PERGAMON

Atmospheric Environment 36 (2002) 2237–2244

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction

Ronald C. Henry[a,]*, Yu-Shuo Chang[a], Clifford H. Spiegelman[b]

[a] *Environmental Engineering Program, Civil Engineering Department, University of Southern California, 3620 South Vermont Avenue, Los Angeles, CA 90089-2531, USA*

[b] *Department of Statistics, Texas A&M University, College Station, TX, USA*

Received 31 July 2001; accepted 17 January 2002

Abstract

The relationship of the concentration of air pollutants to wind direction has been determined by nonparametric regression using a Gaussian kernel. The results are smooth curves with error bars that allow for the accurate determination of the wind direction where the concentration peaks, and thus, the location of nearby sources. Equations for this method and associated confidence intervals are given. A nonsubjective method is given to estimate the only adjustable parameter. A test of the method was carried out using cyclohexane data from 1997 at two sites near a heavy industrial region in Houston, Texas, USA. According to published emissions inventories, 70% of the cyclohexane emissions are from one source. Nonparametric regression correctly identified the direction of this source from each site. The location of the source determined by triangulation of these directions was <500 m from that given in the inventory. Nonparametric regression is a powerful technique that has many potential uses in air quality studies and atmospheric sciences. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Air pollution; Statistics; Data analysis; Wind direction; Nonparametric regression

1. Introduction

The problem addressed here is estimating the wind direction that gives a local maximum in the observed average concentration of an atmospheric species, i.e., finding the directions of peaks in the concentrations. This direction is taken as the direction of the source, assuming that the source is not too distant. Finding the location of a nearby source is important in identification of the causes of local toxic "hot spots" and reconciliation of emission inventories observed concentrations, to give two examples. Somerville et al. (1996) present a classic parametric modeling approach to this problem. In this paper, an alternative nonparametric approach is taken. This approach is related to the kernel density counting procedure proposed by de Haan (1999) for certain air quality models. There is an enormous statistical literature on nonparametric regression. Härdle (1990) gives a very good introduction to this literature and the more practical aspects of the subject. As will be seen, nonparametric regression is a powerful, well-developed method with many possible applications to air quality studies and, indeed, atmospheric sciences in general.

It is very difficult to locate even a single strong peak accurately from a simple scatter plot of the data versus hourly resultant wind direction. This is seen in Fig. 1, a scatter plot of hourly concentrations of cyclohexane observed during 1997 at the Deer Park site near the Houston ship channel in Houston, Texas, USA, an area dominated by large refineries and petrochemical industries. The wind speed and direction were measured at the site. The wind direction is the azimuth angle (measured clockwise from north) that the wind is blowing from.

*Corresponding author. Tel.: +1-213-740-0596; fax: +1-213-744-1426.

E-mail address: rhenry@usc.edu (R.C. Henry).

1352-2310/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S1352-2310(02)00164-4

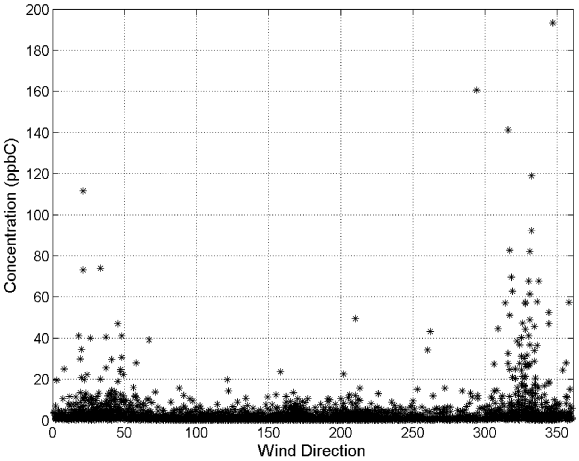

Fig. 1. Hourly cyclohexane measured at Deer Park during 1997 versus the azimuth of the wind direction.

