Article

Air pollution modeling in Austin using Google Street View Cars

Rijul Gosar

Abstract: Air pollution has been one of the leading causes of mortality, not only in developing and underdeveloped countries, but also in developed nations such as the United States. The current paper looks at the use of mobile monitoring to build a Land Use Regression (LUR) model in the city of Austin, Texas. In the summer of 2018, two Google Street View (GSV) vehicles mounted with gas (CO_2, NO, NO_2) and particle (BC, $PM_{2.5}$, UFP) instruments were driven continuously to make a high-resolution map showing variation of these contaminants. The observations were processed in MATLAB and ArcGIS Pro to extract covariates such as Land Cover, Road characteristics, and Population density, which were used to build a population exposure map and the LUR model for two pollutants- CO_2 and UFP. The resulting map showed that sources and population exposure are not necessarily correlated. For the developed LUR model, it was found that the model was able to explain only 44% and 25% of the variability in concentrations for CO_2 and UFP respectively, indicating the need for better covariates to develop the model.

Keywords: air pollution; land use regression; GIS; mobile monitoring

---

1. Introduction

Poor air quality has been a leading cause of premature death across the globe [1,2]. Additionally, the high spatial variability in air quality (sometimes at the block level) makes it more challenging to understand causes and effects of local fluctuations in contaminant concentration. Here, mobile monitoring offers the chance to make highly detailed maps which could help identify hotspots within a city, and consequently sources which need to be controlled for the benefit of the populace. Using this data, a Land Use Regression (LUR) model can be built to help predict concentrations using more readily available information, such as Land Cover, population, and road network characteristics. This study follows a similar campaign that was conducted in Oakland, California [3,4].

Land Use Regression models developed from mobile monitoring measurements serve as the basis for making predictions about unknown, highly variable contaminant levels using easily measurable or available metrics such as Land Cover, Population density, Vegetation Index, Road characteristics, and distance from a major road to name a few. GIS based analysis of such data is the benchmark for developing these simplified models, which serve as a reasonable replacement for more expensive and computationally exhaustive climate models. Eventually, the developed model could then be used to make predictions in locations with similar weather patterns.

2. Data

Two Google Street View (GSV) vehicles were driven around the city of Austin, Texas, over a period of two months in the summer of 2018. The vehicles were mounted with gas and particle instruments, recording data at a 1-Hz frequency. Gas data was collected for Carbon Dioxide (CO_2), Nitric Oxide (NO), and Nitrogen Dioxide (NO_2). Particle instruments measured Particulate Matter ($PM_{2.5}$), Black Carbon (BC), and Ultrafine Particles (UFP). The analysis presented below focuses on CO_2, NO_2, and UFP. ArcGIS Pro has been used for two primary reasons- building a map to show population exposure of contaminants measured, and to extract covariates required for building the LUR model.

Land Cover information is extracted from the ArcGIS Pro living atlas' National Land Cover Database (2011) [6]. The raster, which is available for the entire continental United States is extracted for Austin. It is then vectorized, and this is then used for further analysis. Land cover information from this gives the area of developed land use, vegetation, and water in the domain of interest. These metrics serve as one of the spatial covariates needed for model development.

Population information is obtained from the United States Census Bureau. Block-level data is available for each state [7]. From this, population density in each block is retrieved, which is fed as a covariate in the LUR model. The assumption made here is that the population density is constant throughout the block, which may not be accurate.

Lastly, one of the key driving forces of air quality is the category of road closest to the point of exposure and proximity to the same. Intuitively, people living closer to heavy traffic roads such as freeways are prone to greater exposure, while those living further away are not as exposed. To obtain this road network information for Austin, the OpenStreetMap (OSM) database [8] has been used. This is an open source portal that provides detailed road network information for the entire world, from which the data for Texas (and consequently, Austin) is obtained.

Figure 1 shows the domain of study, the points indicating sample points where concentration measurements were made. The points are equidistant, each 30m apart. The boundary indicates a 1km buffer around the sampling domain.

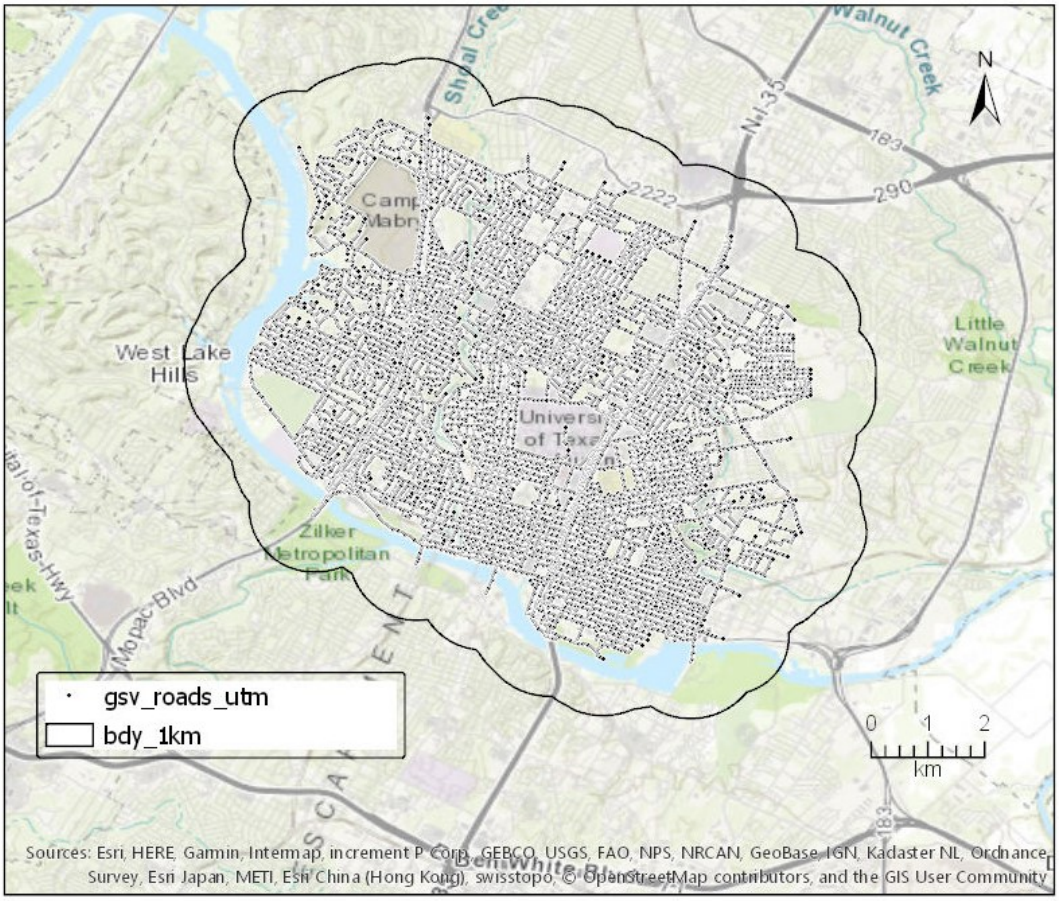

Figure 1. The study region.

2.1. Data pre-processing

Before the data can be visualized in a GIS environment, certain pre-processing steps are necessary. Raw data cleaning is done using MATLAB, the steps for which are not detailed here. The output

of pre-processing gives the median values for concentration for each pollutant sampled along with geographical information (latitude, longitude) which is the input to GIS. Several issues need to be addressed in processing, which are listed below.

