SINUSOIDAL MODELING APPLIED TO SPATIALLY VARIANT TROPOSPHERIC OZONE AIR POLLUTION

BY

NICHOLAS Z. MULLER and PETER C. B. PHILLIPS

COWLES FOUNDATION PAPER NO. 1242

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

2008

http://cowles.econ.yale.edu/

ENVIRONMETRICS
Environmetrics 2008; **19**: 567–581
Published online 5 December 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.897

Sinusoidal modeling applied to spatially variant tropospheric ozone air pollution

Nicholas Z. Muller[1*,†] and Peter C. B. Phillips[2,3,4]

[1]*Department of Economics, Environmental Studies Program, Middlebury College, 303 College Street, Middlebury, VT 05753, U.S.A.*
[2]*Cowles Foundation, Yale University, U.S.A.*
[3]*University of York, U.K.*
[4]*University of Auckland, New Zealand*

SUMMARY

This paper demonstrates how parsimonious models of sinusoidal functions can be used to fit spatially variant time series in which there is considerable variation of a periodic type. A typical shortcoming of such tools relates to the difficulty in capturing idiosyncratic variation in periodic models. The strategy developed here addresses this deficiency. While previous work has sought to overcome the shortcoming by augmenting sinusoids with other techniques, the present approach employs station-specific sinusoids to supplement a common regional component, which succeeds in capturing local idiosyncratic behavior in a parsimonious manner. The experiments conducted herein reveal that a semi-parametric approach enables such models to fit spatially varying time series with periodic behavior in a remarkably tight fashion. The methods are applied to a panel data set consisting of hourly air pollution measurements. The augmented sinusoidal models produce an excellent fit to these data at three different levels of spatial detail. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: air pollution; regional variation; semi-parametric model; sinusoidal function; spatial-temporal data; tropospheric ozone

JEL CLASSIFICATION: C22 ; C23

1. INTRODUCTION

Models based on sinusoidal functions can adequately fit time series that exhibit strong periodic behavior (Bloomfield, 2000). However, such models usually encounter difficulties emulating time series with cyclical behavior that deviates from a fixed periodic structure (Lewis and Ray, 1997). In such cases, some alternative approaches have been proposed to augment sinusoidal models to improve sample period fit and prediction. For instance, Campbell and Walker (1977) employed a model that includes both a deterministic sinusoid and a second-order autoregressive component to describe annual lynx trappings. Dixon and Tawn (1998) constructed a model of sea-level estimation that consists of a

*Correspondence to: N. Z. Muller, Department of Economics, Environmental Studies Program, Middlebury College, 303 College Street, Middlebury, VT 05753, U.S.A.
†E-mail: nmuller@middlebury.edu

Received 27 September 2007
Accepted 20 October 2007

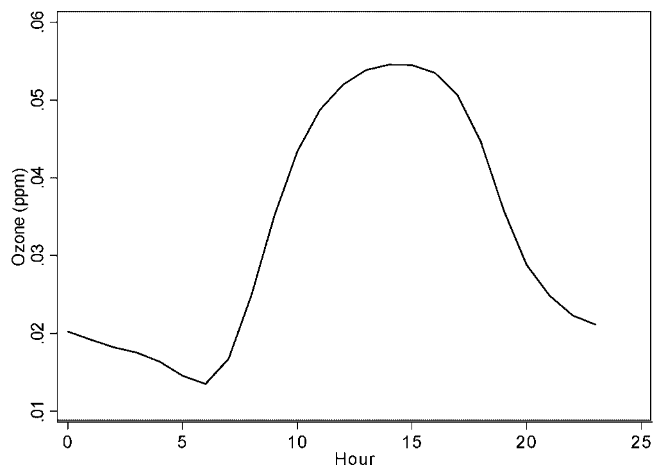

Figure 1. Diurnal ozone cycle

sinusoidal component governing tidal oscillations, a linear model capturing long-term trends, and weather-dependent model to estimate surge.

The present paper develops a new set of statistical tools that are designed to model spatially varying time series which displays some systematic periodic behavior and also manifests characteristics that are station-specific to individual locations. The methodological innovation is to use sinusoidal functions to represent spatiotemporal variation in a semiparametric manner. The technique involves first fitting a finite linear combination of sinusoidal functions to capture the spatially common periodic features of a certain series. This common periodic element may be regarded as parametric and will usually be quite parsimonious. Once this parametric model of common features is determined, it is augmented with a nonparametric component to model idiosyncratic local spatial features, again using sinusoidal functions in the form of a sieve approximation (e.g., Grenander, 1981). This nonparametric model is fitted using local residuals from the common model. Combining the nonparametric and parametric components into a single semiparametric framework provides a mechanism for capturing elements of common variation in spatiotemporal behavior while having the flexibility to emulate a substantial degree of local variation. The advantages of this approach are two-fold. First, the initial sinusoidal specification extracts the common near-periodic element in a complex spatiotemporal process using just a few parameters. Second, the nonparametric component tailors the more rigid common periodic structure to local patterns of variation. This approach resolves a principal drawback of sinusoidal modeling that is cited in the literature (lack of flexibility) and enables the investigator to find common elements of spatiotemporal variation in the data in a parametric manner that increases statistical efficiency. The new approach appears to have broad applicability to spatiotemporal data that manifests some common periodicity but substantial local variations about the common cycle.

We apply this machinery to a panel data set consisting of air pollution measurements in the contiguous United States. Specifically, the data involves measurements of tropospheric ozone (O_3) from the U.S. Environmental Protection Agency's (USEPA) air pollution monitoring network (USEPA 1). This common pollutant exhibits a characteristic unimodal diurnal shape when plotted against the hours in a day (Seinfeld *et al.*, 1998) (see Figure 1). To this daily structure we fit the models outlined above. The modeling approach adopted is well suited to this statistical problem and its various policy applications.

