Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Geocoding and spatio-temporal modeling of long-term $PM_{2.5}$ and NO_2 exposure: the Mexican Teachers' Cohort

Cervantes - Martínez Karla
National Institute of Public Health - Mexico https://orcid.org/0000-0001-5480-5750

Riojas - Rodríguez Horacio
National Institute of Public Health - Mexico

Díaz - Ávalos Carlos
National Autonomous University of Mexico

Moreno - Macías Hortensia
Metropolitan Autonomous University- Mexico

López - Ridaura Ruy
Ministry of Health - Mexico

Stern Dalia
CONACyT - National Institute of Public Health - Mexico

Acosta - Montes Jorge Octavio
Autonomous University of Chihuahua - Mexico

Texcalac - Sangrador José Luis (✉ jtexcalac@insp.mx)
National Institute of Public Health - Mexico

Research Article

Keywords: PM2.5, NO2, air pollution, Generalized Additive Models, exposure assessment, geocoding.

Posted Date: March 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-362282/v1

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Epidemiological studies on the effects of air pollution in Mexico often use the environmental concentrations of monitors closest to the home as exposure proxies, yet this approach disregards the space gradients of pollutants and assumes that individuals have no intra-city mobility. Our aim was to develop high-resolution spatial and temporal models for predicting long-term exposure to $PM_{2.5}$ and NO_2 in a population of ~ 16 500 participants from the Mexican Teachers' Cohort study. We geocoded the home and work addresses of participants. Using information from secondary sources on geographic and meteorological variables as well as other pollutants, we fitted two generalized additive models to predict monthly $PM_{2.5}$ and NO_2 concentrations in the 2004–2019 period. The models were evaluated through 10-fold cross validation. Both showed high predictive accuracy with out-of-sample data and no overfitting (CV RMSE = 0.102 for $PM_{2.5}$ and CV RMSE = 4.497 for NO_2). Participants were exposed to a monthly average of 24.38 (6.78) μg/m^3 of $PM_{2.5}$ and 28.21 (8.00) ppb of NO_2 during the study period. These models offer a solid alternative for estimating $PM_{2.5}$ and NO_2 exposure with high spatio-temporal resolution for epidemiological studies in the Valle de México region.

1. Introduction

In 2015, WHO recognized exposure to ambient air pollutants and its effects on human health as a priority public-health problem requiring further study (WHO, 2015). The rapid spread of urban spaces around the world has increased the number of individuals exposed to pollutant concentrations above the values recommended by the WHO (Riojas-Rodríguez et al., 2016). For instance, the MCMA is the third largest metropolitan area among the member states of the OECD and one of the three megacities in Latin America (OECD, 2016). Historically, this area has registered high rates of air pollution (INE-SEMARNAT, 2011) which have placed the health of its almost 21 million inhabitants permanently at risk (Instituto Nacional de Estadística y Geografía (México)., 2015).

Epidemiological studies have shown that chronic exposure to air $PM_{2.5}$ and NO_2 is associated with increased mortality (Beelen et al., 2014), cardiovascular disease (Dockery, 2001; Hoek et al., 2013), lung cancer (Chen et al., 2008; Hamra et al., 2014; Pope et al., 2002), cardiopulmonary conditions (Krewski et al., 2009) and diabetes (Li et al., 2014), among others. All these studies, which have used a variety of methods to estimate exposure to air pollutants, have developed deterministic and/or probabilistic models as the cornerstone of their analyses (Jerrett et al., 2005). In Mexico, the majority of epidemiological studies on the health effects of air pollutants have estimated exposure to pollutants using proximity methods such as those based on data from the nearest monitor (Barraza-Villarreal et al., 2008; Escamilla-Nuñez et al., 2008; Hernández-Cadena et al., 2009; Rojas-Martinez et al., 2007), city or municipal averages (Carbajal-Arroyo et al., 2011; Téllez-Rojo et al., 2000), IDW interpolation (Riojas-Rodríguez et al., 2014; Trejo-González et al., 2019). One disadvantage of these methodological approaches lies in their failure to capture the spatial variability of exposure caused by local sources, urban topography and local meteorological factors (Jerrett et al., 2005). The strength of these methods is therefore undermined by considerable uncertainty as regards the assignment of exposure. To the best of our knowledge, the literature offers only a handful of epidemiological studies that have employed more sophisticated, high-resolution methods (e.g., based on satellite data) for estimating exposure to ambient air pollutants in the Mexican population (Gutiérrez-Avila et al., 2018; Rosa et al., 2017; Téllez-Rojo et al., 2020). In addition, previous studies have documented that ignoring the daily mobility patterns of the study population can lead to erroneous exposure measurements, thus introducing bias in epidemiological analyses based on exposure estimates at the individual level (Nyhan et al., 2019; Setton et al., 2011). To offset these limitations, we geocoded the home and work addresses of the study population and fitted GAMs to predict $PM_{2.5}$ and NO_2 concentrations at both locations. The principal advantage of the GAMs resides in their use of semiparametric methods to model non-linear functions through penalized splines (Yanosky et al., 2008).

This work represents an improvement over the exposure models developed thus far for the MCMA (Just et al., 2015; Rivera-González et al., 2015; Son et al., 2018). In order to craft a feasible model that could be adapted to conditions of limited information in developing countries, we used the highest-quality and largest amount of information available from secondary sources on air pollutant predictors in Mexico. This study provides a novel methodological tool for the MTC study as well as for subsequent epidemiological studies seeking to assess the effects of $PM_{2.5}$ and NO_2 exposures on health. Our objective was to assign the outdoor $PM_{2.5}$ and NO_2 exposure of ~16 500 participants from an MTC living and working in the MCMA, on a monthly basis over the 2004-2019 period.

2. Material And Methods

We developed spatio-temporal prediction models for $PM_{2.5}$ and NO_2, using the pollutant concentrations at monitoring locations as our dependent variable, and meteorological and geographic variables as well as other pollutants as predictors. For statistical and geostatistical modeling, geoprocessing and spatial analysis, we used R version 3.5.3 software (R Core Team, 2019).

2.1 Study population and geocoding

The MTC study, an ongoing prospective initiative established in 2006 -2008, includes 115 314 female public-school teachers from 12 states in Mexico. A detailed description of the cohort has been published elsewhere (Lajous et al., 2015). Briefly, recruitment was performed in two phases: the first, in 2006, took place in two states (n=27 979), and the second, in 2008, incorporated 10 additional states (n=87 335) including CDMX and EdoMex. The participating teachers answered a reference questionnaire concerning sociodemographic characteristics, reproductive history, diet, lifestyle and health status. We continue to follow-up on the women every 3-4 years to obtain information on disease diagnoses and to update their risk factor profiles. In the first follow-up cycle (2011-2013) the response rate was 83%, and in the second follow-up cycle (2014-2020, ongoing) the response rate was 59%.

Considering the availability and quality of data required to geocode the homes of the teachers, we selected the MCMA as our study area. It was composed of 16 municipalities in CDMX, 59 in EdoMex and one municipality in Hidalgo (Figure 2). Based on the meteorological and air pollutant monitoring stations in place during the study (Figure 2), we delimited the boundaries of the area of analysis using the extreme coordinate position of a 5 km buffer around each monitoring station. Within these boundaries, we defined a grid composed of cells measuring 1 km x 1 km.

We geocoded the HA and WA of each teacher in the MTC. The HAs included ZC, municipalities and states reported by the teachers at baseline. Since information was unavailable on changes of HAs before and after follow-up, we assumed that they were permanent. Based on the criteria of SEPOMEX for valid ZCs, we selected only those containing five digits and verified that the first two corresponded to the codes of the home municipalities reported by the teachers. Finally, we used a geographic layer of SEPOMEX ZCs ("gob.mx," n.d.) to assign the coordinates of the ZC's centroid for each HA. To geocode the WAs, we used the WCs provided by the PNCM at baseline and follow-ups. The geographic coordinates were drawn from the SNIE ("Sistema Nacional de Información de Escuelas," n.d.).

