Environ Ecol Stat (2006) 13:287–298
DOI 10.1007/s10651-004-0012-7

ORIGINAL ARTICLE

The effect of wind direction on ozone levels: a case study

Sreenivasa R. Jammalamadaka · Ulric J. Lund

Received: August 2003 / Revised: June 2004

Abstract This paper provides an illustrative case study on how the wind direction plays an important role in determining the ozone levels, in a suburb of Houston. Circular correlation and circular regression methods are used in the analysis and the primary goal is to illustrate how circular data analytic methods help in analyzing certain environmental issues.

Keywords Circular statistics · Directional data · Ozone level · Wind direction

1 Introduction

The wind direction, unlike many other linear variables such as the wind speed and ozone level, should be dealt with quite differently in statistical analyses. This is because a (2-dimensional) direction is a circular variable which can be represented as a point on the circumference of a circle. The numerical value assigned to it, and therefore, the corresponding statistical analyses based on such numerical values, depend on what is taken to be the zero-direction and whether a clockwise or anti-clockwise sense of rotation is used. Thus, in dealing with circular data like wind directions, one should be careful not to use the standard linear methods like the arithmetic mean and standard deviation, instead using measures and techniques which are independent of this arbitrary zero-direction and sense of rotation. The reader is referred to books such as Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001) for detailed descriptions of such methods. One of the aims of this paper is to increase such awareness by providing a case study that involves air pollutants and wind direction.

S. R. Jammalamadaka (✉)
Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA
e-mail: rao@pstat.ucsb.edu

U. J. Lund
Statistics Department, California Polytechnic State University, San Luis Obispo, CA 93407, USA
e-mail: ulund@calpoly.edu

2 Data source

The Texas Natural Resource Conservation Commission (TNRCC) and local monitoring programs in Texas jointly collected air pollutant and weather data from 1972 to 1996. However, 1995 marked the last year that average wind directions and speeds were measured as well as ozone concentrations. Since the aim of our case study is to analyze the effect of wind direction on ozone levels, we chose 1995 as our study year. The raw data was obtained from the Internet (Texas Natural Resource Conservation Commission 2003a).

The 1995 TNRCC hourly data consists of the following variables: monitoring site; date, and time on a 24 h scale; average ozone level, measured in parts per billion (ppb); average temperature, measured in degrees Fahrenheit; average wind direction, given in degrees clockwise from true north; average wind speed, in miles per hour.

There are 37 monitoring sites included in the 1995 TNRCC data set. Our focus was on the site located in Clute, Texas, due to its relative proximity to the Houston metropolitan area, and our interest in using circular statistics to assess the importance of wind speed and direction in transporting air pollutants. Figure 1 (Texas Natural Resource Conservation Commission 2003b) shows the location of the Clute monitoring site, labelled as C11 on the map. It is situated approximately 50 miles south of Houston.

Since, we had an abundance of data, we decided to omit hourly measurements for which either ozone or weather data was missing. This resulted in 7,608 hourly observations from January 1 to December 31, 1995, which we used in our analysis. Most analyses were done on the daily level. To this effect, the hourly data was averaged to obtain a daily data set as well, consisting of 317 observations for the year 1995. Daily average ozone level, temperature, and wind speed was computed using

Fig. 1 Local monitoring sites in Houston vicinity

the usual sample mean for all non-missing hourly data. Daily average wind direction was computed using the sample mean direction (defined below).

3 Descriptive statistics and graphs

Table 1 presents some descriptive statistics for the daily air quality data set. Sample means and standard deviations are given for the linear variables, ozone level, temperature, and wind speed. The sample mean direction and sample circular dispersion were computed for the circular variable, wind direction.