- Representativity: The biggest challenge in mobile monitoring is relying on one instance of measurement to reliably represent the general trend at a location. To overcome this, each location is repeatedly sampled, to capture the overall behaviour.
- Temporal adjustment: Every location is sampled at a given time of day. This implies that every measurement collected is only an indicator for the location's pollution trend at that instant. To tackle this problem, locations are driven at different times to account for diurnal variability.
- Speed adjustment: Since the vehicle covers unequal distances depending on traffic conditions or speed limits, all raw data points need to be "snapped" to equidistant sampling locations for analysis. Here, all points shown in Figure 1 are 30m apart.

3. Materials and Methods

To build reliable and meaningful maps, ensuring consistency between the different data used is essential. Discrepancies in co-ordinate systems used, and raster or vector representation, can arise due to the multiple data sources. This section focuses on understanding and assimilating the different data used and techniques used for processing.

3.1. Data Analysis

The first step in data analysis is assimilation of all gathered data. Thus, the Land Cover information extracted for Austin, block-level population data, and road network information need to be consolidated for the study domain.

3.1.1. Land Cover data

The land cover information is obtained from the National Land Cover Dataset (NLCD) for 2011, which is available in the Living Atlas in ArcGIS Pro [6]. This gives detailed information regarding the type of land use patterns in the city. This NLCD has to be extracted for the domain of interest, as it is given for the entire continental United States. The desired information is obtained by using the "Extract by Mask" tool, which gives the land cover information for Austin. This raster is then converted to a polygon for performing mathematical calculations using the "raster to polygon" tool. The computed Land Cover polygon is then used for further analysis.

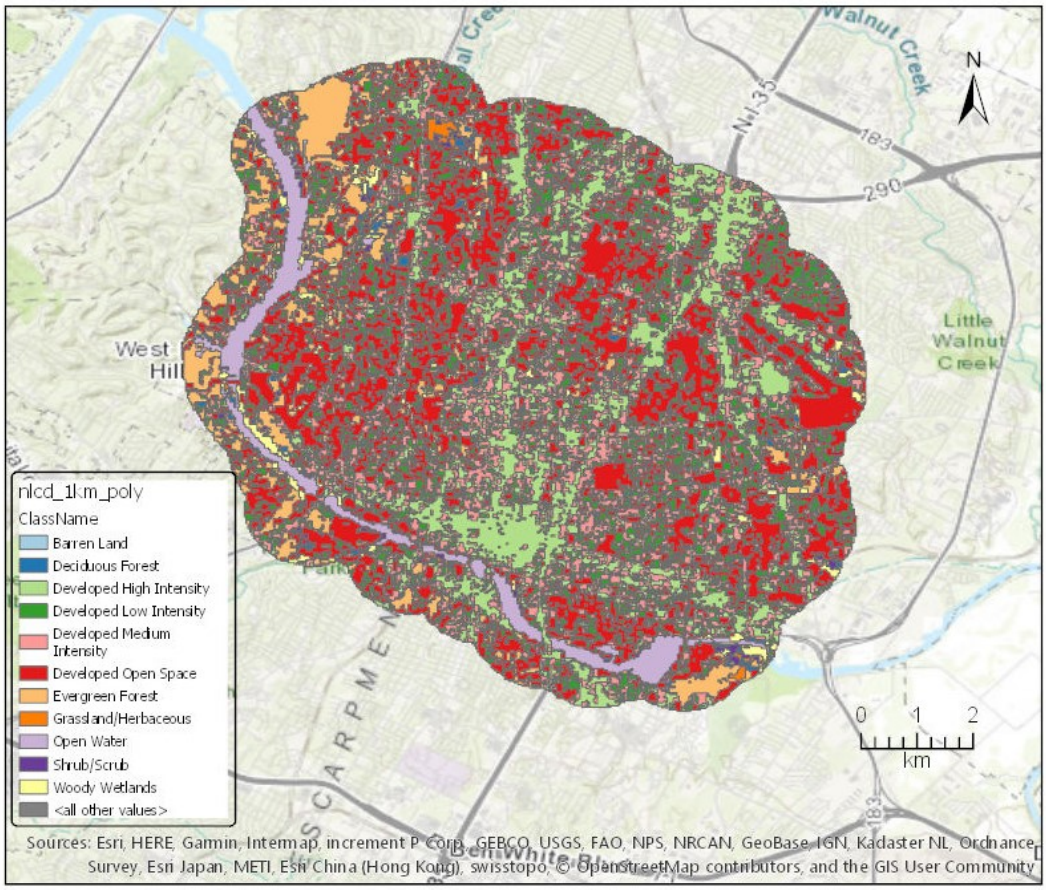

Figure 2. Land Cover polygon for study region

3.1.2. Road information

The road network information is obtained from OpenStreetMap, which has detailed road network information for regions around the world [8]. Available data includes the type of road (freeway, residential, pedestrian, etc), name of the street, whether the street is a one-way route, if the road is restricted to a certain category of vehicles (for example, only cars), and length of each road segment. This helps classify heavy- and light-traffic routes, and identifying the impact of a certain vehicle class on local concentrations. The last step involves associating each road to the corresponding point where the GSV vehicle has sampled. This is done using the "Spatial Join" tool, which adds the road type information to each sample.

The category of roads used in OSM contains many details, most of which are arbitrary. To simplify calculations and reduce computational load, the available classes of roads are grouped into four categories, listed below.

- **Highways**: Motorways and Motorway Links
- **Major Roads**: Primary, Primary link, Secondary, Secondary link, Tertiary, Tertiary link, Service road, and Unclassified
- **Residential Roads**: Residential, and Living street
- **Others**: Tracks, Cycleway, Footway, and Pedestrian