2.2 Meteorological and air pollutant data

We obtained the databases of the hourly measurement of pollutants (PM_{10}, $PM_{2.5}$, SO_2, NO_2 and O_3) and meteorological variables (relative humidity, temperature and wind speed) at the monitoring sites, from 1:00 a.m., January 1st, 2004, to 12:00 a.m., July 31st, 2019. Data were drawn from the monitoring network of the SEDEMA in CDMX. These datasets are permanently available for open consultation on the official websites of both networks ("Dirección de Monitoreo Atmosférico," n.d.).

We calculated the daily estimators (daily averages) for each pollutant and meteorological variable at each monitoring site only when a minimum of 75% of the required observations (18 hours) were available. Otherwise, we entered "missing data." On this basis, we calculated the monthly estimators by averaging at least 23 daily estimators (75%). The seasons of the year, used as a categorical variable, were defined as follows: the cold-dry season, from November to February; the hot-dry season, from March to May; and the rainy season, from June to October.

2.3 Geographic data and other covariables

The geographic profiles of the monitoring sites were characterized using a GIS. We created a geographic layer of the air quality monitoring stations in place from 2004 to 2019 in the study area, based on coordinate data provided by the SINAICA ("INECC | Sistema Nacional de Información de la Calidad del Aire (SINAICA)," n.d.). The altitude variable was provided by the SINAICA in masl (meters above sea level) units and categorized into tertiles as follows: low ≤ 2 300 masl, medium ≥ 2 301 masl and ≤ 2 500 masl, and high ≥ 2 501 masl. The vehicle motorization index (the number of motorized vehicles per 1 000 inhabitants) was calculated annually, at the municipal level, based on population projections from the CONAPO ("Consejo Nacional de Población | Gobierno | gob.mx," n.d.) and the historical databases for registered motorized vehicles (automobiles, passenger buses, cargo trucks and motorcycles) in circulation, according to the INEGI (INEGI, n.d.). This variable was categorized into tertiles as follows: low ≤ 222 vehicles per 1 000 inhabitants, medium ≥ 223 vehicles per 1 000 inhabitants and ≤ 487 vehicles per 1 000 inhabitants, and high ≥ 488 vehicles per 1 000 inhabitants.

2.4 Statistical models

2.4.1. Generalized Additive Models

To ascertain the contribution of each predictor to $PM_{2.5}$ and NO_2 variability at the monitoring station level during the 2004-2019 period, we fitted two independent GAMs (Equation 1).

Equation 1. Generalized Additive Model formula

$$\boldsymbol{E(Y|X_1, X_2, \ldots, X_p) = \alpha_0 + \sum_{j=1}^{p} f_j\,(X_j)}$$

where *f_j(.)* are smooth functions, *Y* is the vector of outcomes and the *X_j* are the vectors of covariates. For each model, we used monthly $PM_{2.5}$ or NO_2 concentrations per monitoring site as the outcome. The vector $PM_{2.5}$ was log-transformed because of its skewed distribution. We only included in the model the covariables which we expected *a priori* to exert a physical influence on the $PM_{2.5}$ and NO_2 levels. The predictors introduced in the adjusted model were the group of other pollutants, the meteorological variables and the geographic variables. To take in to account the correlation between pollutants, we included them using interaction terms. With the view of ensuring a parsimonious model specification, we eliminated each term to assess its contribution and retained only those that improved predictive accuracy and offered statistical significance ($p < 0.05$). Model comparison and selection were performed according to DE, AIC, and RMSE. To test the goodness of fit, we performed a residual diagnosis, allowing us to verify their normality and the absence of any overdispersion patterns. Finally, we performed stepwise selection to test and select the right number of base functions for achieving a smoothed curve that would maximize the fit of our data and reduce the RMSE.

2.4.2. Model validation

To test for possible overfitting in the GAMs and evaluate their predictive accuracy, we carried out a 10-fold cross validation and determined the predictive accuracy of each model based on its CV RMSE.

2.4.3. Model predictions and geostatistical conditional simulation

We used the GAM coefficients to predict the monthly concentrations of $PM_{2.5}$ and NO_2 in all the 1x1-km grid cells during the January 2004-July 2019 period. For each grid cell and for each month of the study period, we carried out 3 000 simulation replicates under the GCS method (ESRI, n.d.). The object was to obtain the mean value of the predictor variables: PM_{10}, O_3, SO_2, NO_2, wind speed, relative humidity and temperature in each cell. As part of GCS, we fitted gaussian and exponential variograms with and without nugget effect to the empirical semivariograms in order to characterize the spatial structures at the unmeasured locations (Equation 2). To this end, we used the Cholesky decomposition of covariance matrix (Chilès and Delfiner, 2012):

Equation 2. Semivariogram formula

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(X_i) - Z(X_i + h)]^2$$

where the degree of spatial dependence among the values of Attribute *Z* in two different locations or points in space was semivariance *γ(h)*, and the distance between the points was lag *h*.

To take into account error ε_{ij} in the GAMs caused by the spatial correlation of the observations, we carried out the same simulation procedure with the residuals and then added them to the predicted values in each grid cell.

2.5 Assigning of cohort participant exposures

To maximize exposure's estimate precision, we considered the intra-city mobility of the teachers. For this purpose, we assigned the predicted monthly values of $PM_{2.5}$ and NO_2 in the grid cell where the HA and WA of each teacher were located each month during the study period. Under the assumption that estimates based on mobility (HWA_{weig}) were closer to actual exposure (individual monitoring), HA and WA exposures were weighted according to the number of hours spent in each address, assuming eight-hour workdays, which is the average working day in Mexico.

Equation 3. Weighting formula

$$HWA_{weig} = 2HA/3 + WA/3$$

To assess the relevance of considering the mobility factor in our study, we compared the HA and WA exposures through the difference between $HA - HWA_{weig}$. The null hypothesis of a difference equaling zero was proved using a mixed-effects model where the teacher was the grouping variable.

3. Results

3.1 Study population and geocoding

Of the 115 314 MTC participants at the national level, 22 743 had their HAs in CDMX or the EdoMex. We were able to geocode the ZC centroids of 95% of them (n=21 594), with 19 547 residing in the MCMA. In this area, the median of teachers per ZC was 23, ranging from 1 to 175 (Figure 2). The areas of the HA ZCs of participants oscillated between 0.005 km^2 and 17.003 km^2.

Of the 19 547 participants with geocoded HAs, we were able to identify and geocode the schools for 17 135 and 16 672 in the MCMA, at the beginning of follow-up (2008) and in the first follow-up cycle (2011-

2013), respectively. We thus geocoded the HA and WA in 2008 and the WA in 2011-2013 of 16 672 participants. Finally, the study population included 16 407 MTC participants who lived and worked in the area of analysis (Figure 1).

3.2 Descriptive statistics

The monthly averages of $PM_{2.5}$, NO_2 and their predictor variables at the monitoring site level are summarized in Table 1. During the 2004-2019 period, a maximum of 11 $PM_{2.5}$ and 17 NO_2 monitoring stations provided data on all the variables required for adjusting the GAMs. The geometric mean and SD of the monthly $PM_{2.5}$ concentrations were 3.18 and 0.27 μg/m^3, respectively. Meanwhile, the mean and SD of the monthly NO_2 concentrations were 28.34 and 8.59 ppb, respectively. Over the study period, the mean ambient temperature oscillated 17 ºC, wind speed averaged 1.85 m/s and humidity reached 53%. The lowest concentrations of both pollutants occurred in the rainy season and the highest in the cold-dry season. The $PM_{2.5}$ concentrations were greatest at the lowest altitudes.

