JOURNAL OF
EMERGING INVESTIGATORS

Article

A land use regression model to predict emissions from oil and gas production using machine learning

Elton Cao[1], Colby Francouer[2,3,4], Jian He[2,3]
[1]Fairview High School, Boulder, CO
[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO
[3]NOAA Chemical Sciences Laboratory, Boulder, CO
[4]Department of Mechanical Engineering, University of Colorado Boulder, Boulder, CO

SUMMARY

Emissions from oil and natural gas (O&G) wells such as nitrogen dioxide (NO_2), volatile organic compounds (VOCs), and ozone (O_3) can severely impact the health of communities located near wells. With the O&G industry growing and 17.6 million people living within a mile from an O&G well, the effect of O&G activity on residents is especially pertinent. In this study, we used O&G activity and wind-carried emissions to quantify the extent to which O&G wells affect the air quality of nearby communities, revealing that NO_2, NO_x, and NO are correlated to O&G activity. We then developed a novel land use regression (LUR) model using machine learning based on O&G prevalence to predict emissions. Many LUR models fail to account for O&G sources, therefore we hypothesized that the inclusion of O&G sources in land use regression models provides an increase in accuracy when predicting emissions. The model performed effectively for NO_2, outperforming past LUR models which did not involve O&G activities. The model makes it possible for not only communities, but also families and individuals, to determine the effect that O&G has on their homes. With current modeling techniques failing to observe the effects of O&G in the face of the growing O&G industry in the U.S., it is crucial that the public is educated on the effect of the O&G industry on their daily lives and has the tools to monitor these effects.

INTRODUCTION

The oil and natural gas (O&G) industry in the U.S. has grown sharply in recent years, with emphasis on the practices of hydraulic fracturing and horizontal drilling (1). As the frequency of such practices increased, the daily natural gas production in the United States also increased from 30 million m^3 per day in 2005 to more than 700 million m^3 per day in 2012, making up 39% of nationwide natural gas production (2). Additionally, between 2008 and 2014, oil production increased by 74% from 1.83 billion barrels a year to 3.18 billion barrels (1). However, these practices of O&G production also bring a multitude of environmental impacts, one of them being the emission of gases like methane into the atmosphere. Recent studies have shown that methane, a primary component of natural gas and a severe greenhouse gas, is emitted into the atmosphere at an estimated 13 million tons per year, which is 60% higher than the EPA national emission inventory made in 2015 (3). Therefore, methane emission numbers from sources like the O&G industry could potentially be more severe than expected. Due to methane being an ozone precursor, which is a pollutant that reacts to form ozone, increased methane emissions can lead to higher levels of tropospheric ozone, which is low lying ozone that can cause harmful health effects upon inhalation (4). O&G operations are also known to be an emission source of volatile organic compounds (VOCs) and nitrogen oxides (NO_x), which are ozone precursors as well (1,5). Increased levels of tropospheric ozone in the air have been associated with harmful effects on human health, such as cardiopulmonary diseases (6). Some VOCs are also carcinogens and can contribute to cardiovascular disease (7).

In the state of Colorado, practices like hydraulic fracturing have emerged both as a prime form of economy and a public health concern (8, 9). With a total of about 55,000 active O&G wells, Colorado has become a hotspot for O&G activity as the state with the 5th highest O&G production in the United States (19). Indeed, areas near O&G fields measured in Northeastern Colorado have already shown an increase in VOCs and other ozone precursors when compared to major US cities (5). Gilman et. al found that the presence of VOCs from natural gas activities, such as propane and ethane, were present at much higher concentrations in the Denver Julesburg region than other urban areas (5). In Colorado's Garfield County, another location of heavy O&G influence, residents living within 800m of an O&G well were found to be subject to the effects of various VOC emissions, such as benzene, a carcinogen (11). Czolowski et al. estimate that 17.6 million people live within a mile from an O&G well, therefore the impacts of O&G emissions are becoming increasingly dangerous (12).

The vast majority of people do not have access to air quality measurement tools, especially in low-income countries, so development of easier methods to model air quality is crucial (14). Monitoring instruments are also extremely costly and impractical for mass usage among citizens (15). Therefore, the development of land use regression (LUR) models which assess the various physical conditions around a location potentially provide an outlet for affordable air quality modeling (15). However, with the modern rise in O&G activities, many LUR models fail to take O&G sources into account (14,15). If an emission is found to be correlated with O&G activity, a LUR model that predicts the corresponding emission that is involved with O&G sources should prove to achieve higher

levels of accuracy than other mainstream LUR models which do not involve O&G sources.

Another issue with many LUR models is the complexity of the variables used. Oftentimes, LUR models utilize a variety of complex variables such as road density, road length, building density, agricultural density, etc. which may not be easily accessible to all people (14,15). These variables may often take a while to gather the necessary information. As a result, in this study, we will also be testing the viability of simpler variables that generalize complex variables into a single variable whose information can be easily gathered from spatial resources. For example, instead of many variables such as population density, road length, and vehicle usage to predict urban emissions, a single urban variable based on a city's population density and coverage was used instead. If such a variable is proven to be viable, the development of LUR models can be streamlined into a much more efficient process.