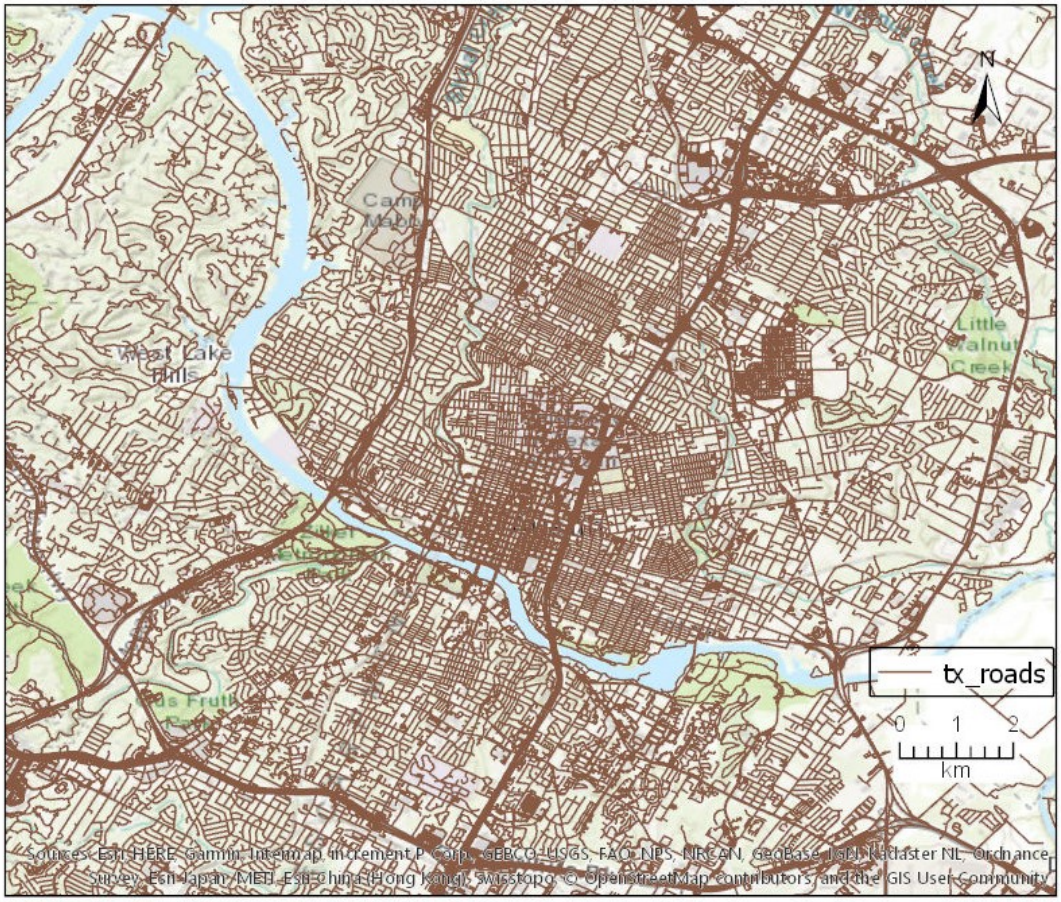

Figure 3. Road network from OpenStreetMap

3.1.3. Population data

The block-level population information is obtained from the United States Census Bureau for the state of Texas [7]. Using this, the relevant information for Austin is extracted. The Bureau gives population in each block, area, and perimeter of the blocks. This is used to compute the population density (# of people/ sq. mile) of each block as the population over the block area, which is then used.

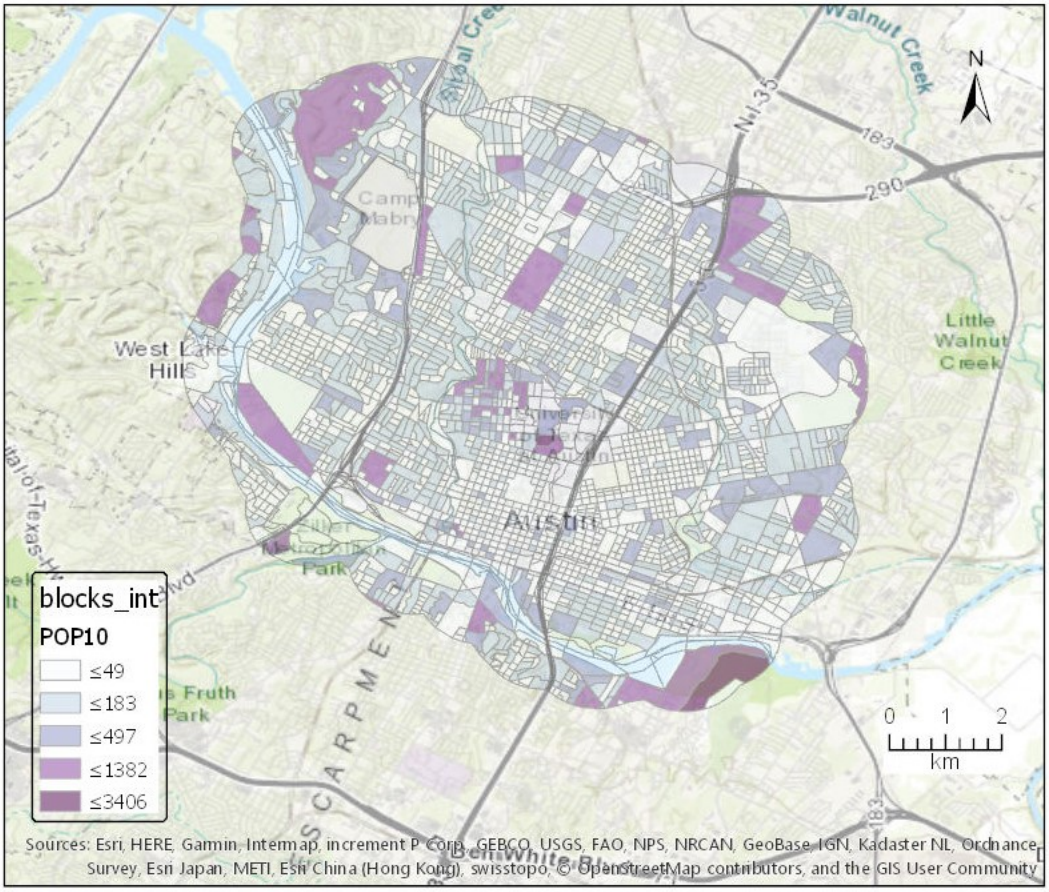

Figure 4. Block level population for study region

3.2. Data processing

Due to the high spatial variability of air pollution, building of an LUR model requires accounting for the distance of the point of exposure from the sampling location. Based on the distance from the point of contaminant release, different covariates can have varying impacts which could affect the concentration. Intuitively, for example, a location very close to a major source (such as a freeway) would primarily be impacted only by the traffic. However, concentration further downwind would also be influenced by the presence of trees, temperature, wind speed to name a few.

To account for this, buffers of multiple radii are created around each sampling point. These buffers have radii of 50m, 100m, 500m, and 1km. Varying buffers enables quantifying the importance of different covariates at different spatial resolution. To add land cover and population data to each buffered data point, the intersect function is used. This adds the land cover types in each buffer size and the number of people living in that buffer using the population density and area of each block within the buffer. This step gives a table of the land cover, road network, and population for each buffer size which is used to build the LUR model.

Due to computational constraints, only the 50m and 100m buffer sizes were able to run on ArcGIS Pro, and consequently the results for the same are presented.

4. Results

4.1. Spatial trends in Austin

The first step in quantifying population exposure is understanding where concentrations are the highest. The figures below show the median values of all data collected over the campaign for the contaminants under study, CO_2, NO_2, and UFP. The figures below show a similar spatial pattern for the pollutants under consideration. All three contaminants are influenced by traffic (or the road network) to a certain extent, with elevated concentrations observed along the major freeways; I-35 in the east and Mopac in the west.