3.3 Generalized Additive Model for PM_5 and predicted values

The $PM_{2.5}$ model predictors were altitude, relative humidity, wind speed, season and year, in addition to the PM_{10} and NO_2 pollutants. Only PM_{10}, relative humidity and wind speed were included as continuous variables with smoothing functions, using penalized splines of up to seven degrees of freedom (Figure 3). The maps in Figure 4 illustrate the average spatial distribution of the monthly $PM_{2.5}$ levels predicted by the model per year on the surface of the area of analysis (the grid). In general, the predicted levels were higher to the north of the MCMA and showed a downward trend until 2014. The map for 2019 considers only measurements until July; therefore, according to Table 1, the rainy season, which yielded the lowest $PM_{2.5}$ levels, was underrepresented. The percentage of deviance explained by the model was high (ED=87.3 %) for the 2004-2019 period (Table 1). The root mean square error in the sample indicated high predictive accuracy with a low level of error (RMSE = 0.100) and confirmed that no overfitting of data existed, as the value was very close to the cross-validation error (CV RMSE = 0.102).

3.4 The Generalized Additive Model for NO_2 and predicted values

The NO_2 model predictors were the vehicle motorization index, wind speed, temperature, season and year, in addition to the PM_{10}, O_3 and SO_2 pollutants. Only PM_{10}, O_3 and wind speed were included as continuous variables with smoothing functions using penalized splines of up to seven degrees of freedom, since the other continuous variables indicated a linear relationship with the dependent variable (Figure 3). The maps in Figure 4 show the average spatial distribution of the monthly levels of NO_2 predicted by the model on the surface of the study area per year. In general, the predicted levels of NO_2 were highest in the center of the MCMA and showed a downward trend until 2016, with a mean (SD) of 16.87 (7.77) ppb in the study area. As was the case with $PM_{2.5}$, only measurements until July 2019 were considered. The percentage of deviance explained by the model was intermediate (ED =74.4 %) for the

2004-2019 period (Table 1). The root mean square error in the sample indicated high predictive accuracy with a low level of error (RMSE = 4.406) and confirmed that no overfitting of data existed, as the value came very close to the cross-validation error (CV RMSE = 4.497).

3.5 Exposure assignment based on a mobility approach

The HAs of all the teachers were different from their WAs. This strengthens the case for considering mobility in estimates of exposure to air pollutants. For the entire period and population analyzed, the ranges of differences in estimated exposure at the homes and workplaces of participants were 0 - 6.01 µg/m^3 for $PM_{2.5}$ and 0 - 14.76 ppb for NO_2. Only 0.11% (n=1,799) of the teachers changed WAs between the first and second follow-up cycles. On average (SD), they traveled 6.285 (6.216) kilometers between their homes and their workplaces (min: 0.0098 km, Q1: 1.846 km, Q2: 4.282 km, Q3: 8.746 km, max: 56.155 km).

We obtained 187 weighted monthly averages of $PM_{2.5}$ and 187 of NO_2 assigned to each teacher as a proxy for long-term exposure. Figure 5 illustrates the annual distribution and principal statistics for the weighted monthly averages (HWA_{weig}) assigned to the 16 407 MTC participants during the 2004-2019 period. For both pollutants, we captured the expected exposure variability attributable to the spatial distribution of the pollutants and the location of the HAs and WAs in the study area. On average (SD), for the entire population and period analyzed, weighted exposures were 24.38 (6.78) µg/m^3 to $PM_{2.5}$ and 28.21 (8.00) ppb to NO_2.

The average exposures to both pollutants were lower at the HAs than at HWA_{weig} ($PM_{2.5}$: HA 24.37 µg/m^3, HWA_{weig} 24.39 µg/m^3; NO_2: HA 28.15 ppb, HWA_{weig} 28.21 ppb). The averages (min-max) of the differences were -0.016 (-2.00 - 1.91) for $PM_{2.5}$ and -0.063 (-4.92 - 4.60) for NO_2. In both cases, the differences between HA and HWA_{weig} did not equal zero and were statistically significant ($p<0.0001$).

4. Discussion And Conclusion

The models developed in our study provide a useful method for predicting monthly exposure to outdoor $PM_{2.5}$ and NO_2 at any location within a big area though the MCMA, with high spatial (1x1 km) and temporal (16 years: 2004–2019) resolution. These models delivered a solid performance and high predictive accuracy, with cross-validation errors slightly higher than those of the models (CV RMSE = 0.102 for $PM_{2.5}$ and CV RMSE = 4.497 for NO_2). A large difference between these values would have suggested that GAMs overfit data, as performance would be lower in the data not included in the adjustment during cross-validation. Because of their high resolution, these models satisfy the latent need of epidemiological studies to predict and assign exposures to ambient air pollutants in numerous MCMA locations with a high degree of precision. Furthermore, they cover a long-time span (16 years) on a small temporal scale (monthly), making it possible to assess the effects of medium- and long-term exposures.

To the best of our knowledge, only three published studies have proposed models for assigning exposure to ambient air pollutants in Mexico (Just et al., 2015; Rivera-González et al., 2015; Son et al., 2018); all of them are focused on Mexico City and/or the MCMA. Rivera et al. provide a comparison of four basic methods based on proximity and interpolation for all criteria pollutants (Rivera-González et al., 2015). The principal limitation of several of these approaches is that they do not capture the spatial variability of exposure (Jerrett et al., 2005). We overcome this limitation by using high spatial resolution, which captures variability to a great extent. Son et al. propose a mixed-effects regression model based on land use for all criteria pollutants. Although they follow the traditional assumption that the relationship between the predictors and the dependent variable is linear, these authors are innovative in that they explore the LASSO method as a statistical technique for fitting models. In addition, they analyze various spatio-temporal scales (hourly, daily, monthly, semi-annual and annual) (Son et al., 2018). As opposed to their model, we propose GAMs, which maximize prediction quality for the dependent variable considering its non-linear relationship with the predictors and incorporate smoothed -not parametric- terms using penalized splines. However, we only explored a monthly exposure scale for two criteria pollutants. Finally, Just et al. developed a more sophisticated mixed-effects regression model, including AOD satellite measurements as their primary $PM_{2.5}$ predictor, over a period of 10 years (Just et al., 2015). While innovative, this method is limited mainly in that it is inapplicable in areas where little or no $PM_{2.5}$ monitoring data exist for its validation, and no satellite measurements can be taken on cloudy days (Bravo et al., 2012; Paciorek and Liu, 2009). We overcame the limitation related to unavailable entry data, an obstacle faced by many developing countries, by using information on predictors from secondary sources that are easy to access and interpret and are permanently monitored. Furthermore, we covered a longer time span (16 years) and modeled NO_2 exposure. Notwithstanding the differences with all the previous studies, in general, we obtained similar results. The predicted spatial patterns for $PM_{2.5}$ were comparable with theirs and, on average, the predicted concentrations for both pollutants ($PM_{2.5}$ and NO_2) were only slightly lower.