As the O&G industry becomes more prevalent, it is important for communities to determine the effect of O&G on their daily lives and health which can be done using LUR models. By developing a simpler O&G based LUR model which can be easily accessed by all people, such a task can be accomplished.

One of the goals of our study is to determine the extent to which the presence of O&G wells are related to increased emissions in the areas surrounding those O&G sites. A spatial analysis of the surrounding O&G wells was conducted to determine this extent. In addition to ozone, other indications of air quality were examined, such as particulate matter, which, along with NO_x, has also been linked to hydraulic fracturing (13).

To verify a correlation between O&G sources and increased emissions and also develop a LUR model, we developed equations to assess the level of O&G activity of an area to test if O&G activity is related to increased emissions. We then created a machine learning LUR model with generalized variables to predict the levels of emissions from O&G activity and other sources, which was compared with non-O&G LUR models to assess the influence of O&G activity as a variable in LUR models and the viability of generalized variables. We hypothesized that, with the growing influence of O&G activity, the inclusion of O&G sources in LUR models provides an increase in accuracy. It was found that emissions were linked to O&G activity, and the inclusion of O&G sources raised accuracy in predictions, particularly for a LUR model predicting NO_2.

RESULTS

In this study, we first investigated whether an emission is related to O&G activity, and then we built LUR models to measure their performance. Across Colorado are monitoring sites managed by the CDPHE (Colorado Department of Public Health and Environment) that measure various emissions. In analyzing the O&G activity surrounding each monitoring site and determining if there is a difference in emissions, the effect of O&G activity on air quality can be discerned. These data are then sent to the EPA where we obtained the data. We collected data from each of the CDPHE monitoring sites that measured wind direction and speed along with the observed emission. We also calculated locational prevalence for every site. Locational prevalence is a measure of the influence of O&G on a location we developed based on the number of wells surrounding the location. For sites that measured wind, we determined the wind prevalence for each observed day. Wind prevalence is another measure of influence of O&G on a location we developed based on the direction and speed of wind and the amount of O&G activity from where the wind comes from. Due to varying percentages of winds from each direction and speeds, wind prevalence varied day by day while locational prevalence remained the same.

We employed linear regression for each parameter using O&G prevalence to test for a correlation between O&G activity and emission type. Before building a model to predict emissions, it was important to first verify that O&G prevalence relates to each emission.

In order to develop a model to predict emissions based on locational prevalence and wind prevalence values of monitoring sites, we used multiple linear regression, a form of machine learning that utilizes multiple independent variables to predict a single dependent variable. In this case, independent variables were locational O&G prevalence, wind O&G prevalence, and urban prevalence. Urban prevalence is a measure of how urban a location is based on the size of nearby cities. Urban prevalence was developed to represent a simplified variable that predicts emissions from urban sources. Each independent variable was standardized for an average at 0 and a standard deviation of 1. These independent variables were used to predict the dependent variable, being the observed emission. Computer predicted data was tested against actual data to determine the accuracy of our model. We conducted regression and machine learning using the Scikit-Learn module in Python. Our emission model was compared with similar LUR models using r^2 and root mean squared error (RMSE).

The regression graphs used all available data from 2021 from the EPA. Certain monitoring sites which displayed large amounts of emissions (>20 ppb NO_2, >0.7 ppm ozone) due to influences like roads or urban areas were removed in order to gain a more accurate regression model as such monitoring sites were often heavily skewed past the capacities of the urban prevalence value.

Nitrogen Dioxide (NO_2)

NO_2 was the primary emission tested in this analysis. As NO_2 is a direct pollutant and heavily emitted from O&G activity, it served as a stable indicator for the effect of O&G on air quality and air quality models (7).

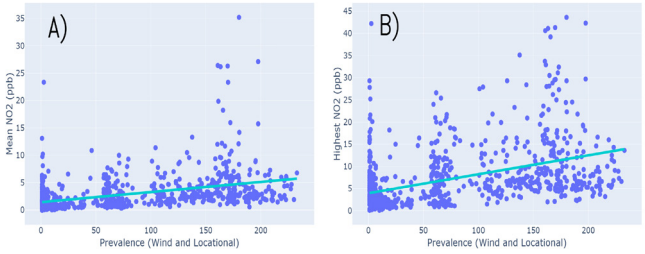

Figure 1: Regression graphs using O&G prevalence to predict NO_2 levels. (A) graphs prevalence equations with mean NO_2 and (B) graphs prevalence equations with highest NO_2 of the day. Mean NO_2 displayed an r^2 value of 0.22 and highest NO_2 displayed an r^2 value of 0.23.

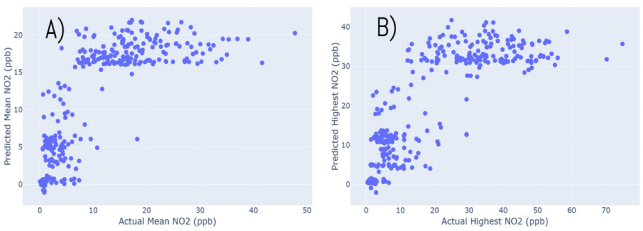

Figure 2: Multiple regression machine learning predictions between actual data and predicted values. (A) shows the prediction results for mean NO_2 and (B) shows the prediction results for highest NO_2 of the day. Mean displayed an r^2 value of 0.60 and highest displayed an r^2 value of 0.69.