The sample mean direction is computed by treating all angular measurements as points on the unit circle, and computing the resultant vector of the unit vectors determined by the data points. The sample mean direction is the direction of this resultant vector, and the the sample mean resultant length provides a measure of concentration of the circular data. That is, for a sample of n observations of a circular variable α, if we define

$$S = \sum_{i=1}^{n} \sin \alpha_i,$$

$$C = \sum_{i=1}^{n} \cos \alpha_i,$$

$$R = \sqrt{C^2 + S^2}$$

then the sample mean direction $\bar{\alpha}$ is the quadrant specific inverse tangent of S/C,

$$\bar{\alpha} = \begin{cases} \arctan(S/C), & \text{if } C \geq 0, \\ \arctan(S/C) + \pi, & \text{if } C < 0, \\ \text{Undefined}, & \text{if } R = 0. \end{cases}$$

The sample mean resultant length is given by $\bar{R} = R/n$. One commonly used measure of circular dispersion, also given in Table 1, is the circular variance defined as $1 - \bar{R}$. This statistic falls in the interval $[0, 1]$, taking values close to 0 when the data is highly concentrated around one direction, and close to 1 for widely dispersed data.

Returning to Table 1, we see that summer months with high-average temperatures, generally also had high-average ozone levels, though July had an unusually low-ozone level. August and September saw the lowest average wind speeds, contributing perhaps to the summer months' high-ozone readings.

Rose diagrams were created to further explore these variables one at a time (univariate) as well as in pairs (bivariate). Figure 2 shows a rose diagram of the average daily wind directions. The winds predominantly came from the east, northeast, and southeast—onshore winds from the Gulf of Mexico.

When restricting to days on which the ozone level was above the 90th percentile, Fig. 3 shows that winds were most frequently coming from the northern two quadrants. Winds were generally coming from the south on days for which the ozone level was below the 10th percentile, as shown in Fig. 4. This may be evidence that indeed ozone was being transported from the Houston region, north of Clute.

Another view of the relationship between wind direction and ozone level is presented in Fig. 5, where a rose diagram of the average daily wind directions is labelled

Table 1 Descriptive statistics for daily data

	Month												
Variable	Annual	January	February	March	April	May	June	July	August	September	October	November	December
Sample size	317	19	12	27	30	27	30	30	31	30	30	20	31
Ozone level													
Mean	28.23	21.15	27.82	28.28	31.05	22.13	38.5	23.02	33.16	38.37	31.98	16.71	19.44
S.D.	15.99	8.04	11.67	9.86	12.08	11.53	18.69	15.9	21.19	21.58	13.24	9.71	6.18
Temperature (Deg Fahrenheit)													
Mean	71.76	59.11	61.66	61.05	68.12	77.67	80.05	83.64	83.24	80.12	70.79	63.47	58.28
S.D.	11.39	7.73	6.15	8.18	6.24	3.51	3.26	2.38	1.43	5.75	4.41	8.34	12.58
Wind speed (Mph)													
Mean	7.58	9.67	7.12	7.56	8.51	9.49	7.53	7.58	5.56	5.55	7.28	7.68	8.18
S.D.	2.90	4.32	2.39	2.36	2.99	2.78	2.71	3.12	1.74	1.63	2.49	2.66	2.58
Wind direction (Deg clockwise from North)													
Circ mean	91.22	256.57	82.41	45.89	106.22	134.36	129.28	203.13	79.17	46.91	49.61	65.39	67.44
Circ disp	0.70	0.79	0.60	0.70	0.43	0.17	0.65	0.62	0.54	0.47	0.41	0.61	0.82