DOI: 10.1002/env

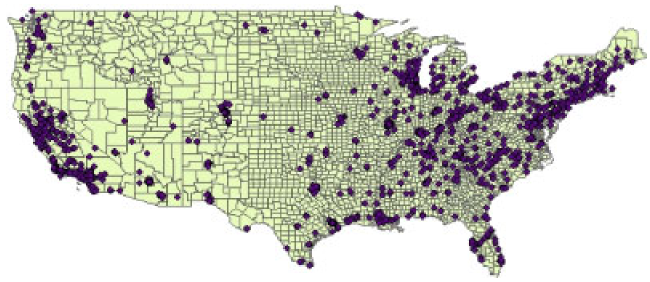

Figure 2. Ozone monitor locations

First, hourly measurements of O_3 do exhibit a fairly regular periodic structure, which suggests that a parametric sinusoidal fit will be generally well suited to the data. Additionally, the specific shape of the time series variation itself varies widely across space. These data therefore provide a suitable context for the application of our semiparametric approach. Second, the O_3 data set is a rich collection of nearly 4 million observations collected in 1996, providing an interesting spatiotemporal setting to test the performance of these new tools.

Finally, this application is in an area of immediate policy relevance. Since tropospheric O_3 produces a variety of deleterious effects on human health (Bell *et al.*, 2004) and welfare, the USEPA has designated O_3 as a criteria air pollutant. This classification stipulates that O_3 is subject to hourly measurement in order to assess regulatory compliance across both time and space. The network of monitors calibrated to measure O_3 consists of scattered observations (see Figure 2). The incomplete spatial coverage of this network has motivated prior efforts to interpolate O_3 readings (Guttorp *et al.*, 1994; Hopkins *et al.*, 1999). These efforts have focused on daily or seasonal average and maximum O_3 concentrations. However, the entire cycle matters because air quality standards have shifted from a 1 h daily maximum structure, adequately described by the previous interpolation methods, to focusing on the maximum 8 h average. The 8 h standard is a moving average. Assessing compliance therefore requires knowledge of the 24 h range of O_3 levels. In response to this shift in policy structure, we work toward a method of spatial interpolation that enables one to predict the entire daily O_3 cycle at points between pollution monitors. This facility would clearly improve the USEPA's ability to make inferences about compliance with the current O_3 standards in locations without measurements.

The method developed in this paper demonstrates the ability of parsimonious sinusoidal models to fit the spatiotemporal O_3 cycle. To use these tools for interpolation purposes, the researcher would begin by employing the spatially-common, regional model to fit the general shape of the O_3 cycle in the area where the prediction is to be made. (This paper shows that parsimonious regional models fit the state-averaged O_3 periodic cycle remarkably well.) In order to make a prediction at a specific location, the next step would involve estimating the amplitude parameters in the model capturing local variability. This could be accomplished using a sample of (perhaps nearby) measurement points where the local model has been estimated and its predictions calibrated against observed data. Thus, once the appropriate regional model is estimated, the task is to interpolate the amplitude parameters of the idiosyncratic model. The interpolated parameters would constitute the idiosyncratic model for the interpolation point, which would be combined with the regional model additively in order to predict the O_3 cycle at the point of interest.

There are numerous strategies to interpolate the idiosyncratic model parameters. For instance, one might regress a collection of parameters from a sample of fitted local models on factors associated

DOI: 10.1002/env

with the O_3 cycle such as temperature, emission intensity, and land use. The predicted values from these regressions would serve as estimates of the idosyncratic model parameters for the unmeasured interpolation point. Further, by interpolating the O_3 cycle at each county in the U.S., the researcher could construct a complete surface of O_3 estimates for the entire U.S. We leave this task to future research.

The results reported herein reveal that, using a sample of locations, the semiparametric modeling methodology fits the observed data in a remarkably tight fashion. In a sample of 10 states, the parametric model deviates from the state average daily O_3 cycle by 1–3%. Using a sample of 10 counties, the semiparametric model generates mean proportional errors of less than 1%. The model also generates an equally close fit to observations of the O_3 cycle at a sample of individual monitors. Further, a series of formal tests provides statistically significant evidence of spatially-variant idiosyncratic processes contributing to the daily O_3 cycle.

2. METHODS

2.1. The model

The parametric model (1) comprises a linear combination of sinusoidal functions and is intended to provide a general representation of the spatiotemporal data over the diurnal cycle:

$$O^s_{t,d} = \beta^s_0 + \sum_{r=1}^{r^*} \left(\beta^s_r \cos(2\pi t\Phi_r)\delta_{tr} + \beta^s_r \sin(2\pi t\Phi_r)\delta_{tr}\right) + \varepsilon_{t,d} \tag{1}$$

where $O^s_{t,d}$ is the Ozone concentration in state (s), for day (d), and hour (t), r is the hourly range, r^* is the number of sinusoidal functions in model (1), δ is Kronecker delta which equals 1 (for $t \in r$), β^s_r is amplitude parameter, for state (s), and hourly range (r), Φ_r is phase parameter, for hourly range (r), and $\varepsilon_{t,d}$ is the stochastic disturbance term.