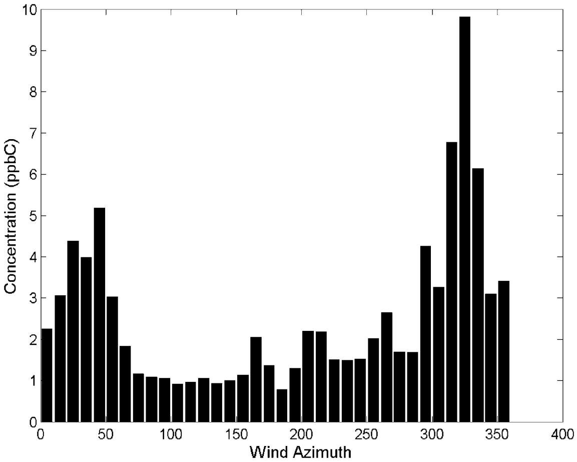

Fig. 2. Bar chart of average cyclohexane in 10° bins starting at zero.

The figure clearly shows high concentrations when the wind comes from between 0° and 50° and between 300° and 350°, but that is about all that can be said.

The usual method of analysis of the data in Fig. 1 is to group the data into bins of width $\Delta\theta$ based on wind direction and calculate the average concentration in each bin. The result can be displayed as a simple bar chart as in Fig. 2 or as a polar chart. Polar charts are not used here because small peaks are forced into an area near the origin, making them hard to see. In Fig. 2 the bins are 10° wide and start at 0°. When the wind speed is low, the direction is not well determined, consequently all hours with wind speed <1 mile/h were excluded from Fig. 2. It is now clear that the data possess several peaks, including a large peak around 330°; but a more precise estimate of the peak location is not possible for reasons discussed next.

Bar plots such as Fig. 2 have major limitations for the problem at hand. The location of the peaks is highly dependent on the choice of the bin size $\Delta\theta$ and the location of the boundaries of the bins. Peaks that are closer together than $2\Delta\theta$ may not be resolved and the location of a peak maximum cannot be estimated to better than $\pm\Delta\theta$, at best. This is less of a problem if $\Delta\theta$ can be made as small as a degree or two. With bins this small, however, almost always there are many bins in the less frequent wind directions that will have too few observations. In practice, even with hourly data for an entire year, bin size can seldom be made $<10°$.

In addition to the large peaks, several small peaks are seen in Fig. 2 at about 170°, 220°, 260°, and 290°. Which, if any, of these are real peaks and which are just random fluctuations in the data? Reliable error bars (or confidence intervals) would help answer these questions. For large peaks, confidence intervals would put bounds on the peak height and provide a measure of the error in peak location.

Based on the above discussion, a method is needed to estimate the location of large peaks more precisely and reliably separate peaks that are close to each other. The method should produce statistical confidence intervals. Finally, any parameters needed by the method, such as bin size in the bar chart, should be estimable by a reproducible, quantitative algorithm, to reduce the subjectivity of the analysis.

Such a method with all these properties exists and is known in the statistical literature as nonparametric regression. The following section will briefly introduce the method. This is followed by the application of the method to cyclohexane data from two sites in Houston, Texas. Cyclohexane is chosen to test the method because 70% of the industrial emissions in the region are known to be from a single source. Thus, the intersection of two lines drawn from each site in the direction of the largest peak should be the location of this source. The success of the method can be judged by how closely the predicted position of the source corresponds to the known location.

2. Nonparametric regression

2.1. Kernel estimators

One obvious improvement that overcomes some of the problems of a simple bar chart is to average over a sliding window of width $\Delta\theta$ centered at θ. Let the observed average concentration for the time period starting at t_i be C_i, where $i = 1, \ldots, n$ observations. Further, let the resultant wind direction for the ith time period be W_i, then the average concentration in the sliding window centered at θ is

$$\bar{C}(\theta) = N^{-1}\sum_{i=1}^{n} K(\theta - W_i)C_i, \tag{1}$$

where $K(x) = 1$, for $x - \Delta\theta/2 \leqslant x \leqslant x + \Delta\theta/2$, and zero otherwise, and N is the number of data points inside the sliding window. Fig. 3 shows the results of this method applied to the same data as Fig. 1. This is certainly an improvement over Fig. 2, but the curve in Fig. 3 is not very smooth, which makes determining peak locations difficult. The problem is that the function K gives equal weight to all the measurements inside the sliding window. A more reasonable approach would be to give less weight to observations near the edges, as shown next.