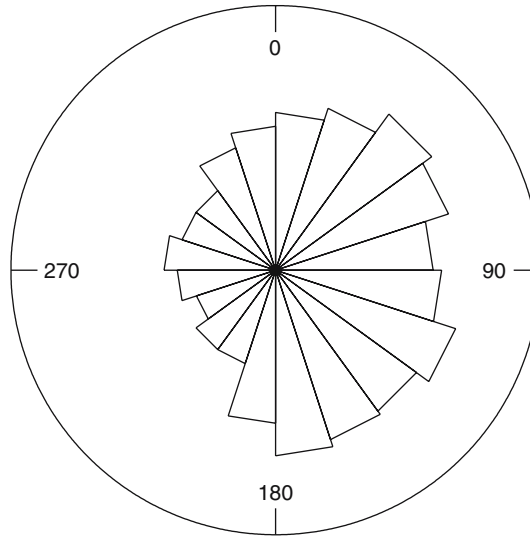

Fig. 2 Average daily wind direction

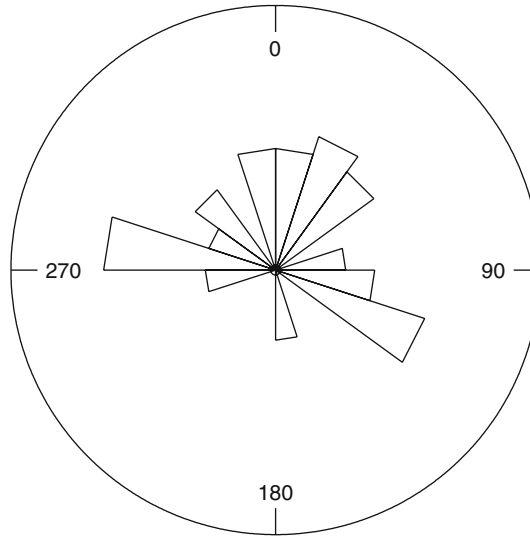

Fig. 3 Average daily wind direction, ozone above 90th percentile

with the average ozone levels. Again, we see that when the wind came from the south, the ozone levels were lower, on average, than when the wind came from the north.

Similar graphs display average temperature and average wind speed by wind direction (Figs. 6, 7). We see that warmer days were associated with winds coming from the south, and these winds were also stronger, on average.

To discern how strong the bivariate relationships were, we used several measures of correlation, depending on what type of variables were considered. That is, different measures of correlation are necessary depending on how many of the two variables are circular variables (none, one, or both). We will refer to these three cases as the linear–linear, linear–circular, and circular–circular cases. For instance, examining the correlation between temperature and ozone level was done using the usual Pearson's

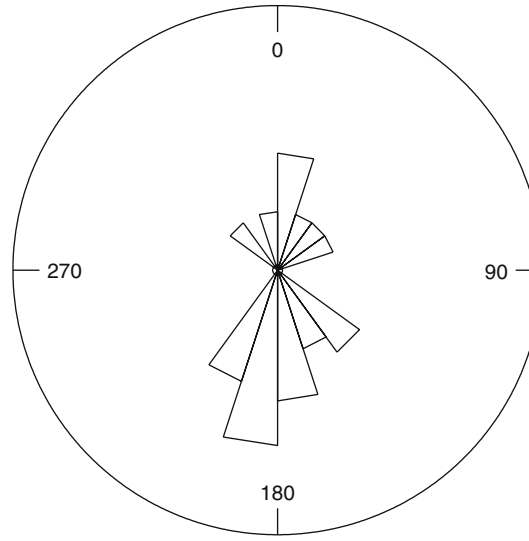

Fig. 4 Average daily wind direction, ozone below 10th percentile

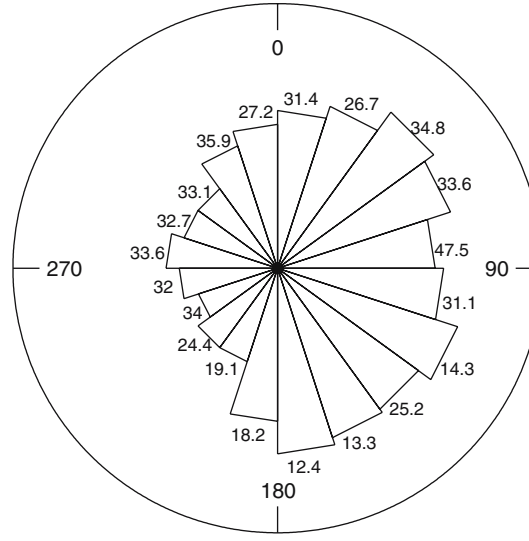

Fig. 5 Average ozone level by daily wind direction

correlation coefficient. However, correlation between ozone level and wind direction would necessitate a linear–circular statistic such as the one described in Mardia (1976).