While Equation (1) is used to model the basic diurnal cycle in series with a strong periodic signature, the general model also allows for some local regional/monitor heterogeneity by means of a nonparametric component which captures variation around the diurnal pattern embodied by Equation (1). In particular, the idiosyncratic process at location (c) is modeled in Equation (2) as a linear combination of sinusoidal functions fitted to the hourly residuals ($\hat{\varepsilon}^c_{t,d} = \hat{O}^s_{t,d} - O^c_{t,d}$), where $\hat{O}^s_{t,d}$ are the hourly predictions from Equation (1), and $O^c_{t,d}$ are hourly observations at location (c) across days (d)[‡]. The model (2) is intended as a trigonometric sieve approximation that approximates the specific (or idiosyncratic) characteristics at location c:

$$\hat{\varepsilon}^c_{t,d} = \gamma^c_0 + \sum_{R=1}^{R^*} (\gamma^c_R \cos(2\pi t\Phi^c_R)\delta_{tr} + \gamma^c_R \sin(2\pi t\Phi^c_R)\delta_{tr}) + u^c_{t,d} \tag{2}$$

where γ^c_R is amplitude parameter, county (c), and hourly range (R), Φ^c_R is phase parameter, and $u^c_{t,d}$ is stochastic disturbance term.

[‡]Guttorp *et al*. (1994) analysis of ozone data examines the residuals due to fitting site-specific AR(2) models to ozone readings. The residuals are then used to estimate the correlation structure among a cluster of sites.

DOI: 10.1002/env

In applications, r^* will usually be small and so the parametric component (1) is a parsimonious representation of the common periodic signature in the series across spatial locations, while R^* will generally be larger so that the component (2) better approximates the individual nonparametric form at location c. In our practical implementation, we find that good approximations are obtained for R^* in the region of 7–10. It is likely that the smoothing parameter R^* will show a broader range of values as the models are applied to more locations. The complete model (3) is therefore semiparametric and incorporates both the parametric part (1) and the nonparametric part (2) to model the O_3 data over time and at different locations:

$$
\begin{aligned}
O_{t,d}^{c} = \beta_0^s + \sum_{r=1}^{r^*} \left(\beta_r^s \cos(2\pi t \Phi_r)\delta_{tr} + \beta_r^s \sin(2\pi t \Phi_r)\delta_{tr}\right) + \gamma_0^c \\
+ \sum_{R=1}^{R^*} \left(\gamma_R^c \cos(2\pi t \Phi_R^c)\delta_{tr}\right) + \left(\gamma_R^c \sin(2\pi t \Phi_R^c)\delta_{tr}\right) + \upsilon_{t,d}^c
\end{aligned}
\tag{3}
$$

where $\upsilon_{t,d}^c$ is a stochastic disturbance term.

2.2. Estimation

(I) In the first stage, estimation focuses on setting appropriate values for the fixed phase parameters, Φ_r, and the number of sinusoidal functions, r^*, that are used in model (1). Once the phase parameters (Φ_r) and the order parameter (r^*) have been identified, the remaining statistical problem of estimating Equation (1) reduces to a linear least squares regression to calculate the amplitude parameters (β_r^s). This step-wise approach is advocated by Damsleth and Spjotvoll (1982). An alternative approach, not pursued here, is to jointly estimate the phase and amplitude parameters by nonlinear regression and use model selection methods to determine the order parameter r^*. In order to determine the number of sinusoidal functions r^* in Equation (1), our approach is to visually inspect the O_3 cycle in the data[§] and find a value of the order parameter that is sufficient to provide a good representation of the diurnal pattern. (Later, we use a similar approach for the determination of R^* in Equation (3)). For the present data set, we found that a value of $r^* \simeq 6$ worked very well. Turning to the phase parameters, we use an automated, iterative approach that tests a range of values for Φ_r on each segment of Equation (1). Both the sine and cosine functions are tested in each segment. We assessed the accuracy of the predicted ($\hat{O}_t^s$) for each segment corresponding to each (Φ_r) value. An algorithm chooses the value of Φ_r that corresponds to the

[§]This visual inspection approach is also suggested by Damsleth and Spjotvoll (1982).

DOI: 10.1002/env

minimum root mean squared error ($\sqrt{\text{MSE}_r}$)[‖] for each hourly segment (r):

$$\sqrt{\text{MSE}_r} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{O}_t^s - O_t^s\right)^2} \tag{4}$$

As an additional diagnostic, we plot the predicted O_3 segments along with the measured O_3 hourly segments against time. This visual inspection provides an important final verification of the choice of Φ_r.

(II) The second stage of estimation solves the least squares minimization problem on the amplitude parameters (β_r^s) using the fixed phase parameters Φ_r identified in stage 1:

$$\min_{\beta^s} S(\beta^s) = \sum_{r=1}^{6}(O_{t,d}^s - \beta_0^s - (\beta_r^s \cos 2\pi t\Phi_r)\delta_{tr} - (\beta_r^s \sin 2\pi t\Phi_r)\delta_{tr})^2 \tag{5}$$

which completes estimation of the parametric model.

(III) In order to tailor the parametric models to capture local behavior patterns, we need to estimate idiosyncratic effects for each locality. This is accomplished by calculating the residuals ($\hat{\varepsilon}_{t,d}^c$) obtained from fitting the state-level model (1) to local O_3 cycles. Then, in manner analogous to stage 1, we visually inspect plots of the ($\hat{\varepsilon}_{t,d}^c$) against time in order to determine a suitable order parameter R^*, the number of sinusoidal components in model (3). In order to accommodate heterogeneous O_3 cycles, residual plots from various regions of the contiguous U.S. are inspected. Markedly different patterns in the residuals necessitate spatially-variant values for the (Φ_R^c). Distinct (Φ_R^c) are identified for the Southeastern states, as well as those in the Midwest, the West and the Northeast. Additionally, in Northeastern and Western states, different phase parameters are specified for the models applied to large urban areas. This spatially nonparametric approach enhances the ability of model (3) to capture local variation in the time series structure.

