APPLICATION OF THE DYNAMIC SPATIAL ORDERED PROBIT MODEL: PATTERNS OF OZONE CONCENTRATION IN AUSTIN, TEXAS

Xiaokun (Cara) Wang
Assistant Professor
Department of Civil and Environmental Engineering
Bucknell University
Lewisburg, PA 17837, USA

Kara M. Kockelman
(corresponding author)
Associate Professor & William J. Murray Jr. Fellow Department of
Civil, Architectural and Environmental Engineering The University
of Texas at Austin
6.9 ECJ, Austin, TX 78712-1076
kkockelm@mail.utexas.edu

The following paper is a pre-print and the final publication can be found in
Transportation Research Record No. 2136:45-56, 2009.
Presented at the 88th Annual Meeting of the Transportation Research Board,
January 2009

ABSTRACT

While a wide variety of transportation data sets involve discrete values scattered across space and time, few techniques presently exist to properly analyze such data. A new dynamic spatial ordered probit model (DSOP) is described here, and its use is demonstrated for a case of ozone concentration categories. Using outputs of photochemical models for the Austin, Texas region over a 24-hour period, the model parameters were estimated using Bayesian techniques, and results illuminate key relationships, many of which are intuitive but generally obscured by complex upstream model systems. Relying on 132 4 km x 4 km surface grid cells as observational units, values are found to exhibit strong patterns of temporal autocorrelation, but appear strikingly random in a spatial context (after controlling for local land cover, transportation, and temperature conditions). While transportation and land cover conditions appear to influence ozone levels, their effects are not as instantaneous, nor as practically significant as the impact of temperature. The DSOP model proposed here is able to accommodate the unusual dynamics and spatial evolution of ordered response categories inherent in the ozone data.

KEY WORDS: spatial autocorrelation, dynamic model, ordered probit, Bayesian estimation, ozone concentration

BACKGROUND

In the study of urban systems, many variables of interest are discrete and ordered in nature. Many also exhibit temporal and spatial dependencies. For example, pavement surface deterioration levels, air pollutant concentration classes, and standard-of-living indices are often described using ordered categories. Such variables also are influenced by various site-specific factors subject to spatial and temporal autocorrelation (across observations in space and time). To understand such phenomena and quantify the effects of influential factors, rigorous statistical methods are needed.

Over the years, various studies have been attempted to recognize spatial and temporal autocorrelations in data analysis. For tackling spatial autocorrelation, two major methods are spatial filtering (e.g., Nelson and Hellerstein [1997], Wear and Bolstad [1998], and Munroe et al. [2001]) and specification of a spatial autoregressive (SAR) process (e.g., Anselin [1988], Anselin and Bera [1998], and Anselin [2003]). For recognizing temporal autocorrelation, time series analysis is widely accepted as a reliable approach. However, few have considered the effects of such autocorrelations in discrete response data analysis. The limited set of published studies in this area focus on binary choice settings, and none recognizes both temporal and spatial autocorrelations simultaneously.

For these reasons, the objective of this study is to illustrate the specification and applicability of the dynamic spatial ordered probit (DSOP) model, a new and powerful approach to spatial data analysis with temporal autocorrelation, as illustrated here using data on ozone concentration levels. The following section motivates this topic, for the case of air quality.

The Importance of Ozone

As a gas in the stratosphere that protects Earth from harmful ultraviolet rays, the ozone layer shields living things. In the troposphere, however, ozone is a powerful oxidizer, harming lung tissue and other materials. Under the National Ambient Air Quality Standards (NAAQS), all Metropolitan Statistical Areas (MSAs) in the United States are required to develop strategies for attaining the standards and accommodate future growth. Thus, planners and policy makers must understand the spatial distribution of air pollutants, like ozone. Currently most studies on ozone concentration projection are based on the modeling of photochemical process. Though such an approach is more insightful, compared to statistical modeling, it is not very convenient for sensitivity analysis, and is not very flexible for adding new variables of interest. In contrast, a rigorous statistical model can be expected to facilitate the understanding of different factors' impacts on ozone concentration more conveniently.

Ozone concentration is usually expressed as a continuous value. For example, the California one-hour ozone standard is set at 0.09 parts per million (ppm) and the eight-hour average ozone standard is 0.070 ppm (BAAQMD, 2005). The U.S. standard was recently reduced to 0.075 ppm (EPA, 2008), and many regions around the U.S. are very anxious to avoid non-attainment status.

Many continuous variables are often made categorical, in order to convey key information more directly to policy makers and the public. This is common in the case of air quality forecast reports for public consumption, which are often indexed as low, moderate and potentially dangerous concentrations. (See, for example, Athanasiadis et al., 2007.)

Of course, many factors can and do influence ozone concentration levels through complex chemical and physical processes. For example, Niemeier et al. (2006) found that for most regions

in the Northern Hemisphere, road traffic intensity is closely associated with local ozone concentrations. They surmised that, if traffic-related emissions per capita in south Asia hit U.S. levels, that continent's surface ozone concentrations would increase by 50 to 100%. Wang et al. (2005) concluded that transportation sources are the main contributor to ozone concentrations, averaging roughly twice the effect of industrial emissions. Friedman et al. (2001) studied changes in commuting behaviors during the1996 Summer Olympic Games in Atlanta and noted how decreased traffic densities were associated with a prolonged reduction in ozone pollution.

Land coverage development and intensity are also important determinants. And, of course, even if the land is not developed for human use, its features need to be classified for calculation of biogenic emissions. These are naturally occurring emissions from vegetation, which can be a strong function of tree type. For example, live oak trees are high emitters of isoprene, a highly reactive, volatile organic compound (VOC) that is a precursor to ozone. In areas such as eastern Texas, where this species is common, biogenic emissions of VOCs dominate the area's emissions inventory (Wiedinmyer, 1999). Another reason for requiring such land coverage information is the calculation of dry deposition rates. Dry deposition refers to the accumulation of particles and gases as they come into contact with soil, water or vegetation on the earth's surfaces. Allen (2002) suggests that during ozone season in Texas, dry deposition is the most important physical removal mechanism for air pollutants. Dry deposition rates for specific pollutants are typically computed according to land cover type. McDonald-Buller et al. (2001) investigated the sensitivity of dry deposition and ozone mixing ratios as a function of land cover classification and noted the importance of establishing accurate, internally consistent land cover data for air quality modeling. Thus, changes to both developed and undeveloped land cover type can significantly alter the magnitude spatial distribution of ozone.

Of course, many other factors also play a role. For example, Guldmann and Kim (2001) suggest that, in addition to land development and transportation characteristics, pollution measurements, meteorological factors and socioeconomic data can and do influence ozone concentrations. Loibl et al. (1994) show how relative altitude and time of day are influential. Pont and Fontan (2001) suggest that though local reduction in traffic is important, advection[1] of ozone is also critical to its concentration.

Obviously, ignoring any of these relevant factors introduces uncertainty in model estimation and prediction. Such variables, if unobserved, can generate both temporal and spatial autocorrelations in model error terms. For example, meteorological factors (such as local wind speeds, rainfall, relative humidity, and temperature), precursors of ozone, and pollution control policies all exhibit positive temporal and spatial dependencies (see, for example, Lin, 2007, and Hancock, 1994). Therefore, it is reasonable to incorporate temporally and spatially lagged term and neighborhood effects in model specification.

In summary, ozone concentration levels are related to numerous factors. Among them, transportation conditions and land use/land cover information appear critical for urbanized region. A statistically rigorous analysis of ozone concentration categories can be achieved via application of an ordered discrete choice model with a temporal lagged item and spatial autocorrelation in error terms. The following sections describe key features of such a model, and its application to the case of data from Austin, Texas.

