1
2 Application of nonparametric regression and statistical
3 testing to identify the impact of oil and natural gas
4 development on local air quality
5
6
7

8 Hanqi Cheng^{a,*}, Mitchell J. Small^a, Natalie J. Pekney^b

9 ^aDepartment of Civil and Environmental Engineering, Carnegie Mellon University,
10 Pittsburgh, PA 15213, USA

11 ^bNational Energy Technology Laboratory, 626 Cochrans Mill Rd., P.O. Box 10940,
12 Pittsburgh, PA 15236, USA
13
14
15
16
17
18

19 Revised submission to
20 Atmospheric Environment
21

22 June 15, 2015
23
24
25
26

27 * Corresponding author.

28 Email address: hanqic@andrew.cmu.edu

29 Tel: +1(412)-951-3988 Fax: (412)-268-7813

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.

Key words: Air pollution; Oil and Natural Gas; Directional analysis; Statistical methods; Nonparametric regression; Block bootstrap

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 (O₃) (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 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 CO₂), diesel trucks used to transport materials, natural gas powered compressor engines (VOCs, NO_x and greenhouse gases including CO₂ 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; Somerville et al. 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 NO₂ 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.

Fig. 1 here

Fig. 1 Locations of the three monitoring sites and upwind oil and gas wells in ANF (From Pekney et al., 2014).

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 PM₁₀, both in $\mu\text{g}/\text{m}^3$) and a wide range of VOCs (in parts per billion, ppb) were determined in one-hour increments. O₃, nitric oxide (NO), nitrogen dioxide (NO₂) and sulfur dioxide (SO₂) 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

186 site, more meteorological data were absent and only two-week pollutant-wind paired data
187 are available for regression analysis. So the regression results of the BRS site were used
188 more for comparison and reference rather than for inference.

189
190 With limited wind speed monitoring data (due to partial malfunctioning of the
191 anemometer, causing wind speeds to be systematically underestimated), it is not possible
192 to incorporate wind speed into the regression model as a second predictor variable.
193 Nonetheless, a qualitative consideration of relative wind speeds reported for different
194 directions may contribute to better interpretation of the concentration variations observed.
195 Appendix A shows wind roses for the three monitoring sites based on uncorrected wind
196 speeds and provides a detailed discussion of these plots based on the regression curves
197 developed later in the paper. No significant mixing or dilution signals associated with
198 strong wind speeds were identified. As such, the regression results obtained in this work
199 are believed to be relatively insensitive to wind speed or wind speed – wind direction
200 correlations.

201
202 The analysis of Pekney et al. indicated that NO_x concentrations at the three sites were not
203 significantly different and were generally much lower than the National Ambient Air
204 Quality Standard. In addition, only 3-4 species of VOCs were detected for more than
205 65% of the time at two or more of the three sites (Pekney et al., 2014). However,
206 considering the baseline concentrations were different at these locations, comparison of
207 absolute values of the basic statistics across sites does not necessarily indicate whether
208 the local impact from O&NG industries exists or not. The statistical tests illustrated
209 below focused on comparison of directional regression curves to identify and confirm the
210 existence of local impact.

211
212 The National Emission Inventory (US EPA, 2011) was used to identify any major
213 emission facilities in the vicinity of the monitoring sites. The information on traffic
214 volumes for the roads in the monitoring counties was obtained from the Pennsylvania
215 Department of Transportation.

216 217 **3. Methods**

218 219 *3.1 Kernel estimator*

220
221 The NPR technique adopted in this work was kernel estimation. A number of studies
222 have demonstrated its validity in SA (Henry et al., 2002; Yu et al., 2004; Henry et al.,
223 2009; Donnelly et al., 2011). Nonparametric kernel regression is effective when

performed on short-term data, which is one major advantage over WSA (Donnelly et al., 2011).

The essential idea of the kernel estimator is to estimate the pollutant concentration at a focal wind direction θ by the locally weighted average over all the observed concentrations in the sample. The weight of the i th observation is determined by the kernel function K and the relative difference between the focal wind direction θ and the i th wind direction W_i . So the estimator of pollutant concentration at wind direction θ is defined as follows,

$$\hat{C}(\theta) = \frac{\sum_{i=1}^n K((\theta - W_i)/h) C_i}{\sum_{i=1}^n K((\theta - W_i)/h)} \quad (1)$$

where n is the sample size and C_i is the i th pollutant concentration corresponding to the wind direction W_i . The relative distance $\theta - W_i$ is scaled by a smoothing parameter/bandwidth h , which determines the smoothness of the resulting curve.