(IV) Once suitable order parameters are obtained, we estimate the coefficients (γ_R^c) in Equation (2) using ordinary least squares:

$$\min_{\gamma^c} S(\gamma^c) = \sum_{R=1}^{7}\left(\hat{\varepsilon}_{t,d}^c - \gamma_0^c - (\gamma_R^c \cos(2\pi t\Phi_R^c)\delta_{tR}\right) - (\gamma_R^c \sin 2\pi t\Phi_R^c)\delta_{tR})^2 \tag{6}$$

(V) Model (2) is appended to (1) additively as in (3) in order to provide local estimates of the O_3 data.

[‖]The state averages are O_3 concentrations for each hour in the day (t) in July, 1996. Thus, the model averages across monitors (n) and across days (d) so that

$$O_t^s = \frac{1}{31}\sum_{d=1}^{31}\frac{1}{N}\sum_{n=1}^{N}(O_{t,d,n}^s)$$

The model (1) allows for amplitude parameter estimates to vary between months since O_3 formation is highly dependent on local climate. Thus, the shape of the daily cycle changes from month to month, as do variables such as temperature, precipitation, and other factors. The findings in this report focus on July measurements to display the methodology and can be implemented in the same manner for other months. As a result we suppress the monthly subscript.

DOI: 10.1002/env

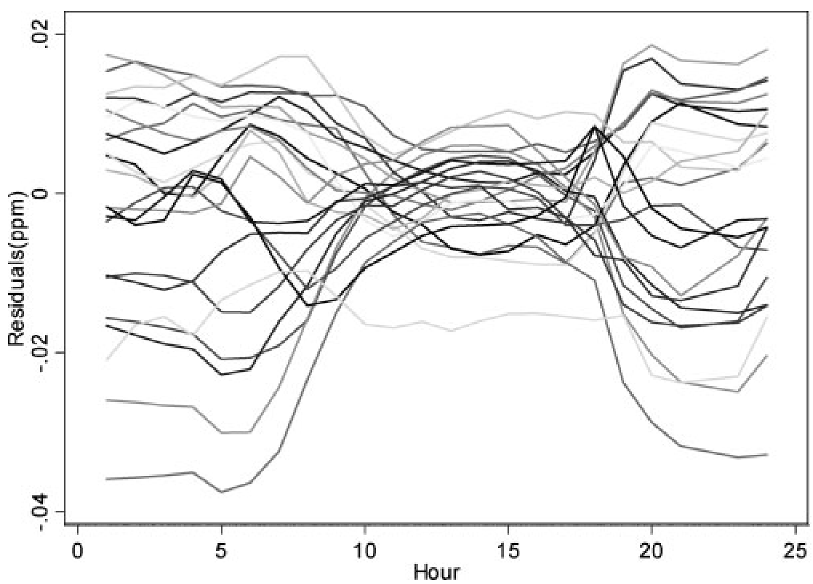

Figure 3. Residuals from local sample

2.3. Model evaluation

In many contexts, evaluating such models entails using the leave-one-out method (Stone, 1974; Hardle, 1990). The foundation of the leave-one-out method is that nearby points bear a strong similarity to one another. Hence, neighboring observations may be used to make predictions; a local sample of measurements at locations (j) is drawn to make inferences about the dependent variable at a point of interest (p). That is, if one supposes that a measured surface is generated by some functional relationship, the leave-one-out method presumes that this function is relatively smooth and continuous within the neighborhood of (p). In contrast to this presumption, the spatially erratic nature of the O_3 time series implies that, in this application, such a function is discontinuous. This largely precludes using the leave-one-out method.

Empirical evidence in the large panel data set used in the present study suggests that the local processes generating O_3 profiles can differ markedly between any two neighboring sites. Since the local deviations from the underlying periodic signature at any two monitors may be very different, the residuals from a collection of (j) local points are generally of little use in trying to model the idiosyncratic process at some given point of interest (p). As an example of this phenomenon, Figure 3 plots the residuals ($\hat{\varepsilon}^c_{t,d}$), calculated as shown in Section 2.1, from a local sample of 20 monitors against time. These monitors are a local sample drawn around a particular monitor (p), whose idiosyncratic effect we hope to estimate[¶]. The residuals corresponding to monitor (p) are shown in Figure 4. Taken together, these plots show that the residuals from a sample of points nearby monitor (p) bear little resemblance to those at (p). Thus, applying the leave-one-out method does not seem appropriate here.

¶In this example the monitor (p) is located in Phoenix, Arizona. The 20 monitors in Figure 3 are those within Maricopa County which encompasses Phoenix.

DOI: 10.1002/env

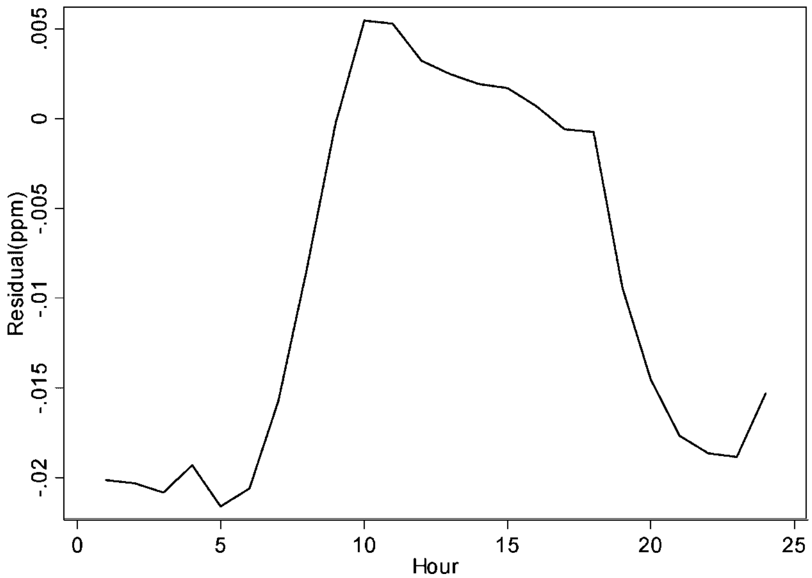

Figure 4. Residuals from monitor (p)