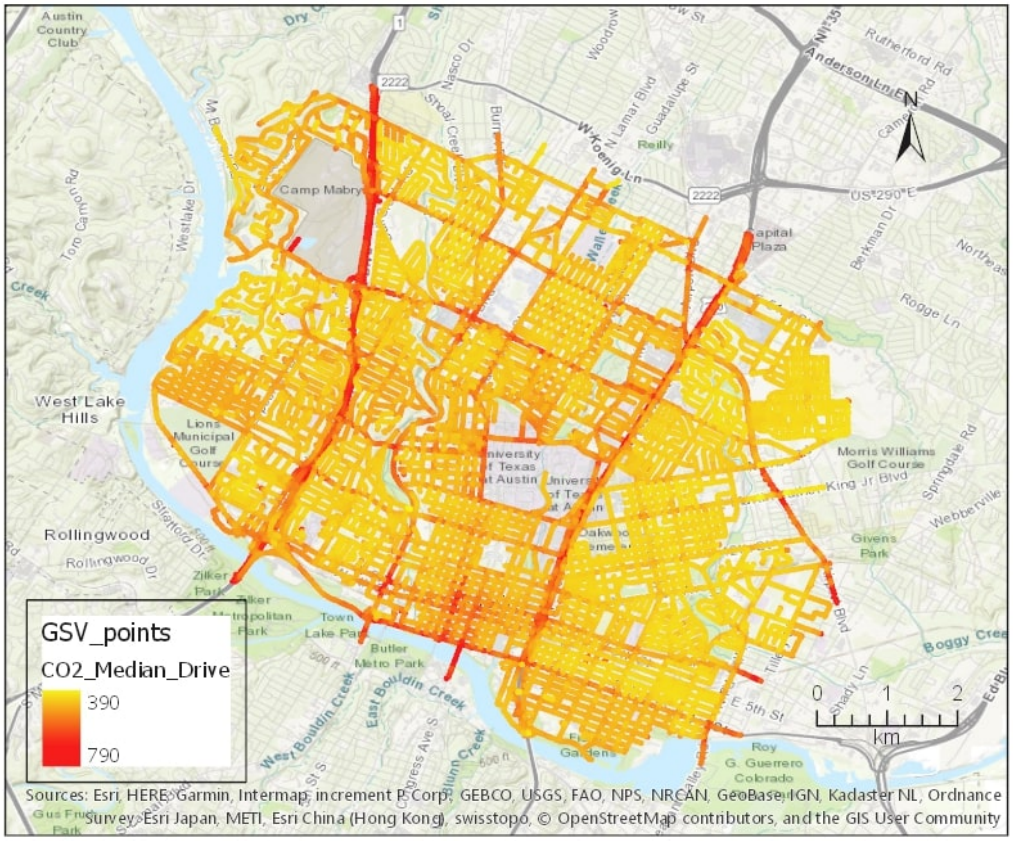

Figure 5. Spatial pattern of Carbon Dioxide, ranging from 390 ppm to 790 ppm

The higher concentrations of CO_2 (Figure 5) along the freeways is an indicator of combustion, which is one of the primary sources of the gas. Lower concentrations in the interiors are approximately 420ppm, which is only marginally above the global averaged concentration, indicating Austin is a relatively "clean" city.

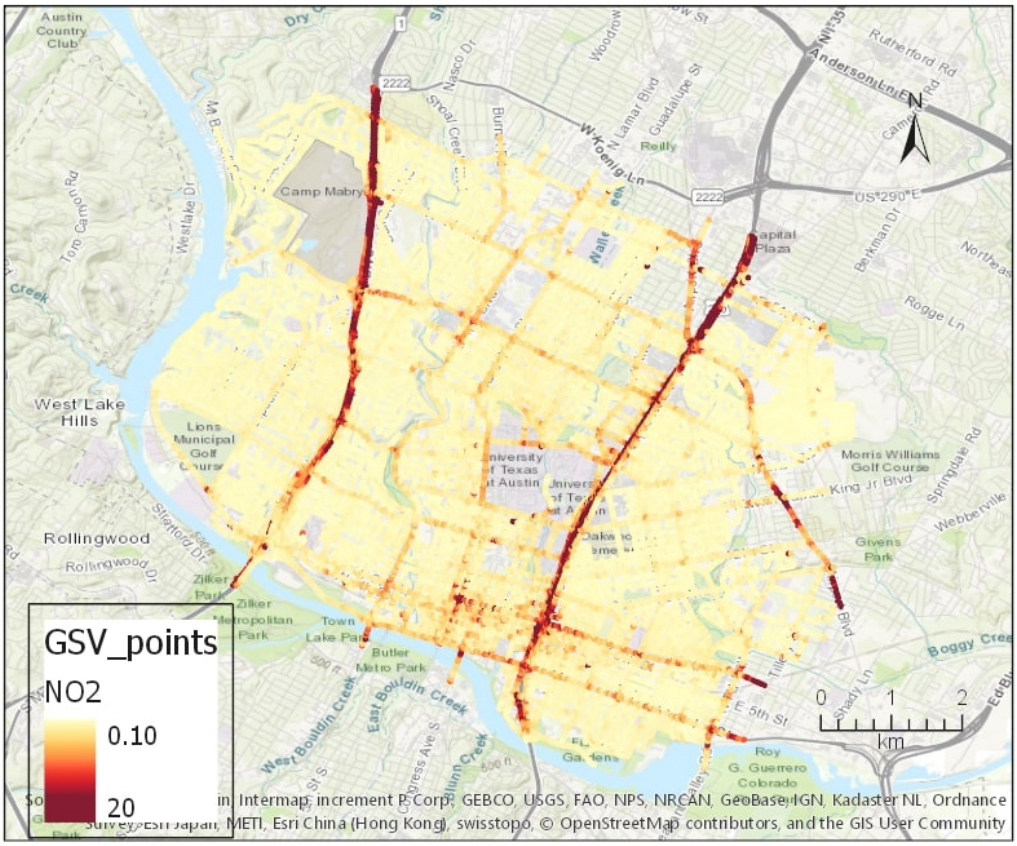

Figure 6. Spatial pattern of Nitrogen Dioxide, ranging from 0.1 ppb to 20 ppb

Looking at NO_2 (Figure 6), concentrations fall relatively rapidly as one moves away from the freeways, which is expected since NO_2 is highly photo-reactive. Since the primary source of NO_2 is vehicular emissions, the freeways stand out clearly in the map above.

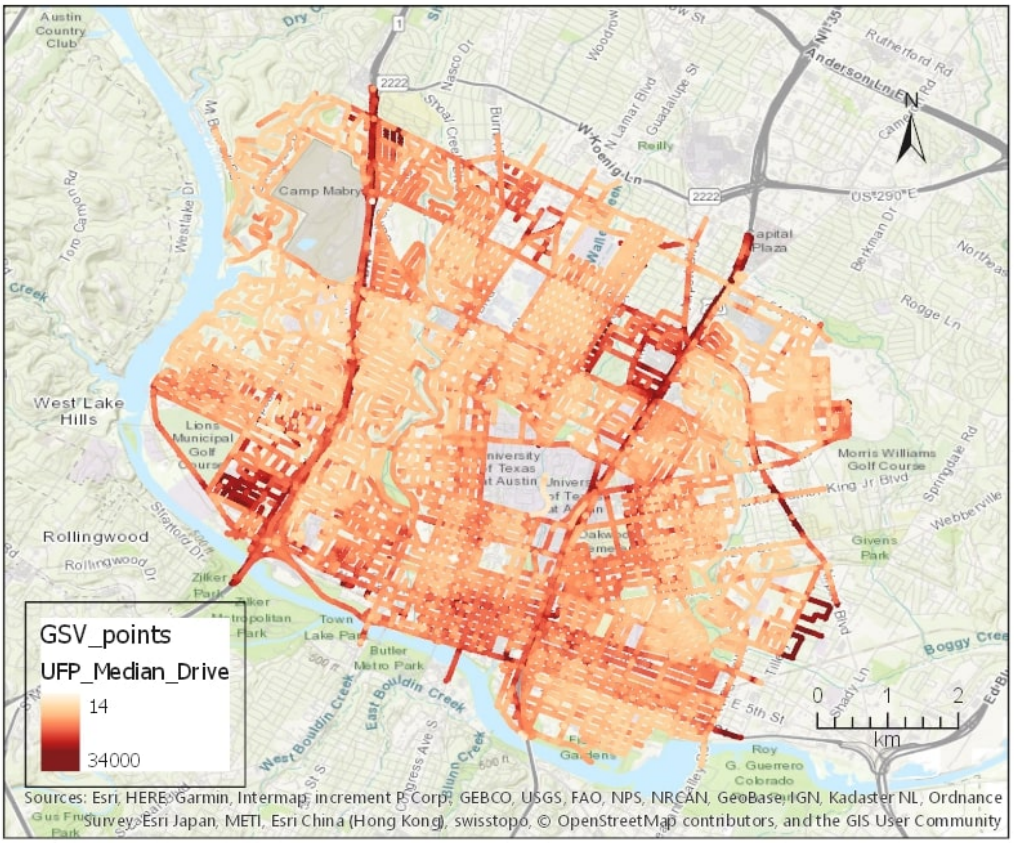

Figure 7. Spatial pattern of Ultrafine Particles, ranging from 14 #/cc to 34,000 #/cc

Figure 7 shows the distribution of Ultrafine Particles through the city. In contrast to the gas pollutants, UFPs are not as strongly influenced by traffic-heavy routes. While vehicular emissions are responsible for UFP concentrations, a significant source is also due to construction activities indicated by the darker region around downtown. Additionally, outdoor cooking (eg, barbecuing) also contributes significantly to UFP levels.

