Atmospheric Environment 119 (2015) 381–392

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Application of nonparametric regression and statistical testing to identify the impact of oil and natural gas development on local air quality

Hanqi Cheng [a,*], Mitchell J. Small [a], Natalie J. Pekney [b]

[a] *Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[b] *National Energy Technology Laboratory, 626 Cochrans Mill Rd., P.O. Box 10940, Pittsburgh, PA 15236, USA*

HIGHLIGHTS

- Air monitoring data were analyzed from one background and two downwind sites.
- Kernel regression was used to identify directional trends in concentration.
- A block bootstrap test for trend significance accounted for autocorrelation.
- A statistically significant regional impact was identified at the background site.
- The downwind sites exhibited statistically significant impacts from local sources.

ARTICLE INFO

Article history:
Received 24 February 2015
Received in revised form
18 June 2015
Accepted 5 August 2015
Available online 8 August 2015

Keywords:
Air pollution
Oil and natural gas
Directional analysis
Statistical methods
Nonparametric regression
Block bootstrap

ABSTRACT

The objective of the current work was to develop a statistical method and associated tool to evaluate the impact of oil and natural gas exploration and production activities on local air quality. Nonparametric regression of pollutant concentrations on wind direction was combined with bootstrap hypothesis testing to provide statistical inference regarding the existence of a local/regional air quality impact. The block bootstrap method was employed to address the effect of autocorrelation on test significance. The method was applied to short-term air monitoring data collected at three sites within Pennsylvania's Allegheny National Forest. All of the measured pollutant concentrations were well below the National Ambient Air Quality Standards, so the usual criteria and methods for data analysis were not sufficient. Using advanced directional analysis methods, test results were first applied to verify the existence of a regional impact at a background site. Next the impact of an oil field on local NO_x and SO_2 concentrations at a second monitoring site was identified after removal of the regional effect. Analysis of a third site also revealed air quality impacts from nearby areas with a high density of oil and gas wells. All results and conclusions were quantified in terms of statistical significance level for the associated inferences. The proposed method can be used to formulate hypotheses and verify conclusions regarding oil and gas well impacts on air quality and support better-informed decisions for their management and regulation.

1. Introduction

Air pollutant emissions from oil and natural gas (O&NG) exploration and production activities are of growing concern as advanced drilling technologies have enabled the rapid expansion of unconventional gas extraction from shale, tight sandstone and coal beds (Moore et al., 2014). As shown in previous studies, the extensive and intensive O&NG activities in Texas, Colorado and Wyoming have caused significant impacts to local and regional air quality, including elevated levels of volatile organic compounds (VOCs) and nitrogen oxides (NO_x), and induced high production of ground-level ozone (O_3) (Katzenstein et al., 2003; Schnell et al., 2009; Gilman et al., 2013; Moore et al., 2014). Since July 28, 2011, the US Environmental Protection Agency (EPA) has promulgated a series of rules and updated standards to regulate emissions from O&NG industries and protect public health and welfare (US EPA, 2014). The current study was concerned with the development of

* Corresponding author.
E-mail address: hanqic@andrew.cmu.edu (H. Cheng).

http://dx.doi.org/10.1016/j.atmosenv.2015.08.016

statistical tools to assess the contribution of O&NG operations to local air pollutant concentrations.

O&NG exploration and production activities involve well development and gas production and processing activities that can be a significant source of a variety of air pollutants. Some examples include diesel-fueled engines of drill rigs and pumps (VOCs, NO_x, fine particulate matter ($PM_{2.5}$) and CO_2), diesel trucks used to transport materials, natural gas powered compressor engines (VOCs, NO_x and greenhouse gases including CO_2 and methane), and fugitive emissions from condensate storage tanks and pipelines (VOCs and methane). These emission sources are widely distributed and highly variable for different operation strategies and different stages of the well lifetime. Additionally, the composition of raw gas varies by reservoir or even by site, resulting in varied profiles of fugitive VOCs emissions.

Due to the high spatiotemporal variability of O&NG emissions, applying traditional source-oriented air quality models to source apportionment (SA) is difficult. An alternative is receptor modeling, based on mass balance and mathematical criteria. Since O&NG operations involve multiple highly variable activities, the output factors of receptor models may mix O&NG emissions with other sources and the impact of O&NG operations becomes difficult to isolate.

To help identify a directional signal in observed concentration data, wind sector analysis (WSA) and parametric regression have been used to fit concentration-wind direction relationships (Somerville et al., 1994, 1996). Wind directions with high pollutant concentrations were identified and the corresponding sources could be recognized. Following the analyses a statistical hypothesis test was often carried out to verify the presence of systematic directionality. Nonetheless, WSA requires long-term monitoring data to guarantee rigorous concentration estimates (Donnelly et al., 2011). Additionally, parametric regression involves strong assumptions about the form of the model, thereby limiting its generality.