We test the fit of the parametric model by comparing its hourly predictions ($\hat{O}_t^s$) to state-averaged hourly observations (O_t^s). The hypothesis is that the state-averages represent the underlying structure of the 24 h O_3 cycle. Model fit is judged according to the mean proportional error (MPE) and the root mean squared error ($\sqrt{\text{MSE}}$). These statistics are calculated as shown in Equations (7) and (8). The error is determined at each of the 24 h in the cycle, and then reported as an average (for each state) as follows:

$$\text{MPE} = \frac{1}{24}\sum_{t=1}^{24}\left(\frac{\text{abs}\left(\hat{O}_t^s - O_t^s\right)}{O_t^s}\right) \tag{7}$$

$$\sqrt{\text{MSE}} = \sqrt{\frac{1}{24}\sum_{t=1}^{24}\left(\hat{O}_t^s - O_t^s\right)^2} \tag{8}$$

We test the fit of the semiparametric model (3) at two different spatial scales: county averages and observations taken from specific pollution monitors. In order to evaluate fit at the county-level, the parametric model is first estimated using all observations from the state containing the county of interest. Next, we compile the county-average O_3 cycle ($O_{t,d}^c$) by averaging across monitors within the county (c) for each day in July, 1996. Then the hourly deviations of the county average from model (1) predictions are calculated:

$$\hat{\varepsilon}_{t,d}^c = \left(\hat{O}_{t,d}^s - O_{t,d}^c\right) \tag{9}$$

The county residuals ($\hat{\varepsilon}_{t,d}^c$) are then regressed on the sinusoidal structure as shown in Equation (2). In order to assess the degree of improvement in fit between model (1) and model (3), we calculate the error

DOI: 10.1002/env

statistics shown in Equations (7) and (8) corresponding to the parametric model (MPE^1, $\sqrt{\text{MSE}^1}$) and after appending the nonparametric model (MPE^3, $\sqrt{\text{MSE}^3}$). Since there are approximately 530 counties with O_3 monitors, we summarize model performance by examining the accuracy of the predictions in a sample of counties from three land-use designations: urban, rural, and suburban counties.

The final test of model performance examines the fit to readings at particular pollution monitors. The experimental structure is the same as for the county-level tests. That is, the appropriate parametric model is first estimated. In order to evaluate fit at the monitor-level, we compile the observed O_3 cycle at monitor (m) for each day in July, 1996. Then the hourly deviations of the monitor cycle from model (1) predictions are calculated:

$$\hat{\varepsilon}^m_{t,d} = \left(\hat{O}^s_{t,d} - O^m_{t,d}\right) \tag{10}$$

The monitor residuals ($\hat{\varepsilon}^m_{t,d}$) are then regressed on the sinusoidal model in Equation (2). Again, the MPE and $\sqrt{\text{MSE}}$ corresponding to the parametric model and after appending the nonparametric element are computed. This reveals the degree of improvement in fit between models (1) and (3) for the monitor data. There are roughly 1000 monitors in the network. We report the fit to a sample of monitors.

2.4. Testing for idiosyncratic processes

In order to formally test for the presence of idiosyncratic effects, we explore whether the amplitude parameters (β^s_r) are significantly different in model (1) fitted to various states, and whether the (γ^c_R) are significantly different in model (3) fitted to various counties. This hypothesis is tested for each hourly segment (r) in Equation (1), and (R) in Equation (2). The test is structured as a two-tailed test with the following null (H_0) and alternative (H_1) hypotheses. Here, Equations (11) and (12) pertain to the tests applied to model (3):

$$H_0 : \hat{\gamma}^i_R = \hat{\gamma}^j_R \tag{11}$$

$$H_1 : \hat{\gamma}^i_R \neq \hat{\gamma}^j_R \tag{12}$$

The test statistic for two counties (i) and (j), denoted ($\tau_{i,j}$), is assumed to be distributed according to Student's t:

$$\tau_{i,j} = \left(\frac{\hat{\gamma}^i_R - \hat{\gamma}^j_R}{\hat{\sigma}_{iR}}\right) \sim t_{0.05}(n-2) \tag{13}$$

where τ is test statistic for counties (i) and (j), γ^i_R is amplitude parameter, at hourly segment (R), and county (i), $\hat{\sigma}_{iR}$ is the standard error estimate for (γ^i_R).

3. RESULTS

Table 1 reports the phase parameters (Φ_r) derived in stage 1 of estimation and subsequently used in the parametric model. Table 1 also shows other aspects of the specification used in model (1). Employing these (Φ_r), the least squares fit to the state average data is remarkably tight. Table 2, which reports

DOI: 10.1002/env

Table 1. Model (1) phase parameters

Hourly segment	$(t \in r)$	Φ_r
$r = 1$ (cos)	1–4	0.9575
$r = 2$ (cos)	5–9	0.9100
$r = 3$ (sin)	10–14	0.9460
$r = 4$ (sin)	15–18	0.9490
$r = 5$ (sin)	19–22	0.9440
$r = 6$ (cos)	23–24	0.9550

Table 2. Model (1) fit

State	MPE (%)	$\sqrt{\text{MSE}}$ (ppb)
Colorado	1	0.33
Wash. D.C.	3	0.56
Idaho	1	0.37
Illinois	2	0.38
Indiana	1	0.36
Louisiana	1	0.35
Michigan	1	0.33
New Jersey	1	0.31
Utah	1	0.50
Washington	2	0.49

results from a sample of 10 states, reveals that model (1) produces a mean proportional error of between 1% and 3%. The $\sqrt{\text{MSE}}$ is less than 1 part per billion (ppb) for each of these states. Figure 5 plots the predicted daily cycle ($\hat{O}_t^s$) from model (1) for Illinois and the observed state average (O_t^s) for Illinois against time. This plot provides additional evidence of the strong fit of the model; the only visually discernible deviation occurs in the early morning hours at the lowest levels of O_3.