There are many possible forms for the kernel function K . However, statistical theoretical calculations have shown that prediction risk is insensitive to this choice. Different kernels often yield numerically similar estimates (Wasserman, 2005). Therefore, this work adopted the Gaussian kernel to estimate pollutant concentration from wind direction, which has been used in a number of the previous studies on SA noted above. The Gaussian kernel is defined in Eq. (2).

$$K(x) = (2\pi)^{-1/2} \exp(-0.5x^2) \quad -\infty < x < \infty \quad (2)$$

The selection of the bandwidth h is critical in kernel regression because h is the only adjustable parameter and determines the degree of smoothness of the final result. A small h results in an under-smoothed curve maintaining too much background variation while a large h results in an over-smoothed curve eliminating systematic signals. One bandwidth selection method frequently adopted in statistical studies is the leave-one-out cross validation (LOOCV), defined as,

$$h^* = \operatorname{argmin}_{h \in (0, 180)} CV(h) = \operatorname{argmin}_{h \in (0, 180)} \sum_{i=1}^n (C_i - \hat{C}_{(-i)})^2 \quad (3)$$

where

$$\hat{C}_{(-i)} = \frac{\sum_{j \neq i} K((W_i - W_j)/h) C_j}{\sum_{j \neq i} K((W_i - W_j)/h)} \quad (4)$$

The optimal bandwidth minimizes the sum of squared error, while the estimate of the i th concentration is obtained by the same kernel regression formula as Eq. (1) but leaving out the i th observation. LOOCV is simply data-driven and makes no assumptions about the distribution of the explanatory variable. In the context of directional data, it is more

feasible than the “plug-in” method used by Xu and Schoenberg (2007) and Donnelly et al. (2011).

While our primary focus was on tests of significance for directionality, the confidence interval of the regression curve also provides useful visual information for source identification and location in a graphical analysis. Previous applications have often employed the point wise confidence interval developed by Härdle (1992). For graphical display purpose, we calculated confidence intervals using the same method. In the context of bootstrap testing, the bootstrap confidence interval in Härdle and Bowman (1988) was adopted to estimate the variance of the regression curve.

3.2 Autocorrelation and block bootstrap

Wind and concentration time series data used in this study both exhibit a high degree of positive serial correlation. Fig. 2 shows the autocorrelation function for the residuals of the kernel smoothing fit of the NO_x data at the HC site. The residuals show strong correlation at a lag of 1 hour. Autocorrelation of residuals can cause the bootstrap confidence intervals to be too narrow and lead to an overestimate of the significance of statistical tests (Zwiers, 1990). To address these problems and improve the accuracy of test results, the block bootstrap method was employed to preserve the serial correlation of the data in the bootstrap samples.

The block bootstrap can work well under very weak assumptions and has a broad range of possible applications in time series analysis (Gentle et al., 2012). Instead of sampling individual data independently, a block of l consecutive data is sampled each time. The re-sample set is obtained by sampling n/l blocks. The blocks can be non-overlapping or overlapping. The overlapping blocks include all possible blocks with length l in the data.

Regarding the selection of block size l , Henry et al. (2009) suggested using 24 hrs because of the common presence of diurnal periodicities. However, as shown in Fig. 2, with a lot of missing data, the current residuals do not indicate a significant 24-hr periodicity. Further, a block length of 24 seems too large for the small data size here and may lead to overestimation of the autocorrelation effect. Hall et al. (1995) showed that the optimal block size depends significantly on the context, being equal to $n^{1/3}$ in the case of variance estimation. This study thus employed $n^{1/3}$ to calculate block size. For the case in Fig. 2, a block size of 9 hr was calculated.

Fig. 2 here

Fig. 2 Autocorrelation function for the residual of kernel smoothing of the HC site NO_x data. The dashed lines indicate rejection region bands for testing of no autocorrelation.

3.3 Statistical testing for kernel regression

This section presents the statistical framework developed for comparison of directional significance between kernel regression curves at two sites. In this application one site was the background HC site; the other was either of the two monitoring sites with potentially local emission sources. The goal of the procedure was to identify and verify the impact of local O&NG exploration and production activities with concurrent characterization of background directionality. Scheme 1 outlines the steps in the statistical testing procedure developed in this study.