NO_2 data were used from each recorded day in 2021 and we calculated the O&G prevalence for each of those days.

We observed a correlation between O&G prevalence and NO_2 levels (**Figure 1**). However, it is a weak correlation when considering the low r^2 values of 0.22 (mean) and 0.23 (highest value). This is likely due to the presence of third-party emissions. Although urban monitoring sites were removed, many of the monitoring sites kept were still subject to urban influence. Without the involvement of urban prevalence, the linear regression has no way to account for the influence of third-party sources such as urban prevalence.

The multiple linear regression models were far more accurate than the linear regression model judging by the r^2 values (**Figure 2**). The multiple linear regression displayed r^2 values of 0.60 (mean) and 0.69 (highest) with an RMSE of 6.36 (mean) and 4.32 (highest), whereas linear regression displayed r^2 values of 0.15 (mean) and 0.17 (highest) when comparing actual values with predicted values. With the addition of urban prevalence along with the separation of wind and locational prevalence, machine learning was as expected able to predict much more accurately than linear regression.

We found that models that predicted overall emissions were more accurate than just O&G sites with urban influenced data removed (**Figure 3a**) and urban prevalence with O&G influenced data removed (**Figure 3b**). This indicates that the improved results are not due to the inclusion of urban data in which urban prevalence easily predicted. It is only through the addition of both sources of data that allowed the O&G model to learn enough to effectively predict emissions.

Since the intended purpose of the O&G model was to predict emissions solely from O&G sources, this can be done by using urban prevalence to determine the amount of emissions to subtract after the O&G model predicts emission

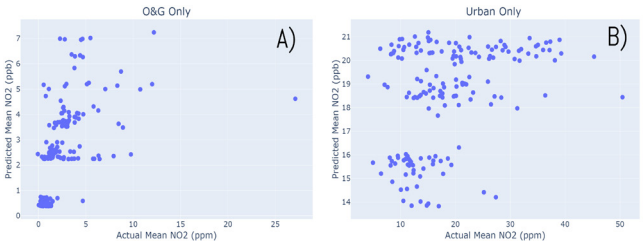

Figure 3: O&G model analysis comparing urban and O&G datasets. (A) displays actual vs. predicted values for a model with urban influences removed and (B) displays the model with O&G influences removed. r^2 value of (A) is 0.33 and r^2 value of (B) is 0.17.

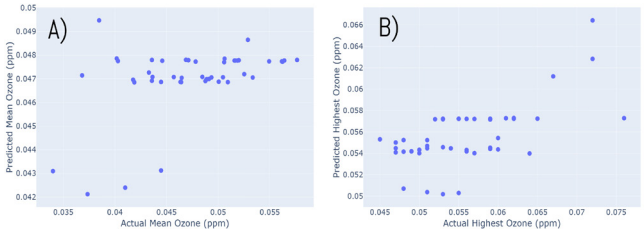

Figure 4: Regression graphs using O&G prevalence to predict ozone levels. (A) graphs prevalence equations with mean ozone and (B) graphs prevalence equations with the highest ozone of the day. Mean ozone displayed an r^2 value of 0.15 and highest ozone displayed an r^2 value of 0.33.

levels, as subtraction before analysis resulted in a much less accurate model.

Ozone (O_3)

Summertime (June, July, August) ozone was used due to ozone reactivity being higher in those months to determine if O&G affects ozone to a dangerous extent. Once again, linear regression was conducted between O&G prevalence (x) and ozone (y). Ozone was measured through ultraviolet absorption. Monitoring sites that measured abnormally high ozone levels were removed, as such high ozone levels indicate that there is heavy influence from another pollution source.

Ozone results, however, failed to show any correlation in linear regression solely based on O&G based equations. This is likely because there are too many factors that affect ozone, as it is a secondary pollutant.

However, multiple variable regression for ozone displayed a correlation between predicted and actual data, but the O&G model is still largely inaccurate in predicting ozone levels when noting the low r^2 levels (**Figure 4A-B**). This is likely because there are too many factors that affect ozone, as it is a secondary pollutant. Different methods will need to be used in order to accurately predict ozone. The amount of data available for ozone analysis was also less than for NO_2 analysis as many monitoring sites were deemed unfit due to abnormally high ozone levels, with five monitoring sites and 1050 measurements for ozone compared to nine monitoring sites and 1650 measurements for NO_2.

Emissions Summary

Regression conducted for PM_{10} and $PM_{2.5}$ (particulate matter of size <10 micrometers and <2.5 micrometers respectively) failed to display any correlation, indicating that the influence of O&G activities on these emissions is minimal or nonexistent. Regression for NO_x and NO showed an expected correlation, albeit a weak one with r^2 values of 0.21 and 0.10 respectively (**Figure 5**). Overall, the results indicate a lack of correlation in PM_{10} and $PM_{2.5}$. However, a correlation exists for NO_x, NO_2, and NO which are the main pollutants from O&G. For ozone, a correlation only existed in the multiple variable regression.