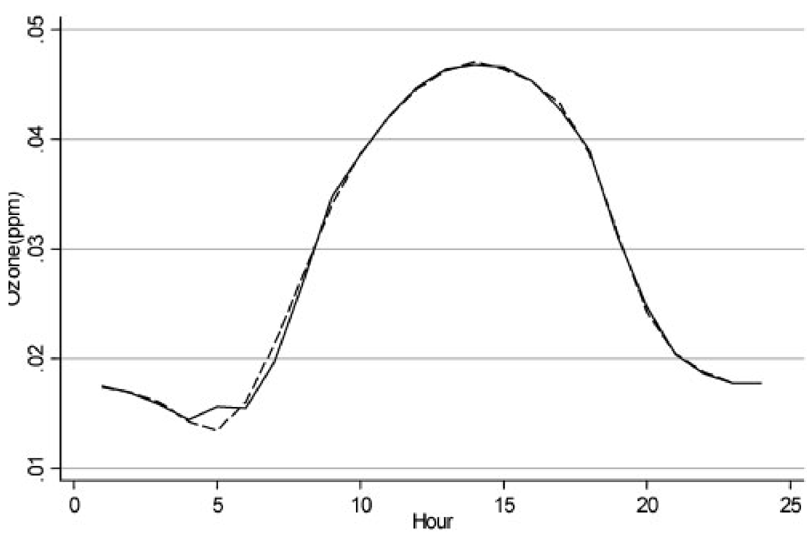

Figure 5. Model 1 fit to Illinois state average (model 1—dash, Observed—line)

DOI: 10.1002/env

Table 3. Model (3) fit: county experiments

County	State	Land use	MPE[1] (%)	MPE[3] (%)	$\sqrt{\text{MSE}^1}$ (ppb)	$\sqrt{\text{MSE}^3}$ (ppb)
Los Angeles	CA	Urban	67	0.0	10	0.4
Harris	TX	Urban	64	0.1	4.8	0.4
Cook	IL	Urban	22	0.1	4.3	0.3
Kings	NY	Urban	26	0.1	5.4	0.6
Westchester	NY	Suburban	11	0.2	4.3	0.6
Orange	CA	Suburban	55	0.1	12.6	0.3
Will	IL	Suburban	9	0.1	2.5	0.4
Oliver	ND	Rural	6	0.0	2.7	0.4
Florence	WI	Rural	27	0.2	6.5	0.6
Hamilton	NY	Rural	21	0.0	5.3	0.4

Table 3 reports the results derived from applying model (3) to a sample of counties. The MPE^1 statistic reveals that, generally, model (1) fails to capture the local O_3 cycle in an adequate fashion. In the four urban counties sampled, the MPE^1 ranges from 22% to 67%. Applying model (1) to these counties generates a $\sqrt{\text{MSE}^1}$ of between 4 and 10 ppb. In the six non-urban counties sampled in this experiment, the parametric model also fails to consistently fit the data; the lowest MPE^1 is 6% and the highest MPE^1 is 55%. Similarly, in this sample the $\sqrt{\text{MSE}^1}$ exhibits substantial variation: from 2.5 to nearly 13 ppb. However, Table 3 shows that model (3) is able to emulate the county-average O_3 data. In each of the 10 counties, the MPE^3 is less than 1%. Further, model (3) reduces the $\sqrt{\text{MSE}^1}$ by roughly an order of magnitude; the $\sqrt{\text{MSE}^3}$ is less than 1 ppb in all of the 10 counties. Figure 6 provides visual evidence of the improvement in fit due to employing model (3). The parametric model predictions (dots) are biased upwards, relative to the county observations, by a significant margin. However, it is evident that model (3) (dashed) fits the county average data (line) quite well.

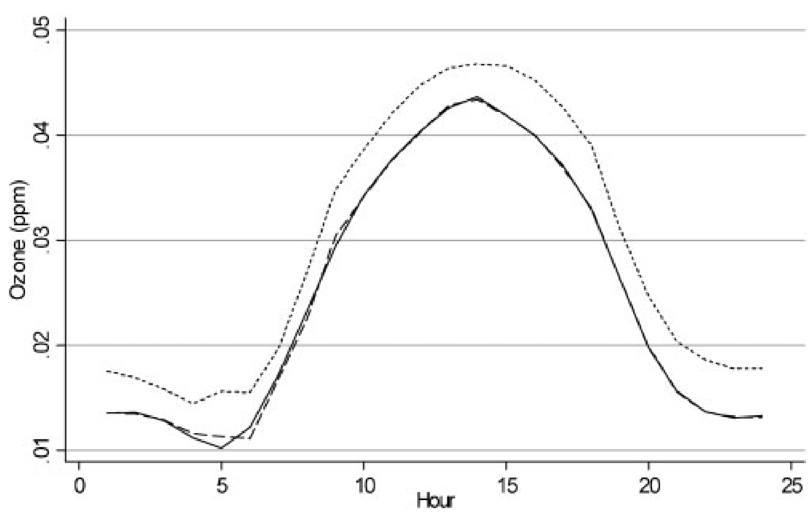

Figure 6. Model (3) fit to county average (Model (1)—dot, Model(3)—dash, Observed—line)