To generalize Eq. (1), it is important that $\Delta\theta$, also called the smoothing parameter, appear explicitly in the equation, so Eq. (1) is rewritten as

$$\bar{C}(\theta, \Delta\theta) = \frac{\sum_{i=1}^{n} K((\theta - W_i)/\Delta\theta)C_i}{\sum_{i=1}^{n} K((\theta - W_i)/\Delta\theta)}, \tag{2}$$

where $K(x) = 1$ for $-\frac{1}{2} \leqslant x \leqslant \frac{1}{2}$, and zero otherwise. Note that the denominator is simply a complicated way of writing N, the number of data points for which $\theta - \Delta\theta/2 \leqslant W_i \leqslant \theta + \Delta\theta/2$. In this form the equation can be generalized by taking $K(x)$ to be any continuous function of x such that

$$\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1. \tag{3}$$

There are many possible choices for K, two of the most often used are:

The Gaussian kernel

$$K(x) = (2\pi)^{-1/2}\exp(-0.5x^2), \quad -\infty < x < \infty, \tag{4}$$

and the Epanechnikov kernel

$$K(x) = 0.75(1 - x^2), \quad -1 \leqslant x \leqslant 1. \tag{5}$$

Both of these kernels will give maximum weight to observations near θ and less weight to observations further away. The major difference between the two is that the Gaussian kernel is defined over an infinite domain and the Epanechnikov kernel is defined over a finite range. For wind direction and other circular data the Gaussian kernel is preferred. For data limited to a finite range the Epanechnikov kernel has less bias at the end points and is preferred, and, under certain conditions, can be shown to converge to the true expected value at the optimal rate.

The primary nonparametric regression estimator for this paper is given by Eq. (2) with the Gaussian kernel. Technically, this is an example of a Nadaraya–Watson estimator, which is known to be consistent, that is, as sample size increases the value of the estimate will converge to the true value (Härdle, 1990, p. 25). Fig. 4 shows the result of using this estimator on the Deer Park cyclohexane data. The plot is much smoother than the moving average plot in Fig. 3. The gray region surrounding the curve in the figure is the 95% confidence interval, which will be discussed later.

The smoothing parameter for Fig. 4 was chosen to produce results somewhat comparable with the 10° bins in Fig. 2. To this end, we define the smoothing parameter in terms of the Full Width at Half Maximum (FWHM), an intuitive measure of the width of the kernel function. It is simply the full width of the peak in K measured at the point where the curve has fallen to half its value at the peak. For the Gaussian kernel, the

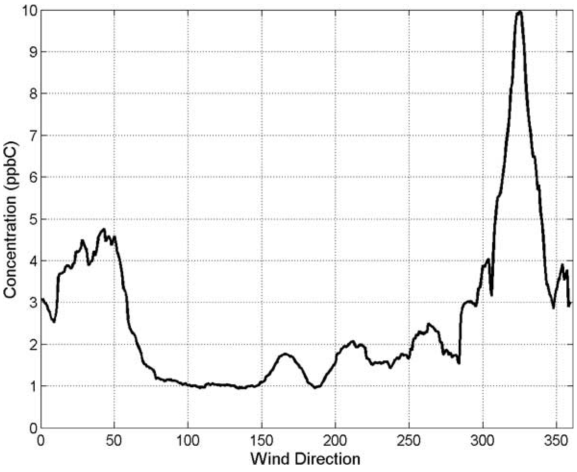

Fig. 3. Moving average cyclohexane concentrations calculated using a 10° wide sliding window.

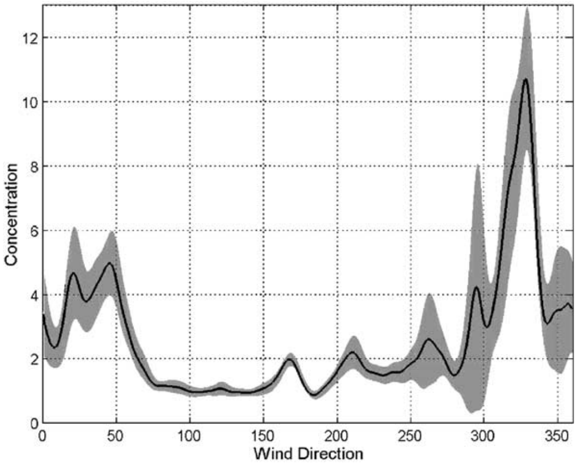

Fig. 4. Nonparametric regression of cyclohexane versus wind direction using a Gaussian kernel with a 10° FWHM. Data with wind speed < 1 mile/h are excluded. The gray region is the 95% confidence interval.