DISCUSSION

In this study, we discovered a link between O&G activity and the different emissions types. In addition, we also created a tool using computer code that makes it possible for a

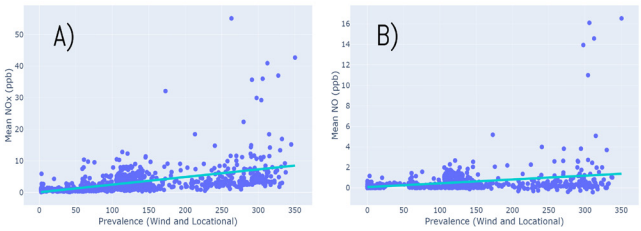

Figure 5: Regression graphs using O&G prevalence to predict NO_x (A) and NO (B) levels. Both graphs show mean emission levels. NO_x displayed an r^2 value of 0.21 and NO displayed an r^2 value of 0.10.

community, or even a family or individual, to determine the effect that O&G has on their homes. With the simple amount of data required to input to the O&G model and the ease of transporting, the O&G model can potentially be published and made available to the public in the future.

In order to determine if influences of O&G are related to air quality emissions and create a LUR model, equations to measure the amount of O&G influence were developed. These equations represent a novel way of assessing land use variables to create a generalized variable.

Our O&G NO_2 model will first be compared to Larkin et al. (14)'s NO_2 model. Larkin et al. (14) created a model to predict the entirety of North America primarily using satellite estimated NO_2 levels, traffic emissions, and other variables. NO_2 LUR models of North American regions are rare at this current time as LUR models are largely used for intra-urban estimates within large cities, not over large regions (18). Therefore, there are few LUR models to compare with.

When comparing each model's predicted values vs. actual values, the North American model had an r^2 value of 0.52 and an RMSE of 5.7, compared to our model's r^2 of 0.60 and RMSE of 6.4. r^2 values indicate that our model fits the data better, yet RMSE values indicate that the North American model had less deviation from the regression line. The North American model was more complex than the O&G model, with more specific road (ex: road length, usage) and urban measurements (ex: population density, building size) along with various other variables such as agriculture. However, despite this, our model achieved a higher r^2 value, the differentiator potentially being the inclusion of O&G sources.

A factor playing a role could also be that the O&G model is fitted to just Colorado and the North American model is for the entire continent. Therefore, it may be worth fitting the O&G model onto different American states to test if such results can be repeated.

It is also worth noting that other groups also generated models for other continents, and the model with most accuracy (South America) is the location in which O&G sources are least developed. Other continents involved heavily in O&G production all displayed lower r^2 values. Considering the jump in O&G production being rather recent, it makes sense that older models may not consider such sources.

Additionally, the purposes of each model should also be assessed when considering their performance. The North American model was adapted for usage in countries without monitoring equipment, especially for professional usage to develop health risk studies (14). Therefore, the users may have the resources to analyze land use variables on a much more specific scale than the O&G model can. This is because the O&G model was made for residential use, to be easily accessed through a website or similar tool by the general population. Therefore, the generalized variables used in the O&G model provide much more attainable variables for such circumstances. When you consider the relative accuracies of both models, it shows that such generalizations can also be used effectively in generating accurate predictions to an acceptable degree, implying the potential strength of generalized variables. Additionally, these generalized variables make it very simple to determine how much emissions come from each source, which may emerge as a valuable source of information in the rise of not only O&G activities but also future pollution sources. As more causes of O&G are discovered and incorporated into this model, the O&G model should only become more accurate in predicting emission levels.

Mavko et al. (15) also created an NO_2 LUR model for Portland, Oregon, which achieved an r^2 value of 0.89, exceeding both the North American model and the O&G model by a large scale. This model, however, was for a much smaller location, and was able to assess much more local and specific variables to generate the LUR model. Since 2008, Oregon has closed all O&G facilities thus this LUR model was completely unaffected by O&G activity (19).

Therefore, regarding the matter of scales of analysis, the North American model, the Oregon model, and the O&G model all show that the usage of LUR models often performs better on smaller scales. As a result, in future applications of machine learning for LUR, it may be beneficial to fit a single model across several smaller regions over a large area that is being examined.

Based on these results from the NO_2 model, the O&G model with generalized variables emerges as a novel method for providing simple and accessible air quality measurement tools to assess the effect of different emission sources on a given location. Other NO_2 LUR models with more data for more variables did not outperform our model, showing the influence of O&G sources on LUR models. As many nations are becoming immersed in the O&G industry, it is important to account for O&G wells in future air quality models (2).

Regarding ozone LUR models, models of such proportions remain to be effectively created. Wolf et al. (21) created an ozone LUR model for a city in Germany which effectively predicted ozone on a much smaller scale, with small scale intra-urban variables, similar to the Oregon model. Yet, from the O&G model's results, it is clear as to why ozone LUR models over large regions are difficult to create. When considering such a large radius in the O&G model, there is a large range of unpredictability when modeling ozone, being a secondary pollutant. This is a good indication of why there is a lack of large-scale ozone LUR models. Again, the benefits of smaller scale LUR models are displayed in the ozone results.