Scheme 1 here

Scheme 1 Steps in the statistical analysis of directionality in sites with potential background and local source influence

3.3.1 Test for directionality at background site

The first step in the statistical method was to characterize the directionality of pollutant concentrations at background sites and assess the presence of potential impact from distant sources. In this work, a full bootstrap test of significance was developed, accounting for the influence of autocorrelation.

The hypotheses for the test were

H_0 : Pollutant concentration shows no directionality at the background site,

H_1 : Pollutant concentration shows directionality at the background site.

To test the existence of association, a permutation test is often employed. The test statistic for nonparametric regression is based on the generalization of the regression statistic Q , defined in the work of Raz (1990).

$$Q = \sum_{i=1}^n (\hat{C}(W_i) - \bar{C})^2 \quad (5)$$

where $\bar{C}$ is the mean of the observed concentrations.

To address the autocorrelation effect on test significance, a refined block permutation test was constructed. The name “block permutation test” originated from the block bootstrap method. Since a permutation test requires sampling without replacement, the non-overlapping version was adopted.

The block permutation algorithm for obtaining the null distribution of the test statistic Q is as follows:

1. Determine the block length l by $l = n^{1/3}$ (Hall et al., 1995).
2. Divide the n consecutive concentrations into $[n/l] + 1$ non-overlapping blocks.
3. Sample the non-overlapping blocks of concentration data randomly without replacement, assigning each to a segment of the observed wind direction data to obtain new, random wind-concentration (W, C') pairs. These pairs randomized the (W, C') relationship, removing any correlation that may have been present in the original dataset. The randomized sequence thus corresponded to the null hypothesis of “no relationship”.
4. Obtain the kernel estimates of concentrations for the new (W, C') pairs. Store the results and calculate the test statistic Q for the new regression.
5. Repeat Step 3 and 4 many times, obtaining the empirical sampling distribution of Q under the null hypothesis.

After the null distribution of the test statistic Q was simulated, calculate the test statistic Q_t for the original regression results. The level of significance (p -value) for the test of hypothesis was calculated by locating Q_t on the permutation distribution:

$$p\text{-value} = \frac{1 + \text{number of values greater than } Q_t \text{ in the null distribution}}{1 + \text{replications}} \quad (6)$$

Racine (1997) suggested 1000 replications for obtaining the tail percentiles. In addition, a 95% null hypothesis band can be generated by joining the 2.5- and 97.5- quantiles of the regression results of the permutations. Displaying the null hypothesis band and the kernel regression curve on the same plot gave a visual comparison and intuitive sense of the significance of the directionality.

The same procedure was used to test for the existence of directionality at the other two monitoring sites with potentially local sources.

3.3.2 Test for comparison of regression curves at two sites

The tests for directionality at two sites may both give low p -values, indicating the existence of directionality at both locations. Under this circumstance it becomes difficult to verify the impact of local sources since the directionality at a suspected locally impacted site can also be attributed to some regional impacts. Assuming the directional variation of regressed concentrations at the background site provides a good reference for the regional impacts, then at a site affected by local sources, the regression curve should contain higher or more peaks than the background site since there exist not only regional

but also local impacts. It might be easy and intuitive to compare the two regression curves graphically. However, random fluctuations can also cause two curves to look different while the inherent relationship is essentially the same. More importantly, while this work focused on the difference in shapes, two curves may differ only by a simple shift due to different baseline concentrations. Recognizing these two issues, a statistical test gives more reliable and quantitative evidence of whether two curves are different.

The hypotheses for the test were

H_0 : The regression curve at a suspected locally impacted site has the same/similar shape as the regression curve at the background site (that is, there are no local impacts),

H_1 : The regression curve at a suspected locally impacted site has more or higher relative peaks than the regression curve at the background site (there are local impacts).

The Kolmogorov-Smirnov distance is a statistic frequently used to test for agreement in the shapes of two curves. It is the maximum difference between the two curves, defined as

$$d = \max |F_1(x) - F_2(x)| \quad (7)$$

A two-sample block bootstrap test was developed for the current work using the Kolmogorov-Smirnov distance as the test statistic.