4.2. Population exposure

Building a map of the concentrations' spatial distribution allows identifying regions where people are most likely to be exposed to higher pollutant levels. By extension, this enables quantifying health impacts of higher concentration exposure and can influence policy making.

The figures below show the population by block in the domain of interest overlapped on the median concentrations of each of the three pollutants- CO_2, NO_2, and UFP. The darker shades represent higher populations within the blocks (blue colour scheme) and higher contaminant concentrations (red colour scheme) while the lighter shades represent the opposite respectively. From the figures, it is evident that people do not tend to live in particularly polluted areas. While there are blocks that seem to be preferred, these need not necessarily be where concentrations are higher. Hence, there seems to be no credible link between higher concentration areas and population exposure.

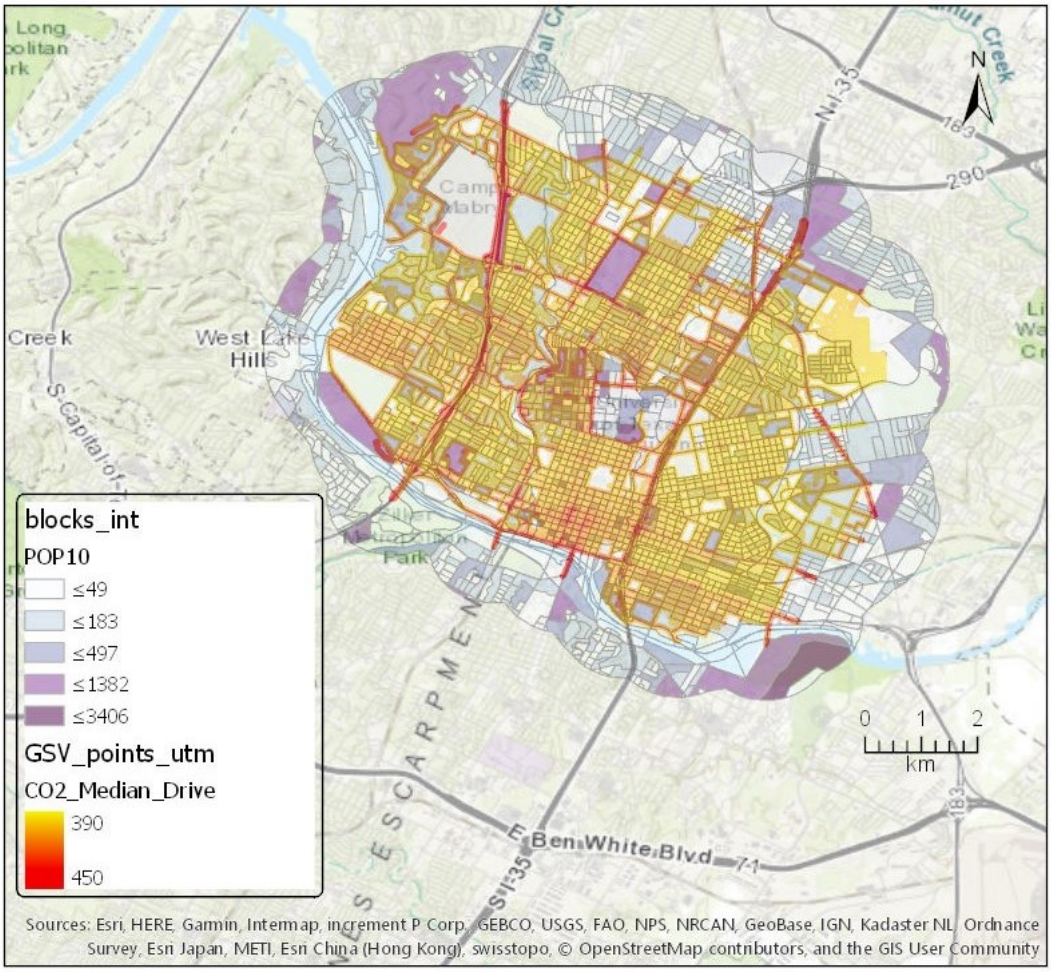

Figure 8. Population exposure to CO_2

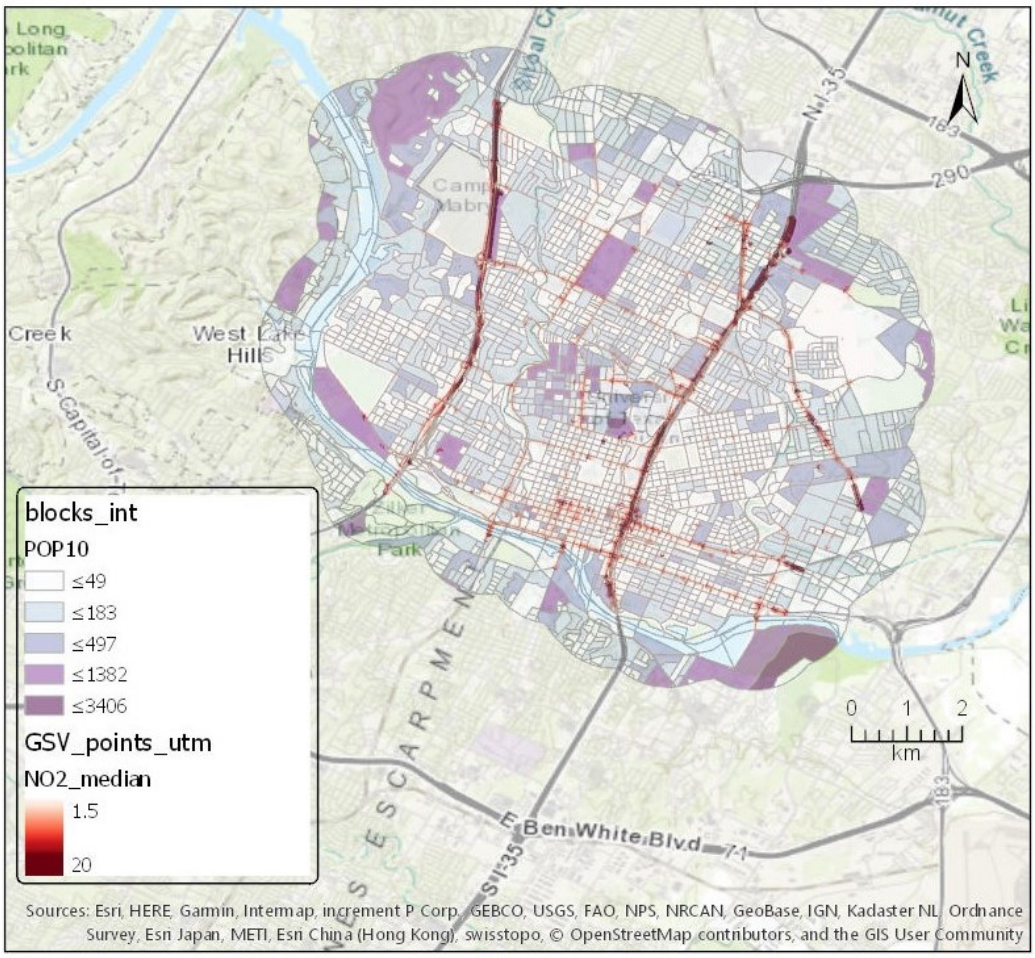

Figure 9. Population exposure to NO_2

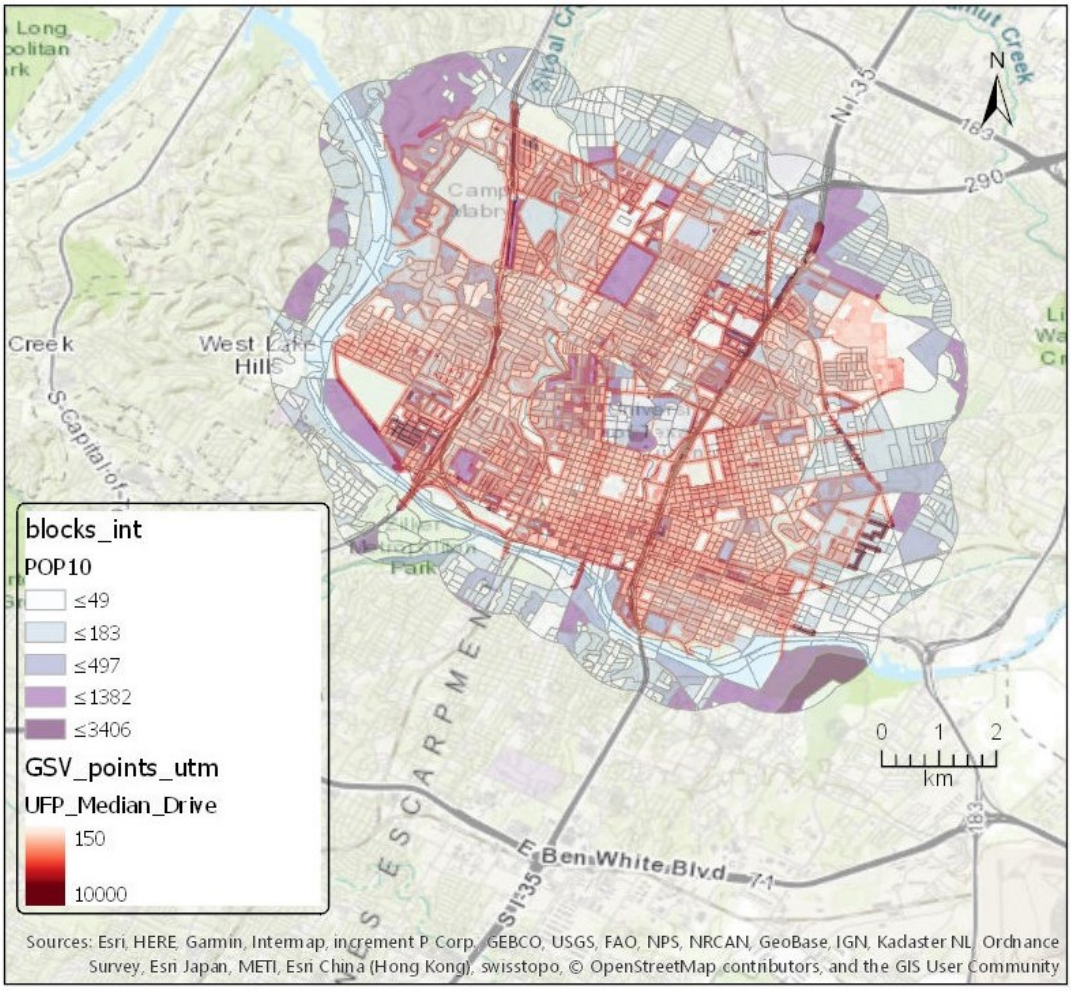

Figure 10. Population exposure to UFP

4.3. Model covariates

The final stage involves extracting the covariates needed to build the Land Use Regression Model. As mentioned above, due to limitations on computational power, only the 50m and 100m buffers were able to run on ArcGIS, the results for which are shown. The steps followed to obtain these values are given below.

1. **Spatial join** of sampling points and road network information
2. **Buffering** of joined features, with radii of 50m and 100m
3. **Intersecting** buffered points with the Land Cover polygon
4. **Intersecting** the resulting polygon with population information
5. **Summarizing** obtained feature class to obtain percentage of land cover and population in each buffer around the sampling point
6. **Exporting** resulting table to Excel to feed into the LUR model

Figure 12 shows a schematic of the aforementioned steps and Figures 13 and 14 show the result of those steps, which are the covariates fed into the LUR model for the 50m and 100m buffers respectively. Each unique Feature ID represents an independent sampling point. The road networks of the particular point are classified using a binary notation- 1 (or true) if that sampling point lies on that kind of road segment, and 0 (or false) otherwise. Land Cover information is represented as a ratio of the area of that land cover type in that buffer divided by the total area of the buffer. Lastly, population is the total

number of people living in that circular buffer calculated using the population density, intersecting block areas, and summing them all together. The 50m buffered points are shown in Figure 11 which consists of concentration, land cover, road network, and population information. The region is zoomed near the Capitol for better clarity. A similar figure (not shown here) is created for the 100m buffered points.