Mardia defines this linear–circular correlation coefficient as the multiple correlation between a linear variable X and the sine and cosine of a circular variable α. The precise formula of the statistic is given by

$$r^2 = \frac{r_{xc}^2 + r_{xs}^2 - 2r_{xc}r_{xs}r_{cs}}{1 - r_{cs}^2},$$

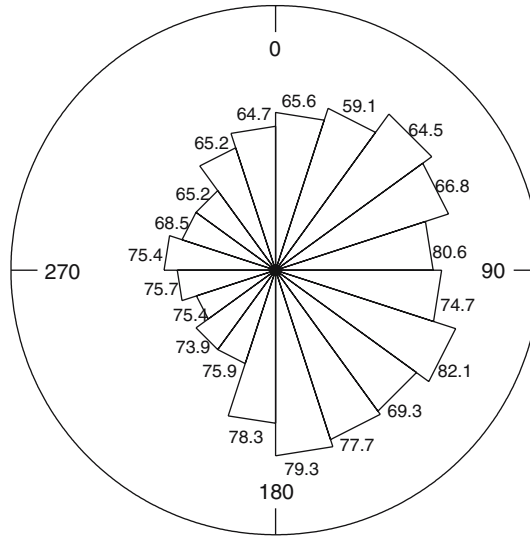

Fig. 6 Average temperature by daily wind direction

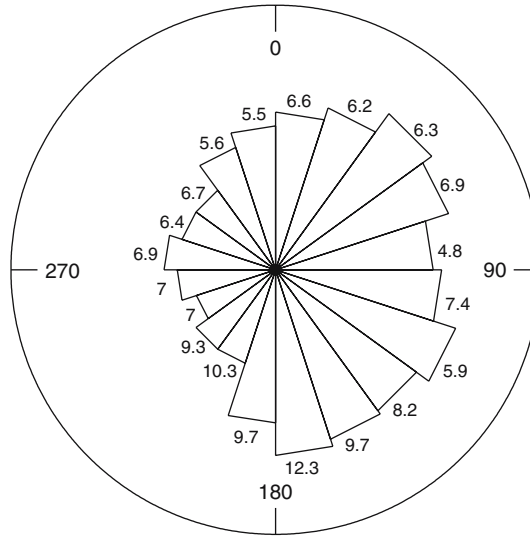

Fig. 7 Average wind speed by daily wind direction

where

$$r_{xc} = \mathrm{corr}(x, \cos\alpha),$$
$$r_{xs} = \mathrm{corr}(x, \sin\alpha),$$
$$r_{cs} = \mathrm{corr}(\cos\alpha, \sin\alpha).$$

Two temporal variables indicating day of year and month of measurement were also coded and converted into circular variables. This was achieved by taking the day of the year and multiplying by $2\pi/365$, and taking month of the year and multiplying by $2\pi/12$. One measure of the association between wind direction and day or month would be the circular–circular correlation (Jammalamadaka and Sarma,

1988; Jammalamadaka and SenGupta 2001, Sect. 8.2) between wind direction and the circular versions of day or month.

In general, the formula for the circular–circular correlation coefficient for n pairs of observations of two angular variables, α and β, with sample mean directions $\bar{\alpha}$ and $\bar{\beta}$, is given by

$$r_{c,n} = \frac{\sum_{i=1}^{n} \sin(\alpha_i - \bar{\alpha}) \sin(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^{n} \sin^2(\alpha_i - \bar{\alpha}) \sin^2(\beta_i - \bar{\beta})}}.$$

If α and β are independent, we would expect a value of $r_{c,n}$ close to 0. If the two angular variables are rotationally dependent, then we would expect a measure of $r_{c,n}$ close to ± 1.

Table 2 presents a summary of the correlations computed for the average daily measurements. Temperature seems to have been the most seasonal variable, as exhibited by the strong linear–circular correlation between temperature and day/month. Ozone

Table 2 Bivariate correlation coefficients

	Ozone	Temperature	Wind speed	Wind direction
Temperature	0.17[a]	–	–	–
Wind speed	−0.38[a]	0.02[a]	–	–
Wind direction	0.12[b]	0.20[b]	0.29[b]	–
Month	0.05[b]	0.67[b]	0.09[b]	−0.28[c]
Day	0.06[b]	0.71[b]	0.08[b]	−0.28[c]

