Regression Model Generation on European NO_2 (Nitrogen Dioxide) Database

Jamal Taghavimehr

University of British Columbia

Abstract

Land Use Regression (LUR) models have multivariate nature that encompasses the pollutant concentrations as their dependent variable and predictor variables as independent such as land use proxies. World Health Organization (WHO) has set a value of 40 μg/m^3 for the annual mean concentration limit in its air quality guidelines for NO_2 according to the adverse health effects arising from this pollutant. I applied LUR model with the most predictive independent variables to estimate NO_2 values for 27 countries across the Europe. I started model building with six independent variables. These variables were Agriculture, Road.Med, Temperature, Transport.Emiss, Industrial.Emiss, and Urban.Emiss. Before building the model, I inspected both non-transformed and log-transformed data distributions using Q-Q plots and Shapiro-Wilk test. Backward elimination method was used for model construction and collinearity between variables was tested via association testing between independent variables. The crude effects estimates were also measured to monitor collinearity between variables during the process of model building. No association between variables was strong and while some variables were not independent; their level of association was not indicative of collinearity. I built the final model with Medium Emission road, Estimated Transport and Industrial Emission. Agriculture was removed from the model for parsimony. Adjusted coefficient of determinations (R^2) of my model and stepwise regression models were 9.6% and 10.24%, respectively. Two models were significantly different with p-value = 0.026. Using my model ,areas with high transport and industrial emissions and adjacent to medium emission roads would have 8.34 μg/m^3 higher NO_2 concentration compared to areas with low transport and industrial emissions with no medium emission roads, on average.

Introduction

Land Use Regression (LUR) model is among some commonly used methods for predictions of air pollution concentrations such as Nitrogen Dioxide (NO_2). These models have a multivariate nature that encompass the pollutant concentrations as their dependent variable and predictor variables as independent such as land use proxies, environmental and socio-economical factors contributing to pollutant fate (1). These variables have the benefit of having less computational requirements and lower need of data input but on the other hand they are less reliable to be applied for large scale by just using variables based on land use only. Some recent efforts have been focused on the ability of these models on assessing the link between the health effects and pollutant exposure risk with emphasis on the high resolution results from these models (2). It is discussed that NO_2 concentrations are attributed to urban activities such as household emissions, transportation and industrial emissions, and agricultural activities including waste and solvent production (3). World Health Organization (WHO) has set a value of 40 µg/m^3 for the annual mean concentration limit in its air quality guidelines for NO_2 according to the adverse health effects arising from this pollutant. In order to account for the adverse environmental impacts of NO_2 and its contribution to both eutrophication and acidification of ecosystems as a result of nitrogen deposition; a more conservative number of 30 µg/m^3 for NO_2 mean annual concentration limit would be proposed (4).

In current study, I applied LUR model with the most predictive independent variables to estimate NO_2 values for 27 countries across the Europe. I used European-NO_2-LUR for this purpose and hypothesized that the presence of medium-emissions road within 100 m buffer, estimated annual NO_2 emissions from transportation infrastructure, estimated annual NO_2 emissions from industrial areas, estimated annual NO_2 emissions from urban areas are the most predictive independent variables and the final model should include them. I started model building with six independent variables. These variables were Agriculture, Road.Med, Temperature, Transport.Emiss, Industrial.Emiss, and Urban.Emiss.

Methods

European NO_2 land use regression dataset has 16 independent variables and 1 dependent variable. It has 1730 observations with 56 missing values for NO_2. The missing values were basically the values below 5 µg/m^3 which are unreliable for some older sampling instruments. Hence 5 µg/m^3 is accounted as Limit of Detection (LOD) although there is no universal LOD for NO_2. The measurements of NO_2 have been done across 26 European countries in 2010 with a range of chemiluminescent samplers. In order to perform regulatory air quality monitoring; these samplers were in compliance with reference methods of European Union.