Recently, nonparametric regression (NPR), especially kernel regression, has been used successfully in source identification/location and apportionment. Henry et al. (2002) found that a regression curve analysis of pollutant concentration on wind direction can be used to effectively locate nearby emission sources. Incorporating wind speed into the model helps to separate out the mix of sources around a monitoring site (Yu et al., 2004; Henry et al., 2009). Applying NPR to NO_2 and wind direction data collected at background sites reveals the presence of regional range transport of pollutants and distant emission sources (Donnelly et al., 2011). Though NPR can produce smoothed regression curves with confidence bands, no statistical framework has been constructed to test for the existence of directionality in the context of NPR. Conclusions were mostly drawn directly from regression plots. In addition, the autocorrelation effects of time series data on NPR was mentioned but not dealt with.

Instead of drawing inferences solely from resulting plots of NPR, the current study aimed to construct a statistical framework of hypothesis tests to give quantitative probability evidence of the air quality impact from local O&NG sources and to differentiate local from regional effects. Nonparametric techniques including kernel regression and bootstrap sampling were combined and employed to provide model tests under relatively unrestrictive assumptions. In addition, an approach for addressing the autocorrelation effects of data on NPR was incorporated as the framework is developed.

The methods were applied to air quality and meteorological data measured at three sites within Pennsylvania's Allegheny National Forest (ANF) by the U.S. National Energy Technology Laboratory (NETL). Analyses of the data can be found in Pekney et al. (2014). As shown below, the developed hypothesis-testing framework was able to provide statistical inference regarding the existence of local O&NG operations and emissions after removal of the regional effect. The results indicate that when background variation (i.e., the regional effect) is non-negligible, this statistical approach works well for short-term (e.g., month-long) monitoring data without emission and baseline concentration information.

2. Data

2.1. Locations

To evaluate the air quality impact from O&NG exploration and production activities, NETL developed a trailer-based autonomous air-monitoring laboratory and deployed it to the Allegheny National Forest in northwest Pennsylvania, historically a productive area of oil and gas wells, from July 2010–June 2011. Meteorological and air quality data were collected at three monitoring sites within the ANF for analysis. The first site was the Kane Experimental Forest (KEF) area in Elk County, which is in the direction of prevailing winds at (henceforth referred to as “downwind of”) the Sackett oilfield. The second site was at the Bradford Ranger Station (BRS) in McKean County, downwind of a large area of historic productivity. Since oil and gas wells have existed in the vicinity of these areas, it is not possible to obtain the baseline concentrations at these two sites. The third site was located at the U.S. Forest Service Hearts Content (HC) campground in Warren County, in an area relatively unimpacted by O&NG development. It was selected to yield reference background pollutant concentrations in the ANF. The number of wells within 10 km of the HC, KEF and BRS sites were about 450, 950 and 3800, respectively. Fig. 1 shows the locations of these three monitoring sites and the oil and gas wells in ANF.

2.2. Air quality and meteorological data and pre-processing

NETL's air-monitoring laboratory had all of its instruments installed in a trailer, operating automatically and collecting data continuously with little user intervention. Particulate matter ($PM_{2.5}$ and PM_{10}, both in $\mu g/m^3$) and a wide range of VOCs (in parts per billion, ppb) were determined in one-hour increments. O_3, nitric oxide (NO), nitrogen dioxide (NO_2) and sulfur dioxide (SO_2) were measured every minute in ppb. Meteorological data (e.g., temperature, humidity, wind direction and wind speed) were also recorded on a minute-averaged basis. Wind data were measured 6 m above the ground. Wind direction is available when wind speed is above the detection limit (i.e., 1 mph, or 0.45 mps). It was measured clockwise from the north (i.e., in the azimuth angle). Generally, abnormal data (e.g., negative concentrations) and data measured during calibration or instrument malfunction were removed in the QA/QC process. The regression analysis was based on hourly-averaged data (minute-averaged data were first transformed into hourly averaged data). Accounting for the circular nature of wind direction data, hourly averaged wind direction was calculated through vector addition, which was adopted in Donnelly et al. (2011).

It was planned that the air-monitoring trailer would stay at each site for 1–2 months. However, due to instrument problems or power outages, either wind or pollutant data were missing for certain periods of time. Therefore, for the HC site, data available for analysis cover a period of approximately six weeks. For the KEF site, more data are accessible. However, as temperature is another important factor that may affect pollutant concentrations (e.g., NO_x), only data measured within a temperature range similar to that of the HC site were included in the analysis, spanning a six-week period. For the BRS site, more meteorological data were absent and only two-week pollutant-wind paired data are available for regression analysis. So