Overall, the comparison between the O&G model and smaller scale models indicated that LUR models may need to be fitted to smaller regions in order to predict with better accuracy. The O&G model, which was fitted over a large region, was much less accurate than smaller region LUR models which were able to better fine tune its predictions. In the future, instead of creating one large scale model, it may be beneficial to split the regions into smaller areas to fit a model piece by piece.

The effect of O&G on ozone emissions is still unclear. The O&G model indicates a weak correlation and a lack of large-scale ozone LUR models make it difficult to assess if O&G influences would benefit a LUR model. Ozone has been shown to be affected by O&G activities (20), yet the amount of effect and how it is affected is unclear.

The major limitation of this study lie in the lack of data currently open to the public. Not only are monitoring sites rather limited throughout Colorado, but the lack of wind data for most of the sites also proved to further limit the variety of sites that could be used in the analysis. The monitoring sites were also mostly clustered in similar areas, further reducing the usefulness of each of these sites due to lack of different data sources. Although each monitoring site provided sufficient and varying data, it would still be beneficial to have more sites to potentially identify further third-party emission sources.

Another issue is the problem with other pollution sources interfering with the O&G pollution. Although we accounted for urban emissions to create a more accurate model, it's still difficult to determine the exact amount of emissions from O&G influences. An ideal method would be to use the VOC signatures of each site to determine the sources of emissions (presence of ethane and propane would indicate emissions from O&G, presence of acetylene would indicate emissions from urban sources). However, only a small number of monitoring sites measure VOCs; it's not enough to create an accurate estimate, and much of VOC data is not disclosed for public usage.

Additionally, the weakness of the wind equation used is that it assumes that emissions blown by wind will follow a straight path to the monitoring site. Wind patterns often do not display such patterns with various changes in direction. Therefore, emissions can be carried to unexpected locations and inaccurately carried to the monitoring site. This equation also does not account for vertical winds, which could also arise as a form of error. However, since the equation is only meant to take a general and spatial analysis of O&G wells and not a specific one, the error should be small over long term analysis.

Both ozone and NO_2 have been linked to inflammatory lung reactions, which can lead to airway diseases, and NO_2 additionally can also worsen asthma symptoms and cause increased death from cardiovascular diseases (21). Therefore, the WHO (World Health Organization) has set thresholds for long term exposure of NO_2 at 40 $\mu g/m^3$, or 21.26 ppb (21).

Among values used in the LUR model of NO_2, 14.54% NO_2 values exceeded this limit, largely concentrated in areas located near highways or large urban areas, such as Denver. However, the O&G model tended to predict too low for these points, predicting values that were under the threshold. Additionally, only seven of the 340 values exceeded this threshold. As a result, when interpreting the results of this model, there should be a reasonable degree of caution.

For ozone, the WHO threshold is at 100 $\mu g/m^3$, or 50.94 ppb (22). In the data we used, 29.54% of ozone values of the LUR model exceeded the threshold, a much higher rate than NO_2. There wasn't a particular pattern to the points to which higher levels of ozone, which is logical considering the unpredictability surrounding the factors that cause the emission of ozone.

The results generally indicate that O&G factors may not affect emissions to a dangerous limit. With most of the points exceeding thresholds being caused by urban emissions, it's difficult to assess the exact influence of O&G. However, with monitoring locations located right in the middle of urban areas and subject to the most urban emissions, it is clear why most of these exceeding points are linked to urban sources. But for O&G wells, there was not a monitoring site which was located particularly close to an O&G well, nor a monitoring site located in the center of O&G activities on the same level that the urban monitoring sites are located in urban centers. Therefore, in order to both better determine if O&G factors may truly affect health to a dangerous level and optimize the LUR model, monitoring data is required for locations exposed heavily to O&G wells.

The O&G prevalence model was also more accurate in predicting highest emission levels of the day. Higher peaks are more evidence that an emission source is producing high amounts of emissions, which in this case can be attributed to O&G sources.

As the original O&G equations were modeled to create urban prevalence, they can be modified to account for different other sources of NO_2 and emissions. Significant power plants, construction sites, and roads can all be potentially accounted for in the future using these equations we have created to develop a much more accurate air quality model. Especially considering that the locational and wind prevalence equations can easily be modified for other sources, it is simple for one to take into account such sources. In this study, such a process was applied to create urban prevalence.

Comparisons between the O&G model and smaller scale models also indicated that LUR models may need to be fitted to smaller regions in order to predict with better accuracy. The O&G model, which was fitted over a large region, was much less accurate than smaller region LUR models which were able to better fine tune its predictions. In the future, instead of creating one large scale model, it may be beneficial to split the regions into smaller areas to fit a model piece by piece.

Regarding ozone, although ozone is a known pollutant from O&G, the O&G model had relatively low accuracy for predicting it. Ozone prediction methods will need to be adjusted in order to generate accurate predictions, with the inclusion of extra variables that may include ozone precursor levels and more specific indicators. A method that may be used in predicting future ozone concentrations is by using NO_2 levels, VOC estimates, and ozone reactivity to predict ozone concentrations for a location. Since ozone is highly dependent on NO_2 and VOC curves (20), this could work much better in providing an accurate estimate of ozone emissions. Although the O&G model is shown to be more suited to NO_2, the inclusion of extra variables that can better filter out emissions from other sources may help the O&G model adapt better to ozone predictions.