The model needs to have high reliability for all Europe and farming activities are widespread in this continent so I decided to include Agriculture in my model. This variable is the total area (m^2) of agricultural land use within a 500m buffer around each station. It is a continuous variable and a surrogate of farming activity within Europe. I disregarded other land use proxies (Industrial and Urban) since I decided to use other better representative variables for urban and industrial activities. I disregarded Station Type and Station Location due to lack of harmonization in information gathering regarding these two variables across Europe. Each country has a different method on information gathering in terms of assigning each station to Station Type and Station Location, so I decided not to include these two variables in the model (5). Transportation is a major source of NO_2 and the presence of major road close to the station might be effective in regards to NO_2 concentrations predictions. I chose Road.Med to include in the model which is indicator of whether there is a medium-emissions road anywhere within 100m buffer around the station. This variable is binary in form of 0 = no and 1 = yes. A binary variable is a specific form of a dichotomous variable in which the categories are opposite of each other. This variable was preferred over other two binary variables (Road.High and Road. Low) since it is less extreme compared to the other two and a good indicator of transportation conditions. Some roads are major highways with very high emissions (Road.High) and some roads are small road with low traffic loads (Road.Low). The model needs a good representative variable of transportation infrastructure which in this case is medium-emissions road. Amongst the environmental variables I chose Temperature which is representative of climatic conditions. In our dataset this variable is annual average temperature at the station location in degrees Celcius. It is a continuous variable. Windspeed is very variable across even smaller areas and might not be a good predictive variable for my model so I decided not to include this variable. Elevation is also less important compared to other variables in our dataset. Transport.Emiss is estimated annual NO_2 emissions from transportation infrastructure during the 2010-2030 period from the GAINS emission database within 800m buffer around the station in tons/year. This variable is continuous and a good indicator of transportation infrastructure within Europe. Industrial.Emiss and Urban.Emiss are also modelled via GAINS database within 2000m buffer around the station in tons/year. Both variables are continuous and indicators of industrial and urban activities within Europe, respectively (http://gains.iiasa.ac.at).

Before building model with six variables including Agriculture, Road.Med, Temperature, Transport.Emiss, Industrial.Emiss, and Urban Emission in R Studio; I replaced missing values with LOD divided by square root of 2. I created a subset of data with NO_2 values above 30 $\mu g/m^3$. I investigated the normality of data distribution using Q-Q plot and ShapiroWilk test. The distribution of data was non-normal so I log-transformed data points. I also investigated normality of log-transformed data with Q-Q plot and ShapiroWilk test. The data was still non-normally distributed. Some studies in applying LUR model have continued to use data with the aforementioned nature so I decided to use log-transformed data (6). As a result of this, I performed multiple linear regression on log-transformed continuous variable. I tested the association between log-transformed dependent variable and each individual variables and built linear regression models with individual independent variables. Following the aforementioned tests, I also investigated the correlation between independent variables to find the highly correlated variables and identify the possible collinearity in the model built. I started model building with all six independent variables and log-transformed dependent variable and performed backward elimination. I removed all variables with p-value higher than 0.2 and ran the model with the remaining variables. The final model was built based on professional judgement considering both R^2 and parsimony.

Results

I built independent variable matrix with all 6 variables and tested the association between variables using correlation test for continuous vs continuous and t-test for binary vs continuous variables (Table 1). For each test of association I hypothesized that variables are independent from each other.

Null hypothesis: Variable A is independent from variable B.

Alternate hypothesis: Variable A is not independent from variable B.

Table 1: Independent variable matrix including p-values from tests of associations. P-values less than 0.05 show significant association between independent variables. Pearson correlation coefficient (R) is also included for test of association between continuous variables.

	Agriculture	Road.Med	Temperature	Transport.Emiss	Industrial.Emiss	Urban.Emiss
Agriculture		t-test	cor.test	cor.test	cor.test	cor.test
Road.Med	**P=5.8x10^{-10}**		t-test	t-test	t-test	t-test
Temperature	P=0.11 R=0.06	P=0.18		cor.test	cor.test	cor.test
Transport.Emiss	**P=1.05x10^{-8} R=-0.23**	P=0.27	**P=0.046 R=0.083**		cor.test	cor.test
Industrial.Emiss	**P=0.0007 R=-0.14**	**P=0.003**	P=0.27 R=0.046	**P=2.2x10^{-16} R=0.398**		cor.test
Urban.Emiss	**P=2.2x10^{-16} R=-0.34**	P=0.63	**P=0.019 R=0.097**	**P=8.44x10^{-9} R=0.237**	**P=3.49x10^{-12} R=-0.284**	

Among independent variables Agriculture and Industrial.Emiss are significantly associated with other variables except Temperature. All correlation coefficients are below 40% which is indicative of weak correlation between all variables even if p-values are below 0.05 which results in rejection of null hypothesis and variables not being independent from each other.

I ran simple linear regression between log-transformed NO_2 and individual independent variables to estimate the crude geometric effects and eventually compare them with geometric effect estimates of variables within multiple regression models to identify collinearity between variables (Table 2).

Table 2: Crude geometric effect estimate with p-value and multiple coefficient of determination (R^2). The association of each independent variable with log-transformed dependent variable (NO_2) is tested and simple linear regression is run to obtain crude geometric effect estimate.