DOI: 10.1002/env

Table 4. Model (3) fit: monitor experiments

Monitor	County	State	Land use	MPE^1 (%)	MPE^3 (%)	$\sqrt{MSE^1}$ (ppb)	$\sqrt{MSE^3}$ (ppb)
6-37-4002	Los Angeles	CA	Urban	30	0.0	10	0.8
48-201-62	Harris	TX	Urban	71	0.1	6.7	0.4
17-31-4002	Cook	IL	Urban	113	0.7	11.9	0.5
4-25-2005	Maricopa	AZ	Urban	20	0.0	9.5	0.7
36-103-4	Suffolk	NY	Suburban	16	0.0	5.7	0.6
6-71-1	San Bernardino	CA	Suburban	15	2.3	13.8	1.8
42-17-12	Bucks	PA	Suburban	0.3	0.2	2.8	0.5
6-109-4	Tuolumne	CA	Rural	30	0.1	14.9	0.8
37-59-2	Davie	NC	Rural	13	0.1	5.5	0.8
45-21-2	Cherokee	SC	Rural	5.1	0.0	2.8	0.5

Model (3) is also tested in terms of fitting the O_3 cycle at particular monitors. The results of this experiment are shown in Table 4. Model (1) is clearly unable to consistently fit the O_3 pattern at the four urban monitors sampled. This is evident in the MPE^1 which ranges from 20% at a monitor near Phoenix to 113% at a monitor in Chicago. At non-urban monitors, the performance of model (1) is inconsistent; the MPE^1 stretches from 0.3% to 30%. In contrast, model (3) fits the local patterns remarkably well. At the four urban monitors, the MPE^3 is less than 1%. Further, the $\sqrt{MSE^3}$ is reduced to less than 1 ppb. At the six non-urban sites, model (3) also performs exceptionally well; the MPE^3 is only greater than 1% at a monitor in San Bernardino, CA. The ability of model (3) to fit local observations of the time series is driven by the model (2) fit to the local residuals. This is evidenced in Figure 7 which shows both the county and monitor residuals and the corresponding predictions from model (2). Figure 7 shows that model (2) is able to capture the idiosyncratic process at two levels of spatial detail.

Table 5 reports the results of the hypothesis tests designed to determine whether the amplitude parameters in model (1) vary across states. The testing results, which are applied to California, Illinois,

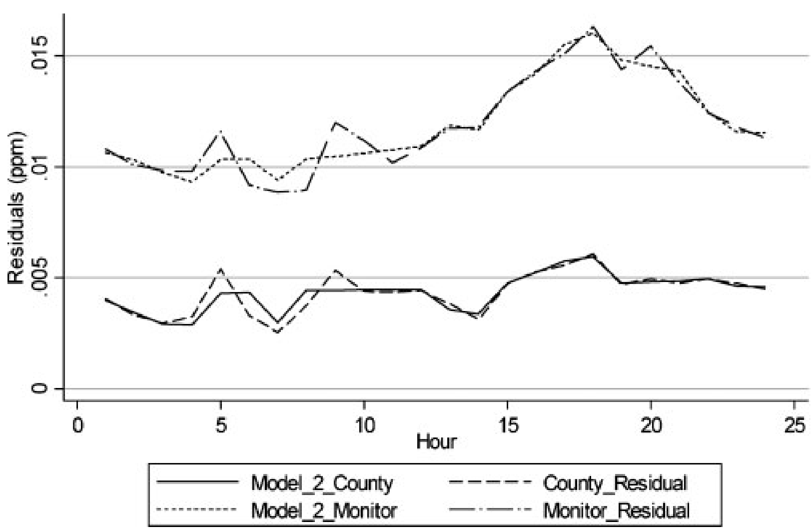

Figure 7. Model (2) fit to county and monitor residuals

DOI: 10.1002/env

Table 5. Testing for heterogeneity in model (1) amplitude parameters: $^{*}p = 0.10$, $^{**}p = 0.05$, and $^{***}p = 0.01$

State	β_1	β_2	β_3	β_4	β_5	β_6
California, Illinois	3.38***	7.5***	17.2**	11.6***	0.29**	0.41
California, New York	5.13***	52***	3.96***	21.5***	2.00***	2.24***
Illinois, New York	1.40	2.97***	8.75**	7.9***	1.60	1.70*

and New York, show strong, consistent evidence that the (β_r^s) vary significantly. This suggests that the amplitude of the periodic structure of the O_3 time series is not spatially homogenous. The test comparing the (β_r^s) estimates using observations from California and Illinois shows that β_1, β_2, and β_4 are significantly different at the 1% level. β_3 and β_5 are significantly different at the 5% level. Only for the test applied to β_6 can we not reject the null hypothesis of equal amplitude parameters across spatial location. The tests comparing the California and New York models indicate that β_1 through β_6 are significantly different at the 1% level. Finally, the tests pertaining to the Illinois and the New York models detect statistically significant evidence of different amplitude parameters for β_2 and β_4 at the 1% level, for β_3 at the 5%, and for β_6 at the 10% level. In the tests applied to β_1 and β_5 we fail to reject the null hypothesis.

To test for the presence of local process effects, we examine whether the amplitude parameters (γ_R^c) estimated in model (2) vary significantly across counties. Results from this testing procedure are reported in Table 6. This test is applied to the models estimated for Cook County, Illinois (encompassing Chicago), Kings County, NY (Brooklyn), and Los Angeles County, CA. For the models applied to Kings County and Los Angeles, β_2, β_3, β_5, and β_7 are significantly different at the 1% level, while β_1, β_4, and β_6 are significantly different at 5%. The test pertaining to Cook County and Kings County reveals that β_2 and β_4 are significantly different at the 1% level, while β_3 and β_5 are significantly different at 5% and 10%, respectively. In the models applied to Chicago and Los Angeles, β_3 shows significant differences at 1%, while β_5 shows significant differences at 5%.