FWHM and smoothing parameter are related by

$$\mathrm{FWHM} = \sqrt{2}\Delta\theta. \quad (6)$$

Thus, since the FWHM is 10, the Gaussian smoothing parameter (usually called the standard deviation) for Fig. 4 is $10/\sqrt{2} = 7.07$. In the following, the FWHM will be used as a more intuitive surrogate for the smoothing parameter.

The connection of the smoothed data plot in Fig. 4 and the usual concept of regression is not immediately obvious. What justification is there for calling this form of data smoothing a regression? The answer is obvious if one writes the simple linear regression formula in an alternate way. The usual simple parametric regression model with one variable is

$$y = ax + b + \varepsilon, \quad (7)$$

where a and b are parameters to be estimated and ε is random error. However, statisticians often prefer to think of regression in terms of the expected value of y given x, using the standard notation the simple straight line regression equation becomes:

$$E(y|x) = ax + b. \quad (8)$$

In this way of looking at regression, Fig. 4 gives the estimated expected value of the concentration (y) given the wind direction (x), and thus can be thought of as a regression model without parameters.

2.2. Choosing the smoothing parameter

The most important decision in nonparametric regression is the choice of the smoothing parameter, or, equivalently, the FWHM. If the FWHM is too large the curve will be too smooth and peaks could be lost or not resolved. If it is too small the curve will have too many small, meaningless peaks dominated by noise or large peaks may resolve into false multiple peaks.

There are several ways to select the best smoothing parameter. This paper applies the cross validation method (Härdle, 1990, p. 152). For each observed wind direction W_j, $j = 1\ldots n$ and associated concentration $C_j = C(t_j)$ use Eq. (2) to estimate the expected concentration, but leaving out the jth observation, i.e.

$$\bar{C}_j(W_j, \Delta\theta) = \frac{\sum_{i\neq j} K((W_j - W_i)/\Delta\theta)C_i}{\sum_{i\neq j} K((W_j - W_i)/\Delta\theta)}. \quad (9)$$

The optimal smoothing parameter is the one that minimizes $V(\Delta\theta)$, the mean squared difference between concentration estimated leaving out one observation and the observed concentration:

$$V(\Delta\theta) = \sum_{j=1}^{n}(C_j - \bar{C}_j(W_j, \Delta\theta))^2. \quad (10)$$

For the Deer Park data in Fig. 4, the minimum of V occurs at a FWHM of 7. This is close to the value of 10 used in Fig. 4 but indicates that Fig. 4 may be somewhat over-smoothed.

2.3. Confidence intervals

The confidence intervals in Fig. 4 are calculated from formulae based on the asymptotic normal distribution of the kernel estimates. The sample estimate of the variance of the asymptotic distribution of $\bar{C}(\theta, \Delta\theta)$ is given by (Härdle, 1990, pp. 98–101)

$$s^2(\theta) = \frac{c_K\bar{\sigma}(\theta)}{n\Delta\theta\bar{f}(\theta)},$$

where

$$c_K = \int_{-\infty}^{\infty} K^2(x)\,\mathrm{d}x = \frac{1}{2\sqrt{\pi}}, \text{ for a Gaussian } K,$$

$$\bar{f}(\theta) = (n\Delta\theta)^{-1}\sum_{i=1}^{n} K\left(\frac{\theta - W_i}{\Delta\theta}\right),$$

and

$$\bar{\sigma}^2(\theta) = (n\bar{f}(\theta))^{-1}\sum_{i=1}^{n} K\left(\frac{\theta - W_i}{\Delta\theta}\right)(C_i - \bar{C}(\theta, \Delta\theta))^2. \quad (11)$$

Thus, if c_a is the $(100 - a)$-quantile of the normal distribution (1.96 for $a = 0.025$) the confidence limits on $\bar{C}(\theta, \Delta\theta)$ are given by

$$\bar{C}(\theta, \Delta\theta) \pm c_a s(\theta). \quad (12)$$

If $a = 0.025$, then the expression above gives a two-sided 95% confidence interval. This is shown as the gray shaded area of Fig. 4. From these confidence intervals, it is obvious that the small peak near 170 is real but the peaks near 260 and 290 are not. The peak near 220 is not an obvious call and requires further analysis.