[1] Advection refers to the transport of something from one region to another. Ozone's advection is predominantly horizontal, following weather system patterns (Noguchi et al., 2006)

MODEL SPECIFICATION AND ESTIMATION

Wang (2007) and Wang and Kockelman (2008a) explain the dynamic ordered probit (DSOP) model's specification and estimation process in detail; and Wang and Kockelman (2008b) use the DSOP model to analyze land development intensity levels over time (for purposes of anticipating land use change). This section simply summarizes the specification, to show how the model incorporates spatial, temporal and discrete features of the dataset.

The model starts with specification of the latent (unobserved) response variable U_{kt} , where the subscript indicates individual i ($i=1,\ldots,N$) in period t (t=1, 2, …, T). Each individual is observed T times, making the total number of observations as NT. Each latent variable U_{it} is a function of the unobserved variable from previous period U_{it-1}, and other explanatory variables X_{it}. Therefore, the specification is as follows:

$$U_{it} = \lambda U_{it-1} + \boldsymbol{X}_{it}'\boldsymbol{\beta} + \theta_i + \varepsilon_{it}\ ,\ t = 1,\ldots,T \tag{1}$$

where λ is the temporal autocorrelation coefficient and X_{it} is a $Q\times 1$ vector of explanatory variables. β is the set of corresponding parameters. The remaining (uncontrolled/latent) information is composed of two parts: θ_i which captures the individual-specific random components for individual i, and time-variant individual effect ε_{it} which is allowed to be heteroscedastic with variance υ_i .

Furthermore, θ_i values can exhibit spatial autorcorrelation, so that

$$\theta_i = \rho\sum_{j=1}^{M} w_{ij}\theta_j + u_i\ ,\ i = 1,\ldots,M \tag{2}$$

where weight w_{ij} is an exogenous indicator of contiguity (1 for contiguous and 0 otherwise), ρ_t is the spatial coefficient, and u_i (which is iid normal with zero mean and variance σ^2) stands for the part of individual specific effect that is not influenced by its neighbors. The vector of regional effects is thus a function of the weight matrix $\boldsymbol{W}$, with w_{ij} as its elements.

The observed response variable, y_{ikt}, is a censored form of the unobserved response variable:

$$y_{it} = s \quad \text{if } \gamma_{s-1} < U_{it} < \gamma_s \text{ for } s = 1,\ldots,S \tag{3}$$

That is, the possible outcomes have potential integer values between 1 to S , which are determined by the value of latent variable U_{it} and the unknown boundaries $\gamma_0 < \gamma_1 < \cdots < \gamma_{S-1} < \gamma_S$. ($\gamma_0 = -\infty$ and $\gamma_S = +\infty$).

In the ordered probit setting, the likelihood function can be easily derived as follows:

$$Likelihood = \prod_{t=1}^{T}\prod_{i=1}^{N}\sum_{s=1}^{S}\delta\left(y_{it}=s\right)\cdot\delta\left(\gamma_{s-1} < U_{it} < \gamma_s\right) \tag{4}$$

where $\delta(A)$ is an indicator function that equals 1 when event A is true and 0 otherwise.

As explained in Wang (2007) and Wang and Kockelman (2008a), Bayesian MCMC methods are used to estimate all unknown parameters, providing valuable distribution information for all estimators (rather than simply means and standard deviations, as in the case of classical methods).

This model recognizes regional effects, spatial heterogeneity, spatial autocorrelation, and temporal autocorrelation in a latent setting with ordered categorical responses. The general framework can reduce to several simpler specifications, for cases of special interest – such as when the dataset exhibits no temporal autocorrelation (i.e., individuals' current responses do not rely on prior states) and responses are homoskedastic.

Due to a somewhat limited sample size and no obvious arguments for heteroskedastic tendencies across cell ozone levels, a single variance is used ($\upsilon_i = \upsilon$).

DATA DESCRIPTION

Ozone concentration levels were derived from continuous values originally prepared for an EPA project, and provided by Dr. Elena McDonald-Buller at the University of Texas at Austin (CAPCO et al., 2004). Using the ENVIRON's ® CAMx photochemical model, many emissions inventories and a variety of behavioral assumptions, the researchers developed hourly ozone concentration estimates for a high-ozone episode, using meterological data for the September 13-20, 1999 period.

In the Capital Area Planning Council (CAPCO) study, there are three levels of spatial resolution, and the finest is 4 km. This resolution area covers a 360 km x 432 km area (i.e., 90 x 108 grid cells) and includes all major urban centers within southern Texas and the Texas Gulf Coast. In this study, hourly data for just one day (September 13, 1999) was selected, and available transportation and land cover information (derived by Wang 2007) limited the scope to the Austin region, a 44 km x 48 km study area containing 132 4 km x 4 km grid cells. Thus, the resulting dataset is a 132 (N) × 24 (T) panel with values indicating ozone concentrations in parts per million (ppm).

The rule for defining ozone concentration levels should be flexible and adaptable to the user's needs. In addition, every category needs to contain enough observations so that each is well represented. Here, the values were categorized into 5 groups: values below 0.035 are assigned Level 1, values between 0.035 and 0.04 are Level 2, those between 0.04 and 0.45 are Level 3, those between 0.045 and 0.05 are Level 4, and those above 0.05 are categorized as Level 5.[2]

Figure 1 illustrates the continuous ozone concentration values and their corresponding levels using data between 4 and 5pm on September 13, 1999 as an example. Table 1 shows the changing trend of ozone concentration levels during the 24 hours: the levels are higher during daytime, especially in the afternoon, and lowest at night and in the early morning.

Austin's neighborhoods' temperature information comes from the same EPA project datasets, provided by Dr. McDonald-Buller. Table 2 illustrates the distribution and changes in temperatures over the 132 cells and 24 hours.

As noted earlier, local traffic and land use/land cover conditions may influence local ozone concentrations. Ideally, traffic counts and VMT by hour by cell would be available for use. Such variables were not readily available (by time of day or all network links), so the total length of street centerlines (per grid cell) was used as a proxy for local VMT levels.

Land cover type influences ozone concentration because it contributes to both ozone generation (biogenic or anthropogenic) and deposition. Residential, commercial, transportation and industrial land (i.e., "developed" lands) may be categorized together, since they mainly

[2] While the current non-attainment threshold is 0.08 ppm, the sample data do not contain such high concentrations.

contribute to anthropogenic emissions and their land cover materials may offer similar dry-deposition rates. Treed areas, brush, and agricultural land all contribute biogenic emissions and are expected to have similar dry deposition rates, so they may be aggregated as "vegetation." Barren land and water, though having quite different dry deposition rates, only account for a small proportion of the land in the study area, and so have been grouped together, as "undeveloped land", in order to avoid possible multi-collinearity issues.

The land cover information comes from year-2000 satellite data provided by Tufts University's Dr. Parmenter (at 30 meter resolution). The satellite data come from September 3, 1999 – very close to the model day. Hence, seasonality differences can be ignored. Based on computer-aided classification of the satellite images (using both supervised and unsupervised methods), fractions of the three aggregate land cover types described above were calculated. In addition, in order to account for variations in human activities and the effect of daylight (which can be influential to both ozone generation and deposition) across different times of day, these transportation and land cover fractions were interacted with several time-of-day indicators. Total street length was multiplied by an indicator for peak travel hours (i.e., 7:00 to 10:00 and 16:00 to 19:00 h) and non-peak hours (any other time of day); developed land was interacted with working hours (i.e., 8:00 to 17:00 h) and non-work hours; and, because plant activity is strongly influenced by the presence of daylight, vegetated lands were interacted with a "day time" indicator (6:00 through 18:00 h) an night-time indicator (18:00 through 6:00 h). The fraction of undeveloped land was used as the base case.