Independent Variable	Test of Association	P-value	Crude Geometric Effect Estimate *=P<0.05	Multiple R^2
Agriculture	cor.test	0.0007	**0.9999699*** **	0.0199
Road. Med	t-test	$3.65x10^{-10}$	**1.164514*** **	0.0661
Temperature	cor.test	0.516	1.002054	0.0007
Transport.Emiss	cor.test	0.00053	**1.000001*** **	0.0206
Industrial.Emiss	cor.test	0.0254	**0.9999998***	0.0087
Urban.Emiss	cor.test	0.00037	**1.000004*** **	0.0218

All variables except Temperature were highly associated with log-transformed NO_2 and had significant crude geometric effect on dependent variable. Multiple R^2 for Temperature was also the lowest ($R^2 = 0.0007$).

I chose backward model building to build my final model from the six independent variables. In this method, I started to remove variable with the highest p-value and continued variable eliminations until all variables with p-values below 0.2 remained in the model. I monitored adjusted R^2 and p-value achieved from model. Moreover, I monitored the changes in significance level and direction of effect for each variable and compared the effect estimates with crude effects of each variable to track major changes to identify possible collinearity between variables (Table 3). Temperature and Urban.Emiss had p-values larger than 0.2. They both were removed in first and second steps. Model 3 had Agriculture, Road.Med, Transport.Emiss, and Industrial.Emiss. Although p-value for Agriculture was below 0.05 (p=0.026) and significant but the level of significance was different than the crude estimate. I decided to remove Agriculture from the model to have a more parsimonious model. I eliminated Agriculture and obtained the best model with R^2=0.096 (Model 4). Models 5 and 6 were created for parsimony and lost more variability than expected. I also ran stepwise regression and R software suggested Model 3 as the best model. I ran ANOVA between these two models and they were significantly different. The information of both models is also included in Table 3. Model 3 is suggested by stepwise regression and I obtained Model 4 from backward elimination. F values for Model 3 and 4 are 17.43 and 21.44, respectively.

Table 3: Backward model building started with removing variables with highest p-value until all variables with p-values below 0.2 remained in the model. Independent variable elimination continued for parsimony and R^2 was monitored in order not to lose the majority of variability explained by the model. The effect estimates with different significance level are highlighted in blue color.

Variables	Adjusted R^2	F vs. Previous	Agriculture	Road.Med	Temp	Transport .Emiss	Industrial .Emiss	Urban .Emiss
Model 1: Agriculture Road.Med Tempreture Transport.Emiss Industrial.Emiss Urban.Emiss	0.1024		0.9999834· (Changed significance)	**1.1451775*****	1.0026118	**1.0000009** (Changed significance)**	**0.9999998** (Changed significance)**	1.0000015 (Changed significance)
Model 2: Agriculture Road.Med Transport.Emiss Industrial.Emiss Urban.Emiss	0.1028	0.3914	0.9999845 (Changed significance)	**1.1444536*****		**1.0000009** (Changed significance)**	**0.9999998** (Changed significance)**	1.0000015 (Changed significance)
Model 3 (Also suggested by stepwise regression in R): Agriculture Road.Med Transport.Emiss Industrial.Emiss	0.1024	0.2650	**0.9999805* (Changed significance)**	**1.1397059*****		**1.000001*****	**0.9999998*****	
Model 4 (The final model): Road.Med Transport.Emiss Industrial.Emiss	0.096	0.026		**1.1481168*****		**1.0000011*****	**0.9999997*****	
Model 5: Agriculture Road.Med Transport.Emiss	0.084	0.0004	**0.9999827***	**1.152297*****		**1.000006** (Changed significance)**		
Model 6: Road.Med Transport.Emiss	0.079	0.0007		**1.159389*****		**1.000001** (Changed significance)**		

Discussion

The dataset that I used has measurements from 26 countries across Europe and as mentioned before there might be discrepancies in data collections and assigning types and locations to the NO_2 stations, hence building a model which has capability in explaining high variability in dependent variable is not easy. I built the most parsimonious model which is able to explain 9.6% variability in dataset. This is not a very high R^2 and it is important to consider that this land use regression model is not expected to have very high coefficient of determination for all Europe. The low reliability in large scale is a downside of these models. In the final model, Industrial and Transport Emission variables are correlated but their correlation coefficient is below 40% which is low and not strong indicator of their collinearity. On the other hand, after removing industrial emission from the model no strong changes were observed in effect estimate of Transport Emission when compared with its crude estimate effect. Before making the final version of model Agriculture was preferred to be removed since its removal presented changes in level of significance for other variables, while Industrial Emission removal had no effect of effect estimate in terms of significance level and change of direction. The effects of all variables in final model were highly significant with p-values less than 0.001. My final model was more parsimonious compared to stepwise regression model provided by R Studio with a 0.6% drop in coefficient of determination (R^2) which is negligible. The model estimates for low outcome and high outcome scenarios are as follows.