The bootstrapping algorithm for obtaining the null distribution of the test statistic d is as follows:

1. Obtain the means of the pollutant concentrations in the two samples respectively. To normalize the samples, divide the concentrations by their own mean so the baseline difference is eliminated.
2. Obtain the kernel regression estimates $\hat{C}_b(W_i)$ of normalized concentrations $C_b(W_i)$ at the background site, using Eq. (1).
3. Generate residuals $\hat{\epsilon}_i = C_b(W_i) - \hat{C}_b(W_i)$ for all $i = 1, \dots, n$. Then re-center them around zero.
4. Draw blocks of residuals with length l ($l = n^{1/3}$) randomly with replacement and make up a new bootstrap residual sample $\{\hat{\epsilon}_i^*\}$ with a size of n .
5. Generate a new bootstrap data sample by $C_n(W_i) = \hat{C}_b(W_i) + \hat{\epsilon}_i^*$.
6. Obtain kernel regression estimates $\hat{C}_n(W_i)$ for the new data sample $C_n(W_i)$ by Eq. (1) and calculate the test statistic $d = \max |\hat{C}_n(W_i) - \hat{C}_b(W_i)|$.
7. Repeat Steps 4~6 many times, obtaining the null distribution of the test statistic d . Steps 3~5 provide the basic procedure for constructing bootstrap confidence intervals and characterizing the variance of the regression curve.

After the null distribution of d was simulated, the kernel estimates $\hat{C}_l(W_i)$ for the normalized sample $C_l(W_i)$ at the suspected locally impacted site were obtained. Since the alternative hypothesis only addresses whether the curve at the local site has more or higher peaks (i.e., it is a one-sided test), the test statistic for the two curves was calculated as Eq. (8),

$$d_t = \max \left(\hat{C}_l(W_i) - \hat{C}_b(W_i) \right) \quad (8)$$

which is the maximum amount by which the local curve exceeds the background curve (in contrast to the absolute difference between the two curves). The level of significance can be calculated by locating d_t on the null distribution:

$$p\text{-value} = \frac{1 + \text{number of values greater than } d_t \text{ in the null distribution}}{1 + \text{replications}} \quad (9)$$

Since a relatively simple statistic was used and may lead to greater uncertainties, the sampling distribution of d was simulated using 10,000 replications to obtain a more reliable bootstrap p -value.

Comparisons between sites might also evaluate differences in the distribution or central tendency of concentrations, irrespective of the wind direction of the measurements. In the following section we consider possible insights from such a test by applying a nonparametric Mann-Whitney U-test to compare medians at downwind vs. background sites.

To supplement our evaluation of potential local and regional sources, we implemented the HYSPLIT back-trajectory model using the three monitoring sites as the receptor points (Draxler and Rolph, 2014). The resulting backward trajectory plots at each site for a two-week monitoring period are presented in Appendix B. The trajectories were smooth, exhibiting no abrupt changes in direction on a local scale. This result indicated that local sources may be effectively located using wind directions measured at the site.

4. Results and discussion

4.1 Comparison of NO_x variations across the three monitoring sites

The kernel estimates of NO_x concentrations as a function of wind direction for the HC, KEF and BRS site data are displayed graphically in Figs. 3 (a)-(c), respectively. A 95% point-wise confidence interval and 95% null hypothesis band representing no directionality are shown together with the regression curve.

Fig. 3 (a) shows the result at the background HC site. The directional variation exhibited by the kernel regression of NO_x concentrations was relatively small. Identification of a single, directional peak was difficult, but in general higher NO_x concentrations were observed when winds blew from the south. Note that northwest to north winds corresponded to below average NO_x concentrations, which can occur due to the absence of major sources located in these directions from the site, or impacts from other meteorological conditions (e.g., enhanced dispersion by high wind speed) (Donnelly et al., 2011). The block permutation did produce a much wider null hypothesis band than the regular permutation method. This made it difficult to determine the presence of directionality graphically. Table 1 summarizes the *p*-values for all the statistical tests implemented in this work. For NO_x estimates at the HC site, the block permutation test for no directionality gave a low *p*-value of 0.022. The method and data indicate the presence of directionality of NO_x concentration at the background site, most likely resulting from the regional impacts of distant emission sources. Additionally, the average NO₂/NO_x ratio over the monitoring period is 0.96, which provides good evidence that fresh combustion sources of NO_x were not present.