In summary, the dataset used for the ozone concentration model contains 132 observational units (grid cells) over 24 hours, providing a total of 3,168 data points. Explanatory variables include temperature, street lengths interacted with indicators for peak/non-peak hours, percentages of developed land interacted with indicators for work/non-work hours, and percentages of vegetated land interacted with indicators for day/night time conditions. Table 3 summarizes definitions and statistics of all these variables. The mean and standard deviation of the ozone concentration levels imply that the dependent variable values are well balanced (i.e., each level has adequate observations, also shown by Table 1). The large standard deviations of all explanatory variables (as compared to the mean values) indicate substantial data variability, which can be useful from the standpoint of statistically significant identification of parameter values.

MODEL ESTIMATION AND RESULTS

This section presents the methods and results of DSOP model estimation using Austin's CAMx-predicted ozone concentration categories. Parameter estimates, their marginal effects and model predictions disclose some interesting findings, which may help researchers and planners better understand air quality dynamics.

As previously noted, each grid cell serves as its own region, and region-specific effects are assumed homoskedastic (over space). In addition, all variances of individual-specific error terms can be set to equal 1.0, so these errors all follow standard normal distributions. While most of the posterior distributions are standard and can be conveniently generated using routines built in commercial mathematical analysis packages, the spatial coefficient ρ's distribution is non-standard and had to be generated using numerical methods. The vector of threshold parameters, γ, follows a multidimensional truncated normal distribution, and these truncations co-vary. More

information on these and other estimation details can be found in Wang (2007) and Wang and Kockelman (2008a).

As a standard estimation procedure, the model for ozone concentration level was initiated with diffuse priors, and the total number of iterations used was 8,000. As Figure 2 suggests, after 4000 runs all traces become stable, indicating convergence. Therefore, the first 4000 runs were omitted (as a "burn in" sample), and all inferences were drawn from results in iterations 4001 to 8000.

Table 4 shows parameter estimates based on the final 4000 runs. The estimation suggests that temperature has a statistically significant (and positive) effect on ozone concentration levels, as expected. Interestingly, during peak travel hours, the total length of streets in the area (in this case equal to road density, since area is constant across grid cells) has no statistically significant effect. However, during non-peak hours, higher road density is associated with higher levels. This somewhat counterintuitive phenomenon may be explained by a delay in the photochemical process for ozone generation and deposition: the process may require several hours to develop.

Two other factors to consider are the fraction of developed land and vegetation. The fraction of developed land has the same effect during work and non-work hours. Vegetated land also has nearly the same effect – both day and night. These results indicate that, while land cover has a significant role, its effect is not instantaneous, possibly due to the time needed for the photochemical process.

The estimation also shows that λ has a fairly high value and is statistically significant, indicating that the latent dependent variable of the previous period plays an important role. The ρ value is close to zero, slightly negative (on average), and statistically insignificant, implying that the control variables adequately explain any spatial clustering in ozone concentration levels.

The values of estimated regional-specific errors (θ_i), and their statistical significance (t-statistic greater than 1.64) are shown, in Figure 3. As indicated by the low ρ value returned by the model, the θ values appear randomly distributed across space. (This result is supported by a Moran's I of -0.05, with a Z score of just 0.4.[3])

Posterior distributions of all parameters are shown in Figure 4. While the distributions for threshold parameters are multimodal as in the case of the development intensity results, the overall intervals are fairly narrow, offering statistically significant estimates. Of course, it should be mentioned that the response data are generated using a model (CAMx). Thus, to a large extent the estimation process is simply recovering key features and influential factors in that model. The primary objective of this paper is model demonstration; actual ozone readings would be needed to understand key factors from a chemical and physical standpoint.

Marginal Impacts of Explanatory Variables

With models of discrete response data, evaluation of marginal effects is an important tool for interpretation of results. Marginal effects indicate the effect that a one-unit change in an

[3] Moran's I is an important spatial statistic. A Moran's I value near +1.0 indicates clustering while a value near –1.0 indicates dispersion. The Z score indicates whether or not the null hypothesis "there is no spatial clustering" can be rejected. At a significance level of 0.05, a Z score less than -1.96 or greater than 1.96 indicates statistical significance. (ESRI, 2005)

explanatory variable has on the probability of different discrete outcomes[4]. Marginal effects were calculated here during the Gibbs Sampling routine for each observation in each time period (see Wang [2007] and Wang and Kockelman [2008b] for more details). The average values of these results are summarized in Table 5.

One interesting result is the switch in signs of effects across ozone classes, presenting a "jumpy" pattern: Levels 1 and 4 share a consistent direction of change that opposes all others. This example highlights the fact that marginal effects for intermediate levels cannot be inferred directly from parameter signs when multiple observational units are involved.

Table 5 values quantify how different variables are estimated to influence ozone concentration levels, thereby illuminating their practical significance. A variable's practical significance often relates to its statistical significance, but is generally more relevant for model application and inference.

First, by increasing temperatures one degree centigrade, the probabilities of Levels 1 and 4 are expected to fall by 3.8% and 1.7%, respectively; and probabilities of Levels 2, 3 and 4 are estimated to increase by 1.2%, 0.6%, and 3.7%, respectively. Considering that temperature can change by more than 10 degrees in a day, its effect is quite impressive.

The effect of street length, during both peak and non-peak hours, is negligible: even when lengths are increased by 20 km (roughly the current average), the corresponding change in different levels' probabilities is less than 1.3%. This result suggests that traffic local intensities (as approximated using road density) may not influence ozone concentrations in Austin. This conclusion is somewhat counterintuitive (see, e.g., Niemeier et al., 2006, Wang et al., 2005, and Friedman et al., 2001) but is probably due to the fact that the variable of total centerline miles of roadway assigns a high value to dense local street networks (often with light traffic conditions) and low values to single major freeway corridors (generally carrying a great deal of traffic).

The fraction of developed land has a significant (negative) influence. If this fraction increases by just 1%, probabilities of the highest and lowest ozone categories (Levels 1 and 5) are estimated to change by 0.9% (when computed using sample averages). The fraction of vegetated land has a similar effect: an increase of 1% suggests a 0.7% increase in Level 1 concentrations and a 0.7% decrease in Level 5 concentrations. While developed and vegetated lands may be expected to contribute more to regional ozone formation, they also may assist in the process of local ozone deposition. Thus, their net effect, when compared to barren land and water, may be to decrease local ozone concentration levels. Though more insightful reasons for explaining the effects of land cover cannot be given here due to insufficient understanding of the photochemical process, the statistical relationships provided by the model estimation are helpful enough for planners to associate land cover with air quality.

Model Prediction

The model can be used for prediction, as in the following scenario: cell/region temperatures are set to those at 0:00 to 1:00 h on September 13, the fraction of developed land in each grid cell is assumed to be 1.2 times that of its current value, and vegetated land is set to 80 percent of its current value. The "previous period" is 23:00 to 24:00 h on September 13, so we are predicting just 1 hour forward in time.

[4] Of course, when sizable correlations exist among control variables and uncontrolled factors, such variables tend to proxy for the effects of missing variables, resulting in biased parameter estimates.

Graphs (a) and (b) in Figure 5 show the most likely predicted ozone concentration levels and an "uncertainty index" for each cell. The uncertainty index is simply entropy (based on summing the (negative) product of predicted probabilities for each response level and their natural log: $-p_i \ln(p_i)$, as described in Wang and Kockelman (2008a)). Ozone concentration levels generated by CAMx for 23:00-24:00 on September 13 and 0:00-1:00 on September 14 are also shown, as graphs (c) and (d), for comparison.

The prediction suggests almost no effect of land cover, as one might expect (since ozone formation is such a regional phenomenon). Part of the reason is that developed land and vegetated land effects are estimated to be very similar. (So increases in one offset reductions in the other). Since land cover variations ultimately have a negligible effect in this data set, and temporal dependencies (on prior latent dependent variable values) are strong, one expects a prediction pattern somewhere between patterns shown in graphs (c) and (d), as is clearly the case here. To some extent, this comparison validates the model.