Low outcome scenario:

NO_2 ($\mu g/m^3$) = exp[$0.1381_{\text{Medium Emission Road}}$ x 0 + $1.081\text{x}10^{-6}$ x $5000_{\text{Transport Emission}}$ (tons/yr) - $2.754\text{x}10^{-7}$ x $750_{\text{Industrial Emission}}$ (tons/yr)]

Geometric mean of NO_2 with no medium emission road within 100m buffer, estimated transportation emission of 5000 tons/year, and estimated industrial emissions of 750 tons/year in one year is 38.39 [36.99 , 39.83] $\mu g/m^3$.

High outcome scenario:

NO_2 ($\mu g/m^3$) = exp[$0.1381_{\text{Medium Emission Road}}$ x 1 + $1.081\text{x}10^{-6}$ x $250000_{\text{Transport Emission}}$ (tons/yr) - $2.754\text{x}10^{-7}$ x $750000_{\text{Industrial Emission}}$ (tons/yr)]

Geometric mean of NO_2 with medium emission road within 100m buffer, estimated transportation emission of 250000 tons/year, and estimated industrial emissions of 750000 tons/year in one year is 46.73 [41.88 , 52.14] $\mu g/m^3$.

Estimated NO_2 concentration under high outcome scenario is 8.34 $\mu g/m^3$ higher than NO_2 concentration under low outcome scenario. It means that areas with high transport and industrial emissions and adjacent to medium emission roads will have 8.34 $\mu g/m^3$ higher NO_2

concentration compared to areas with low transport and industrial emissions with no medium emission roads, on average.

Compared to the study done by Henderson et al., 2007 (7) there was no seasonal information in my dataset. Seasonal variations have strong implications in NO_2 LUR studies. As mentioned before, lack of harmonization in data collection and assigning location and type to the stations was a limitation in this study. A harmonized way of data collection is needed for all Europe to include these two import variables in the model. Future studies should consider this to increase the reliability of the model.

References

1. Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X, et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. Atmos Environ [Internet]. Pergamon; 2013 Jun 1 [cited 2019 Mar 4];72:10–23. Available from: https://www-sciencedirect-com.ezproxy.library.ubc.ca/science/article/pii/S1352231013001386

2. Vienneau D, de Hoogh K, Beelen R, Fischer P, Hoek G, Briggs D. Comparison of land-use regression models between Great Britain and the Netherlands. Atmos Environ [Internet]. Pergamon; 2010 Feb 1 [cited 2019 Mar 4];44(5):688–96. Available from: https://www.sciencedirect.com/science/article/pii/S1352231009009601?via%3Dihub

3. European Commission. The 2015 Ageing Report: Economic and Budgetary Projections for the 28 EU Member States (2013-2060). Directorate-General for Economic and Financial Affairs, EC, Economic Policy Committee of the European Communities. Publications Office, Luxembourg. [Internet]. 2015. Available from: http://ec.europa.eu/economy_finance/publications/european_economy/2015/ee3_en.htm

4. World Health Organization. Air quality guidelines : global update 2005 : particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. World Health Organization; 2006. 484 p.

5. UN. Environmental Performance Reviews: The Former Yugoslav Republic of Macedonia - Google Books [Internet]. United Nations. 2002 [cited 2019 Mar 5]. Available from: https://books.google.ca/books?id=-1uKDF1WnRQC&pg=PA82&lpg=PA82&dq=harmonization+station+location+no2&source=bl&ots=UZfb7h-o07&sig=ACfU3U1T2rjsYPGvFgZFFUX8qiv-drEdRg&hl=en&sa=X&ved=2ahUKEwi38NDQzuzgAhWJjp4KHepRAjUQ6AEwBXoECAkQAQ#v=onepage&q=harmonization%252

6. Sahsuvaroglu T, Arain A, Kanaroglou P, Finkelstein N, Newbold B, Jerrett M, et al. A Land Use Regression Model for Predicting Ambient Concentrations of Nitrogen Dioxide in Hamilton, Ontario, Canada. J Air Waste Manage Assoc [Internet]. Taylor & Francis Group ; 2006 Aug 27 [cited 2019 Mar 5];56(8):1059–69. Available from: https://www.tandfonline.com/doi/full/10.1080/10473289.2006.10464542

7. Sarah B. Henderson †, Bernardo Beckerman ‡,§,∥, Michael Jerrett ‡,§,∥ and, Michael Brauer* †. Application of Land Use Regression to Estimate Long-Term Concentrations of Traffic-Related Nitrogen Oxides and Fine Particulate Matter. American Chemical Society ; 2007 [cited 2019 Mar 7]; Available from: https://pubs.acs.org/doi/abs/10.1021/es0606780