The results of our analyses demonstrate the relevance of using meteorological variables and other pollutants as the principal $PM_{2.5}$ y NO_2 predictors. This could indicate that, even with secondary monitoring data from local networks, it is possible to generate highly predictive models. We therefore confirm that, in spite of limited information conditions, it is possible to obtain feasible methods for predicting outdoor exposures to air pollutants. However, it is important to include geographic predictors given the nature of the contaminants of interest and the geographic characteristics of the study area. In the two models we developed, the geographic predictors indicated smaller coefficients than the other predictors. In addition to altitude and the vehicle motorization index, we generated and tested other geographic variables. The linear meters of the different types of avenues in buffers of 100, 300 and 500 meters around the monitoring stations did not prove significant and therefore were not included in the final models. We tested the UV-B radiation variable and found a significant relationship with NO_2; however, we excluded it from the final model because we were unable to perform a semivariographic analysis owing to insufficient monitoring sites in the study area (only 0–6 stations during the 2004–2019 period). This caused an increase of 0.583 units of CV RMSE in the NO_2 model (CV RMSE including UV-B = 3.914).

A recent study which analyzed and compared the individual home-work mobility patterns of ~ 400 000 individuals corroborated that ignoring mobility causes an erroneous classification when estimating health effects (Nyhan et al., 2019). The principal strength of our study was lay in considering home-work mobility to estimate $PM_{2.5}$ and NO_2 exposures. It is expected that, based on the findings of Nyhan et al. (Nyhan et al., 2019), this will help reduce measurement errors that could lead the association measurements of subsequent epidemiological studies to a null value. Another strength of our study concerned the method adopted to ascertain the spatial distribution of the predictor variables at unmeasured sites. Although no existing method can offer an exact estimate of an unknown reality, GCS considers the dispersion of the phenomenon analyzed and allows for obtaining one of the possible executions of a random function (surface). The simulated variables thus present the same variability and correlation characteristics as the observed data (measured according to their mean, their variance and the semivariogram) (Deutsch and Journel, 1998; Goovaerts, 1997). Among other available methods, we could have chosen IDW interpolation or Kriging; however, despite their extensive use in earth and environmental sciences, they provide only a smooth image of reality. GCS assumptions include seasonality and a normal distribution of data (Deutsch and Journel, 1998; Goovaerts, 1997).

We are conscious that the precision and measurement error inherent in predictive models such as the one we developed depends not only on mobility, but also on the residential history of the participants and a precise geocoding of the addresses. Prior studies have suggested that the geocoding technique can influence health effect estimates when using fine-scale exposure models (Jacquemin et al., 2013). We therefore recognize two sources of error. First, we assumed that the HAs of cohort participants were permanent throughout follow-ups. Nonetheless, Pimienta et al. (Pimienta Lastra and Toscana Aparicio, 2019) registered a low level of inter-municipal migration within the MCMA (4.8%) during the 2010–2015 period. This could suggest that the study population was unlikely to change addresses, a hypothesis similar to our assumption. Second, we geocoded the centroids of the ZCs instead of the exact location of each HA. As the ZCs were spread across a wide-ranging area (0.005–17.003 km^2), it is expected that the estimated exposure was more precise for participants residing in the smaller ZC areas. Finally, we assume that the long-term outdoor exposure assigned constitutes a reasonable representation of exposure in interior spaces and microenvironments.

Our models were subject to various limitations. First, the delimitation of the study area and population was restricted by the quality of data (the geographic layer of ZCs) obtained from secondary sources of information to geocode the HAs of the MTC participants. Although the cohort was extended across 12 Mexican states, we were able to partially include only two of them in this study. Second, given the suboptimal quality of the data, we did not include information to characterize the sources of pollution attributable to land use and vehicular traffic emissions. Nonetheless, as a proxy for the latter, we used a vehicle motorization index created with official data from secondary sources regarding registered vehicles at the municipal level. This variable proved significant in the NO_2 but not the $PM_{2.5}$ model. This is consistent with our theoretical framework, given that NO_2 is generated primarily by the combustion of vehicles.

It is crucial for environmental epidemiology to rely on methods created and refined to estimate exposure to ambient air pollutants, which transcend the limitation of unavailable entry data, particularly in developing countries. In conclusion, our results suggest that the models developed in this study offer a solid alternative that can be used in epidemiological studies of the MCMA to predict $PM_{2.5}$ and NO_2 exposure with high spatio-temporal resolution for the 2004–2019 period. Beyond the MTC context, these models can be replicated in other cities with similar sources of information. Our results support the relevance of developing exposure models with high predictive accuracy based on available secondary data from official sources of information. Future research efforts should therefore focus on expanding the models developed thus far to other Mexican cities.

Declarations

ACKNOWLEDGEMENTS

This study was conducted with support from the national scholarship program of the Mexican Postgraduate Quality Program run by CONACYT. The support was granted to the first author during her doctoral studies.

CONFLICT OF INTEREST

The authors declare having no conflict of interest.

References

Barraza-Villarreal, A., Sunyer, J., Hernandez-Cadena, L., Escamilla-Nuñez, M.C., Sienra-Monge, J.J. et al., 2008. Air pollution, airway inflammation, and lung function in a cohort study of Mexico City schoolchildren. Environ. Health. Perspect. 116, 832-838.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G. et al., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. Lancet 383, 785-795.

Bravo, M.A., Fuentes, M., Zhang, Y., Burr, M.J., Bell, M.L., 2012. Comparison of exposure estimation methods for air pollutants: Ambient monitoring data and regional air quality simulation. Environ. Res. 116, 1-10.

Carbajal-Arroyo, L., Miranda-Soberanis, V., Medina-Ramón, M., Rojas-Bracho, L., Tzintzun, G. et al., 2011. Effect of PM(10) and O(3) on infant mortality among residents in the Mexico City Metropolitan Area: a case-crossover analysis, 1997-2005. J. Epidemiol. Community. Health. 65, 715-721.

Chen, H., Goldberg, M.S., Villeneuve, P.J., 2008. A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases. Rev. Environ. Health. 23, 243-297.

Chilès, J.-P., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty, Edición: 2. ed. Wiley, Hoboken, N.J.

Consejo Nacional de Población | Gobierno | gob.mx. URL https://www.gob.mx/conapo (accessed 12.4.19).

Deutsch, C.V., Journel, A.G., 1998. GSLIB: geostatistical software library and user's guide, Version 2.0, 2nd ed. ed. Oxford University Press, New York.

Dirección de Monitoreo Atmosférico. URL http://www.aire.cdmx.gob.mx/default.php (accessed 5.8.19).

Dockery, D.W., 2001. Epidemiologic evidence of cardiovascular effects of particulate air pollution. Environ. Health. Perspect. 4, 483-486.

Escamilla-Nuñez, M.-C., Barraza-Villarreal, A., Hernandez-Cadena, L., Moreno-Macias, H. et al., 2008. Traffic-related air pollution and respiratory symptoms among asthmatic children, resident in Mexico City: the EVA cohort study. Respir. Res. 9, 74.

ESRI, n.d. Key concepts of geostatistical simulation. ArcGIS. URL https://desktop.arcgis.com/es/arcmap/latest/extensions/geostatistical-analyst/key-concepts-of-geostatistical-simulation.htm (accessed 1.7.16). URL http://www.gob.mx/ (accessed 7.4.17).

Goovaerts, P., 1997. Geostatistics for natural resources evaluation, Applied geostatistics series. Oxford University Press, New York.

Gutiérrez-Avila, I., Rojas-Bracho, L., Riojas-Rodríguez, H., Kloog, I., Just, A.C., Rothenberg, S.J., 2018. Cardiovascular and cerebrovascular mortality associated with acute exposure to PM2.5 in Mexico City. Stroke. 49, 1734-1736.

Hamra, G.B., Guha, N., Cohen, A., Laden, F., Raaschou-Nielsen, O. et al., 2014. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. Environ. Health. Perspect. https://doi.org/10.1289/ehp.1408092

Hernández-Cadena, L., Holguin, F., Barraza-Villarreal, A., Del Río-Navarro, B.E., Sienra-Monge. et al., 2009. Increased levels of outdoor air pollutants are associated with reduced bronchodilation in children with asthma. Chest. 136, 1529-1536.