[a] Linear–linear correlation (Pearson's correlation coefficient)
[b] Linear–circular correlation (Mardia 1976)
[c] Circular–circular correlation (Jammalamadaka and Sarma, 1988)

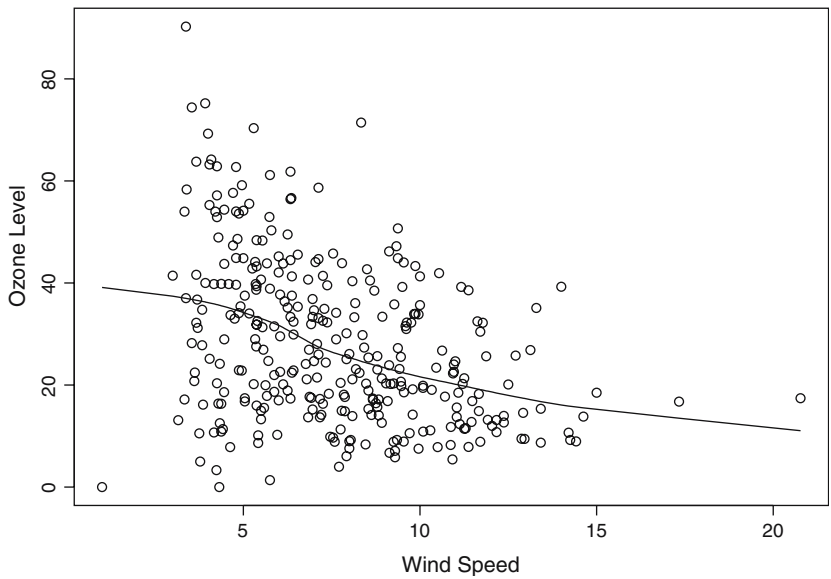

Fig. 8 Ozone level by wind speed; scatter plot *smoothing line*

was most strongly correlated with wind speed. The negative correlation coefficient and the scatter plot smoother shown in Fig. 8 indicate that lower ozone levels were associated with higher wind speeds. We saw above that stronger winds generally came from the south. Therefore, the negative correlation between wind speed and ozone level may again be due to the fact that Houston is located to the north of Clute.

The relatively low correlations between wind direction and day/month are not surprising if we recall that the statistic computed measures the degree that two variables are related by a rotation, and we don't expect these variables to have such a relationship, but rather a seasonal fluctuation.

4 Regression models

Two regression models were fitted using the Texas air quality data. One model used the daily measurements to predict ozone level from the remaining variables in the data set: temperature, wind speed, wind direction, month of measurement. This model had a linear response variable and a combination of two linear and one circular predictor variables. To also illustrate a regression model in which the response variable is circular, we chose to use the hourly data and modeled wind direction as a function of time of day.

Initially, a multiple linear regression model was used to predict ozone level from wind direction. This was a multiple linear regression, because to accommodate the circular covariate wind direction, we included the sine and cosine of wind direction as predictor variables, instead of the wind direction variable itself. A graph of ozone level versus wind direction, along with the fitted regression line, is given in Fig. 9. Again, we can see that ozone levels were lowest on average when winds were coming from the southern quadrants (around π radians). However, there is still a substantial