It is also worth improving the O&G model to account for more emissions sources which can eventually lead to creating a universal air quality measurement tool. With this research, the public will not only be able to determine the effect of O&G pollution on their homes, but also from urban sources, road sources, power plants, agriculture, etc.. By using a spatial analysis of surrounding sources along with a simple wind measurement, this is a powerful tool that may be utilized by anyone with access to a computer.

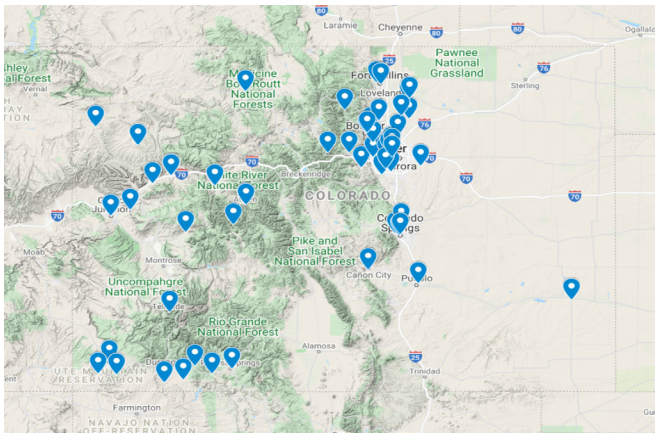

Figure 6. Map of monitoring sites across Colorado. Built using Google maps, it can be found in the following URL: https://www.google.com/maps/d/u/0/edit?hl=en&mid=1XUhtaOJhpa3HbIMyhDXrAEP48-Q_8RE&ll=39.31800047048807%2C-106.00521116294118&z=7

MATERIALS AND METHODS

Monitoring Sites and Data

Monitoring sites and data were obtained from the EPA at the following link: https://aqs.epa.gov/aqsweb/airdata/download_files.html. The EPA's data contains data uploaded from the CDPHE. Methods of emission measurements are therefore consistent with the standards of the CDPHE and the EPA. The parameters that are used in this study are nitrogen oxides (NO_x), nitrogen dioxide (NO_2), nitric oxide (NO), ozone (O_3), PM_{25}, and PM_{10}. Monitoring sites are located across Colorado, largely concentrated in urban locations (**Figure 6**).

Equations and Modeling

We created equations based on different buffer distances that assess the various circumstances involved in O&G emissions. The equations generated numerical values that tell the influence of O&G for a given location. These equations represent a novel way of assessing land use variables to create a generalized variable. To our understanding, this is the first attempt at creating such generalized variables to use with a LUR model.

The first equation involves the proximity that a location has on nearby wells, which will be referred to as "locational prevalence." The following equation was created to predict emissions based on wells in close proximity:

$$P=2A+B+0.5C+0.2D+0.01E$$

where P is the "O&G prevalence" based on well distances and number, A is the number of O&G sites within an 800 m radius of the monitoring site, B is the number of sites within an 800-2000 m radius, C is the number of sites within a 2-3 km radius, D is the number of sites within a 3-10 km radius, and E is the number of sites within a 10-20 km radius. Higher values of O&G prevalence indicate a higher concentration of O&G in the area, whereas lower values of O&G prevalence indicate lower concentrations of O&G activity. We used the Python programming language to calculate this O&G prevalence. The code for this project can be found at the following Github link: eltonc01/OG-ML-Study (github.com).

Areas within 800 m from O&G sites have significantly increased exposure to air emissions (11), hence the distance of A in the equation. A U.S. EPA report regarding the dilution of toxic air contaminants also found that in areas within 800 m from the source of emissions was 0.1 g/m^3 per g/s (16). In areas within 2000 m of the source recorded 0.015 g/m^3 per g/s, and in areas within 3000 m, there was a dilution of 0.007 g/m^3 per g/s (16). A study regarding proximity of natural gas wells and the effect of emissions based on different buffer distances also reported that respiratory symptoms were more frequent among locations <1 km away from the source compared with locations >2 km away (16). Based on these data, we have chosen the distances of B and C in the equation. For benzene (a VOC), it was determined that beyond 3 km, dilutions were two orders of magnitude less than the 800 m radius and were all relatively equal (16). These dilutions determined the distance of D in the equation. The value of D is to take into account of O&G wells are on the border of affecting the location of the monitoring site. Lastly, the distance of E is to add on to the spatial representation of the prevalence of O&G wells. Although these distances may not have a large or direct impact on the routines of the examined site, they are still useful in order to gain an idea of the amount of O&G production in an area.

In general, the coefficients of the equations have been chosen to reflect the impact that a well may have based on the distance from the examined site. However, although locations closer to O&G sites have displayed much larger emissions than distances further away, the coefficients of the equation do not directly reflect this difference in emissions. Having large coefficients can significantly alter the data, so in order to lessen this impact, the coefficients are more focused on a spatial representation of O&G sites in the general vicinity. In the case that data regarding the O&G wells have been improperly collected by the COGCC, the changing of coefficients would also help reduce the error in the calculations, so that these errors play less of a role in changing the O&G prevalence. Essentially, the coefficients help take a more conservative method of building the equation in order to counteract the heavy variance involved with predicting emissions.