Hoek, G., Krishnan, R.M., Beelen, R., Peters, A., Ostro, B. et al., 2013. Long-term air pollution exposure and cardio- respiratory mortality: a review. Environ. Health. Glob. Access. Sci Source. 12, 43.

INECC | Sistema Nacional de Información de la Calidad del Aire (SINAICA). INECC. URL http://sinaica.inecc.gob.mx (accessed 6.11.13).

INEGI, I.N. de E. y G., n.d. Instituto Nacional de Estadística y Geografía. INEGI. URL https://www.inegi.org.mx/ (accessed 12.4.19).

INE-SEMARNAT, 2011. Cuarto almanaque de datos y tendencias de la calidad del aire en 20 ciudades mexicanas (2000-2009)., Primera. ed. Instituto Nacional de Ecología, México, D.F.

Instituto Nacional de Estadística y Geografía (México)., 2015. México - Población - Encuestas, 2015. https://www.inegi.org.mx/programas/intercensal/2015/default.html#Documentacion

Jacquemin, B., Lepeule, J., Boudier, A., Arnould, C., Benmerad. et al., 2013. Impact of geocoding methods on associations between long-term exposure to urban air pollution and lung function. Environ. Health. Perspect. 121, 1054-1060.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou. et al., 2005. A review and evaluation of intraurban air pollution exposure models. J. Expo. Anal. Environ. Epidemiol. 15, 185-204.

Just, A.C., Wright, R.O., Schwartz, J., Coull, B.A., Baccarelli, A.A. et al., 2015. Using high-resolution satellite aerosol optical depth to estimate daily PM $_{2.5}$ geographical distribution in Mexico City. Environ. Sci. Technol. 49, 8576-8584.

Krewski, D., Jerrett, M., Burnett, R.T., Ma, R., Hughes, E. et al., 2009. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. Res. Rep. Health. Eff. Inst. 5-136.

Lajous, M., Ortiz-Panozo, E., Monge, A., Santoyo-Vistrain, R., García-Anaya, A. et al., 2015. Cohort profile: the mexican teacher's cohort (MTC). Int. J. Epidemiol. https://doi.org/10.1093/ije/dyv123

Li, C., Fang, D., Xu, D., Wang, B., Zhao, S. et al., 2014. Main air pollutants and diabetes-associated mortality: a systematic review and meta-analysis. Eur. J. Endocrinol. Eur. Fed. Endocr. Soc. 171, R183-190.

Nyhan, M.M., Kloog, I., Britter, R., Ratti, C., Koutrakis, P., 2019. Quantifying population exposure to air pollution using individual mobility patterns inferred from mobile phone data. J. Expo. Sc.i Environ. Epidemiol. 29, 238-247.

OECD, 2016. The Economic Consequences of Outdoor Air Pollution. Org. for Economic Cooperation & Development, Paris.

Paciorek, C.J., Liu, Y., 2009. Limitations of remotely sensed aerosol as a spatial proxy for fine particulate matter. Environ. Health. Perspect. 117, 904-909.

Pimienta-Lastra, R., Toscana-Aparicio, A., 2019. Migración intermunicipal permanente de la Zona Metropolitana del Valle de México 2010-2015. Cienc. Sum. 26, 1-19.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D.et al., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA J. Am. Med. Assoc. 287, 1132-1141.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Riojas-Rodríguez, H., Álamo-Hernández, U., Texcalac-Sangrador, J.L., Romieu, I., 2014. Health impact assessment of decreases in PM10 and ozone concentrations in the Mexico City Metropolitan Area: A basis for a new air quality management program. Salud. Públ. Méx. 56, 579-591.

Riojas-Rodríguez, H., da Silva, A.S., Texcalac-Sangrador, J.L., Moreno-Banda, G.L., 2016. Air pollution management and control in Latin America and the Caribbean: implications for climate change. Rev. Panam. Salud. Públ. Pan. Am. J. Public. Health. 40, 150-159.

Rivera-González, L.O., Zhang, Z., Sánchez, B.N., Zhang, K., Brown. et al., 2015. An assessment of air pollutant exposure methods in Mexico City, Mexico. J. Air Waste. Manag. Assoc. 65, 581-591.

Rojas-Martinez, R., Perez-Padilla, R., Olaiz-Fernandez, G., Mendoza-Alvarado, L., Moreno-Macias. et al., 2007. Lung function growth in children with long-term exposure to air pollutants in Mexico City. Am. J. Respir. Crit. Care. Med. 176, 377-384.

Rosa, M.J., Just, A.C., Guerra, M.S., Kloog, I., Hsu, H.-H.L. et al., 2017. Identifying sensitive windows for prenatal particulate air pollution exposure and mitochondrial DNA content in cord blood. Environ. Int. 98, 198-203.

Setton, E., Marshall, J.D., Brauer, M., Lundquist, K.R., Hystad, P. et al., 2011. The impact of daily mobility on exposure to traffic-related air pollution and health effect estimates. J. Expo. Sci. Environ. Epidemiol. 21, 42-48.

Sistema Nacional de Información de Escuelas URL http://snie.sep.gob.mx/SNIESC/ (accessed 12.4.19).

Son, Y., Osornio-Vargas, Á.R., O'Neill, M.S., Hystad, P., Texcalac-Sangrador, J.L. et al., 2018. Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. Sci. Total. Environ. 639, 40-48.

Téllez-Rojo, M.M., Romieu, I., Ruiz-Velasco, S., Lezana, M.A., Hernández-Avila, M.M., 2000. Daily respiratory mortality and PM10 pollution in Mexico City: importance of considering place of death. Eur. Respir. J. 16, 391-396.

Téllez-Rojo, M.M., Rothenberg, S.J., Texcalac-Sangrador, J.L., Just, A.C., Kloog, I. et al., 2020. Children's acute respiratory symptoms associated with PM2.5 estimates in two sequential representative surveys from the Mexico City Metropolitan Area. Environ. Res. 180, 108868.

Trejo-González, A.G., Riojas-Rodriguez, H., Texcalac-Sangrador, J.L., Guerrero-López, C.M. et al., 2019. Quantifying health impacts and economic costs of PM2.5 exposure in Mexican cities of the National Urban System. Int. J. Public. Health. https://doi.org/10.1007/s00038-019-01216-1

WHO, 2015. Health and the environment: addressing the health impact of air pollution (68 World Health Assembly, Report by the Secretariat No. A68/18). World Health Organization, Geneva, Switzerland.

Yanosky, J.D., Paciorek, C.J., Schwartz, J., Laden, F., Puett, R., Suh, H.H., 2008. Spatio-temporal modeling of chronic PM10 exposure for the Nurses? Health Study. Atmos. Environ. 42, 4047-4062.