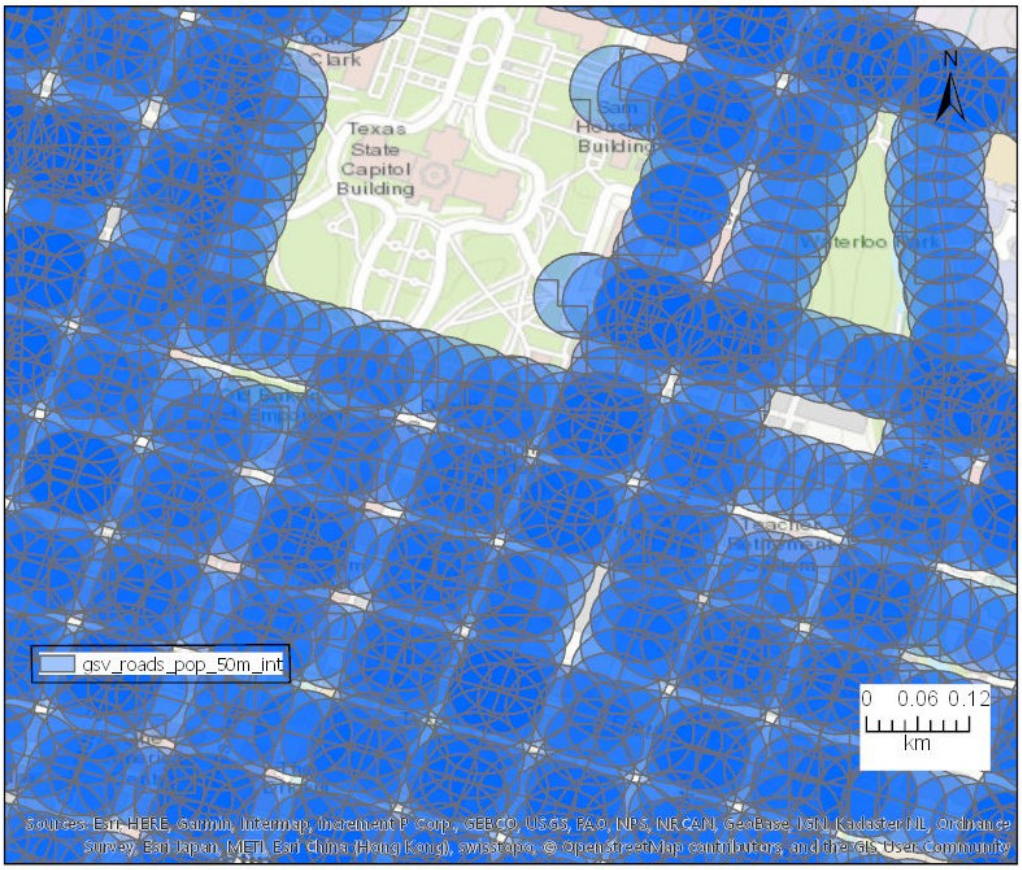

Figure 11. Sampling points with 50m buffers

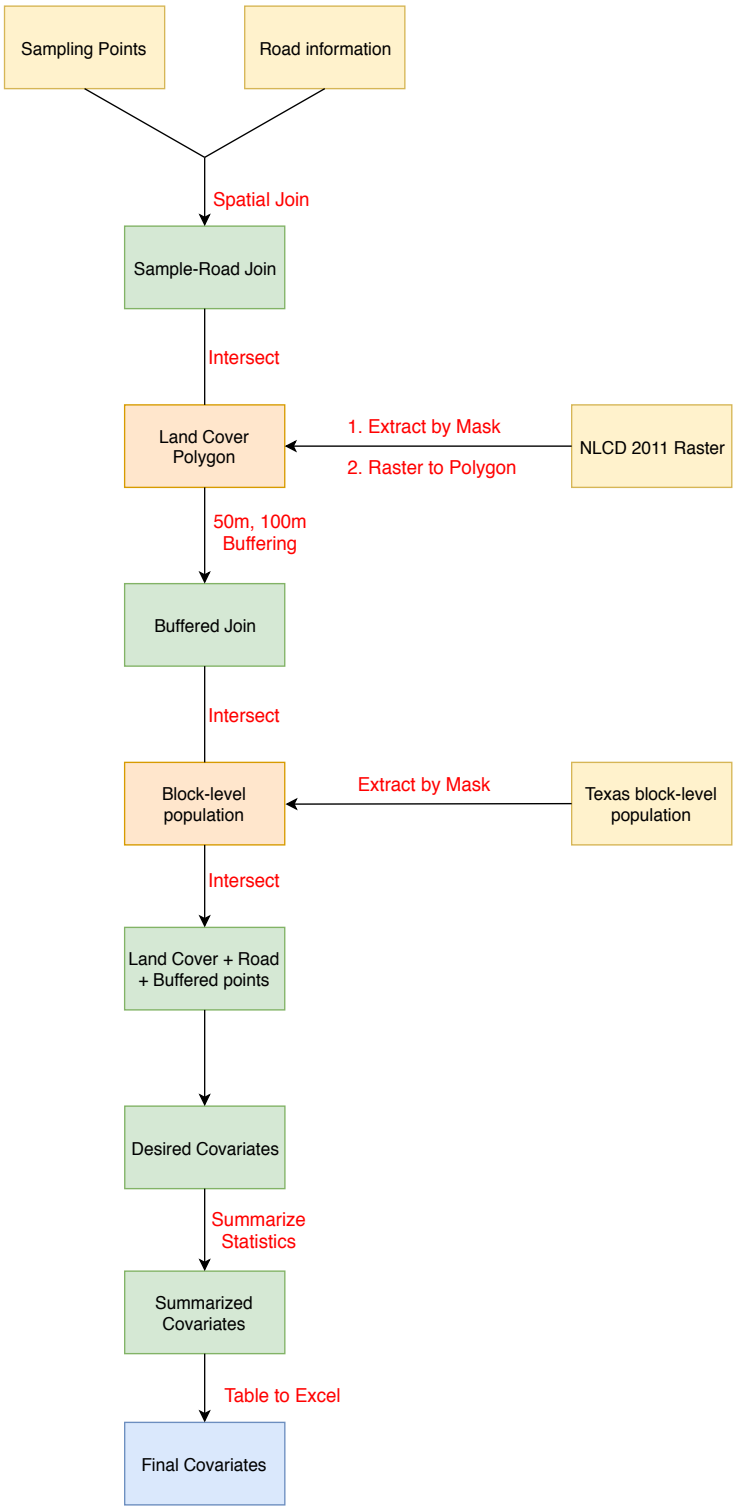

Figure 12. Schematic of steps

FID_gsv_roads	MEAN_CO2	MEAN_UFP	NLCD_DevHigh_50	NLCD_DevMed_50	NLCD_DevLow_50	NLCD_DevOpen_50	OSM_hwy	OSM_major	OSM_res	OSM_others	Pop_50
1	413.246	5666.000	0.267	0.549	0.153	0.031	0	1	0	0	3.513
2	420.808	7990.000	0.198	0.283	0.459	0.060	0	0	1	0	7.422
3	409.003	6613.333	0.195	0.651	0.154	0.000	0	1	0	0	1.768
4	418.133	8007.500	0.217	0.667	0.115	0.000	0	1	0	0	1.753
5	415.661	6290.000	0.339	0.533	0.128	0.000	0	1	0	0	1.730
6	427.990	7144.643	0.292	0.431	0.164	0.113	0	1	0	0	5.755
7	419.260	8112.000	0.056	0.229	0.628	0.087	0	0	1	0	7.449
8	409.880	5295.000	0.353	0.428	0.115	0.104	0	1	0	0	1.706
9	419.113	5316.667	0.254	0.321	0.138	0.288	0	1	0	0	1.681
10	410.660	5550.000	0.111	0.338	0.151	0.400	0	1	0	0	1.656
11	409.720	5770.000	0.024	0.431	0.216	0.329	0	1	0	0	1.632
12	409.720	5770.000	0.024	0.431	0.216	0.329	0	1	0	0	1.632
13	411.700	5760.000	0.000	0.439	0.236	0.325	0	1	0	0	1.617
14	411.700	5760.000	0.000	0.439	0.236	0.325	0	1	0	0	1.617
15	413.725	5815.000	0.011	0.409	0.229	0.352	0	1	0	0	1.615
16	418.354	5915.833	0.000	0.191	0.663	0.146	0	0	1	0	8.178
17	409.705	5820.000	0.001	0.376	0.260	0.363	0	1	0	0	1.663
18	407.055	5820.000	0.000	0.307	0.257	0.436	0	1	0	0	1.747
19	407.483	5845.000	0.000	0.133	0.311	0.556	0	1	0	0	1.884
20	429.728	7824.250	0.094	0.347	0.263	0.297	0	1	0	0	6.087