2.4. Bias and serial correlation in nonparametric regression

This type of nonparametric regression has some obvious drawbacks, chief among these being bias. Since the data is being smoothed, the peaks usually will not be quite as high or sharp as in reality. This bias is an inevitable result of the smoothing. Bias can be estimated by simply using the output curve in Fig. 4 as the input to the nonparametric regression. The bias estimate is the difference between the twice-smoothed curve and the once-smoothed curve. This estimate for bias is called the plug-in bias estimate. Calculated this way, the bias in Fig. 4 is <10% at the peaks, and much less than this elsewhere. Because the bias is small in the examples considered here, it will be ignored in the rest of the paper.

Atmospheric concentration data, especially the hourly averages considered here, often have a high degree of serial correlation induced by the effects of meteorology and diurnal emission patterns. Fig. 5 shows the

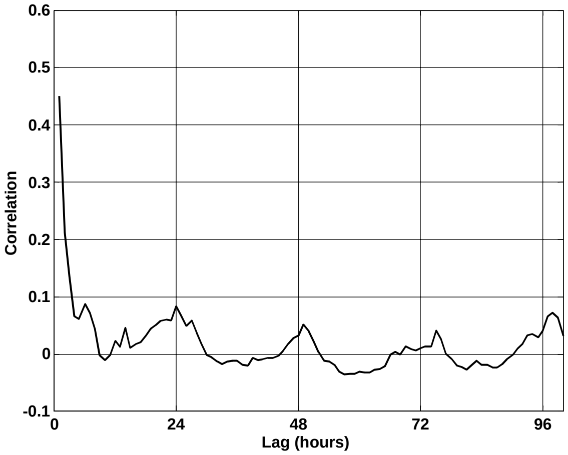

Fig. 5. Autocorrelation function for the hourly concentration data used in Fig. 4.

autocorrelation function for the concentration data used in Fig. 4. The correlation with the next hour is 0.45 and rapidly drops off rising again to 0.0835 at a lag of 24 h. The autocorrelation function for the residuals of the nonparametric regression in Fig. 4 is very similar as is the autocorrelation function for the Clinton Dr. data. This type of positive serial correlation may make the actual confidence intervals larger than those estimated by the methods of the previous section. Unlike bias, due to the fact that the data is not evenly spaced because only strong winds are used to estimate peaks, there is no simple way to estimate the effects of serial correlation on the confidence intervals. One must simply realize that the estimated confidence intervals are almost surely a bit too small. The possible effects of this on the results in this paper are discussed next.

In this paper, the confidence intervals are used in two ways; first, in a subjective way to judge if a peak is merely noise or not. A small peak that has small confidence intervals is likely to be a real peak, while a peak with large confidence intervals is likely to be noise. If the estimated confidence intervals are too small because of serial correlation, the worst result would be the possible acceptance of a peak as real when it is not. The second use of confidence intervals in this paper is to make a rough estimate of the uncertainty in the peak location by drawing a horizontal line at the peak. The intersection of this line with the upper confidence interval curve gives the lower and upper limits on the peak location. If the estimated confidence limits are too small, then the range for the peak locations will also be too small. The worst possible result is that the true source may lie outside where it is predicted to be. However, as shown below, this was not found to be a problem in the application of these methods to the Houston cyclohexane data, where the true locations of the peaks were known.

For this paper, which uses peak locations to determine the locations of an emission source, a more significant question is if serial correlation in the data increases bias in the estimate of the peak location. One can show that this is not the case and thus it has no effect on the estimates of source location that are the main subject of this paper and the following sections.

3. Application to 1997 Houston cyclohexane data

3.1. Data

Concentration data for cyclohexane and other volatile organic compounds (VOCs) for the year 1997 were obtained from the US Environmental Protection Agency's (EPA) Photochemical Assessment Monitoring Stations (PAMS) database. Hourly concentrations from an automated gas chromatograph from two sites were available. The two sites, Deer Park and Clinton Drive, are shown in Fig. 8 in context with nearby 1997 emissions of VOCs taken from the EPA's AIRS database. This database does not have emissions of individual species, unlike the Air Toxics Emissions Inventory. Extracted from the air toxics inventory, Table 1 lists all the emissions of cyclohexane for the year 1997 in Harris County, Texas, which includes the city of Houston and the Houston Ship Channel, an area of major petroleum refining and petrochemical industries. Table 1 shows that one company, Phillips Petroleum, is the source of almost 70% of the emissions in the inventory. Thus, one would expect that high concentrations of cyclohexane would be associated with wind directions at the sites coming from this facility. Lines drawn in the direction of the maximum cyclohexane concentrations from the two sites should intersect near the location given for the source in the inventory. The accuracy of the nonparametric regression can be judged by how close the estimated position is to the putative position.