Table 1

Summary of *p*-values for all bootstrap tests in this work

Pollutant \ Test	Test for directionality			Test for comparison of two curves at HC and	
	HC	KEF	BRS	KEF	BRS
NO _x	0.022	<0.001	<0.001	0.0015	0.046
SO ₂	0.073	<0.001	0.13	0.074	0.481
Ethane	0.003	0.058	N/A	0.434	N/A

Fig. 3 (b) shows the result for the KEF site, which is downwind of the Sackett oilfield. In contrast to Fig. 3 (a), the regression curves in Fig. 3 (b) departed significantly from the null hypothesis band, with two peaks for southwest and southeast winds, respectively. The directionality at the KEF site was statistically significant at the 99% confidence level. Visual comparison of the two regression curves in Fig. 3 (a) and (b) indicated that they were very different. The two-sample block bootstrap test for comparison of the two regression curves produced a *p*-value of 0.0015. Thus with more than 99% confidence the regression curve at the KEF site had higher and more peaks than that at the HC site, supporting the argument that there were emission sources of NO_x in the vicinity of KEF that appeared to have a noticeable impact on the air quality at this location. The average NO₂/NO_x ratio at the KEF site is 0.70. This relatively low percentage of NO₂ in NO_x also indicates closer combustion emissions.

Fig. 3 (c) shows the result at the BRS site, which is downwind of a large area of historic productivity. Appropriately paired wind and concentration data are available for this site for only 12 days out of the total sampling period of seven weeks. A broad range of winds, clockwise from the south to the northwest, gave higher NO_x concentrations than from the remaining sectors. This observation was consistent with two areas of dense wells located close to the site. The variation was not significant from the figure due to a wide and relatively asymmetric null hypothesis band. The “abnormal” band resulted mainly from the low density of data in some wind directions. Nonetheless, the *p*-value of the block permutation test is less than 0.001, indicating the existence of directionality with more than 99% confidence. The following two-sample test to compare the regression curve with that at the background site gave a *p*-value of 0.046. Therefore, at the 95.4% confidence level, the NO_x variation at the BRS site is indicated to result from both local and regional impacts.

Fig. 3 here

Fig. 3 Kernel regression results for NO_x concentrations at the (a) HC (b) KEF (c) BRS site

4.2 Comparison of SO₂ variations across the three monitoring sites

Fig. 4 (a)-(c) give the kernel regression of SO₂ concentrations on wind direction at the HC, KEF and BRS sites, respectively. As shown in Fig. 4 (a), at the background site, south to southwest winds produced the highest SO₂ concentrations. However, the majority of the curve was within the null hypothesis band. The *p*-value of the block permutation test for no directionality is 0.073; this is not a very low value (and there was only 92.7% confidence to say that SO₂ had directionality at the HC site). Nonetheless, comparing Fig. 4 (a) with Fig. 3 (a) revealed that at the same location, the regression plots for SO₂ and NO_x were very similar. These pollutants may share similar distant sources.

In Fig. 4 (b), the departure of the regression curve from the null hypothesis band was obvious. There were two peaks pointing towards the southwest and southeast of the site, respectively. This directional pattern was very similar to that in Fig. 3 (b). Again, the same sources may dominate SO₂ and NO_x at the KEF site. However, the relative heights of the peaks in the two curves were different. The southeast peak was higher for NO_x and the southwest peak was higher for SO₂. In addition, the southeast peak in the SO₂ curve was not very significant since only the upper confidence interval was beyond the null hypothesis band. This observation suggests that the potential source located southeast of the site was a greater relative contributor of NO_x.

The test for no directionality produced more than 99% confidence that the SO₂ directional variation existed. The difference between the KEF and HC curves was graphically clear. However, the test to compare these two curves only gave a *p*-value of 0.074. The existence of local impacts was thus confirmed with only 92.6% confidence. In contrast to the NO_x case, the magnitude of the SO₂ residuals was comparable to the magnitude of the estimated SO₂ concentrations. These residuals resulted in greater uncertainties in the estimate of the variance of the regression curve, and thus a greater variance for the null distribution of the K-S distance. Therefore, when the magnitude of errors is very significant, test conclusions tend to be conservative.

The SO₂ directional variation in Fig. 4 (c) was very insignificant. The regression line was completely within the null hypothesis band. Generally, the concentrations associated with south winds were higher. This trend was very similar to that at the background site, reflecting some regional impacts. The test for no directionality gave a *p*-value of 0.13; this value was not low enough to reject the null hypothesis. Therefore, there was only little, if any, SO₂ directional variation at the BRS site. In practice it may not be necessary to carry out the two-sample bootstrap test to compare the BRS and HC curves since the permutation test shows there is no statistically significant contributor of SO₂ at the BRS site. Nonetheless, to verify our conclusion here and examine the performance of the two-sample test, the directionality at the BRS site was compared with that at the HC site; the two-sample block bootstrap test gave a *p*-value of 0.481. This suggests the presence of little or no local impact at the BRS site. Though a large area of gas wells was surrounding the south of the site, it had no apparent impact on the local SO₂ concentration. A stronger verification of this inference will require a larger dataset at the BRS site.