Graph (b) shows that higher uncertainties are associated with higher levels of ozone, but even the highest uncertainty is only around 0.75. A closer look at the data shows that the uncertainty is mainly caused by confusion or ambiguity between Levels 2 and 3. Given the expectation that the predicted pattern should lie between cases (c) and (d) – which are dominated by Levels 2 and 3, such confusion is understandable.

SUMMARY AND CONCLUSIONS

This study uses a dynamic spatial ordered probit (DSOP) model to analyze ozone concentration levels across Austin, Texas, recognizing the discrete nature of the observed response values (necessitating the use of a latent model structure), their temporal dependence (on prior period latent responses), and spatial autocorrelation (in unobserved error terms). The results reveal a highly continuous process, where ozone concentration levels during any given hour are primarily determined by the values from the previous hour, with a 0.66 mean temporal lag coefficient. Temperature is another very influential predictor, while transportation and land cover variables prove far less helpful (as to be expected, given the source of the data). In addition, their mild effects are not instantaneous. The coefficient on the spatial error matrix was close to zero in this case, further implying that the temporal lagged utility and temperature adequately explain the changing ozone levels predicted by the CAMx model.

The ozone dataset includes a total of 132 grid cells at 4 km space resolution, and the estimation suggests no spatial autocorrelation. However, sample size may be too small and grid cell sizes too large to discern clustering and other patterns of spatial autocorrelation. More importantly, both dependent and explanatory variables in the ozone dataset are derived from CAMx model predictions, rather than actual ozone measurements across Austin. In the future, predictions for a larger-scale area should be tried, and if possible, observed ozone concentration data should be collected and used, though such sites are generally few[5]. The results from using real data should be compared to those from the CAMx model to further validate this statistical method.

[5] In practice, sampling sites are very scarce. For example, the study area has only two such sites on September 13, 1999 (and seven in June 2007) (TCEQ 2007). For San Francisco Bay Area basin, the number is 23 in 1999 and 22 in 2007 (CARB, 2007).

The primary purpose of this paper is simply a demonstration of a new model for a relatively common style of data of interest to transportation analysts, regional scientists, demographers, planners, chemists, and others. Though the data set can be further refined, the application of Wang and Kockelman's (2008) DSOP model to air quality discloses some interesting findings. More importantly, it illustrates the potential of spatial statistical methods for analyzing a variety of interesting problems. In the current context of ozone data, such methods can serve as a supplemental approach to the existing large-scale, complex modeling of photochemical processes. And extensions to three-dimensional and multinomial (rather than ordered) responses are of great interest.

ACKNOWLEDGEMENTS

The authors thank the U.S. Environmental Protection Agency (EPA) for funding this research (under STAR grant No. RD83183918), along with support from the Benjamin Stevens Graduate Fellowship in Regional Science. The authors also are grateful to Dr. Elena McDonald-Buller and her teammates for provision of the ozone data, Dr. Jim LeSage for offering useful discussions relating to analytical methods, and Ms. Annette Perrone for her administrative support.

REFERENCES

Allen, D.T., Durrenberger, C. and TNRCC Technical Analysis Division (2002) "Accelerated science evaluation of ozone formation in the Houston Galveston area: Photochemical air quality modeling." Technical paper. Accessed May 10, 2007: http://www.utexas.edu/research/ceer/texaqsarchive/pdfs/Modeling02_17_02.PDF.

Anselin, L. and Bera, A. (1998). "Spatial dependence in linear regression models with an introduction to spatial econometrics." In Ullah, A. and Giles, D. E.A. (Eds.), *Handbook of Applied Economic Statistics.* New York: Marcel Dekker.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Press.

Anselin, L. (2003) "Spatial externalities, spatial multipliers and spatial econometrics." *International Regional Science Review* 26(2): 153-166.

Athanasiadis, I. N., Karatzas, K. D. and Mitkas, P. A. (2007) "Classification techniques for air quality forecasting." Working paper. Accessed May 10, 2007: http://issel.ee.auth.gr/ktree/Documents/Root%20Folder/ISSEL/Publications/AKM_besai_KARATZAS_revised.doc.

Bay Area Air Quality Management District (BAAQMD) (2005) "Bay Area 2005 ozone strategy and draft environmental impact report summary." Technical report. Accessed May 10, 2007: http://www.baaqmd.gov/pln/plans/ozone/2003_workgroup/os_deirsummary.pdf.

CAPCO (Capital Area Planning Council), The University of Texas at Austin and ENVIRON International Corporation (2004) *Development of the September 13-20, 1999 Base Case Photochemical Model for Austin's Early Action Compact*. Technical report. Accessed May 10, 2007:

http://www.capco.state.tx.us/CAPCOairquality/NOV_30/eac_basecase_milestone_complete(revised2).pdf.

EPA (Environmental Protection Agency) (2008) National Ambient Air Quality Standards (NAAQS): Ozone (O3) Standards. Accessed June 29, 2008: http://www.epa.gov/ttn/naaqs/standards/ozone/s_o3_index.html.

ESRI (2005) ArcView GIS software, version 3.2. Environmental Systems Research Institute Inc., Redlands California.

Friedman, M. S., Powell, K. E., Hutwagner, L., Graham, L. M. and Teague, W. G. (2001) "Impact of changes in transportation and commuting behaviors during the 1996 Summer Olympic Games in Atlanta on air quality and childhood asthma." *Journal of American Medical Association* 285(7): 897-905.

Guldmann, J.M. and Kim, H.Y. (2001) "Modeling air quality in urban areas: A cell-based statistical approach." *Journal of Planning Literature* 16(1): 80-163.

Lin, C. Y. C. (2007) "A spatial econometric approach to measuring pollution externalities: An application to ozone smog." Working paper. Accessed May 10, 2007: http://www.des.ucdavis.edu/faculty/Lin/airqual_ext_paper.pdf.

Loibl, W., Winiwarter, W., Kopsca, A., Zueger, J. and Baumann, R. (1994) "Estimating the spatial distribution of ozone concentrations in complex terrain." *Atmospheric Environment* 28(16): 2557–2566.

McDonald-Buller, E.C., Wiedinmyer, C., Kimura, Y. and Allen, D. T. (2001) "Effects of land use data on dry deposition in a regional photochemical model." *Journal of Air and Waste Management Association* 51(8): 1211-1218.

Munroe, D., Southworth, J. and Tucker, C. M. (2001) "The dynamics of land-cover change in western Honduras: Spatial autocorrelation and temporal variation". *Conference Proceedings. American Agricultural Economics Association.* AAEA-CAES 2001 Annual Meeting. Accessed July 10, 2004: http://agecon.lib.umn.edu/cgi-bin/pdf_view.pl?paperid=2611

Nelson, G. C., and Hellerstein, D. (1997). "Do roads cause deforestation: Using satellite images in econometric analysis of land use". *American Journal of Agricultural Economics* 79: 80-88.

Niemeier, U., Granier, C., Kornblueh, L., Walters, S. and Brasseur, G. P. (2006) "Global impact of road traffic on atmospheric chemical composition and on ozone climate forcing." *Journal of Geophysical Research* 111(9): D09301.01- D09301.13.

Noguchi, K., Imamura, T., Oyama, K.I., and Bodeker, G.E. (2006). "A global statistical study on the origin of small-scale ozone vertical structures in the lower stratosphere." *Journal of Geophysical Research* 111: D23105, doi:10.1029/2006JD007232.

Pont, V. and Fontan, J. (2001) "Comparison between weekend and weekday ozone concentration in large cities in France." *Atmospheric Environment* 35(8): 1527-1535.

Wang, X., Carmichael, G., Chen, D., Tang, Y. and Wang, T. (2005) "Impacts of different emission sources on air quality during March 2001 in the Pearl River Delta (PRD) region." *Atmospheric Environment* 39(29): 5227–5241.