Figure 13. 50m buffered covariate table

FID_gsv_roads	MEAN_CO2	MEAN_UFP	NLCD_DevHigh_100	NLCD_DevMed_100	NLCD_DevLow_100	NLCD_DevOpen_100	NLCD_OpenWater_100	OSM_hwy	OSM_major	OSM_res	OSM_others
1	413.246	5666.000	0.218	0.333	0.189	0.141	0.120	0	1	0	0
2	420.808	7990.000	0.168	0.182	0.363	0.165	0.121	0	0	1	0
3	409.003	6613.333	0.249	0.355	0.130	0.134	0.132	0	1	0	0
4	418.133	8007.500	0.214	0.416	0.124	0.097	0.149	0	1	0	0
5	415.661	6290.000	0.237	0.401	0.120	0.098	0.143	0	1	0	0
6	427.990	7144.643	0.180	0.308	0.274	0.149	0.089	0	1	0	0
7	419.260	8112.000	0.093	0.170	0.384	0.237	0.116	0	0	1	0
8	409.880	5295.000	0.234	0.414	0.116	0.104	0.123	0	1	0	0
9	419.113	5316.667	0.238	0.420	0.108	0.120	0.088	0	1	0	0
10	410.660	5550.000	0.207	0.411	0.135	0.166	0.052	0	1	0	0
11	409.720	5770.000	0.166	0.398	0.160	0.226	0.018	0	1	0	0
12	409.720	5770.000	0.166	0.398	0.160	0.226	0.018	0	1	0	0
13	411.700	5760.000	0.111	0.408	0.151	0.297	0.000	0	1	0	0
14	411.700	5760.000	0.111	0.408	0.151	0.297	0.000	0	1	0	0
15	413.725	5815.000	0.073	0.392	0.167	0.346	0.000	0	1	0	0
16	418.354	5915.833	0.049	0.100	0.423	0.347	0.080	0	0	1	0
17	409.705	5820.000	0.068	0.316	0.142	0.456	0.000	0	1	0	0
18	407.055	5820.000	0.072	0.214	0.169	0.541	0.000	0	1	0	0
19	407.483	5845.000	0.044	0.191	0.201	0.564	0.000	0	1	0	0
20	429.728	7824.250	0.161	0.328	0.303	0.197	0.012	0	1	0	0

Figure 14. 100m buffered covariate table

The table shown above is the final output of all the GIS steps described earlier. It consists of the spatial covariates that are used to build the Land Use Regression Model for predicting concentrations. The 50m buffered LUR model is built in python using the Simple Linear Regression Algorithm which returns a cross-validation correlation coefficient (r^2) of 0.44 for CO_2 and 0.25 for UFP respectively. This implies that 44% and 25% of the variability in concentrations of CO_2 and UFP can be explained using the current model, respectively. The 100m buffered LUR model built using the same algorithm returns a cross-validation correlation coefficient of 0.43 for CO_2 and 0.23 for UFP. The lower values of r^2 for the larger buffer could be due to the fact that factors such as wind direction and speed, temperature, and humidity impact concentrations as they move away from the source, which is not accounted for in the current model. While this is not a very high correlation coefficient for the model, it is a starting point in trying to understand the factors influencing air quality in a growing city.

5. Conclusion

Land Use Regression models offer an opportunity to model concentrations in areas where it is difficult to setup long-term stationary sites. Based on short bursts of measurements through a location, these models attempt to replicate daily and even seasonal patterns in pollutant concentrations. Additionally, mobile monitoring allows sampling a large spatial domain with very high resolution, which is not possible in stationary sampling.

From the spatial trend maps shown above, the road category variable seems to have the maximum impact on the pollutant concentrations for the city of Austin. Consequently, this fact shows up in the LUR model building stage, where the maximum contribution to the correlation coefficient comes from the road network classification, with population and land cover contributing marginally. This gradient is most pronounced for CO_2, indicating that vehicular emissions is one of the primary sources for this pollutant. The lower correlation for UFPs shows that traffic, while significant, is not a primary source of ultrafines.

Lastly, the analysis helps to understand the exposure of people to higher concentrations. From the exposure maps shown above, we find that people do not tend to live where they are exposed to the highest concentrations, i.e., near major sources. While this preliminary comparison indicates that the sources and exposure points are not the same, further analysis could look at the economic divide, age distribution, or racial factors which could contribute to disproportionate exposure.

Abbreviations

The following abbreviations are used in this manuscript:

LUR	Land Use Regression
GSV	Google Street View
UFP	Ultrafine Particles
CO_2	Carbon Dioxide
NO	Nitric Oxide
NO_2	Nitrogen Dioxide
BC	Black Carbon
$PM_{2.5}$	Particulate Matter smaller than 2.5 μm
NLCD	National Land Cover Database
NDVI	Normalized Difference Vegetation Index
OSM	OpenStreetMap
UTM	Universal Transverse Mercator

References

1. Michael Brauer, Greg Freedman, Joseph Frostad, Aaron van Donkelaar, Randall V. Martin, Frank Dentener, Rita van Dingenen, Kara Estep, Heresh Amini, Joshua S. Apte, Kalpana Balakrishnan, Lars Barregard, David Broday, Valery Feigin, Santu Ghosh, Philip K. Hopke, Luke D. Knibbs, Yoshihiro Kokubo, Yang Liu, Stefan Ma, Lidia Morawska, José Luis Texcalac Sangrador, Gavin Shaddick, H. Ross Anderson, Theo Vos, Mohammad H. Forouzanfar, Richard T. Burnett, and Aaron Cohen. **Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013.** *Environmental Science & Technology* **2016**, *50* (1), 79-88, doi: 10.1021/acs.est.5b03709.
2. Joshua S. Apte, Julian D. Marshall, Aaron J. Cohen, and Michael Brauer. **Addressing Global Mortality from Ambient PM2.5.** *Environmental Science & Technology* **2018**, *49* (13), 8057-8066, doi: 10.1021/acs.est.5b01236.
3. Joshua S. Apte, Kyle P. Messier, Shahzad Gani, Michael Brauer, Thomas W. Kirchstetter, Melissa M. Lunden, Julian D. Marshall, Christopher J. Portier, Roel C.H. Vermeulen, and Steven P. Hamburg. **High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data.** *Environmental Science & Technology* **2017**, *51* (12), 6999-7008, doi: 10.1021/acs.est.7b00891.
4. Kyle P. Messier, Sarah E. Chambliss, Shahzad Gani, Ramon Alvarez, Michael Brauer, Jonathan J. Choi, Steven P. Hamburg, Jules Kerckhoffs, Brian LaFranchi, Melissa M. Lunden, Julian D. Marshall, Christopher J. Portier, Ananya Roy, Adam A. Szpiro, Roel C. H. Vermeulen, and Joshua S. Apte. **Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression.** *Environmental Science & Technology* **2018**, *52* (21), 12563-12572, doi: 10.1021/acs.est.8b03395.
5. Joshua S. Apte, Michael Brauer, Aaron J. Cohen, Majid Ezzati, and C. Arden Pope, III. **Ambient PM2.5 Reduces Global and Regional Life Expectancy.** *Environmental Science & Technology* **2018**, *5* (9), 546-551, doi: 10.1021/acs.estlett.8b00360.
6. United States Geological Survey. National Land Cover Database, **2011**. https://catalog.data.gov/dataset/usgs-national-land-cover-dataset-nlcd-downloadable-data-collection
7. United States Census Bureau. TIGER/Line® with Selected Demographic and Economic Data, **2010**. https://www.census.gov/geo/maps-data/data/tiger-data.html
8. OpenStreetMap. Texas, North America **2018**. https://download.geofabrik.de/north-america.html