Table

Variable	Time-varying	Model	
		$PM_{2.5}$ mean (SD) n = 779	NO_2 mean (SD) n = 1 298
Monitoring stations included 2004-2019 period (min - max)		3 - 11	6 - 17
$PM_{2.5}$ (mg/m^3)	Yes	25.05 (6.90)	---
NO_2 (ppb)	Yes	29.83 (8.02)	28.34 (8.59)
O_3 (ppb)	Yes	---	27.95 (7.14)
PM_{10} (mg/m^3)	Yes	49.67 (17.15)	50.41 (18.66)
SO_2 (ppb)	Yes	---	6.93 (4.23)
Temperature (ºC)	Yes	---	16.71 (2.18)
Wind speed (meters/second)	Yes	1.90 (0.50)	1.81 (0.47)
Relative humidity (%)	Yes	52.54 (11.40)	---
Season [a]	No		
Rainy (ref)		19.73 (3.90)	23.42 (6.35)
Cold-dry		29.12 (6.30)	33.13 (8.62)
Hot-dry		27.90 (5.88)	28.83 (7.39)
Motorization index (vehicles / 1 000 ha) [a]	Yes		
High (ref)		---	27.60 (7.47)
Medium		---	29.35 (8.77)
Low		---	27.52 (9.68)
Altitude (meters) [a]	Yes		
High (ref)		19.80 (4.53)	---
Medium		24.96 (6.10)	---
Low		26.36 (7.20)	---
Year [a]			
2004 (ref)		28.32 (5.75)	33.97 (7.15)
2005		29.26 (9.94)	34.02 (8.86)
2006		29.09 (6.41)	33.14 (7.22)
2007		28.16 (4.92)	33.04 (7.64)
2008		26.80 (7.10)	31.73 (7.17)
2009		24.07 (5.88)	30.70 (7.25)
2010		23.63 (8.01)	32.14 (8.49)
2011		24.87 (7.95)	28.06 (8.88)
2012		23.39 (4.49)	27.78 (8.77)
2013		25.33 (7.47)	25.98 (7.96)
2014		23.22 (6.14)	26.06 (7.47)
2015		24.96 (5.75)	23.92 (7.21)
2016		24.39 (6.97)	24.80 (7.66)
2017		24.66 (7.75)	25.74 (8.25)
2018		24.05 (5.91)	25.72 (7.30)
2019		24.15 (6.37)	22.20 (6.67)
Statistics		Model	
		$PM_{2.5}$	NO_2
DE		87.3 %	74.4 %
RMSE		0.100	4.406
CV RMSE		0.102	4.497

Table 1. Descriptive statistics for final GAM predictor variables, as well as bias and precision statistics from cross-validation (CV)

Note: ^a variable included as categorical; however, for descriptive purposes, their mean (SD) is presented for the dependent variable in each category

Abbreviations: DE = deviance explained, RMSE = root mean square error, and CV = cross validation

Figures

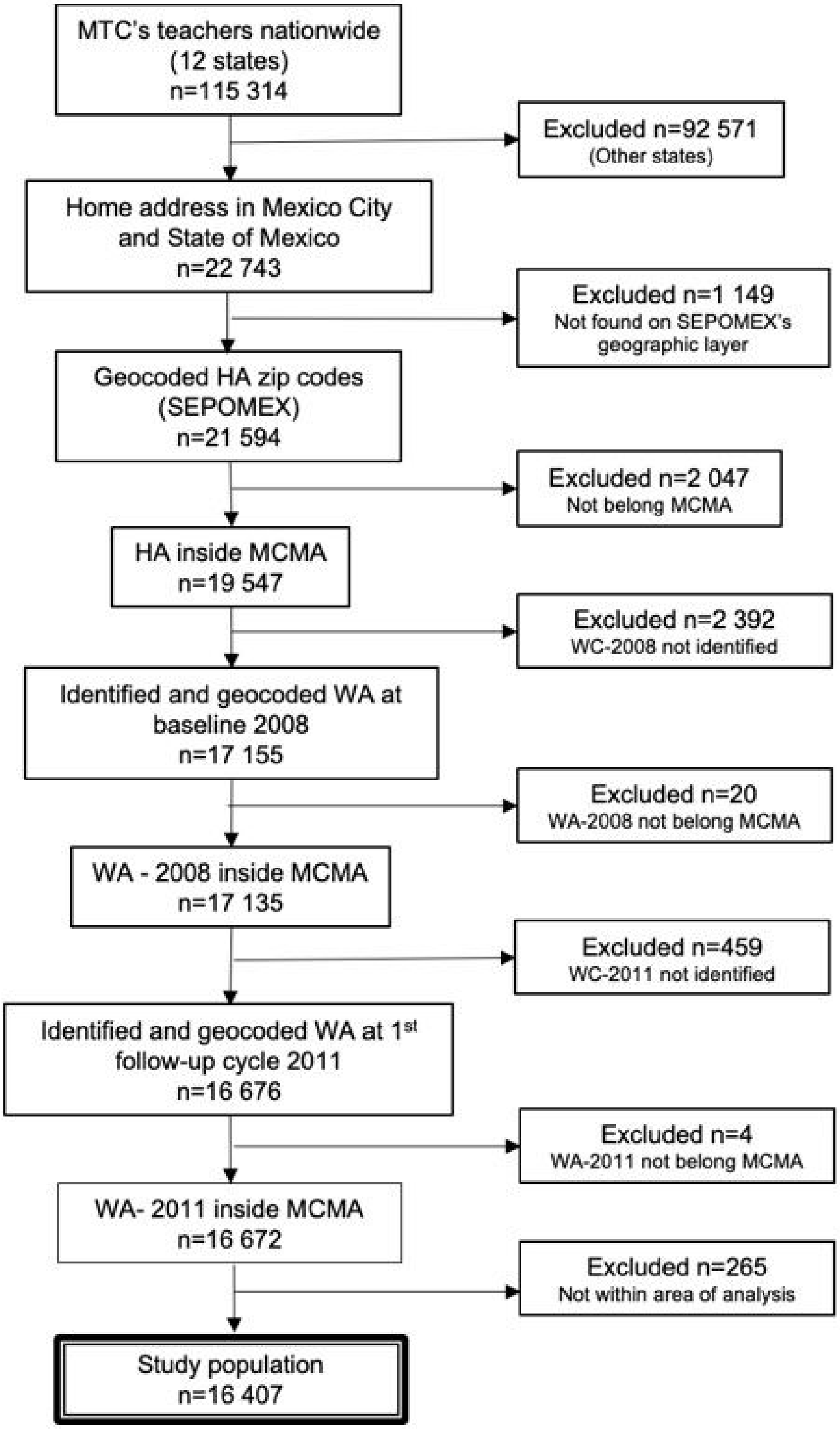

MTC's teachers nationwide (12 states) n=115 314
Excluded n=92 571 (Other states)
Home address in Mexico City and State of Mexico n=22 743
Excluded n=1 149 Not found on SEPOMEX's geographic layer
Geocoded HA zip codes (SEPOMEX) n=21 594
Excluded n=2 047 Not belong MCMA
HA inside MCMA n=19 547
Excluded n=2 392 WC-2008 not identified
Identified and geocoded WA at baseline 2008 n=17 155
Excluded n=20 WA-2008 not belong MCMA
WA - 2008 inside MCMA n=17 135
Excluded n=459 WC-2011 not identified
Identified and geocoded WA at 1st follow-up cycle 2011 n=16 676
Excluded n=4 WA-2011 not belong MCMA
WA- 2011 inside MCMA n=16 672
Excluded n=265 Not within area of analysis
Study population n=16 407

Figure 1

Flow diagram of geocoding process and identification of study population. Abbreviations: MTC = Mexican Teachers’ Cohort, MCMA = Mexico City Metropolitan Area, SEPOMEX = Mexican Postal Service, HA = Home Address, WA = Work Address and WC = workplace code