Wang, X. (2007) *Capturing Patterns of Spatial and Temporal Autocorrelation in Ordered Response Data: A Case Study of Land Use and Air Quality Changes in Austin, Texas.* Ph.D. Dissertation, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin.

Wang, X. and Kockelman, K. (2005) "Occupant injury severity using a heteroscedastic ordered logit model: distinguishing the effects of vehicle weight and type." *Transportation Research Record* 1908: 195-204.

Wang, X. and Kockelman, K. (2008a) "The Dynamic Spatial Ordered Probit Model: Methods for Capturing Patterns of Spatial and Temporal Autocorrelation in Ordered Response Data, using Bayesian Estimation." Under review for publication in the Journal of Regional Science.

Wang, X. and Kockelman, K. (2008b) "Application of the Dynamic Spatial Ordered Probit Model: Patterns of Land Development Change in Austin, Texas." Under review for publication in the Papers inRegional Science.

Wear, D. N. and Bolstad, P. (1998) "Land-use changes in southern Appalachian landscapes: Spatial analysis and forecast evaluation." *Ecosystems* 1: 575-594.

Wiedinmyer, C. (1999) *Biogenic Hydrocarbons in Texas: Source Characterization and Chemistry.* Ph.D. Dissertation, The University of Texas at Austin.

LIST OF TABLES AND FIGURES

Table 1 Frequency of Ozone Concentration Levels throughout the Day

	Number of Grid Cells with Different Ozone Concentration Levels				
Hour	1	2	3	4	5
0	2	130	0	0	0
1	33	99	0	0	0
2	55	77	0	0	0
3	68	64	0	0	0
4	93	39	0	0	0
5	80	52	0	0	0
6	88	44	0	0	0
7	115	17	0	0	0
8	119	13	0	0	0
9	64	65	3	0	0
10	0	46	86	0	0
11	0	0	32	100	0
12	0	0	0	103	29
13	0	0	1	110	21
14	0	0	1	88	43
15	0	0	0	64	68
16	0	0	0	79	53
17	0	0	4	58	70
18	0	0	0	17	115
19	0	0	0	9	123
20	0	0	6	58	68
21	0	7	66	36	23
22	1	43	87	1	0
23	7	87	38	0	0
Total	725	783	324	723	613

Table 2 Frequency of Temperatures throughout the Day

Hour	Number of Grid Cells at a Given Temperature												
°C	15*	16	17	18	19	20	21	22	23	24	25	26	27
°F	59.0	60.8	62.6	64.4	66.2	68.0	69.8	71.6	73.4	75.2	77.0	78.8	80.6
0	0	0	0	72	60	0	0	0	0	0	0	0	0
1	0	0	1	111	20	0	0	0	0	0	0	0	0
2	0	0	6	126	0	0	0	0	0	0	0	0	0
3	0	3	121	8	0	0	0	0	0	0	0	0	0
4	0	60	72	0	0	0	0	0	0	0	0	0	0
5	0	111	21	0	0	0	0	0	0	0	0	0	0
6	51	81	0	0	0	0	0	0	0	0	0	0	0
7	7	109	16	0	0	0	0	0	0	0	0	0	0
8	0	0	0	94	38	0	0	0	0	0	0	0	0
9	0	0	0	0	4	128	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	132	0	0	0	0	0
11	0	0	0	0	0	0	0	0	108	24	0	0	0
12	0	0	0	0	0	0	0	0	0	0	132	0	0
13	0	0	0	0	0	0	0	0	0	0	132	0	0
14	0	0	0	0	0	0	0	0	0	0	0	107	25
15	0	0	0	0	0	0	0	0	0	0	0	132	0
16	0	0	0	0	0	0	0	0	0	0	0	132	0
17	0	0	0	0	0	0	0	0	0	0	16	116	0
18	0	0	0	0	0	0	0	0	0	47	85	0	0
19	0	0	0	0	0	16	116	0	0	0	0	0	0
20	0	0	0	0	2	130	0	0	0	0	0	0	0
21	0	0	0	74	58	0	0	0	0	0	0	0	0
22	0	0	0	132	0	0	0	0	0	0	0	0	0
23	0	56	76	0	0	0	0	0	0	0	0	0	0
Total	58	420	313	617	182	274	116	132	108	71	365	487	25

Table 3 Data Description for Ozone Analysis

Variable	Description	Min	Max	Mean	Std. Dev.
OZONE	Ozone concentration level (1 = 0 to 0.035 ppb; 2 = 0.035 to 0.04 ppb; 3 = 0.04 to 0.045 ppb, 4 = 0.045 to 0.05 ppb; 5 = above 0.05 ppb)	1.00	5.00	2.91	1.47
TEMP	Temperature (Centigrade)	15.35	27.05	21.01	3.73
PEAKTRAF	Total length of street centerline (km) × indicator for peak travel hour (7:00 to 10:00 and 16:00 to 19:00)	0.00	208.93	19.99	37.76
NONPTRAF	Total length of street centerline (km) × indicator for non-peak hour hour	0.00	208.93	39.99	45.32
WKDEV	Percentage of developed land (%) × indicator for work hours (8:00 to 17:00)	0.00	93.43	14.61	22.66
NWKDEV	Percentage of developed land (%) × indicator for non-work hours	0.00	93.43	20.46	24.48
DTVEG	Percentage of vegetation (%) × indicator for daytime (6:00 to 18:00)	0.00	98.59	33.81	35.69
NTVEG	Percentage of vegetation (%) × indicator for nighttime	0.00	98.59	28.61	35.02
UNDEV	Percentage of undeveloped land (%) - used as base case	0.00	30.92	2.51	5.52

values below 0.035 are assigned Level 1, values between 0.035 and 0.04 are Level 2, those between 0.04 and 0.45 are Level 3, those between 0.045 and 0.05 are Level 4, and those above 0.05 are categorized as Level 5.

Table 4 Estimation Results for Model of Ozone Concentration Levels

Variable	Mean	Std. Dev.	t-stat.
TEMP	3.164E-01	1.410E-02	22.44
PEAKTRAF	1.300E-03	1.900E-03	0.68
NONPTRAF	4.900E-03	1.900E-03	2.58
WKDEV	-7.390E-02	5.200E-03	-14.21
NWKDEV	-7.360E-02	4.800E-03	-15.33
DTVEG	-6.020E-02	3.100E-03	-19.42
NTVEG	-5.910E-02	2.700E-03	-21.89
λ	6.583E-01	1.230E-02	53.52
ρ	-2.700E-03	1.874E-01	-0.01
σ^2	9.550E-02	2.480E-02	3.85
γ_1	-1.219E+00	8.710E-02	-13.99
γ_2	9.792E-01	5.800E-02	16.88
γ_3	2.462E+00	7.690E-02	32.02
γ_4	4.770E+00	1.040E-01	45.86

Table 5 Marginal Effects of Covariates on Ozone Concentration Levels

	Marginal Effect (10^{-2})				
Variable	Level 1	Level 2	Level 3	Level 4	Level 5
TEMP	-3.806	1.175	0.562	-1.653	3.723
PEAKTRAF	-0.016	0.005	0.002	-0.007	0.016
NONPTRAF	-0.061	0.019	0.009	-0.026	0.059
WDEV	0.886	-0.274	-0.131	0.385	-0.867
NWDEV	0.885	-0.273	-0.131	0.384	-0.865
DLVEG	0.722	-0.223	-0.107	0.314	-0.707
NTVEG	0.710	-0.219	-0.105	0.308	-0.694