Fig.4 here

Fig. 4 Kernel regression results for SO₂ concentrations at the (a) HC (b) KEF (c) BRS site

4.3 Comparison of Ethane variations across the three monitoring sites

During the monitoring period a variety of VOCs were measured. However, for most species only a relatively small percentage of the samples were above the detection limit. Additionally, to choose a VOC species for regression analysis, it should be a major component of the raw natural gas. Two possible candidates are ethane and propane. The correlation coefficients of ethane and propane concentrations at the three sites are 0.954, 0.940 and 0.945, respectively. Since these two species are highly correlated, the regression results should be very similar and ethane was chosen for the regression and directional testing analysis. Sufficient wind speed data for this analysis are available only

for the HC and KEF sites, and results of this analysis are presented in Appendix C. At the HC site, the ethane directionality was clearly displayed and statistically significant. At the KEF site, the directionality was not as significant and there was little support for the presence of local impacts.

4.4 Identification of the regional and local impacts

The block permutation test results indicate statistically significant evidence that NO_x, SO₂ and ethane concentrations at the HC site were all affected by regional scale impacts. Additionally, it is shown that the same distant sources may dominate the three pollutants in the background sector location(s). The Pittsburgh metropolitan area with a high density of population and commercial/manufacturing activities is located approximately 150km to the southwest of the ANF area. With no other direct emissions or major roads in the vicinity of the HC site, the analysis suggests that the directional variations in NO_x, SO₂ and ethane concentrations were mainly contributed by the Pittsburgh metropolitan area and other possible distant sources.

For pollutants NO_x and SO₂, the two-sample bootstrap test reveals that their concentrations at the KEF site were affected by direct local emissions in addition to the regional impacts. Furthermore, the two pollutants may share the same local sources with similar peak locations. Since the southeast peak in the curve for SO₂ did not depart enough from the null hypothesis band to yield statistical significance, the potential source located to the southeast of the KEF site was a significant contributor of NO_x, but not SO₂. Based on the 2011 National Emission Inventory, the most probable facility source is the Domtar-Johnsonburg Paper Mill, located 14 km to the southeast of the KEF site, with an annual NO_x emission of 973 tons. Two small townships are also located close to the southeast of the site. The nearby roads all have relatively low traffic volume, so they are unlikely to have a noticeable effect on the air quality at the monitoring location.

The only possible source located to the southwest of the KEF site is the Sackett oil field. On the map it is an area with a high density of wells. The distance between the center of the area and the monitoring location is approximately 14 km. There was no other high density of wells, facilities, or major roads in the vicinity of the KEF site. This analysis suggests that oil industry activity did have an impact on the local air quality, with noticeable increases in NO_x and SO₂ concentrations. However, oil and natural gas industries are usually not considered a significant source for SO₂. Some possible on-site sources include diesel-powered drilling rigs and pumps, flaring of excess natural gas, oil field preprocessors and fuels used by workers living on site.

Statistical testing also indicates that NO_x concentrations at the BRS site were affected by both local and regional impacts. The higher concentrations of NO_x were consistent with two areas of dense wells, located to the northwest and southwest of the site, respectively. With only several days of data, greater uncertainties exist in the BRS site analysis.

4.5 Analysis methods

In the analysis to identify the impact of O&NG industries on local air quality, the combination of nonparametric regression and statistical testing was found to provide a number of advantages. First, the statistical testing enabled information from a small dataset to be more rigorously evaluated and interpreted. In previous studies, at least a one-year long dataset was used for NPR analysis. With more data, wind speed can also be incorporated into the model as the second predictor variable. In this work, the largest dataset is less than six weeks. As such, the model could only consider one predictor variable and the estimates contained higher uncertainties. However, hypothesis testing can offer quantitative evidence of the existence of regional and local impacts when graphical analysis is equivocal. Second, the methodology developed here enabled comparison between two sites with different baseline concentrations. This advantage can be demonstrated from Table 2. Table 2 lists the *p*-values of the two-sample block bootstrap test and the Mann-Whitney U test. The second test was used to compare the median concentrations of two samples. As can be seen from Table 2, the Mann-Whitney test gave strong evidence that the median concentration of NO_x at the HC site was higher than that at the KEF or BRS site. Therefore, simply comparing the absolute value of concentrations at two sites does not reveal clear and complete information about the impact from local sources. However, as illustrated above, the two-sample block bootstrap test was able to identify directional differences between two concentration curves and give statistical evidence about local impacts. Last, autocorrelation effects were fully considered in the methodology. The statistical significance will be overestimated if the autocorrelation of the data is not addressed. With the application of block bootstrap, the null hypothesis band was wider and some “false” peaks were eliminated.