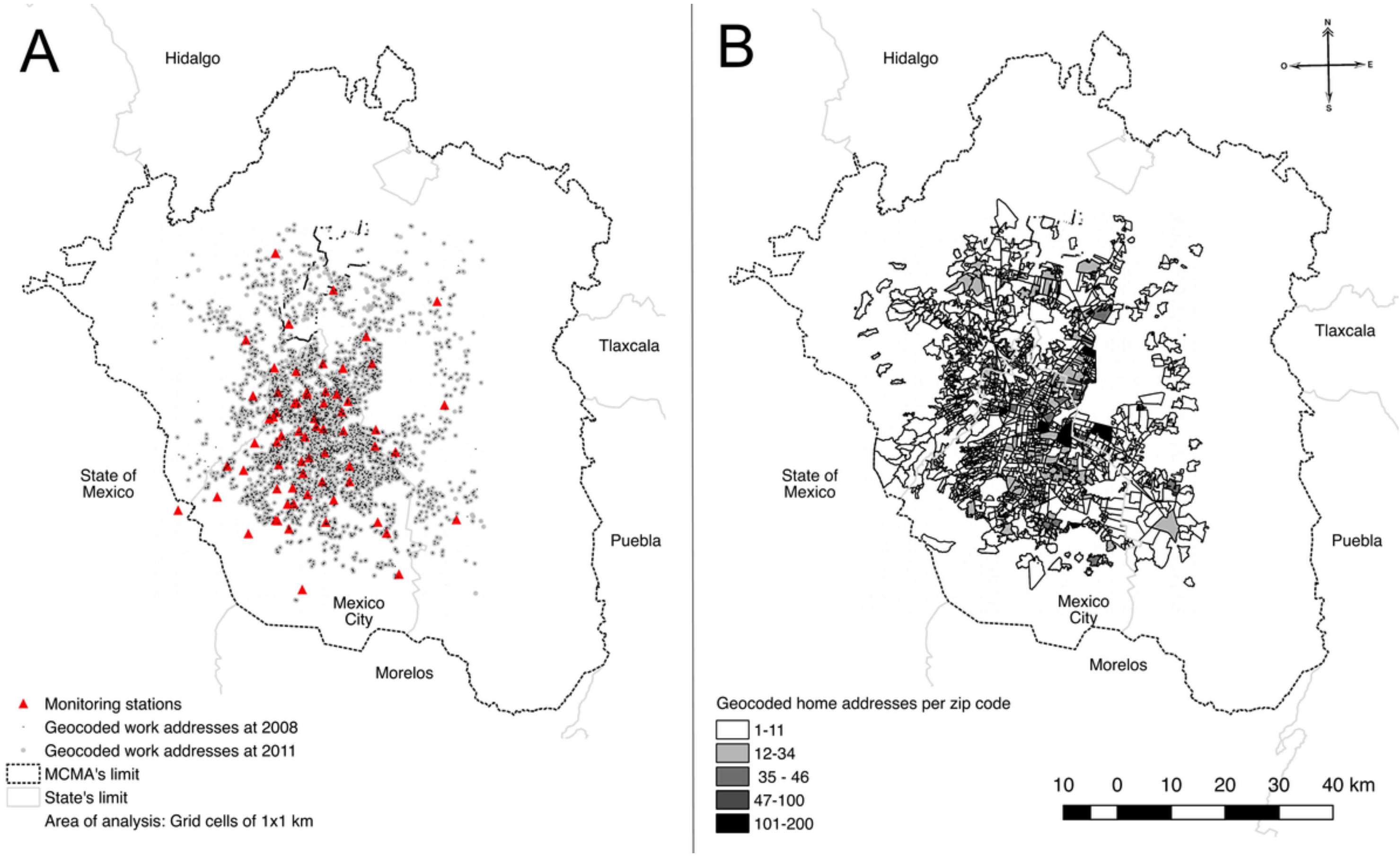


Figure 2

Area of analysis and geocoded addresses. A: location of work addresses at baseline (2008) and first follow-up cycle (2011-2013). B: location of home addresses and number of teachers per zip code.

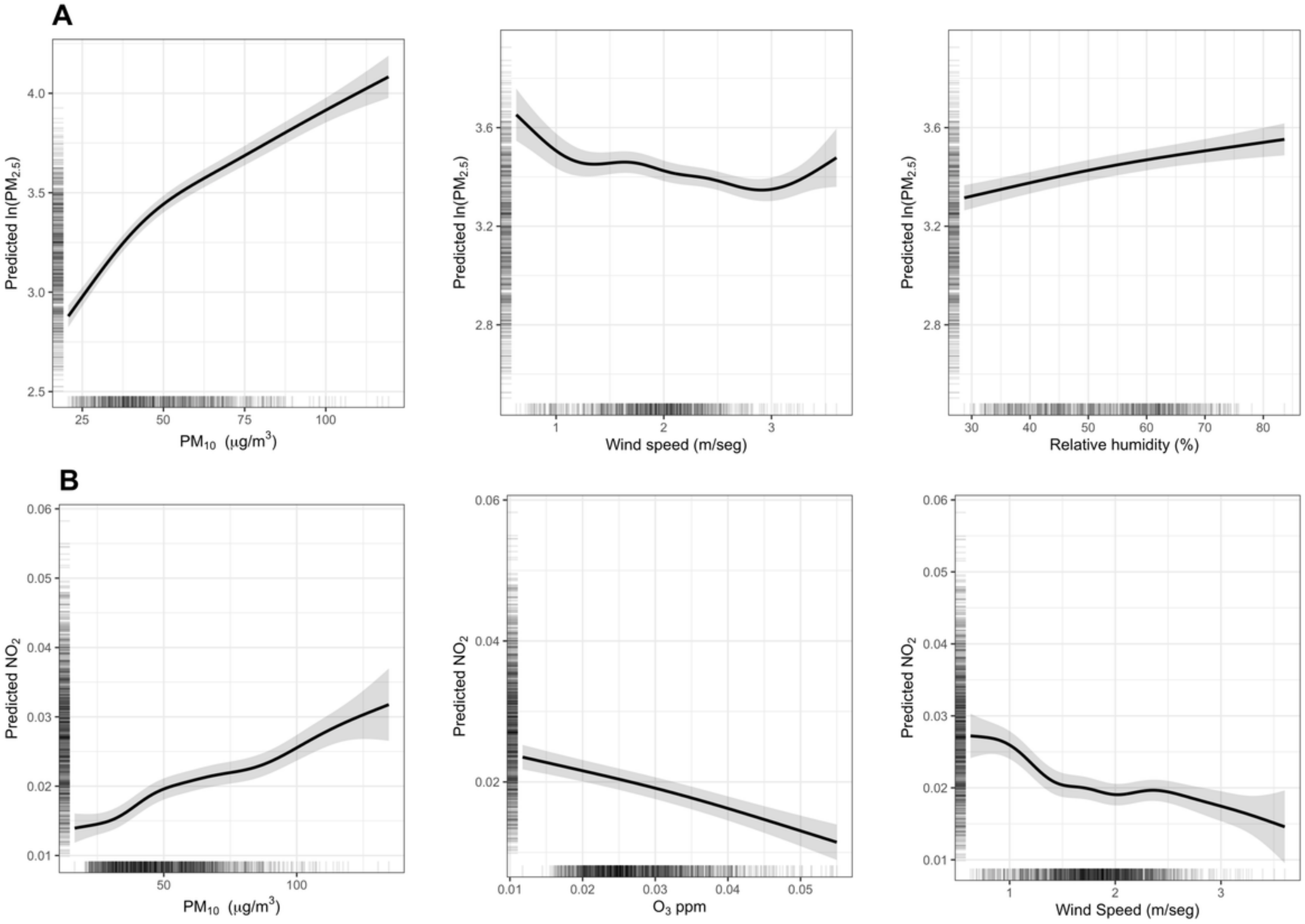


Figure 3

Smooth plots from the final models showing non-linear effects of selected predictors and associated predicted PM2.5 (uptrend) and NO2 (downtrend) concentrations. Note: CI at two standard errors above and below the mean