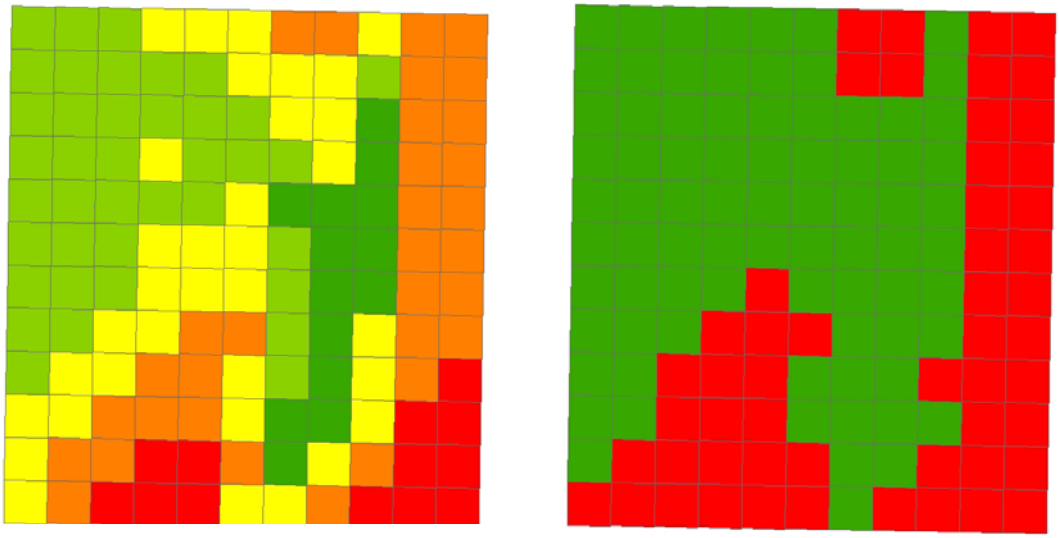
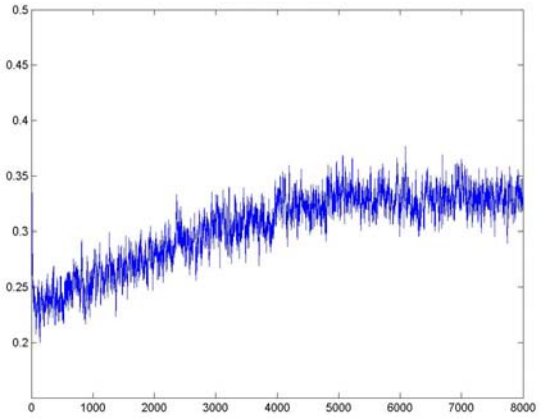
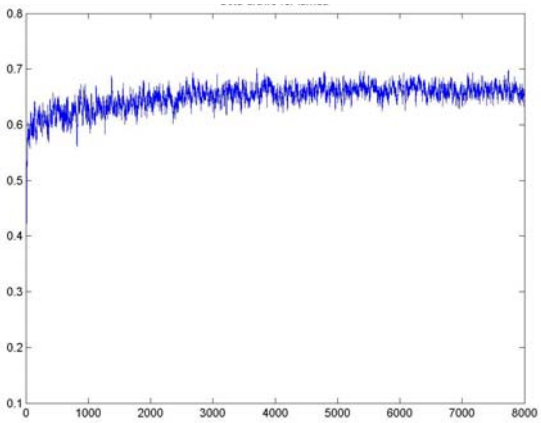


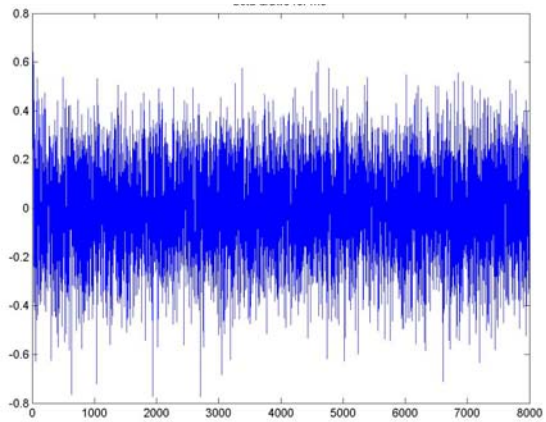
Figure 1 Ozone Concentration Values and Corresponding Levels (4 to 5pm on Monday, September 13, 1999)



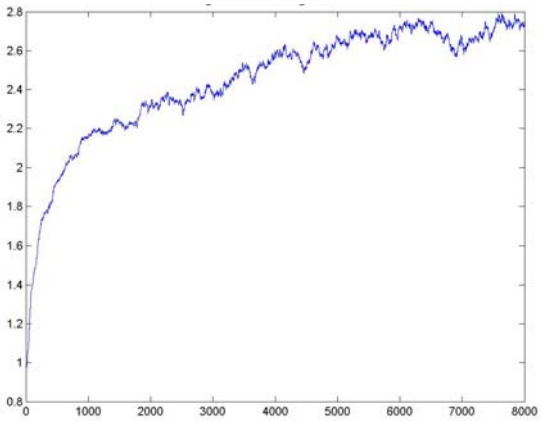
(a) Trace of β_{TEMP}



(b) Trace of λ



(c) Trace of ρ



(d) Trace of γ_3

Figure 2 Convergence Patterns of Ozone Concentration Level Parameter Estimation

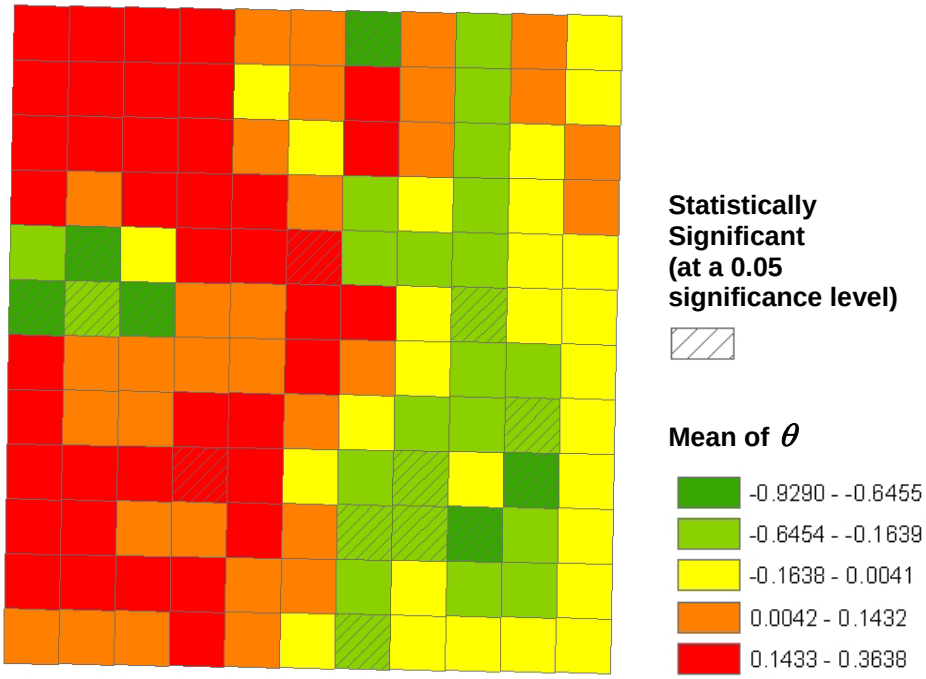
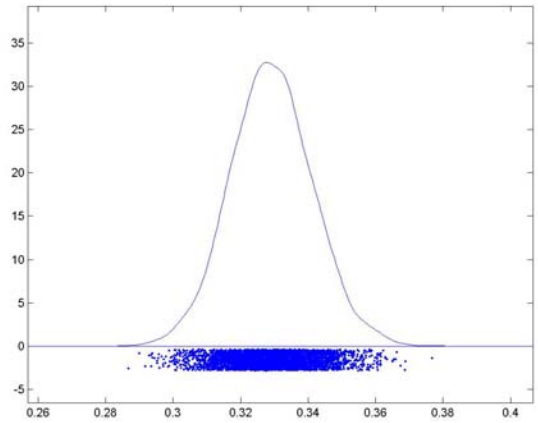
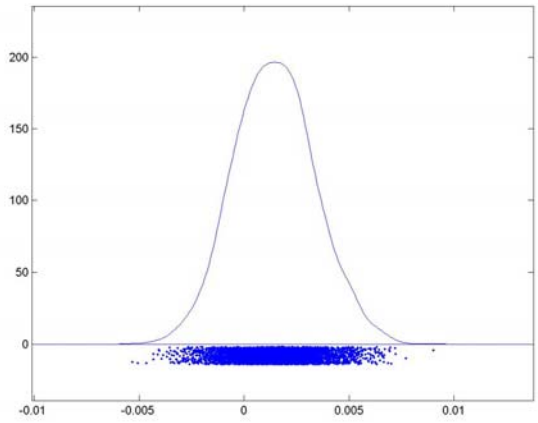