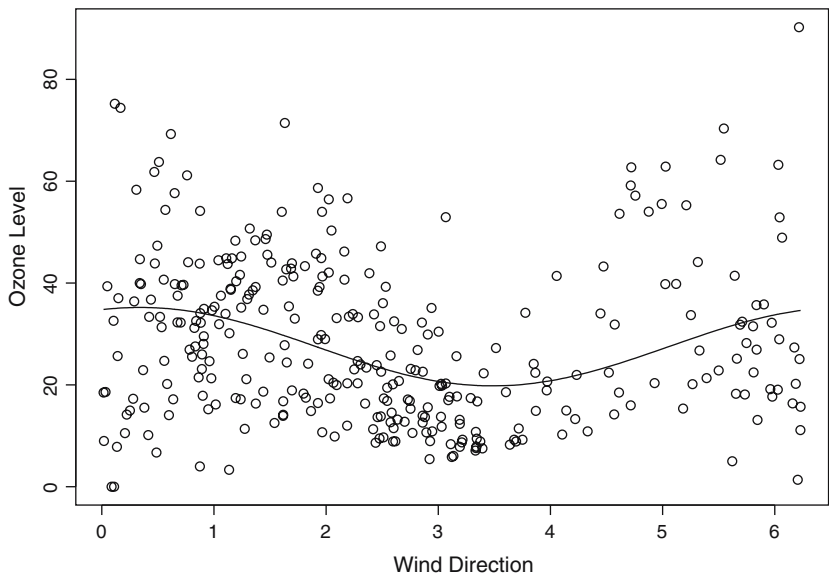

Fig. 9 Ozone level by wind direction; multiple *linear regression line*

Table 3 Multiple regression predicting ozone level

Coefficient	Value	Std. error	*t*-value	*p*-value
Intercept	−25.91	11.18	−2.32	0.0211
Temperature	0.88	0.15	5.75	0.0000
Wind speed	−1.18	0.33	−3.61	0.0004
sin(wind direction)	2.74	1.23	2.22	0.0273
cos(wind direction)	11.30	1.60	7.05	0.0000
sin(month)	7.23	1.76	4.11	0.0001
cos(month)	2.75	1.85	1.49	0.1376

amount of unexplained variation ($R^2 = 0.12$) in the data, and we also see evidence of heteroscedasticity in this model fit.

By adding the other covariates, temperature, wind speed, and the seasonal component month, the amount of explained variation is more than doubled ($R^2 = 0.31$), yielding a significantly improved model fit ($p < 0.001$). The estimated regression coefficients for the model are given in Table 3. Here, we see that increased temperatures were associated with higher ozone levels, and higher wind speeds were associated with lower ozone levels. The interpretation of the regression coefficients for the sine and cosine terms are not as transparent as for the linear predictor variables.

Next, for the purpose of illustrating the circular–circular regression model due to Jammalamadaka and Sarma (1993) (see also Jammalamadaka and SenGupta 2001, Sect. 8.6), we shifted through the data set to find a combination of circular variables that exhibited reasonable level of association. There was a strong relationship between hour of the day and wind direction during the days of June 7 through 10, 1995. For this subset of hourly data, the circular–circular correlation coefficient between wind direction and time of day was −0.67.

The regression model, we are considering here essentially fits two trigonometric polynomials predicting the sine and cosine of the response variable, in this case wind direction. If we let θ denote the observed wind direction, and let α denote the hour the observation was taken, then the mth-order trigonometric polynomial predicting wind direction from time of day would fit the following two multiple linear regression models:

$$\cos(\theta) = \sum_{k=0}^{m} (A_k \cos k\alpha + B_k \sin k\alpha) + \epsilon_1,$$

$$\sin(\theta) = \sum_{k=0}^{m} (C_k \cos k\alpha + D_k \sin k\alpha) + \epsilon_2.$$

The appropriate order m of trigonometric polynomial is determined by testing the significance of the $m + 1$st terms in both regression models. If neither model requires the $m + 1$st terms, mth-order models are used.

For example, the first and second-order models' regression coefficients are shown in Table 4, along with their p-values for assessing the need for higher order models. We can see that third-order models are not necessary ($p = 0.3314$ and 0.5045), whereas there was a benefit to the second-order model versus the first ($p = 0.0156$ and 0.0342). A graph of the data along with the first and second-order models is presented in Fig. 10.

Table 4 Circular-circular regression predicting wind direction

	Response	
Coefficient	cos(wind direction)	sin(wind direction)
Intercept	−0.8298	0.5317
First-order terms		
cos(hour)	0.0415	0.0587
sin(hour)	−0.08954	−0.1335
p-value for second-order	0.0156	0.0342
Second-order terms		
cos(hour)	−0.0055	0.0009
sin(hour)	−0.0255	−0.0371
p-value for third-order	0.3314	0.5045