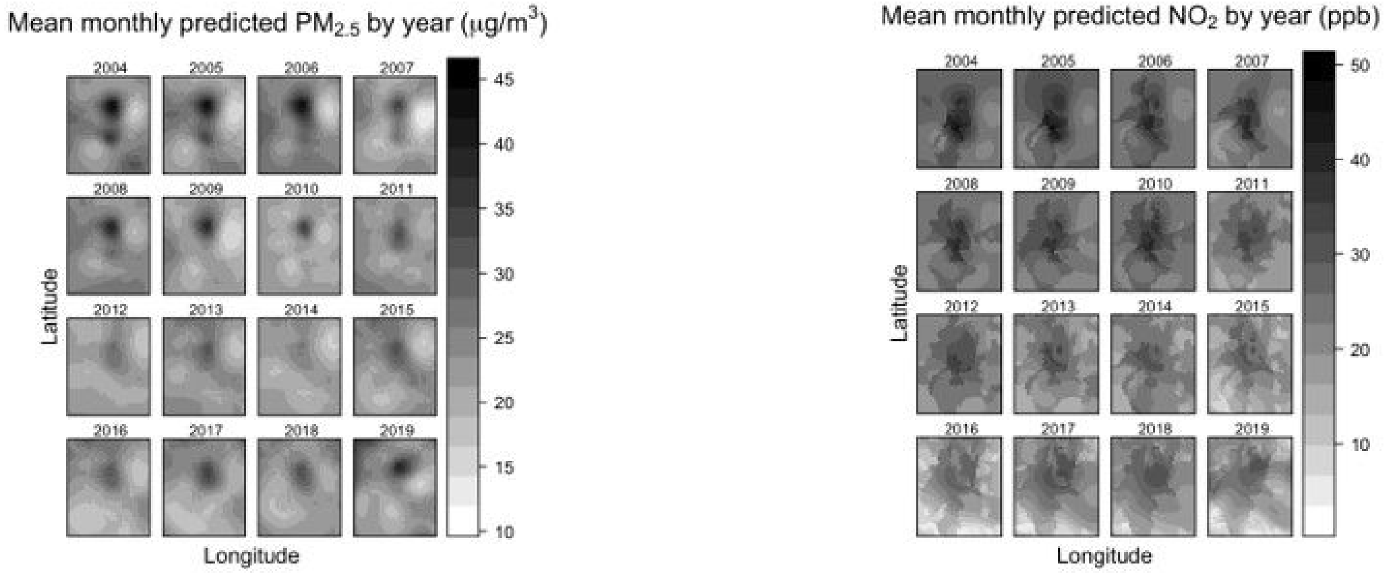
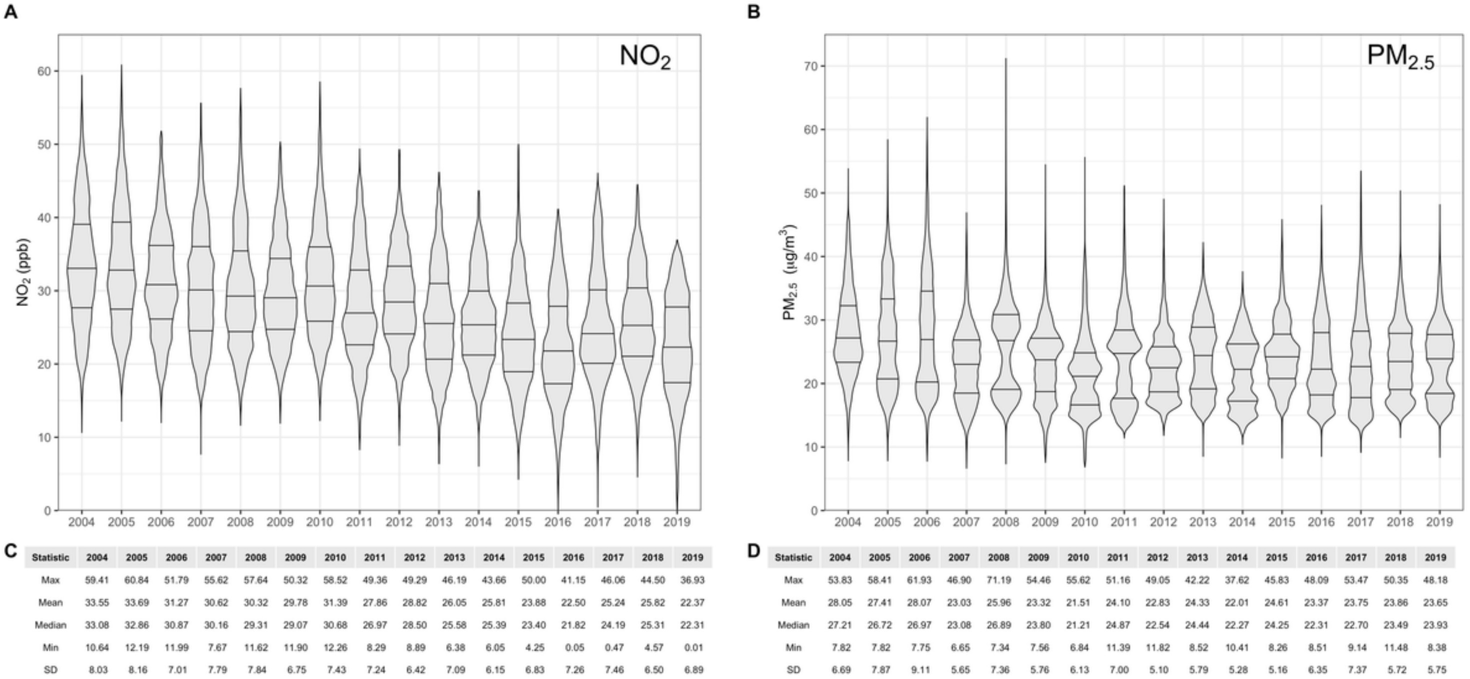


Figure 4

Raster of the annual average of the monthly PM2.5 and NO2 predicted concentrations on the surface of the area of analysis (grid made up of 1x1-km cells) 2004-2019



Statistic	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Max	59.41	60.84	51.79	55.62	57.64	50.32	58.52	49.36	49.29	46.19	43.66	50.00	41.15	46.06	44.50	36.93
Mean	33.55	33.69	31.27	30.62	30.32	29.78	31.39	27.86	28.82	26.05	25.81	23.88	22.50	25.24	25.82	22.37
Median	33.08	32.86	30.87	30.16	29.31	29.07	30.68	26.97	28.50	25.58	25.39	23.40	21.82	24.19	25.31	22.31
Min	10.64	12.19	11.99	7.67	11.62	11.90	12.26	8.29	8.89	6.38	6.05	4.25	0.05	0.47	4.57	0.01
SD	8.03	8.16	7.01	7.79	7.84	6.75	7.43	7.24	6.42	7.09	6.15	6.83	7.26	7.46	6.50	6.89

Statistic	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Max	53.83	58.41	61.93	46.90	71.19	54.46	55.62	51.16	49.05	42.22	37.62	45.83	48.09	53.47	50.35	48.18
Mean	28.05	27.41	28.07	23.03	25.96	23.32	21.51	24.10	22.83	24.33	22.01	24.61	23.37	23.75	23.86	23.65
Median	27.21	26.72	26.97	23.08	26.89	23.80	21.21	24.87	22.54	24.44	22.27	24.25	22.31	22.70	23.49	23.93
Min	7.82	7.82	7.75	6.65	7.34	7.56	6.84	11.39	11.82	8.52	10.41	8.26	8.51	9.14	11.48	8.38
SD	6.69	7.87	9.11	5.65	7.36	5.76	6.13	7.00	5.10	5.79	5.28	5.16	6.35	7.37	5.72	5.75

Figure 5

Violin plot and descriptive statistics for assignment of monthly PM2.5 (B and D) and NO2 (A and C) weighted exposures (HWAweig) by year. Note: First, second and third quartiles estimated for n=16,407 teachers.