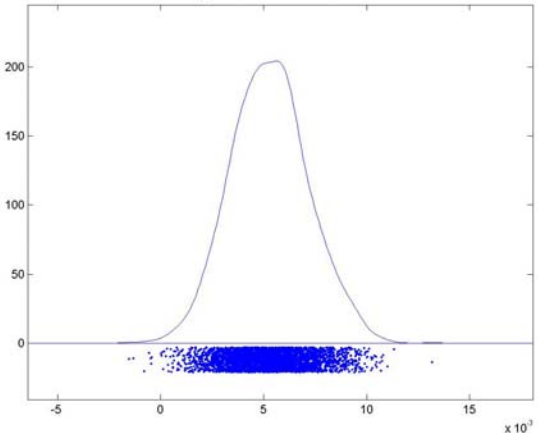
Figure 3 Distribution of Regional-Specific Error Term Estimates (θ) for Ozone Concentration Levels



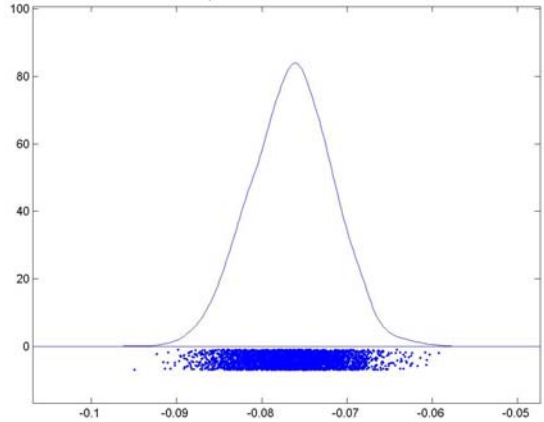
(a) Posterior Distribution of β_{TEMP}



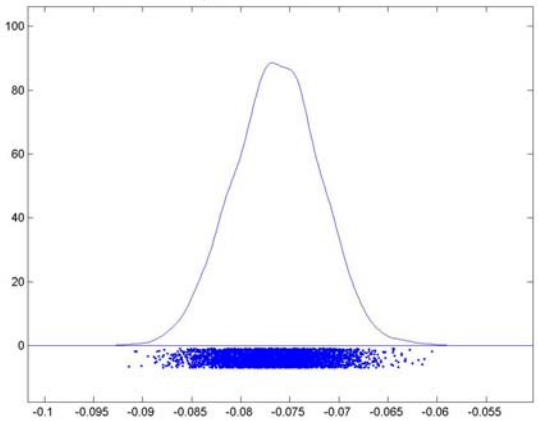
(b) Posterior Distribution of $\beta_{PEAKTRAF}$



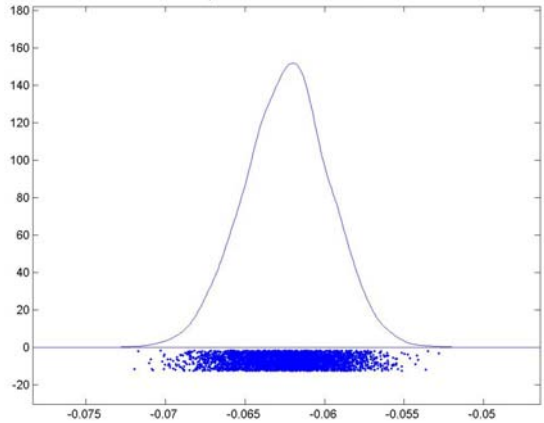
(c) Posterior Distribution of $\beta_{NONPTRAF}$



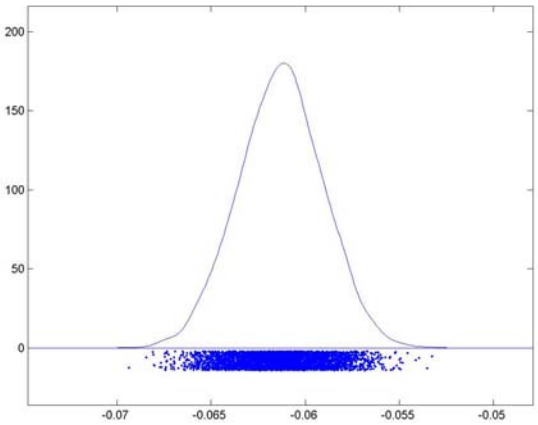
(d) Posterior Distribution of β_{WKDEV}



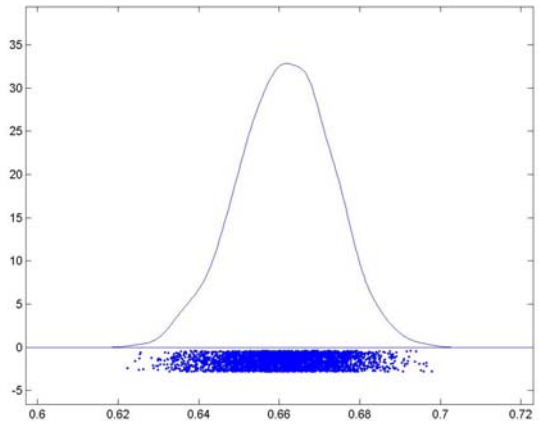
(e) Posterior Distribution of β_{NWKDEV}