3.2. Results

Figs. 6 and 7 are the result of nonparametric regression of cyclohexane on wind direction at the two sites using optimal FWHM values. Much of the wind data at the Clinton Drive site was missing. Thus, the wind direction and speed data from Deer Park were used in the analysis of both sites. The two sites are only 14.33 km apart and the terrain is very flat, so it is expected that the wind data at one site will serve for both. Fig. 6 is for Deer Park but it is not the same as Fig. 4 for two reasons. Only data with wind speed >6

Table 1
Emissions of cyclohexane for 1997 in Harris Co., TX

Facility name	Release type	Total release (lbs/year)	Percent of total	Latitude	Longitude	Accuracy (m)	Deer Park		Clinton Drive	
							Azimuth	Distance (km)	Azimuth	Distance (km)
Phillips Petroleum Co.	STACK	167 000	58.74	29.74167	95.17556	50	330.27	9.25	83.25	7.91
Phillips Petroleum Co.	FUGITIVE	33 000	11.61	29.74167	95.17556	50	330.27	9.25	83.25	7.91
Exxonmobil Baytown Refinery	STACK	15 282	5.38	29.73944	95.00694	80	56.33	14.05	88.33	24.15
Exxonmobil Baytown Refinery	FUGITIVE	3777	1.33	29.73944	95.00694	80	56.33	14.05	88.33	24.15
Enichem Americas Inc.	STACK	16 402	5.77	29.77194	95.01694	11000	43.24	15.65	79.44	23.56
Lyondell-Citgo Refinery	FUGITIVE	8472	2.98	29.71806	95.23000	50	298.79	11.23	123.13	3.11
Lyondell-Citgo Refinery	STACK	7509	2.64	29.71806	95.23000	50	298.79	11.23	123.13	3.11
Shell Chemical	STACK	7000	2.46							
Valero Refining Co.	STACK	6628	2.33	29.72333	95.25306	20	296.43	13.48	161.34	1.17
Valero Refining Co.	FUGITIVE	2323	0.82	29.72333	95.25306	20	296.43	13.48	161.34	1.17
Westhollow Tech. Center	STACK	6365	2.24	29.725	95.63333	11000	277.34	49.19	268.63	36.36
Millennium Petrochemical Inc.	FUGITIVE	4360	1.53	29.71389	95.06833	80	49.40	7.60	96.72	18.34
Crown Central Refinery	FUGITIVE	3594	1.26	29.72389	95.20833	50	308.00	9.84	102.60	4.81
Fmc Corp.	FUGITIVE	2576	0.91	29.6325	95.04140	80	116.11	9.33	118.25	23.65
Total emissions		284 288								

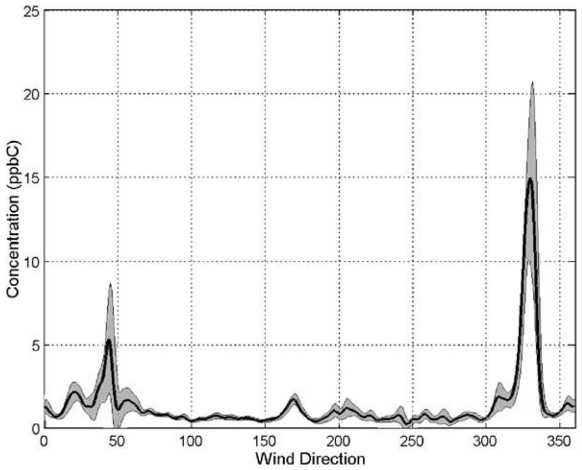

Fig. 6. Nonparametric regression of cyclohexane at Deer Park using a Gaussian kernel with a FWHM of 5. Data are restricted to periods with wind speed >6 miles/h (about 1 h travel time from the largest source to the site). The gray region is the 95% confidence interval.