Table 2

Comparison of *p*-values for two-sample bootstrap test and the Mann-Whitney U test

Pollutant	Site 1	Site 2	Two-sample bootstrap test	Mann-Whitney U test
			H ₀ : regression curve at site 1 is the same as that at site 2 (no local impact) H ₁ : regression curve at site 1 has more or higher peaks than that at site 2	H ₀ : median concentration at site 1 equals that at site 2 H ₁ : median concentration at site 1 is greater than that at site 2
NO _x	KEF	HC	0.0015 (reject H ₀)	≈ 1 (retain H ₀)

NO _x	BRS	0.046 (reject H ₀)	≈ 1 (retain H ₀)
SO ₂	KEF	0.074 (reject H ₀ on 90% CL)	(reject H ₀)
Ethane	KEF	0.434 (retain H ₀)	(reject H ₀)

5. Conclusions

The impact of oil and natural gas exploration and production activities on local air quality in Pennsylvania's Allegheny National Forest were analyzed using nonparametric regression of pollutant concentrations on wind direction and bootstrap testing for the existence of regional/local impacts. For the background site, test results reveal statistically significant evidence of the existence of regional impacts. For the site downwind of a currently producing oil field with active operations, test results indicate the existence of impact from the oil field on local NO_x and SO₂ concentrations after removing the regional effect. Another local source of NO_x was also identified during the analysis. For the last site, regression analysis was not fully implemented due to a high proportion of missing data. Nonetheless, the impact of high densities of gas wells on local NO_x concentrations was clearly supported.

Statistical testing was shown to be an effective method to identify the local impacts in nonparametric regression results with a quantified significance level. It can be applied successfully to short-term monitoring data. The block bootstrap was incorporated into the methodology to address autocorrelation effects of time series data and prevent overestimation of test significance. The method worked well with concentrations well below the National Ambient Air Quality Standards. Therefore, it is a powerful tool to identify the impact from existing O&NG operations even when relative standards and regulations are not violated. This is extremely important for areas where future expansion of O&NG exploration and production activities is expected. The possible impacts can be effectively estimated and corresponding control strategies can be identified.

Acknowledgements

This project was funded by the US Department of Energy, National Energy Technology Laboratory (NETL), as part of the Regional University Alliance (RUA), RES contract DE-FE0004000. The authors would like to acknowledge the assistance of Ralph Perron, Jim McCloskey, Ernie Wiltsie, and Susan Stout of the U.S. Forest Service Allegheny National Forest and the Northern Research Station in establishing the monitoring locations, and James Sams of the NETL for preparing a figure used in the manuscript. The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model. The views and opinions expressed herein are solely those of the authors and do not necessarily state or reflect those of the United States government or any agency thereof.