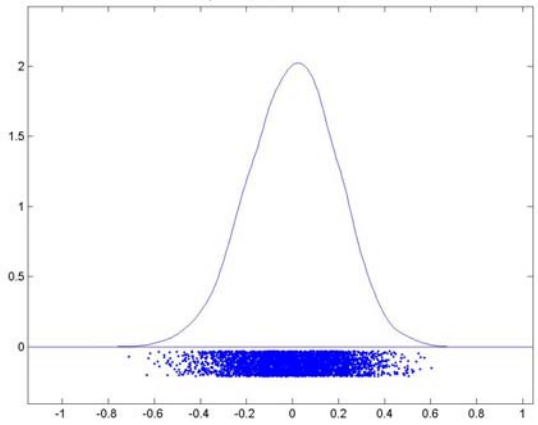
(f) Posterior Distribution of β_{DTVEG}



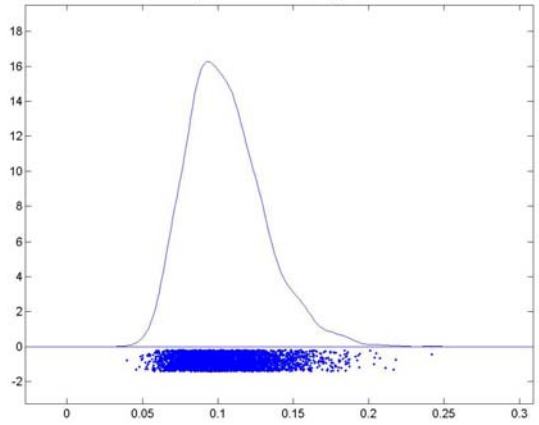
(g) Posterior Distribution of β_{NTVEG}



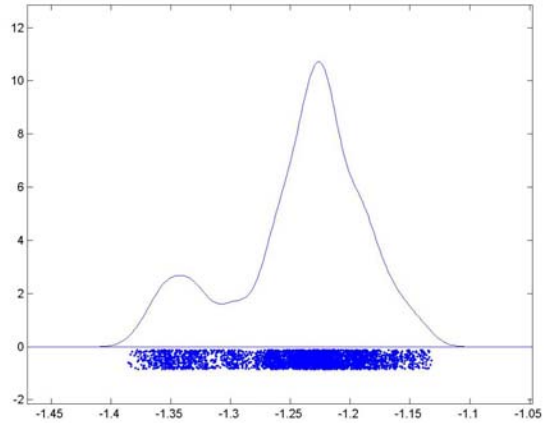
(h) Posterior Distribution of λ



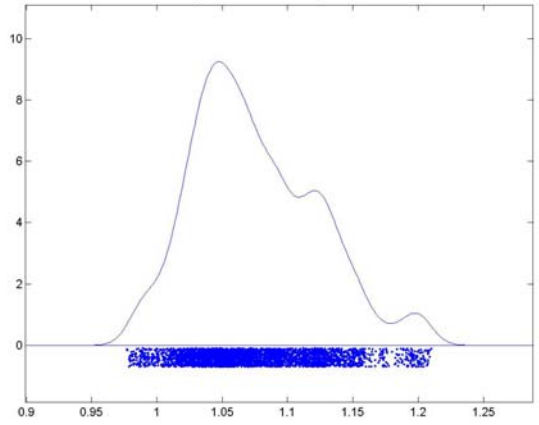
(i) Posterior Distribution of ρ



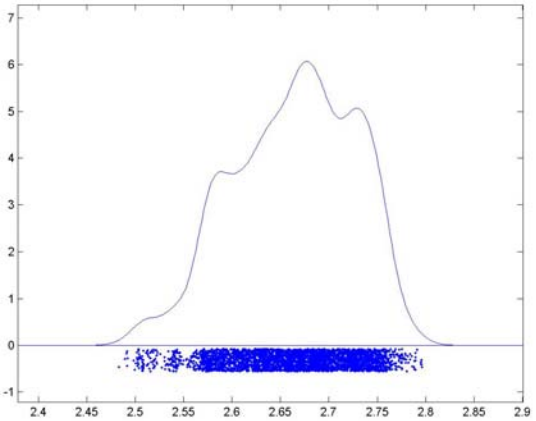
(j) Posterior Distribution of σ^2



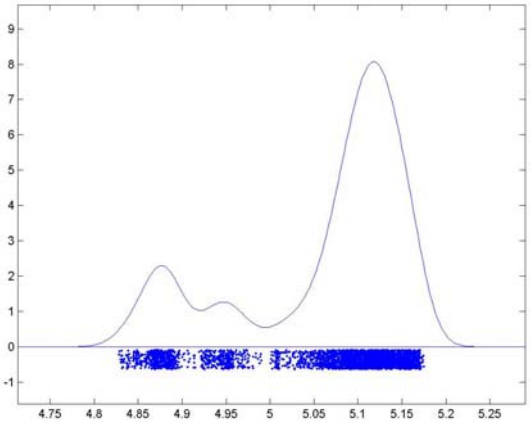
(k) Posterior Distribution of γ_1



(l) Posterior Distribution of γ_2

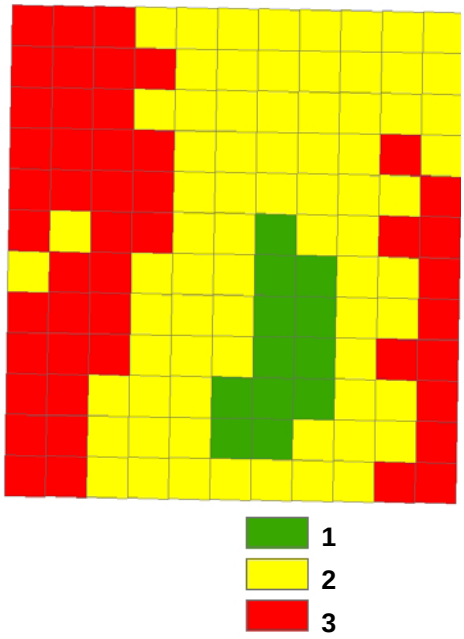


(m) Posterior Distribution of γ_3

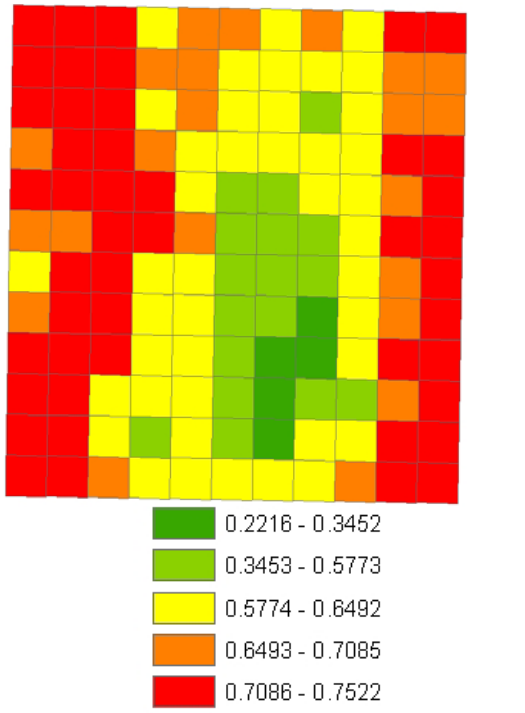


(n) Posterior Distribution of γ_4

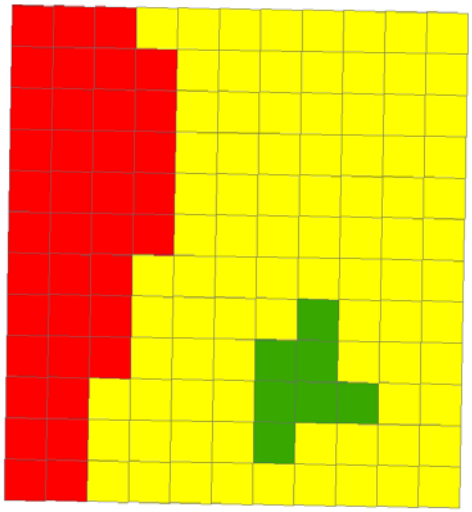
Figure 4 Posterior Distributions of Ozone Concentration Level Model Parameters



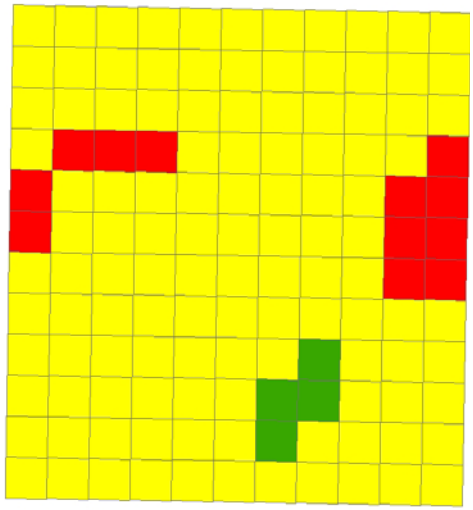
(a) Most Likely Ozone Concentration Levels



(b) Prediction Uncertainty (Entropy)



(c) Ozone Concentration Levels on Sept. 13, 1999 (11pm to midnight)



(d) Ozone Concentration Levels on Sept. 14, 1999 (midnight to 1am)

Figure 5 Prediction and Comparison of Ozone Concentration Levels

(Note: No Levels 4 and 5 exist at these points in time.)