The second equation takes wind into account, which can blow emissions from farther O&G wells to an observed location. The first equation alone cannot entirely be accurate in predicting emissions, so the following equation was developed to account for wind:

$$P = \sum^{Directions} \frac{15x(\frac{s}{2} - (5z - 25))}{2L}$$

where P is the "O&G prevalence" based off wind, L is the distance from nearby O&G activity, z is the wind speeds (km), s is the prevalence strength of the nearby O&G activity, and x is the % of winds blowing in the direction being measured. The equation is a sum of the eight different cardinal directions (north, northeast, east, southeast, south, southwest, west, northwest), which represents the direction the wind is blowing.

Distance from nearby O&G activity, or L, was calculated by finding the center point of nearby O&G wells in the region described by the cardinal directions. Each region can be described as a "pie slice," in which we can find the center point of O&G wells for distances between 20 km and 50 km away. This distance value is chosen to build from the first equation, which stops inputting values beyond 20 km. The center point of O&G wells will be calculated by finding the

Population	Radius (km)	Strength
500k+	15	10
400-500k	12	8
300-400k	8	6
100-300k	6	5
50-100k	4	3
10-50k	3	2
5-10k	2	1
1-5k	1	0.1

Table 1: Table of urban prevalence radius and strength based on city population.

average latitudes and longitudes of all wells in each "pie slice." Once the center points are calculated, the distances will be measured in kilometers.

Prevalence strength of the O&G activity nearby, or 's,' was calculated by counting the number of O&G wells in the "pie slice."

Lastly, percent of winds measured, or x, is the percentage of measured winds that blew in the cardinal direction that is being used. This was collected by finding the percentage of winds blowing in each direction based on hourly measurements for the year of 2021.

Each equation was developed and put into action using Python.

Urban Prevalence

Monitoring sites located at or near urban locations such as cities were subject to much higher levels of emissions like NO_2 and NO_x, indicating the large influence of such urban emissions. Therefore, locational and wind prevalence equations have been modified to account for urban influence, creating an "urban prevalence" value. As the equations account for a spatial and wind analysis of emissions, they should be able to be applied to other sources of emissions.

Population of each city determined the radius of influence along with the emission strength of each city (**Table 1**). For example, if a monitoring site was located at the black 'x', then the model would determine the amount of influence from Denver that factors into the equation (**Figure 7**). The red circle represents Denver's radius of influence, and the amount of overlap would be counted towards the equation. Strength acts as a multiplier of the overall result—the more populous a city is, the denser it will be, creating more potential for emissions. As radius increases outwards, strength also decreases. This setup effectively creates a generalized variable which is based on the principle that as a city's influence recedes the further you are.

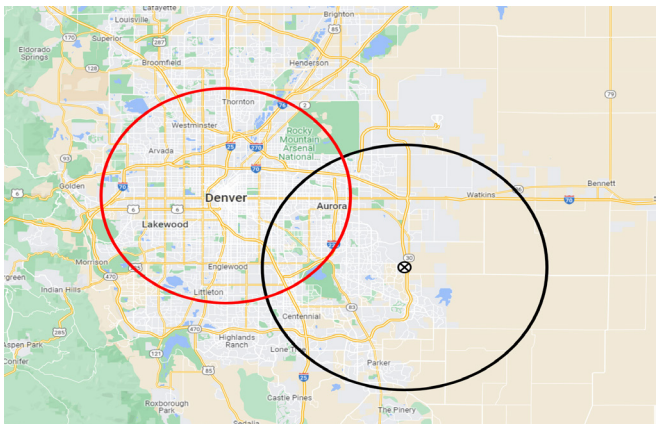

Figure 7. Example of urban prevalence equation application. The red circle represents Denver's area of influence, and the black 'x' and circle indicate a monitoring site and its 50 km area of influence. Image was created using Google Maps.

ACKNOWLEDGEMENTS

We would like to thank Dr. Paul Strode (Fairview High School) for guiding our endeavors in this project, providing valued resources, and sparking our passions for science.