Table 6. Testing for heterogeneity in model (2) amplitude parameters: $^{*}p = 0.10$, $^{**}p = 0.05$, and $^{***}p = 0.01$

Counties	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
Los Angeles, Cook	0.26	1.27	6.12***	0.87	2.26**	0.29	1.29
Los Angeles, Kings	2.09**	2.57***	12.0***	2.54**	4.28***	2.36**	10.5***
Cook, Kings	0.91	88.5***	2.43**	3.25***	1.79*	1.59	0.84

Table 7. Model (2) phase parameters

Hourly segment	$(t \in r)$	2nd Stage (Φ_R)	Southeast	West	Midwest	Northeast urban	Northeast
$R = 1$ (sin)	1–3	Φ_1	0.970	0.970	0.970	0.870	0.895
$R = 2$ (cos)	4–7	Φ_2	0.975	0.910	0.910	0.900	0.925
$R = 3$ (cos)	8–12	Φ_3	0.965	0.930	0.930	0.953	0.946
$R = 4$ (cos)	13–15	Φ_4	0.965	0.927	0.927	0.967	0.963
$R = 5$ (cos)	16–18	Φ_5	0.970	0.973	0.973	0.970	0.970
$R = 6$ (cos)	19–21	Φ_6	0.975	0.980	0.980	0.952	0.952
$R = 7$(cos)	22–24	Φ_7	0.960	0.970	0.958	0.956	0.956

DOI: 10.1002/env

4. CONCLUSIONS

This paper demonstrates that models consisting of sinusoidal functions can be used to fit spatiotemporal data in which there is considerable variation in the periodic structure. While it is a recognized shortcoming that such models typically have difficulty capturing idiosyncratic variation, the semiparametric strategy developed in the present paper successfully addresses the deficiency. Past work has sought to overcome the shortcoming by augmenting sinusoidal models with other modeling forms such as autoregression and deterministic trends. The approach developed here uses instead a parametric sinusoidal structure at the aggregate level and combines this common structure with a flexible sieve sinusoidal form to capture local idiosyncratic effects.

The empirical application reveals that this semiparametric approach can model spatiotemporal data with a variable periodic signature in a remarkably tight fashion. Using panel data of hourly air pollution measurements at monitors located throughout the United States, the sinusoidal semiparametric model produces an excellent fit at three successive levels of spatial detail. The state experiments show that the parametric component of the model is able to mimic state average measurements, thereby giving an underlying common periodic structure to the data. The county experiments show how the models replicate local idiosyncratic variation. This particular scale is a crucial test of model accuracy for policy purposes since the USEPA enforces its air quality standards at the county level. Thus, if the methods are to be used for interpolation purposes and policy analysis, there must be an adequate fit to county level readings. Finally, the monitor level experiments emphasize the method's inherent flexibility as it is able to match the observed O_3 time series at particular locations with a mean proportional error of less than 2.5%.

From a practitioner's perspective, the utility in these models lies in their ability to predict O_3 diurnal signatures at points not currently measured by the USEPA's network. Prior interpolation models have focused on daily maximum values, seasonal averages, and daily averages. However, the USEPA's shift from a 1 h standard to an 8 h standard makes it necessary to interpolate the entire daily O_3 cycle. One way to accomplish this within our framework is to functionalize the idiosyncratic effects on covariates of readily observable variables that are plausibly associated with O_3 measurements. Such covariates might include temperature, precipitation, and wind speed, as well as population and local land-use data. Then, since the local phase parameters are known, the local amplitude parameters can be regressed on these covariates to furnish predictions of the idiosyncratic process effects at a given location where there are no current O_3 measurements.

ACKNOWLEDGEMENTS

Phillips acknowledges support from the NSF under Grant Nos. SES 04-142254 and SES 06-47086. Muller would like to thank the Glaser Progress Foundation for supporting this research.

REFERENCES

Bell ML, McDermott A, Zeger SL, Samet JM, Dominici F. 2004. Ozone and short-term mortality in 95 US urban communities 1987–2000. *JAMA* **292**(19): 2372–2378.

Bloomfield P. 2000. *Fourier Analysis of Time Series* (2nd edn). John Wiley and Sons: New York.

Campbell MJ, Walker AM. 1977. A survey of statistical work on the MacKenzie river series of annual Canadian lynx trappings for the years 1821–1934 and a new analysis. *Journal of the Royal Statistical Society. Series A* **140**(4): 411–431.

Damsleth E, Spjotvoll E. 1982. Estimation of trigonometric components in time series. *Journal of the American Statistical Association* **77**(378): 381–387.
Dixon MJ, Tawn JA. 1999. The effect of non-stationarity on extreme sea-level estimation. *Applied Statistics* **48**(2): 135–151.
Grenander U. 1981. *Abstract Inference*. John Wiley and Sons: New York.
Guttorp P, Meiring W, Sampson PD. 1994. A space-time analysis of ground-level ozone data. *Environmetrics* **5**(3): 241–254.
Hardle W. 1990. *Applied Nonparametric Regression*. Econometric Society Monograph No. 19, Cambridge University Press: Cambridge, U.K.
Hopkins LP, Ensor KB, Rifai HS. 1999. Empirical evaluations of ambient ozone interpolation procedures to support exposure models. *Journal of the Air and Waste Management Association* **49**(7): 839–846.
Lewis PAW, Ray BK. 1997. Modeling long-range dependence, non-linearity, and periodic phenomena in sea surface temperatures using TSMARS. *Journal of the American Statistical Association* **92**(439): 881–893.
Seinfeld JH, Pandis SN. 1998. *Atmospheric Chemistry and Physics From Air Pollution to Climate Change*. Wiley-Interscience: New York.
Stone M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (methodological)* **36**(2): 111–147.
USEPA 1. http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm

DOI: 10.1002/env