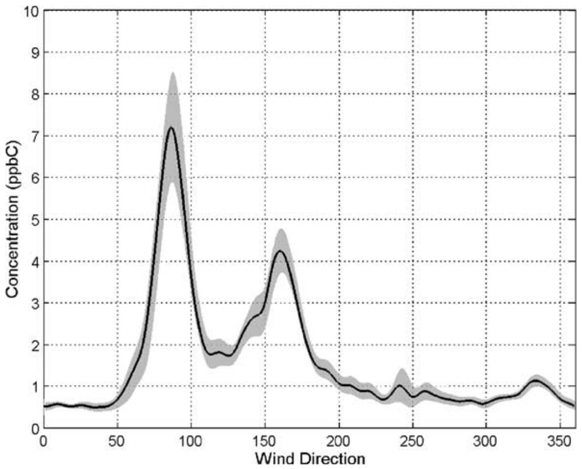

Fig. 7. Nonparametric regression of cyclohexane at Clinton Drive using a Gaussian kernel with a FWHM of 10. Data are restricted to periods with wind speed >5 miles/h (about 1 h travel time from the largest source to the site). The gray region is the 95% confidence interval.

miles/h (9.66 km/h) were used, instead of all data >1 mph as for Fig. 4. The optimal FWHM for data with wind speed >6 mph was calculated to be 5, significantly smaller than 10 used in Fig. 4. The data are restricted to periods with wind speed >6 mph because the Deer Park site is 9.25 km from the Phillips source and the clearest impact will be seen if the travel time from the source to the sampler is <1 h (Fig. 8). For the same reason, the data from Clinton Drive were restricted to hours where the wind speed is >5 miles/h (8 km/h) since the source is 7.91 km from the site. Using the known location of the source to restrict the data is not a case of unacceptable circular reasoning (using the source location to estimate the source location). Data for all wind speed >1 mph could have been used to estimate the location of the peaks and get an approximate location of the source. From this the approximate distance of the source to the sites can be estimated, which would lead to the same result without prior knowledge of the source location.

Table 2 gives the wind direction and the expected concentration of the four largest peaks in Figs. 6 and 7. The nonparametric regression calculations were carried out for each whole degree from 0 to 359. Thus, to get the entries in Table 2 it was necessary to interpolate to find the peak locations with greater precision than 1°. If the data are equally spaced with spacing h and local maximum $C(x_0)$, then the interpolated maximum is at $x_0 + ph$, where

$$p = \frac{(C(x_0 - h) - C(x_0 + h))}{2(C(x_0 - h) - 2C(x_0) + C(x_0 + h))}. \tag{13}$$

The interpolated maximum concentration is given by

$$C(x_0 + ph) \approx 0.5p(p-1)C(x_0 - h) + (1 - p^2)C(x_0) + 0.5p(p+1)C(x_0 + h). \tag{14}$$

Both these formulae are from Abramowitz and Stegun (1972).

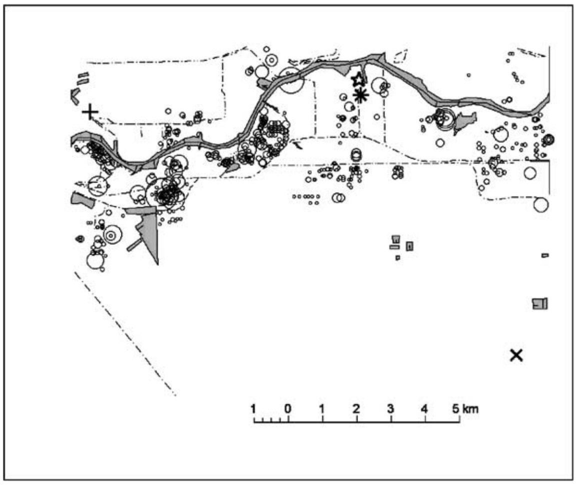

Fig. 8. The location of the Phillips Petroleum source in the inventory is shown as a star; the estimated location is marked as *. The Deer Park site is the × and the Clinton Drive site is the +. The gray shaded areas are water bodies. The circles are VOC sources with the area proportional to the annual emissions. Dash-dotted lines are railroads.