Received: July 5, 2022
Accepted: September 20, 2022
Published: March 24, 2023

REFERENCES

1. Allen, David T. "Emissions From Oil and Gas Operations in the United States and their Air Quality Implications." *Journal of the Air & Waste Management Association*, vol. 66, no. 6, 1 June 2016, pp. 549-575. doi:10.1080/10962247.2016.1171263.
2. Jackson, Robert B., et al. "The Environmental Costs and Benefits of Fracking." *Annual review of Environment and Resources*, vol. 39, no. 1, 11 Aug 2014, pp. 327-362. doi:10.1146/annurev-environ-031113-144051
3. Zhang, Yuzhong, et al. "Quantifying Methane Emissions from the Largest Oil-Producing Basin in the United States from Space." *Science Advances*, vol. 6, no. 17, 22 Apr 2020, doi:10.1126/sciadv.aaz5120.
4. Anenberg, Susan C., et al. "Global Air Quality and Health Co-Benefits of Mitigating Near-Term Climate Change through Methane and Black Carbon Emission Controls." *Environmental Health Perspectives*, vol. 120, no. 6, 1 June 2012, pp. 831-839. doi:10.1289/ehp.1104301.
5. Gilman, Jessica B., et al. "Source Signature of Volatile Organic Compounds from Oil and Natural Gas Operations in Northeastern Colorado." *Environmental Science & Technology*, vol. 47, no. 3, 14 January 2013, pp. 1297-1305. doi:10.1021/es304119a.
6. Jerrett, Michael, et al. "Long-Term Ozone Exposure and Mortality." *New England Journal of Medicine*, vol. 360, no. 11, 12 March 2009, pp. 1085-1095. doi:10.1056/NEJMoa0803894
7. Li, Hugh Z., Matthew D. Reeder, and Natalie J. Pekney. "Quantifying Source Contributions of Volatile Organic Compounds Under Hydraulic Fracking Moratorium." *Science of the Total Environment*, vol. 732, 11 May 2020, doi:10.1016/j.scitotenv.2020.139322.
8. Mayer, Adam. "Risk and Benefits in a Fracking Boom: Evidence from Colorado." *The Extractive Industries and Society*, vol. 3, no. 3, 8 August 2016, pp. 744-753. doi:10.1016/j.exis.2016.04.006.
9. Thompson, Chelsea R., et al. "Influence of Oil and Gas Emissions on Ambient Atmospheric Non-Methane Hydrocarbons in Residential Areas of Northeastern

Colorado NMHC in Residential Areas of Northeastern Colorado." *Elementa: Science of the Anthropocene*, vol. 3, 14 November 2015, doi:10.12952/journal.elementa.000035.
10. U.S. Energy Information Administration (EIA) (7/1/2022), from eia.gov/energyexplained/oil-and-petroleum-products/where-our-oil-comes-from.php
11. McKenzie, Lisa M., et al. "Human Health Risk Assessment of Air Emissions from Development of Unconventional Natural Gas Resources." *Science of the Total Environment*, vol. 424, 22 March 2012, pp. 79-87. doi:10.1016/j.scitotenv.2012.02.018.
12. Czolowski, Eliza D., et al. "Toward Consistent Methodology to Quantify Populations in Proximity to Oil and Gas Development: a National Spatial Analysis and Review." *Environmental Health Perspectives*, vol. 125, no. 8, 23 August 2017, doi:10.1289/EHP1535.
13. McDermott-Levy, Ruth, Nina Kaktins, and Barbara Sattler. "Fracking, the Environment, and Health." *AJN The American Journal of Nursing*, vol. 113, no. 6, June 2013, pp. 45-51. doi:10.1097/01.NAJ.0000431272.83277.f4.
14. Larkin, Andrew, et al. "Global Land Use Regression Model for Nitrogen Dioxide Air Pollution." *Environmental Science & Technology*, vol. 51, no. 12, May 2017, pp. 6957-6964. doi:10.1021/acs.est.7b01148.
15. Mavko, Matthew E., Brian Tang, and Linda A. George. "A Sub-Neighborhood Scale Land Use Regression Model for Predicting NO_2." *Science of the Total Environment*, vol. 398, no. 1-3, April 2008, pp. 68-75. doi:10.1016/j.scitotenv.2008.02.017.
16. Rabinowitz, Peter M., et al. "Proximity to Natural Gas Wells and Reported Health Status: Results of a Household Survey in Washington County, Pennsylvania." *Environmental Health Perspectives*, vol. 123, no. 1, 1 January 2015, pp. 21-26. doi:10.1289/ehp.1307732.
17. Gately, Conor K., et al. "Urban Emissions Hotspots: Quantifying Vehicle Congestion and Air Pollution Using Mobile Phone GPS Data." *Environmental Pollution*, vol. 229, 30 June 2017, pp. 496-504. doi:10.1016/j.envpol.2017.05.091.
18. Beelen, Rob, et al. "Development of NO_2 and NO_x Land Use Regression Models for Estimating Air Pollution Exposure in 36 Study Areas in Europe–The ESCAPE Project." *Atmospheric Environment*, vol. 72, February 2013, pp. 10-23. doi:10.1016/j.atmosenv.2013.02.037
19. U.S. Energy Information Administration (EIA) (7/1/2022), from eia.gov/state/analysis.php?sid=OR
20. Francoeur, Colby B., et al. "Quantifying Methane and Ozone Precursor Emissions from Oil and Gas Production Regions across the Contiguous US." *Environmental Science & Technology*, vol. 55, no. 13, 23 June 2021, pp. 9129-9139. doi:10.1021/acs.est.0c07352.
21. Wolf, Kathrin, et al. "Land Use Regression Modeling of Ultrafine Particles, Ozone, Nitrogen Oxides and Markers of Particulate Matter Pollution in Augsburg, Germany." *Science of the Total Environment*, vol. 579, January 2017, pp. 1531-1540. doi:10.1016/j.scitotenv.2016.11.160.