References

- Donnelly, A., Misstear, B., Broderick, B., 2011. Application of nonparametric regression methods to study the relationship between NO₂ concentrations and local wind direction and speed at background sites. *Science of the Total Environment* 409(6), 1134-1144, DOI: 10.1016/j.scitotenv.2010.12.001.
- Draxler, R. R. and Rolph, G. D., 2014. HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) Model access via NOAA ARL READY Website (<http://www.arl.noaa.gov/HYSPLIT.php>). NOAA Air Resources Laboratory, College Park, MD.
- Gentle, J. E., Härdle, W. K., Mori, Y. (Eds.), 2012. *Handbook of Computational Statistics*. Springer, New York.
- Gilman, J. B., Lerner, B. M., Kuster, W. C., de Gouw, J. A., 2013. Source signature of volatile organic compounds from oil and natural gas operations in northeastern Colorado. *Environmental Science & Technology* 47(3), 1297-1305, DOI: 10.1021/es304119a.
- Hall, P., Horowitz, J. L., Jing, B., 1995. On blocking rules for the bootstrap with dependent data. *Biometrika* 82(3), 561-574, DOI: 10.1093/biomet/82.3.561.
- Härdle, W., Bowman, A. W., 1988. Bootstrapping in nonparametric regression: local adaptive smoothing and confidence Bands. *Journal of the American Statistical Association* 83(401), 37-41, DOI: 10.1080/01621459.1988.10478572.
- Härdle, W., 1992. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.
- Henry, R. C., Chang, Y. S., Spegelman, C. H., 2002. Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction. *Atmospheric Environment* 36, 2237-2244, DOI: 10.1016/S1352-2310(02)00164-4.
- Henry, R., Norris, G. A., Vedantham, R., Turner, J. R., 2009. Source region identification using kernel smoothing. *Environmental Science & Technology* 43, 4090-4097, DOI: 10.1021/es8011723.
- Katzenstein, A. S., Doezema, L. A., Simpson, I. J., Black, D. R., Rowland, F. S., 2003. Extensive regional atmospheric hydrocarbon pollution in the southwestern United States. *Proceedings of the National Academy of Science of the United States of America* 100(21), 11975-11979, DOI: 10.1073/pnas.16352581000.
- Moore, C. W., Zielinska, B., Pétron, G., Jackson, R. B., 2014. Air impacts of increased natural gas acquisition, processing, and use: a critical review. *Environmental Science & Technology* 48(15), 8349-8359. DOI: 10.1021/es4053472.
- Pekney, N. J., Hammack, R., Velsoki, G., Tamlia, J., 2014. Measurement of atmospheric pollutants associated with oil and natural gas exploration and production activities in Pennsylvania's Allegheny National Forest. *Journal of the Air and Waste Management Association* 64(9), 1062-1072.

- Pennsylvania Department of Transportation. Traffic volume maps (Updated for 2012).
<http://www.dot.state.pa.us/Internet/Bureaus/pdPlanRes.nsf/infoBPRTrafficInfoTrafficVolumeMap?openform> (last access Feb 20, 2014)
- Racine, J., 1997. Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics* 15, 369-378, DOI: 10.1080/07350015.1997.10524714.
- Raz, J., 1990. Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *Journal of the American Statistical Association* 85 (409), 132-138, DOI: 10.1080/01621459.1990.10475316.
- Schnell, R. C., Oltmans, S. J., Neely, R. R., Endres, M. S., Molenaar, J. V., White, A. B., 2009. Rapid photochemical production of ozone at high concentrations in a rural site during winter. *Nature Geoscience* 2(2), 120-122, DOI: 10.1038/ngeo415.
- Somerville, M. C., Mukerjee, S., Fox, D. L., Stevens, R. K., 1994. Statistical approaches in wind sector analyses for assessing local source impacts. *Atmospheric Environment* 28(21), 3483-3493, DOI: 10.1016/1352-2310(94)90006-X.
- Somerville, M. C., Mukerjee, S., Fox, D. L., 1996. Estimating the wind direction of maximum air pollutant concentration. *Environmetrics* 7, 231-243, DOI: 10.1002/(SICI)1099-095X(199603)7:2<231::AID-ENV207>3.0.CO;2-I.
- US EPA. The 2011 National Emission Inventory.
<http://www.epa.gov/ttnchie1/net/2011inventory.html> (last access Feb 9, 2015)
- US EPA, 2014. Oil and Natural Gas Air Pollution Standards: Regulatory Actions. Washington, DC: Office of Air and Radiation, US Environmental Protection Agency.
<http://www.epa.gov/airquality/oilandgas/actions.html> (last access Feb 9, 2015)
- Wasserman, L., 2005. *All of Nonparametric Statistics*, Springer Texts in Statistics. Springer, New York.
- Xu, H., Schoenberg, F. P., 2007. Kernel regression of directional data with application to wind and wildfire data in Los Angeles County, California. Department of Statistics, UCLA.
- Yu, K., Cheung, Y., T., C., Henry, R. C., 2004. Identifying the impact of large urban airports on local air quality by nonparametric regression. *Atmospheric Environment* 38, 4501-4507, DOI: 10.1016/j.atmosenv.2004.05.034.
- Zwiers, F. W., 1990. The effect of serial correlation on statistical inferences made with resampling procedures. *Journal of Climate* 3, 1452-1461, DOI: 10.1175/1520-0442(1990)003<1452:TEOSCO>2.0.CO;2.

*If the background site shows no directionality while the other monitoring site does, a judgment regarding the existence of local sources can be made without the comparison of two curves.

a

b

C

a

b

C