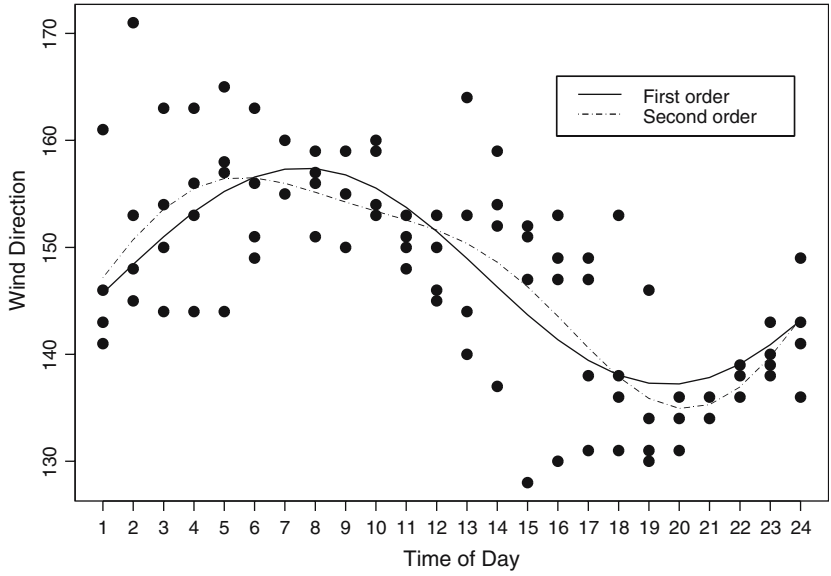

Fig. 10 Wind direction by time of day (*hour*); circular–circular regression lines

5 Conclusion

Using published data on air pollutants and weather from the TNRCC, it is shown here that the wind direction plays a significant role in how much ozone is transported from Houston, Texas to one of its neighboring towns. This case study illustrates the power of circular data analytic tools in such environmental studies.

All calculations and graphs presented herein were obtained using S-PLUS 6.1 with the aid of the circular statistics library CircStats. CircStats is available for download from the Internet (Lund 2003), and also is enclosed with the text by Jammalamadaka and SenGupta (2001).

References

Jammalamadaka SR, Sarma YR (1988) A correlation coefficient for angular variables. In: Matusita K (ed) Statistical theory and data analysis II. North Holand, Amsterdam, pp 349–364
Jammalamadaka SR, Sarma YR (1993) Circular regression. In: Matusita K (ed) Statistical theory and data analysis. VSP, Utrecht, pp 109–128
Jammalamadaka SR, SenGupta A (2001) Topics in circular statistics. World Scientific Publication, NJ
Lund UJ (2003) Internet location: http://statweb.calpoly.edu/lund. Site accessed August, 2003
Mardia KV (1976) Linear-circular correlation and rhythmometry. Biometrika 63:403–405
Mardia KV, Jupp PE (2000) Directional statistics. Wiley, New York
Texas Natural Resource Conservation Commission. (2003a) Internet location: http://www.tnrcc.state.tx.us/air/monops/Data.html. Site accessed August, 2003
Texas Natural Resource Conservation Commission, (2003b) Internet location: http://www.tnrcc.state.tx.us/cgi-bin/monops/select_summary?region12.gif. Site accessed August, 2003

Biographical sketches

S. Rao Jammalamadaka is Professor of Statistics at the University of California, Santa Barbara. He received his Ph.D. in statistics in 1969 from the Indian Statistical Institute, working with C.R. Rao, and held faculty positions around the world before joining the University of California in 1976. His current research interests include non-parametric statistical inference, limit distribution theory and asymptotic efficiencies of test procedures, directional data analysis, and goodness of fit tests. He has written books on the topics of circular statistics, introductory statistics, and linear models.

Ulric J. Lund is Assistant Professor in the Department of Statistics at California Polytechnic State University, San Luis Obispo. He did his graduate studies at the University of California, Santa Barbara under the direction of S. Rao Jammalamadaka, receiving his Ph.D. in statistics in 1998. After a visiting faculty position at Western Washington University in 1999, he worked as a statistical consultant in the San Francisco area for several years, specializing in environmental and public health risk analyses. His interests lie in circular statistics, statistical computing, and statistical analysis of public health and accident related data.