3.3. Comparison to known sources

The sources in Table 1 can now be compared with the peaks in Table 2. A measure of the uncertainty of the peak locations is helpful in this comparison. Table 2 gives very conservative ranges for the peak locations that are calculated by drawing a horizontal line through the peak and reading its intersection with the upper confidence boundary. Not surprisingly, at both sites the largest peak has an azimuth consistent with the location of the largest source in the inventory. For Clinton Drive, the second largest source in the inventory also lies in the azimuth range of the largest peak in Fig. 7. This may explain why this peak is so broad. The second largest peak for Clinton Drive is at 160, which corresponds well with Valero Refining in the inventory. Valero only accounts for 4.79% of the emissions but it is located only 1.17 km from the monitoring site. It seems reasonable to associate this peak with this source. The remaining two small peaks in Fig. 7 do not correspond to any source in Table 1. These could be emissions associated with nonindustrial sources such as roadways. At Deer Park, the location of the second largest peak in Fig. 6 corresponds closely with Enichem Americas in Table 1. However, the location of the source is given an accuracy of only 11 km, so one cannot definitely associate this peak with this source. The remaining two small peaks in Fig. 6 do not correspond to any sources in Table 1.

Table 2
Largest peaks in the nonparametric regression of cyclohexane on wind direction, Figs. 6 and 7

	Maximum	Deer Park		Maximum	Clinton Drive	
		Azimuth	Azimuth range		Azimuth	Azimuth range
Peak 1	14.953	329.12	325.64–332.68	7.197	86.43	80.56–92.76
Peak 2	5.391	43.72	40.68–46.96	4.251	160.04	153.90–166.21
Peak 3	2.197	21.60	15.51–25.01	1.147	332.63	326.37–340.20
Peak 4	1.775	168.89	165.87–171.44	1.027	240.95	235.12–248.86

3.4. Location of the largest source

The location of the Deer Park and Clinton Drive sites as given by the AIRS database are 29.6694°N, 95.1281°W, and 29.7333°N, 95.2569°W, see Fig. 8. Using these positions and the azimuth of the largest peaks the estimated location of the source is 29.7378°N, 95.1751°W. (All calculations involving longitude and latitude and distances were performed using functions from the Matlab Mapping Toolbox.) The distance between this and the location given in Table 1 is 0.436 km. As seen from the Deer Park site this distance corresponds to an error of 2.7°, or 3.2° as seen from Clinton Drive. The best that could be done from the bar chart in Fig. 2 would be to estimate the location to ±20°. It is fair to say that the nonparametric regression method is a major improvement and that the source location predicted by the nonparametric regression is in good agreement with the inventories. Now that this technique has been validated, it may be applied to other chemical species in the air that are not dominated by a single source. The results will help reconcile the emissions inventories to observed concentrations.

4. Discussion

In obvious extension of this method, the wind direction analysis presented here can be applied to the source contributions estimated from receptor models. The results should be sharper that using chemical species that usually have many sources. Previous work on comparison of the emissions inventory for the Houston ship channel relied on simple bar charts to determine the direction of sources (Henry et al., 1997). Later work of this type should benefit from the methods presented here. Nonparametric regression techniques could also be used in other air quality applications, such as trend analysis of time series or simply providing data smoothing for exploratory analysis of air quality data.

This work has demonstrated the usefulness of nonparametric regression of air quality data on wind direction. Nonparametric regression allows for accurate determination of the wind direction of maximum concentration. However, ground level concentrations are function of factors other than wind direction. For elevated sources, ground level concentrations can be a complex function of emission rates, wind speed and atmospheric stability. Indeed, scenarios could be constructed where the direction of the maximum average concentration does not correspond to the direction of the source. Such situations are probably rare, but a method that included other parameters such as wind speed would be desirable. Simultaneous nonparametric regression of concentrations on wind direction and wind speed is possible and could help throw light on, among other things, the distance to the sources and whether the sources are ground level or elevated. This will be the subject of a sequel to this paper. More can be done with the confidence intervals to determine if a peak is real or noise and to estimate the variance of the peak location.

References

Abramowitz, M., Stegun, I., 1972. Handbook of Mathematical Functions, Dover Publications, New York, p. 879 and p. 881.

de Haan, P., 1999. On the use of density kernels for concentration estimations within particulate and puff dispersion models. Atmospheric Environment 33, 2007–2021.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Henry, R.C., Spiegelman, C.H., Collins, J.F., EunSug Park, 1997. Reported emissions of organic gases are not consistent with observations. Proceedings of the National Academy of Sciences of the United States of America 94, 6596–6599.

Somerville, M.C., Mukerjee, S., Fox, D.L., 1996. Estimating the wind direction of maximum air pollutant concentrations. Environmentrics 7, 231–243.