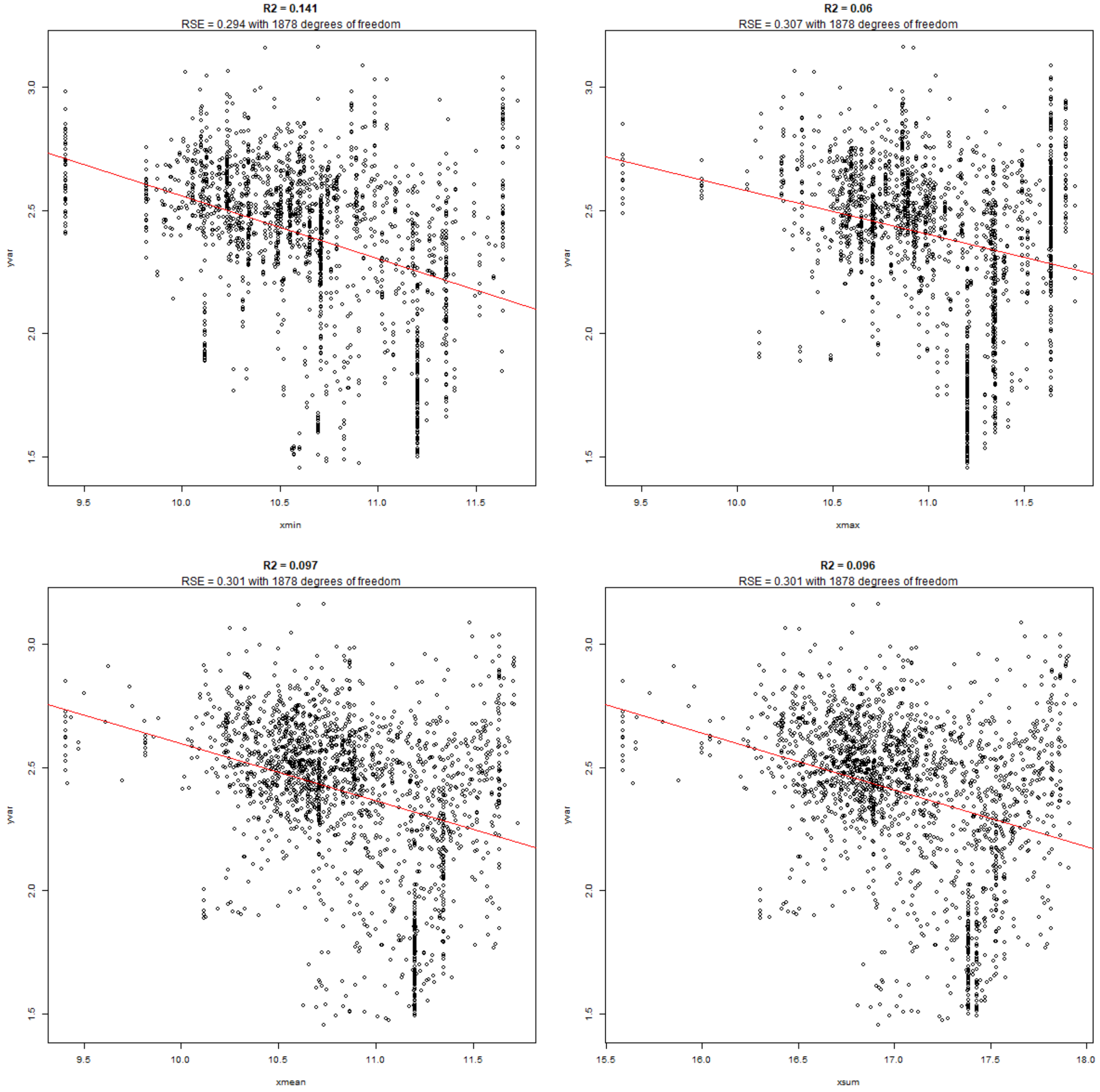
Openshaw, S. (1981). The modifiable areal unit problem. *Quantitative geography: A British view*, 60-69.

Organiación Mundial de Salud (OMS) (2021) Ambient (outdoor) air pollution Key facts https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Appendix 1: Results of the OLS regression for NO2 and GHI at 48m resolution and 1km x 1km grid. NO2 (yaxis) vs (Clockwise from top left) Min GHI; Max GHI; Mean GHI; Total GHI.



Appendix 2: Results of the OLS regression for PM2.5 and GHI at 48m resolution and 1km x 1km grid. NO2 (yaxis) vs (Clockwise from top left) Min GHI; Max GHI; Mean GHI; Total GHI.


R2 = 0.141
RSE = 0.294 with 1878 degrees of freedom
yvar
xmin
R2 = 0.06
RSE = 0.307 with 1878 degrees of freedom
yvar
xmax
R2 = 0.097
RSE = 0.301 with 1878 degrees of freedom
yvar
xmean
R2 = 0.096
RSE = 0.301 with 1878 degrees of freedom
yvar
xsum